Optimization

GD

SGD

mini batch SGD

X, y - ℓ штук

$\ell \gg 1$

B - batch size

L - Loss function

w / Θ - параметры

$a_w(x) \to \hat{y}$

$$L = \frac{1}{B} \sum_i^B (y_i - \hat{y}_i)^2 \to \min_w$$

$\hat{y}_i = a_w(x_i)$

t - шаг

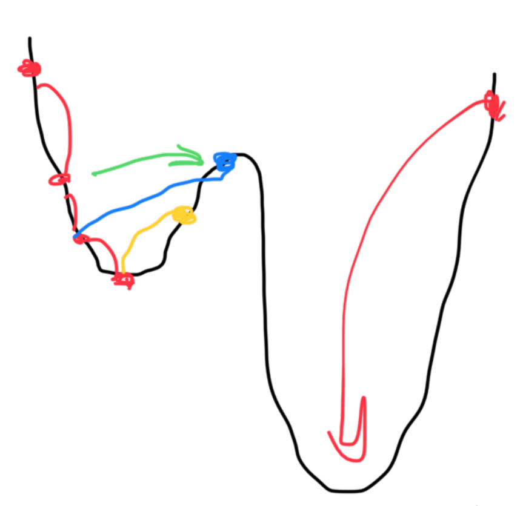
L^t

SGD:

$g_t = \nabla_w L^t(w^{t-1})$

$w^t = w^{t-1} - \eta_t g_t$

η_t - lr



$m_t = \mu \cdot m_{t-1} + g_t$

$w_t = w_{t-1} - \eta_t m_t$

momentum

$lr \sim 10^{-2} \to 10^{-7} = \eta$

$m_0 = 0$

$0 \le \mu \le 1$ $\quad \mu = 0{,}8 - 0{,}9$

↳ typ. value

$L \to L^* = L + \lambda \|w\|_2^2$

$\nabla_w \|w\|_2^2 = 2w$

$g_t \to g_t^* = \nabla_w L^t(w_{t-1}) + \underbrace{\lambda \cdot 2}_{\lambda} w_{t-1}$

λ PyTorch

weight_decay

Adam

(AdaGrad, RMSProp)

1) Momentum

2) Adaptive learning rate

$g_t = \nabla_w L(w_{t-1}) + \boxed{\lambda w_{t-1}}$

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \to$ элементный

$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \leftarrow$ степень шаг

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$w_t = w_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$

поэлементно

защита от ÷0

$\boxed{m_0 = v_0 = 0}$

$\beta_1 = 0{,}9$

$\beta_2 = 0{,}999$

$\varepsilon = 10^{-8..9}$

$m_t = (1-\beta_1) g_t + \beta_1 m_{t-1} =$

$= (1-\beta_1) g_t + \beta_1 (1-\beta_1) g_{t-1} + \beta_1^2 \cdot m_{t-2} = (1-\beta_1) g_t + \beta_1(1-\beta_1) g_{t-1} +$

$+ \ldots + \beta_1^{t-1} (1-\beta_1) g_1 + \cancel{\beta_1^t m_0} =$

$= \underbrace{\sum_{i=0}^{t-1} \beta_1^i (1-\beta_1) g_{t-i}}$

$\hat{m}_t = \frac{m_t}{1-\beta_1^t} = \frac{1-\beta_1}{1-\beta_1^t} \sum_{i=0}^{t-1} \beta_1^i g_{t-i} =$

$= \frac{\cancel{1-\beta_1}}{\cancel{(1-\beta_1)}(\sum_{k=0}^{t-1} \beta_1^k)} \sum_{i=0}^{t-1} \beta_1^i g_{t-i} \boxed{=}$

$(1-\beta_1)(1+\beta_1+\beta_1^2) \ominus$

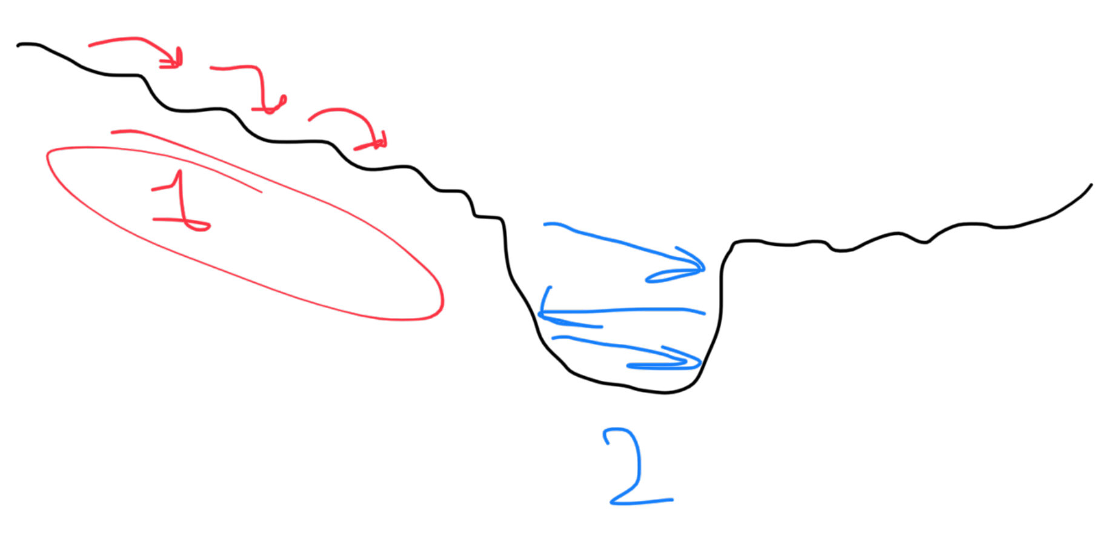
$\ominus\ 1 + \cancel{\beta_1} + \cancel{\beta_1^2} - \cancel{\beta_1} - \cancel{\beta_1^2} - \beta_1^3$

$\boxed{=} \frac{\sum_{i=0}^{t-1} \beta_1^i g_{t-i}}{\sum_{k=0}^{t-1} \beta_1^k} = \left\{ \alpha_i = \frac{\beta_1^i}{\sum_{k=0}^{t-1} \beta_1^k}, \right.$

$\left. \alpha_i > 0;\ \sum_{i=0}^{t-1} \alpha_i = 1 \right\} = \underbrace{\sum_{i=0}^{t-1} \alpha_i g_{t-i}} =$

$= \mathbb{E} g$

$\hat{v}_t = \mathbb{E} g^2 = \mathrm{Var}(g) + (\mathbb{E} g)^2$



1: $\mathrm{Var}(g) \to 0$

$w_t = w_{t-1} - \eta_t \underbrace{\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}}$

$\to \frac{\mathbb{E} g}{\sqrt{\cancel{\mathrm{Var}(g)} + (\mathbb{E} g)^2} + \varepsilon} \approx 1$

2. $\mathrm{Var}(g) \,!= 0$

$\frac{\mathbb{E} g}{\sqrt{\mathrm{Var}(g) + (\mathbb{E} g)^2} + \varepsilon}$

не выкинуть $\Rightarrow$ $\eta_t \cdot *$

$* < 1$

AdamW

$g_t = \nabla_w L^t(w_{t-1})$

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_1) g_t^2$

$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$w_t = w_{t-1} \boxed{(1 - \eta_t \lambda)} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$

weight decay $\quad \frac{\eta_t \lambda}{\sqrt{\hat{v}_t}}$

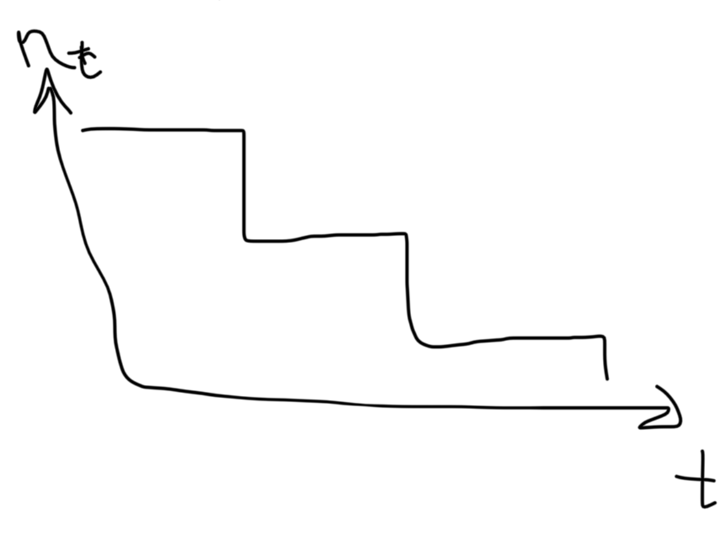
LBFGS-B

η_t

Constant

$\eta_t = \eta_0$

Step LR



$\eta_0 : 1 \le t < t_0$

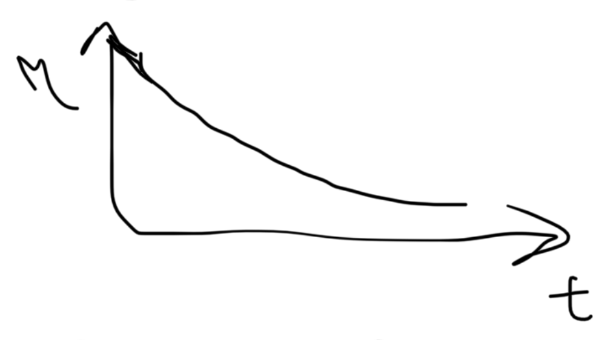
$\eta_1 : t_0 \le t < t$

⋮

Exponential LR

$\eta_t = \gamma \eta_{t-1}$

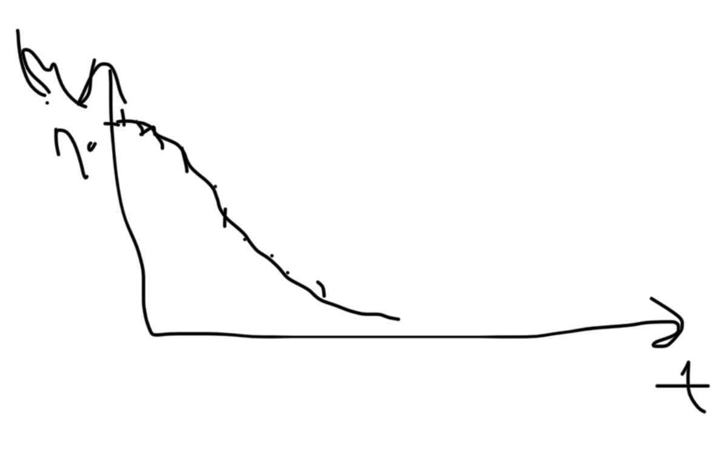
$\gamma < 1$



Cosine LR

$\eta_t = \eta_0 \cdot \frac{1}{2}\left(1 + \cos \frac{\pi t}{T}\right)$

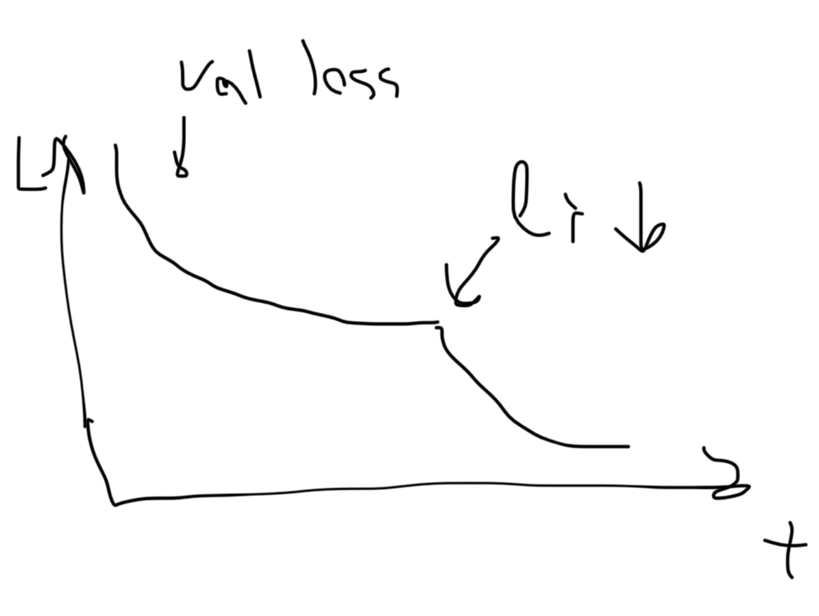
T = epochs



1, 2, 3, 4

Linear warmup

Reduce LR on plateau



Cosine with Restarts