Optimization

GD - градиентный спуск

SGD - стохастический градиентный спуск

mini batch SGD - мини батч SGD

обычно размер степени двойки

Основные обозначения

x, y - ℓ штук

ℓ много больше 1

B - batch size

L - loss function

w / θ - параметры

$a_w(x)$ - модель

$\hat{y}$ - предсказания модели

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

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

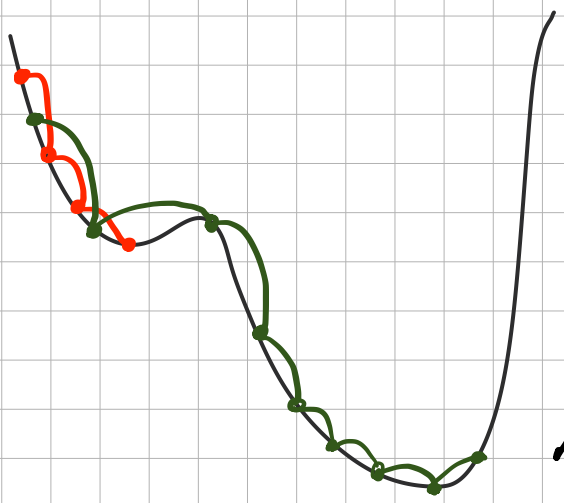
t - шаг оптимизации

SGD

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

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

SGD + momentum



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

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

μ - значение момента (гиперпараметр)

Значения гиперпараметров

$lr \in (1e^{-2}; 1e^{-7})$ - η

$m_0 = 0$

$\mu \in [0; 1]$ - momentum обычно 0,8 - 0,9

+ weight decay

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

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

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

λ - weight_decay

Adam

Используем идеи:

1) Моментум
2) Adaptive Learning Rate

$$g_t = \nabla_w L(w_{t-1}) + \underbrace{\lambda w_{t-1}}_{\text{weight decay}}$$

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

$$V_t = \beta_2 V_{t-1} + (1-\beta_2) \cdot g_t^2 \leftarrow \text{моментум}$$

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

$$\hat{V}_t = \frac{V_t}{1-\beta_2^t}$$

$$w_t = w_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{V}_t} + \varepsilon}$$ защита от деления на ноль

$$m_0 = V_0 = 0$$

$$\beta_1 = 0{,}9 \qquad \beta_2 = 0{,}999 \qquad \varepsilon = 10^{-9}$$

Разложение Adam

раскладываем по формуле

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

раскладываем по формуле

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

$$= (1-\beta_1)\, g_t + \beta_1 (1-\beta_1)\cdot g_{t-1} + \beta_1^2 \cdot (1-\beta_1)\cdot g_{t-2} + \beta_1^2 \cdot m_{t-3} =$$

Если продолжим раскладывать до m_0, то получим:

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

$$\dots + \beta_1^{t-1}(1-\beta_1)\cdot g_1 + \beta_1^t \underbrace{m_0}_{\to 0} =$$

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

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

$$= \frac{1-\beta_1}{(1-\beta_1)\left(\sum_{k=0}^{t-1} \beta^k\right)} \cdot \sum_{i=0}^{t-1} \beta_1^i\, g_{t-i}$$

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

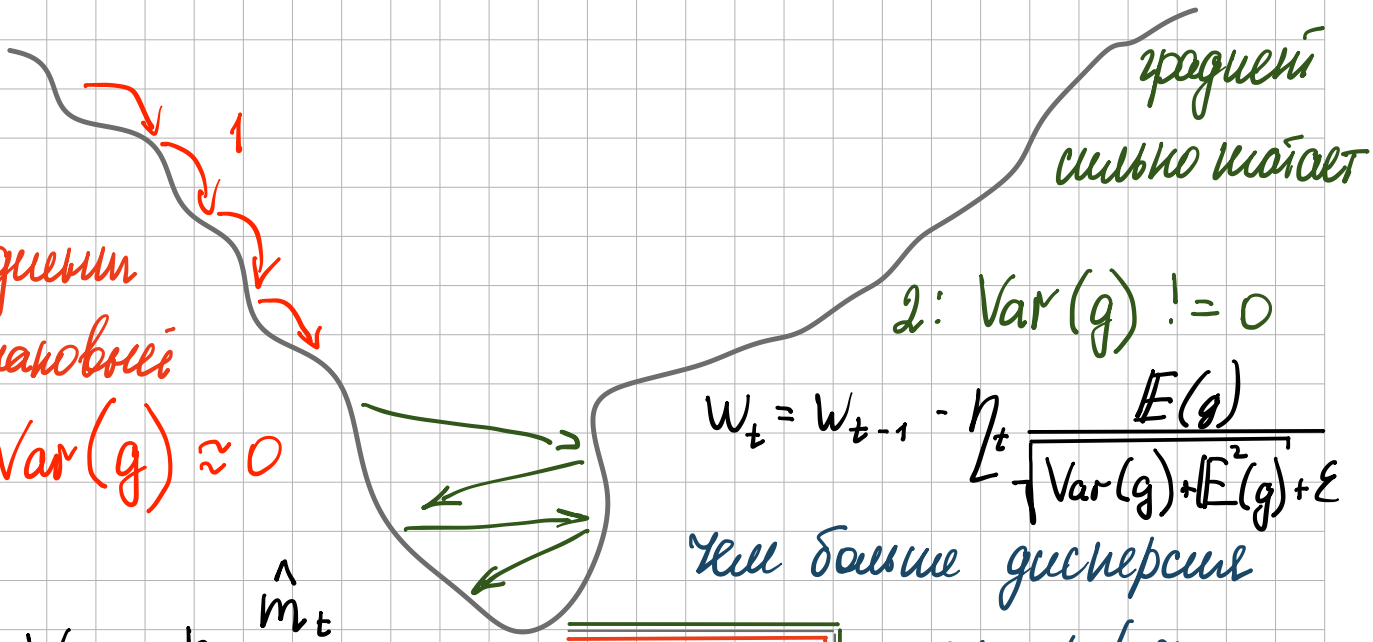
$$\alpha_i = \frac{\beta_1^i}{\sum_{k=0}^{t-1} \beta_1^k}$$

$$\alpha_i > 0$$

$$\sum_{i=0}^{t-1} \alpha_i = 1$$

$$= \mathbb{E}(g)$$

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



$$W_t = W_{t-1} - \eta_t \frac{\mathbb{E}(g)}{\sqrt{Var(g) + \mathbb{E}^2(g)} + \varepsilon}$$

Чем больше дисперсия тем короче будем шаг.

$$W_t = W_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{V}_t} + \varepsilon} =$$

$$= W_{t-1} - \eta_t \frac{\mathbb{E}(g)}{\sqrt{Var(g) + \mathbb{E}^2(g)} + \varepsilon} \approx$$

$$\approx W_{t-1} - \eta_t$$

AdamW

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

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

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

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

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

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

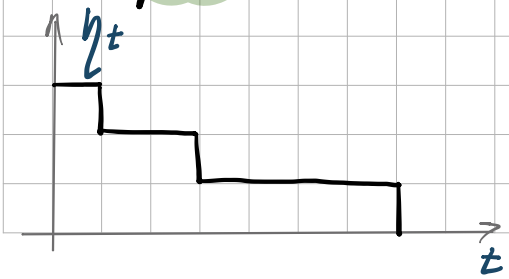
$\tilde{\lambda}$ ↑ weight_decay

Изменение шага обучения

Constant

$$\eta_t = \eta_0$$

StepLR



η_0: $1 \leq t < t_0$

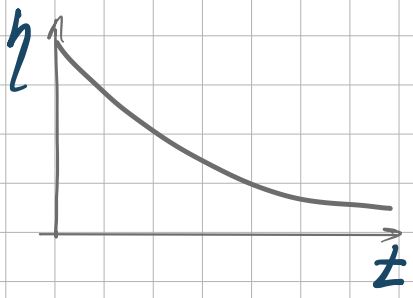
η_1: $t_0 \leq t < t_1$

⋮

ExponentialLR

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

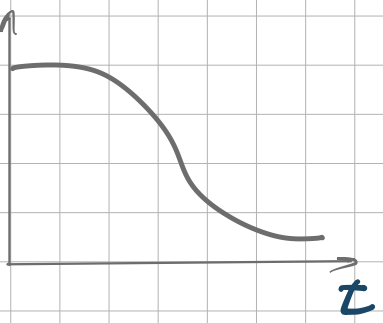
$$0 < \gamma < 1$$



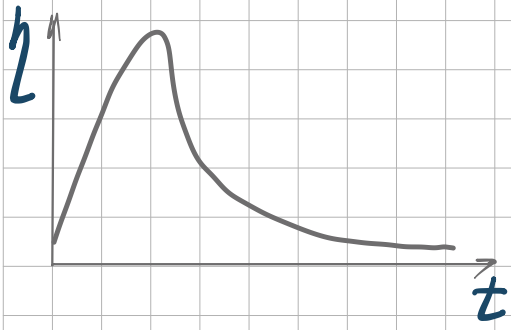
Cosine LR

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

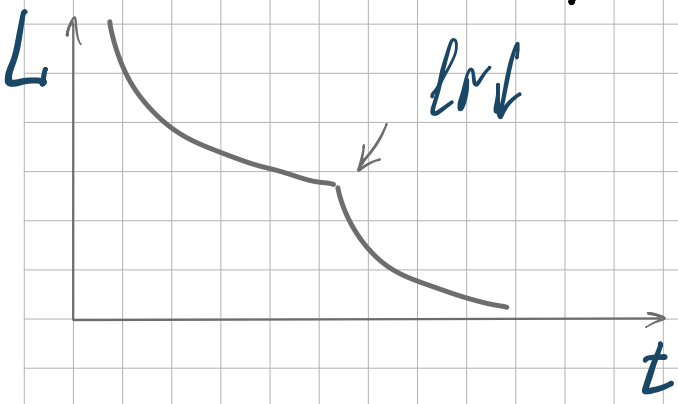
T = epochs



Linear warmup



Reduce LR on plateau



Cosine with Restarts