∂

$$\hat{y}_i = F(x_i, \theta)$$

$$\mathcal{T} = \{(x_i, y_i)\}_N$$

$$\mathcal{L} = \mathcal{L}(\theta, \{(x_i, y_i)\})$$

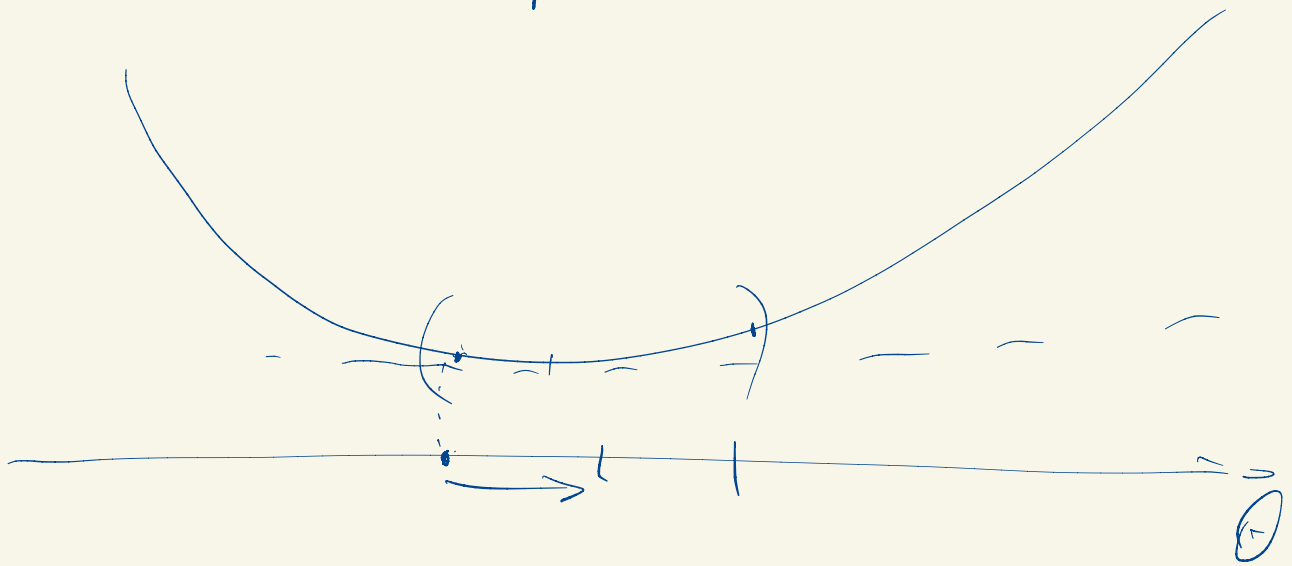
$\vec{\nabla_\theta \mathcal{L}}$ $\vec{\theta}$

① $\theta_0 \sim \mathcal{N}(0, 1)$ η - шаг обучения learning rate

② $\vec{g} = \nabla_\theta \mathcal{L}(\theta_t) = \nabla_\theta \mathcal{L}_t = \nabla_\theta \mathcal{L}(\theta_t, \mathcal{T})$

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_t$$

③ $\mathcal{L} \le \varepsilon \Rightarrow$ stop $\Rightarrow \theta^*$



SGD — Стохастический градиентный спуск

Stochastic gradient descent

① $\theta_0 \sim \mathcal{N}(0, 1)$

η - шаг обучения (learning rate)

n - batch size

② $g = \nabla_\theta \mathcal{L}(\theta_t, D)$ — D - подвыборка $\mathcal{T}$ размера n

$\theta_{t+1} = \theta_t - \eta g$

③ $\mathcal{L} \le \varepsilon \Rightarrow$ stop $\Rightarrow \theta^*$

---

① $n^* = k n$

$\eta^* = k \eta$