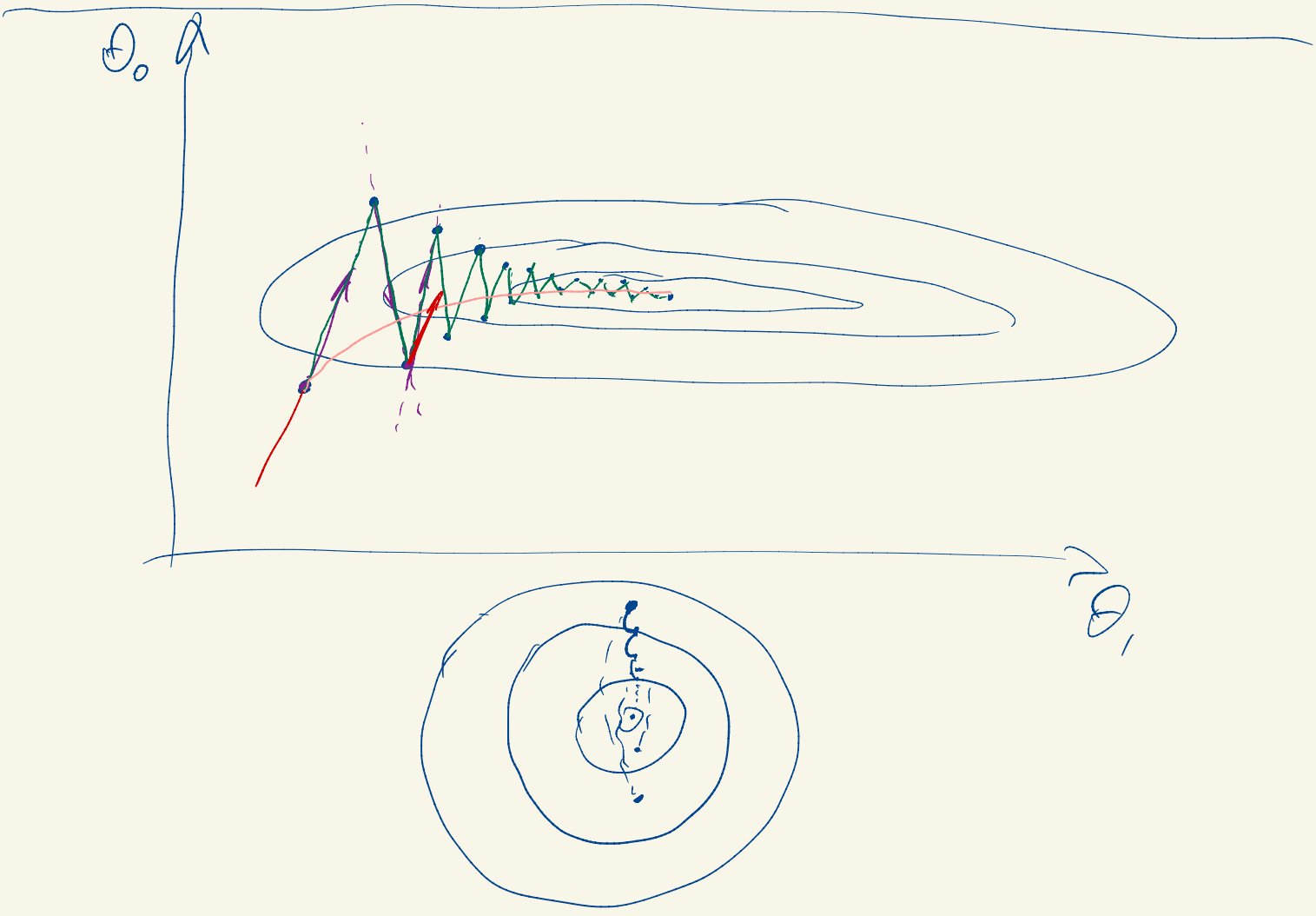
θ_0

$$\theta_t = \theta_{t-1} - \eta \nabla_\theta \mathcal{L}\left(\{x_i, y_i\}_{B_{t-1}}, \theta_{t-1}\right)$$

итерация:

Эпоха: обучение на всём наборе данных (1 итерация)

10'000 bs: 10



$$\theta_t = \theta_{t-1} - \eta \nabla_\theta \mathcal{L}(B_{t-1}, \theta_{t-1}) \quad | \text{-SGD}$$

$$\begin{cases} m_t = (1-\beta) \nabla_\theta \mathcal{L}(B_{t-1}, \theta_{t-1}) + \beta m_{t-1} \\ \theta_t = \theta_{t-1} - \eta m_t \end{cases}$$

SGD + momentum

импульсный метод

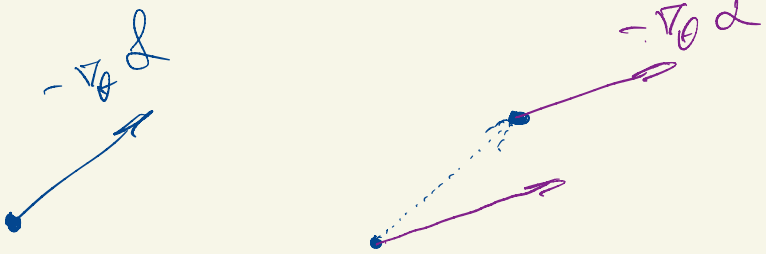
$\beta = 0 \Rightarrow$ SGD + momentum $=$ SGD

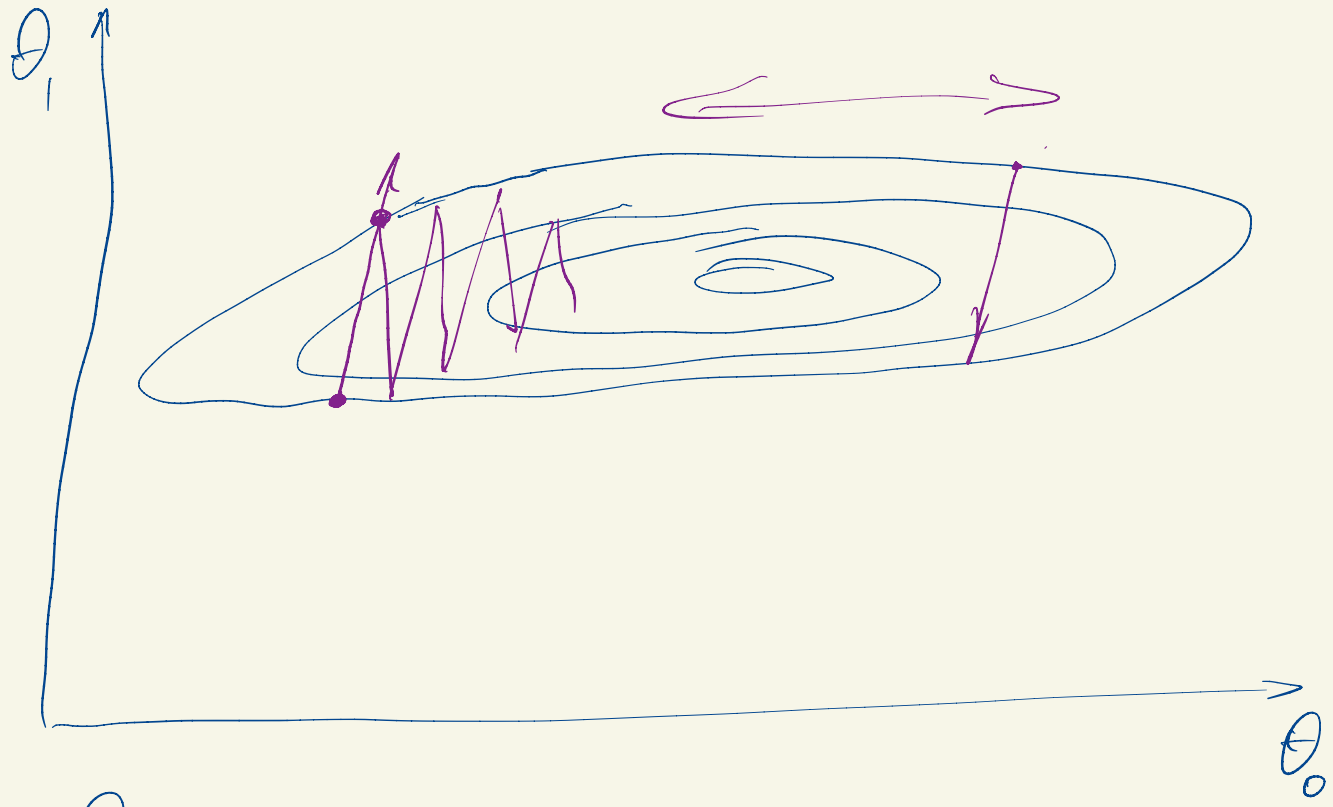
$\beta = 1 \Rightarrow$ скорее всего, оптимизация не сойдётся

Nesterov momentum

$$m_t = (1-\beta) \nabla_\theta \mathcal{L}(B_{t-1}, \theta_{t-1}) + \beta m_{t-1} \quad \text{- SGD+ momentum}$$

$$m_t^{\text{N}} = (1-\beta) \nabla_\theta \mathcal{L}(B_{t-1}, \theta_{t-1} - \eta m_{t-1}) + \beta m_{t-1}$$





$$\theta_t = \theta_{t-1} - \eta\, m_t$$

$$S_t = \beta S_{t-1} + (1-\beta)\left(\nabla_\theta \mathcal{L}(\theta_{t-1})\right)^2 \text{ — поэлементно}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\nabla \mathcal{L}(\theta_{t-1})}{\sqrt{S + \varepsilon}}$$

(element-wise)

Адаптивный градиент
AdaGrad
RMSProp

AdaGrad + Momentum = AdaM

$$m_t = (1-\beta_1)\,\nabla_\theta \mathcal{L}(\theta_{t-1}, B_{t-1}) + \beta_1 m_{t-1}$$

$$S_t = \beta_2 S_{t-1} + (1-\beta_2)\left(\nabla_\theta \mathcal{L}(\theta_{t-1}, B_{t-1})\right)^2$$

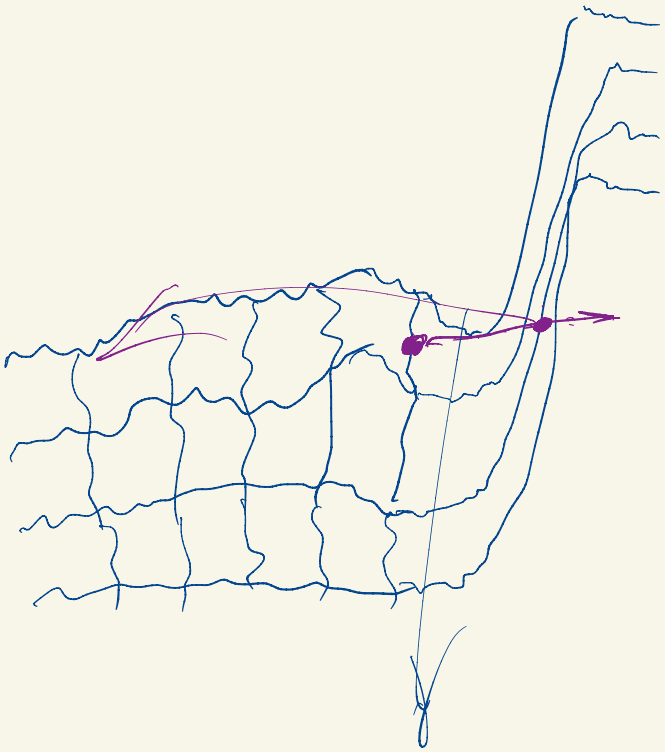
$$\theta_t = \theta_{t-1} - \eta \frac{m_t}{\sqrt{S_t + \varepsilon}}$$

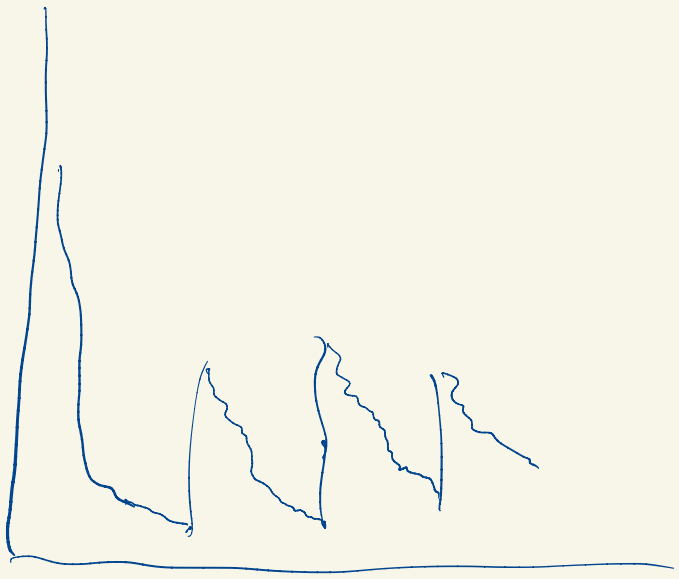
$$S_t^* = \max(S_t, S_{t-1})$$

AmsGrad

$\beta_1 = 0{,}9$

$\beta_2 = 0{,}99$





Gradient clipping

$$g_t = \nabla_\theta \mathcal{L}(B_t, \theta_t)$$

① global norm clipping

$$g_t^* = \frac{g_t}{\|g_t\|} \cdot S_g \qquad S_g = 10$$

② component-wise clipping

$$g_{t_i}^* = \begin{cases} g_{t_i}, & \text{if } g_{t_i} \le L_g \\ L_g, & \text{if } g_{t_i} > L_g \end{cases}$$