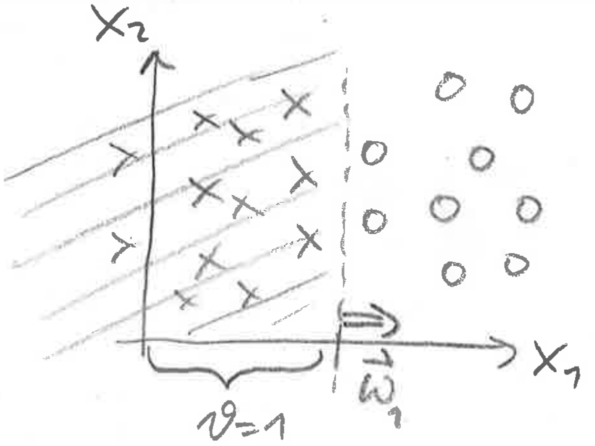
Blackboard 5-1a: Saddle Points

1 neuron : input space (data) parameter space



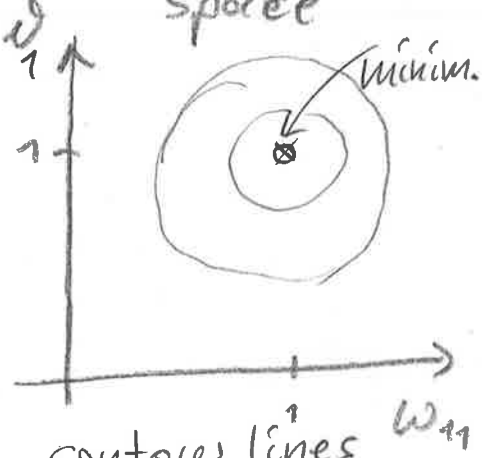
$\vec{w}_1 = (w_{11}, w_{12})$

$= (1, 0)$

$\vartheta_1 = 1$

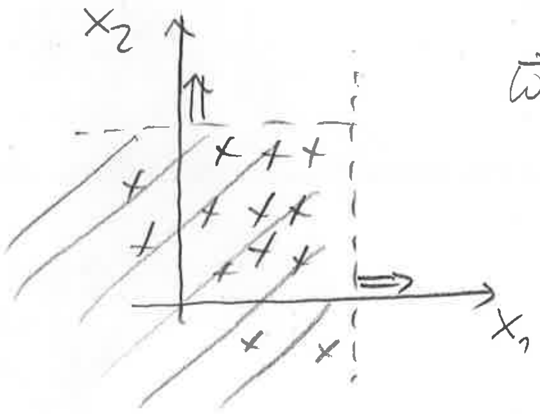
normalized

$\|\vec{w}\| = 1$



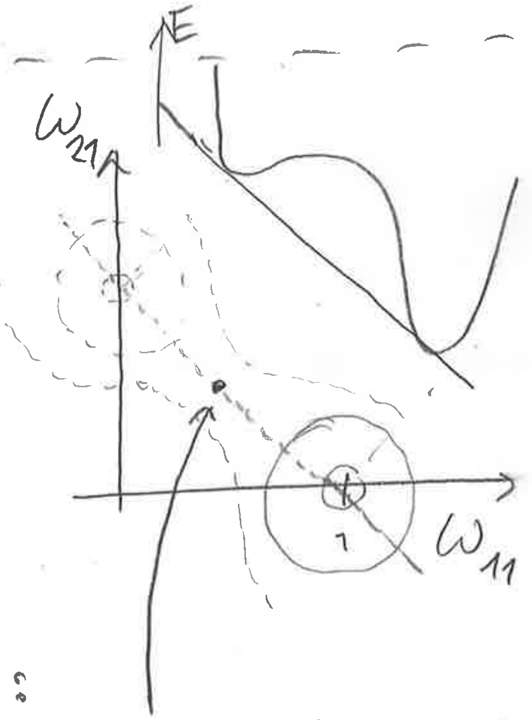
contour lines of error f.
projection to $w_{12} = 0$

2 neurons



$\vec{w}_1 = (1, 0)\ ;\ \vartheta_1 = 1$

$\vec{w}_2 = (0, 1)\ ;\ \vartheta_2 = 1$



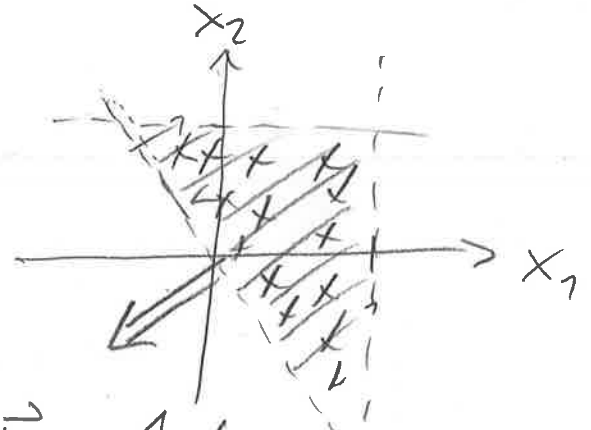
permutation of weight vectors:

$\vec{w}_1 \rightleftarrows \vec{w}_2$

"2nd neuron implements first hyperplane and viceversa"

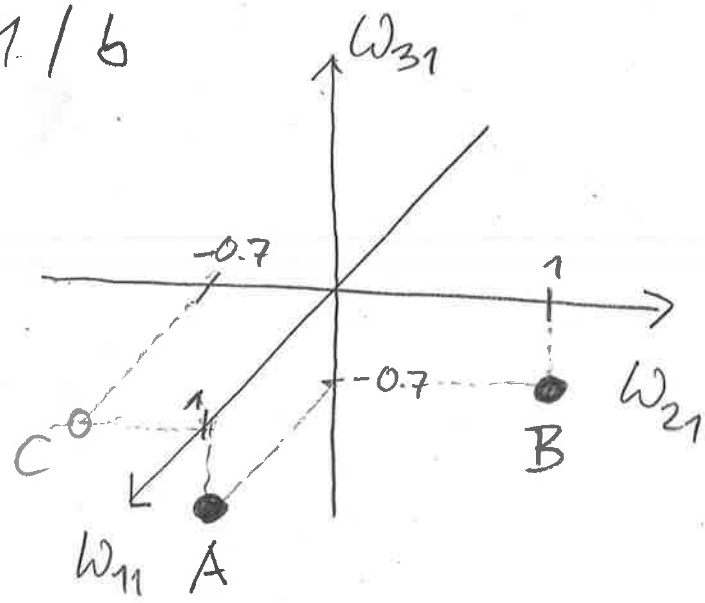
Saddle between minima

Blackboard 5.1/6



$\vec{w}_3 = \frac{1}{\sqrt{2}}(-1,-1)$

$\vartheta_3 = 0$



Attention: parameters $w_{12}, w_{22}, w_{32}, \vartheta_1, \vartheta_2, \vartheta_3$ are **not** plotted.
Idea: adjusted to minimum **given** w_{11}, w_{21}, w_{31}

$$A = (1, 0, -0.7) \quad (w_{11}, w_{21}, w_{31})$$

first permutation $\vec{w}_1 \leftrightarrows \vec{w}_2$

$$B = (0, 1, -0.7)$$

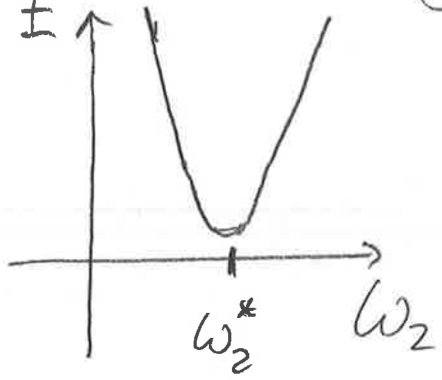
Second mutation, starting from A: $\vec{w}_2 \rightleftarrows \vec{w}_3$

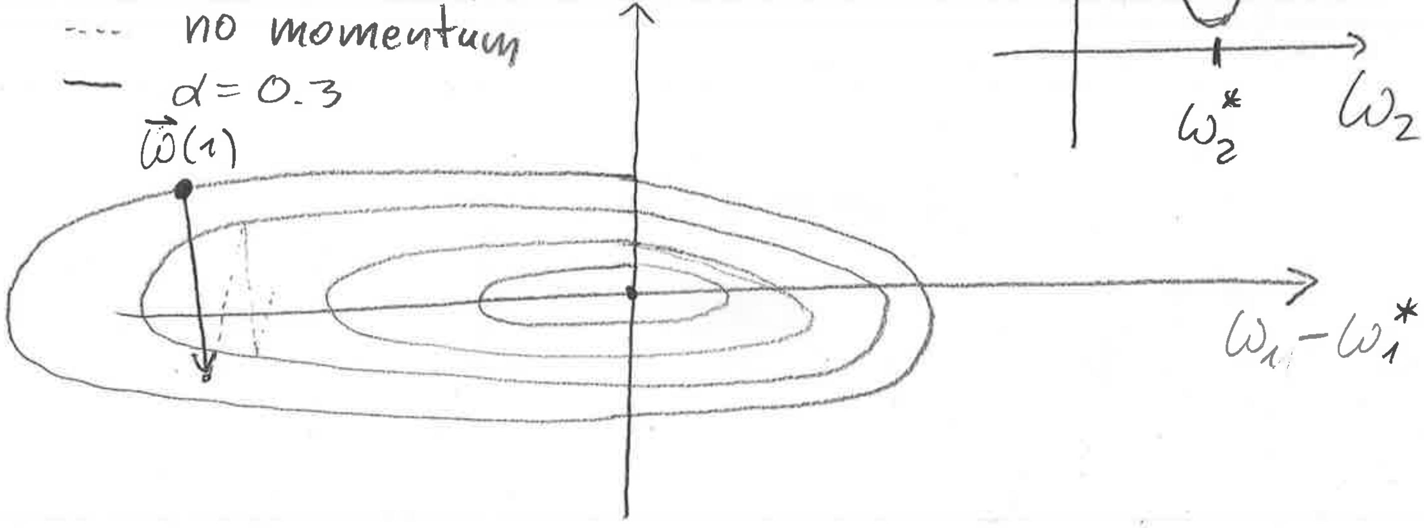
$$C = (1, -0.7, 0)$$

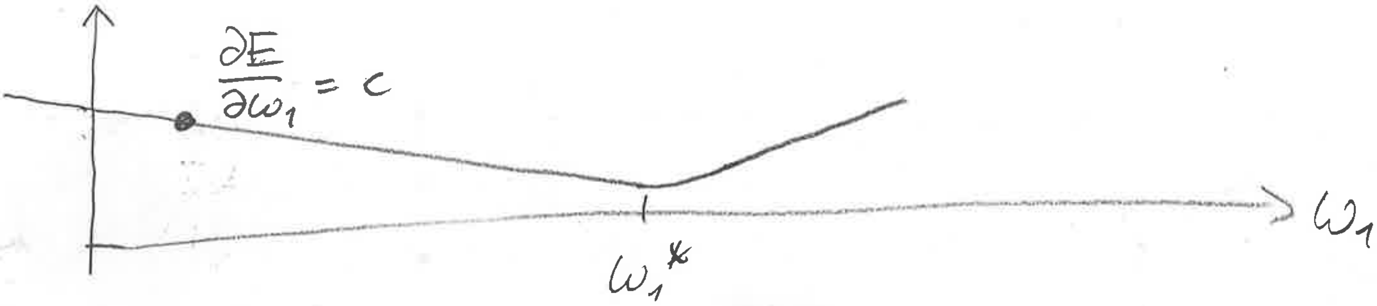
etc.

saddle points between all permutations!

Blackboard 5.2: Momentum E







$$\Delta\omega_1(1) = -\gamma \cdot \frac{\partial E}{\partial \omega_1} = -\gamma c$$

$$\Delta\omega_1(2) = -\gamma c + \alpha \cdot \Delta\omega_1(1) = -\gamma c\,(1+\alpha)$$

$$\Delta\omega_1(3) = -\gamma c + \alpha \cdot \Delta\omega_1(2) = -\gamma c - \alpha \cdot \gamma c\,(1+\alpha)$$

$$= -\gamma c\,[1 + \alpha + \alpha^2]$$

$\downarrow$

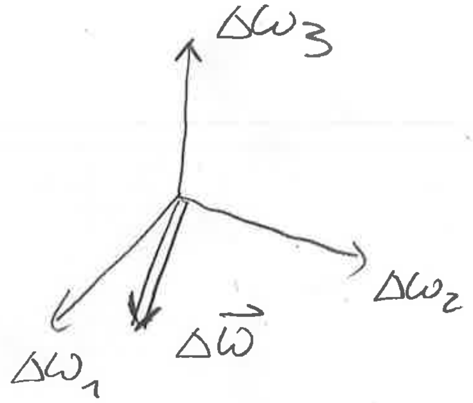
$$\Delta\omega_1(n) = -\gamma c\,[1 + \alpha + \ldots + \alpha^{n-1}]$$

$\downarrow$

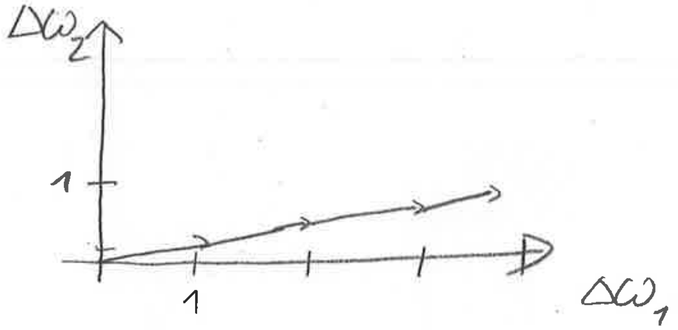
$$\Delta\omega_1(\infty) = -\underbrace{\frac{\gamma}{1-\alpha}} \cdot \overset{\uparrow}{c}$$

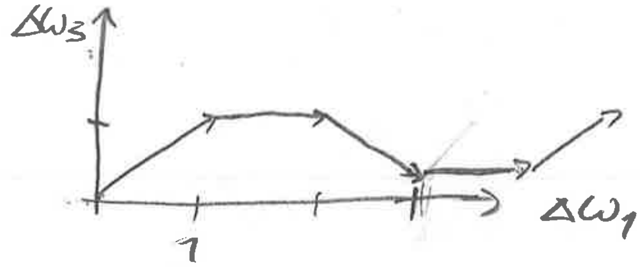
$$= -\gamma_{eff} \; \frac{\partial E}{\partial \omega_1}$$

Blackboard 5.3: stochastic gradients



projections 2D





		ratio
$\langle \Delta w_1 \rangle = 1$	; $\sqrt{\langle \Delta w_1^2 \rangle} \approx 1.01$	$= 1$
$\langle \Delta w_2 \rangle = 0.1$	; $\sqrt{\langle \Delta w_2^2 \rangle} = 0.1$	$= 1$
$\langle \Delta w_3 \rangle = 0.05$	; $\sqrt{\langle \Delta w_3^2 \rangle} = \sqrt{0.5} \approx 0.7$	≈ 0.07

$\Rightarrow$ smaller steps in "noisy" directions

<u>note:</u> absolute size of gradient irrelevant