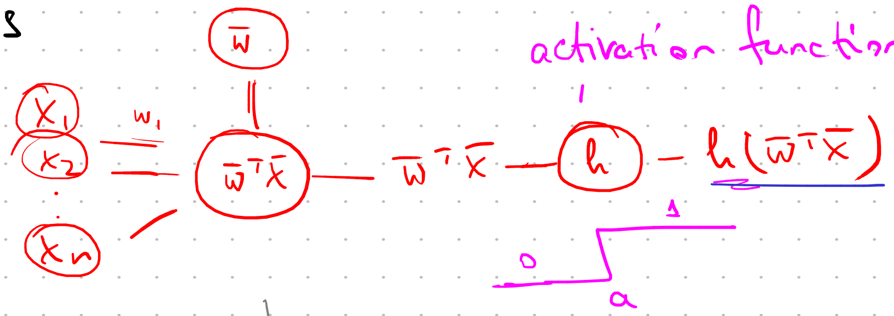
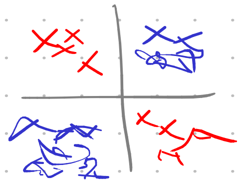
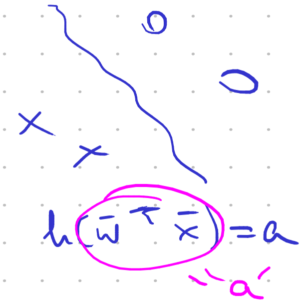
① ANNs



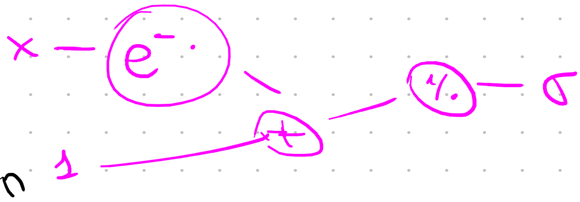




② Activation functions

$$\mathrm{ReLU}(x) = \max(0, x)$$

$$\sigma(x) = 1/(1+e^{-x})$$



$$S_\beta(x_1 \ldots x_n) = \frac{x_1 \cdot e^{\beta x_1} + \ldots + x_n \cdot e^{\beta x_n}}{e^{\beta x_1} + \ldots + e^{\beta x_n}}$$

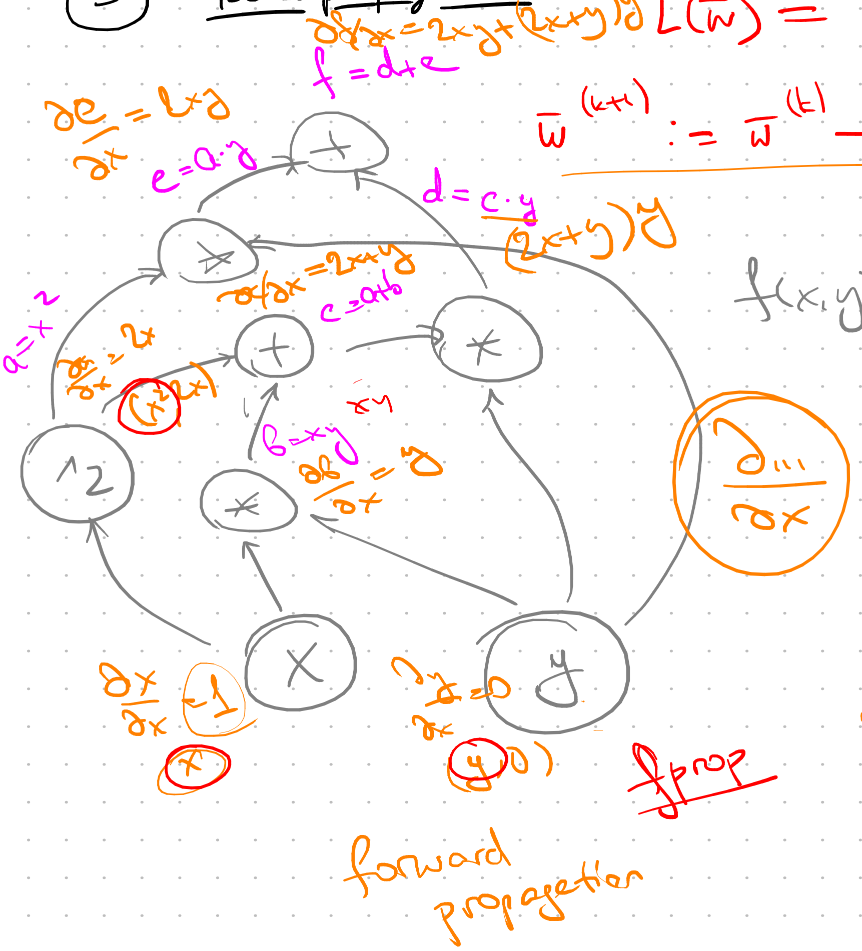
$$-\log p(\bar{x}_n \mid \bar{w})$$

③ Backpropagation

$$L(\bar{w}) = \sum_{n=1}^{N} \ell(\bar{x}_n, \bar{w}) \xrightarrow{\bar{w}} \min$$

$$\bar{w}^{(k+1)} := \bar{w}^{(k)} - \alpha_k \nabla_{\bar{w}} L(\bar{w}) \Big|_{\bar{w}^{(k)}}$$

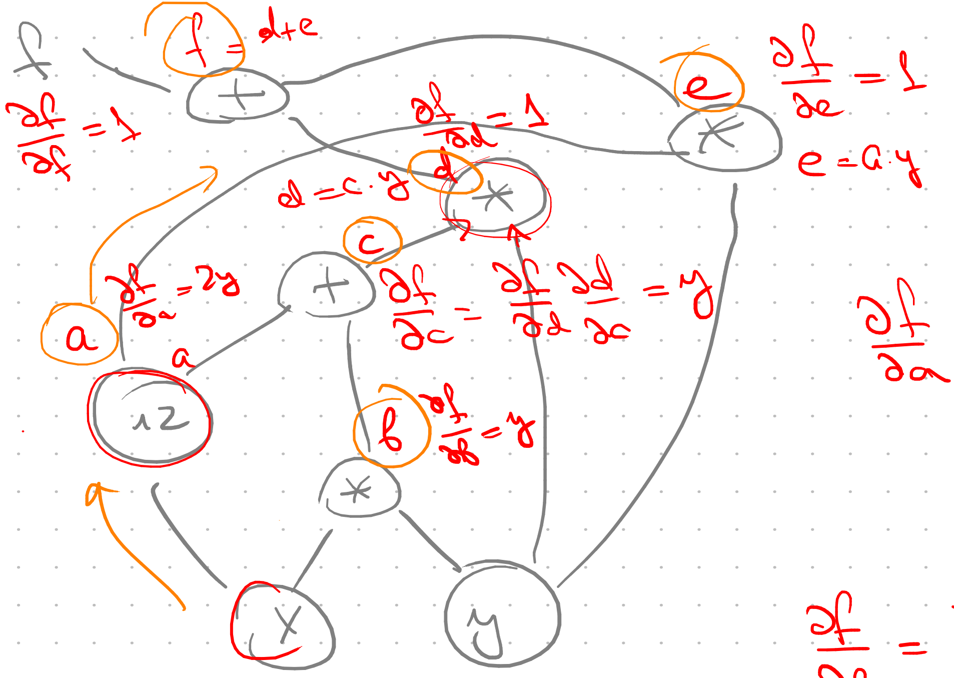
$$f(x,y) = (xy + x^2)y + x^2 y$$



$\partial f/\partial x$

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

$$\frac{\partial d}{\partial x} = \frac{\partial d}{\partial c}\frac{\partial c}{\partial x} = y \cdot \frac{\partial c}{\partial x}$$



$\partial f / \partial \ldots$

$$\frac{\partial f}{\partial a} = \overset{1}{\frac{\partial f}{\partial e}}\frac{\partial e}{\partial a} + \overset{y}{\frac{\partial f}{\partial c}}\frac{\partial c}{\partial a} =$$

$$= y + y = 2y$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} = y$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x} + \frac{\partial f}{\partial b}\frac{\partial b}{\partial x} = 4xy + y^2$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial y} + \frac{\partial f}{\partial d}\frac{\partial d}{\partial y} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial y} = xy + x^2 + xy + x^2$$

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

4 Stochastic gradient descent

$$F(\bar{x}) \xrightarrow{\bar{x}} \min \qquad \bar{x}^{(k+1)} := \bar{x}^{(k)} - \alpha \nabla_{\bar{x}} F\Big|_{\bar{x}^{(k)}}$$

$$f(\bar{x}) \simeq F(\bar{x}^{(k)}) + (\bar{x} - \bar{x}^{(k)}) \cdot \underbrace{\nabla_{\bar{x}} F(\bar{x}^{(k)})}_{\bar{g}_k} + \frac{1}{2}(\bar{x} - \bar{x}^{(k)})^T H_k (\bar{x} - \bar{x}^{(k)})$$

$$\bar{g}_k + H_k(\bar{x} - \bar{x}^{(k)}) = 0$$

Newton's method

$$\boxed{\bar{x} = \bar{x}^{(k)} - H_k^{-1} \cdot \bar{g}_k}$$

$\nabla f \Big|_{\bar{x}^{(k)}}$

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)$$

H_k^{-1} − $N \times N$

quasi Newton algorithms

L-BFGS

$\cancel{\bar{g}_k}, \bar{x}_k \rightsquigarrow$ low-rank $\approx H^{-1}$

$$\bar{g}_k = \sum_{n=1}^{N} \nabla_{\bar{x}} f(\bar{x}, \bar{y}_n)$$

SGD - stochastic GD

$$q(\bar{y}) = \mathrm{Unif}(D)$$

$D \sim p_{data}$

$q(\bar{y}) = p_{data}$

$$F(\bar{x}) = \mathbb{E}_{q(\bar{y})}\left[f(\bar{x}, \bar{y})\right] \xrightarrow{\bar{x}} \min$$

$$G(\bar{x}) = \mathbb{E}_{q(\bar{y})}\left[\nabla_{\bar{x}} f(\bar{x}, \bar{y})\right]$$

m - mini-batch

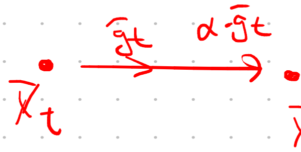
$$\tilde{F}(\bar{x}) = \frac{1}{m}\sum_{i=1}^{m} f(\bar{x}, \bar{y}_i) \quad \text{где} \quad \bar{y}_i \sim q(\bar{y})$$

$$\hat{g}(\bar{x}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\bar{x}} f(\bar{x}, \bar{y}_i)$$

SGD

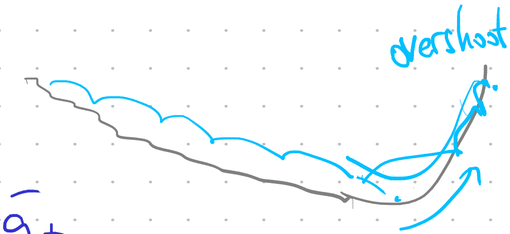
$$\bar{x}_{k+1} := \bar{x}_k - \alpha_k \cdot \hat{\bar{g}}_k$$

5 SGD problems & solutions



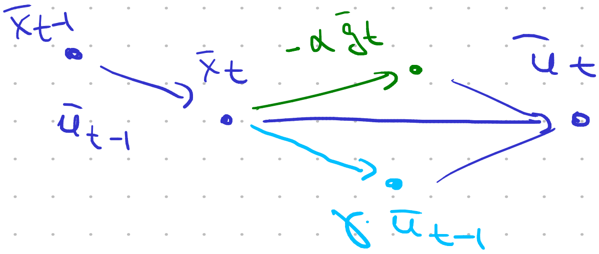
$$\bar{x}_{t+1} = \bar{x}_t - \alpha \cdot \bar{g}_t$$

$$\bar{u}_t = \bar{x}_{t+1} - \bar{x}_t = -\alpha \bar{g}_t$$



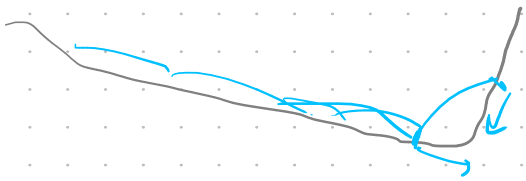
1) SGD w/momentum

$$\gamma = 0.9, 0.95$$

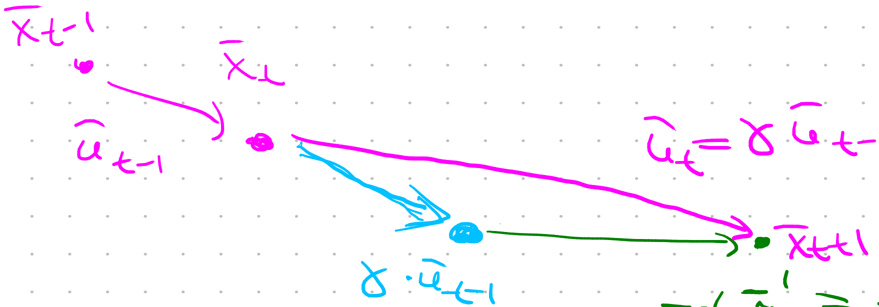


$$\bar{u}_t = \gamma \cdot \bar{u}_{t-1} - \alpha \bar{g}_t$$

$$\bar{x}_{t+1} = \bar{x}_t + \bar{u}_t$$

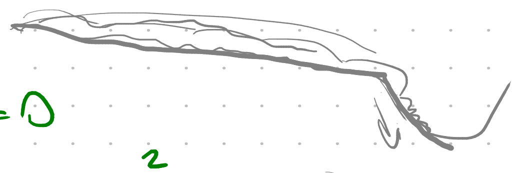


2) Nesterov accelerated gradient (NAG)



$$\bar{u}_t = \gamma \bar{u}_{t-1} - \alpha \bar{g}'_t$$

$$-\alpha \cdot \bar{g}'_t = -\alpha \cdot \nabla_{\bar{x}} F(\bar{x}_t + \gamma \cdot \bar{u}_{t-1})$$

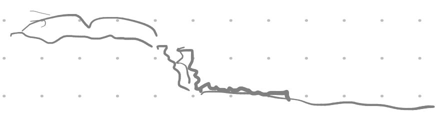


3) Adaptive SGD

- Adagrad

$$G_{0,i} = 0$$

$$G_{t,i} = G_{t-1,i} + g_{t,i}^2$$



$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i}+\varepsilon}} \cdot g_{t,i}$$

c c 1 $\sqrt{M/c^2}$ M/c

- **RMSProp** $G_{t,i} := \gamma G_{t-1,i} + (1-\gamma) g^2_{t,i}$

- **Adadelta** $\bar{g}$ – M/c, $\bar{x}$ – сек, $f(\bar{x})$ – меры)

$\bar{x} = \bar{x} - \alpha \cdot \bar{g}$

c c 1 M/c

$\bar{x}_{t+1} = \bar{x}_t - H_t^{-1} \bar{g}_t$

c c $(M/c^2)^{-1} \cdot M/c$

$$x_{t+1,i} = x_{t,i} - \alpha \cdot \frac{\sqrt{R_{t,i}+\varepsilon}}{\sqrt{G_{t,i}+\varepsilon}} g_{t,i}, \quad R_{t,i} = \rho \cdot R_{t-1,i} + (1-\rho) u^2_{t,i}$$

Lion
Muon

- **Adam**

$$G_{t,i} = \beta_2 G_{t-1,i} + (1-\beta_2) g^2_{t,i}$$

$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i}+\varepsilon}} \cdot m_{t,i} \qquad m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$$

$\beta_1 = 0.9 \quad \beta_2 = 0.999 \quad \varepsilon = 10^{-8}$

- **AdamW**

L_2: $L(\bar{w}) + \alpha \cdot \|\bar{w}\|^2$

weight decay: $\bar{w} := \underbrace{(1-\eta)}\bar{w} - \alpha \cdot \nabla_{\bar{w}} L(\bar{w})$

$$= \bar{w} - \underbrace{\eta \bar{w}} - \alpha \nabla_{\bar{w}} L(\bar{w}) =$$

$$= \bar{w} - \alpha \cdot \nabla_{\bar{w}} \left(L(\bar{w}) + \frac{\eta}{2\alpha} \cdot \|\bar{w}\|^2 \right)$$