1) Введение

posterior distribution, likelihood, prior distribution

$$p(\bar\theta|D) = \frac{p(D|\bar\theta)\,p(\bar\theta)}{p(D)}$$

парам. модели, данные

loss functions

ML: $-\log p(D|\bar\theta) \xrightarrow{\bar\theta} \min$ / max

MAP: $\log p(\bar\theta|D) = \log p(D|\bar\theta) + \log p(\bar\theta) + \text{Const} \xrightarrow{\bar\theta} \max$

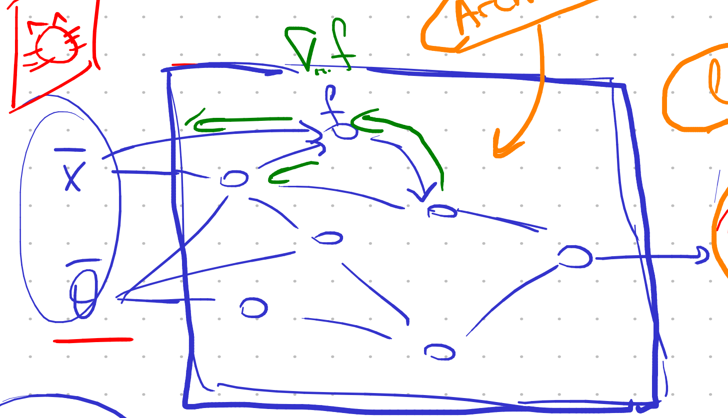
regularizer

D — $\bar\theta$ — $L(\bar\theta, D) = -\log p(\bar\theta|D) \xrightarrow{\bar\theta} \min$

Architecture

loss functions

VAE



$L(\bar\theta, \bar x) \xrightarrow{\bar\theta} \min$

$-\log p(\text{„cat"})$

$\nabla_{\bar\theta} L(\bar\theta, \bar x)$, $\nabla_{\bar x} L(\bar\theta, \bar x)$

SGD

$$L(\bar\theta) = \frac{1}{N}\sum_n L(\bar\theta, \bar x_n) = \underbrace{\mathbb{E}_{\bar x\sim D}[L(\bar\theta, \bar x)]}_{\approx \frac{1}{m}\sum_{i=1}^{m} L(\bar\theta, \bar x_i)} \xrightarrow{\bar\theta} \min$$

$N = 10^6$

4, 8 $m \ll N$

2) Регуляризация / dropout

$$\log p(\bar\theta|D) = \text{Const} + \underbrace{\log p(D|\bar\theta)}_{\text{loss}} + \underbrace{\log p(\bar\theta)}_{\text{regularizer}}$$

$+\alpha \sum_i \theta_i^2$ L_2-reg.

$\Leftrightarrow p(\bar\theta) = \mathcal{N}(\bar\theta|0, cI)$

$\mathbb{E} = p \cdot \mathbb{E}[z_i]$

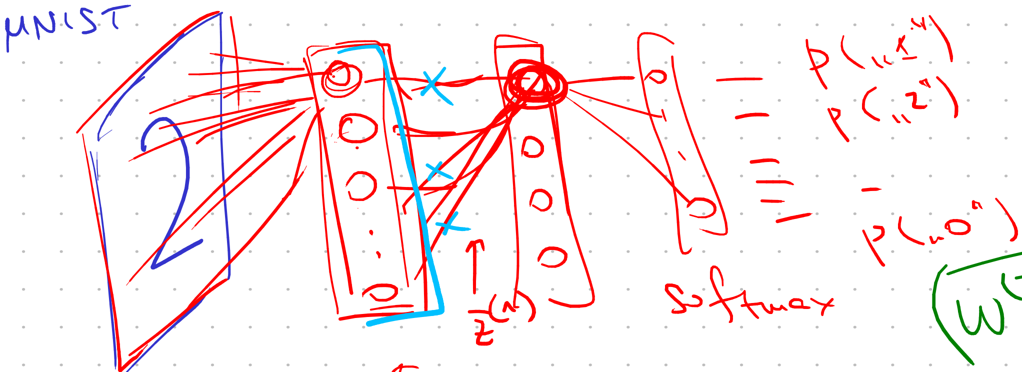
$0,\ 1-p$; $1,\ p$

$h(\bar z^T \bar w)$

z_1, $\cancel{z_2}$, ..., $\cancel{z_k}$

$p \cdot z_1$ — $p \cdot z_k$

MNIST



min L

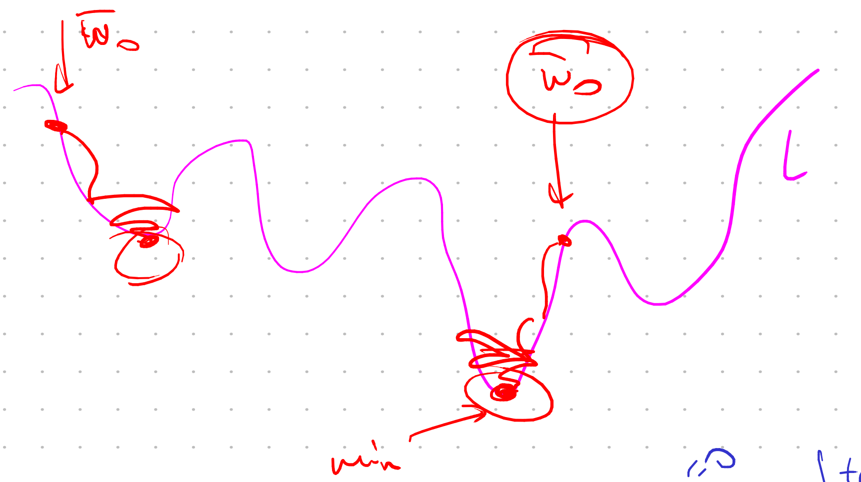
$$\left(W^{(2)} A^{-1}\right)\left(A W^{(1)}\right)$$

$$\bar{x} \rightsquigarrow W^{(1)}\bar{x} \to \underbrace{h\left(W^{(1)}\bar{x}\right)}_{\text{"id"}} \rightsquigarrow \underbrace{W^{(2)} W^{(1)}}_{\min L}\bar{x} \rightsquigarrow \dots$$

ReLU

③ Инициализация весов

unsupervised pretraining

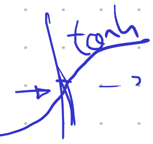


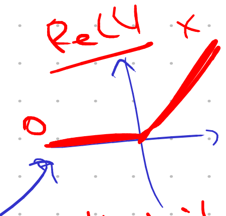
$$\bar{y} = \underbrace{W^{(k)} W^{(k-1)} \dots W^{(1)}}\bar{x}$$

$$\|Wx\| \approx \alpha \|\bar{x}\|$$

≈ 0

$E[y]$





He init

Kaiming He

$$\bar{x} \equiv \Sigma - \underline{\bar{w}^T\bar{x}} = y - \underline{h} - h(y)$$

$$\mathrm{Var}[y] = \underline{\sum_i \mathrm{Var}[y_i]} = \sum_i \underline{\mathrm{Var}[w_i x_i]}$$

$$\mathrm{Var}[w_i x_i] = \dots = \underbrace{\cancel{E[x_i]^2}}_{=0} \mathrm{Var}[w_i] + \cancel{E[w_i]^2} \mathrm{Var}[x_i] + \mathrm{Var}[w_i]\,\mathrm{Var}[x_i]$$

$$\mathrm{Var}[y_i] = \mathrm{Var}[w_i]\,\mathrm{Var}[x_i]$$

$$\underbrace{\mathrm{Var}[y]} = \underbrace{n\,\mathrm{Var}[w_i]}\cdot\underbrace{\mathrm{Var}[x_i]}$$

$$\boxed{\mathrm{Var}[w_i] = O(1/n)}$$

$$W_i \sim \mathrm{Unif}\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right]\right)$$

2010

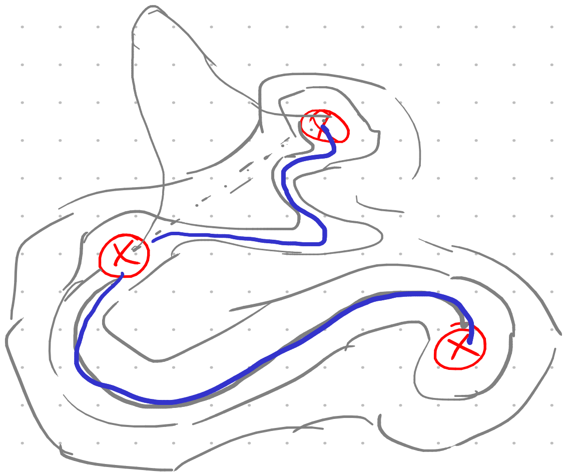
Xavier Glorot

Bengio

$$\mathrm{Var}(\mathrm{Unif}([a,b])) = \frac{(b-a)^2}{12}$$

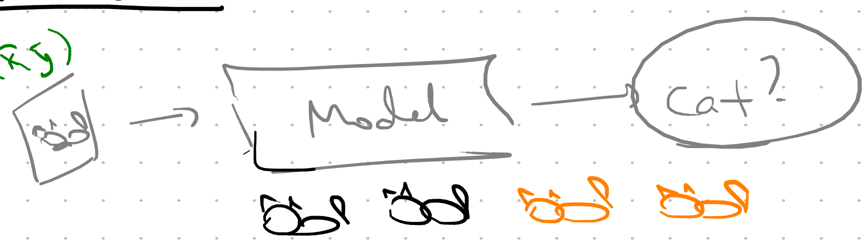
$$\mathrm{Var} \sim \frac{1}{3n}$$

Xavier init

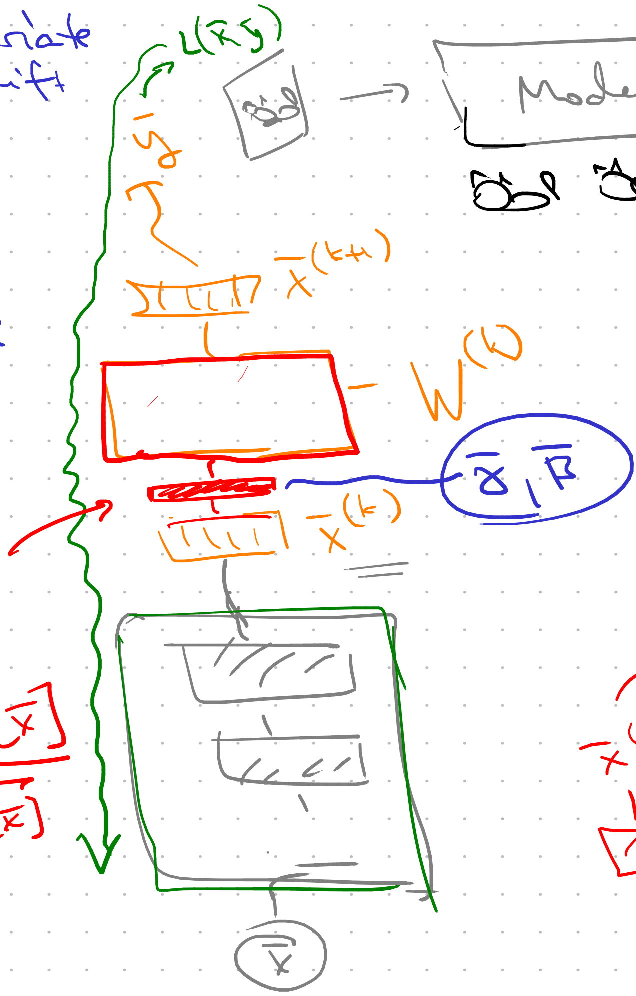


4 Batch normalization

covariate shift

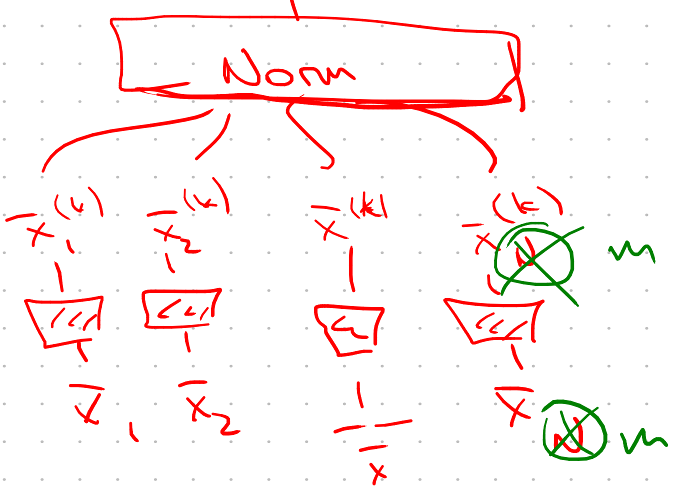


internal covariate shift



$$\bar x \to \frac{\bar x - E[\bar x]}{\sqrt{\mathrm{Var}[\bar x]}}$$

$$\hat{x}^{(k)} = \frac{\bar x^{(k)} - \frac{1}{m}\sum \bar x_i^{(k)}}{\sqrt{\mathrm{Var}[\bar x_n^{(k)}]}}$$



$$BN(\bar x) = \frac{\bar x - E_m[\bar x]}{\sqrt{\mathrm{Var}_m[\bar x]}} \odot \underbrace{\bar\gamma}_{\text{scale}} + \underbrace{\bar\beta}_{\text{shift}}$$