$X \rightarrow h^{(\ell=1)} \rightarrow h^{(\ell=2)} \cdots h^{(\ell)} \rightarrow$
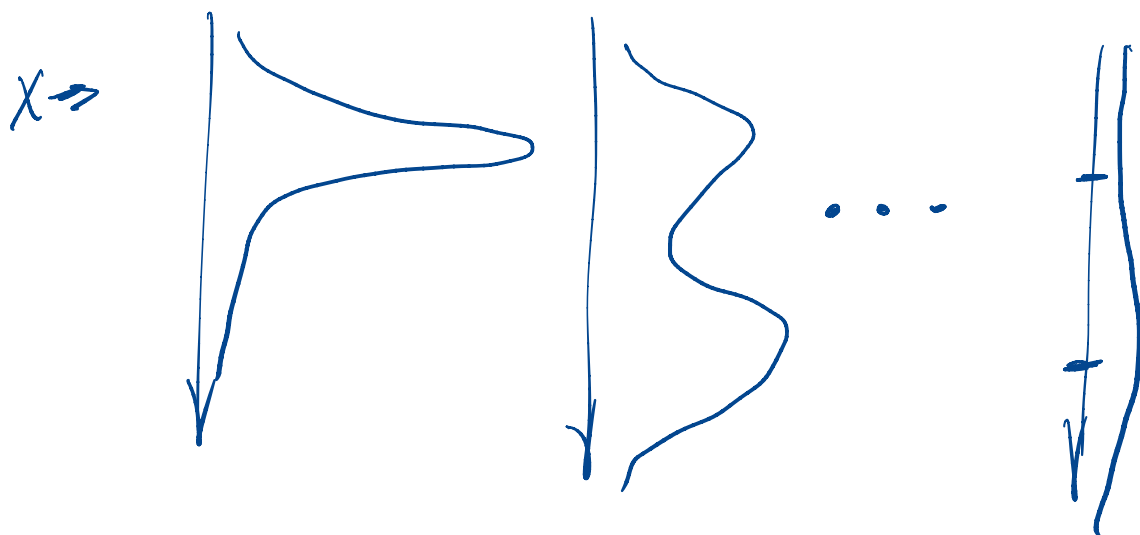
$$\theta_{t+1} = \theta_t - \eta \, g(B, \theta)$$

SGD:
$$\theta_{t+1} = \theta_t - \eta \, \nabla_\theta \mathcal{L}(B, \theta)$$

$X \rightarrow$
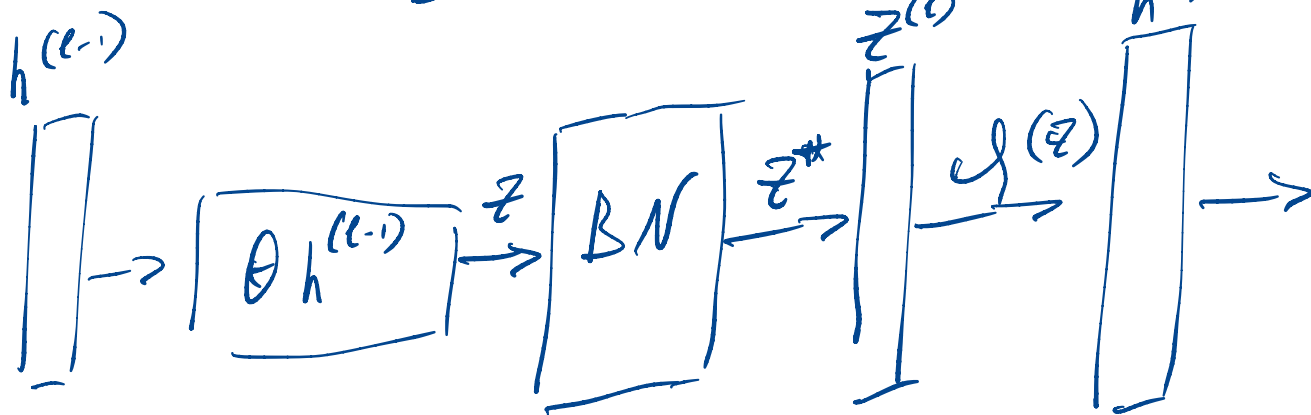
Batch normalization
Пакетная нормализация

$$Z^{(\ell)} = \Theta \, h^{(\ell-1)}$$

$$Z^{*(\ell)} = \frac{Z - \mu_B(Z)}{\sigma_B(Z)} \Rightarrow \begin{array}{l} \mu(Z^*) = 0 \\[2mm] \sigma(Z^*) = 1 \end{array}$$

$$\mu_t^* = \mu_{t-1}^* \beta + \mathop{E}_B(Z) \qquad\qquad \mu_t^* = \mathop{E}_B(Z)$$

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

$$Z^* = \frac{Z - \mu_t^*}{\sigma_t^* + \varepsilon} \cdot \gamma + \xi \qquad\qquad \begin{array}{l} \gamma = 1 \\ \xi = 0 \end{array}$$
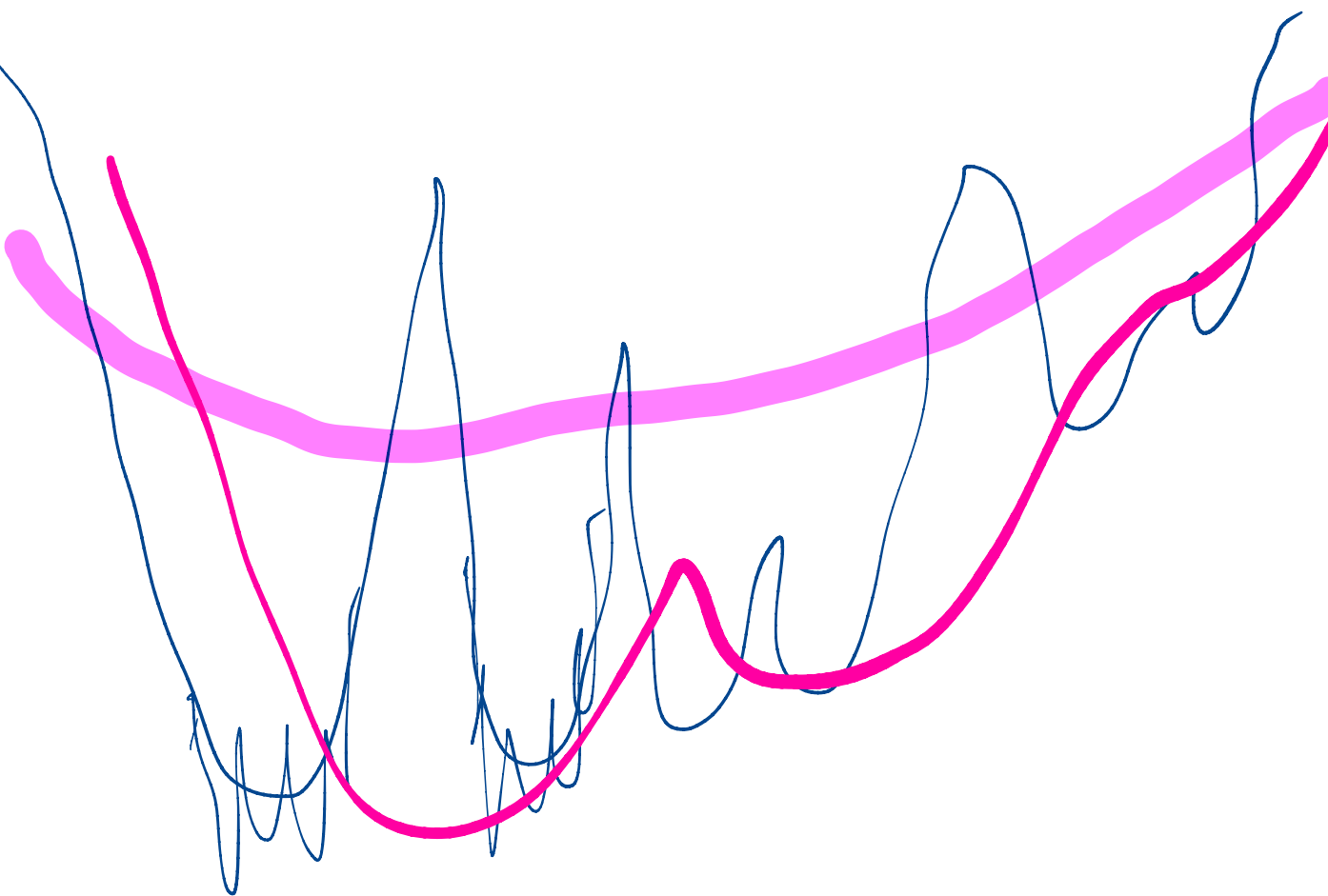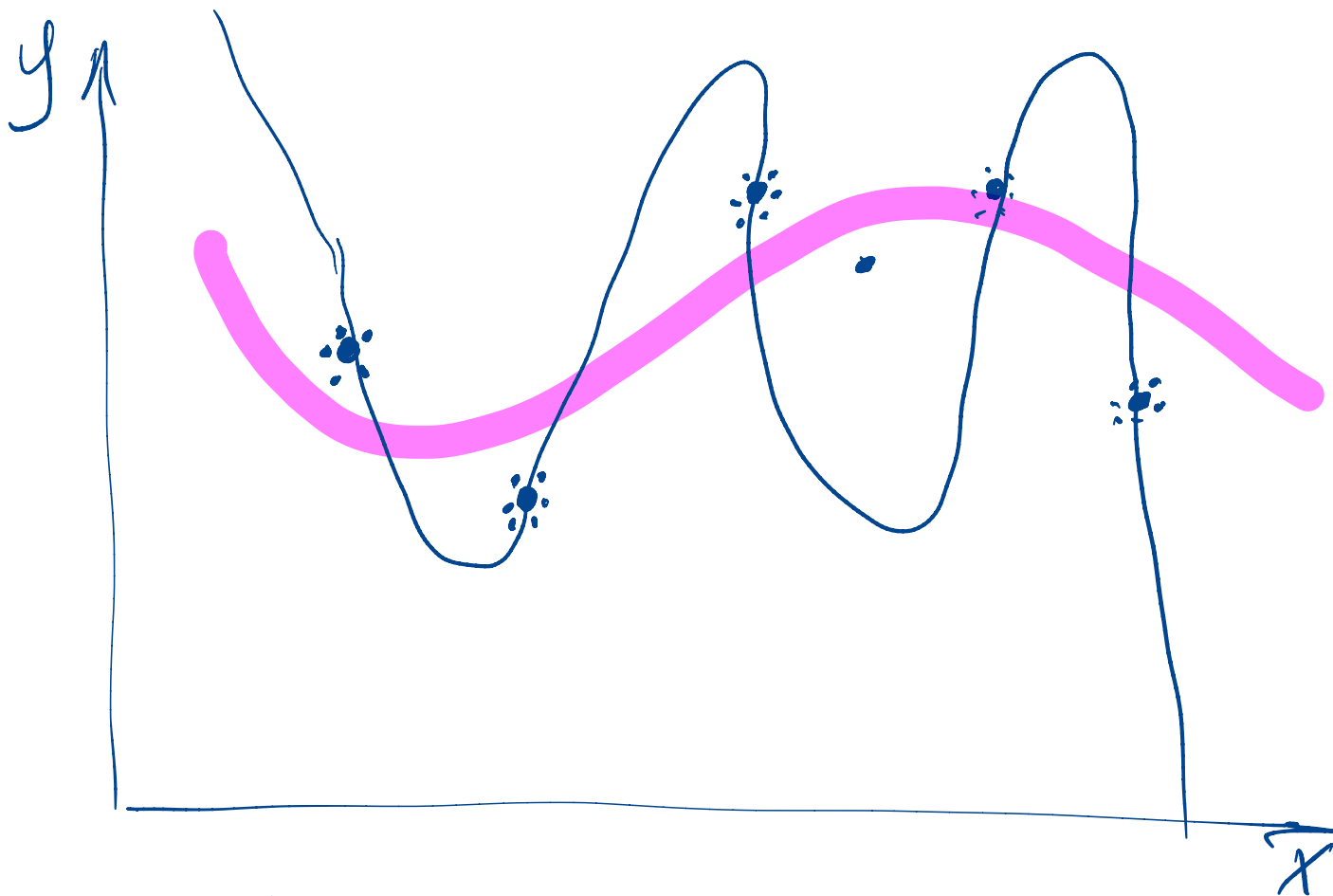
При обучении ① оптимизируются γ, ξ

② „накапливаются" μ^*, σ^*

При вычислении: $M^*, \sigma^*, \gamma', \xi$ — не меняются

$\mathcal{L}(\theta, \beta)$

Искусственное дополнение данных

Data augmentation

i.i.d.