SGD - stochastic gradient descent

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \to \min_{x \in R^p}$$

$$\nabla f(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x) \quad - \text{ gradient}$$

вместо $\nabla f(x_k)$ использовать g_k

$$\mathbb{E}\, g_k = \nabla f(x_k) \qquad \text{хотелось бы}$$

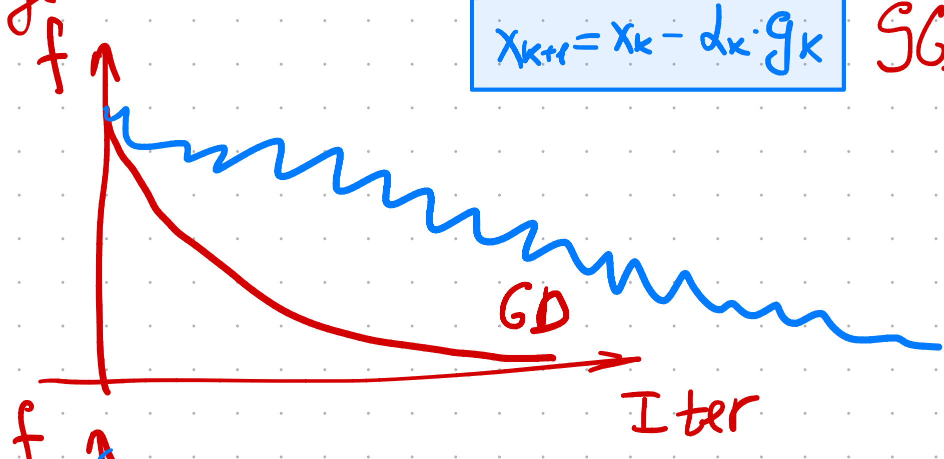
$$g_k = \nabla f_j(x_k) \leftarrow j \in [1, \dots N] \text{ выбирается случ. образом}$$

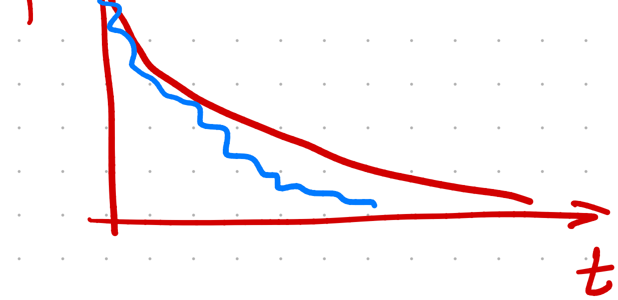
$$\mathbb{E}\, g_k = \sum_{i=1}^{N} \frac{1}{N} \cdot \nabla f_i(x_k) = \nabla f(x_k)$$

стох. градиент.

дешевле в N раз

$$\boxed{x_{k+1} = x_k - \alpha_k \cdot g_k} \quad \text{SGD}$$





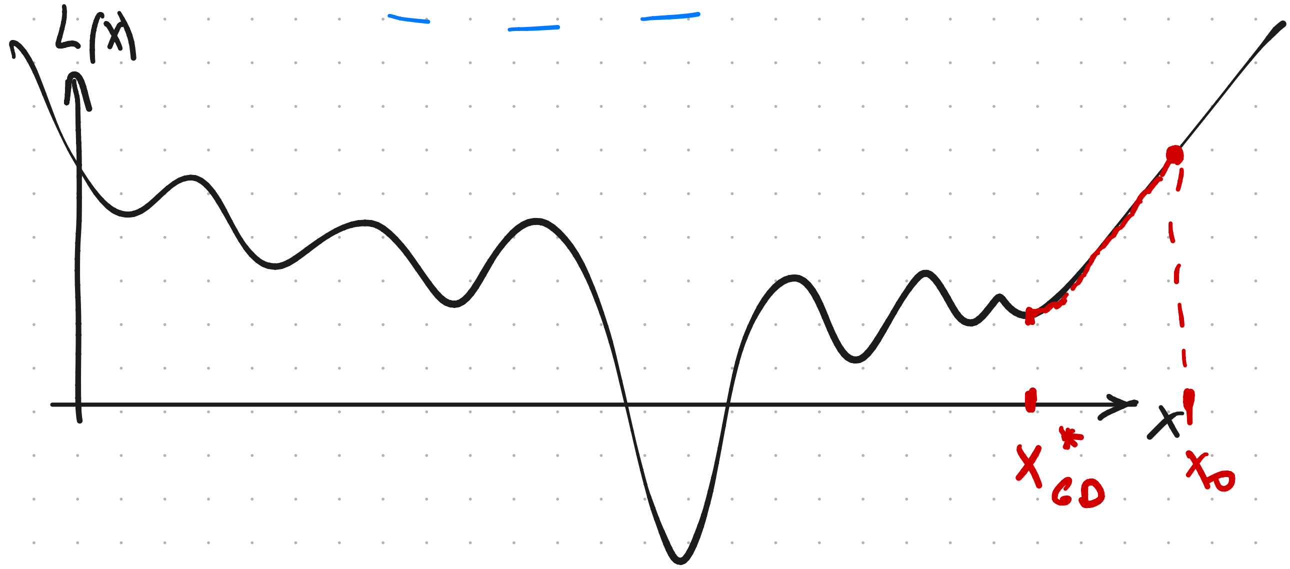
mini batch SGD

$$\boxed{g_k = \frac{1}{b}\sum_{j=1}^{b} \nabla f_{i_j}(x_k)}$$

выбрали b штук

$b = 1 \to$ SGD

$b = N \to$ GD



$b \ll N$

Вычисление g_k надо параллелить

$\nabla f_2(x_k)$ $\quad \nabla f_5(x_k)$ $\quad \nabla f_{228}(x_k)$

$g_k = \frac{1}{3}(\cdot + \cdot + \cdot)$

GPU

$\boxed{\cancel{b = 32}}$

$b = 100$

3p float

$\cancel{3}p \cdot 3\cancel{2}$ бит $= 16$ Гб $=$

$= \cancel{16} \cdot 10^9 \cdot \cancel{8}$

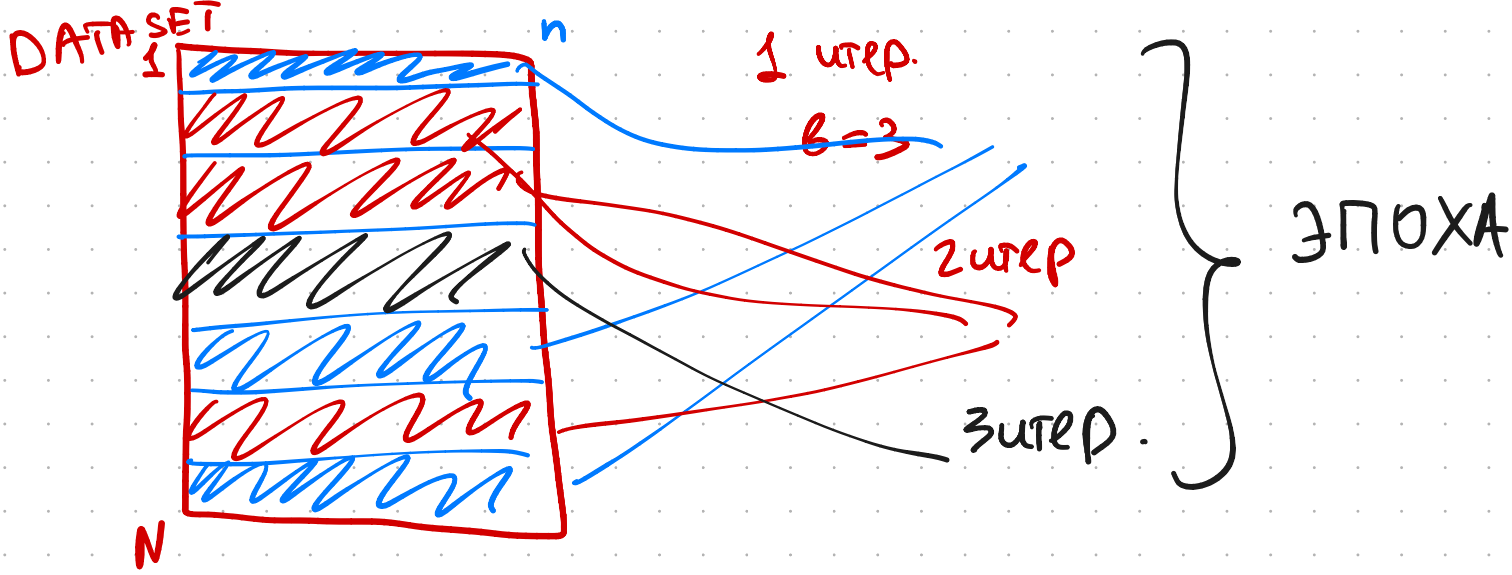
$p \approx 10^9$

$p \cdot \cancel{16} \cdot \cancel{32}$ бит $\approx \cancel{16 \cdot 8} \cdot 10^9$

100

$p = 10^7 - 10^8$

10^8



Как изменится время, затраченное на 1 эпоху при $b\uparrow\uparrow$?

MSE: $$L = \frac{1}{N}\sum_{i=1}^{N}\left(y_{pred}(W, D^i) - y_{true}^i\right)^2$$

$$f_i = \left(y_{pred}(w, D^i) - y_{true}^i\right)^2$$

$$\nabla_w f_i = 2\left(y_{pred}(w, D^i) - y_{true}^i\right)\cdot \frac{\partial y_{pred}}{\partial w}(w, D^i)$$

$$W_5 \cdot W_4 \cdot W_3 \cdot W_2 \cdot W_1 \cdot d_{input}$$