Стохастическая градиентная оптимизация

$$\Theta^* = \underset{\Theta}{\arg\min}\, \mathcal{L}(\mathcal{T}, \Theta)$$

① Θ_0

② (Adam) $\beta_1, \beta_2, \eta,$ C - предикат остановки

B, $t = 0$

③ 1) $\{X, Y\}^B - \mathcal{T}^{(B)}$

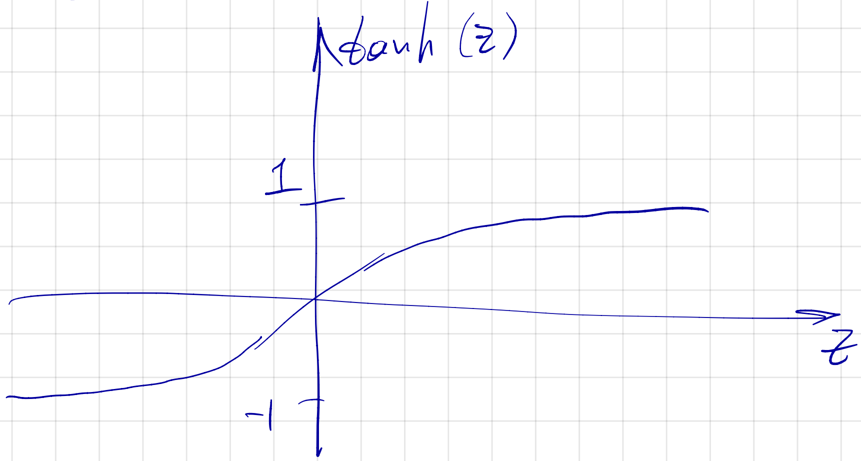
2) $\mathcal{L} = \mathcal{L}(\mathcal{T}^{(B)}, \Theta^{(t)})$

3) $g = \nabla_\Theta \mathcal{L}$ (Adam) $\Rightarrow g^*$

4) $\Theta^{(t+1)} = \Theta^{(t)} - \eta g^*$

5) C? $\Rightarrow$ stop

$\sigma(z) = \tanh(z)$



$$\hat{y} = \varphi(\ldots(\theta^{(2)}\sigma(\theta^{(1)}\sigma(\underbrace{\theta^{(0)}x}_{z^{(0)}})))\ldots)$$

$$\mathrm{Var}\, z_i = \mathrm{Var}\, \theta_i h_i = \mathrm{Var}\, \theta_i \;\; \mathrm{Var}\, h_i$$

$$z^{(j)} = \sum_{i=1}^{m} \theta_i h_i$$

$$\mathrm{Var}\, z^{(j)} = \sum_{i=1}^{m} \mathrm{Var}\,\theta_i \; \mathrm{Var}\, h_i = m \,\mathrm{Var}\, \theta_i \;\; \mathrm{Var}\, h_i$$

$\sigma = \tanh$

$$\mathrm{Var}\, \overbrace{\sigma(z^{(j)})}^{h} = \mathrm{Var}\, z^{(j)}$$

$$\mathrm{Var}\, h^{(l+1)} = m \underbrace{\mathrm{Var}\, \theta^{(l+1)}}_{c} h^{(l)}$$

$$\mathrm{Var}\, h^{(l+k)} = m^k c^k \;\mathrm{Var}\, h^{(l)}$$

$$(mc) \approx 1$$

$$\mathrm{Var}\,\theta^{(l)} \approx \frac{1}{m^{(l)}}$$

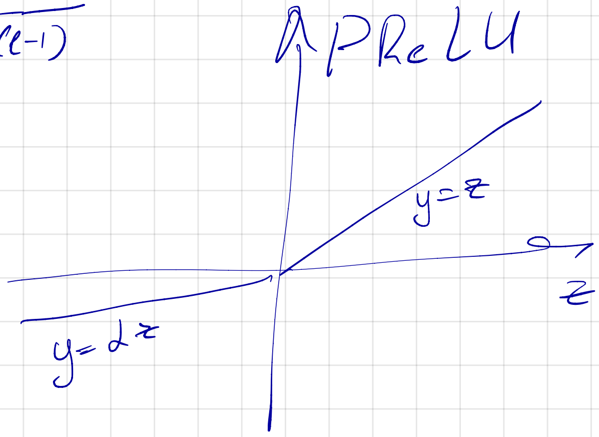
$$\theta \sim \mathcal{N}\left(0, \sigma^2 = \frac{1}{m}\right)$$

$$\theta \sim U\left(-\frac{\sqrt{m}}{2}; \frac{\sqrt{m}}{2}\right)$$

Kaiming He 2015

ReLU: $\mathrm{Var}\,\theta^{(l)} = \dfrac{2}{n^{(l-1)}}$ $\quad n$ – ширина слоя $(l-1)$

PReLU: $\mathrm{Var}\,\theta^{(l)} = \dfrac{2}{(1+\alpha^2)\, n^{(l-1)}}$


PReLU
y=z
z
y=αz