Inhaltsübersicht für heute:

Freie Nichtlineare Optimierung

Orakel, lineares/quadratisches Modell

Optimalitätsbedingungen

Das Newton-Verfahren

Line-Search-Verfahren

Inhaltsübersicht für heute:

Freie Nichtlineare Optimierung

Orakel, lineares/quadratisches Modell

Optimalitätsbedingungen

Das Newton-Verfahren

Line-Search-Verfahren

Freie Nichtlineare Optimierung

Verfahren zur Minimierung glatter Funktionen ohne Nebenbedingungen,

$$\min_{x \in \mathbb{R}^n} f(x), \quad f : \mathbb{R}^n \to \mathbb{R} \quad \text{hinreichend glatt}$$

Hinreichend glatt bedeutet, dass f so oft stetig differenzierbar sein soll, wie für das jeweilige Verfahren erforderlich.

Ziele für die Verfahren:

- Finde ein lokales Minimum (sogar weniger: finde ein $\bar{x}$, das die notwendigen Opt.-Bed. 1. Ordnung erfüllt, s. dort)
- Schnelle Konvergenz in der Nähe lokaler Optima
- Der Rechenaufwand soll möglichst klein bleiben
- Numerische Stabilität und hohe Genauigkeit

Anwendungen:

- Nichltineare kleinste Quadrate Probleme
- Als Löser für Optimierungsprobleme mit Nebenbedingungen [s. Barriere-, Straf- und augmentierte Lagrange-Verfahren]

In welcher Form soll f für die Verfahren zugänglich sein?

Inhaltsübersicht für heute:

Orakel allgemein und Orakel 0. Ordnung

In vielen Anwendungen ist die Funktion f nicht analytisch verfügbar, sondern ergibt sich z.B. aus der Lösung eines Systems partieller Differentialgleichungen oder einer Simulation.

Daher setzen allgemeine Optimierungsverfahren nur eine Unterroutine voraus, die das Verfahren nach dem Wert der Funktion und eventuell auch nach Ableitungsinformation in dem jeweils betrachteten Punkt befragen kann $\rightarrow$ „Orakel“.

Ist die Funktion analytisch gegeben, erzeugen Modellierungssprachen wie AMPL, GAMS, ... automatisch entsprechende Orakel/Unterroutinen, die Wert und Ableitungsinformation liefern.

Orakel allgemein und Orakel 0. Ordnung

In vielen Anwendungen ist die Funktion f nicht analytisch verfügbar, sondern ergibt sich z.B. aus der Lösung eines Systems partieller Differentialgleichungen oder einer Simulation.

Daher setzen allgemeine Optimierungsverfahren nur eine Unterroutine voraus, die das Verfahren nach dem Wert der Funktion und eventuell auch nach Ableitungsinformation in dem jeweils betrachteten Punkt befragen kann $\rightarrow$ „Orakel“.

Ist die Funktion analytisch gegeben, erzeugen Modellierungssprachen wie AMPL, GAMS, ... automatisch entsprechende Orakel/Unterroutinen, die Wert und Ableitungsinformation liefern.

Ein **Orakel 0. Ordnung** berechnet für gegebenes $x \in \mathbb{R}^n$ nur den Funktionswert $f(x)$, aber keine Ableitungsinformation.
Verfahren für glatte Funktionen benötigen Ableitungsinformation und approximieren diese numerisch durch vielfache Funktionsaufrufe (s. später).

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

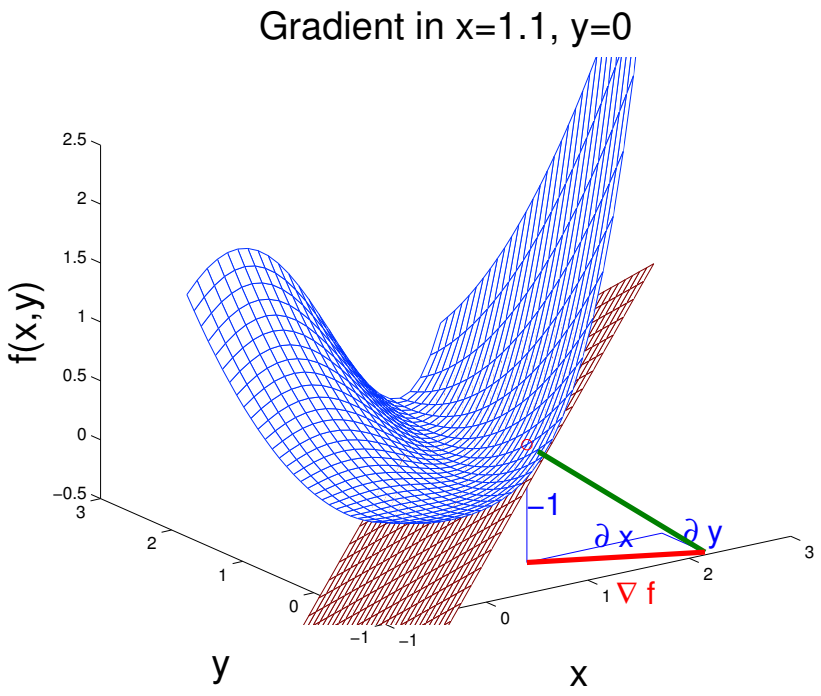
Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



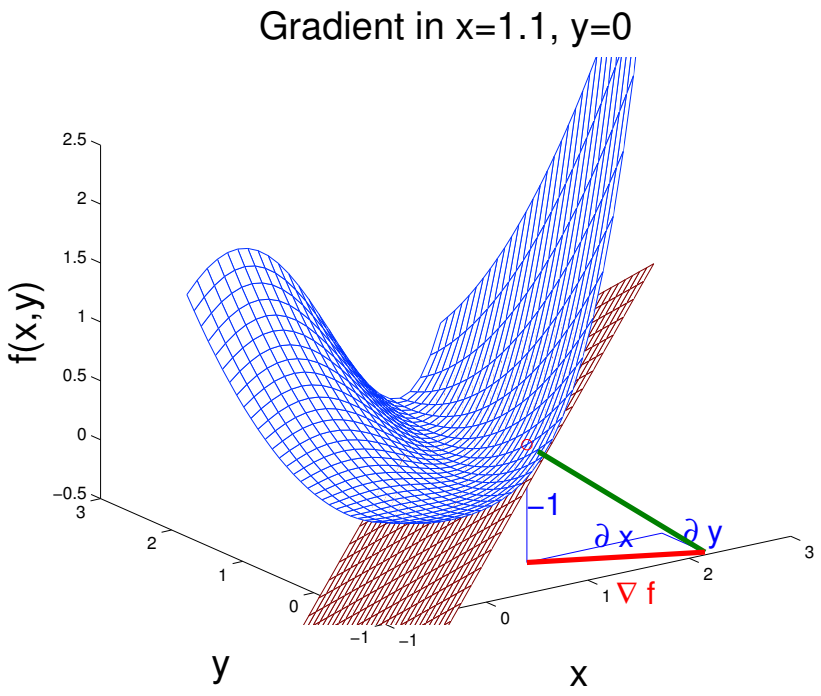
- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Die Tangentialebene bildet das lineare Modell von f um $\bar{x}$,

$$\bar{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine gute Näherung an f: für glattes f erfüllt ∇f

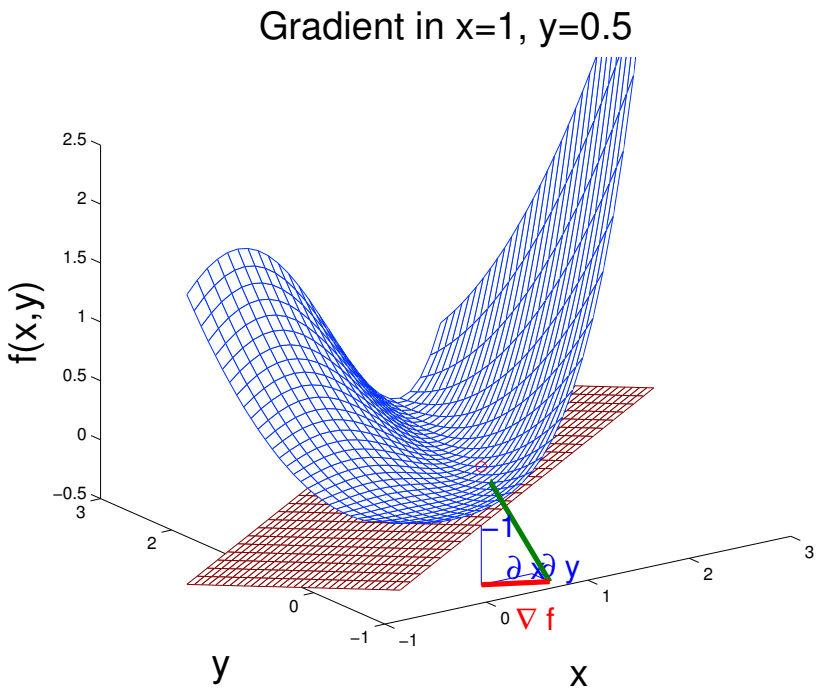
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x})}{\|x - \bar{x}\|} = \lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x})^T h}{\|h\|} = 0.$$

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Die Tangentialebene bildet das lineare Modell von f um $\bar{x}$,

$$\bar{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine gute Näherung an f: für glattes f erfüllt ∇f

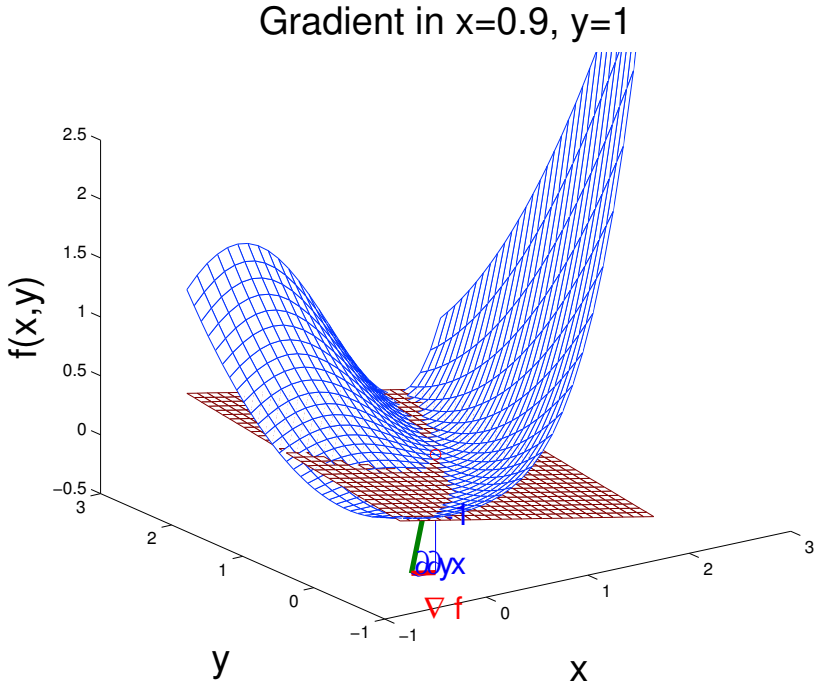
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x})}{\|x - \bar{x}\|} = \lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x})^T h}{\|h\|} = 0.$$

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Die Tangentialebene bildet das lineare Modell von f um $\bar{x}$,

$$\bar{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine gute Näherung an f: für glattes f erfüllt ∇f

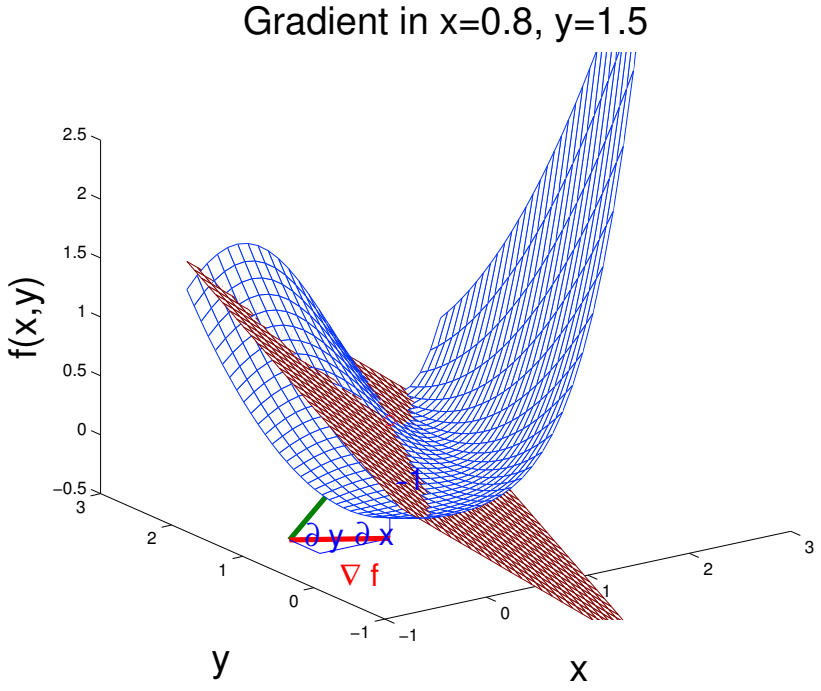
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x})}{\|x - \bar{x}\|} = \lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x})^T h}{\|h\|} = 0.$$

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Die Tangentialebene bildet das lineare Modell von f um $\bar{x}$,

$$\bar{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine gute Näherung an f: für glattes f erfüllt ∇f

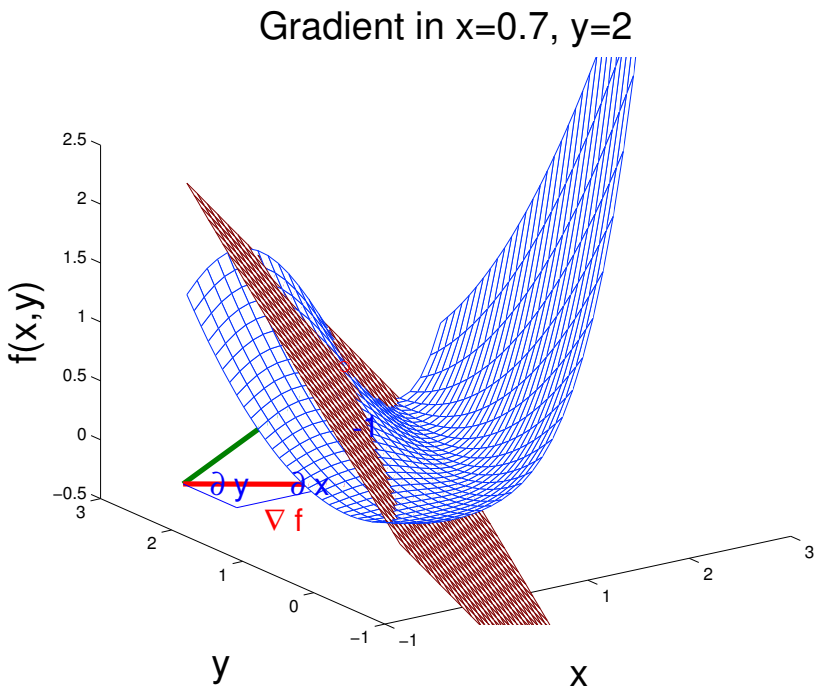
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x})}{\|x - \bar{x}\|} = \lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x})^T h}{\|h\|} = 0.$$

Orakel 1. Ordnung: $f(x)$, $\nabla f(x)$

Für $\bar{x} \in \mathbb{R}^n$ wird Funktionswert $f(\bar{x})$ und Gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ berechnet.

Der Gradient:

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$



- $\nabla f(x)$ zeigt in Richtung des steilsten Anstiegs von f in x.
- $\|\nabla f(x)\|$ misst die Größe des Anstiegs.
- $\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$ ist der Normalvektor zur Tangentialebene an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Die Tangentialebene bildet das lineare Modell von f um $\bar{x}$,

$$\bar{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine gute Näherung an f: für glattes f erfüllt ∇f

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x})}{\|x - \bar{x}\|} = \lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x})^T h}{\|h\|} = 0.$$

Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x})$$

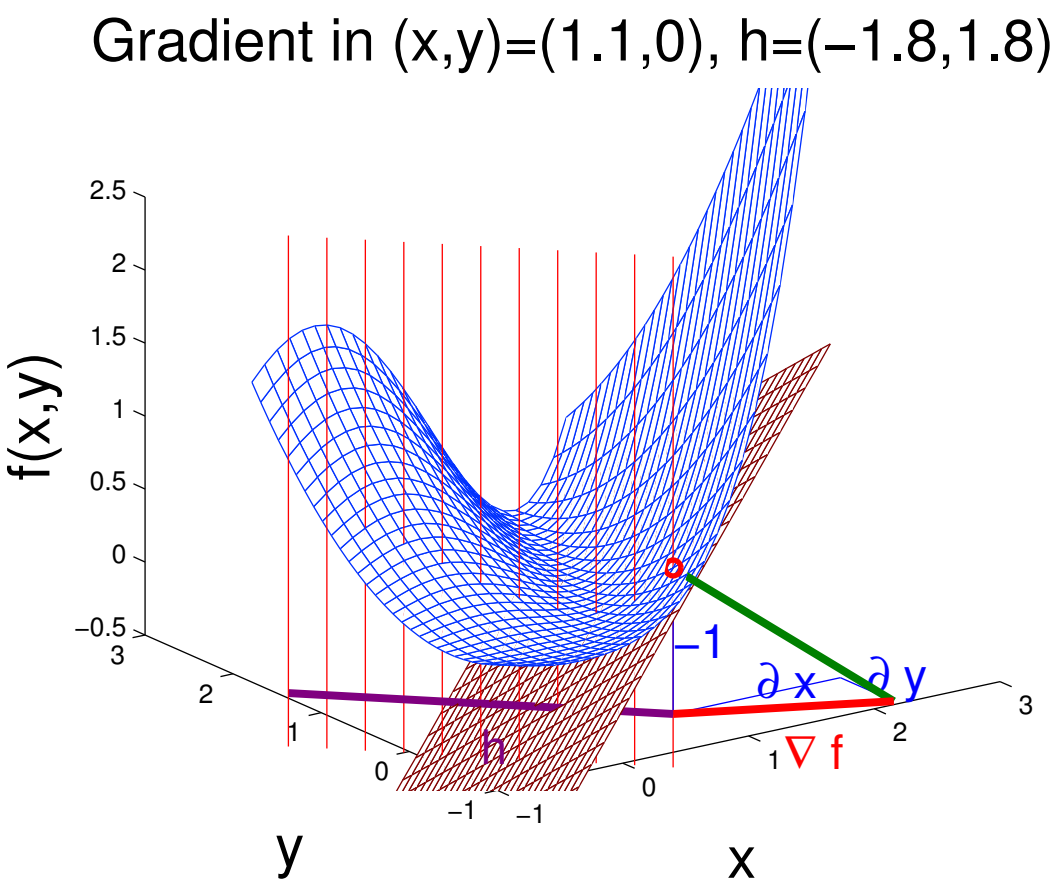
[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]

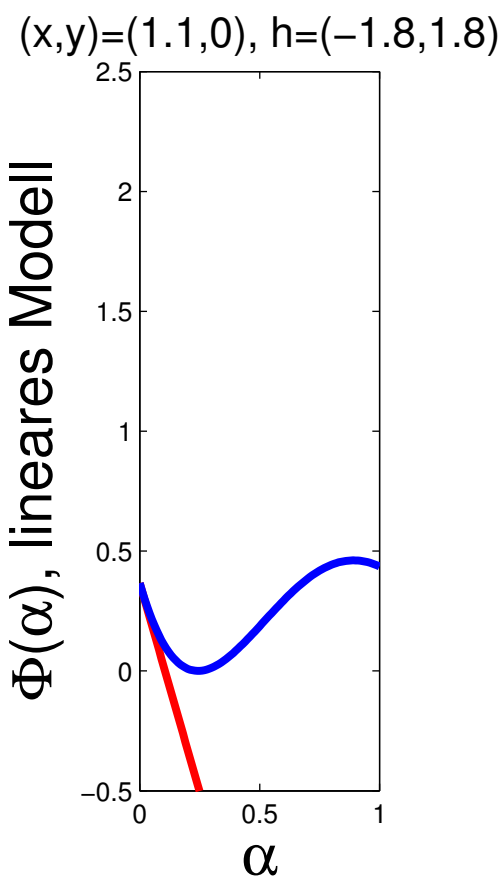
Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x})$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



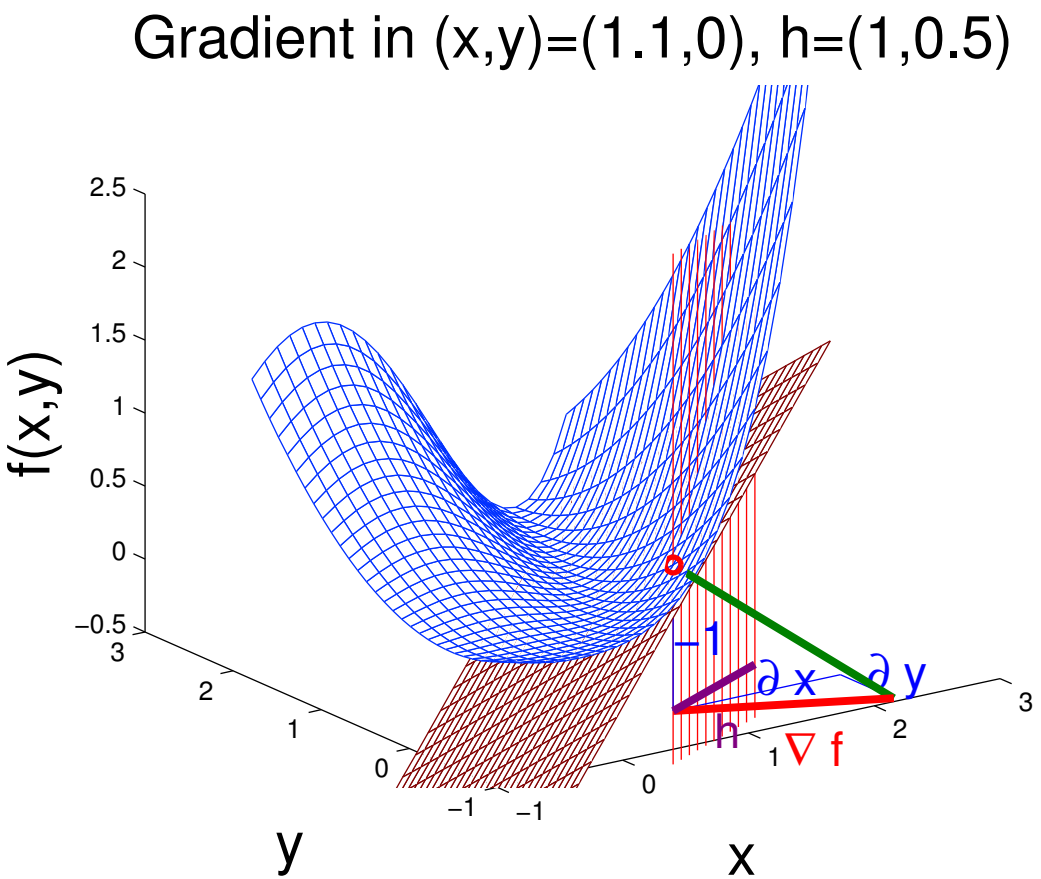


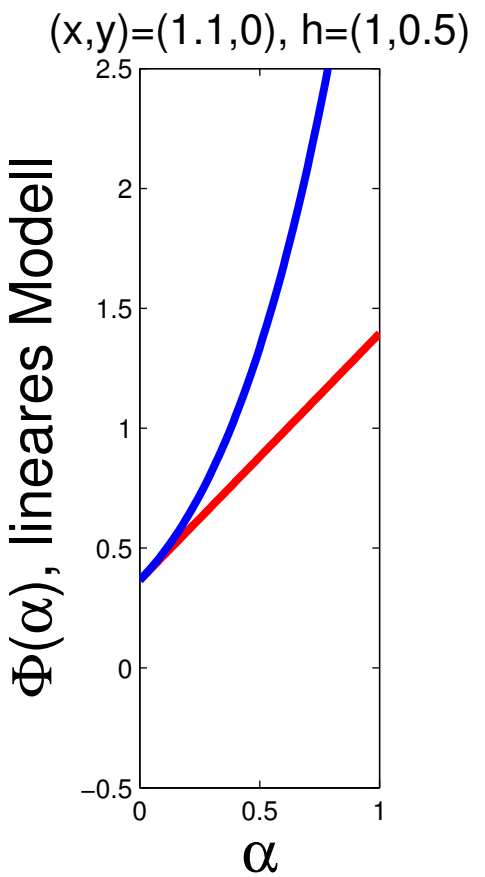
Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x})$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



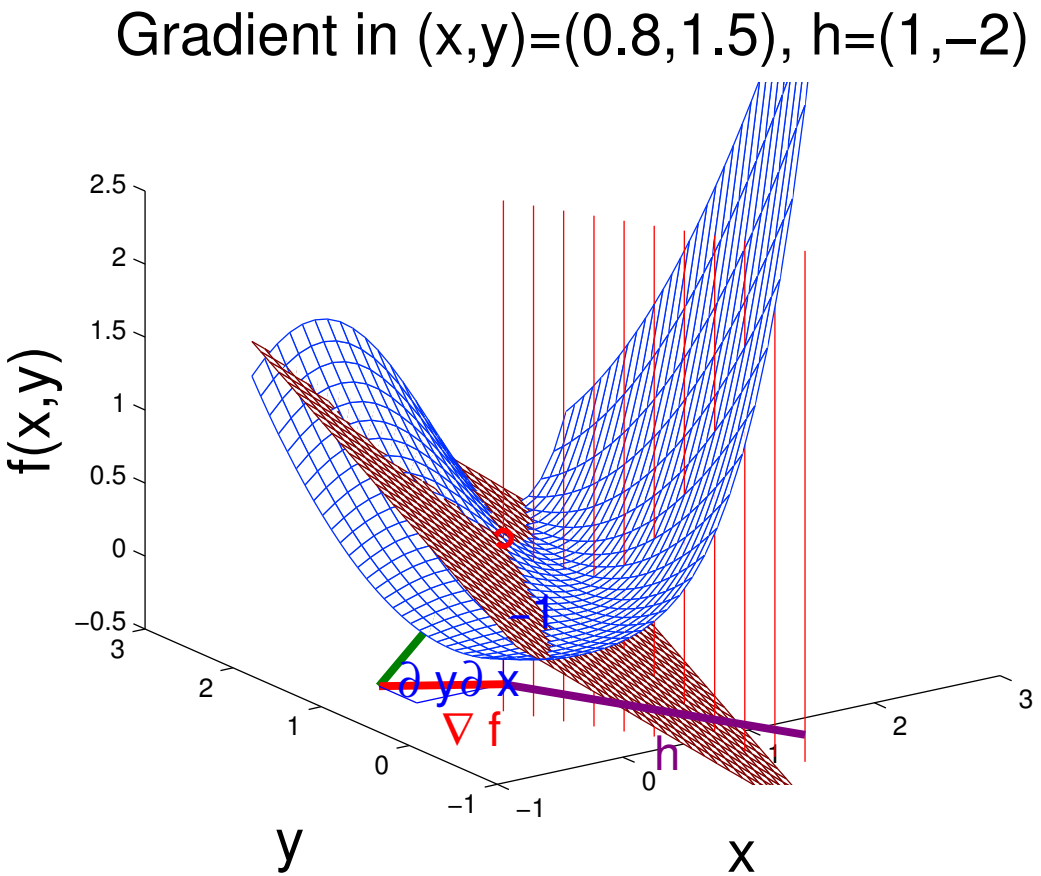


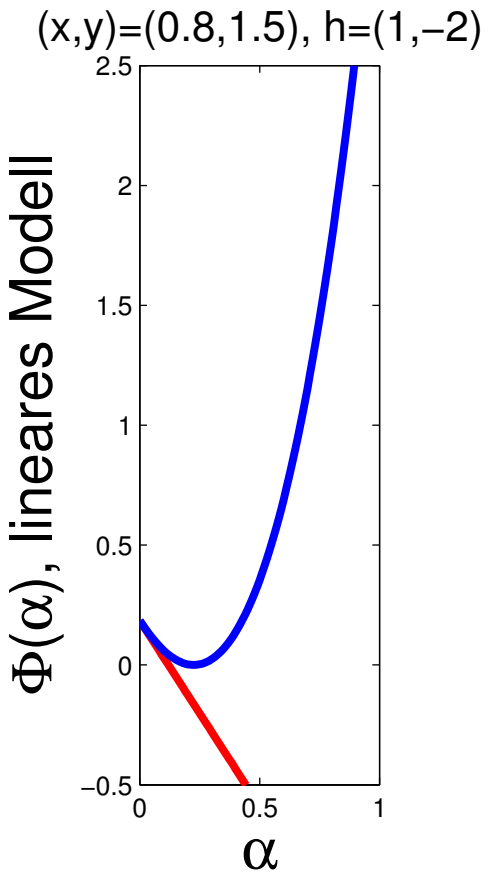
Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x})$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



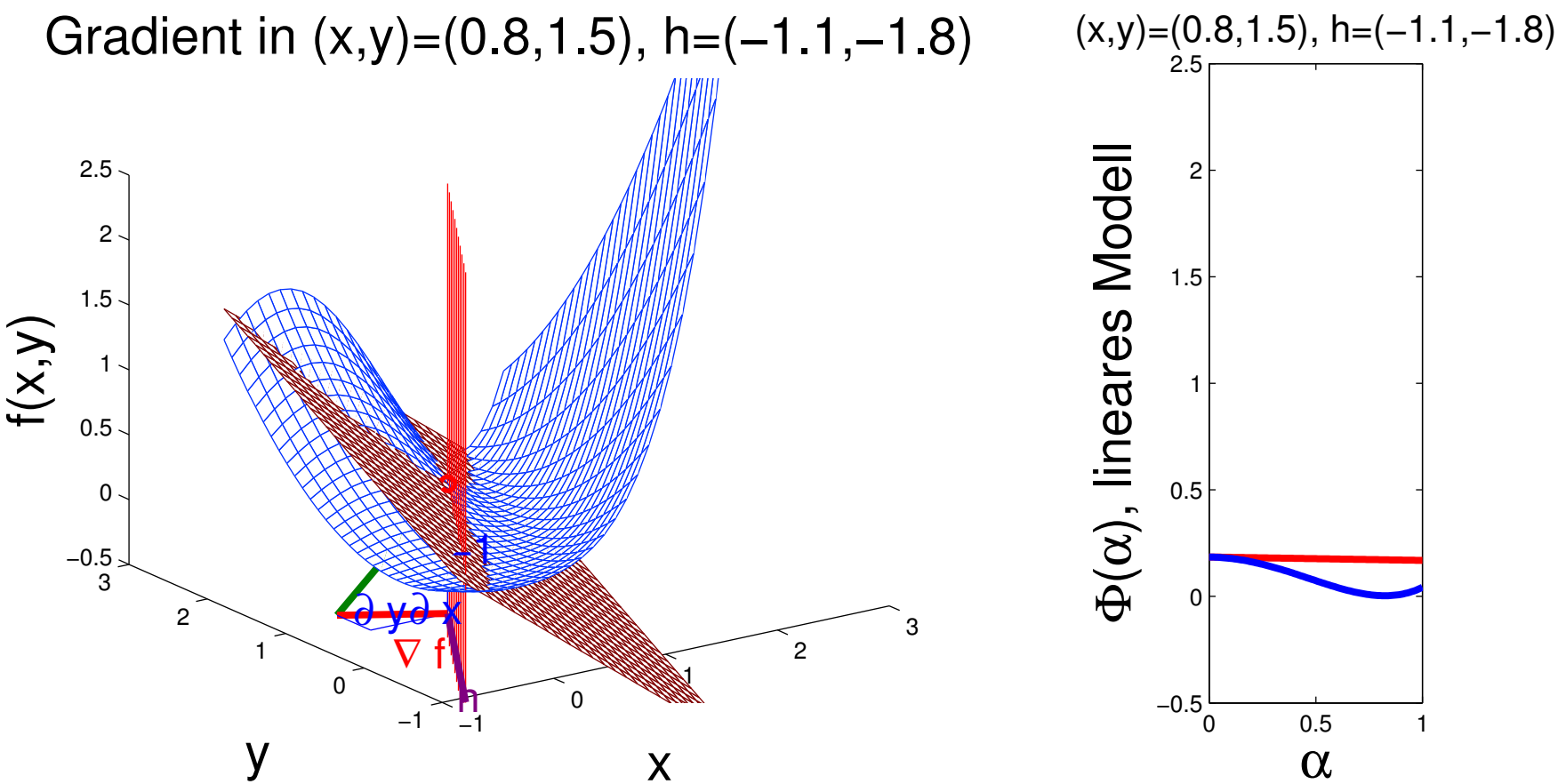


Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x})$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]

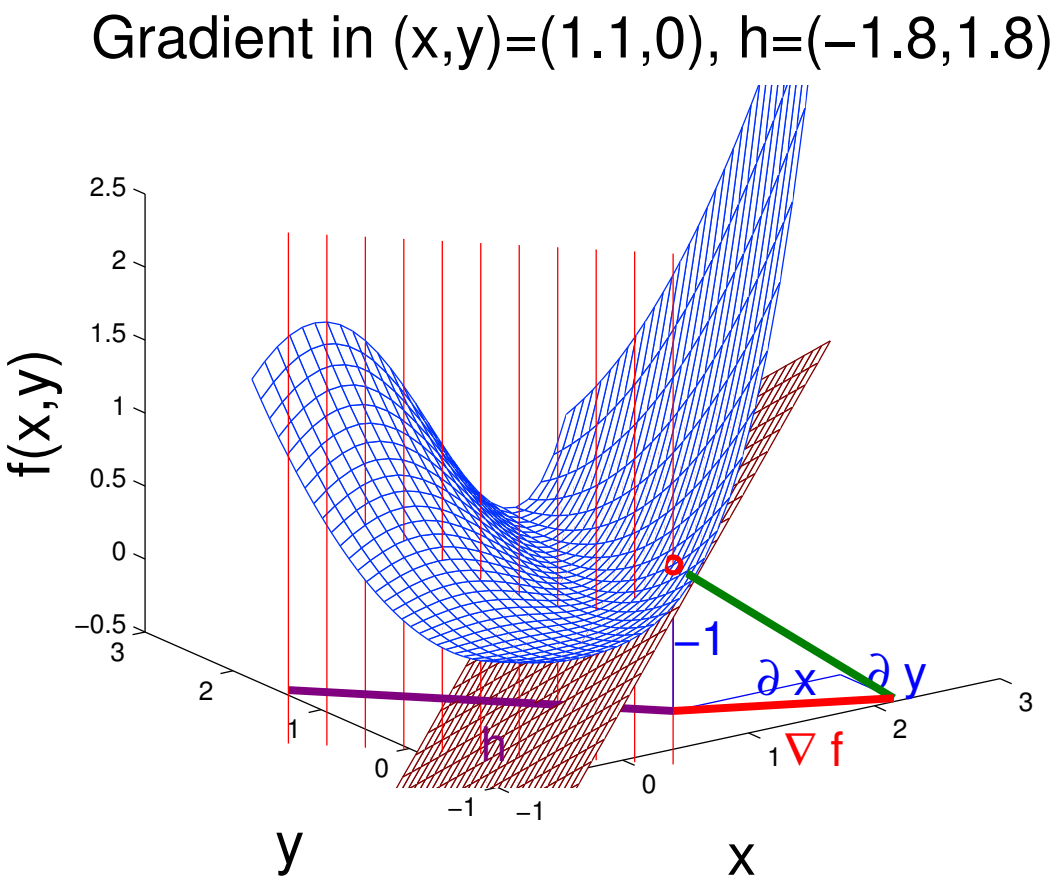


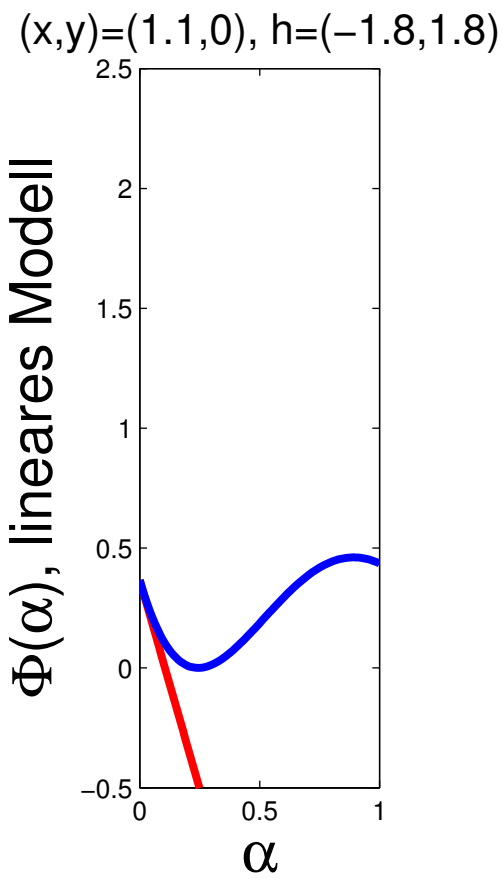
Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x}) \qquad [D_{\lambda h} f(\bar{x}) = \lambda D_h f(\bar{x})]$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



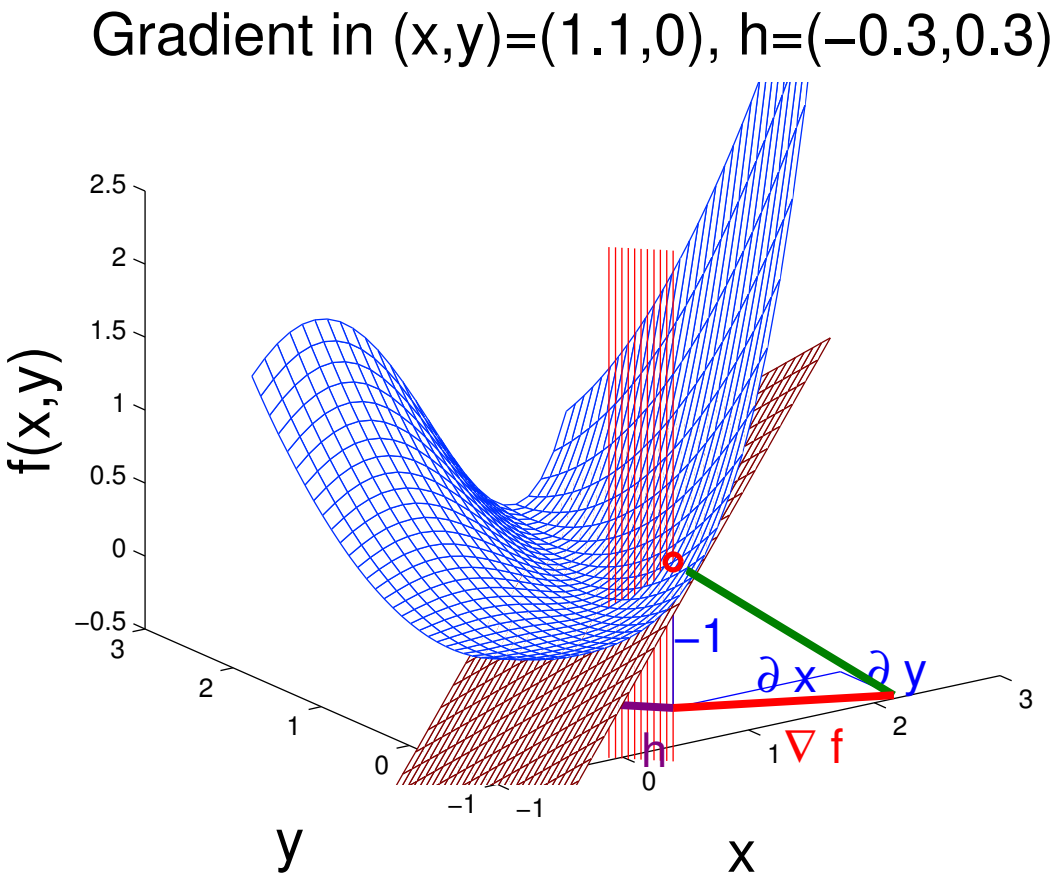


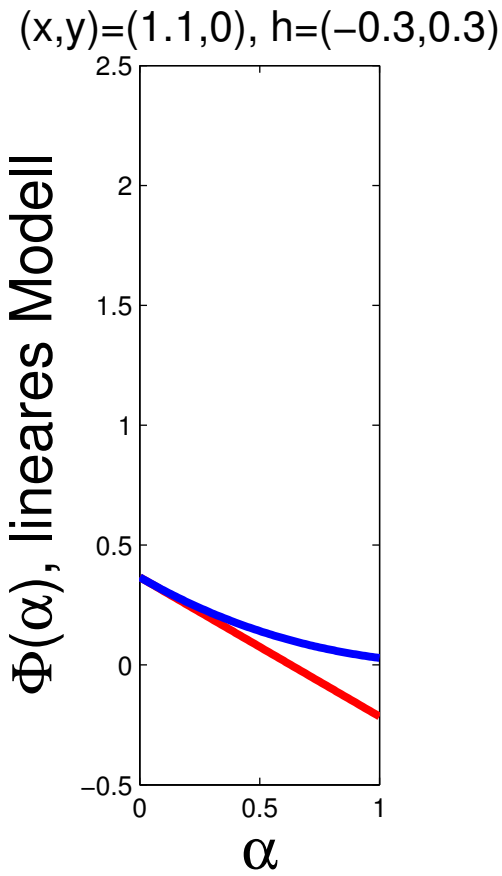
Gradient und Richtungsableitung, lineares Modell

Das lineare Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ den gleichen Anstieg wie f in $\bar{x}$: die Richtungsableitung von f in $\bar{x}$ in Richtung h,

$$\nabla f(\bar{x})^T h = \lim_{\alpha \downarrow 0} \frac{f(\bar{x} + \alpha h) - f(\bar{x})}{\alpha} =: D_h f(\bar{x}) \qquad [D_{\lambda h} f(\bar{x}) = \lambda D_h f(\bar{x})]$$

[= Ableitung $\frac{d}{d\alpha}\Phi(0)$ der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]





Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

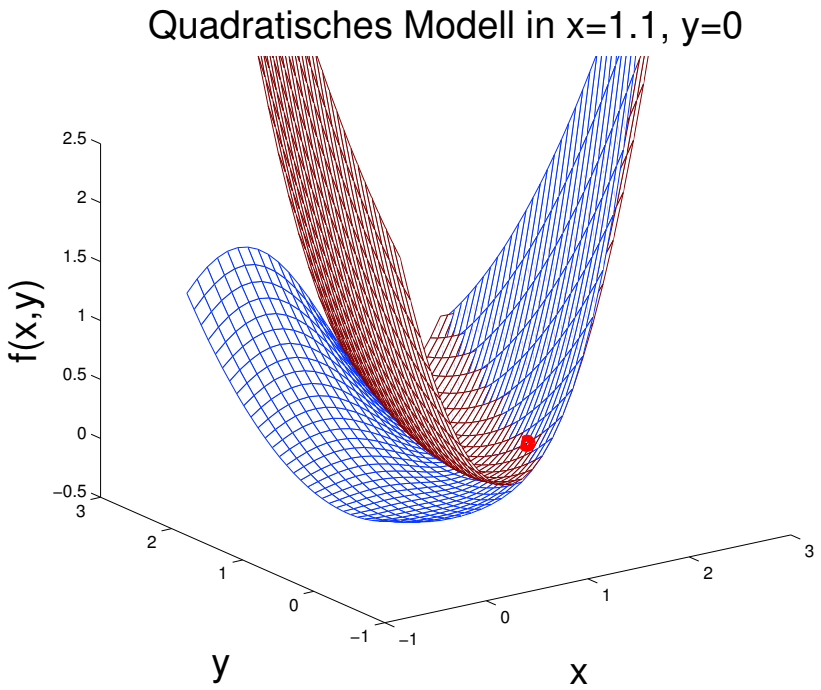
Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- $\nabla^2 f(x)$ ist symmetrisch, falls f zweimal stetig differenzierbar ist.
- $\nabla^2 f(x)$ ist die Krümmung von f in x.
- Das quadrat. Modell aus $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ schmiegt sich in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$ an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f an.



$f(\bar{x})$, $\nabla f(\bar{x})$ und $\nabla^2 f(\bar{x})$ bilden das quadratische Modell von f um $\bar{x}$,

$$\breve{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine sehr gute Näherung an f: für glattes f erfüllt $\nabla^2 f$

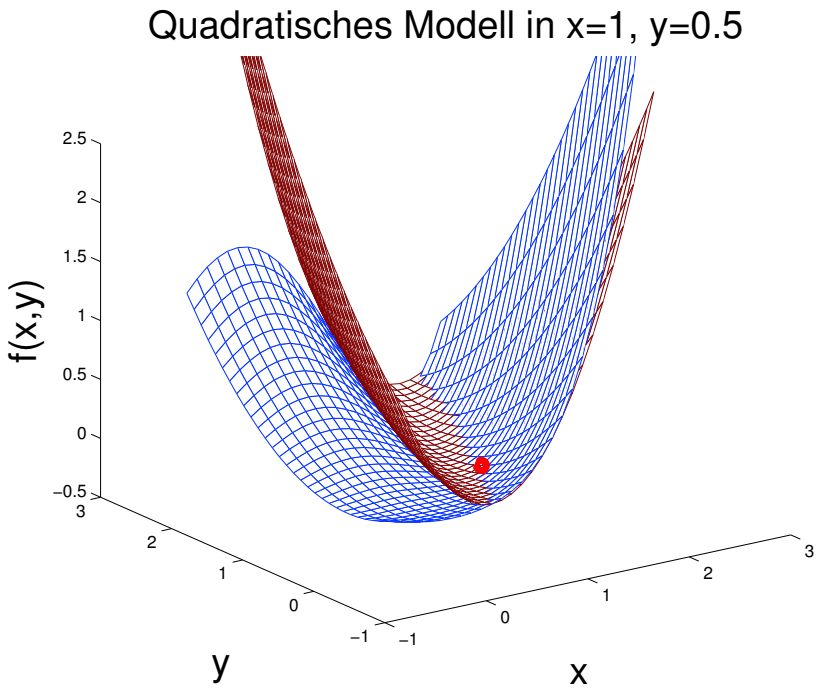
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x}) - \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$

Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- $\nabla^2 f(x)$ ist symmetrisch, falls f zweimal stetig differenzierbar ist.
- $\nabla^2 f(x)$ ist die Krümmung von f in x.
- Das quadrat. Modell aus $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ schmiegt sich in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$ an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f an.



$f(\bar{x})$, $\nabla f(\bar{x})$ und $\nabla^2 f(\bar{x})$ bilden das quadratische Modell von f um $\bar{x}$,

$$\breve{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine sehr gute Näherung an f: für glattes f erfüllt $\nabla^2 f$

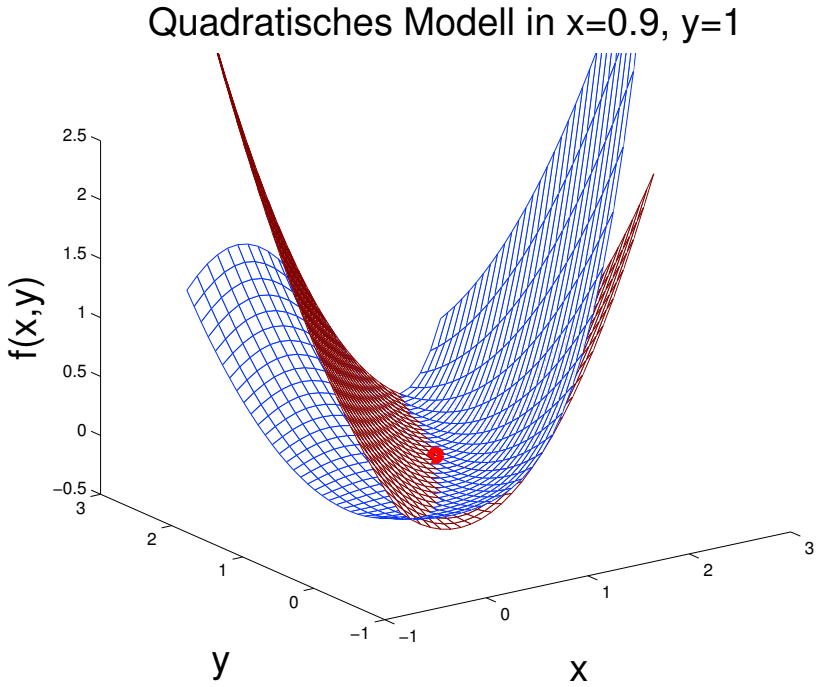
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x}) - \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$

Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- $\nabla^2 f(x)$ ist symmetrisch, falls f zweimal stetig differenzierbar ist.
- $\nabla^2 f(x)$ ist die Krümmung von f in x.
- Das quadrat. Modell aus $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ schmiegt sich in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$ an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f an.



$f(\bar{x})$, $\nabla f(\bar{x})$ und $\nabla^2 f(\bar{x})$ bilden das quadratische Modell von f um $\bar{x}$,

$$\breve{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine sehr gute Näherung an f: für glattes f erfüllt $\nabla^2 f$

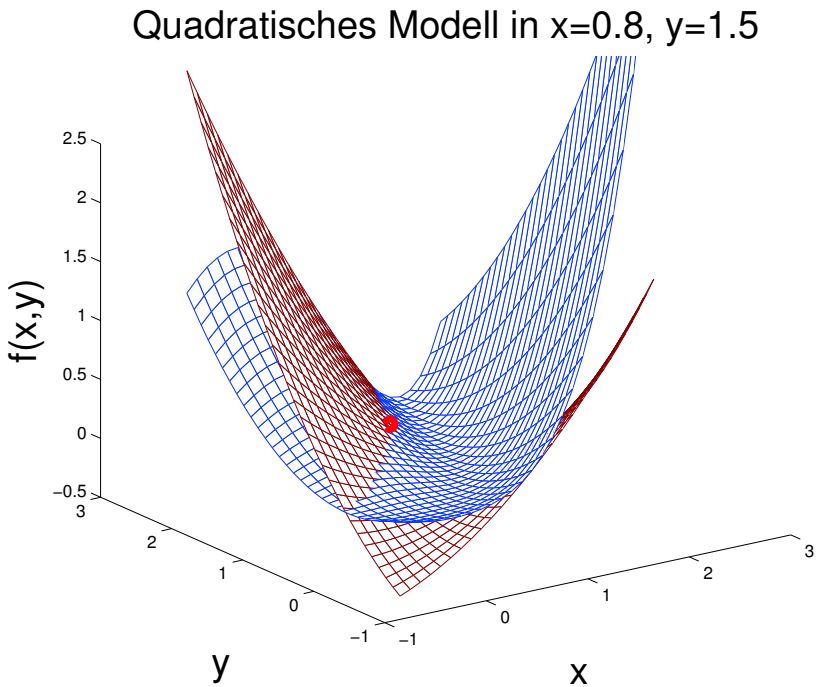
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x}) - \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$

Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- $\nabla^2 f(x)$ ist symmetrisch, falls f zweimal stetig differenzierbar ist.
- $\nabla^2 f(x)$ ist die Krümmung von f in x.
- Das quadrat. Modell aus $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ schmiegt sich in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$ an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f an.



$f(\bar{x})$, $\nabla f(\bar{x})$ und $\nabla^2 f(\bar{x})$ bilden das quadratische Modell von f um $\bar{x}$,

$$\breve{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine sehr gute Näherung an f: für glattes f erfüllt $\nabla^2 f$

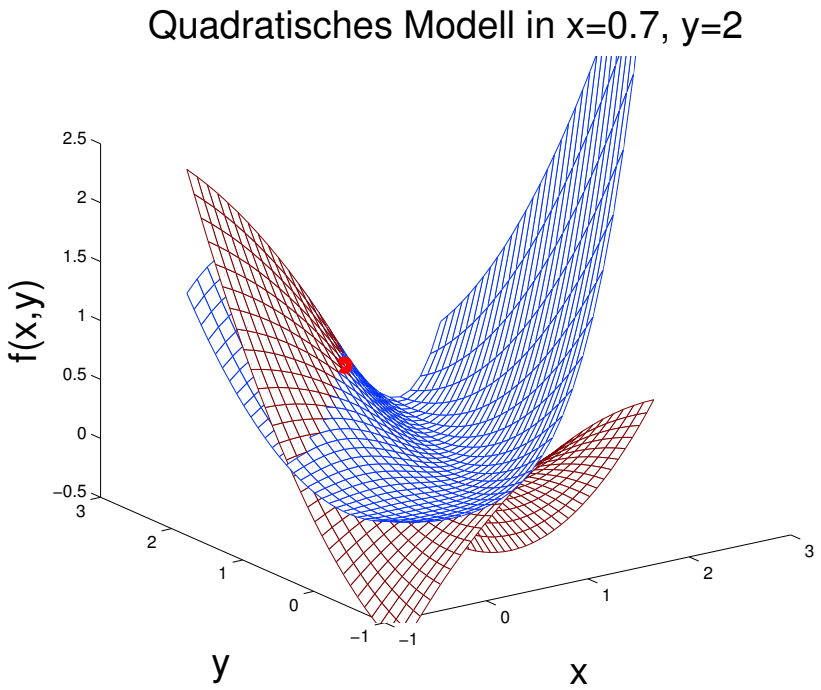
$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x}) - \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$

Orakel 2. Ordnung: $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$

Für $\bar{x} \in \mathbb{R}^n$ werden $f(\bar{x})$, $\nabla f(\bar{x})$ und Hessematrix $\nabla^2 f(\bar{x})$ (2. Abl.) berechnet.

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

- $\nabla^2 f(x)$ ist symmetrisch, falls f zweimal stetig differenzierbar ist.
- $\nabla^2 f(x)$ ist die Krümmung von f in x.
- Das quadrat. Modell aus $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$ schmiegt sich in $\begin{bmatrix} x \\ f(x) \end{bmatrix}$ an den Graphen $\left\{ \begin{bmatrix} x \\ f(x) \end{bmatrix} : x \in \mathbb{R}^n \right\}$ von f an.



$f(\bar{x})$, $\nabla f(\bar{x})$ und $\nabla^2 f(\bar{x})$ bilden das quadratische Modell von f um $\bar{x}$,

$$\breve{f}_{\bar{x}}(x) := f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}).$$

Für x nahe bei $\bar{x}$ ist es eine sehr gute Näherung an f: für glattes f erfüllt $\nabla^2 f$

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x})^T (x - \bar{x}) - \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x})}{\|x - \bar{x}\|^2} = 0.$$

Quadratisches Modell in Richtung h

Das quadratische Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ die gleiche Steigung und Krümmung wie f in $\bar{x}$.

$$\breve{f}_{\bar{x}}(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \nabla f(\bar{x})^T h + \frac{\alpha^2}{2} h^T \nabla^2 f(\bar{x}) h.$$

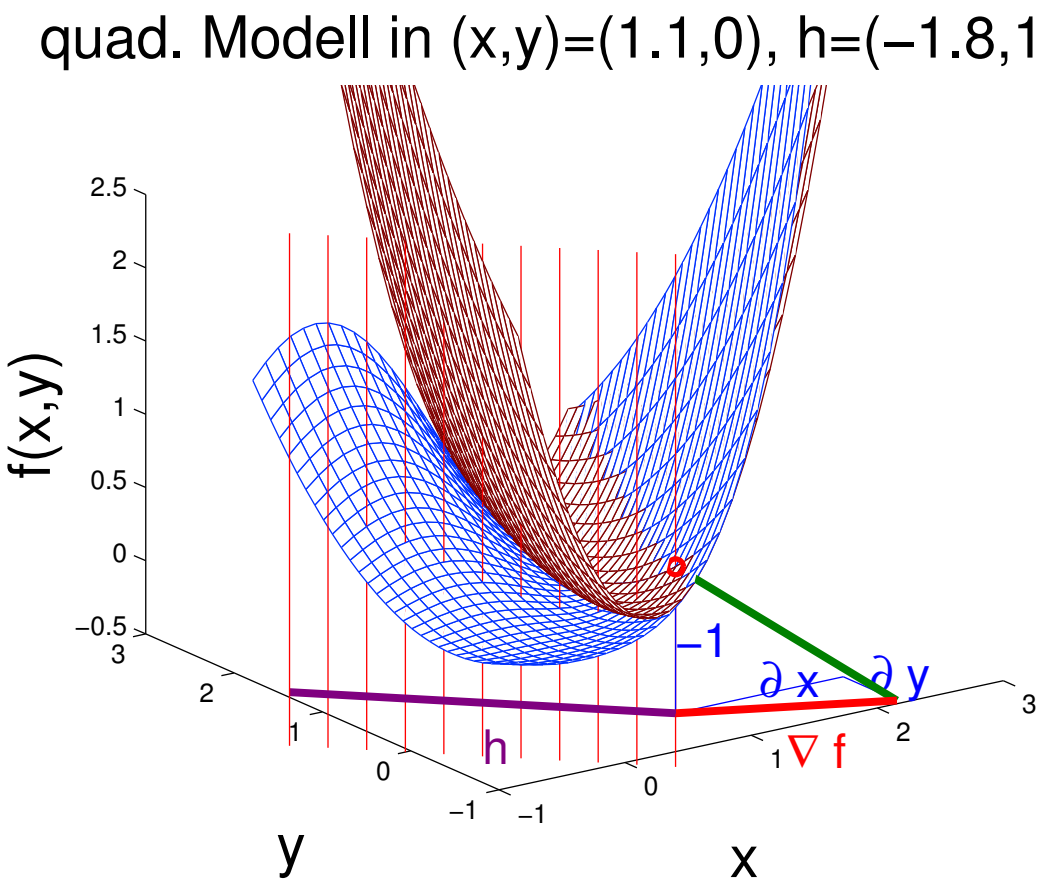
[Taylor-Entw. 2. Ord. der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]

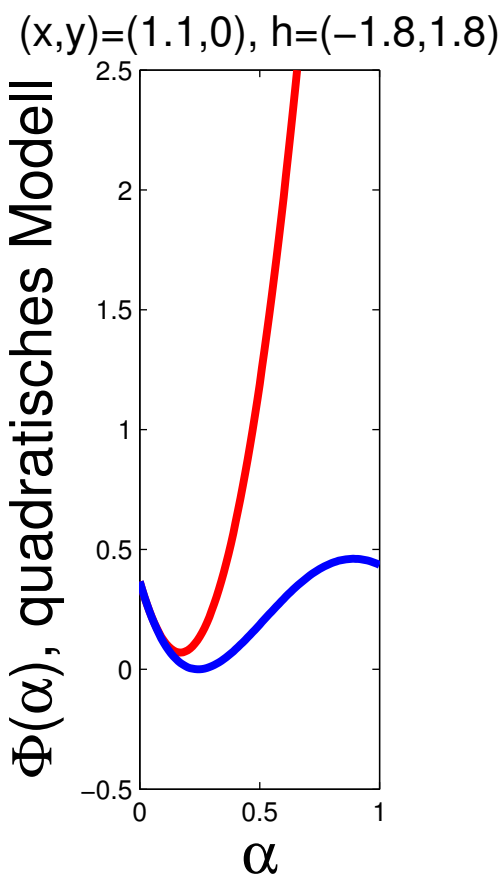
Quadratisches Modell in Richtung h

Das quadratische Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ die gleiche Steigung und Krümmung wie f in $\bar{x}$.

$$\breve{f}_{\bar{x}}(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \nabla f(\bar{x})^T h + \frac{\alpha^2}{2} h^T \nabla^2 f(\bar{x}) h.$$

[Taylor-Entw. 2. Ord. der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



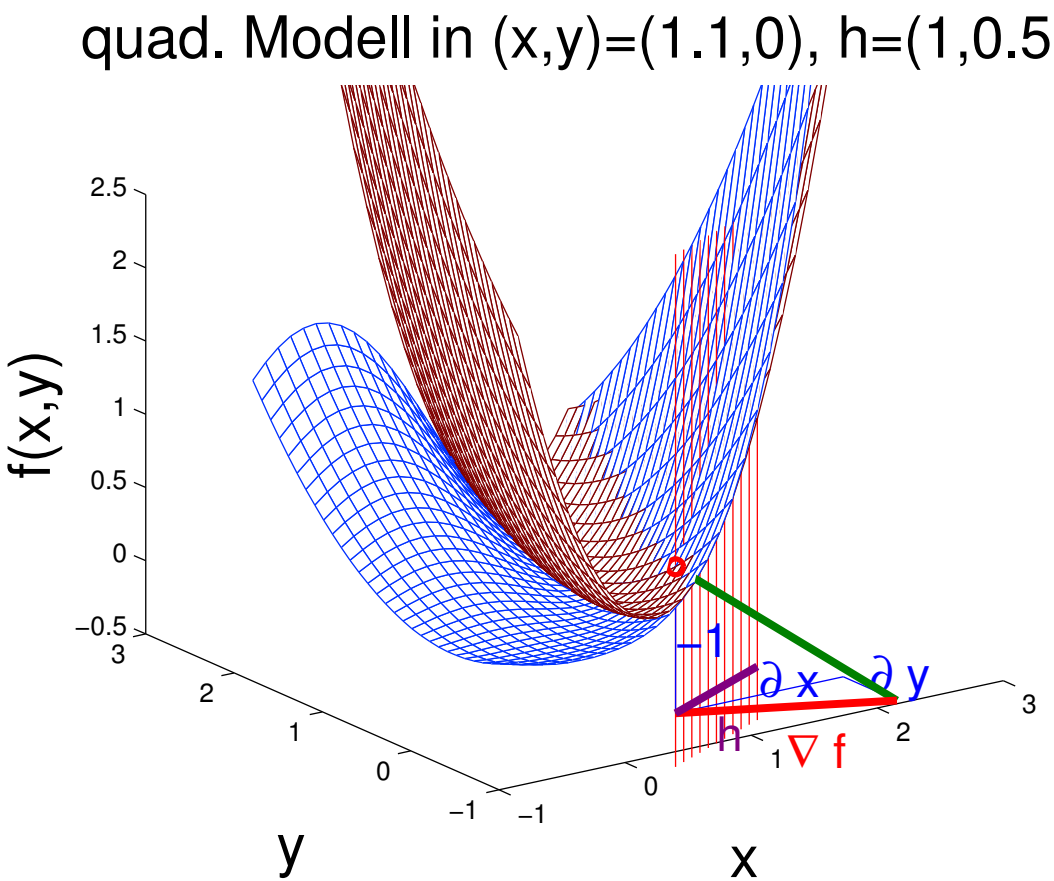


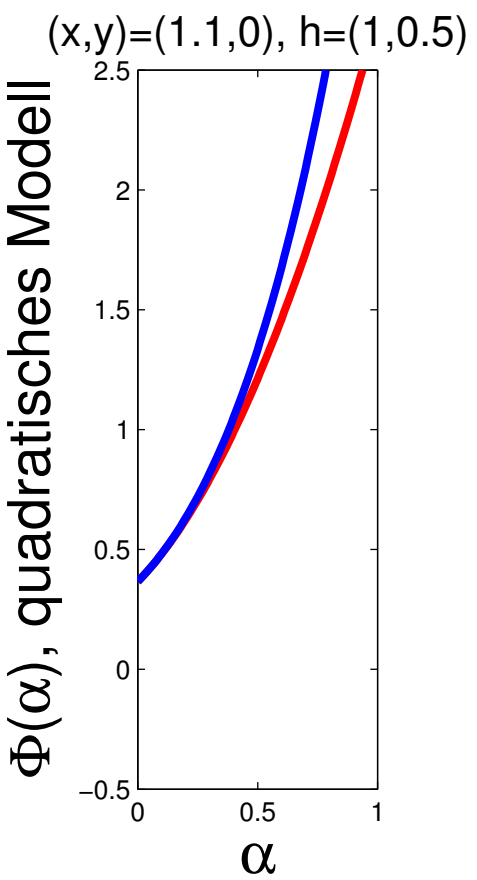
Quadratisches Modell in Richtung h

Das quadratische Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ die gleiche Steigung und Krümmung wie f in $\bar{x}$.

$$\breve{f}_{\bar{x}}(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \nabla f(\bar{x})^T h + \frac{\alpha^2}{2} h^T \nabla^2 f(\bar{x}) h.$$

[Taylor-Entw. 2. Ord. der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



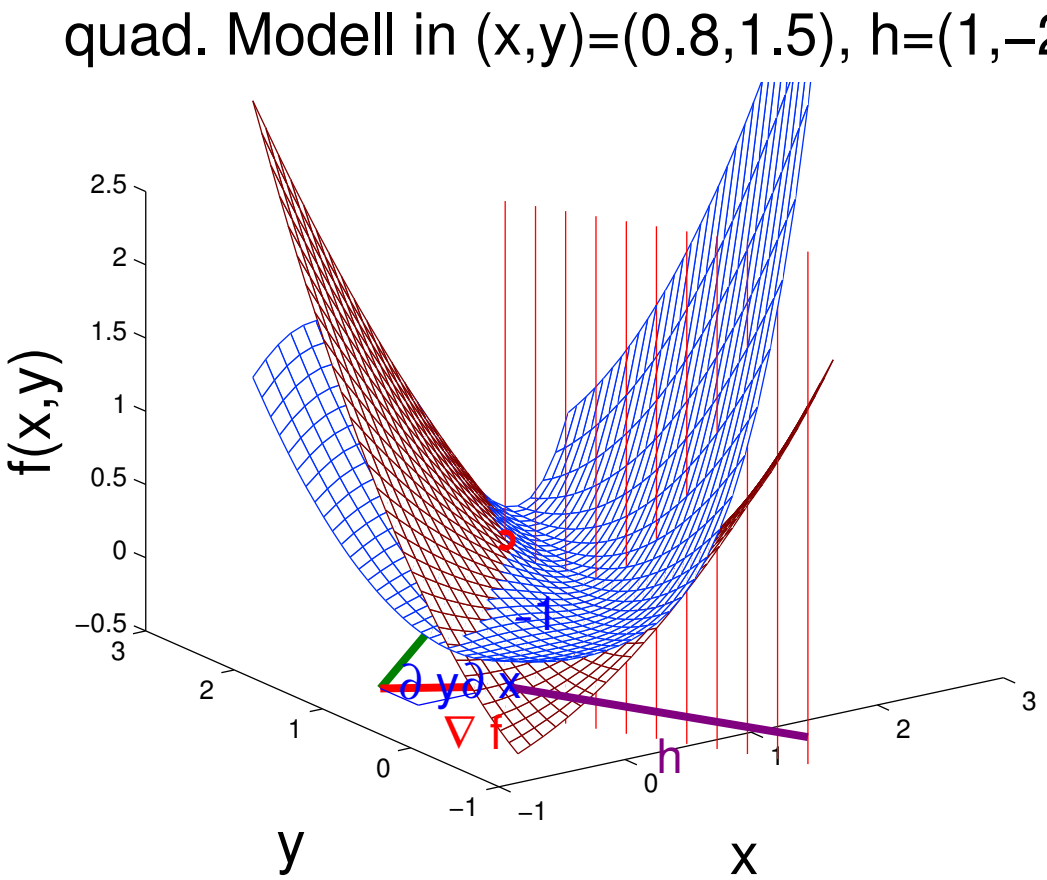


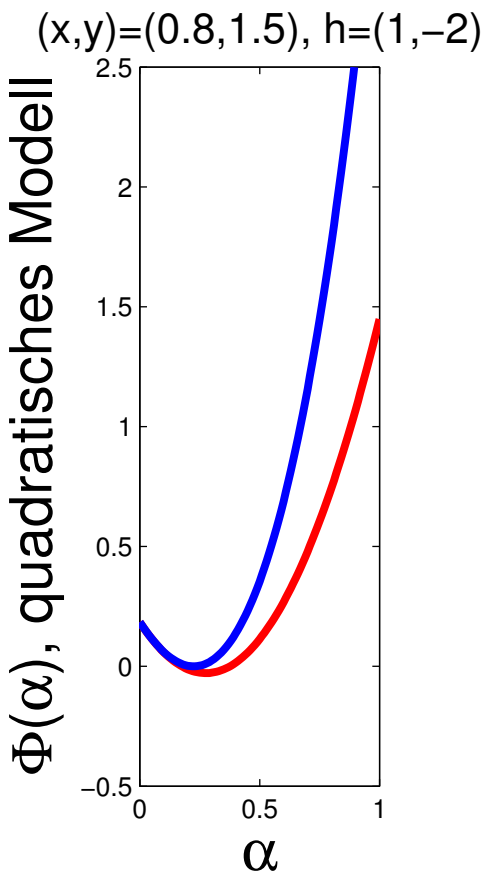
Quadratisches Modell in Richtung h

Das quadratische Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ die gleiche Steigung und Krümmung wie f in $\bar{x}$.

$$\breve{f}_{\bar{x}}(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \nabla f(\bar{x})^T h + \frac{\alpha^2}{2} h^T \nabla^2 f(\bar{x}) h.$$

[Taylor-Entw. 2. Ord. der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]



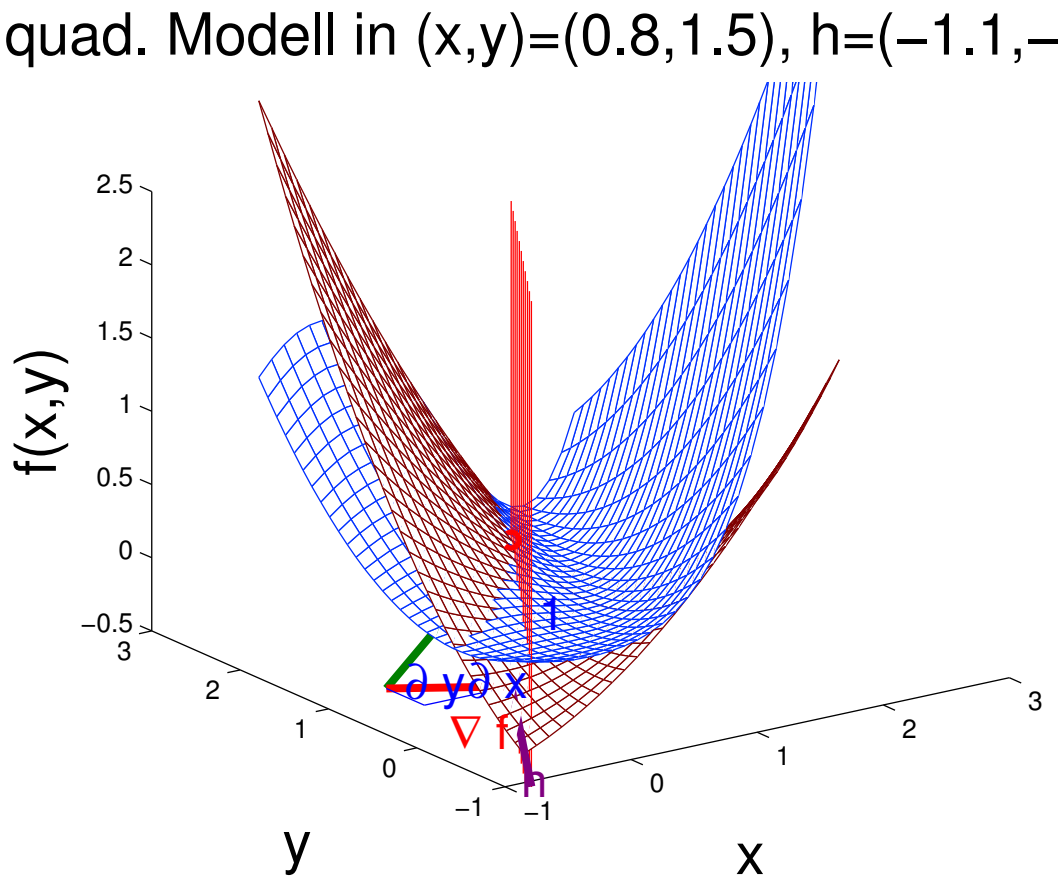


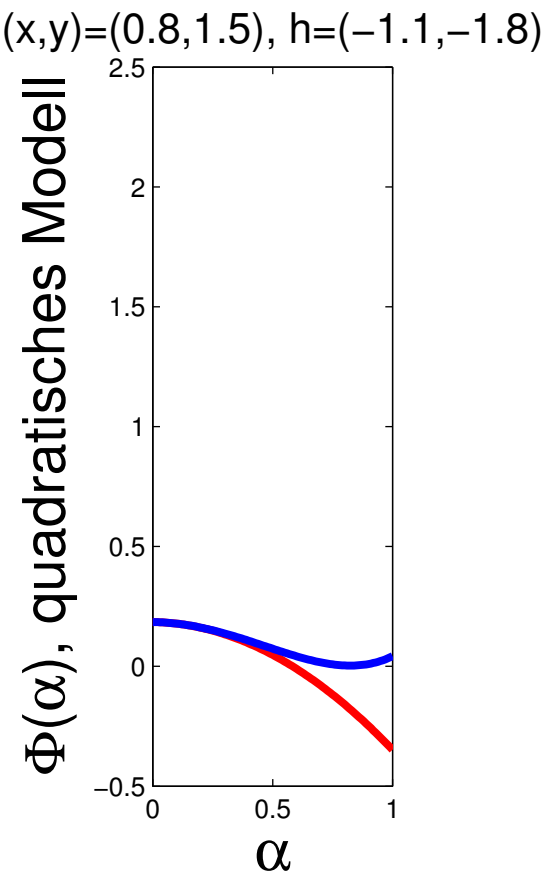
Quadratisches Modell in Richtung h

Das quadratische Modell von f in $\bar{x}$ hat in jede Richtung $h \in \mathbb{R}^n$ die gleiche Steigung und Krümmung wie f in $\bar{x}$.

$$\breve{f}_{\bar{x}}(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \nabla f(\bar{x})^T h + \frac{\alpha^2}{2} h^T \nabla^2 f(\bar{x}) h.$$

[Taylor-Entw. 2. Ord. der 1-D Funktion $\Phi : \mathbb{R} \to \mathbb{R}$, $\Phi(\alpha) := f(x + \alpha h)$]





Der Satz von Taylor/Mittelwertsatz

Satz (Taylor/Mittelwertsatz)

Sei f oft genug stetig differenzierbar und $\bar{x}, h \in \mathbb{R}^n$, dann

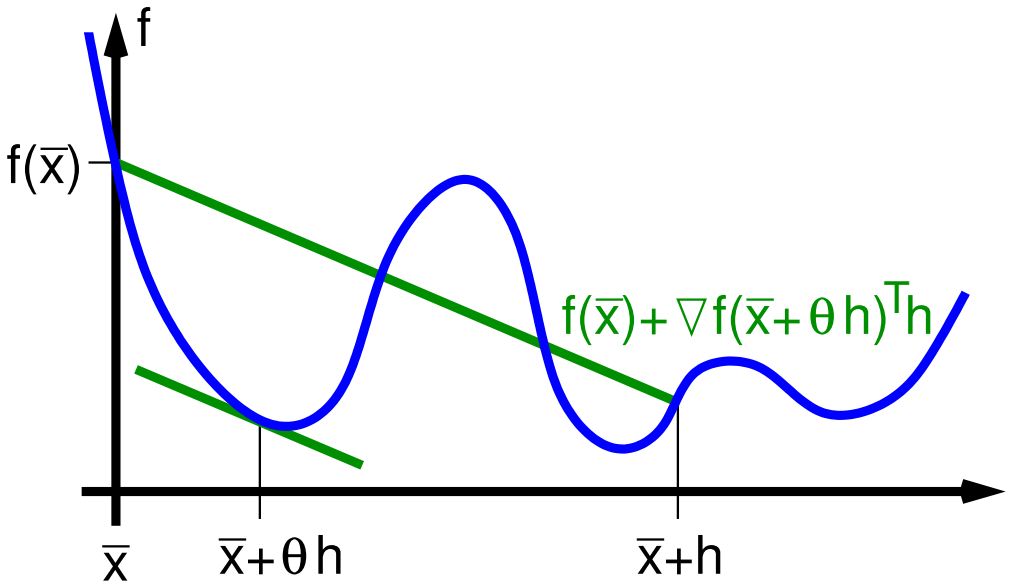
$$\exists \theta_1 \in (0,1): \quad f(\bar{x}+h) = f(\bar{x}) + \nabla f(\bar{x}+\theta_1 h)^T h,$$

$$\exists \theta_2 \in (0,1): \quad f(\bar{x}+h) = f(\bar{x}) + \nabla f(\bar{x})^T h + \tfrac{1}{2} h^T \nabla^2 f(\bar{x}+\theta_2 h) h,$$

$$\exists \theta_3 \in (0,1): \quad f(\bar{x}+h) = f(\bar{x}) + \nabla f(\bar{x})^T h + \tfrac{1}{2} h^T \nabla^2 f(\bar{x}) h + \tfrac{1}{6} \nabla^3 f(\bar{x}+\theta_3 h)[h,h,h]$$

[∇^3 steht für die 3. Ableitung] „Taylor-Entwicklung von f um $\bar{x}$“

Illustration des ersten Falls des Mittelwertsatzes:



Lipschitz-Stetigkeit, „Klein-**o**-Notation“

Eine Funktion $G : \mathbb{R}^n \to \mathbb{R}^m$ heißt Lipschitz-stetig auf einer Menge $S \subseteq \mathbb{R}^n$, falls es eine Konstante $L > 0$ gibt mit $\|G(x) - G(y)\| \leq L\|x - y\| \ \forall x, y \in S$.
[$\Rightarrow$ Die Werte können sich nur bei größerem Abstand stark ändern!]

Lipschitz-Stetigkeit, „Klein-**o**-Notation“

Eine Funktion $G : \mathbb{R}^n \to \mathbb{R}^m$ heißt Lipschitz-stetig auf einer Menge $S \subseteq \mathbb{R}^n$, falls es eine Konstante $L > 0$ gibt mit $\|G(x) - G(y)\| \leq L\|x - y\| \ \forall x, y \in S$.
[$\Rightarrow$ Die Werte können sich nur bei größerem Abstand stark ändern!]

Jede auf $\mathbb{R}^n$ stetig differenzierbare Funktion ist für jedes $\bar{x} \in \mathbb{R}^n$ und $\rho > 0$ auf der ρ-Kugel um $\bar{x}$, $B_\rho(\bar{x}) := \{x \in \mathbb{R}^n : \|x - \bar{x}\| \leq \rho\}$ Lipschitz-stetig.

- Ist ∇f um $\bar{x}$ für großes ρ Lipschitz-stetig mit kleinem L, dann ist das lineare Modell auf B_ρ eine gute Näherung an f.
- Ist $\nabla^2 f$ um $\bar{x}$ für großes ρ Lipschitz-stetig mit kleinem L, dann ist das quadratische Modell auf B_ρ eine gute Näherung an f.

Lipschitz-Stetigkeit, „Klein-**o**-Notation“

Eine Funktion $G : \mathbb{R}^n \to \mathbb{R}^m$ heißt Lipschitz-stetig auf einer Menge $S \subseteq \mathbb{R}^n$, falls es eine Konstante $L > 0$ gibt mit $\|G(x) - G(y)\| \leq L\|x - y\| \ \forall x, y \in S$.
[$\Rightarrow$ Die Werte können sich nur bei größerem Abstand stark ändern!]

Jede auf $\mathbb{R}^n$ stetig differenzierbare Funktion ist für jedes $\bar{x} \in \mathbb{R}^n$ und $\rho > 0$ auf der ρ-Kugel um $\bar{x}$, $B_\rho(\bar{x}) := \{x \in \mathbb{R}^n : \|x - \bar{x}\| \leq \rho\}$ Lipschitz-stetig.

- Ist ∇f um $\bar{x}$ für großes ρ Lipschitz-stetig mit kleinem L, dann ist das lineare Modell auf B_ρ eine gute Näherung an f.
- Ist $\nabla^2 f$ um $\bar{x}$ für großes ρ Lipschitz-stetig mit kleinem L, dann ist das quadratische Modell auf B_ρ eine gute Näherung an f.

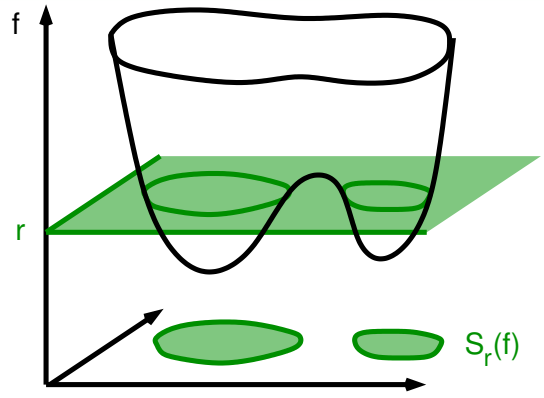
Aus dem Satz von Taylor und der Lipschitz-Stetigkeit von $\nabla^k f$ folgt
$f(\bar{x} + h) = f(\bar{x}) + \nabla f(\bar{x})^T h + \mathbf{o}(\|h\|)$,
$f(\bar{x} + h) = f(\bar{x}) + \nabla f(\bar{x})^T h + \frac{1}{2} h^T \nabla^2 f(\bar{x}) h + \mathbf{o}(\|h\|^2)$

Das Landau-Symbol $\mathbf{o}(g(y))$ steht immer als Ersatz für eine nicht weiter interessierende Funktion $g'(y)$ mit der Eigenschaft $\lim_{y \to 0} \frac{g'(y)}{g(y)} \to 0$, also ein g', das schneller klein wird als g.
Bei Folgen $g(y^{(k)})$ steht es für ein $g'(y^{(k)})$ mit $\lim_{k \to \infty} \frac{g'(y^{(k)})}{g(y^{(k)})} \to 0$.

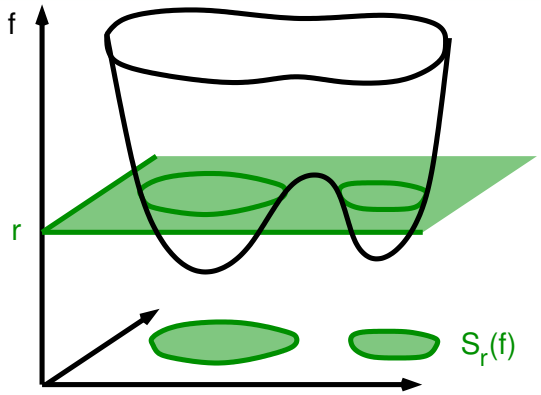
Niveaumengen und Niveaulinien

Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$



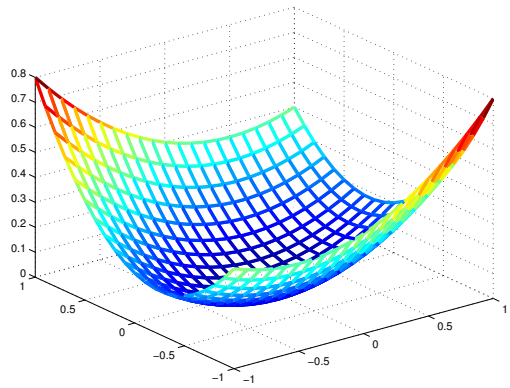
Niveaumengen und Niveaulinien



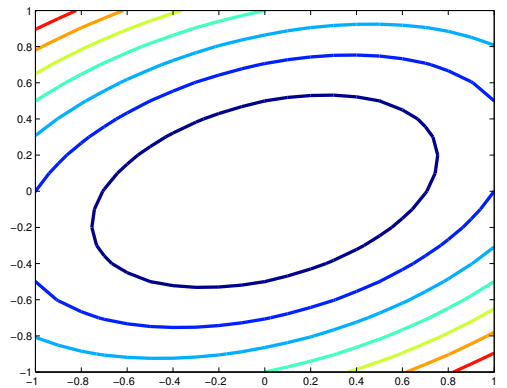
Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$

Funktionsdarstellungen über Niveaulinien [„Linien“]
$N_r(f) := \{x \in \mathbb{R}^n : f(x) = r\}$ (*contour plots*)
helfen, Verfahren zu illustrieren.
[Höhenlinien in Landkarten, Wetterkarten]



Beachte: Der Gradient ist immer orthogonal zur Niveaulinie, denn für $x, x+h \in N_r(f)$ gilt $0 = f(x+h) - f(x) = \nabla f(x)^T h + \mathbf{o}(\|h\|)$.

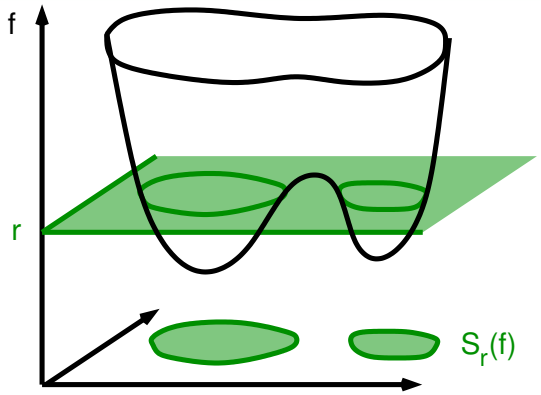


Bsp: quadratische Funktion

$$\tfrac{1}{2} x^T Q x + q^T x + d$$

mit Q positiv definit.

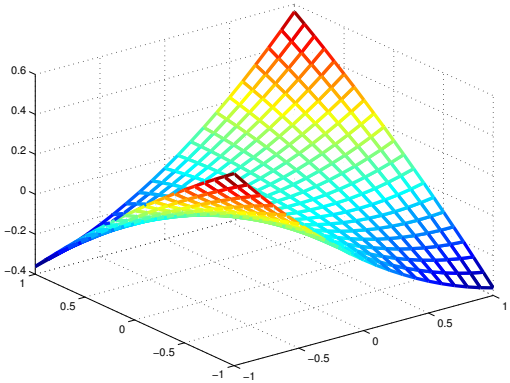
Niveaumengen und Niveaulinien



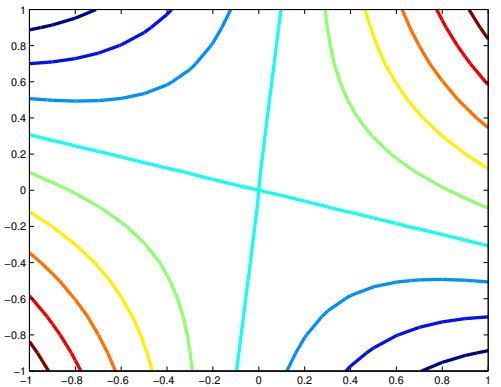
Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$

Funktionsdarstellungen über Niveaulinien [„Linien“]
$N_r(f) := \{x \in \mathbb{R}^n : f(x) = r\}$ (*contour plots*)
helfen, Verfahren zu illustrieren.
[Höhenlinien in Landkarten, Wetterkarten]



Beachte: Der Gradient ist immer orthogonal zur Niveaulinie, denn für $x, x+h \in N_r(f)$ gilt $0 = f(x+h) - f(x) = \nabla f(x)^T h + \mathbf{o}(\|h\|)$.

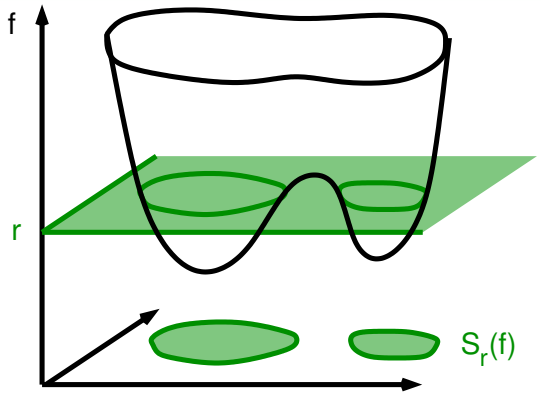


Bsp: quadratische Funktion

$$\tfrac{1}{2} x^T Q x + q^T x + d$$

mit Q indefinit.

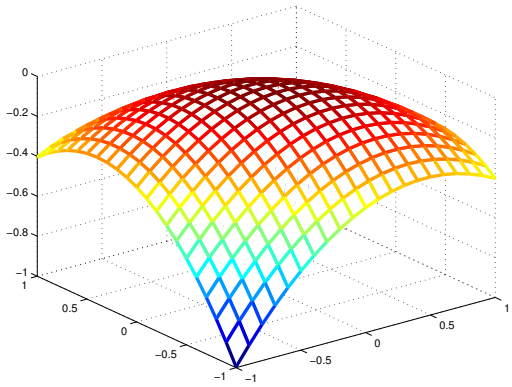
Niveaumengen und Niveaulinien



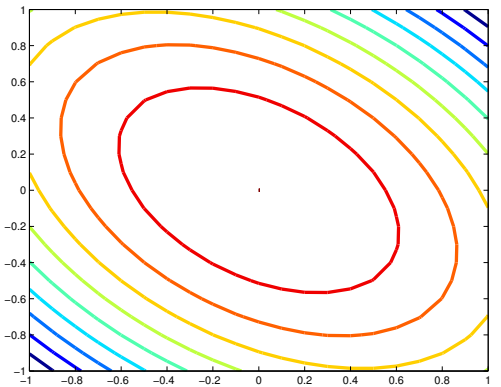
Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$

Funktionsdarstellungen über Niveaulinien [„Linien“]
$N_r(f) := \{x \in \mathbb{R}^n : f(x) = r\}$ (*contour plots*)
helfen, Verfahren zu illustrieren.
[Höhenlinien in Landkarten, Wetterkarten]



Beachte: Der Gradient ist immer orthogonal zur Niveaulinie, denn für $x, x+h \in N_r(f)$ gilt $0 = f(x+h) - f(x) = \nabla f(x)^T h + \mathbf{o}(\|h\|)$.



Bsp: quadratische Funktion

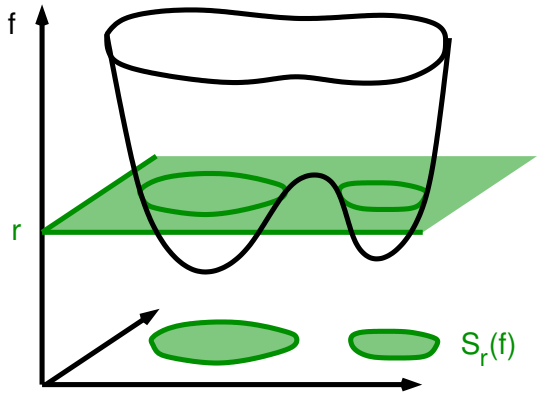
$$\tfrac{1}{2} x^T Q x + q^T x + d$$

mit Q negativ definit.

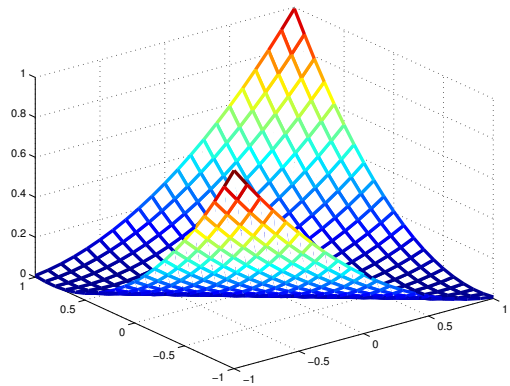
Niveaumengen und Niveaulinien

Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

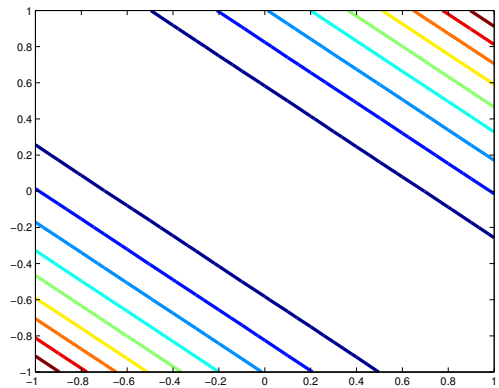
$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$



Funktionsdarstellungen über Niveaulinien [„Linien“]
$N_r(f) := \{x \in \mathbb{R}^n : f(x) = r\}$ (*contour plots*)
helfen, Verfahren zu illustrieren.
[Höhenlinien in Landkarten, Wetterkarten]



Beachte: Der Gradient ist immer orthogonal zur Niveaulinie, denn für $x, x+h \in N_r(f)$ gilt $0 = f(x+h) - f(x) = \nabla f(x)^T h + \mathbf{o}(\|h\|)$.

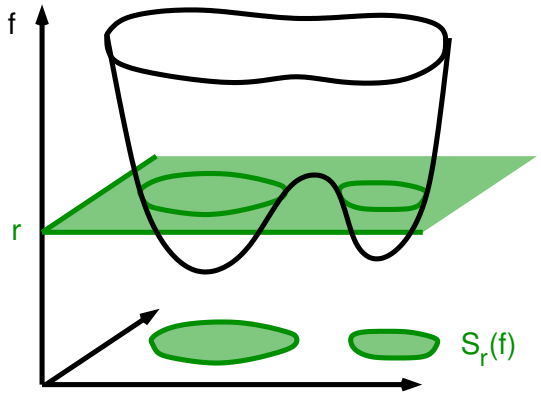


Bsp: quadratische Funktion

$$\tfrac{1}{2} x^T Q x + q^T x + d$$

mit Q positiv semidefinit.

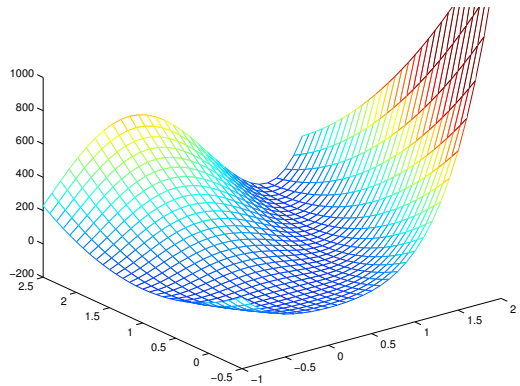
Niveaumengen und Niveaulinien



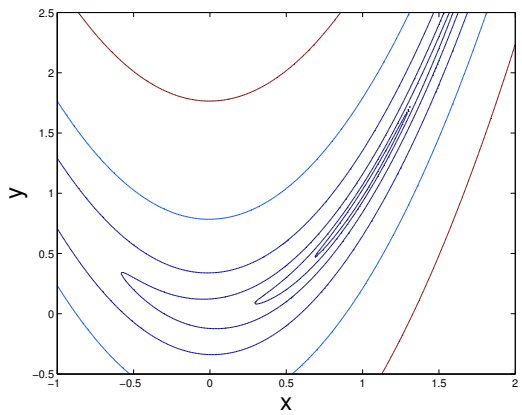
Verfahren der freien nichtlinearen Optimierung suchen nur Punkte mit besserem Zielfunktionswert als dem derzeitigen, also nur Punkte aus der Niveaumenge von f zu einem Wert $r \in \mathbb{R}$,

$$S_r(f) := \{x \in \mathbb{R}^n : f(x) \leq r\}.$$

Funktionsdarstellungen über Niveaulinien [„Linien“]
$N_r(f) := \{x \in \mathbb{R}^n : f(x) = r\}$ (*contour plots*)
helfen, Verfahren zu illustrieren.
[Höhenlinien in Landkarten, Wetterkarten]



Beachte: Der Gradient ist immer orthogonal zur Niveaulinie, denn für $x, x+h \in N_r(f)$ gilt $0 = f(x+h) - f(x) = \nabla f(x)^T h + \mathbf{o}(\|h\|)$.



Bsp: Rosenbrock-Funktion (*banana shape*)

$$f(x,y) = \tfrac{1}{4}[(y-x^2)^2 + \tfrac{1}{100}(1-x)^2]$$

Minimum wird in (1,1) angenommen.

Inhaltsübersicht für heute:

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Ein Punkt $\bar{x}$ mit $\nabla f(\bar{x}) = 0$ heißt stationärer Punkt von f.

Stationarität ist notwendig, aber i.A. nicht hinreichend für Minimalität!
Bsp: $x = 0$ ist stationärer Punkt von $f(x) = x^3$ oder $f(x) = -x^2$.

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Ein Punkt $\bar{x}$ mit $\nabla f(\bar{x}) = 0$ heißt stationärer Punkt von f.

Stationarität ist notwendig, aber i.A. nicht hinreichend für Minimalität!
Bsp: $x = 0$ ist stationärer Punkt von $f(x) = x^3$ oder $f(x) = -x^2$.

Ausnahme: Für konvexes f ist jeder stationäre Punkt globales Minimum!

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Ein Punkt $\bar{x}$ mit $\nabla f(\bar{x}) = 0$ heißt stationärer Punkt von f.

Stationarität ist notwendig, aber i.A. nicht hinreichend für Minimalität!
Bsp: $x = 0$ ist stationärer Punkt von $f(x) = x^3$ oder $f(x) = -x^2$.

Ausnahme: Für konvexes f ist jeder stationäre Punkt globales Minimum!

Bsp: $f(x) = \frac{1}{2} x^T Q x + q^T x + c$ mit $Q \succ 0$ ist (streng) konvex.
Mit $\nabla f(x) = Qx + q$ bestimmt $\nabla f(x^*) = 0$ das Minimum x^* eindeutig,

$$x^* = -Q^{-1} q.$$

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Ein Punkt $\bar{x}$ mit $\nabla f(\bar{x}) = 0$ heißt stationärer Punkt von f.

Stationarität ist notwendig, aber i.A. nicht hinreichend für Minimalität!
Bsp: $x = 0$ ist stationärer Punkt von $f(x) = x^3$ oder $f(x) = -x^2$.

Ausnahme: Für konvexes f ist jeder stationäre Punkt globales Minimum!

Geometrisch bedeutet $\nabla f(x) = 0$, dass die Tangentialebene an f in x „waagrecht“ liegt.
Ist sie nicht waagrecht, kann man sicher noch ein wenig hinunterrutschen.

Notwendige Optimalitätsbedingung 1. Ordnung

Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Minimum einer glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, so gilt

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

Sonst wäre (setze $h = -\nabla f(\bar{x})$ in Taylor) für α klein genug

$$f(\bar{x} - \alpha \nabla f(\bar{x})) = f(\bar{x}) - \alpha \underbrace{\nabla f(\bar{x})^T \nabla f(\bar{x})}_{= \|\nabla f(\bar{x})\|^2 > 0} + \mathbf{o}(\alpha) < f(\bar{x}).$$

Ein Punkt $\bar{x}$ mit $\nabla f(\bar{x}) = 0$ heißt stationärer Punkt von f.

Stationarität ist notwendig, aber i.A. nicht hinreichend für Minimalität!
Bsp: $x = 0$ ist stationärer Punkt von $f(x) = x^3$ oder $f(x) = -x^2$.

Ausnahme: Für konvexes f ist jeder stationäre Punkt globales Minimum!

Konsequenz für Optimierungsverfahren:
Ist $\nabla f(x) \neq 0$, so kann man die Funktion in Richtung $-\nabla f(x)$ immer verbessern (u.U. nur für sehr kleine Schrittweite).
Die Schrittrichtung $h = -\nabla f(x)$ heißt steilster Abstieg (*steepest descent*).

Notwendige Optimalitätsbedingung 2. Ordnung

Satz (Notwendige Optimalitätsbedingung 2. Ordnung)

Ist $\bar{x} \in \mathbb{R}^n$ ein lokales Min. einer hinr. glatten Funktion $f : \mathbb{R}^n \to \mathbb{R}$, gilt

$$\nabla f(\bar{x}) = 0 \quad \textit{und} \quad \nabla^2 f(\bar{x}) \succeq 0.$$

Sonst gibt es ein $h \in \mathbb{R}^n$ mit $h^T \nabla^2 f(\bar{x}) h < 0$, Taylor ergibt für kleine α

$$f(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T h}_{=0 \ (\nabla f(\bar{x})=0)} + \frac{\alpha^2}{2} \underbrace{h^T \nabla^2 f(\bar{x}) h}_{<0} + \mathbf{o}(\alpha^2) < f(\bar{x}).$$

Die Bedingung ist wieder nur notwendig und i.A. nicht hinreichend.

Bsp: $f(x, y) = x^2 - y^4$, $\nabla^2 f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & -12y^2 \end{bmatrix}$ für $(x, y) = (0, 0)$.

Konsequenz für Optimierungsverfahren:
Ist zwar $\nabla f(\bar{x}) = 0$ aber $\lambda_{\min}(\nabla^2 f(\bar{x})) < 0$, so kann f in Richtung eines Eigenvektors zu $\lambda_{\min}$ verbessert werden.

Hinreichende Optimalitätsbedingung 2. Ordnung

Satz (Hinreichende Optimalitätsbedingung 2. Ordnung)

Gilt für ein $\bar{x} \in \mathbb{R}^n$ *sowohl* $\nabla f(\bar{x}) = 0$ *als auch* $\nabla^2 f(\bar{x}) \succ 0$,
so ist $\bar{x}$ *ein lokales Minimum von* f.

Denn für beliebiges $h \in \mathbb{R}^n \setminus \{0\}$ und α klein genug gilt mit Taylor

$$f(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T h}_{=0 \ (\nabla f(\bar{x})=0)} + \frac{\alpha^2}{2} \underbrace{h^T \nabla^2 f(\bar{x}) h}_{>0} + \mathbf{o}(\alpha^2) > f(\bar{x}).$$

Hinreichende Optimalitätsbedingung 2. Ordnung

Satz (Hinreichende Optimalitätsbedingung 2. Ordnung)

Gilt für ein $\bar{x} \in \mathbb{R}^n$ *sowohl* $\nabla f(\bar{x}) = 0$ *als auch* $\nabla^2 f(\bar{x}) \succ 0$,
so ist $\bar{x}$ *ein lokales Minimum von* f.

Denn für beliebiges $h \in \mathbb{R}^n \setminus \{0\}$ und α klein genug gilt mit Taylor

$$f(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T h}_{=0 \ (\nabla f(\bar{x})=0)} + \frac{\alpha^2}{2} \underbrace{h^T \nabla^2 f(\bar{x}) h}_{>0} + \mathbf{o}(\alpha^2) > f(\bar{x}).$$

Die Bedingung ist hinreichend, aber nicht notwendig: $f(x) = x^4$ in $x = 0$.
In der Praxis ist sie erstaunlich oft erfüllt.

Hinreichende Optimalitätsbedingung 2. Ordnung

Satz (Hinreichende Optimalitätsbedingung 2. Ordnung)

Gilt für ein $\bar{x} \in \mathbb{R}^n$ *sowohl* $\nabla f(\bar{x}) = 0$ *als auch* $\nabla^2 f(\bar{x}) \succ 0$,
so ist $\bar{x}$ *ein lokales Minimum von* f.

Denn für beliebiges $h \in \mathbb{R}^n \setminus \{0\}$ und α klein genug gilt mit Taylor

$$f(\bar{x} + \alpha h) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T h}_{=0 \ (\nabla f(\bar{x})=0)} + \frac{\alpha^2}{2} \underbrace{h^T \nabla^2 f(\bar{x}) h}_{>0} + \mathbf{o}(\alpha^2) > f(\bar{x}).$$

Die Bedingung ist hinreichend, aber nicht notwendig: $f(x) = x^4$ in $x = 0$.
In der Praxis ist sie erstaunlich oft erfüllt.

Konsequenz für Optimierungsverfahren:
In der Nähe eines lokalen Minimums sieht die Funktion wie eine konvexe quadratische Funktion aus, das quadratische Modell ist dort eine gute Approximation!

Bemerkungen

- In Optimalitätsbedingungen für Maximierungsprobleme muss man nur $\nabla^2 f \succeq 0 (\succ 0)$ durch $\nabla^2 f \preceq 0 (\prec 0)$ ersetzen, der Rest bleibt gleich.

Bemerkungen

- In Optimalitätsbedingungen für Maximierungsprobleme muss man nur $\nabla^2 f \succeq 0 (\succ 0)$ durch $\nabla^2 f \preceq 0 (\prec 0)$ ersetzen, der Rest bleibt gleich.
- Stationäre Punkte sind entweder lokale Minima, lokale Maxima oder Sattelpunkte (manche Richtungen führen aufwärts, manche abwärts).

Bemerkungen

- In Optimalitätsbedingungen für Maximierungsprobleme muss man nur $\nabla^2 f \succeq 0 (\succ 0)$ durch $\nabla^2 f \preceq 0 (\prec 0)$ ersetzen, der Rest bleibt gleich.
- Stationäre Punkte sind entweder lokale Minima, lokale Maxima oder Sattelpunkte (manche Richtungen führen aufwärts, manche abwärts).
- Alle Optimierungsverfahren versuchen eine Folge zu erzeugen, die gegen einen stationären Punkt konvergiert. In der Nähe eines stationären Punktes soll die Konvergenz möglichst quadratisch sein, wie beim Newton-Verfahren.

Inhaltsübersicht für heute:

Das Newton-Verfahren

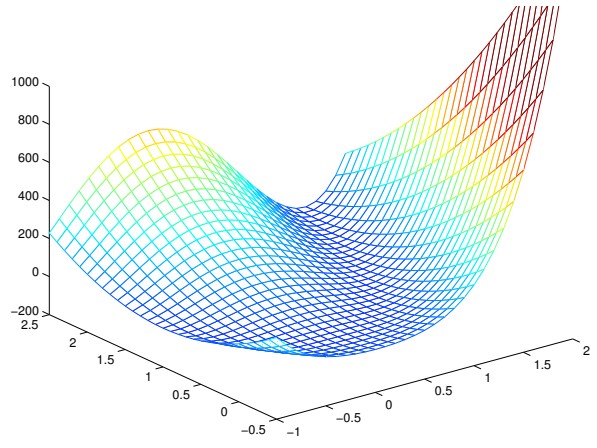
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

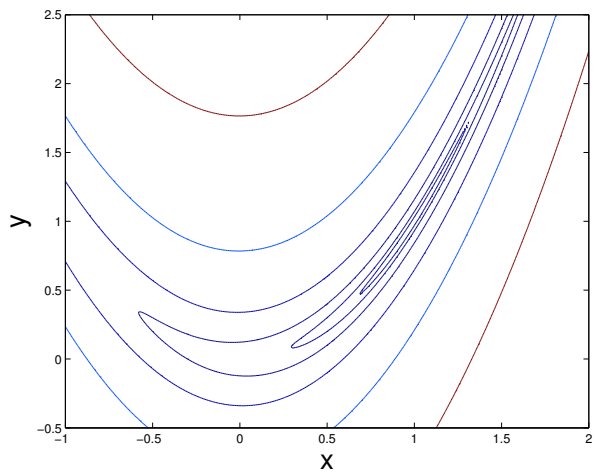
Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$





Das Newton-Verfahren

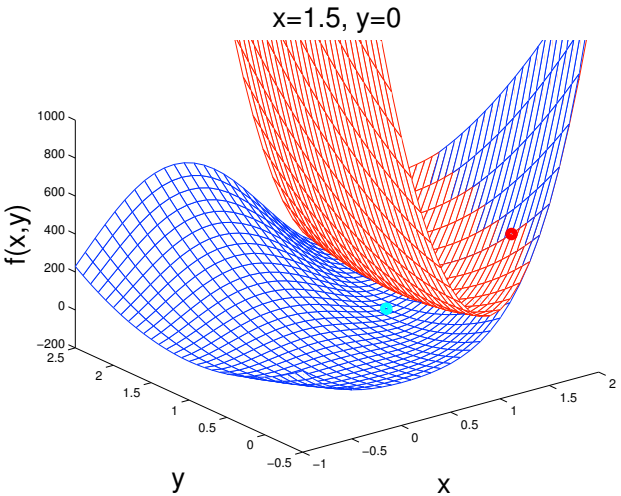
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

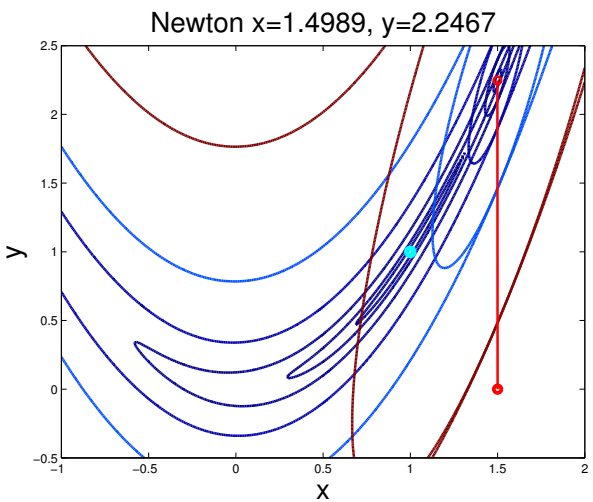
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Das Newton-Verfahren

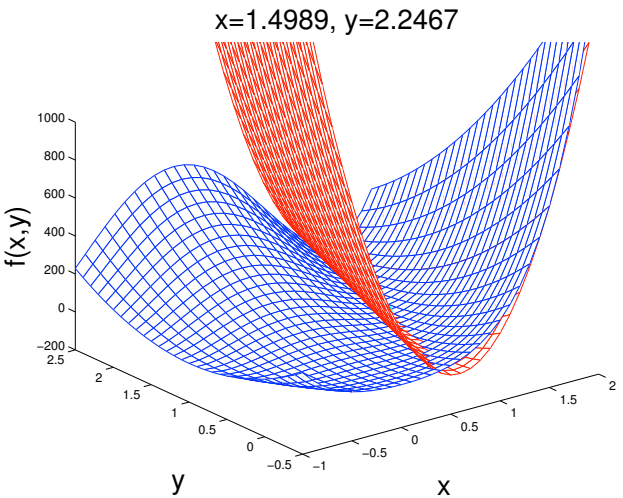
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

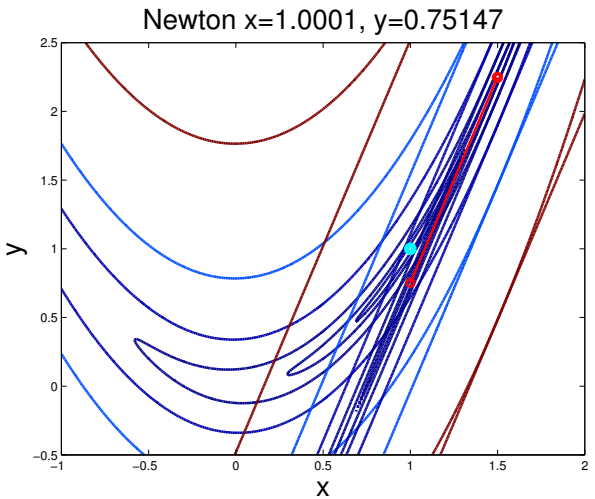
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Das Newton-Verfahren

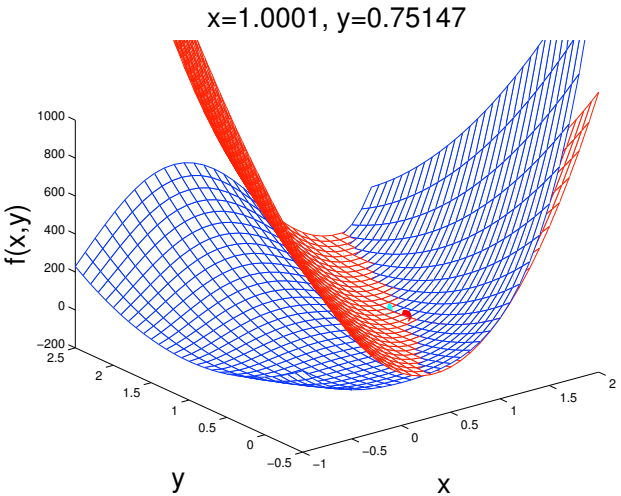
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

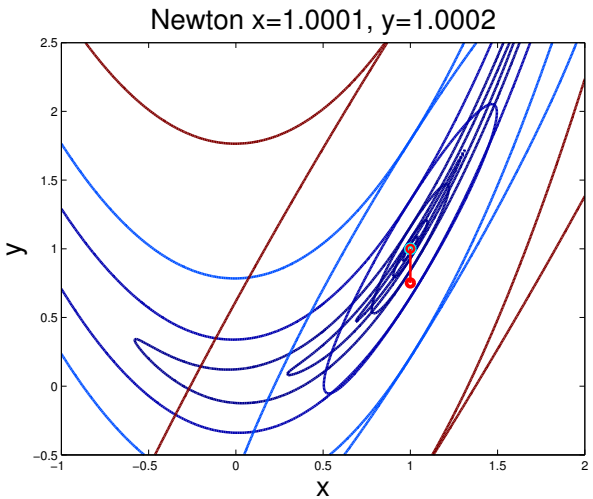
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Das Newton-Verfahren

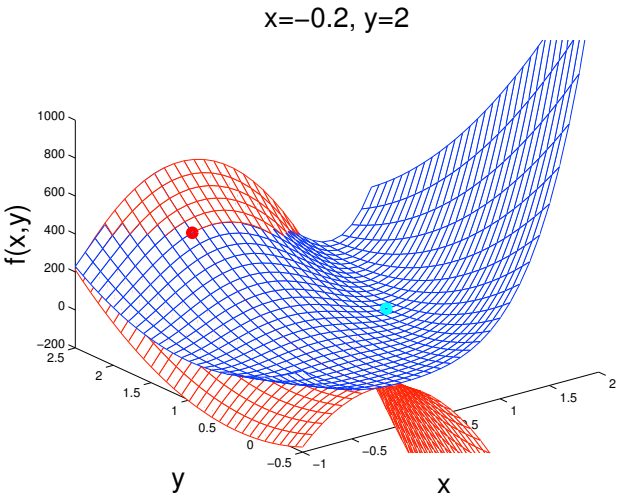
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

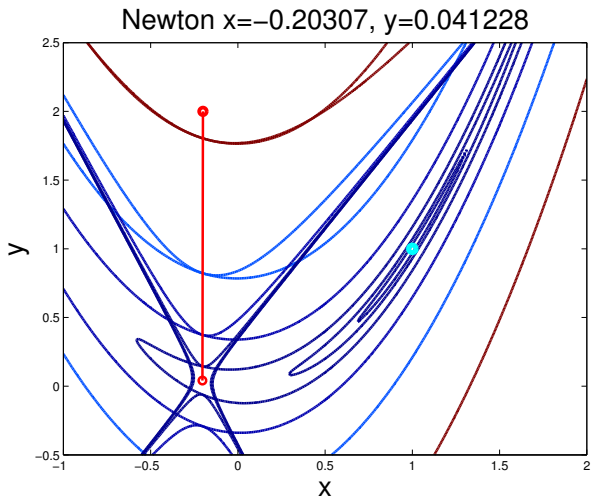
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Das Newton-Verfahren

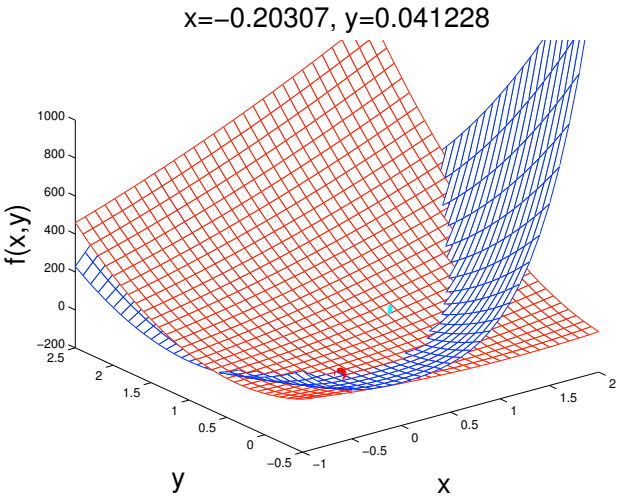
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

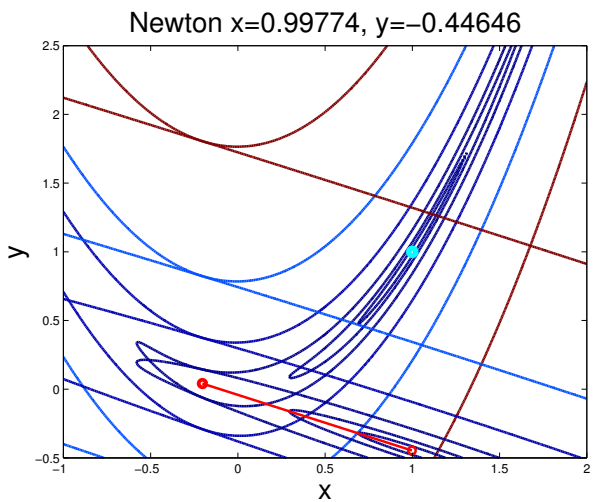
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Das Newton-Verfahren

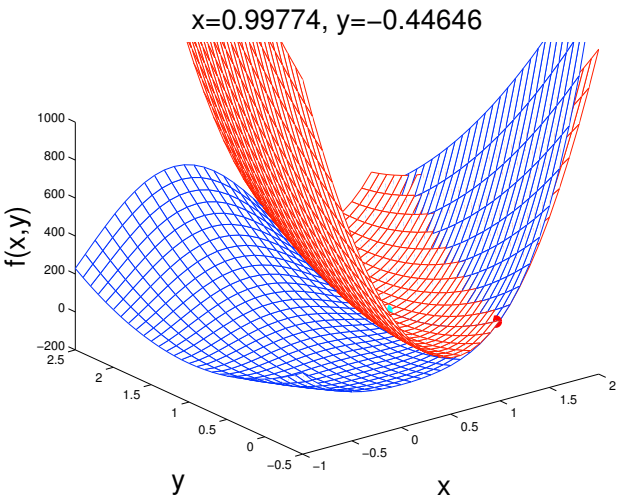
[Eigentlich sucht es Nullstellen von Funktionen $F : \mathbb{R}^n \to \mathbb{R}^m$ (s. später), hier suchen wir ein x mit $\nabla f(x) = 0$.]

Ist x^* ein lokales Minimum mit $\nabla^2 f(x^*) \succ 0$ und ändert sich $\nabla^2 f$ nicht zu schnell ($\nabla^2 f$ Lipschitz-stetig), gilt $\nabla^2 f(x) \succ 0$ für alle x nahe bei x^*.

Für jedes $x^{(k)}$ nahe bei x^* ist dann das quadratische Modell streng konvex,

$$f(x^{(k)}) + \nabla f(x^{(k)})^T h + \tfrac{1}{2} h^T \nabla^2 f(x^{(k)}) h$$

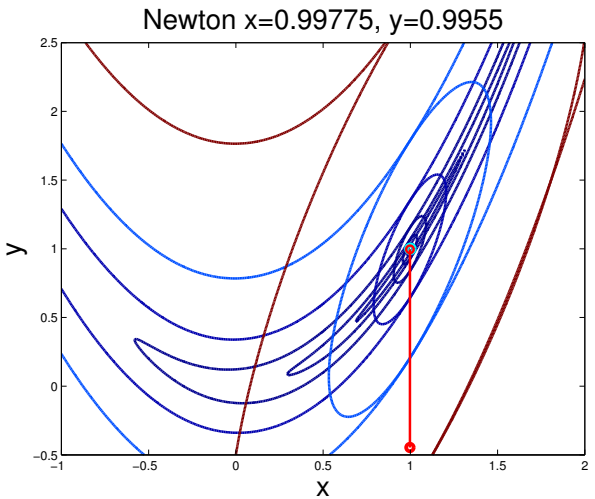
$$[= c + q^T h + \tfrac{1}{2} h^T Q h]$$



Das Newton-Verfahren wählt als nächsten Punkt $x^{(k+1)} = x^{(k)} + h$ den stationären Punkt des quadratischen Modells (= das Minimum falls $\nabla^2 f(x^{(k)}) \succ 0$),

$$h_N^{(k)} := -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}). \quad [= -Q^{-1} q]$$

$h_N^{(k)}$ ist der Newton-Schritt und ist für $\nabla^2 f(x^{(k)})$ regulär definiert.



Satz (lokal-quadratische Konvergenz des Newton-Verfahrens)

Sei f zweimal stetig differenzierbar, x^* ein lokales Minimum, das die hinreichenden Optimalitätsbedingungen erfüllt, $\nabla^2 f$ sei Lipschitz-stetig in einer Umgebung von x^*. Für jeden nahe genug an x^* gelegenen Startpunkt $x^{(0)}$ gilt für die Folge

$$x^{(k+1)} := x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

1. Die $x^{(k)}$ konvergieren quadratisch gegen x^*, d.h.,

$$\exists K \in \mathbb{N}, c > 0 : \|x^{(k+1)} - x^*\| \leq c\|x^{(k)} - x^*\|^2 \text{ für } k > K.$$

2. Die Gradienten-Normen $\|\nabla f(x^{(k)})\|$ konvergieren quadratisch gegen 0.

Satz (lokal-quadratische Konvergenz des Newton-Verfahrens)

Sei f zweimal stetig differenzierbar, x^* ein lokales Minimum, das die hinreichenden Optimalitätsbedingungen erfüllt, $\nabla^2 f$ sei Lipschitz-stetig in einer Umgebung von x^*. Für jeden nahe genug an x^* gelegenen Startpunkt $x^{(0)}$ gilt für die Folge

$$x^{(k+1)} := x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

1. Die $x^{(k)}$ konvergieren quadratisch gegen x^*, d.h.,

$$\exists K \in \mathbb{N}, c > 0 : \|x^{(k+1)} - x^*\| \leq c\|x^{(k)} - x^*\|^2 \text{ für } k > K.$$

2. Die Gradienten-Normen $\|\nabla f(x^{(k)})\|$ konvergieren quadratisch gegen 0.

- Der Satz gilt nur lokal und gibt keine Konvergenzgarantie für weit entfernte Startpunkte (das kann selbst für konvexes f fehlschlagen).
- Die Funktionswerte $f(x^{(k)})$ müssen keineswegs monoton fallen.
- Startet die Folge in der Nähe eines anderen stationären Punktes (Maximum oder Sattelpunkt), konvergiert die Folge zu diesem.
- Kommt man ins Gebiet quadratischer Konvergenz, verdoppelt sich pro Iteration die Anzahl korrekt berechneter Stellen.

Bemerkungen

- Um von lokalen zu „globalen“ Minimierungsverfahren zu kommen, wird in Globalisierungsstrategien $f(x^{(k+1)}) < f(x^{(k)})$ gefordert, siehe z.B. Line-Search- und Trust-Region-Verfahren.

Bemerkungen

- Um von lokalen zu „globalen“ Minimierungsverfahren zu kommen, wird in Globalisierungsstrategien $f(x^{(k+1)}) < f(x^{(k)})$ gefordert, siehe z.B. Line-Search- und Trust-Region-Verfahren.
- Die Bestimmung von $\nabla^2 f$ ist oft zu aufwendig bzgl. Rechenzeit und Speicherbedarf. Meist setzen Verfahren daher nur Orakel 1. Ordnung voraus und approximieren $\nabla^2 f$ lokal zur Konvergenzverbesserung. Damit sind aber die Bedingungen 2. Ordnung nicht gut überprüfbar.

Bemerkungen

- Um von lokalen zu „globalen“ Minimierungsverfahren zu kommen, wird in Globalisierungsstrategien $f(x^{(k+1)}) < f(x^{(k)})$ gefordert, siehe z.B. Line-Search- und Trust-Region-Verfahren.
- Die Bestimmung von $\nabla^2 f$ ist oft zu aufwendig bzgl. Rechenzeit und Speicherbedarf. Meist setzen Verfahren daher nur Orakel 1. Ordnung voraus und approximieren $\nabla^2 f$ lokal zur Konvergenzverbesserung. Damit sind aber die Bedingungen 2. Ordnung nicht gut überprüfbar.
- Ein nichtlineares Optimierungsverfahren heißt global konvergent, wenn es für jede nach unten beschränkte Funktion und jeden Startpunkt $x^{(0)}$ eine Punktfolge $x^{(k)}$ mit $\|\nabla f(x^{(k)})\| \to 0$ erzeugt. [Das kann auch $\|x^{(k)}\| \to \infty$ bedeuten, etwa für $f(x) = \frac{1}{x}$.]

Bemerkungen

- Um von lokalen zu „globalen“ Minimierungsverfahren zu kommen, wird in Globalisierungsstrategien $f(x^{(k+1)}) < f(x^{(k)})$ gefordert, siehe z.B. Line-Search- und Trust-Region-Verfahren.
- Die Bestimmung von $\nabla^2 f$ ist oft zu aufwendig bzgl. Rechenzeit und Speicherbedarf. Meist setzen Verfahren daher nur Orakel 1. Ordnung voraus und approximieren $\nabla^2 f$ lokal zur Konvergenzverbesserung. Damit sind aber die Bedingungen 2. Ordnung nicht gut überprüfbar.
- Ein nichtlineares Optimierungsverfahren heißt global konvergent, wenn es für jede nach unten beschränkte Funktion und jeden Startpunkt $x^{(0)}$ eine Punktfolge $x^{(k)}$ mit $\|\nabla f(x^{(k)})\| \to 0$ erzeugt. [Das kann auch $\|x^{(k)}\| \to \infty$ bedeuten, etwa für $f(x) = \frac{1}{x}$.]
- Durch die Bedingung $f(x^{(k+1)}) < f(x^{(k)})$ hoffen die Verfahren im Konvergenzfall ein Minimum gefunden zu haben, manchmal ist es jedoch ein Sattelpunkt. Für den Anwender ist das meist leicht zu erkennen, für das Verfahren nicht $\to$ besseren Startpunkt wählen.

Bemerkungen

- Um von lokalen zu „globalen“ Minimierungsverfahren zu kommen, wird in Globalisierungsstrategien $f(x^{(k+1)}) < f(x^{(k)})$ gefordert, siehe z.B. Line-Search- und Trust-Region-Verfahren.
- Die Bestimmung von $\nabla^2 f$ ist oft zu aufwendig bzgl. Rechenzeit und Speicherbedarf. Meist setzen Verfahren daher nur Orakel 1. Ordnung voraus und approximieren $\nabla^2 f$ lokal zur Konvergenzverbesserung. Damit sind aber die Bedingungen 2. Ordnung nicht gut überprüfbar.
- Ein nichtlineares Optimierungsverfahren heißt global konvergent, wenn es für jede nach unten beschränkte Funktion und jeden Startpunkt $x^{(0)}$ eine Punktfolge $x^{(k)}$ mit $\|\nabla f(x^{(k)})\| \to 0$ erzeugt. [Das kann auch $\|x^{(k)}\| \to \infty$ bedeuten, etwa für $f(x) = \frac{1}{x}$.]
- Durch die Bedingung $f(x^{(k+1)}) < f(x^{(k)})$ hoffen die Verfahren im Konvergenzfall ein Minimum gefunden zu haben, manchmal ist es jedoch ein Sattelpunkt. Für den Anwender ist das meist leicht zu erkennen, für das Verfahren nicht $\to$ besseren Startpunkt wählen.
- Wir nutzen die Kurzschreibweise f_k für $f(x^{(k)})$, ∇f_k für $\nabla f(x^{(k)})$ und $\nabla^2 f_k$ für $\nabla^2 f(x^{(k)})$.

Inhaltsübersicht für heute:

Line-Search-Verfahren

Schematischer Ablauf von Line-Search-Verfahren:

1. Rufe das Orakel für $x^{(k)}$ auf $\rightarrow f_k$, ∇f_k, (vielleicht auch $\nabla^2 f_k$).
2. Ist $\|\nabla f_k\|$ klein genug, STOP.
3. Abstiegsrichtung: Wähle $h^{(k)} \in \mathbb{R}^n$ mit $\nabla f_k^T h^{(k)} < 0$.
4. Line-Search: Finde eine Schrittweite $\alpha_k \geq 0$ mit $f(x^{(k)} + \alpha_k h^{(k)})$ „ausreichend“ kleiner als f_k
5. Setze $x^{(k+1)} := x^{(k)} + \alpha_k h^{(k)}$, $k \leftarrow k+1$, gehe zu 1.

Zwei Hauptaufgaben:

- Bestimmung einer Abstiegsrichtung
- Bestimmung einer Schrittweite (Line-Search)

Abstiegsrichtung (für $\bar{x}$ mit $\nabla f(\bar{x}) \neq 0$)

Eine Richtung $h \in \mathbb{R}^n$ heißt Abstiegsrichtung für f in $\bar{x}$, falls $\nabla f(\bar{x})^T h < 0$.

Abstiegsrichtung (für $\bar{x}$ mit $\nabla f(\bar{x}) \neq 0$)

Eine Richtung $h \in \mathbb{R}^n$ heißt Abstiegsrichtung für f in $\bar{x}$, falls $\nabla f(\bar{x})^T h < 0$.

Die meisten Algorithmen bestimmen h für ein $B \succ 0$ in der Form

$$h := -B^{-1}\nabla f(\bar{x}), \quad \text{denn } \nabla f(\bar{x})^T h = -\underbrace{\nabla f(\bar{x})^T B^{-1} \nabla f(\bar{x})}_{>0} < 0.$$

Beispiele (s. später zu Vor- und Nachteilen):

- $B = I$: steilster Abstieg $h = -\nabla f(\bar{x})$ (steepest descent).
- $B = \nabla^2 f(\bar{x})$ Newton-Richtung (Abstiegsrichtung für $\nabla^2 f(\bar{x}) \succ 0$)
- $B = [\nabla^2 f(\bar{x}) + \lambda I] \succ 0$ modifizierte Newton-Richtung
- $B \succ 0$ als Approximation von $\nabla^2 f(\bar{x})$ Quasi-Newton-Richtung

Abstiegsrichtung (für $\bar{x}$ mit $\nabla f(\bar{x}) \neq 0$)

Eine Richtung $h \in \mathbb{R}^n$ heißt Abstiegsrichtung für f in $\bar{x}$, falls $\nabla f(\bar{x})^T h < 0$.

Die meisten Algorithmen bestimmen h für ein $B \succ 0$ in der Form

$$h := -B^{-1}\nabla f(\bar{x}), \quad \text{denn } \nabla f(\bar{x})^T h = -\underbrace{\nabla f(\bar{x})^T B^{-1} \nabla f(\bar{x})}_{>0} < 0.$$

Beispiele (s. später zu Vor- und Nachteilen):

- $B = I$: steilster Abstieg $h = -\nabla f(\bar{x})$ (steepest descent).
- $B = \nabla^2 f(\bar{x})$ Newton-Richtung (Abstiegsrichtung für $\nabla^2 f(\bar{x}) \succ 0$)
- $B = [\nabla^2 f(\bar{x}) + \lambda I] \succ 0$ modifizierte Newton-Richtung
- $B \succ 0$ als Approximation von $\nabla^2 f(\bar{x})$ Quasi-Newton-Richtung

---

Für globale Konvergenz der Line-Search Verfahren ist nur wichtig, dass die Richtungen nicht orthogonal zur steilsten Abstiegsrichtung werden:

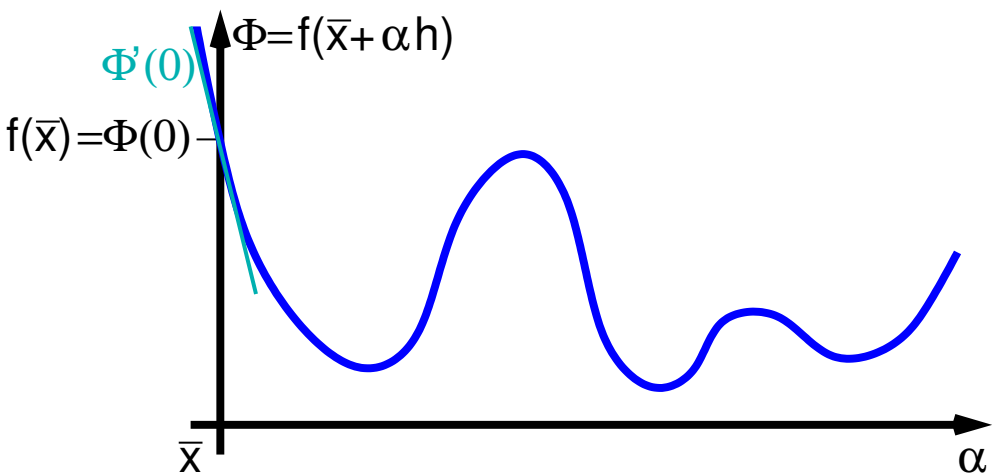
$$\exists \delta > 0 : \frac{-\nabla f_k}{\|\nabla f_k\|}^T \frac{h^{(k)}}{\|h^{(k)}\|} = \cos \angle(-\nabla f_k, h^{(k)}) \geq \delta > 0 \text{ für } k > 0.$$

Das ist erfüllt, falls $\frac{\lambda_{\max}(B_k)}{\lambda_{\min}(B_k)} < \kappa$ für ein $\kappa > 0$ und gilt z.B. für $B = I$ oder Newton-Richtung in der Nähe von x^* unter den Vor. des Newton-Satzes.

Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

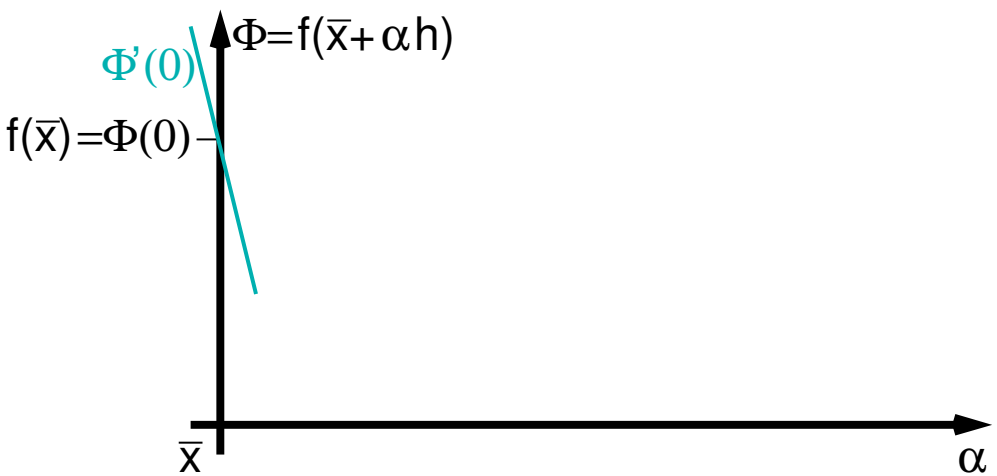


Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

Anfangs sehr wenig Information: $\Phi(0) = f(\bar{x})$, Ableitung $\Phi'(0) = \nabla f(\bar{x})^T h$



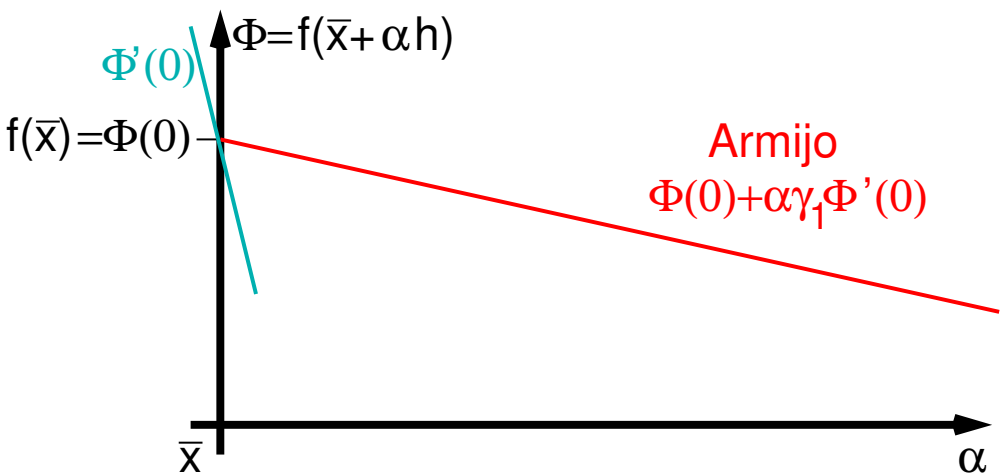
Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

Anfangs sehr wenig Information: $\Phi(0) = f(\bar{x})$, Ableitung $\Phi'(0) = \nabla f(\bar{x})^T h$
Ein $\bar{\alpha}$ mit ausreichendem Abstieg (sufficient decrease) erfüllt:

1. Mindestanteil $0 < \gamma_1 < 1$ an dem durch $\Phi'(0)$ „versprochenen“Abstieg:
   $\Phi(\bar{\alpha}) \leq \Phi(0) + \bar{\alpha}\gamma_1\Phi'(0)$ (Armijo-Bedingung) [für kleine α erfüllt]



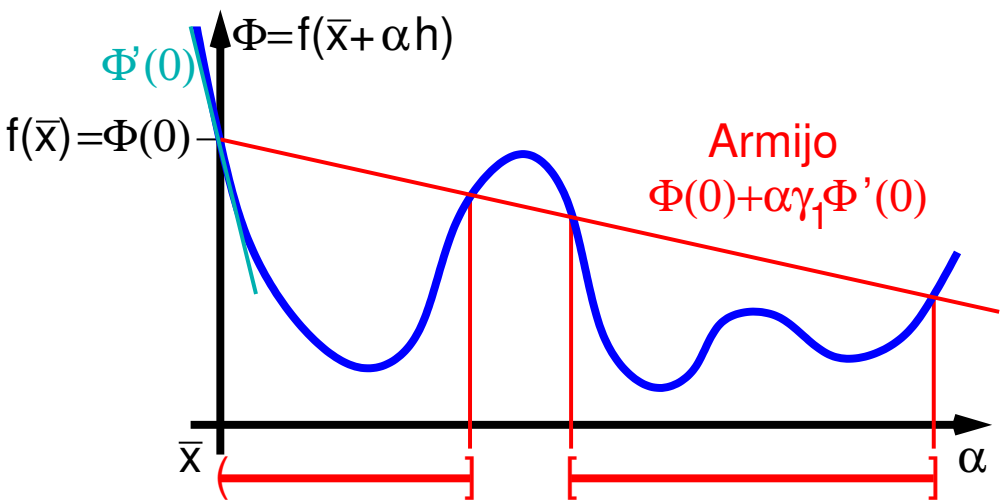
Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

Anfangs sehr wenig Information: $\Phi(0) = f(\bar{x})$, Ableitung $\Phi'(0) = \nabla f(\bar{x})^T h$
Ein $\bar{\alpha}$ mit ausreichendem Abstieg (sufficient decrease) erfüllt:

1. Mindestanteil $0 < \gamma_1 < 1$ an dem durch $\Phi'(0)$ „versprochenen“Abstieg:
   $\Phi(\bar{\alpha}) \leq \Phi(0) + \bar{\alpha}\gamma_1\Phi'(0)$ (Armijo-Bedingung) [für kleine α erfüllt]



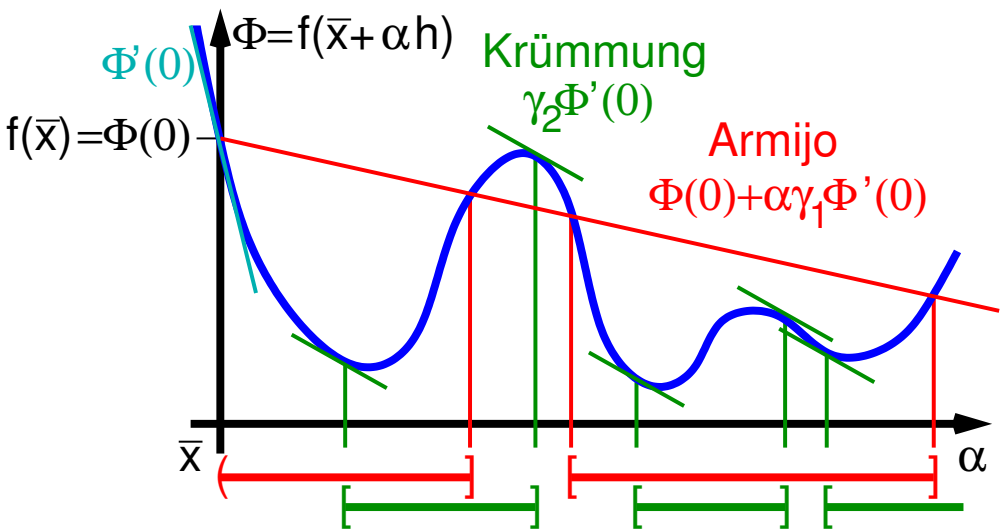
Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

Anfangs sehr wenig Information: $\Phi(0) = f(\bar{x})$, Ableitung $\Phi'(0) = \nabla f(\bar{x})^T h$
Ein $\bar{\alpha}$ mit ausreichendem Abstieg (sufficient decrease) erfüllt:

1. Mindestanteil $0 < \gamma_1 < 1$ an dem durch $\Phi'(0)$ „versprochenen“Abstieg:
   $\Phi(\bar{\alpha}) \leq \Phi(0) + \bar{\alpha}\gamma_1\Phi'(0)$ (Armijo-Bedingung) [für kleine α erfüllt]
2. An der Stelle $\bar{\alpha}$ ist der Abstieg Φ' schlecht $(0 < \gamma_1 < \gamma_2 < 1)$:
   $\Phi'(\bar{\alpha}) \geq \gamma_2\Phi'(0)$ (Krümmungs-Bedingung) [$\nabla f^T h$ stark geändert]



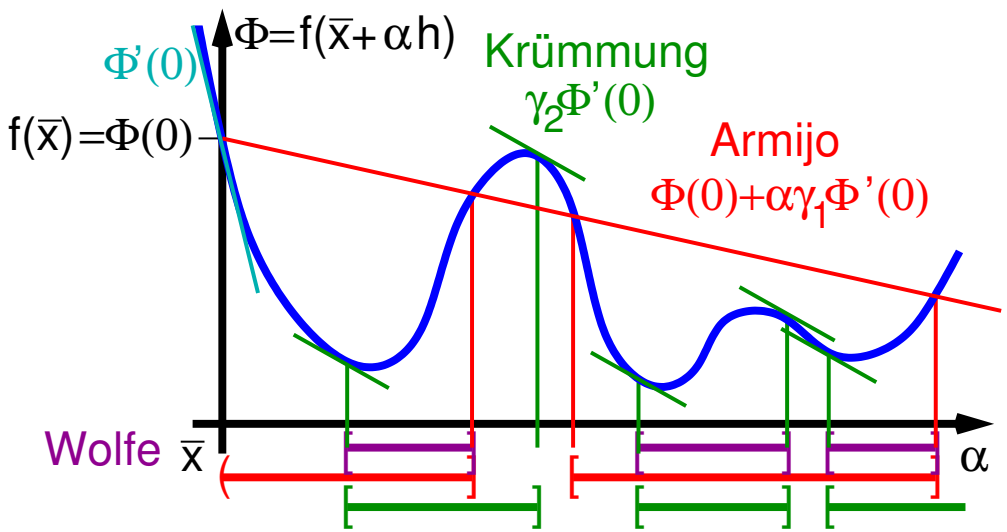
Line-Search für Abstiegsrichtung h

Bestimme Schrittweite $\bar{\alpha} \geq 0$ als **Näherung** zu $\min_{\alpha \geq 0} \Phi(\alpha) := f(\bar{x} + \alpha h)$.

Berechnung von $\bar{\alpha} \in \operatorname{Argmin}_{\alpha \geq 0} \Phi(\alpha)$ (exakter Line-Search) wäre sinnlos aufwendig, da die Richtung h meist weit am Optimum vorbeiführt.

Anfangs sehr wenig Information: $\Phi(0) = f(\bar{x})$, Ableitung $\Phi'(0) = \nabla f(\bar{x})^T h$
Ein $\bar{\alpha}$ mit ausreichendem Abstieg (sufficient decrease) erfüllt:

1. Mindestanteil $0 < \gamma_1 < 1$ an dem durch $\Phi'(0)$ „versprochenen“Abstieg:
   $\Phi(\bar{\alpha}) \leq \Phi(0) + \bar{\alpha}\gamma_1\Phi'(0)$ (Armijo-Bedingung) [für kleine α erfüllt]
2. An der Stelle $\bar{\alpha}$ ist der Abstieg Φ' schlecht $(0 < \gamma_1 < \gamma_2 < 1)$:
   $\Phi'(\bar{\alpha}) \geq \gamma_2\Phi'(0)$ (Krümmungs-Bedingung) [$\nabla f^T h$ stark geändert]



Armijo- und Krümmungs-Bedingung gemeinsam heißen Wolfe-Bedingungen und Schrittweiten, die diese erfüllen, garantieren ausreichenden Abstieg.
$[\gamma_1 = 10^{-4}, \gamma_2 \in \{0.1, 0.9\}]$

Wolfe-Bedingungen und globale Konvergenz

Für $0 < \gamma_1 < \gamma_2 < 1$ erfüllt Schrittweite α_k die Wolfe-Bedingungen, wenn

$$\begin{array}{rcll} f(x^{(k)} + \alpha_k h^{(k)}) & \leq & f_k + \alpha_k \gamma_1 \nabla f_k^T h^{(k)} & \text{(Armijo)} \\ \nabla f(x^{(k)} + \alpha_k h^{(k)})^T h^{(k)} & \geq & \gamma_2 \nabla f_k^T h^{(k)} & \text{(Krümmung)} \end{array}$$

Armijo sichert Abstieg, Krümmung eine Mindestschrittweite, falls ∇f Lipschitz-stetig ist. Beides ist mit einem Orakel 1. Ordnung überprüfbar. Solche Schrittweiten gibt es immer, wenn f nach unten beschränkt ist.

Wolfe-Bedingungen und globale Konvergenz

Für $0 < \gamma_1 < \gamma_2 < 1$ erfüllt Schrittweite α_k die Wolfe-Bedingungen, wenn

$$\begin{array}{rcll} f(x^{(k)} + \alpha_k h^{(k)}) & \leq & f_k + \alpha_k \gamma_1 \nabla f_k^T h^{(k)} & \text{(Armijo)} \\ \nabla f(x^{(k)} + \alpha_k h^{(k)})^T h^{(k)} & \geq & \gamma_2 \nabla f_k^T h^{(k)} & \text{(Krümmung)} \end{array}$$

Armijo sichert Abstieg, Krümmung eine Mindestschrittweite, falls ∇f Lipschitz-stetig ist. Beides ist mit einem Orakel 1. Ordnung überprüfbar. Solche Schrittweiten gibt es immer, wenn f nach unten beschränkt ist.

Satz (Globale Konvergenz von Line-Search-Verfahren)

Sei f nach unten beschränkt. Für den Startpunkt $x^{(0)}$ sei ∇f auf der Niveaumenge $\{x \in \mathbb{R}^n : f(x) < f_0\}$ Lipschitz-stetig. Garantiert ein Line-Search-Verfahren $-\frac{\nabla f_k^T h^{(k)}}{\|\nabla f_k\| \|h^{(k)}\|} \geq \delta$ für ein $\delta > 0$ sowie die Wolfe-Bedingungen für die Schrittweiten α_k, dann gilt $\|\nabla f_k\| \to 0$.

Wolfe-Bedingungen und globale Konvergenz

Für $0 < \gamma_1 < \gamma_2 < 1$ erfüllt Schrittweite α_k die Wolfe-Bedingungen, wenn

$$\begin{array}{rcll} f(x^{(k)} + \alpha_k h^{(k)}) & \leq & f_k + \alpha_k \gamma_1 \nabla f_k^T h^{(k)} & \text{(Armijo)} \\ \nabla f(x^{(k)} + \alpha_k h^{(k)})^T h^{(k)} & \geq & \gamma_2 \nabla f_k^T h^{(k)} & \text{(Krümmung)} \end{array}$$

Armijo sichert Abstieg, Krümmung eine Mindestschrittweite, falls ∇f Lipschitz-stetig ist. Beides ist mit einem Orakel 1. Ordnung überprüfbar. Solche Schrittweiten gibt es immer, wenn f nach unten beschränkt ist.

Satz (Globale Konvergenz von Line-Search-Verfahren)

Sei f nach unten beschränkt. Für den Startpunkt $x^{(0)}$ sei ∇f auf der Niveaumenge $\{x \in \mathbb{R}^n : f(x) < f_0\}$ Lipschitz-stetig. Garantiert ein Line-Search-Verfahren $-\frac{\nabla f_k^T h^{(k)}}{\|\nabla f_k\| \|h^{(k)}\|} \geq \delta$ für ein $\delta > 0$ sowie die Wolfe-Bedingungen für die Schrittweiten α_k, dann gilt $\|\nabla f_k\| \to 0$.

Vorsicht: Man hofft auf Konvergenz gegen ein Minimum, aber sowohl $\|x^{(k)}\| \to \infty$ als auch Konvergenz gegen einen Sattelpunkt sind nicht ausgeschlossen!

Bestimmung der Schrittweite in der Praxis

- Ziel ist, mit möglichst wenig Funktionsauswertungen einen Wolfepunkt zu finden.
- Die vorhergehende Schrittweite dient meist als Startwert, beim allerersten Mal nutzt man gerne $\alpha = \frac{1}{\|h\|}$.
- Der nächsten Kandidat wird z.B. über kubische Interpolation, die neue und alte Funktionswerte und Ableitungen nutzt, bestimmt.
- Jeder Fehl-Versuch erlaubt, das Suchintervall zu verkleinern.
- Eine solide und effiziente Implementation, die auch mit numerischen Schwierigkeiten umgehen kann, ist sehr schwer und aufwendig.
- Entscheidend für den Erfolg ist vor allem die Schrittrichtung!

AUSNAHME: Für Newton-ähnliche Richtungen wird immer Schrittweite 1 zuerst probiert und nur auf Armijo getestet. Solange Armijo nicht erfüllt ist, reduziert man die Schrittweite (durch Interpolation oder einfaches Backtracking, d.h., Multiplikation der Schrittweite mit einem Faktor $0 < \sigma < 1$).