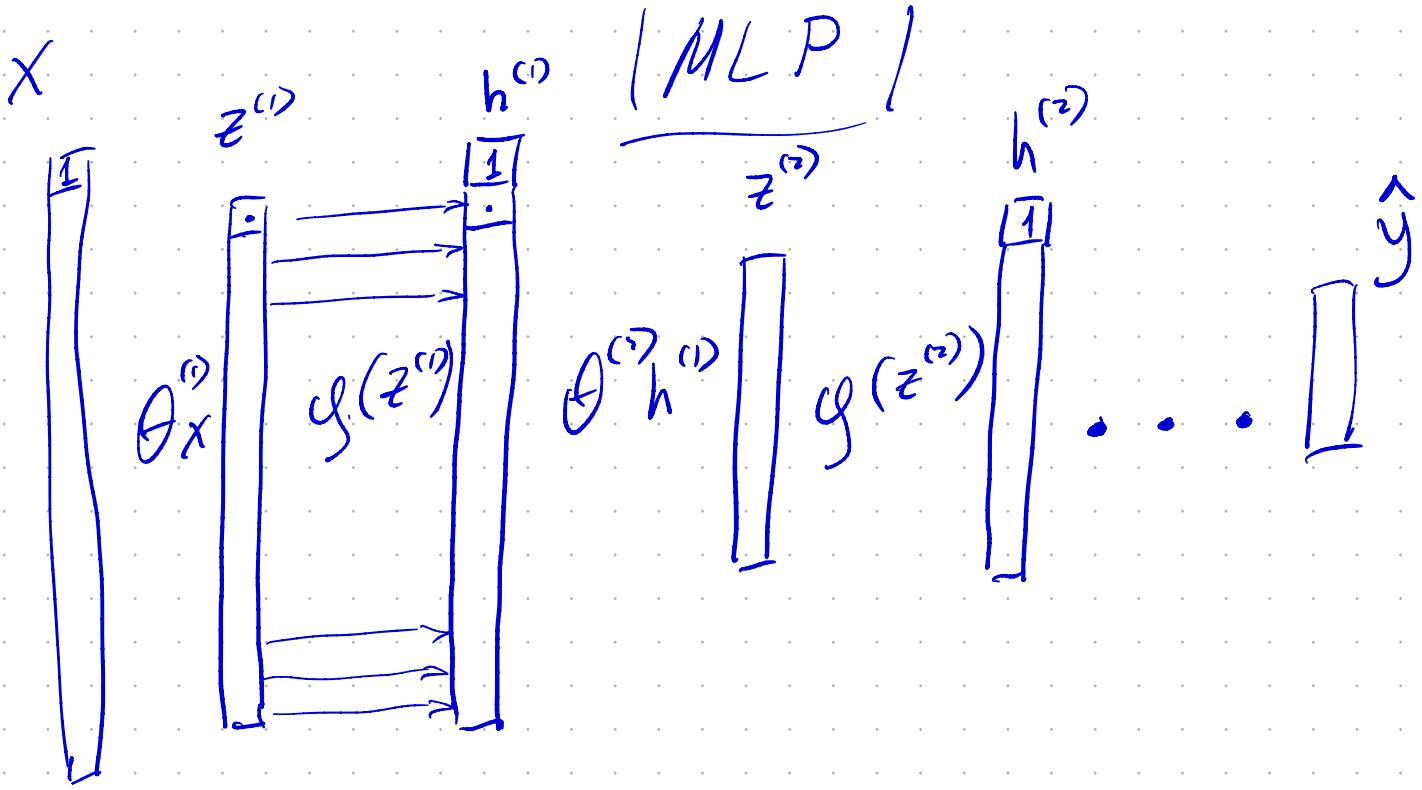
MLP



$$\hat{y} = \psi\left(\dots \varphi\left(\theta^{(2)} \varphi^{(1)}\left(\theta^{(1)} x\right)\right)\right)$$

$h^{(i)}$ – вектор скрытых представлений
hidden representations
embeddings

карты активации

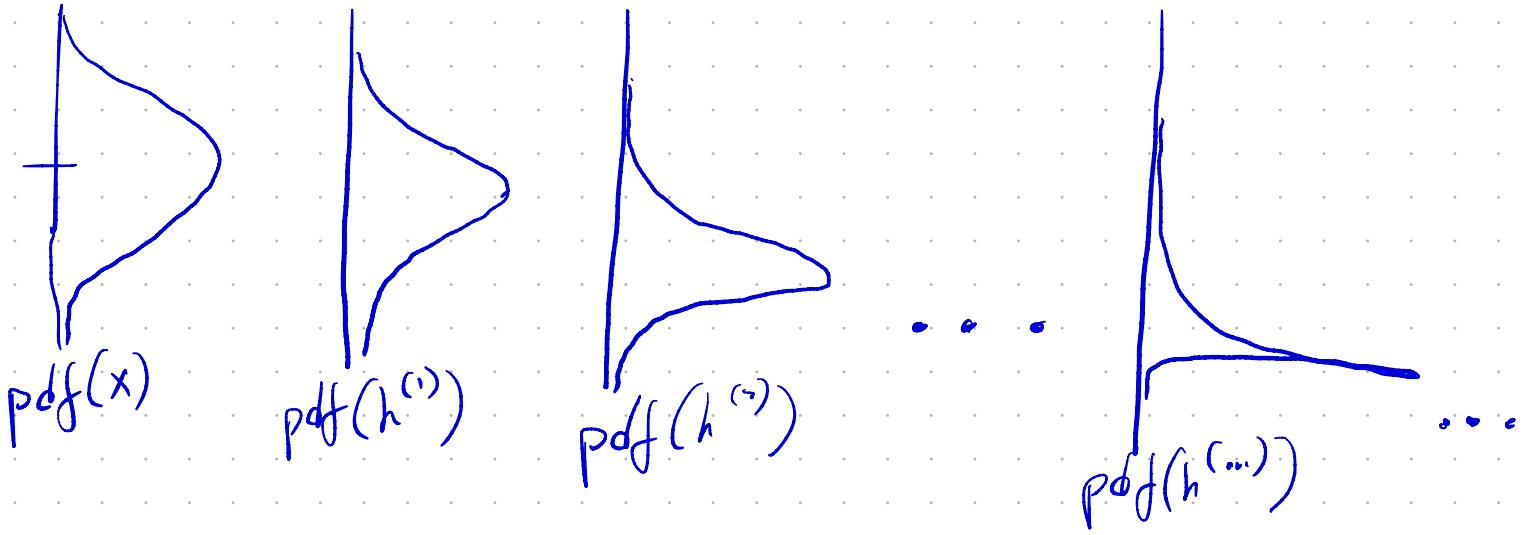
$\hat{y} \in \mathbb{R}^1 \quad \mathbb{R}^d$

ψ: $I(z^{(L)})$

$\sigma(z^{(L)})$

$\mathrm{Softmax}(z^{(L)})$

$$Var(h^{(l)}) \sim k \, Var(\theta^{(l+1)}) Var(h^{(l+1)})$$



$k \, Var(\theta^{(...)}) < 1$

$Var(\hat{y}) \simeq 0$

$\hat{y} = const$

$k \, Var(\theta^{(...)}) > 1$

$Var(\hat{y}) \to \infty$

$\left| \frac{\partial \hat{y}}{\partial x} \right|_{ij} \to \infty$

$k \cdot Var(\theta^{(...)}) \simeq 1$

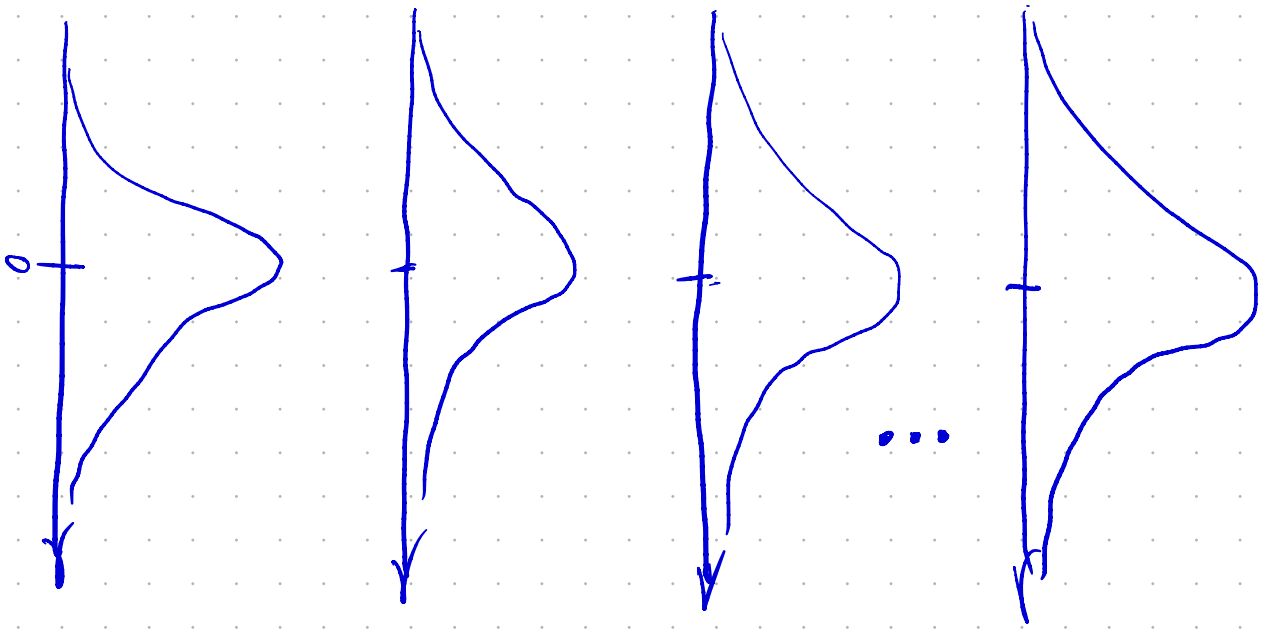
$Var(\theta^{(...)}) \simeq \frac{1}{k}$

Вычисление, применение НС, inference

$$\hat{y} = F_{MLP}(X)$$

Обучение, оптимизация (ф-ии потерь), training, learning

$$\theta^* = \arg\min \mathcal{L}(F_{MLP}(X), Y)$$

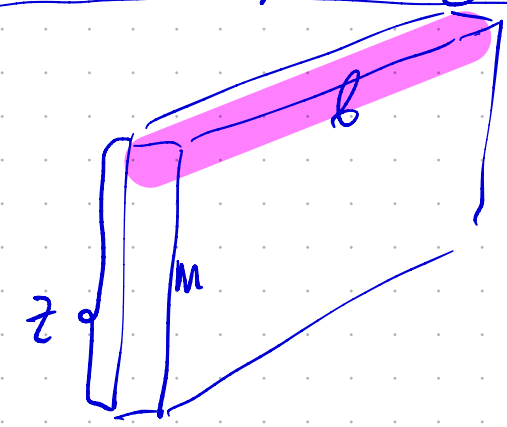


Batch Normalization

Пакетная нормализация

$$z^{(l)} = \theta h^{(l-1)}$$

$$z^* = \frac{z - \mu_B(z)}{\sigma_B(z)}$$



$z \in \mathbb{R}^m$

$\mu \in \mathbb{R}^m$

$\sigma_z \in \mathbb{R}^m$

$$\mu(z^*) = 0$$

$$\sigma(z^*) = 1$$

$$\mu_t^* = \mu_{t-1}^* \beta + \mathbb{E}_{B_t}(z)$$

$$\sigma_t^* = \sigma_{t-1}^* \alpha + \sigma_{B_t}(z)$$

$$z^* = \frac{z - \mu_t^*}{\sigma_t^* + \varepsilon} \cdot \gamma + \xi \qquad \begin{matrix} \gamma = 1 \\ \xi = 0 \end{matrix}$$

При обучении оптимизируются γ, ξ

"накапливаются" μ^*, σ^*

При вычислении: $\mu^*, \sigma^*, \gamma, \xi$ фиксированы и не меняются