Gradient descent.

1) $\Theta_0 = \Theta_{init} \sim N(\cdot\,\cdot)$. $(x, y) \in D$ $|D| = N$

2) $\Theta_n = \Theta_{n-1} - \alpha \frac{1}{N}\sum_{i=0}^{N} \nabla_\Theta L(f(x_i,\Theta_{n-1}), y_i)$

D

---

1) $\Theta_0 = \Theta_{init} \sim N(\cdot\,\cdot)$. SGD

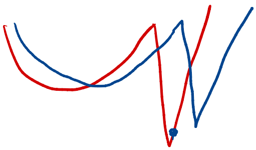
2) d is random subset of D of fixed size $\underline{k}$.

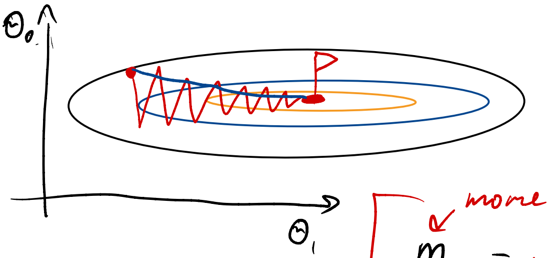
mini batch

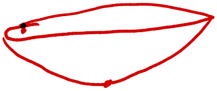
3) $\Theta_n = \Theta_{n-1} - \alpha \frac{1}{k}\sum_{(x,y) \in d} \nabla_\Theta L(f(x_i,\Theta_{n-1}), y_i)$

1) От k зависит „уровень шума".

2) $k = N$ $\alpha = \alpha_0$. $k = N/5$ $\alpha = \alpha_0/5$.

3) 



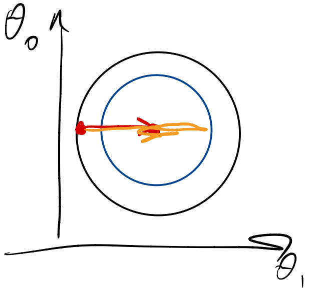


$\boxed{\beta = 0.9}$

momentum

$$m_t = \nabla_{\theta_{t-1}} L (1-\beta) + m_{t-1}\cdot\beta$$

$$\theta_t = \theta_{t-1} - \alpha m_t$$



SGD momentum

$$m_t = \nabla_\theta L(\theta_{t-1}) \cdot (1-\beta) + m_{t-1}\cdot\beta$$

$$m_t = \nabla_\theta L(\theta_{t-1} + \beta m_{t-1}) \cdot (1-\beta) + m_{t-1}\cdot\beta$$

lookahead gradient

Nesterov momentum

$$v_t = \beta v_{t-1} + (1-\beta)\left(\nabla_\theta L(\theta_{t-1})\right)^2$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\nabla_\theta L(\theta_{t-1})}{\sqrt{v_t + \epsilon}}$$

AdaGrad
RmsProp

AdaGrad + Momentum

AdaM.

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

$$V_t = \beta_2 V_{t-1} + (1-\beta_2)\left(\nabla_\theta L(\theta_{t-1})\right)^2.$$

$$\theta_t = \theta_{t-1} - \alpha \frac{m_t}{\sqrt{V_t} + \epsilon}$$

AMsgrad.

$$\hat{V}_t = \beta_2 V_{t-1} + (1-\beta_2)\left(\nabla_\theta L(\theta_{t-1})\right)^2.$$

m_t

$$V_t = \max(\hat{V}_t, V_{t-1}).$$

Nadam

AdaBelief.

$$\boxed{V_t = V_{t-1} \cdot \beta_2 + (1-\beta_2)\left(m_t - \nabla_\theta L(\theta_{t-1})\right)^2.}$$

---

n - number of ops.

~2n $O(n)$.

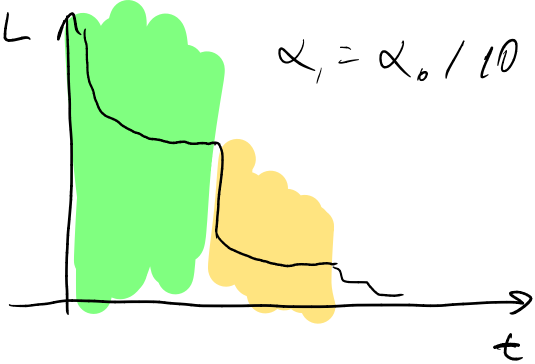
backprop.

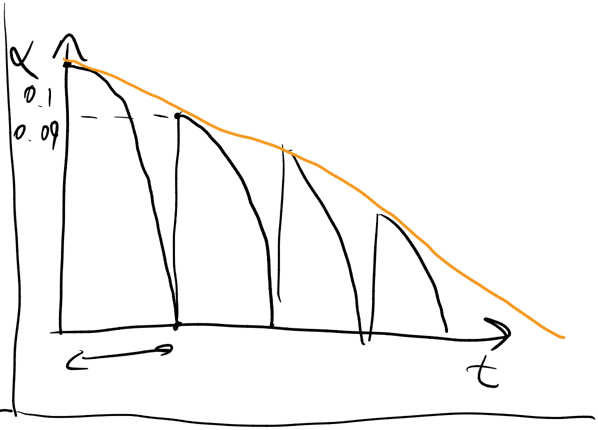
P - num of params $|\theta|$

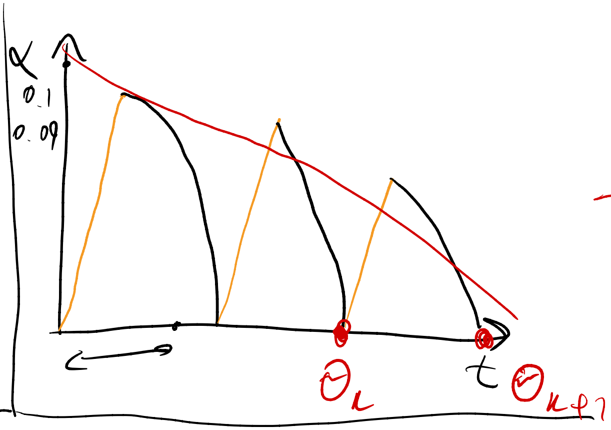
$\underbrace{2n} + 2p$ - momentum

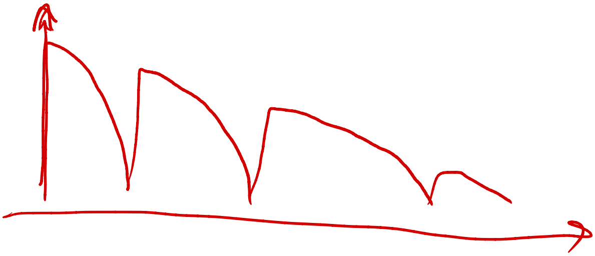
$\underbrace{2n} + 3p$ - adam

$2n + 2p$ — Nesterov.









$$\Theta = \frac{1}{n} \sum \Theta_i$$