

## • THEMA

- Lineare Regression
  - Vergleich Perzeptron
  - Bsp. Hauspreis Vorhersage
  - Kostenfunktion  $J(\omega)$  – Visuelle Intuition
  - Gradientenvektor , Gradientenabstieg
  - Polynomiale Regression
- Skalierung der Merkmale

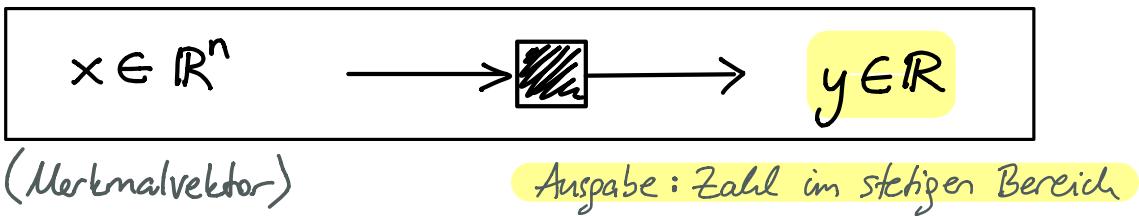
## • WICHTIG

- $\mathcal{H} = \left\{ w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n : w_i \in \mathbb{R} \right\}$
- $\Delta$  = Gradientenabstieg

## • LINEARE REGRESSION - INTRO

- Das Problem

Regression bedeutet,  
 $\hat{y}$  ist reellwertig.  
 (nach Abu-Mostafa)



- D : Die Daten sind Eingabe-Ausgabe Paare

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$$

- H : Die Hypothesen sind lineare Funktionen

$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n$$

- Linear in  $w$  !!
- Finde beste Gewichte → Finde  $h^*$

- L : Die Verlustfunktion (Loss) : Squared Error

$$L(\hat{y}, y) = (\hat{y} - y)^2$$

Fehlerquadrat

Die "Distanz" zwischen  $\hat{y}$  und  $y$

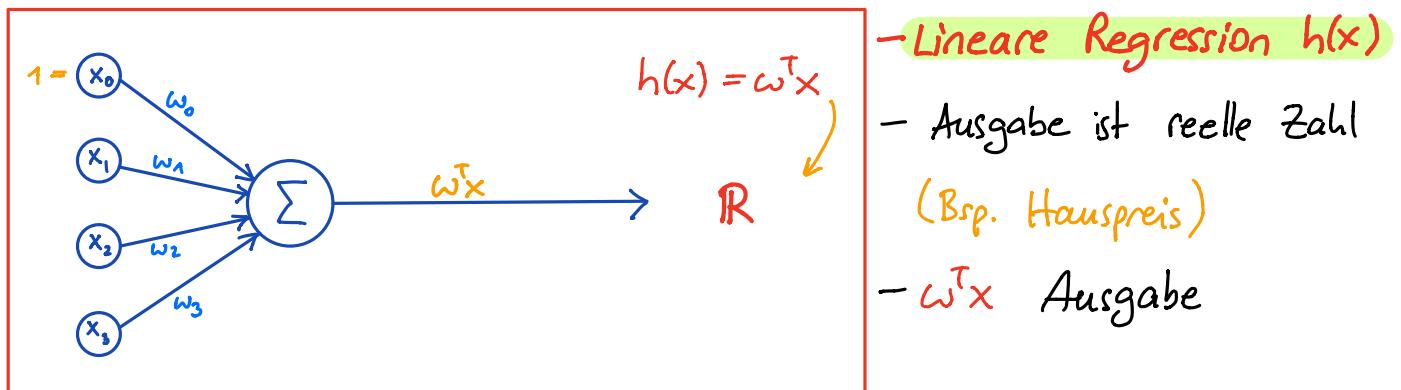
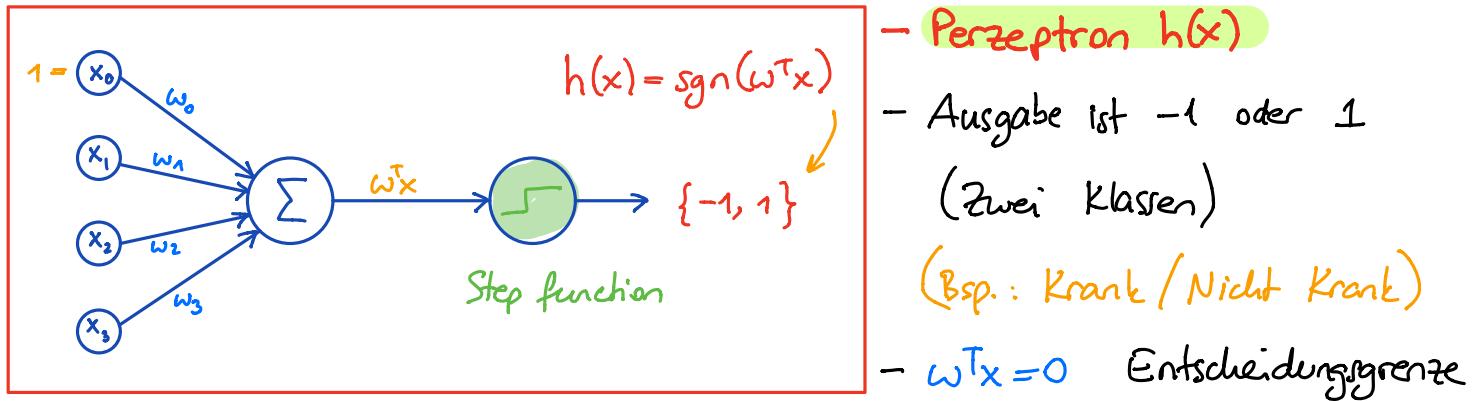
- J : Die Kostenfunktion (Cost) : Mean Squared Error (MSE)

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (h_w(x^{(i)}) - (y^{(i)}))^2$$

◦ Der durchschnittliche Fehlerquadrat

- A : Gradient Descent (Gradientenabstieg)

- "GRAPH"ISCHER VERGLEICH MIT PERZEPTRON



↑  
Input  $\vec{x}$

↑  
Output  $y$

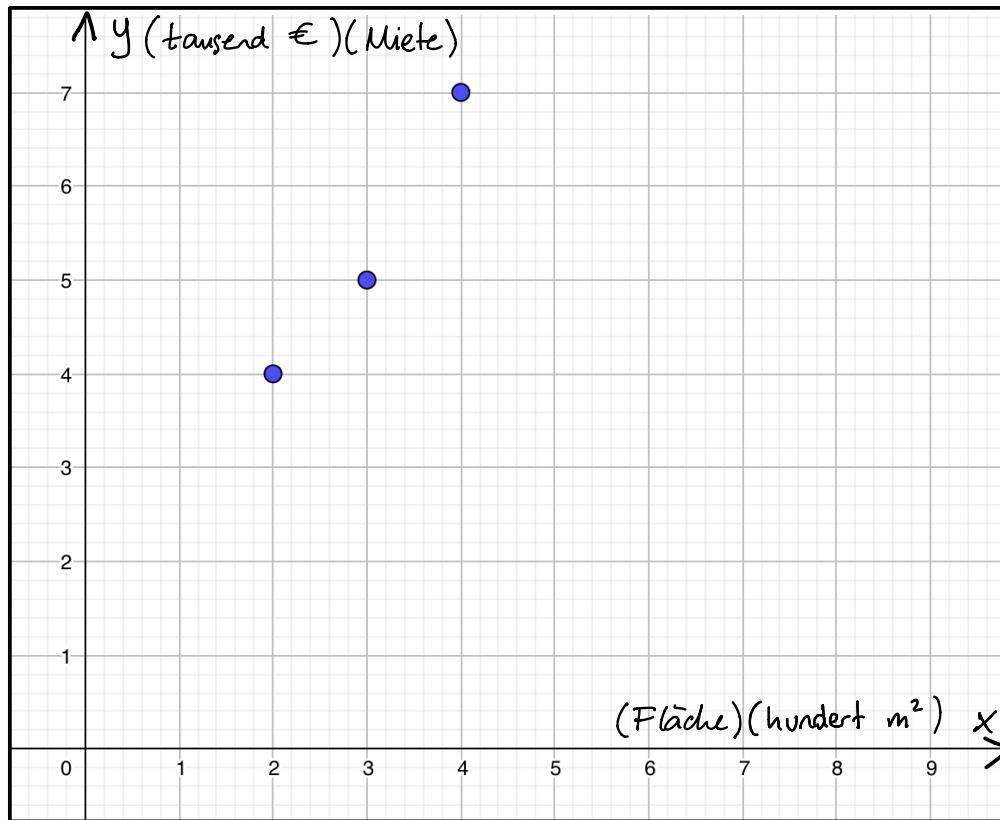
BSP.

## Hauspreis Vorhersage (ein Merkmal)

- Trainingsdaten:

$$\mathcal{D} = \{(2, 4), (3, 5), (4, 7)\}$$

Nr. i	Fläche (Hundert m <sup>2</sup> ) $x^{(i)}$	Miete (Tausend €) $y^{(i)}$
1	2	4
2	3	5
3	4	7



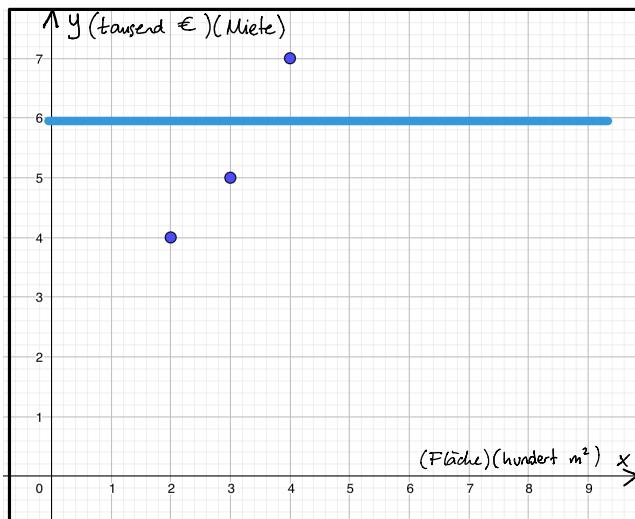
Desmos Worksheet: <https://www.desmos.com/calculator/s9yfddyz3>

- $\mathcal{H} : h_w(x) = \omega_0 + \omega_1 x$

Suche: "best fit"

- Einige zufällige Gewichtsvektoren und ihre "Performanz"  
(Je kleiner die Kosten, desto besser die Hypothesenfunktion)

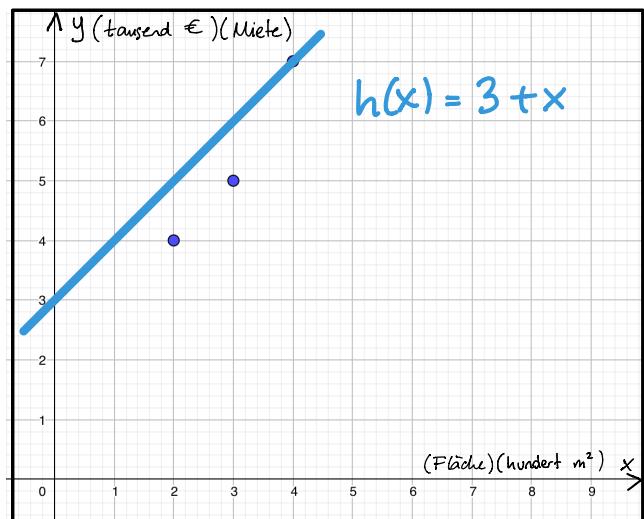
a)  $(\omega_0, \omega_1) = (6, 0)$



$$MSE = J(6, 0) = 1$$

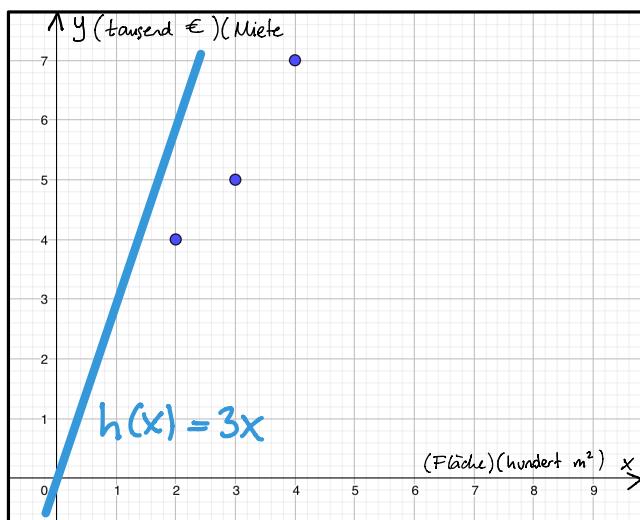
$$= \frac{1}{6} [(4-6)^2 + (5-6)^2 + (7-6)^2]$$

b)  $(\omega_0, \omega_1) = (3, 1)$



$$MSE = J(3, 1) = 0.\overline{3}$$

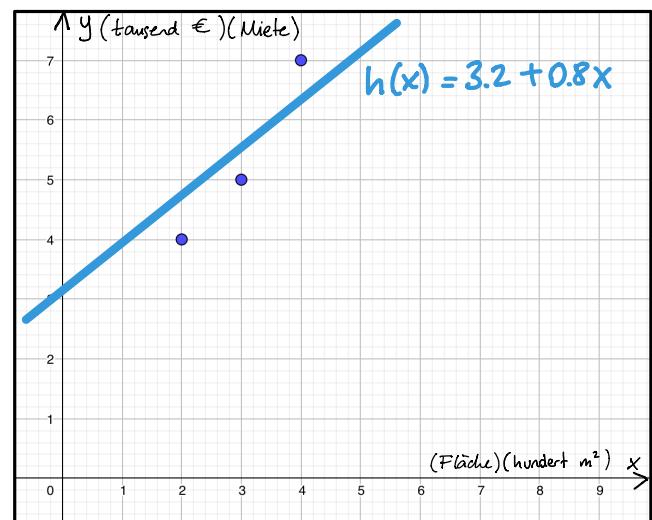
b)  $(\omega_0, \omega_1) = (0, 3)$



$$MSE = J(0, 3) = 7.5$$

$$= \frac{1}{6} [(6-4)^2 + (9-5)^2 + (12-7)^2]$$

d)  $(\omega_0, \omega_1) = (3.2, 0.8)$



$$MSE = J(3.2, 0.8) = 0.22\overline{6}$$

Minimal?

## • THE MEAN SQUARED ERROR (MSE)

Durchschnittliche Summe der Fehlerquadrate

$$J(\omega) = \frac{1}{2m} \sum_{i=1}^m (h_{\omega}(x^{(i)}) - (y^{(i)}))^2$$

b)  $(\omega_0, \omega_1) = (3, 1)$   $h(x) = 3 + x$

j	$x^{(i)}$	$y^{(i)}$	$h(x^{(i)})$	Fehler $h(x^{(i)}) - y^{(i)}$	Fehlerquadrat $(h(x^{(i)}) - y^{(i)})^2$
1	2	4	5	1	1
2	3	5	6	1	1
3	4	7	7	0	0

$$J(3, 1) = \frac{1}{6} \sum_{i=1}^3 (h(x^{(i)}) - y^{(i)})^2 = \frac{1}{6} (1+1+0) = 0.3$$

d)  $(\omega_0, \omega_1) = (3.2, 0.8)$   $h(x) = 3.2 + 0.8x$

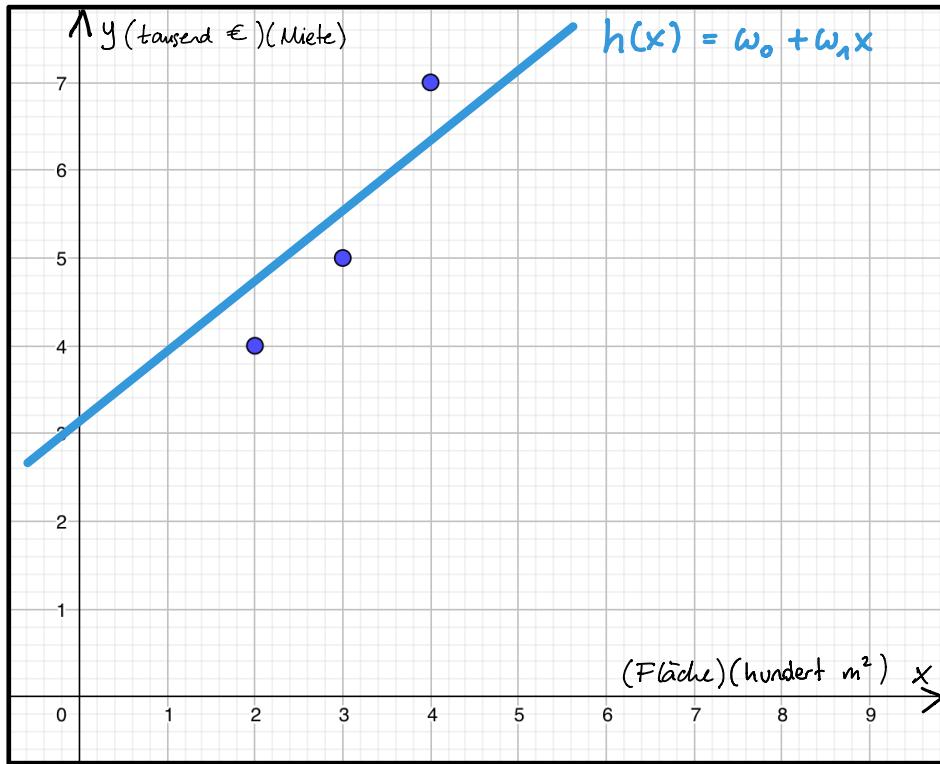
j	$x^{(i)}$	$y^{(i)}$	$h(x^{(i)})$	Fehler $h(x^{(i)}) - y^{(i)}$	Fehlerquadrat $(h(x^{(i)}) - y^{(i)})^2$
1	2	4	4.8	0.8	(0.64)
2	3	5	5.6	0.6	(0.36)
3	4	7	6.4	-0.6	(0.36)

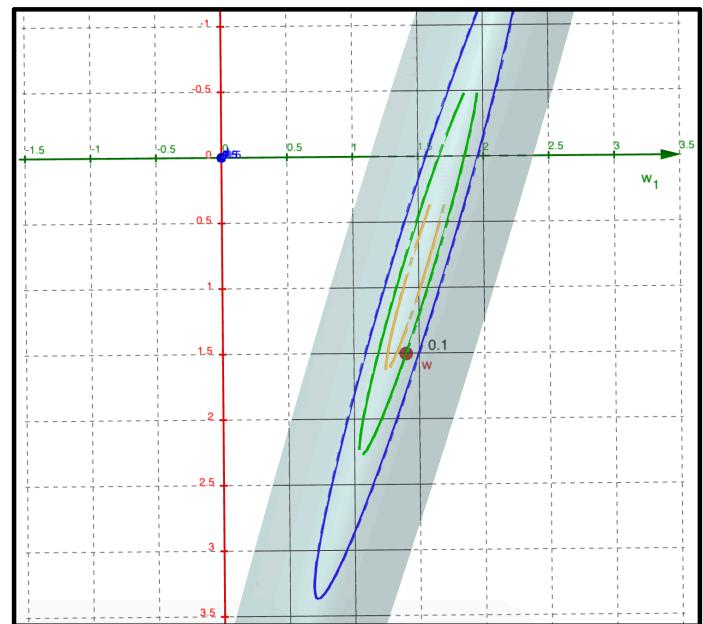
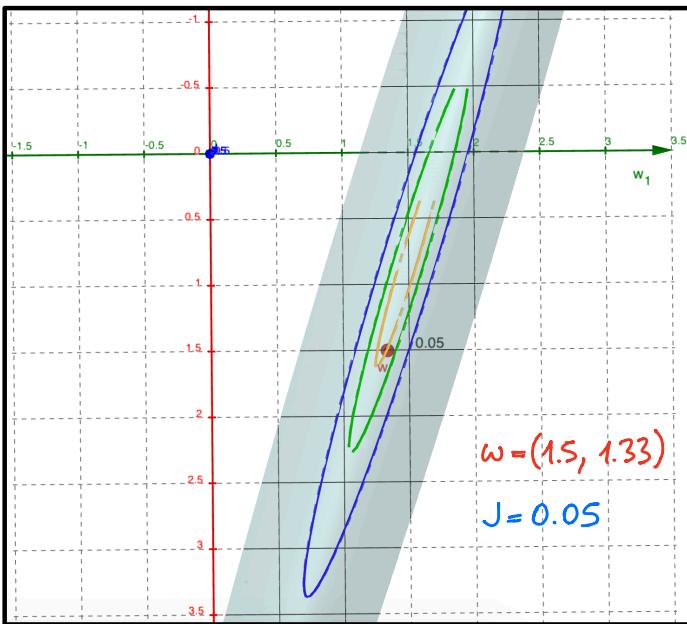
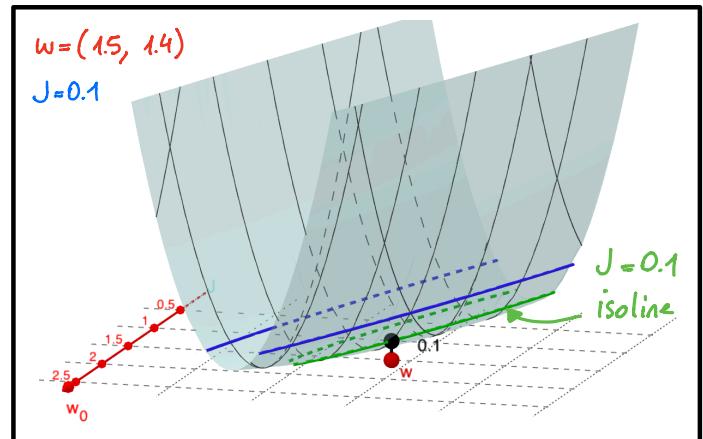
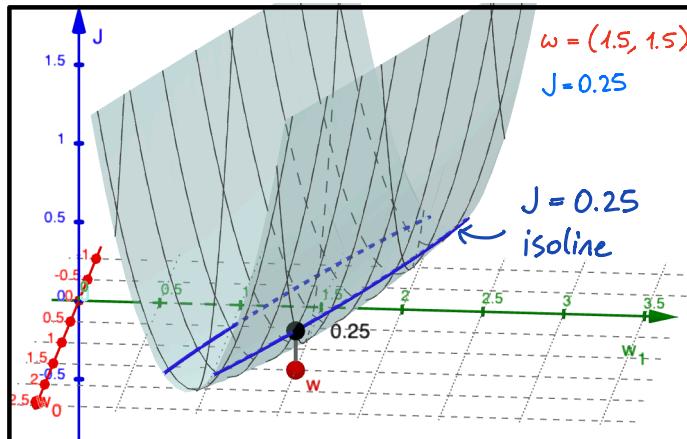
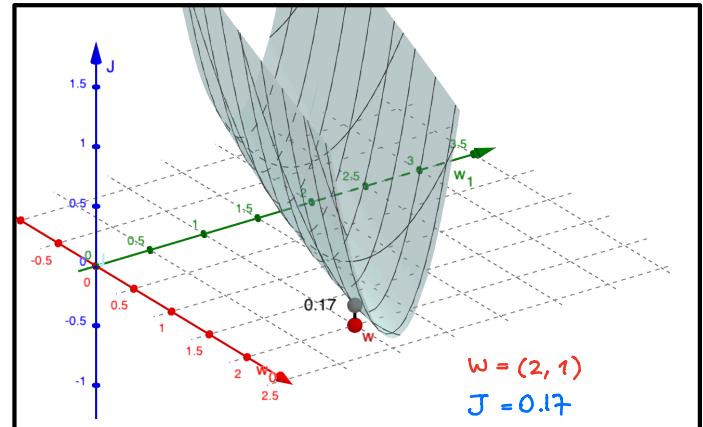
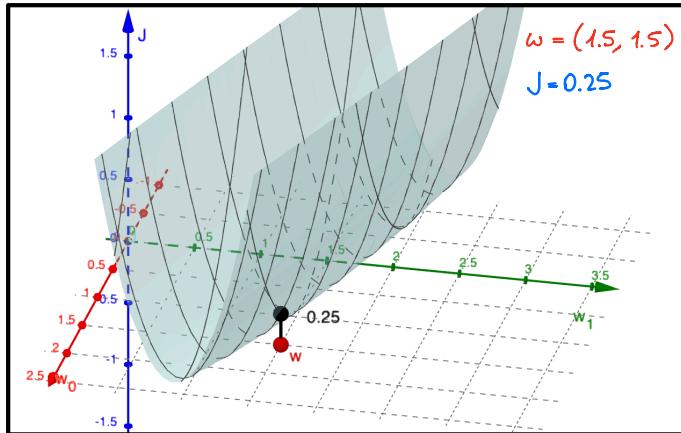
$$J(3.2, 0.8) = \frac{1}{6} \sum_{i=1}^3 (h(x^{(i)}) - y^{(i)})^2 = \frac{1}{6} (1.36) \cong 0.23$$

Minimum?

- Visuelle Intuition für  $J(\omega)$

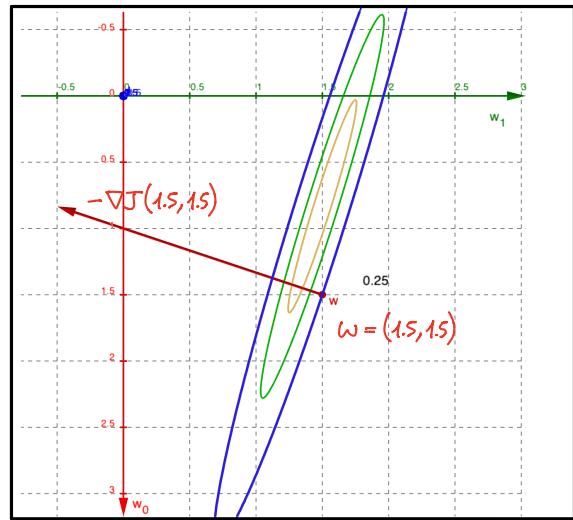
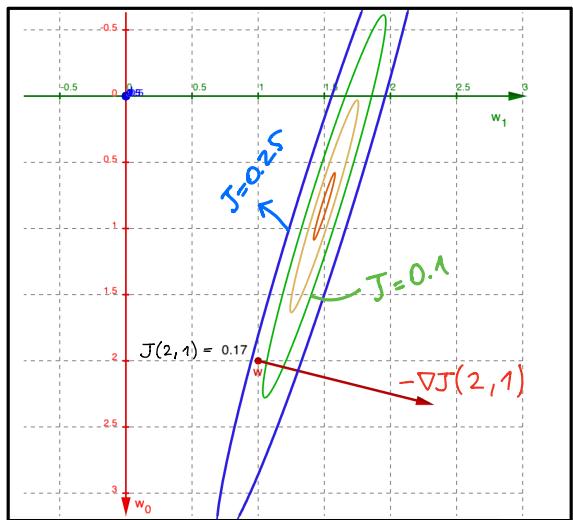
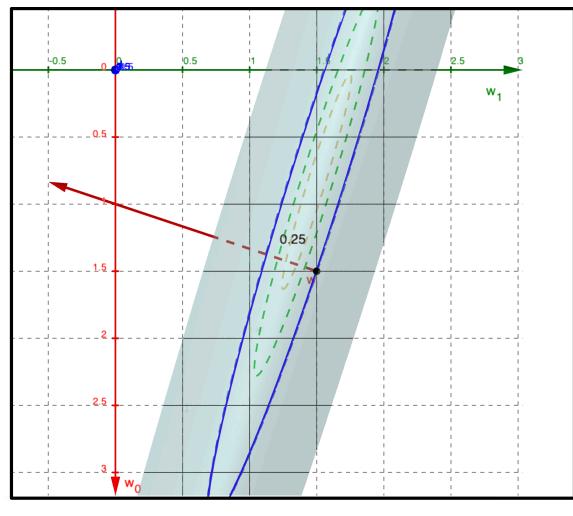
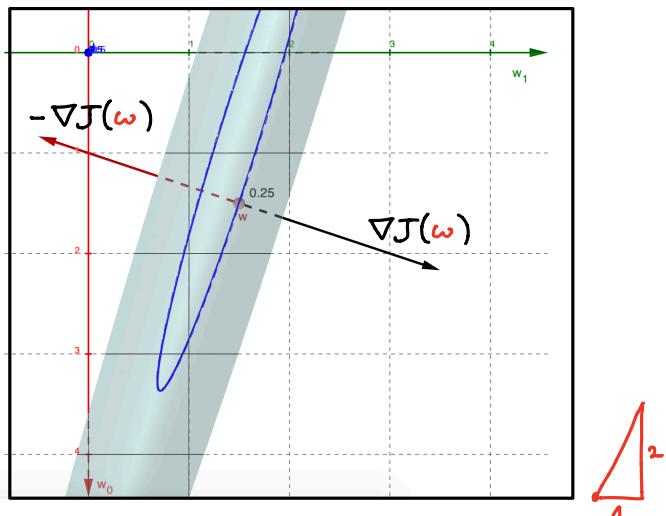
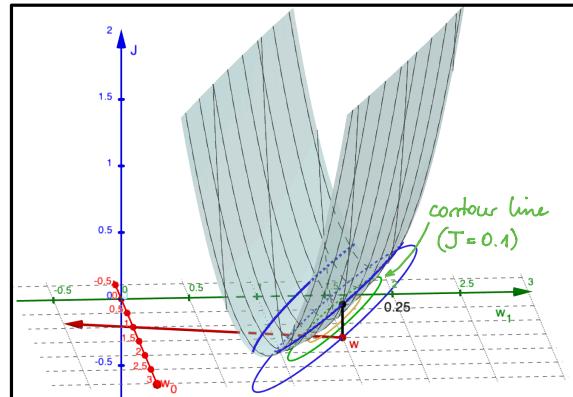
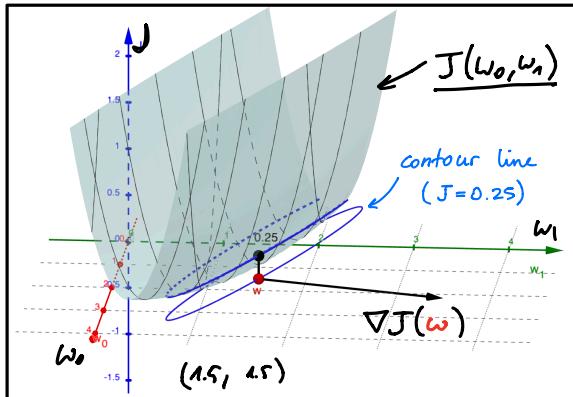
$$\begin{aligned}
 J(\omega_0, \omega_1) &= \frac{1}{6} \left[ (\omega_0 + \omega_1 x^{(1)} - y^{(1)})^2 + (\omega_0 + \omega_1 x^{(2)} - y^{(2)})^2 \right. \\
 &\quad \left. + (\omega_0 + \omega_1 x^{(3)} - y^{(3)})^2 \right] \\
 &= \frac{1}{6} \left[ (\omega_0 + 2\omega_1 - 4)^2 + (\omega_0 + 3\omega_1 - 5)^2 + (\omega_0 + 4\omega_1 - 7)^2 \right]
 \end{aligned}$$





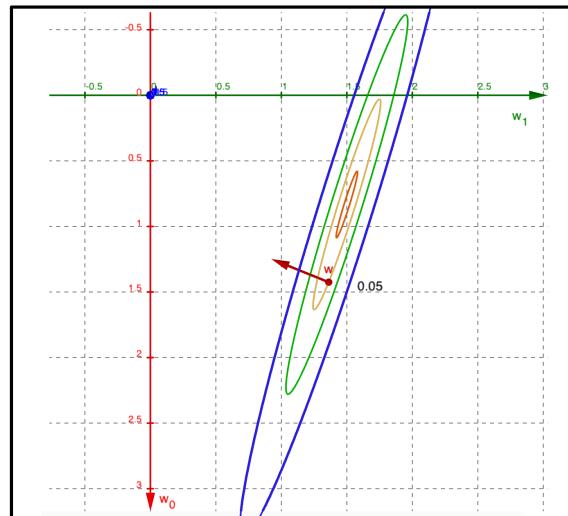
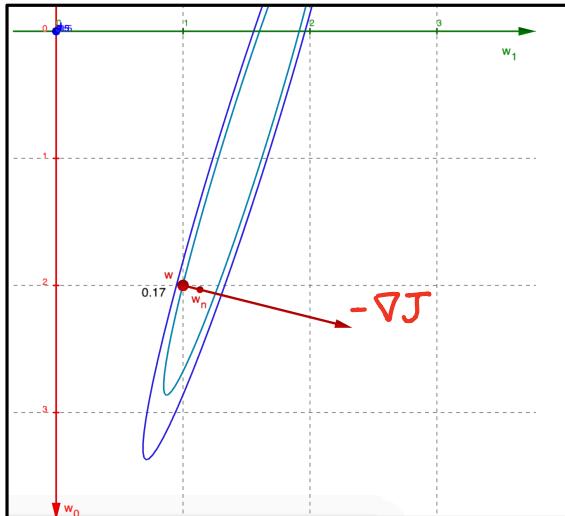
- $J$  bleibt konstant entlang einer Konturlinie (Niveau Linie / Isoline )

## • Gradientenvektor $\nabla J$ - Visuelle Intuition



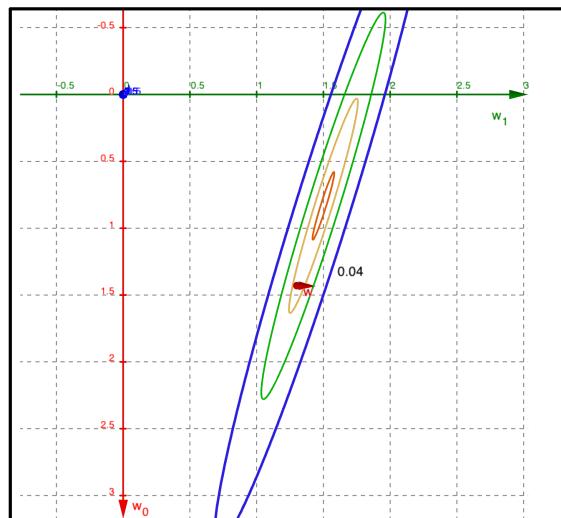
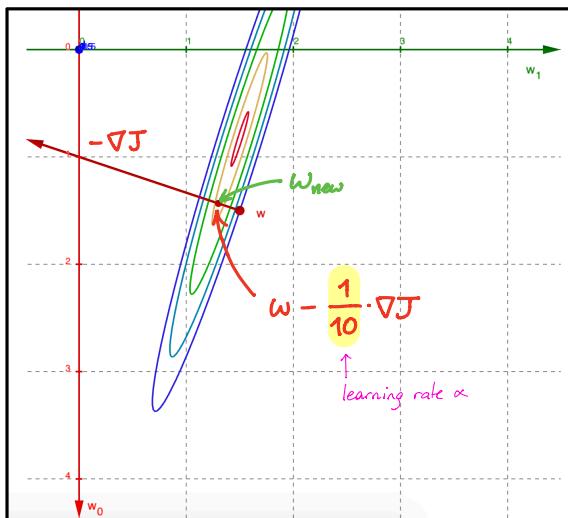
- Der Vektor  $\nabla J$  ist eine Funktion von  $w$ .
- Gradientenvektor  $\nabla J(w)$  zeigt in die Richtung des steilsten Aufstiegs, in jedem Punkt  $w$ .

- Die Steigung der Kostenfunktion  $J$  in jedem Punkt  $w$  ist
  - Null in Richtung der Konturlinie
  - maximal in Richtung des Gradienten  $\nabla J$
- Der Gradientenvektor ist orthogonal zu der Konturlinie.

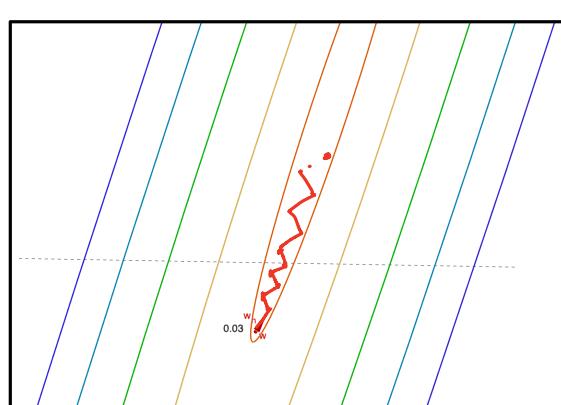
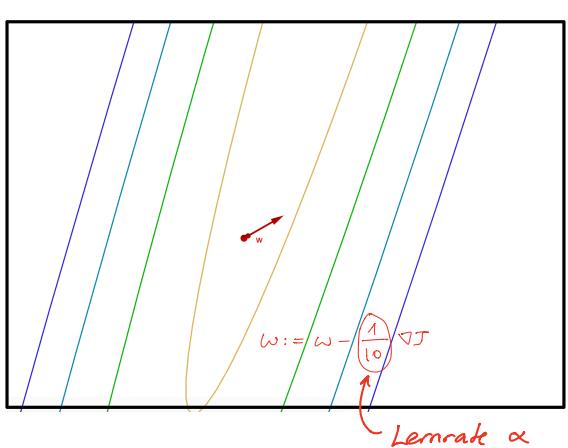
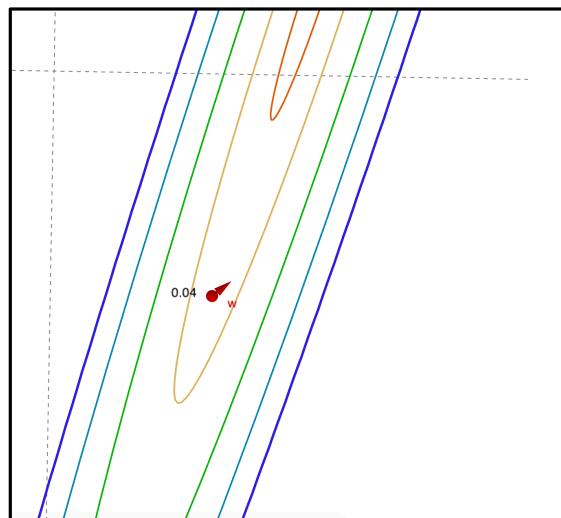
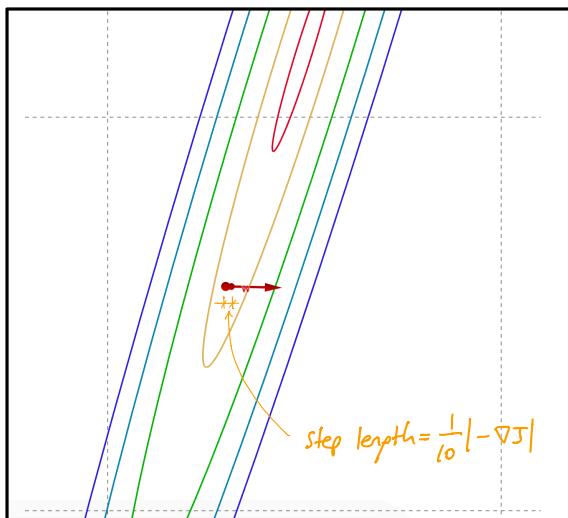


- In jedem Punkt  $w$ , bewege Dich
  - parallel zu der Konturlinie, um auf gleicher Höhe zu bleiben,
  - senkrecht zu der Konturlinie, in Richtung  $\nabla J$   
für den steilsten Aufstieg,
  - senkrecht zu der Konturlinie, in Richtung  $-\nabla J$   
für den steilsten Abstieg,
- Die Länge des Vektors  $\nabla J$  ist die steilste Steigung in dem Punkt.
- Je näher der Punkt  $w$  am Minimum ist, desto kleiner ist die Steigung.

## • A : Das Lernen guter Gewichte – Visuelle Intuition



- Mache einen "kleinen" Schritt in Richtung  $-\nabla J$   
z.B.  $1/10$  von  $|\nabla J|$



# • Der Gradientenabstieg Algorithmus (A)

① Starte mit zufälligem  $\omega$

② Berechne den Gradientenvektor im aktuellen Punkt  $\omega$

$$\nabla J|_{\omega}$$

(Richtung des steilsten Aufstiegs)

③ Gehe einen "kleinen" Schritt in Richtung  $-\nabla J|_{\omega}$

$$\omega := \omega - \alpha \cdot \nabla J|_{\omega}$$

THE  
LEARNING

$$\begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{bmatrix}_{\text{neu}} := \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{bmatrix}_{\text{alt}} - \alpha \cdot \begin{bmatrix} \frac{\partial J}{\partial \omega_0} \\ \frac{\partial J}{\partial \omega_1} \\ \vdots \\ \frac{\partial J}{\partial \omega_n} \end{bmatrix}_{\omega}$$



•  $\alpha$  : Lernrate / Schrittweite

$\alpha$  Hyperparameter  
(e.g.  $\alpha=0.1$ )

• Bemerkung: Alle Gewichte  $\omega_0, \omega_1, \dots, \omega_n$  werden gleichzeitig aktualisiert (simultan).

- Gradientenvektor für die Lineare Regression (m Beispiele  
n Merkmale)

- Die partiellen Ableitungen

$$J(\omega) = \frac{1}{2m} \sum_{i=1}^m \left( h_{\omega}(x^{(i)}) - (y^{(i)}) \right)^2$$

$$J(\omega) = \frac{1}{2m} \sum_{i=1}^m \left[ \left( \sum_{j=0}^n \omega_j x_j^{(i)} \right) - y^{(i)} \right]^2$$

$$\boxed{\omega_0 + \omega_1 x_1^{(i)} + \omega_2 x_2^{(i)} + \dots + \omega_j x_j^{(i)} + \dots + \omega_n x_n^{(i)}}$$

$$\Rightarrow \boxed{\frac{\partial J}{\partial \omega_j} = \frac{1}{m} \sum_{i=1}^m \left[ \left( \sum_{j=0}^n \omega_j x_j^{(i)} \right) - y^{(i)} \right] \cdot x_j^{(i)}} \quad (j=0, \dots, n)$$

↑  
Partielle Ableitung von  $h(x)$  nach  $\omega_j$

- Die Gewichtsaktualisierungen

$$\boxed{\omega_j := \omega_j - \frac{\alpha}{m} \sum_{i=1}^m \left[ h(x^{(i)}) - y^{(i)} \right] \cdot x_j^{(i)}}$$

$$(j=0, 1, \dots, n)$$

## BSP.

- $J(\omega_0, \omega_1) = \frac{1}{6} \left[ (\omega_0 + 2\omega_1 - 4)^2 + (\omega_0 + 3\omega_1 - 5)^2 + (\omega_0 + 4\omega_1 - 7)^2 \right]$

- $\frac{\partial J}{\partial \omega_0} = \frac{1}{6} \left[ 2(\omega_0 + 2\omega_1 - 4) \cdot 1 + 2(\omega_0 + 3\omega_1 - 5) \cdot 1 + 2(\omega_0 + 4\omega_1 - 7) \cdot 1 \right]$

$$\frac{\partial J}{\partial \omega_1} = \frac{1}{6} \left[ 2(\omega_0 + 2\omega_1 - 4) \cdot 2 + 2(\omega_0 + 3\omega_1 - 5) \cdot 3 + 2(\omega_0 + 4\omega_1 - 7) \cdot 4 \right]$$

$$\Rightarrow \nabla J = \begin{bmatrix} \partial J / \partial \omega_0 \\ \partial J / \partial \omega_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{6}(6\omega_0 + 18\omega_1 - 32) \\ \frac{1}{6}(18\omega_0 + 58\omega_1 - 102) \end{bmatrix}$$

- Die Gewichtsaktualisierung

- Aktuelle Gewichte :  $(\omega_0, \omega_1) = (3, 1)$  ( $\omega_{\text{alt}}$ )

- Gradient an dieser Stelle :

$$\nabla J \Big|_{(3,1)} = \begin{bmatrix} \frac{1}{6}(6\omega_0 + 18\omega_1 - 32) \\ \frac{1}{6}(18\omega_0 + 58\omega_1 - 102) \end{bmatrix}_{(3,1)} = \begin{bmatrix} 4/6 \\ 10/6 \end{bmatrix}$$

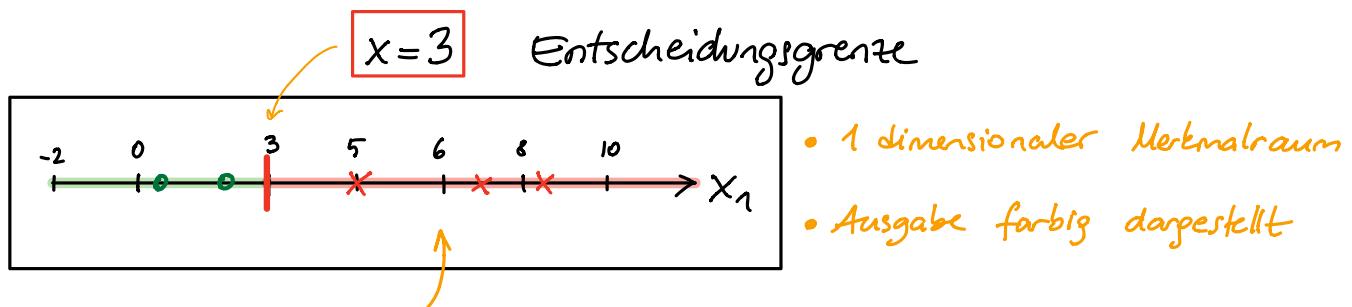
- Neue Gewichte :

$$\omega_{\text{neu}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} - (0.1) \cdot \begin{bmatrix} 1/3 \\ 5/3 \end{bmatrix} \approx \begin{bmatrix} 2.93 \\ 0.83 \end{bmatrix} (\omega_{\text{neu}})$$

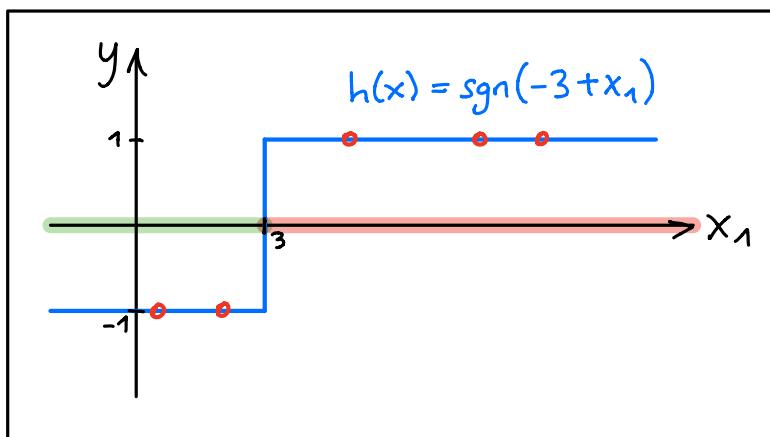
- Frage: Sind die neuen Gewichte bestimmt besser als die alten?

## • Regression - Style Visualisierung des Perzeptrons (zum Vergleich)

- $h(x) = \text{sgn}(-3 + x_1) = \text{sgn}(\omega^T x)$   $(\omega_0, \omega_1) = (-3, 1)$
- $\omega^T x = -3 + x_1 \geq 0 \Rightarrow x_1 \geq 3 \Rightarrow 1: \text{Hoch}$   
 $\omega^T x = -3 + x_1 < 0 \Rightarrow x_1 < 3 \Rightarrow -1: \text{Niedrig}$



- Visualisiert im Regressionsstil



- 1 dimensionaler Merktraum
- Ausgabe mittels  $y$ -Achse dargestellt  
 $y \in \{-1, 1\}$

## • Skalierung der Merkmale (Feature Scaling)

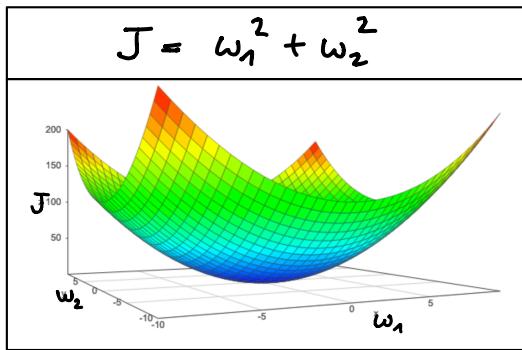
Der Gradientenabstieg kann schneller konvergieren, wenn die Merkmale in ähnlichen Wertebereichen liegen. WARUM?

- $h(x) = \omega_1 x_1 + \omega_2 x_2$

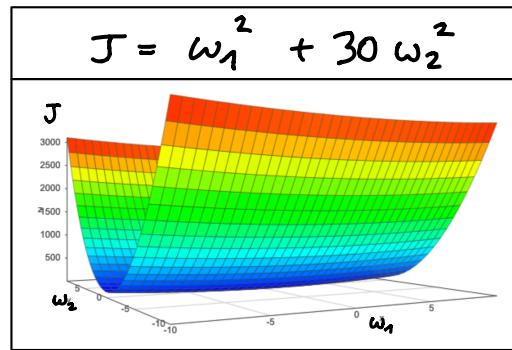
( sei  $\omega_0 = 0$  )

$$J(\omega_1, \omega_2) = \frac{1}{2} (\omega_1 x_1 + \omega_2 x_2 - y)^2$$

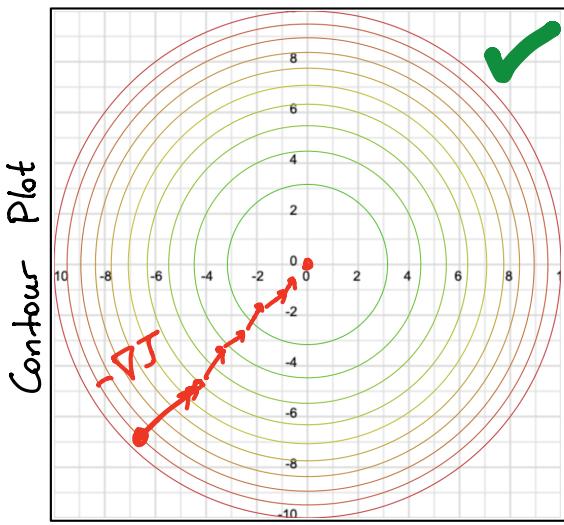
( ein Beispiel )



Quelle: <https://academo.org/demos/3d-surface-plotter>

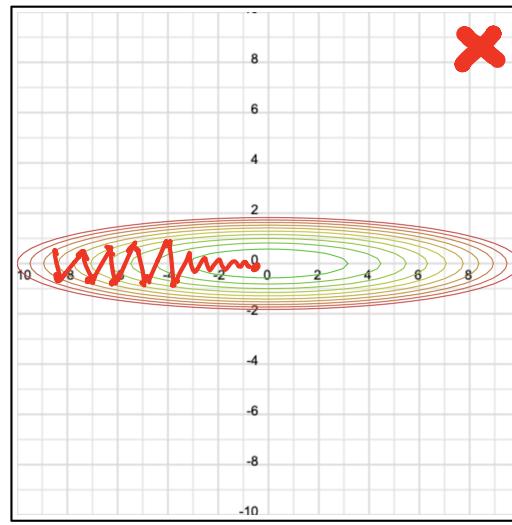


Quelle: <https://academo.org/demos/3d-surface-plotter>



Quelle: <https://academo.org/demos/contour-plots>

Schnellere Konvergenz



Quelle: <https://academo.org/demos/contour-plots>

Langsamere Konvergenz

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \checkmark$$

$$\Rightarrow J = \frac{1}{2} (\omega_1^2 + \omega_2^2 + \dots)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 10 \end{bmatrix} \quad \times$$

$$\Rightarrow J = \frac{1}{2} (\omega_1^2 + 100\omega_2^2 + \dots)$$

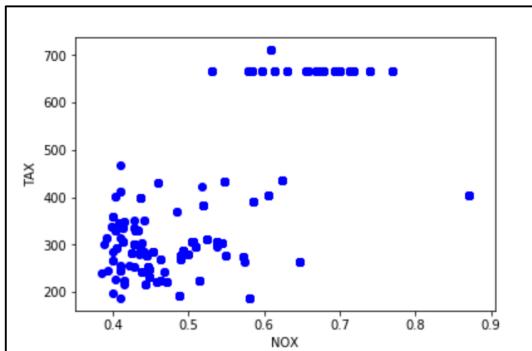
- Standardisierung :

$$x' = \frac{x - \mu}{\sigma}$$

$\mu$ : Mittelwert

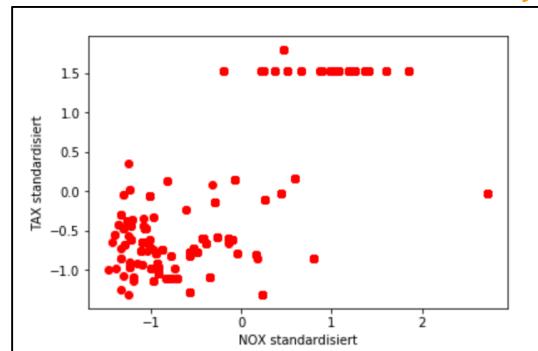
$\sigma$ : Standardabweichung

BSP.



$$NOX \in [0.4, 0.9]$$

$$TAX \in [187, 711]$$



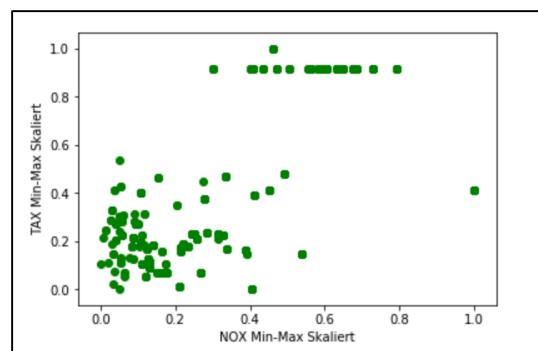
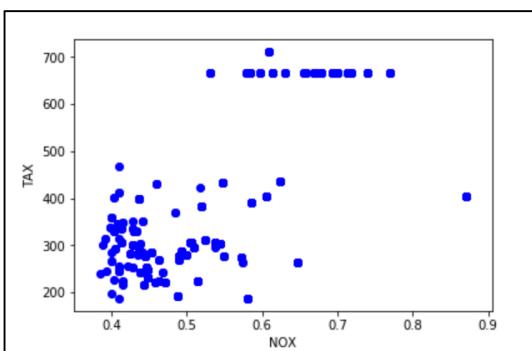
$$NOX' \in [-1.5, 2.8]$$

$$TAX' \in [-1.4, 1.8]$$

- Min-Max Skalierung :

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

BSP.



$$NOX' \in [0, 1]$$

$$TAX' \in [0, 1]$$

- Normalisierung mittels L1 oder L2 Norm (seltener)

- L1 :

$$x_j = \frac{x_j}{\sum_{j=0}^n x_j}$$

- L2 :

$$x_j = \frac{x_j}{\sqrt{\sum_{j=0}^n x_j^2}}$$

Anwendung nicht auf Merkmale, sondern auf Beispiele !

BSP.

UE

## • Minimierung der Kostenfunktion - Analytische Lösung

$$J(w_0, w_1) = \frac{1}{6} \left[ (w_0 + 2w_1 - 4)^2 + (w_0 + 3w_1 - 5)^2 + (w_0 + 4w_1 - 7)^2 \right]$$

$$\frac{\partial J}{\partial w_0} = \frac{1}{6} \left[ 2 \cdot (w_0 + 2w_1 - 4) \cdot 1 + 2 \cdot (w_0 + 3w_1 - 5) \cdot 1 + 2 \cdot (w_0 + 4w_1 - 7) \cdot 1 \right] = 0$$

$$\Rightarrow \frac{1}{6} (6w_0 + 18w_1 - 32) = 0$$

$$\begin{aligned} \frac{\partial J}{\partial w_1} &= \frac{1}{6} \left[ 2 \cdot (w_0 + 2w_1 - 4) \cdot 2 + 2 \cdot (w_0 + 3w_1 - 5) \cdot 3 + 2 \cdot (w_0 + 4w_1 - 7) \cdot 4 \right] = 0 \\ &= \frac{1}{6} [18w_0 + 58w_1 - 102] = 0 \end{aligned}$$

$$\begin{array}{l|l} \begin{array}{l} 6w_0 + 18w_1 - 32 = 0 \\ 18w_0 + 58w_1 - 102 = 0 \end{array} & \left| \begin{array}{ccc|c} 6 & 18 & 32 \\ 18 & 58 & 102 \end{array} \right. \\ & \left| \begin{array}{ccc|c} 6 & 18 & 32 \\ 0 & 4 & 6 \end{array} \right. \end{array} \quad 4w_1 = 6 \Rightarrow \boxed{w_1 = \frac{3}{2}} \\ & \qquad \qquad \qquad = 1.5 \end{math>$$

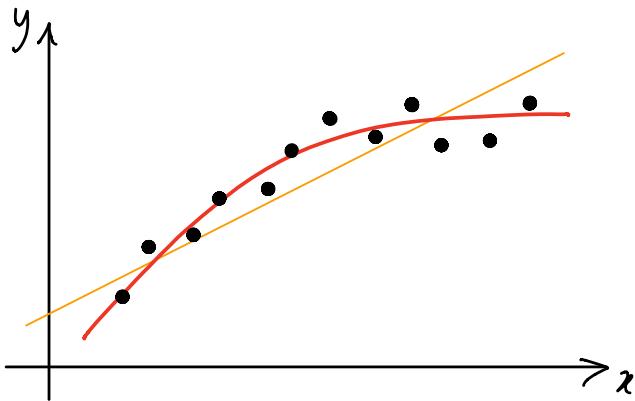
$$6w_0 + 18 \cdot \frac{3}{2} = 32 \quad \Rightarrow 6w_0 + 18 \cdot \frac{3}{2} = 32$$

$$6w_0 = 32 - 27 = 5 \quad \Rightarrow \quad w_0 = \frac{5}{6} \quad \Rightarrow \quad 6w_0 = 32 - 27 = 5$$

$$\Rightarrow (w_1, w_0) = \left( \frac{3}{2}, \frac{5}{6} \right) \quad \Rightarrow \boxed{w_0 = \frac{5}{6}} = 0.8\overline{3}$$

$$\Rightarrow \boxed{h(x) = \frac{3}{2}x + \frac{5}{6} =: g(x)} \quad \text{Beste Gerade } h^*(x)$$

## • Polynomiale Regression



- Der Zusammenhang zwischen  $x$  und  $y$  scheint hier nicht linear zu sein
- Die rote Kurve passt sich besser an die Daten an, als die gelbe Gerade.
- Unter Einführung von nichtlinearen Termen ( $x^2, \sqrt{x}, \dots$ ) (als neue Merkmale) kann man "flexible" Kurven an die Daten anpassen, wie zum Beispiel:

$$h(x) = w_0 + w_1 \underbrace{x}_1 + w_2 \underbrace{x^2}_2 + w_3 \underbrace{\sqrt{x}}_3 \rightarrow \text{lineare Regression mit drei Merkmalen}$$

- $x^2$  und  $\sqrt{x}$  sind Zahlen,  $h(x)$  ist immer noch linear in  $w_i$ !

Die Kostenfunktion ist weiterhin quadratisch in  $w_i$  und hat ein globales Minimum, das wie bei linearer Regression mittels Gradientenabstieg berechnet werden kann.