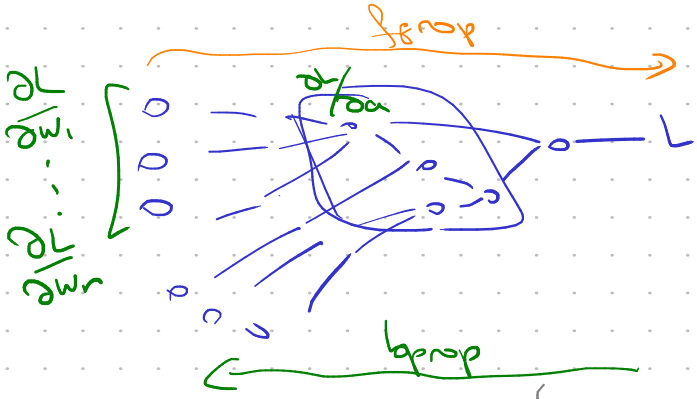
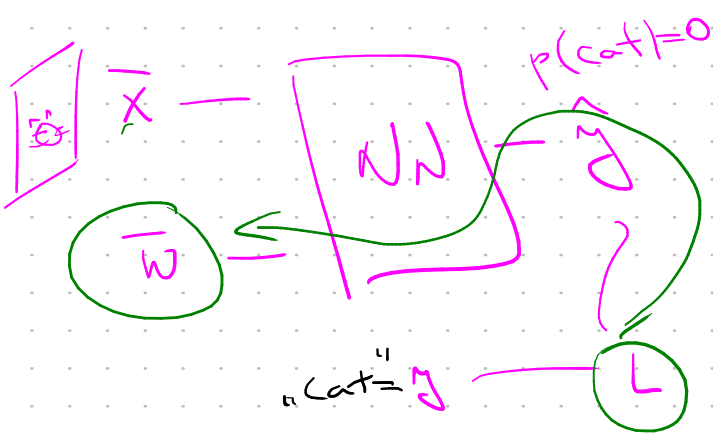


$L(\bar{w}) \xrightarrow{\bar{w}} \min$

① GD и SGD

GD: $\bar{w}_0, \bar{w}_1, \dots \to \bar{w}_{k+1} = \bar{w}_k - \eta \cdot \nabla_{\bar{w}} L(\bar{w}_k)$



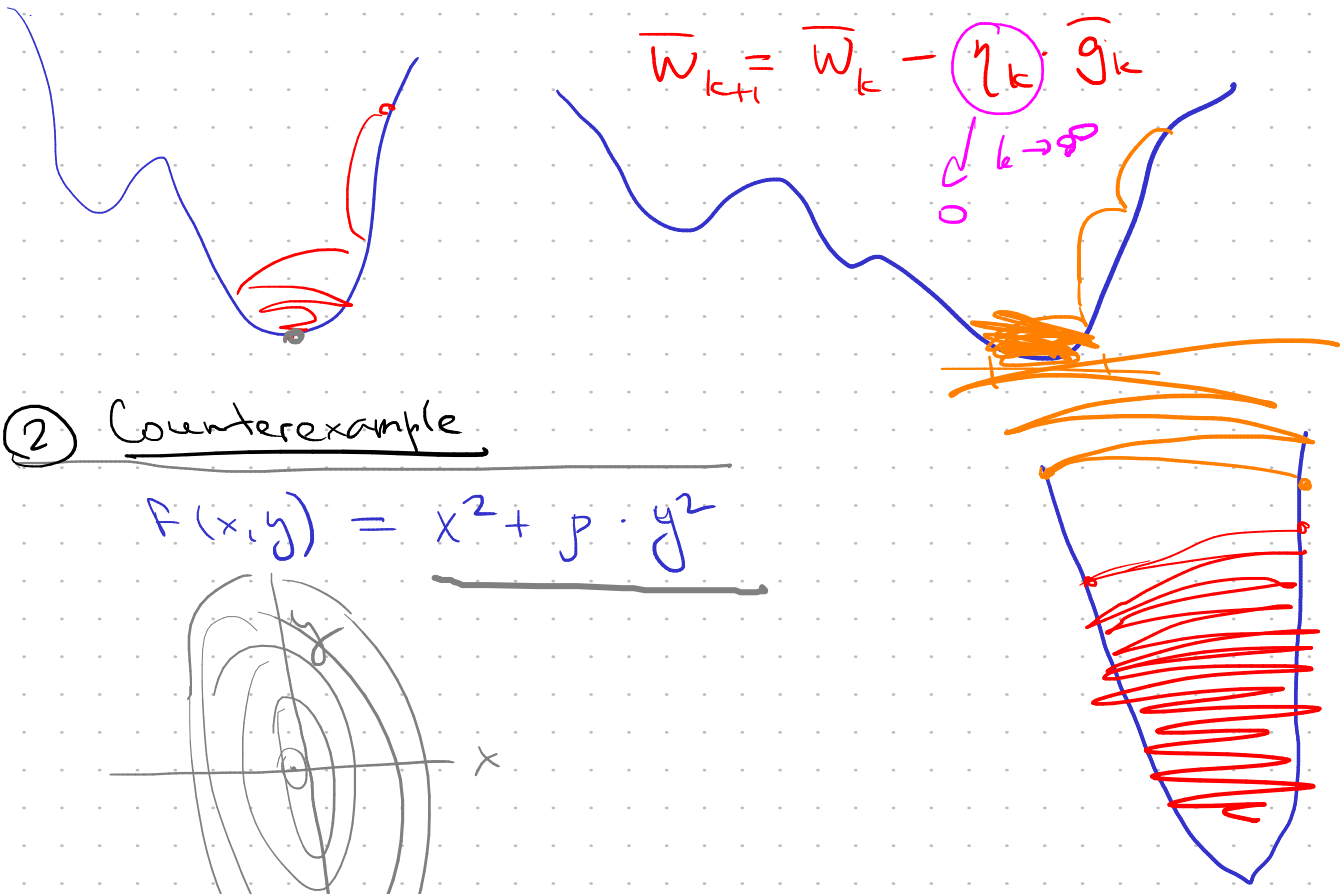


$$L(\bar{w}) = \frac{1}{N} \sum_{n=1}^{N} \ell(\bar{w}, \bar{x}_n, y_n)$$

SGD: $\bar{g}_k = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\bar{w}} \ell(\bar{w}, \bar{x}_i, y)$

$\bar{w}_{k+1} = \bar{w}_k - \eta_k \bar{g}_k$

$\eta_k \to 0$

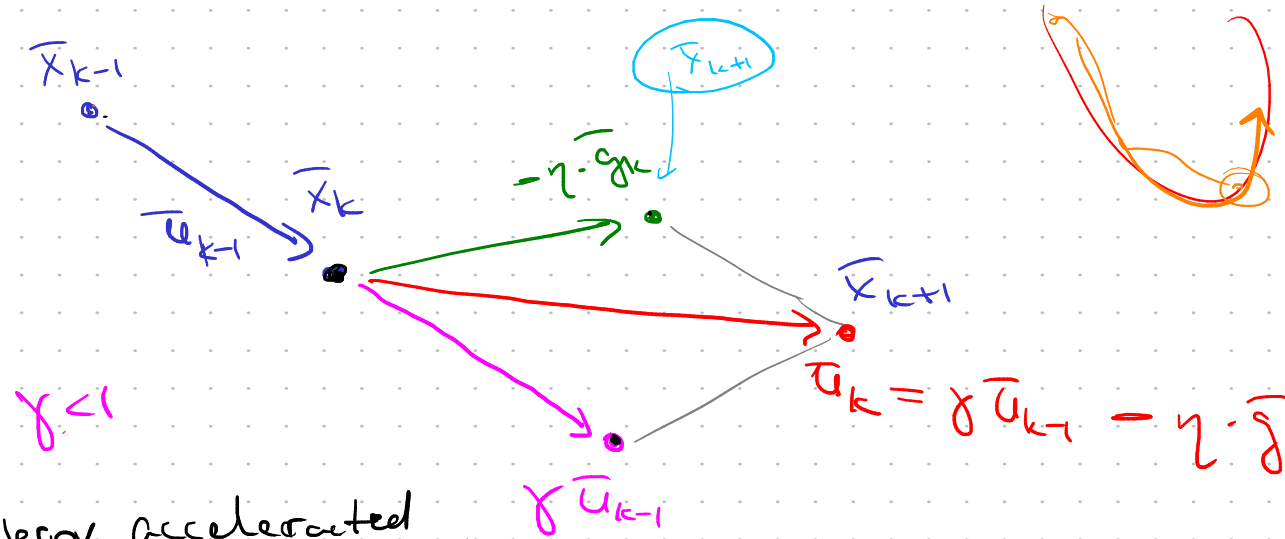


② Counterexample

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

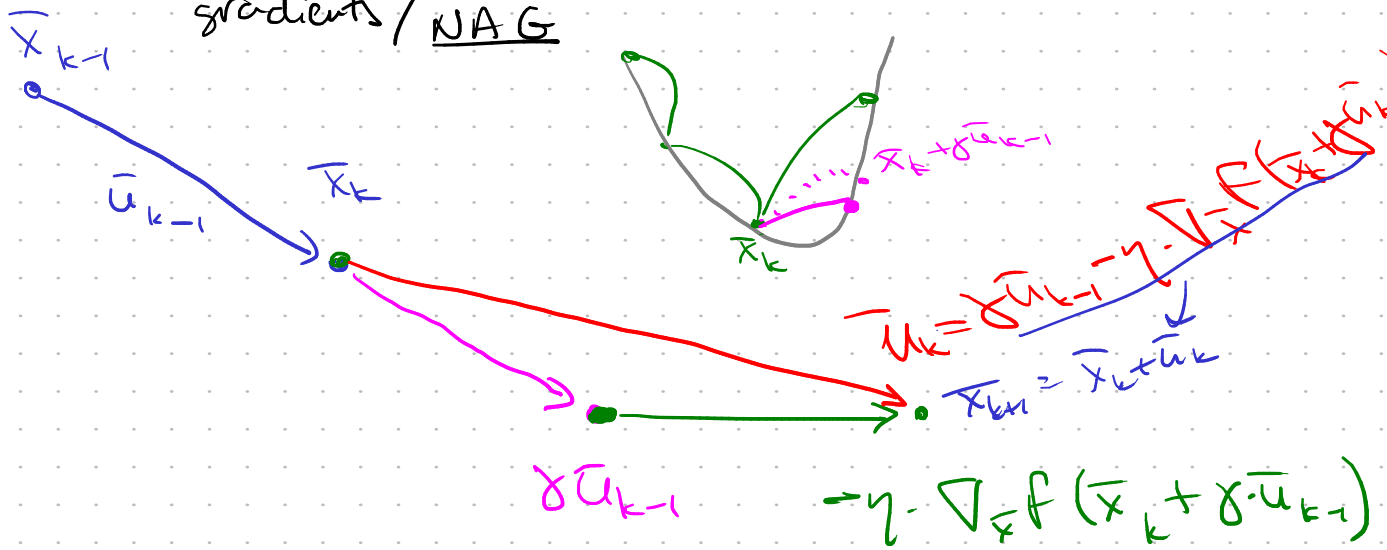
③ Momentum

$f(\bar{x}) \xrightarrow{\bar{x}} \min$



$\bar{u}_k = \gamma \bar{u}_{k-1} - \eta \cdot \bar{g}_k$

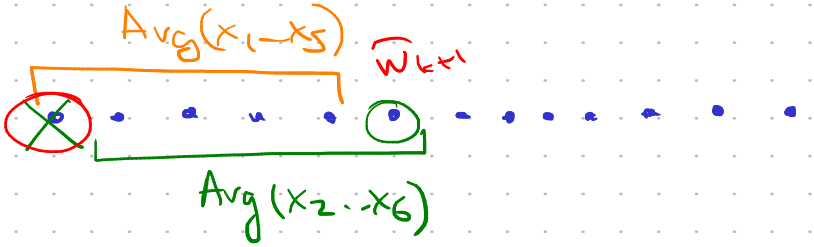
Nesterov accelerated gradients / NAG

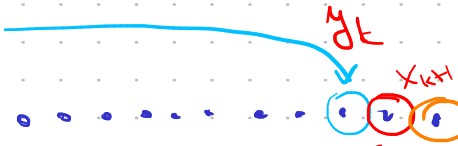


④ EMA (exponential moving average)

$\bar{w}_1 \; \bar{w}_2 \; \bar{w}_3 \; \ldots \; \bar{w}_k \ldots$

$x_1 \; x_2 \; x_3 \; x_4 \ldots \; x_k \; x_{k+1} \ldots$





$y_{k+2} = \beta \cdot x_{k+2} + (1-\beta) y_{k+1} =$

$= \beta \cdot x_{k+2} + (1-\beta) \cdot \beta \cdot x_{k+1} +$

$+ (1-\beta)^2 y_k =$

$y_{k+1} = \beta \cdot x_{k+1} + (1-\beta) y_k$

$= \beta x_{k+2} + (1-\beta) \cdot \beta x_{k+1} + (1-\beta)^2 \beta x_k + (1-\beta)^3 y_{k-1}$

$$y_k = \beta \cdot x_k + (1-\beta)\cdot\beta\, x_{k-1} + (1-\beta)^2\beta\, x_{k-2} + (1-\beta)^3\beta\, x_{k-3} + \dots$$

$$\dots (1-\beta)^k \cdot \beta \cdot x_0$$

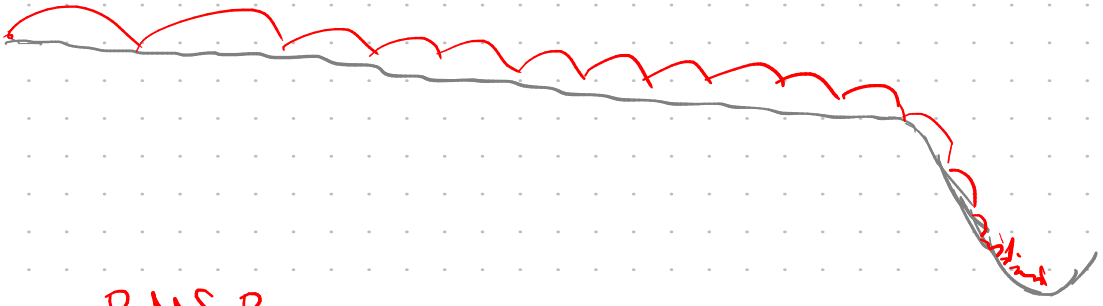
5 Adaptive SGD algorithms

– AdaGrad $\quad F(\bar{w})$

$\bar{g}_t$ – стох. градиент

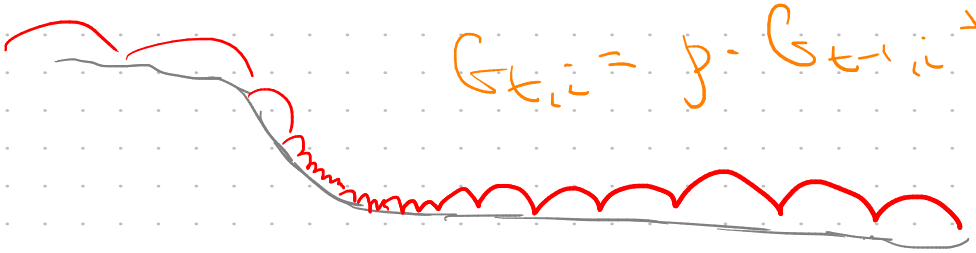
$$w_{t+1,i} = w_{t,i} - \frac{\eta}{\sqrt{G_{t,i}+\varepsilon}} \cdot g_{t,i}$$

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



– RMS Prop

$$G_{t,i} = \beta \cdot G_{t-1,i} + (1-\beta)\cdot g_{t,i}^2$$



– Adam

$$w_{t+1,i} = w_{t,i} - \frac{\eta}{\sqrt{G_{t,i}+\varepsilon}} \cdot m_{t,i}$$

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

$$m_{t,i} = \beta_1 \cdot m_{t-1,i} + (1-\beta_1)\, g_{t,i}$$

$\beta_1 = 0.9$

$\beta_2 = 0.995$

AdamW

Muon / Lion