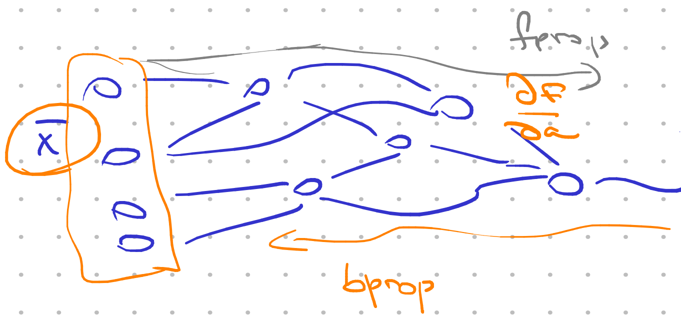
Computational graph



$\nabla_{\bar{x}} F(\bar{x})$

$F(\bar{x})$ $\quad \log p(D|\bar{w}) \xrightarrow{\bar{w}} \max$

$F(\bar{x}) \to \min \qquad \bar{x} := \bar{x} - \eta \cdot \nabla_{\bar{x}} F(\bar{x})$

$H = \nabla_{\bar{w}} \nabla_{\bar{w}} F(\bar{w})$

quasi-Newton methods $\quad$ L-BFGS

$$\boxed{F(\bar{w}) = \frac{1}{N} \sum_{n=1}^{N} L(\bar{x}_n, \bar{w})} \qquad q(\bar{y}) = \mathrm{Unif}(D)$$

$$\nabla_{\bar{w}} F = \frac{1}{N} \sum_n \nabla_{\bar{w}} L$$

$$\mathbb{E} f \approx \frac{1}{N} \sum_n f(\bar{x}_n)$$

Stochastic gradient descent

$$F(\bar{x}) = \mathbb{E}_{q(\bar{y})} f(\bar{x}, \bar{y}) \longrightarrow \min$$

$$G(\bar{x}) = \nabla_{\bar{x}} F(\bar{x})$$

$m << N$

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

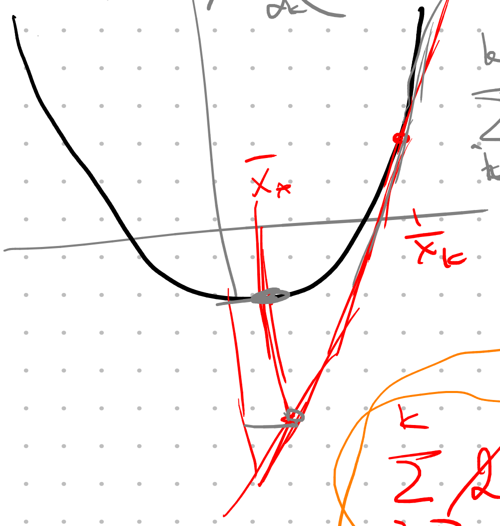
$$\bar{x}_{k+1} = \bar{x}_k - \alpha_k \hat{g}_k$$

$$\mathbb{E}\Big[\|\bar{x}_{k+1} - \bar{x}_*\|^2 = \|\bar{x}_k - \alpha_k \hat{g}_k - \bar{x}_*\|^2 =$$

$$= \|\bar{x}_k - \bar{x}_*\|^2 - 2\alpha_k \hat{g}_k^T (\bar{x}_k - \bar{x}_*) + \alpha_k^2 \|\hat{g}_k\|^2 \Big]$$

$$\mathbb{E} \|\bar{x}_{k+1} - \bar{x}_*\|^2 = \mathbb{E} \|\bar{x}_k - \bar{x}_*\|^2 - 2\alpha_k \underbrace{\bar{g}_k^T (\bar{x}_k - \bar{x}_*)}_{\geq \alpha_k (F(\bar{x}_k) - F(\bar{x}_*))} + \alpha_k^2 \mathbb{E} \|\bar{g}_k\|^2$$

$$\alpha_k f(\bar x_k) \geq \alpha_k\left(f(\bar x_k) + (\bar x_* - \bar x_k)^T \bar g_k\right)$$



$$\sum_{k=0}^{k} \left| \; \mathbb{E}\|\bar x_{k+1}-\bar x_*\|^2 - \mathbb{E}\|\bar x_k-\bar x_*\|^2 \leq -2\alpha_k\left(f(\bar x_k)-f(\bar x_*)\right) + \alpha_k^2\,\mathbb{E}\|\hat g_k\|^2 \right.$$

$$\sum_{i=0}^{k} \cancel{2}\alpha_i\left(f(\bar x_i)-f(\bar x_*)\right) \leq \frac{\mathbb{E}\|\bar x_0-\bar x_*\|^2 - \cancel{\mathbb{E}\|\bar x_{k+1}-\bar x_*\|^2}}{2} + \frac{\sum_{i=0}^{k}\alpha_i^2\,\mathbb{E}\|\hat g_i\|^2}{2}$$

$$f\left(\frac{\sum \alpha_i x_i}{\sum \alpha_i}\right) \leq \frac{\sum \alpha_i f(\bar x_i)}{\sum \alpha_i}$$

$$\underbrace{\sum\alpha_i \cdot f\left(\frac{\sum \alpha_i x_i}{\sum \alpha_i}\right)} - \sum\alpha_i \cdot f(\bar x_*) \leq \frac{1}{2}\left(\dots\right)$$

$$\mathbb{E} f(\bar x_k') - f(\bar x_*) \leq \frac{\mathbb{E}\|\bar x_0-\bar x_*\|^2 + \sum_{i=0}^{k}\alpha_i^2\,\mathbb{E}\|\hat g_i\|^2}{2\sum_{i=0}^{k}\alpha_i}$$

$$\|\bar x_0-\bar x_*\| \leq R, \quad \mathbb{E}\|\hat g_k\|^2 \leq G^2$$

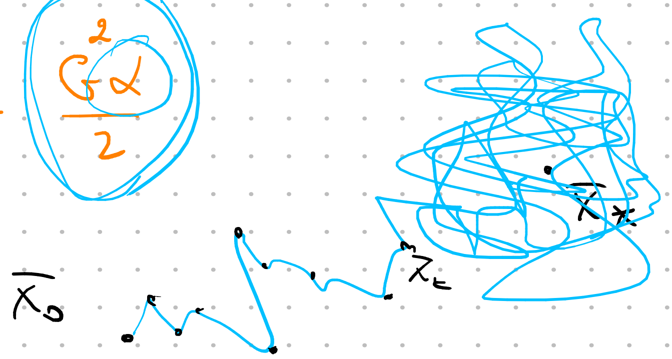
$$\boxed{\mathbb{E} f(\bar x_k') - f(\bar x_*) \leq \frac{R^2 + G^2\sum_{i=0}^{k}\alpha_i^2}{2\sum_{i=0}^{k}\alpha_i}}$$

$$\sum_{i=0}^{k}\alpha_i^2 < \infty \qquad \sum_{i=0}^{k}\alpha_i \to \infty$$

$$\alpha_i = \frac{1}{i}$$

$$\alpha_i = \alpha$$

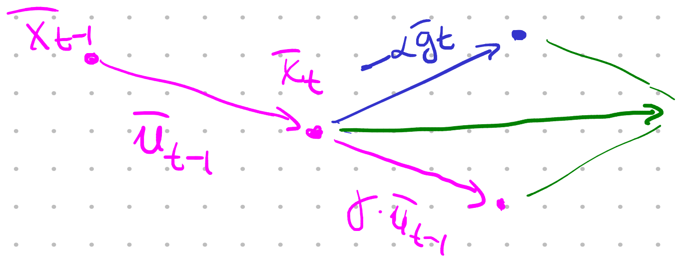
$$\mathbb{E} f_k - f(\bar x_*) \leq \overset{\to 0}{\frac{R^2}{2\alpha k}} + \frac{G^2\alpha}{2}$$



Варианты SGD

$$\bar x_{t+1} = \bar x_t - \alpha \bar g_t$$

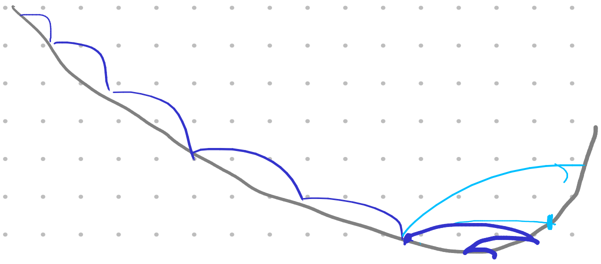
SGD with momentum

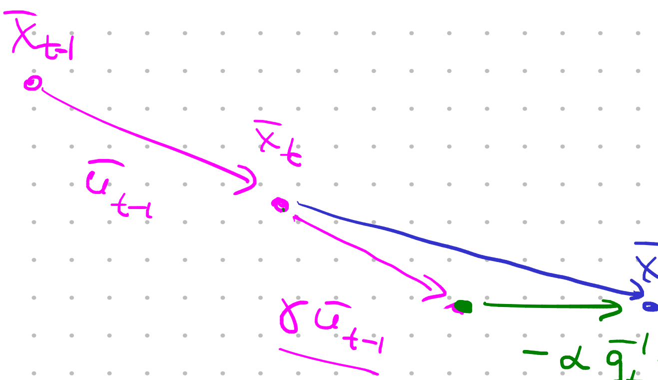


$\gamma = 0.99, 0.995$ $\quad \bar u_t$

$$\bar x_{t+1} = \bar x_t + (\gamma \bar u_{t-1} - \alpha \bar g_t)$$

Nesterov accelerated gradient (NAG)





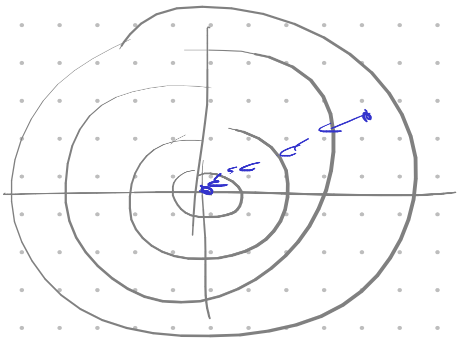
$$\bar x_{t+1} = \bar x_t + \gamma \bar u_{t-1} - \alpha \bar g_t'$$

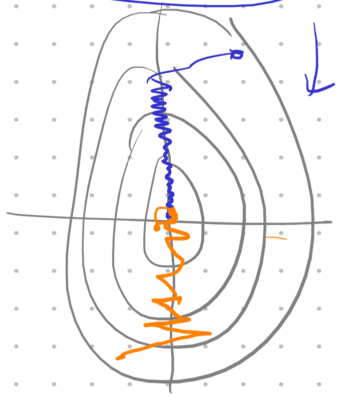
$$-\alpha \bar g_t' = -\alpha \nabla_{\bar x} f(\bar x_t + \gamma \bar u_{t-1})$$

$$f(\bar x) = x_1^2 + x_2^2$$

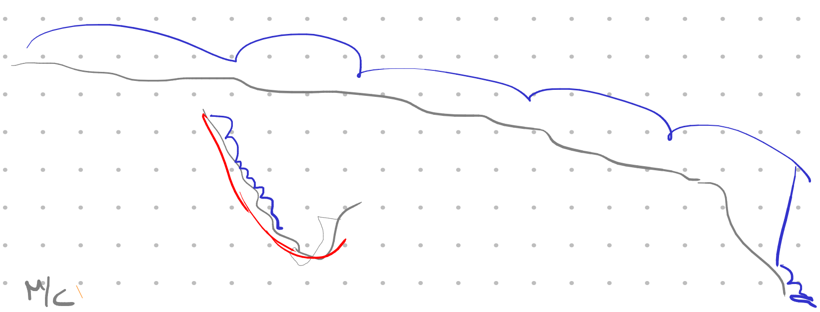
0.001

$$x_1^2 + 0.1 x_2^2$$





Adaptive SGD



Adagrad

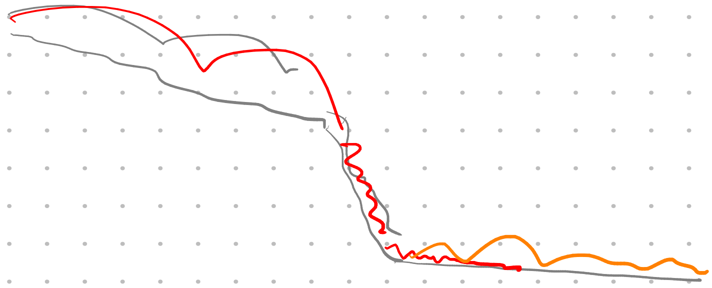
$$G_0 = 0$$

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

c c м/с

$$\bar x_{t+1} = \bar x_t - \frac{\alpha}{\sqrt{G_t + \varepsilon}} \odot \bar g_t$$

м/с



RMS Prop

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

$f(\bar{x})$ – метры
$\bar{x}$ – секунды

Adadelta

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

с с м/с

$$R_{t_i} = \rho R_{t-1,i} + (1-\rho)\, u^2_{t_i}$$

$$\bar{x}_{t+1} = \bar{x}_t - \frac{1}{2} H^{-1} \cdot \bar{g}_t$$

с с $(\text{м}/\text{с}^2)^{-1}$ · м/с

$$\underline{x_{t+1}} = \underline{\bar{x}_t} - \alpha \cdot \frac{\sqrt{R_t + \varepsilon}}{\sqrt{G_t + \varepsilon}} \odot \bar{g}_t$$

Adam

$$\bar{x}_{t+1} = \bar{x}_t - \frac{\alpha}{\sqrt{G_t + \varepsilon}} \cdot \bar{m}_t \text{, где}$$

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

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

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