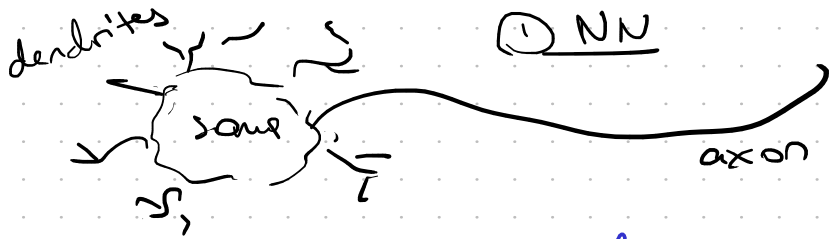
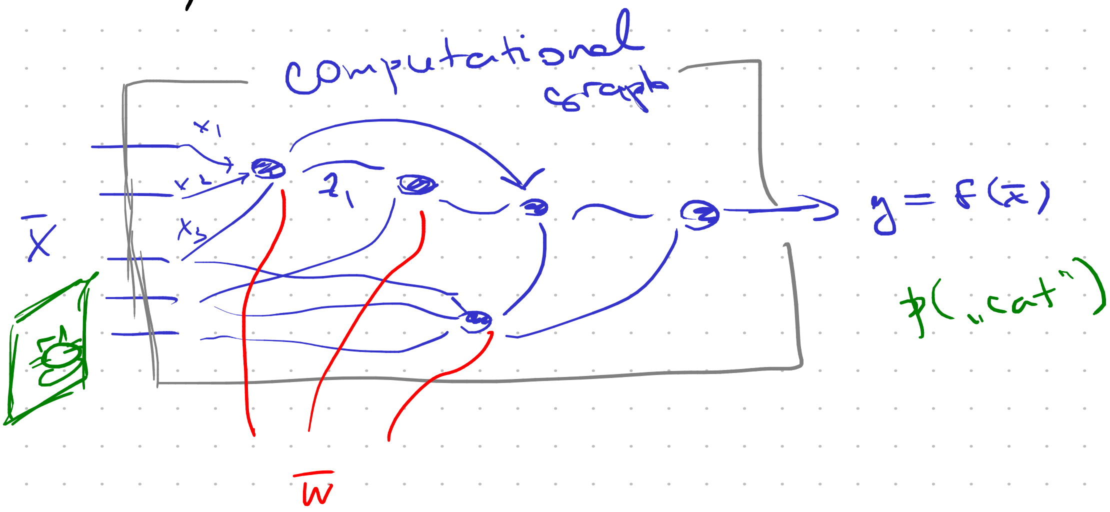
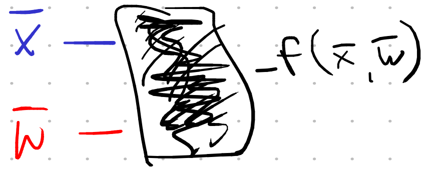
① NN

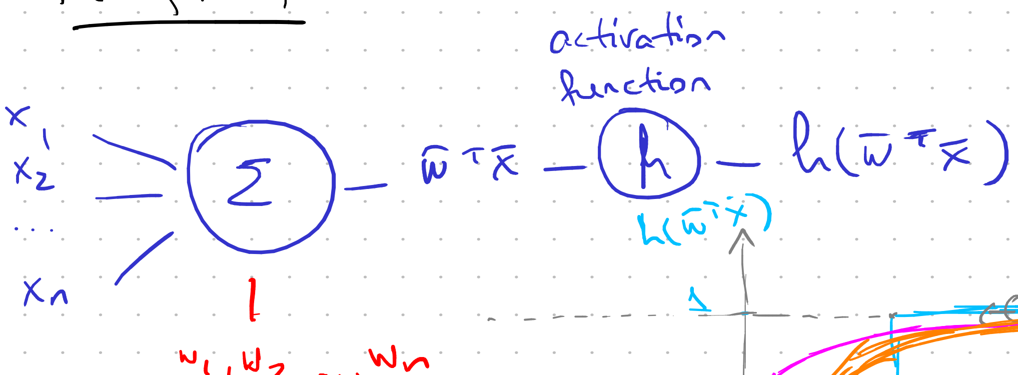


Connectionism



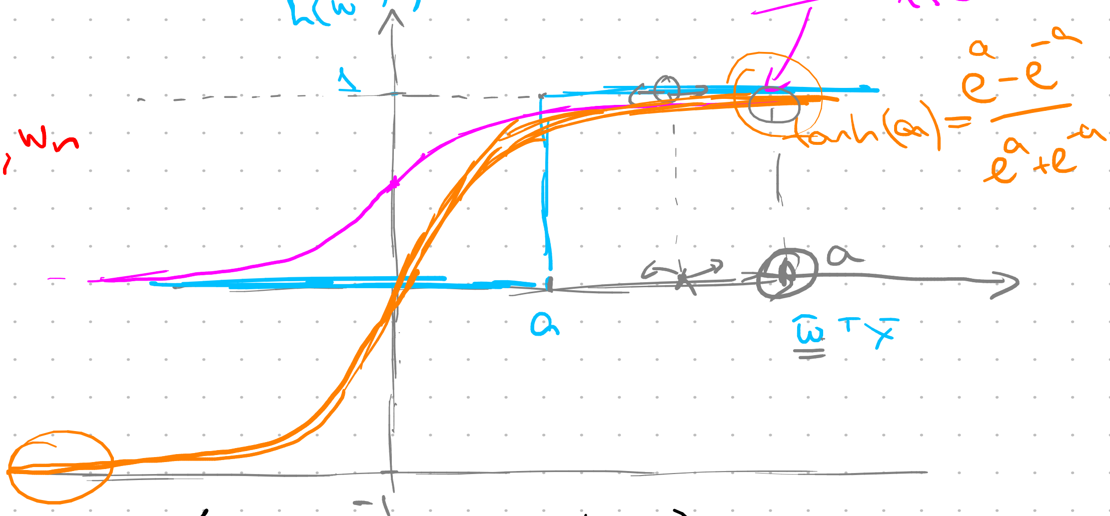


② Perceptron

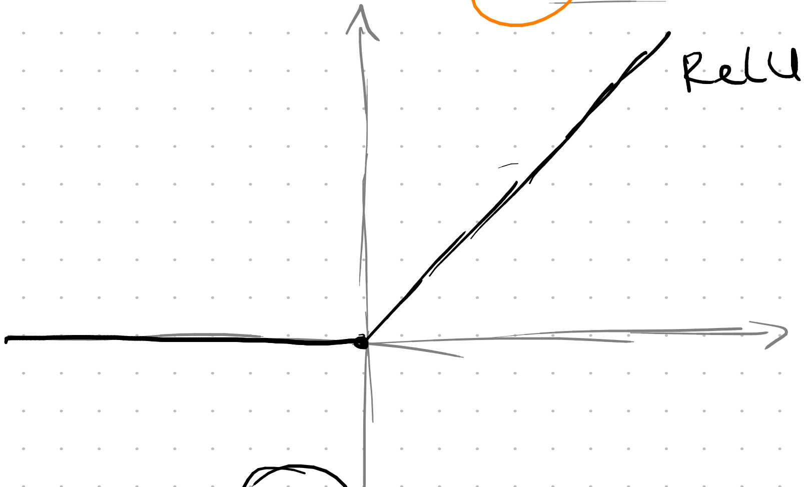


$$\sigma(a) = \frac{1}{1+e^{-a}}$$

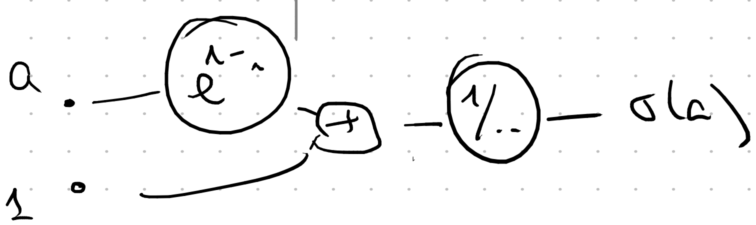
$$\tanh(a) = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}$$

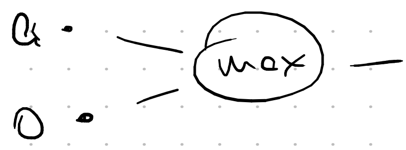


$$ReLU(a) = \max(0, a)$$



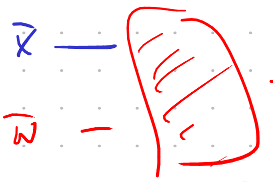
h(a)





③ Backpropagation

Gradient descent / ascent



$$y = f(\bar{x}, \bar{w})$$

$$\bar{w} \to \min$$

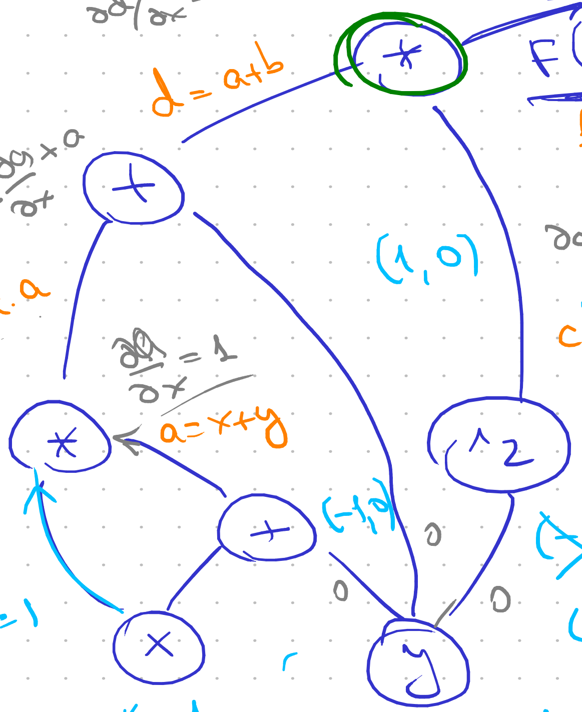
$$\nabla_{\bar{w}} f(\bar{x}, \bar{w}) = \left(\frac{\partial f}{\partial w_1}, \dots, \frac{\partial f}{\partial w_n} \right)$$

$$F(\bar{x}) \to \min$$

$$\bar{x}^{(0)}, \dots, \bar{x}^{(k+1)} := \bar{x}^{(k)} - \eta \cdot \nabla_{\bar{x}} F(\bar{x})$$

$$F(x,y) = y^2 \cdot (x(x+y)+y)$$

$$\nabla F = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y} \right)$$



$$\frac{\partial \dots}{\partial x}$$

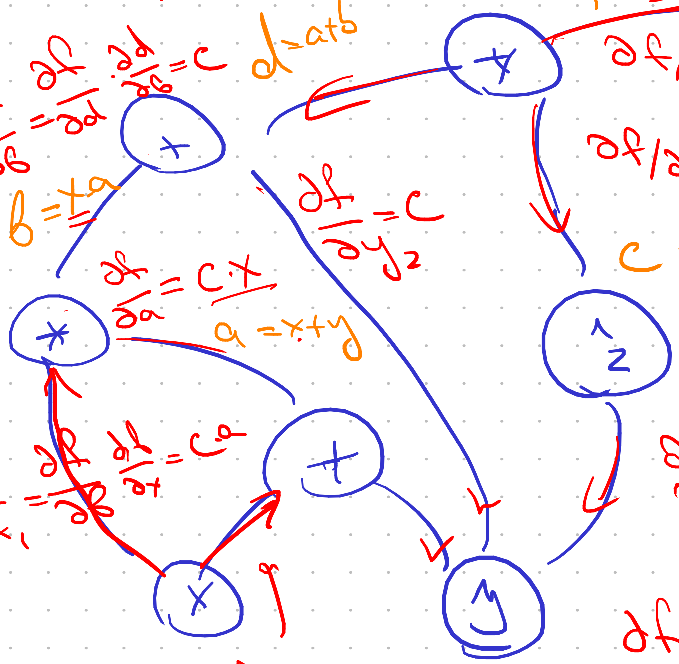
forward propagation

f prop

$$\partial f / \partial x$$

$$\frac{\partial f}{\partial x} = c \cdot \frac{\partial d}{\partial x} + \frac{\partial c}{\partial x} \cdot d = y^2 \cdot (x+a) =$$

$$= y^2 (x + x + y)$$



$$\frac{\partial f}{\partial \dots}$$

$$\boxed{\text{backpropagation}}$$

$$\frac{\partial f}{\partial y_2} = \frac{\partial f}{\partial c} \frac{\partial c}{\partial y} = 2y \cdot d$$

$$\frac{\partial f}{\partial x} = \frac{\partial a}{\partial x} \frac{\partial f}{\partial a} + \frac{\partial b}{\partial x} \frac{\partial f}{\partial b} = c \cdot a + c \cdot x =$$

$$= y^2 (2x + y)$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial y} + \frac{\partial f}{\partial d}\frac{\partial d}{\partial y} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial y} = \partial f/\partial y_1 + \partial f/\partial y_2 + \partial f/\partial y_3$$

$$= c\cdot x + c + 2y\cdot d = y^2(x+1) + 2y(x(x+y)+y)$$

PyTorch / Tensorflow / ... $\rightarrow$ automated differentiation library

(3) Gradient descent

first-order optimization / second-order optim.?

$$F(\bar{x}) \approx f(\bar{x}_0) + \nabla F^T(\bar{x}-\bar{x}_0) + \frac{1}{2}(\bar{x}-\bar{x}_0)^T H (\bar{x}-\bar{x}_0)$$

quasi-Newton methods $\qquad$ l-BFGS $\qquad$ $\boxed{F(\bar{x})}$ $\boxed{\nabla F(\bar{x})}$

$D = \{(\bar{x}_n, y_n)\}$

$$L(\bar{w}) = \frac{1}{N}\sum_{n=1}^{N} l(\bar{w}, \bar{x}_n, y_n) = \mathbb{E}_{(\bar{x},\bar{y})\sim Unif(D)}\left[l(\bar{w}, \bar{x}, \bar{y})\right]$$

Stochastic optimization

$$F(\bar{x}) = \mathbb{E}_{q(\bar{y})}\left[f(\bar{x},\bar{y})\right] \approx \frac{1}{m}\sum_{i=1}^{m} f(\bar{x}, \bar{y}_i), \quad \bar{y}_i \sim q(\bar{y})$$

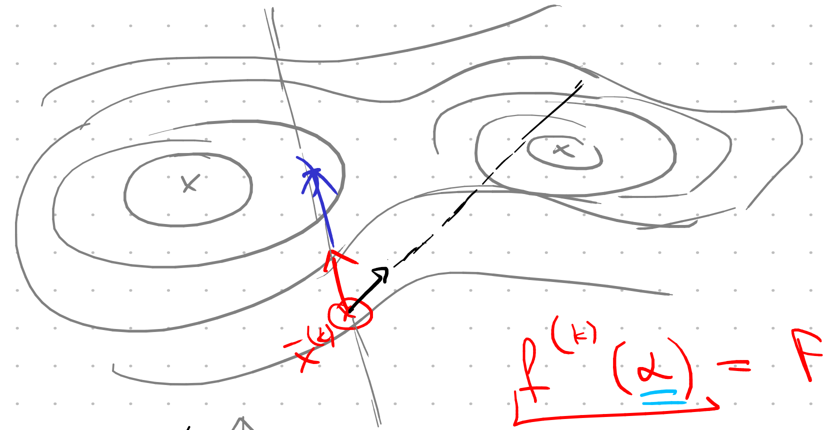
$F(\bar{w})$ | $\hat{F}(\bar{w}) \approx F(\bar{w})$

$\nabla_{\bar{w}} F$ | $\hat{G}(\bar{w}) \approx \nabla_{\bar{w}} F$

Stochastic grad descent

4 Stochastic gradient descent

$$\bar{x}^{(k+1)} := \bar{x}^{(k)} - \eta \cdot \nabla_{\bar{x}} f(\bar{x})$$

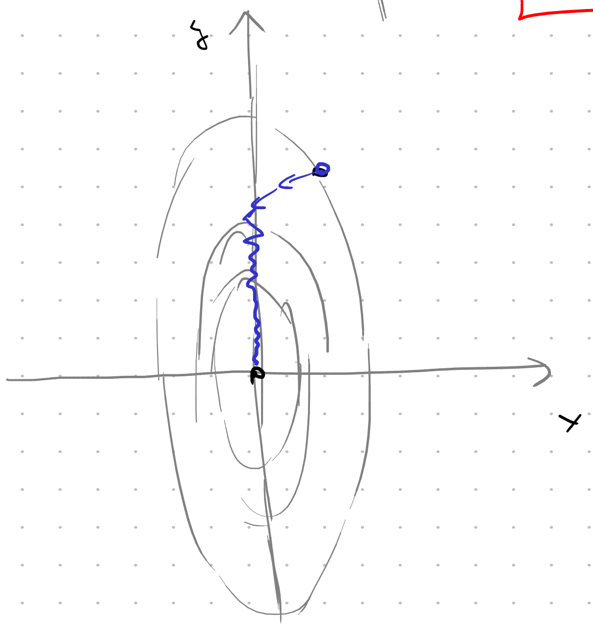


$$f^{(k)}(\alpha) = f\left(\bar{x}^{(k)} - \alpha \nabla_{\bar{x}} f \big|_{\bar{x}^{(k)}}\right)$$

$\alpha \to \min$

$\frac{1}{10}, \frac{1}{100}$

$$f(x,y) = x^2 + \beta \cdot y^2$$



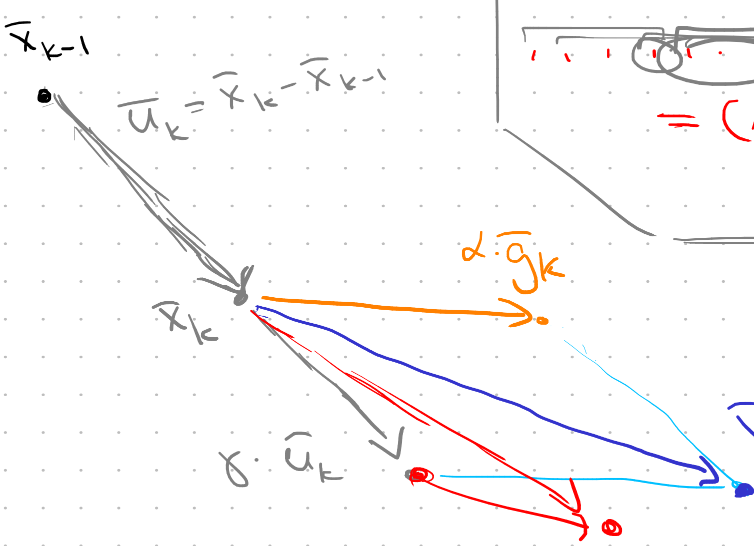
$\bar{x}_k, \bar{x}_{k+1}, \dots, \quad \bar{g}_k = \nabla f \big|_{\bar{x}_k}$

- momentum

EMA - exp. moving avg

$$y_{k+1} = \gamma \cdot y_k + (1-\gamma) x_k =$$

$$= (1-\gamma) x_k + \gamma \cdot (1-\gamma) x_{k-1} + \gamma^2 (1-\gamma) x_{k-2} + \dots$$



$\gamma = 0.95, 0.98$

$$\bar{x}_{k+1} = \bar{x}_k + \gamma \cdot \bar{u}_k + \alpha \cdot \bar{g}_k$$

Adam
AdamW
NadamW

- Nesterov acc_grad

m/c $\quad g = \left(\frac{\partial f}{\partial x}\right)$

$$\bar{x}_{k+1} = \overbrace{\bar{x}_k}^{c} - \alpha \cdot \nabla \overbrace{F}^{M}$$

M/c

$$\bar{x}_{k+1} = \bar{x}_k - \frac{\alpha}{\sqrt{G = \sum g^2}} \cdot \nabla F$$

$$c := c - 1 \cdot M/c$$

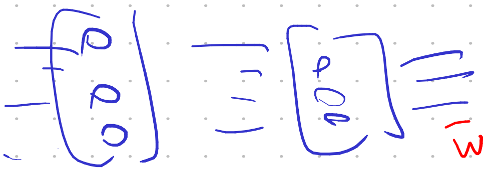
$$c := c - \frac{1}{\sqrt{(M/c)^2}} \cdot M/c$$

$$= c - 1$$

Newton:

$$\bar{x}_{k+1} := \bar{x}_k - H^{-1} \cdot \nabla_{\bar{x}} F$$

$$c := c - \left(M/c^2\right)^{-1} \; M/c$$



$\bar{w}_*$

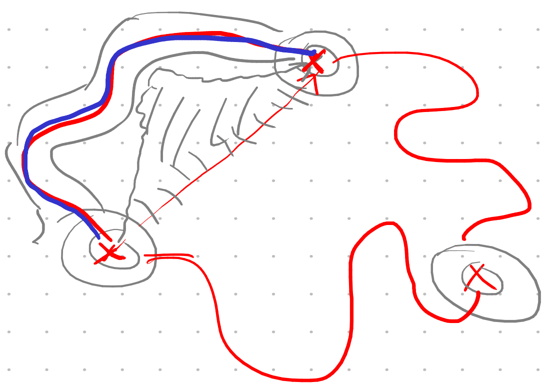
$$\bar{x} \to W_1 \bar{x} \to W_2 W_1 \bar{x}$$

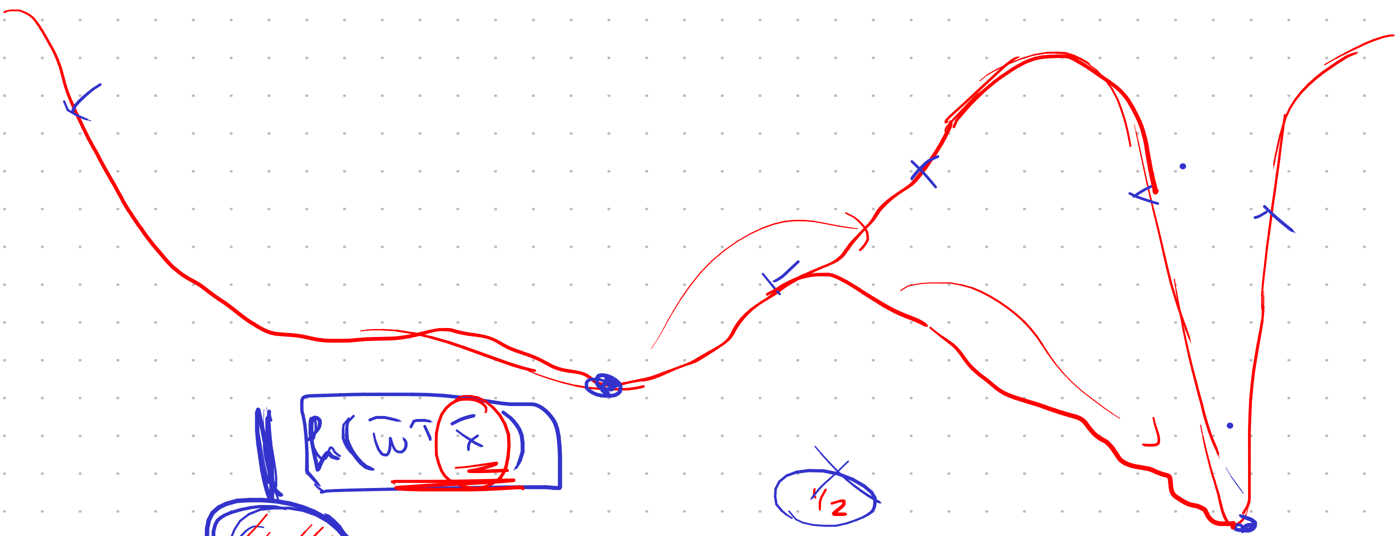
$$(W_2 \cdot A)(A^{-1} W_1)$$

$\bar{w}'_*$

1. Ветров

$\bar{w}$





$h(\bar{w}^T \bar{x})$

x_1 - x_n

x_i: