1. θ_0 (M) $\mathcal{L}(\theta_i, B)$

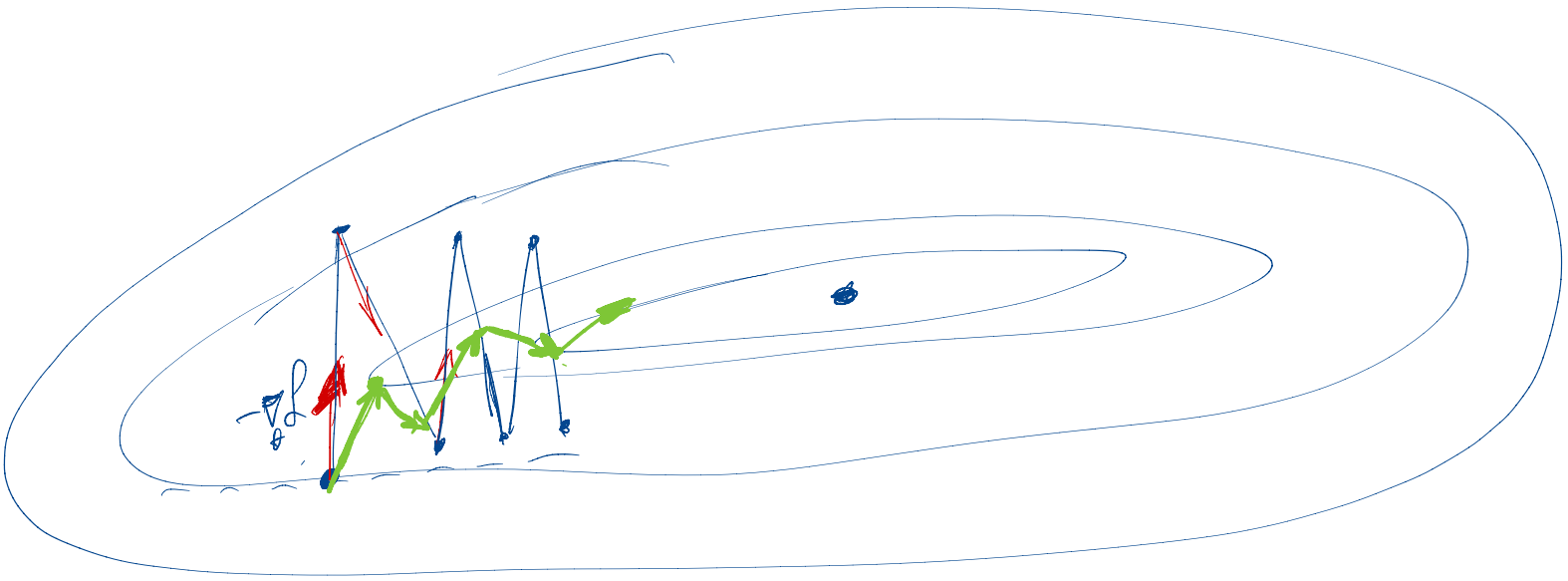
$\bar{g} = \nabla_\theta \mathcal{L}$ (M)

$\theta_{t+1} = \theta_t - \eta \bar{g}_t$ $\quad \bar{g}_t = \nabla_\theta \mathcal{L}(\theta_t, B_t)$

Stochastic Gradient Descent

SGD

(LM)



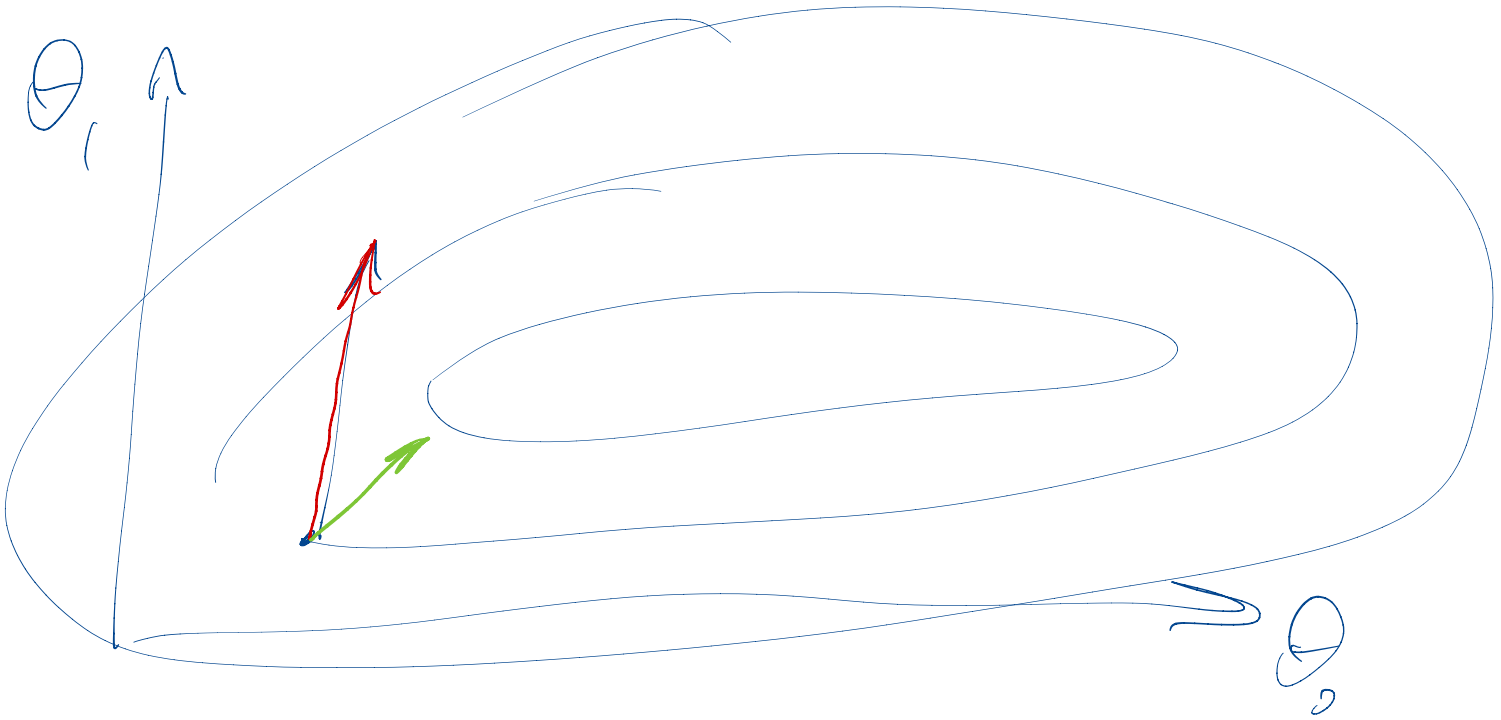
Моментум (Импульсный метод)

$\theta_t : M$

$\bar{g}_t = \nabla_\theta L$ (M)

$\bar{m}_t = (1-\beta)\,\bar{g}_t + \beta\,\bar{m}_{t-1}$ (M) $\beta = 0,9$

$\theta_{t+1} = \theta_t - \zeta\,\bar{m}_t$ (3M)



$$S_t = \left(\bar{g}_t\right)^2$$

$$S_t = \beta S_{t-1} + (1-\beta)\left(\bar{g}_t\right)^2$$

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

Адаптивный градиент (AdaGrad)

SGD + AdaGrad + Momentum

Adam

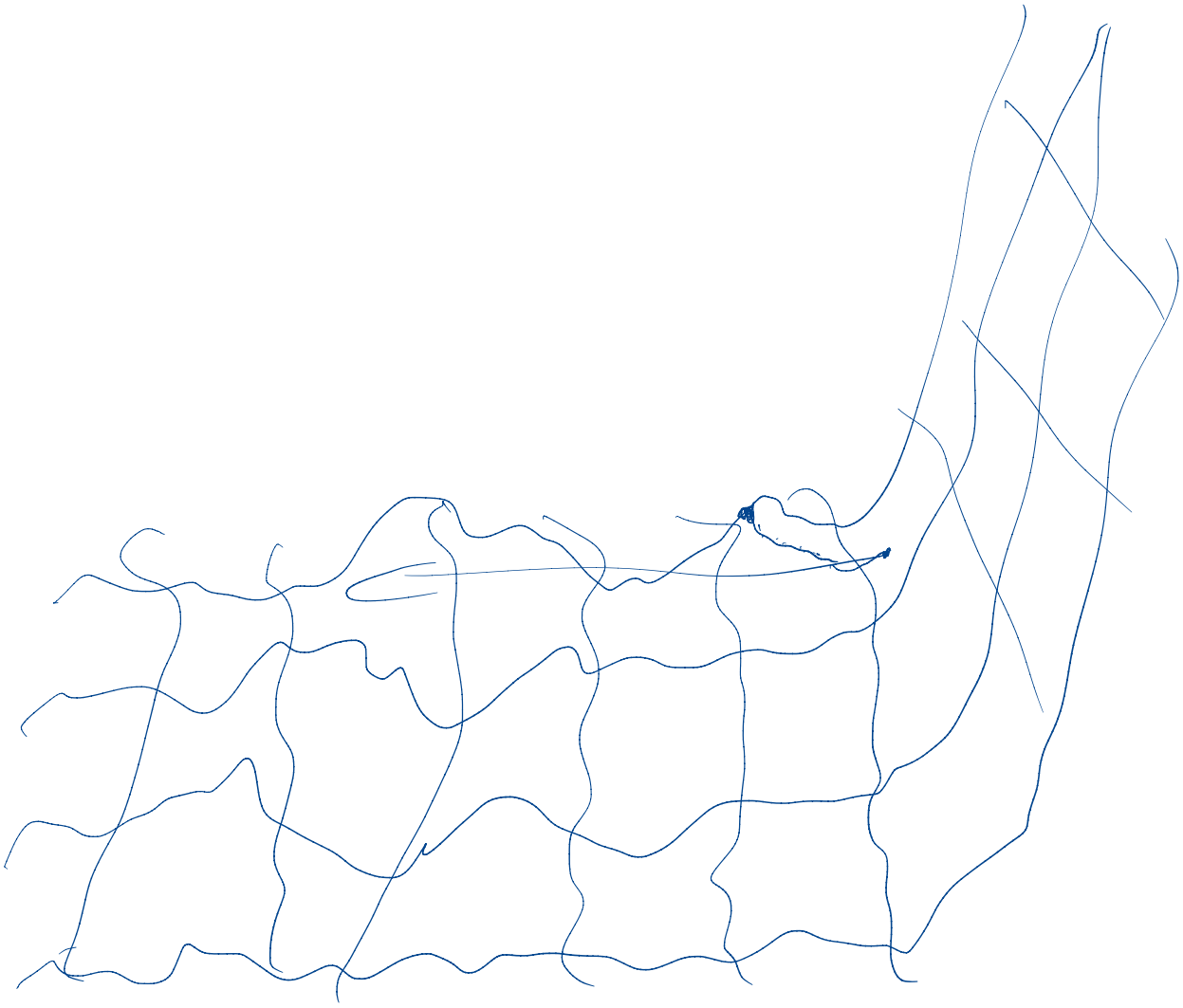
θ_0 (M)

$\bar{m}_t = (1-\beta_1)\,\bar{g}(\theta_t, \mathcal{B}_t) + \beta_1 \bar{m}_{t-1}$ (M, M)

$s_t = \beta_2 s_{t-1} + (1-\beta_2)\,\bar{g}_t^2$ (M)

(4 M)

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



Gradient clipping

$$\bar{g}^* = \frac{\bar{g}}{\|g\|} \cdot s_g$$

$$g_t^* = \begin{cases} g_t, & \text{if } \|g\| \le L \\ L g, & \text{if } \|g\| > L \end{cases}$$