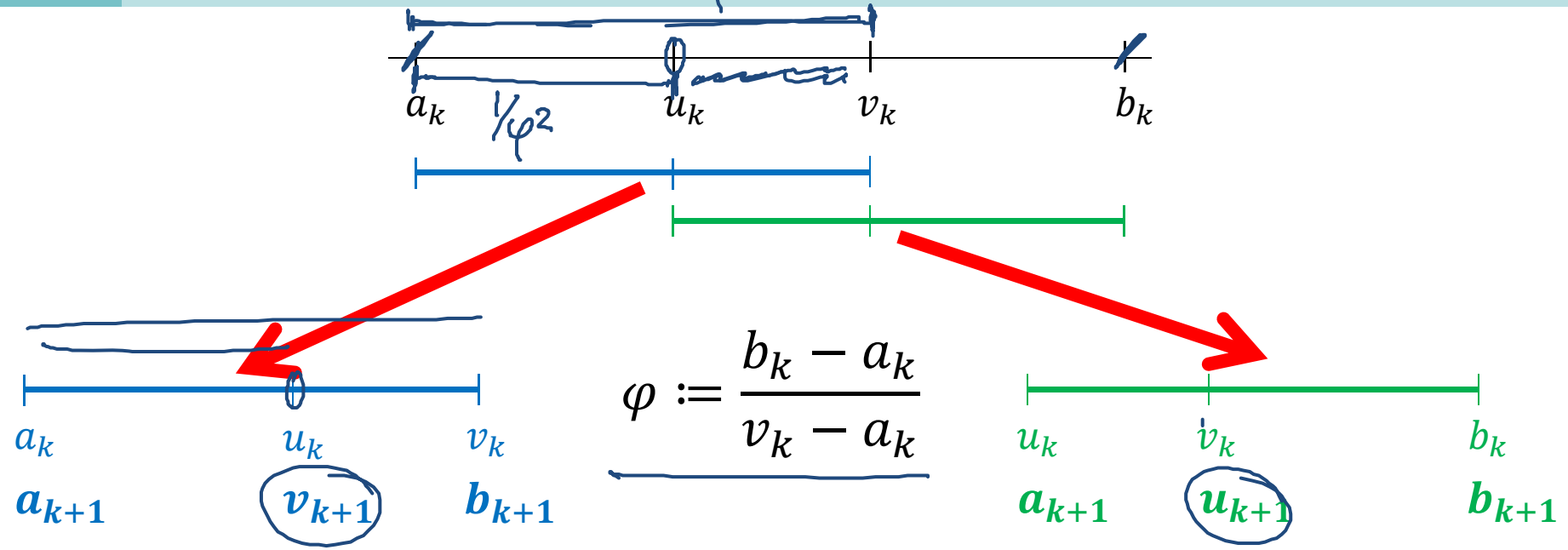
Optimierung

Vorlesung 3

Optimierung ohne Nebenbedingungen
Newton- und Quasi-Newton-Verfahren

Golden Ratio Search



$$\varphi := \frac{b_k - a_k}{v_k - a_k}$$

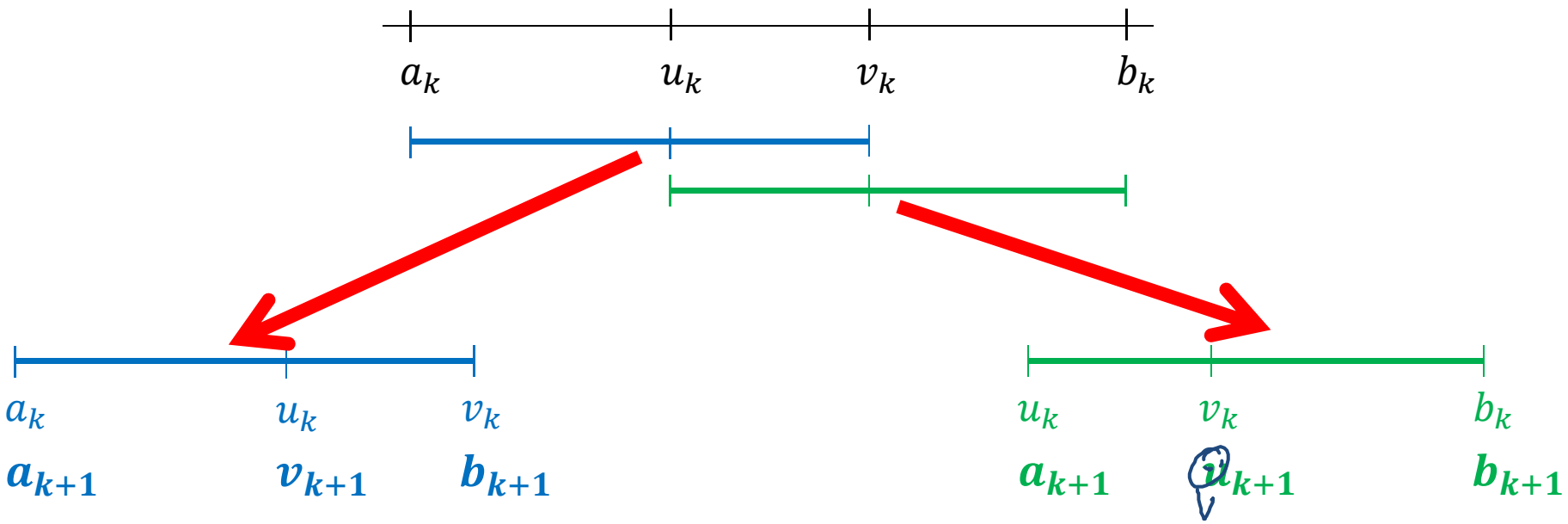
$$\frac{v_k - a_k}{u_k - a_k} = \frac{b_{k+1} - a_{k+1}}{v_{k+1} - a_{k+1}} = \varphi \Rightarrow \frac{b_k - a_k}{u_k - a_k} = \varphi^2$$

$$\rightarrow \frac{b_k - u_k}{v_k - u_k} = \frac{b_{k+1} - a_{k+1}}{u_{k+1} - a_{k+1}} = \varphi^2$$

$$b_k - u_k = (b_k - a_k)\left(1 - \frac{1}{\varphi^2}\right), \qquad v_k - u_k = (b_k - a_k)\left(\frac{1}{\varphi} - \frac{1}{\varphi^2}\right)$$

$$\varphi^2 = \frac{b_k - u_k}{v_k - u_k} = \frac{1^2 - 1/\varphi^2}{1/\varphi - 1/\varphi^2} = \frac{(1 + 1/\varphi)(1 - 1/\varphi)}{1/\varphi\,(1 - 1/\varphi)} = \varphi + 1$$

Golden Ratio Search



$$\frac{b_k - a_k}{v_k - a_k} = \varphi \quad \Rightarrow \quad \frac{b_k - a_k}{u_k - a_k} = \varphi^2 \quad \Rightarrow \quad \varphi + 1 = \varphi^2$$

$$\boldsymbol{\varphi + 1 = \varphi^2 \ (\varphi > 1)} \quad \Rightarrow \quad \boldsymbol{\varphi = \frac{1 + \sqrt{5}}{2} \approx 1.618}$$

Newton-Verfahren

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

$f: \mathbb{R}^n \to \mathbb{R}$

- Finde stationären Punkt x^*: $\nabla f(x^*) = 0$

Iteratives Verfahren

$x_{k+1} = x_k + d_k$

- Berechne Sequenz $x_0, x_1, x_2, \ldots,$
- Für $i \to \infty$, x_i sollte zu stationärem Punkt konvergieren

Taylor-Approximation 1. Ordnung von ∇f

$f'(x+c) \approx f'(x) + c f''(x)$

$$\nabla f(x_k + d_k) \approx \nabla f(x_k) + H(x_k) \cdot d_k$$

– Hesse'sche Matrix $H(x_k) = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(x_k) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x_k) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x_k) & \cdots & \frac{\partial^2}{\partial x_n^2} f(x_k) \end{pmatrix}$

– Taylor (1. Ordn.): $\nabla f(x_k + d_k) = \begin{pmatrix} \vdots \\ \frac{\partial}{\partial x_i} f(x_k + d_k) \\ \vdots \end{pmatrix} \approx \begin{pmatrix} \vdots \\ \frac{\partial}{\partial x_i} f(x_k) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ d_k^\top \cdot \nabla \left(\frac{\partial}{\partial x_i} f(x_k) \right) \\ \vdots \end{pmatrix}$

Newton-Verfahren

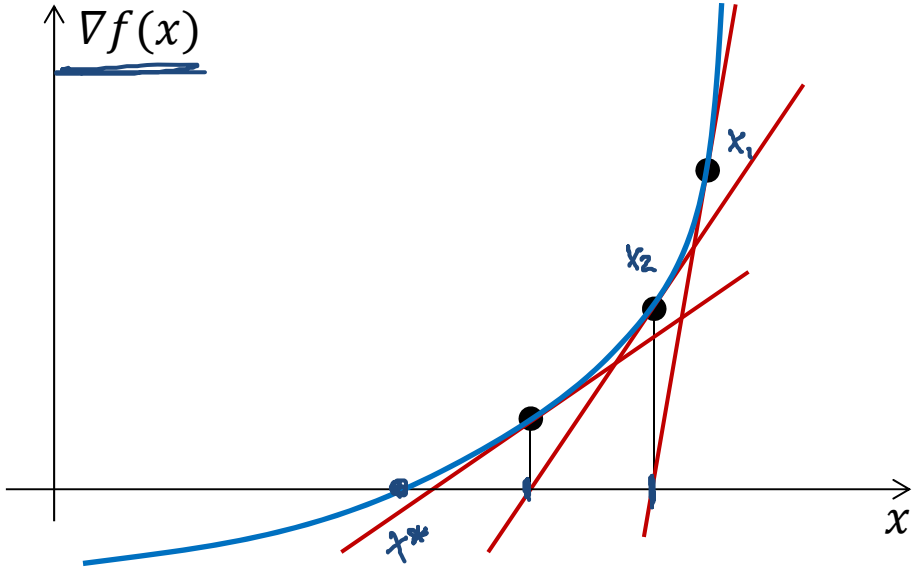
Taylor-Approximation 1. Ordnung von ∇f

$$\nabla f(x_k + d_k) \approx \nabla f(x_k) + H(x_k) \cdot d_k$$

Newton-Verfahren

$$0 \stackrel{!}{=} \nabla f(x_k + d_k) \approx \nabla f(x_k) + H(x_k) \cdot d_k$$

- Setze $\nabla f(x_k) + H(x_k) d_k = 0$ und berechne $\boldsymbol{x_{k+1} = x_k + d_k}$
- Löse lineares Gleichungssystem für d_k: $\boldsymbol{d_k = -H(x_k)^{-1} \cdot \nabla f(x_k)}$



Newton-Verfahren (alternative Sicht)

Taylor-Approximation 2. Ordnung von f

$$f(x_k + d_k) \approx f(x_k) + d_k^\top \cdot \nabla f(x_k) + \frac{1}{2} d_k^\top \cdot H(x_k) \cdot d_k$$

- $H(x_k)$: Hesse'sche Matrix an der Stelle x_k

Newton-Verfahren

- Neuer Wert x_{k+1}: minimiere 2. Ordnung Taylor-Approximation

$$\boldsymbol{d_k = \arg\min_{d \in \mathbb{R}^n} \left\{ f(x_k) + d^\top \nabla f(x_k) + \frac{1}{2} d^\top H(x_k) d \right\}}$$

$$\boldsymbol{x_{k+1} = x_k + d_k}$$

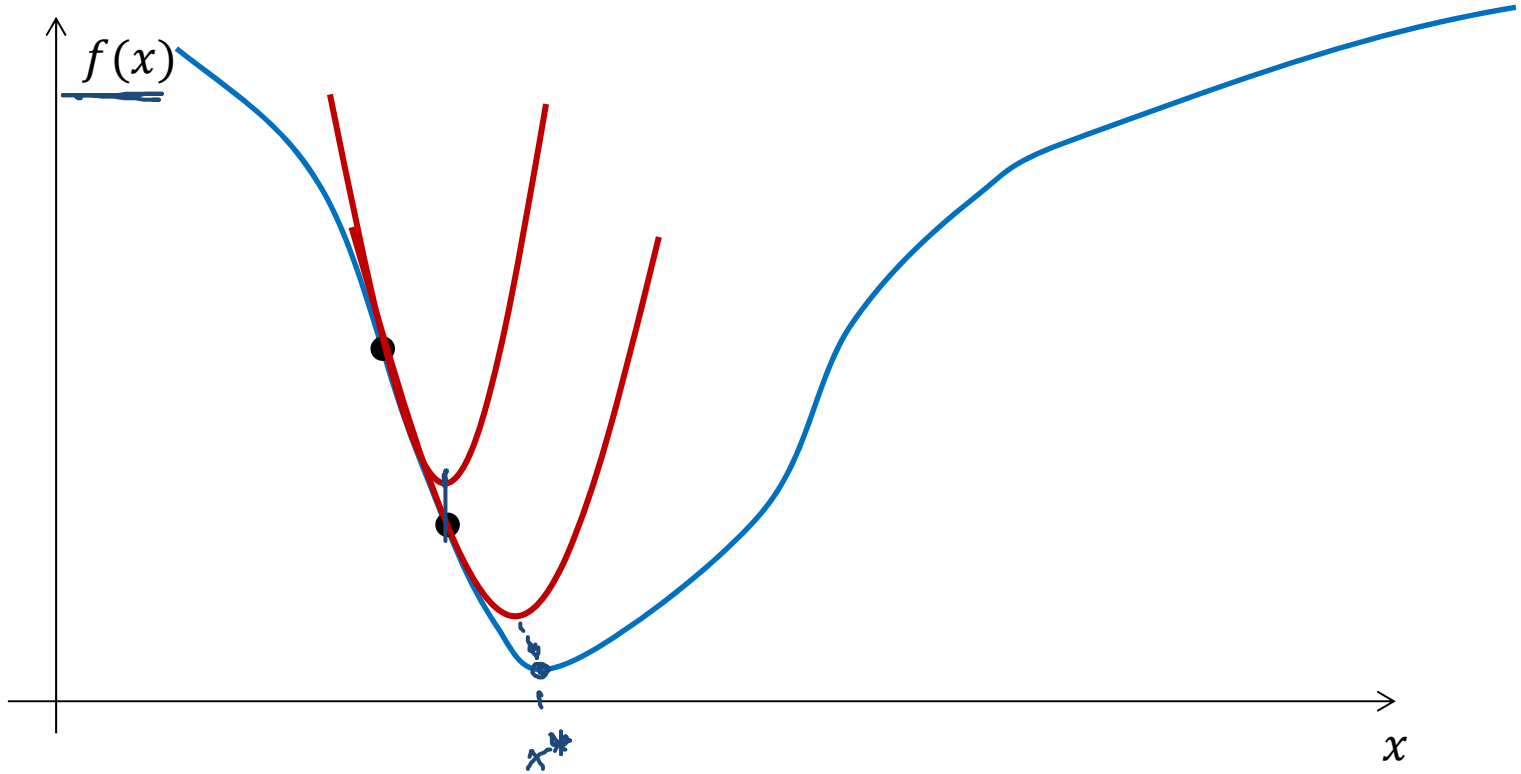
Newton-Verfahren (alternative Sicht)

Taylor-Approximation 2. Ordnung von f

$$f(x_k + d_k) \approx f(x_k) + d_k^\top \cdot \nabla f(x_k) + \frac{1}{2} d_k^\top \cdot H(x_k) \cdot d_k$$

Newton-Verfahren

- Neuer Wert x_{k+1}: minimiere 2. Ordnung Taylor-Approximation



Newton: Beispiel 1

$$f(x) = 5x - \ln x$$

- Gradient: $\nabla f(x) = f'(x) = 5 - {}^1/_x\,, \qquad f'(x^*) = 0 \implies x^* = {}^1/_5$
- Hesse'sche Matrix: $H(x) = f''(x) = {}^1/_{x^2}$
- Newton-Richtung: $d_k = -H(x_k)^{-1}\nabla f(x_k) = -\frac{f'(x_k)}{f''(x_k)} = x_k - 5x_k^2$

$$x_{k+1} = x_k + d_k = 2x_k - 5x_k^2$$

k	x_k	x_k
1	0.05	0.3
2	0.087500000000000	0.15
3	0.136718750000000	0.1875
4	0.179977416992188	0.19921875
5	0.197995480848476	0.199996948242188
6	0.199979909514856	0.199999999953434
7	0.199999997981862	0.200000000000000

Newton: Beispiel 2

$\ln(2x) = \ln(x) + \ln 2$

$$f(x,y) = \ln(1 + x + y) + \ln 2x + \ln 5y$$

- $\nabla f(x,y) = \begin{pmatrix} \frac{1}{1+x+y} + \frac{1}{x} \\ \frac{1}{1+x+y} + \frac{1}{y} \end{pmatrix}$, $\nabla f(x^*, y^*) = 0 \implies x^* = -\frac{1}{3}, y^* = -\frac{1}{3}$
- $H(x,y) = -\begin{pmatrix} \left(\frac{1}{1+x+y}\right)^2 + \frac{1}{x^2} & \left(\frac{1}{1+x+y}\right)^2 \\ \left(\frac{1}{1+x+y}\right)^2 & \left(\frac{1}{1+x+y}\right)^2 + \frac{1}{y^2} \end{pmatrix}$
- Newton-Richtung: $-H(x,y)d_k = \nabla f(x,y)$ ← Newton-Schritt
- Startwerte: $x_1 = -0.05$, $x_2 = -0.75$

$$\begin{pmatrix} 425 & 25 \\ 25 & 25 + {}^{16}/_{9} \end{pmatrix} \cdot \begin{pmatrix} d_{k,x} \\ d_{k,y} \end{pmatrix} = \begin{pmatrix} -15 \\ 5 - {}^{4}/_{3} \end{pmatrix} \implies d_k = \begin{pmatrix} -0.0458677686 \\ 0.1797520661 \end{pmatrix}$$

$x_{k+1} = x_k + d_k$

Fortschritt beim Newton-Verfahren

Abstiegsrichtung

- Eine Richtung d ist eine Abstiegsrichtung für einen Wert x_k, falls

$$d^\top \nabla f(x_k) < 0$$

  - Beispiel: Neg. Gradient $d = -\nabla f(x_k)$

$$-\nabla f^\top \nabla f < 0$$

- Taylor-Approximation: $f(x_k + \varepsilon \cdot d) \approx f(x_k) + \varepsilon \cdot d^\top \nabla f(x_k) < f(x_k)$

Newton-Richtung $\boldsymbol{d_k = -H(x_k)^{-1} \nabla f(x_k)}$

$$\forall y: y^\top H(x) y > 0$$

- Falls $H(x_k)$ positiv definit ist, ist d_k eine Abstiegsrichtung

$$d_k^\top \nabla f(x_k) < 0$$

$$(AB)^\top = B^\top A^\top$$

$$H^{-1\top} = H^{-1}$$

$$(-H^{-1}\nabla f)^\top \nabla f = -\nabla f^\top H^{-1} \nabla f$$

$$= -\nabla f^\top H^{-1} H H^{-1} \nabla f$$

$$= -(H^{-1}\nabla f)^\top H (H^{-1}\nabla f)$$

$$= -v^\top H v < 0$$

$v^\top H v > 0$

Fortschritt beim Newton-Verfahren

Konvexe Funktionen $f(x)$:

- Newton-Richtung ist eine Abstiegsrichtung
- Newton-Verfahren konvergiert zu lokalem Minimum

Konvexe, quadratische Funktionen $f(x)$:

- Newton-Schritt geht zu Minimum der quadratischen Approx. von $f(x)$
- Verfahren erreicht lokales Minimum in einem Schritt!

Allgemein:

- Konvergiert sobald man *nahe genug* bei einem lokalen Minimum ist
- Kann mit ***line search*** (wie beim Gradientenverfahren) kombiniert werden:

Newton-Richtung: $d_k = -H(x_k)^{-1} \cdot \nabla f(x_k)$

Bestimme $x_{k+1} = x_k + \tau_k \cdot d_k$, so dass $f(x_k + \tau_k d_k)$ minimal ist

Konvergenz

- Wie schnell kommt man zum lokalen Minimum?

Konvergenz von Folgen $s_1, s_2, \ldots,$ $\lim_{k\to\infty} s_k = \bar{s}$

Lineare Konvergenz:

$$\limsup_{k\to\infty} \frac{|s_{k+1} - \bar{s}|}{|s_k - \bar{s}|} = C < 1$$

$s_k = 0.9^k$

Superlineare Konvergenz:

$$\limsup_{k\to\infty} \frac{|s_{k+1} - \bar{s}|}{|s_k - \bar{s}|} = 0$$

$s_k = 1/k!$

Quadratische Konvergenz:

$$\limsup_{k\to\infty} \frac{|s_{k+1} - \bar{s}|}{|s_k - \bar{s}|^2} = C$$

$s_k = 0.9^{2^k}$

Konvergenz der wichtigsten Verfahren

Gradientenverfahren (steilster Abstieg)

- lineare Konvergenz

Newtonverfahren

- quadratische Konvergenz
- minimiert quadratische Funktionen in 1 Schritt

Konjugierte Gradienten

- minimiert n-dimensionale quadratische Funktionen in n Schritten
- quadratische Konvergenz alle n Schritte (superlinear)

Quasi-Newtonverfahren

- superlineare Konvergenz
- einfacher als Newton-Verfahren

Lineare Konvergenz des Gradientenverfahrens

Annahme: streng konvexe, quadratische Fkt. $f(x) = \frac{1}{2} x^\top A x - b^\top x$ (pos. def.)

- A symm. pos. definit ($\forall x: x^\top A x > 0$, alle Eigenwerte reell und positiv)
- Eindeutiges Minimum bei x^*, so dass $\nabla f(x^*) = 0$

$$0 = \nabla f(x^*) = Ax^* - b \implies x^* = A^{-1} b$$

- Funktionswert bei x^*: $f(x^*) = \frac{1}{2} b^\top A^{-1} A A^{-1} b - b^\top A^{-1} b = -\frac{1}{2} b^\top A^{-1} b$

1 Iterationsschritt an Stelle x

- Abstiegsrichtung $d = -\nabla f(x) = b - Ax$
- Neuer Iterationswert $x' = x + \tau \cdot d$, wobei $\tau = \arg\min_\alpha f(x + \alpha \cdot d)$

$$f(x+\alpha d) = \frac{1}{2}(x+\alpha d)^\top A (x+\alpha d) - b^\top (x+\alpha d)$$

$$= \frac{1}{2}\left(x^\top A x + \alpha d^\top A x + \alpha x^\top A d + \alpha^2 d^\top A d\right) - b^\top x - \alpha b^\top d \qquad (\alpha d^\top b)$$

$$= f(x) + \alpha d^\top (Ax - b) + \frac{1}{2} d^\top A d \cdot \alpha^2 = -\alpha d^\top d + \frac{1}{2}\alpha^2 d^\top A d + f(x) \qquad (Ax - b = -d)$$

$$\frac{\partial}{\partial \alpha} f(x+\alpha d) = \alpha\, d^\top A d - d^\top d = 0 \implies \tau = \frac{d^\top d}{d^\top A d}$$

Lineare Konvergenz des Gradientenverfahrens

Quadratische konvexe Funktion

$$f(x) = \frac{1}{2} x^\top A x - b^\top x, \qquad x^* = A^{-1} b, \qquad f(x^*) = -\frac{1}{2} b^\top A^{-1} b$$

1 Iterationsschritt an Stelle x (neuer Wert x')

$$d = -\nabla f(x) = b - Ax, \qquad x' = x + \tau \cdot d, \qquad \tau = \frac{d^\top d}{d^\top A d}$$

Konvergenzrate C

$$C = \max_{x \text{ nahe genug bei } x^*} \frac{f(x') - f(x^*)}{f(x) - f(x^*)}$$

$$f(x') = f(x + \tau d) = f(x) + \frac{1}{2}\tau^2 d^\top A d - \tau d^\top d$$

$$= f(x) + \frac{1}{2}\frac{(d^\top d)^2}{d^\top A d} - \frac{(d^\top d)^2}{d^\top A d}$$

$$= f(x) - \frac{1}{2}\frac{(d^\top d)^2}{d^\top A d}$$

Lineare Konvergenz des Gradientenverfahrens

Quadratische konvexe Funktion

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x, \qquad x^* = A^{-1}b, \qquad f(x^*) = -\frac{1}{2}b^\top A^{-1}b$$

$$d = b - Ax, \qquad x' = x + \tau d, \qquad \tau = \frac{d^\top d}{d^\top Ad}, \qquad f(x') = f(x) - \frac{1}{2}\frac{(d^\top d)^2}{d^\top Ad}$$

Konvergenzrate C

$$C = \frac{f(x') - f(x^*)}{f(x) - f(x^*)} = \frac{f(x) - \frac{1}{2}\frac{(d^\top d)^2}{d^\top Ad} + \frac{1}{2}b^\top A^{-1}b}{f(x) + \frac{1}{2}b^\top A^{-1}b}$$

$$\frac{\alpha+\beta}{\alpha} = 1 + \frac{\beta}{\alpha}$$

$$= 1 - \frac{\frac{1}{2}\frac{(d^\top d)^2}{d^\top Ad}}{f(x) + \frac{1}{2}b^\top A^{-1}b} = 1 - \frac{\frac{1}{2}\frac{(d^\top d)^2}{d^\top Ad}}{\frac{1}{2}x^\top Ax - b^\top x + \frac{1}{2}b^\top A^{-1}b}$$

$$\frac{1}{2}x^\top AA^{-1}Ax = \frac{1}{2}(Ax)^\top A^{-1}(Ax)$$

$$\frac{1}{2}(Ax)^\top A^{-1}(Ax) - b^\top x + \frac{1}{2}b^\top A^{-1}b = \frac{1}{2}(Ax-b)^\top A^{-1}(Ax-b) = \frac{1}{2}d^\top A^{-1}d$$

$$\frac{1}{2}x^\top A A^{-1}(-b)$$

Lineare Konvergenz des Gradientenverfahrens

Konvergenzrate C

$$C = \frac{f(x') - f(x^*)}{f(x) - f(x)} = 1 - \frac{\frac{1}{2}\frac{(d^\top d)^2}{d^\top A d}}{\frac{1}{2} d^\top A^{-1} d} = 1 - \frac{d^\top d}{d^\top A d} \cdot \frac{d^\top d}{d^\top A^{-1} d}$$

- Eigenwerte von A : $\mathbf{0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n}$ (positiv und reell)
- Eigenwerte von A^{-1}: $0 < {}^1\!/_{\lambda_n} \leq {}^1\!/_{\lambda_{n-1}} \leq \cdots \leq {}^1\!/_{\lambda_1}$
- Grösster Eigenwert $\mu_{\max}$ einer reellen, symm. $n \times n$-Matrix M:

$$\mu_{\max} = \max_{x \in \mathbb{R}^n} \frac{x^\top M x}{x^\top x}$$

Rayleigh-Koeffizient

- Deshalb: $\frac{d^\top A d}{d^\top d} \leq \lambda_n$, $\frac{d^\top A^{-1} d}{d^\top d} \leq {}^1\!/_{\lambda_1}$ $\Longrightarrow$ $C \leq 1 - {}^{\lambda_1}\!/_{\lambda_n}$

Konditionszahl von A

- Genauer: $C \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2$

Quadratische Konvergenz des Newton-Verfahren

Ziel: Die Sequenz $x_1, x_2, \ldots$ konvergiert quadratisch gegen x^*

Matrix-Norm $\|\boldsymbol{M}\|$ einer Matrix $\boldsymbol{M}$

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

- $\forall M, M', x\colon\ \|Mx\| \leq \|M\| \cdot \|x\|, \quad \|M \cdot M'\| \leq \|M\| \cdot \|M'\|$

Annahmen: Für x, y nahe genug bei x^* gilt $\|\boldsymbol{H}(\boldsymbol{x}) - \boldsymbol{H}(\boldsymbol{y})\| \leq \boldsymbol{L} \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$,
sowie $\left\|\boldsymbol{H}(\boldsymbol{x})^{-\boldsymbol{1}}\right\| \leq \boldsymbol{C}$ für Konstanten $L, C \geq 0$

Lemma:

$$\nabla f(z) - \nabla f(x) = \int_0^1 \left[H\big(x + t(z-x)\big)\right](z-x)dt$$

Quadratische Konvergenz des Newton-Verfahren

Lemma:

$$\nabla f(z) - \nabla f(x) = \int_0^1 \left[H\big(x + t(z-x)\big)\right](z-x)dt$$

Beweis:

- Definiere $\phi(t) \coloneqq \nabla f\big(x + t(z-x)\big)$
- Ableitung $\phi'(t) = \left[H\big(x + t(z-x)\big)\right](z-x)$

Quadratische Konvergenz des Newton-Verfahren

Ziel: Die Sequenz $x_1, x_2, \ldots$ konvergiert quadratisch gegen x^*

Lemma:

$$\nabla f(z) - \nabla f(x) = \int_0^1 \left[H\big(x + t(z-x)\big)\right](z-x)dt$$

■

$$x' - x^* = x - H(x)^{-1}\nabla f(x) - x^*$$

Quadratische Konvergenz des Newton-Verfahren

Matrix-Norm $\|\boldsymbol{M}\|$ einer Matrix $\boldsymbol{M}$

- $\forall M, M', x\colon\ \|Mx\| \leq \|M\| \cdot \|x\|, \quad \|M \cdot M'\| \leq \|M\| \cdot \|M'\|$

Annahmen: Für x, y nahe genug bei x^* gilt $\|\boldsymbol{H}(\boldsymbol{x}) - \boldsymbol{H}(\boldsymbol{y})\| \leq \boldsymbol{L} \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$,
sowie $\left\|\boldsymbol{H}(\boldsymbol{x})^{-\boldsymbol{1}}\right\| \leq \boldsymbol{C}$ für Konstanten $L, C \geq 0$

$$x' - x^* = H(x)^{-1} \int_0^1 \left[H\big(x + t(x^* - x)\big) - H(x)\right](x^* - x)dt$$

$$\|x' - x^*\| \leq \|H(x)^{-1}\| \int_0^1 \left\|H\big(x + t(x^* - x)\big) - H(x)\right\| \cdot \|x^* - x\,\|dt$$

Gradientenverfahren vs. Newton-Verfahren

Gradientenverfahren

- Lineare Konvergenz
  - Benötigt i.A. deutlich mehr Schritte (vor allem, wenn die Hesse'sche Matrix schlecht konditioniert ist)
- 1. Ableitung nötig
- 1 Iterationsschritt:
  - Evaluation Gradient + line search
- "Billige" Schritte, "schlechte" Konv.
- Bessere Konvergenz:

  Konjugierte Gradientenverfahren

Newton-Verfahren

- Quadratische Konvergenz
  - konvergiert in wenigen Schritten (bei quadr. Problem in 1 Schritt)
- 1. & 2. Ableitung nötig
- 1 Iterationsschritt:
  - Evaluation Gradient + Hesse'sche Matrix
  - n-dim. lineares Gl.-system
- "Teure" Schritte, "gute" Konvergenz
- Einfachere Iterationen:

  Quasi-Newton-Verfahren

Quasi-Newton-Verfahren

- Newton-Verfahren + Line Search:

$$x_{k+1} = x_k + \tau_k d_k, \qquad \text{wobei } d_k = -H(x_k)^{-1}\nabla f(x_k)$$

- **Idee:** Ersetze $H(x_k)^{-1}$ durch eine Matrix D_k, welche einfacher berechnet werden kann

- Taylor-Approximation: $\nabla f(x_{k+1}) - \nabla f(x_k) \approx H(x_{k+1})(x_{k+1} - x_k)$

  - Idee: D_k^{-1} sollte dies auch erfüllen...

- Quasi-Newton Bedingung:

$$\boldsymbol{D_{k+1} \cdot q_k = p_k}, \qquad \text{wobei } \boldsymbol{p_k = x_{k+1} - x_k},\ \boldsymbol{q_k = \nabla f(x_{k+1}) - \nabla f(x_k)}$$

Quasi-Newton-Verfahren

Quasi-Newton-Verfahren:

$$\boldsymbol{x_{k+1} = x_k + \tau_k \cdot d_k,} \qquad \textbf{wobei} \quad \boldsymbol{d_k = -D_k \nabla f(x_k)}$$

- τ_k wird typischerweise durch line search ermittelt

Quasi-Newton Bedingung:

$$D_{k+1} \cdot q_k = p_k, \qquad \text{wobei} \quad p_k = x_{k+1} - x_k, \;\; q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

Berechnung von $\boldsymbol{D_{k+1}}$:

- Alte Matrix + Korrekturterm: $D_{k+1} = D_k + C_k$

$$(D_k + C_k)q_k = p_k \quad \Longrightarrow \quad C_k q_k = p_k - D_k q_k$$

- Zum Beispiel ($\phi \in [0,1]$ ist ein Parameter)

$$C(\phi) = \frac{pp^\top}{p^\top q} - \frac{Dqq^\top D}{q^\top Dq} + \phi \eta v v^\top, \qquad \text{wobei } v = \frac{p}{p^\top q} - \frac{Dq}{\eta}, \eta = q^\top Dq$$

Quasi-Newton-Verfahren

Quasi-Newton-Verfahren:

$$x_{k+1} = x_k - \tau_k \cdot D_k \nabla f(x_k)$$

$$D_{k+1} = D_k + C_k, \qquad C_k q_k = p_k - D_k q_k$$

- Parameter $\phi = 0$ (Davidson-Fletcher-Powell, erstes Quasi-Newton-Verf.)

$$C(0) = \frac{pp^\top}{p^\top q} - \frac{Dqq^\top D}{q^\top Dq}$$

- Parameter $\phi = 1$ (Broyden-Fletcher-Goldfarb-Shanno)

$$C(1) = \frac{pp^\top}{p^\top q}\left[1 + \frac{q^\top Dq}{p^\top q}\right] - \frac{Dqp^\top + pq^\top D}{p^\top q}$$

- Quasi-Newton-Verfahren konvergieren oft superlinear
- Für quadratische Funktionen in n Schritten
  - entspricht dann dem konjugierten Gradientenverfahren