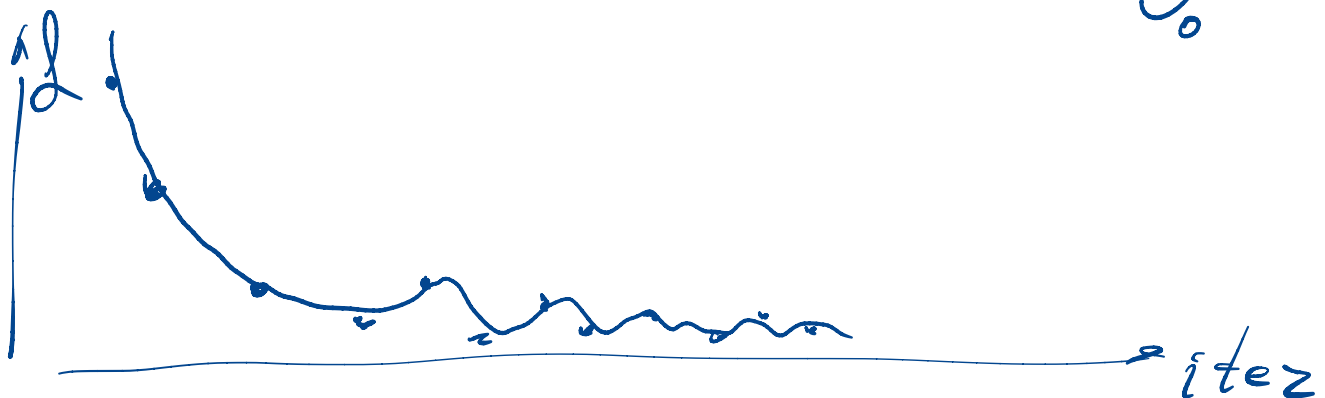
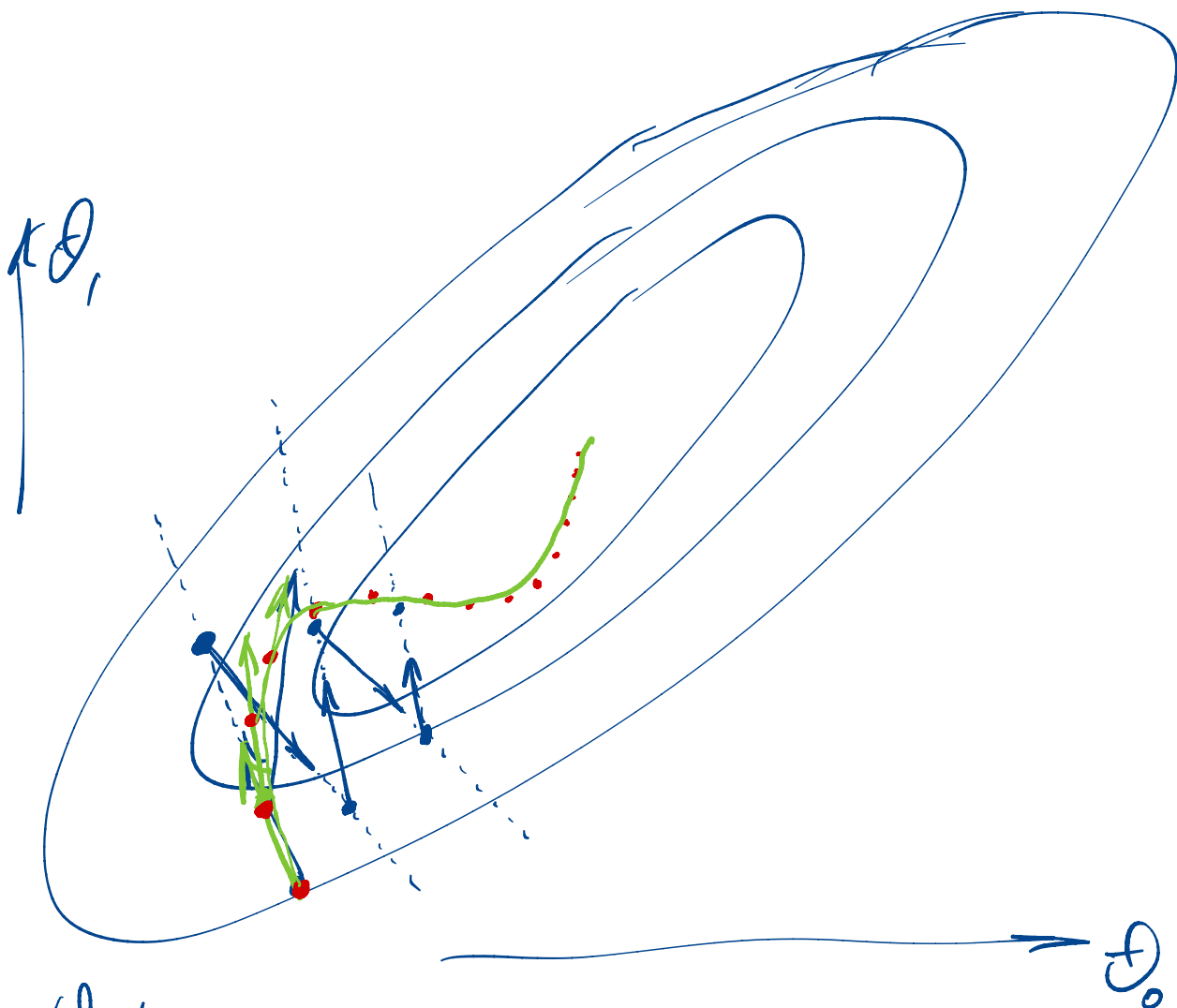


SGD

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \nabla_{\Theta} \mathcal{L}(\mathcal{D}, \Theta^{(t)})$$

$\eta = \text{learning rate}$



Импульсная метод (Momentum)

$$\vec{g}^{(t)} = \nabla_{\theta} \mathcal{L}(\beta, \theta^{(t)})$$

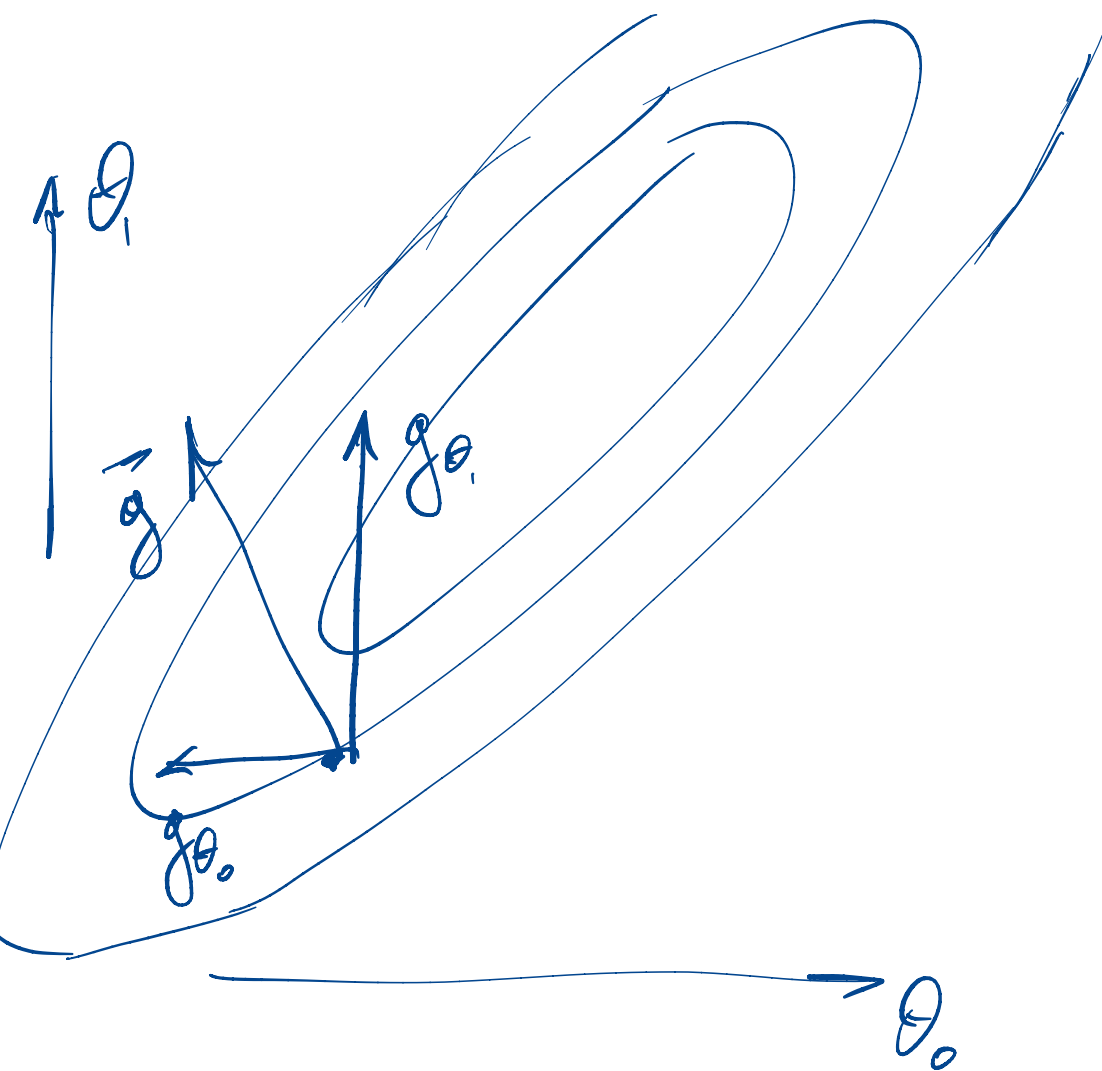
$$\vec{m}^{(t+1)} = \beta \vec{m}^{(t)} + (1 - \beta) \vec{g}^{(t)}$$

$$\beta \sim 0,9 - 0,99$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \vec{m}^{(t)}$$

Эффекты:

- ① Сглаженная оптимизация
- ② Потенциально более быстрая оптимизация
- ③ Потенциально повышенная обобщающая способность.



Может наблюдаться слабый дисбаланс масштабов \vec{g} по разным θ_i

$$g = \nabla_{\theta} L \quad g = \frac{g}{\|g\|}$$

$$S^{(t+1)} = \beta_2 S^{(t)} + (1 - \beta_2) g^2$$

$$g' = \frac{g}{\sqrt{S + \epsilon}}$$

$$\epsilon \sim 10^{-8}$$

$$\beta_2 \sim 0.999$$

AdaGrad

SGD + AdaGrad Momentum

Adam

$$g^{(t)} = \nabla_{\theta} \mathcal{L}(B, \theta^{(t)})$$

$$m^{(t)} = m^{(t-1)} \beta_1 + (1 - \beta_1) g^{(t)}$$

$$s^{(t)} = s^{(t-1)} \beta_2 + (1 - \beta_2) (g^{(t)})^2$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{m^{(t)}}{\sqrt{s^{(t)} + \epsilon}}$$