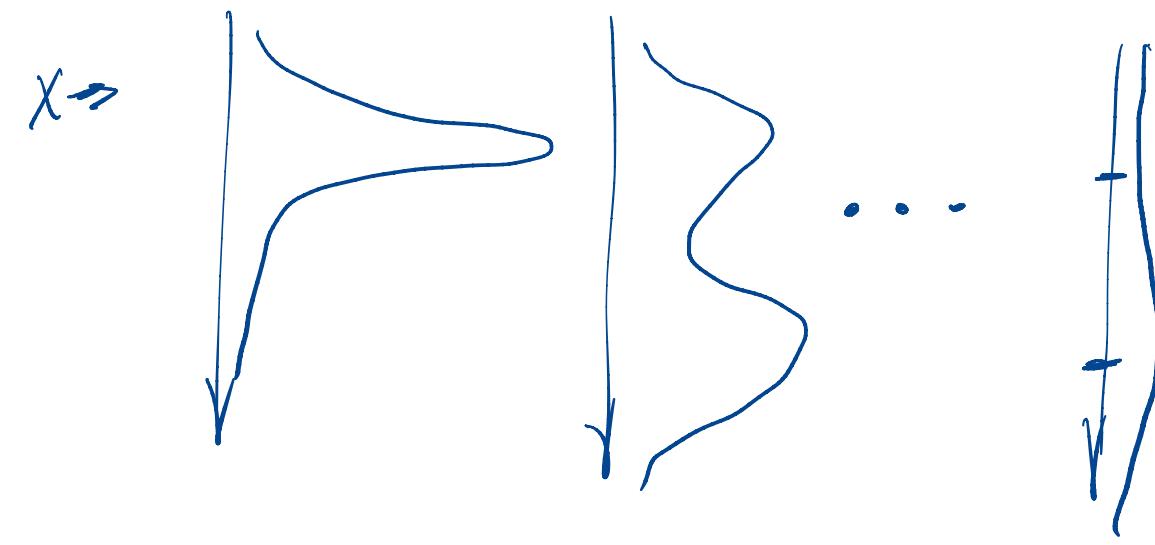


$$\theta_{t+1} = \theta_t - \gamma g(B, \theta)$$

SGD:

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



Batch normalization

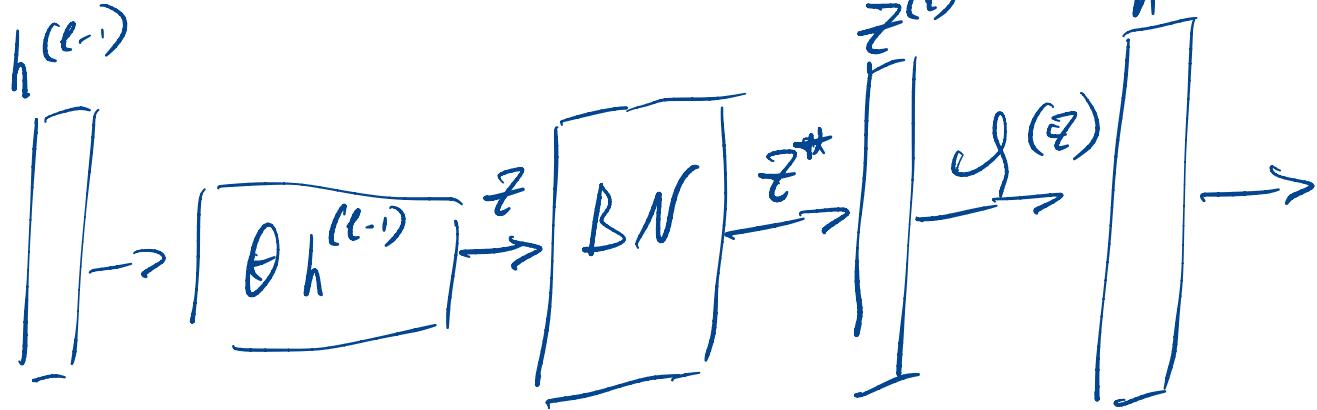
Нормализация

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

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

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

$$G(z^*) = 1$$



$$\mu_t^* = \mu_{t-1}^* \beta + E_B(z)$$

$$\mu_t^* = E_B(z)$$

$$G_t^* = G_{t-1}^* \lambda + G_B(z)$$

$$z^* = \frac{z - \mu_t^*}{G_t^* + \epsilon} \cdot \gamma + \xi$$

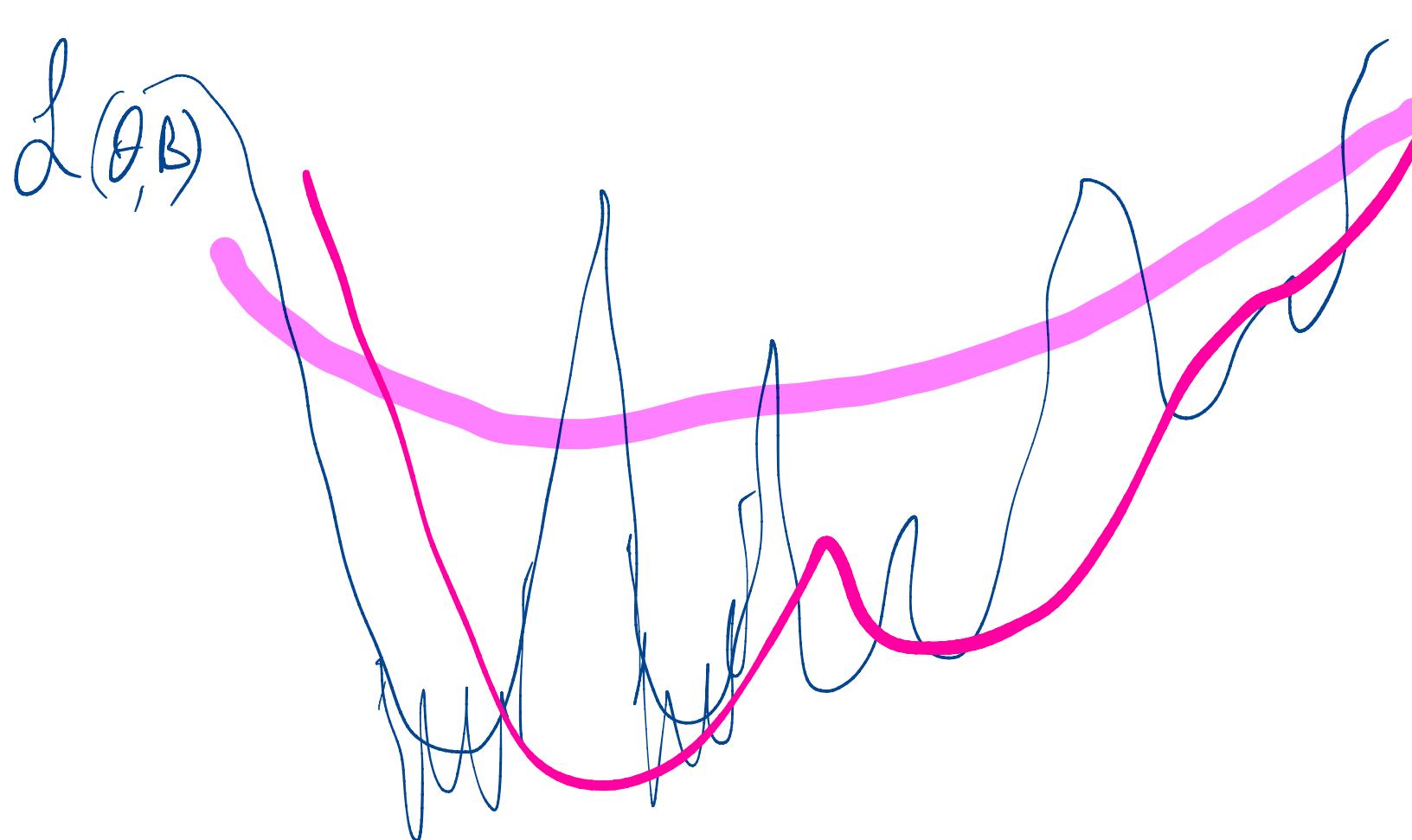
$\gamma \neq 0$
 $\xi = 0$

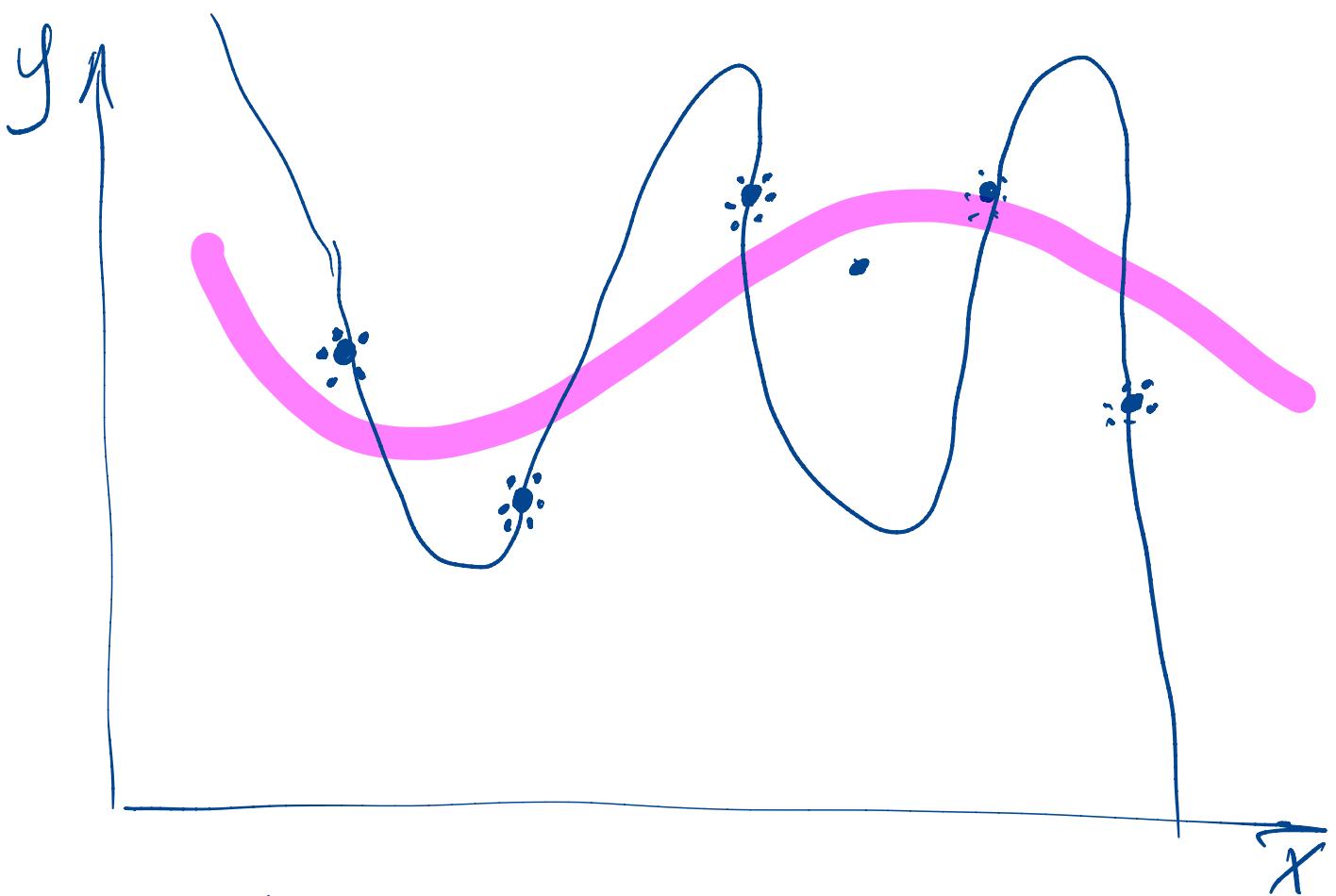
Также обозначим

① оптимизируемые γ, ξ

② "накалывающиеся" μ^*, G^*

The conditions: $\mu^*, \sigma^*, \gamma, \beta$ - are independent





Искусственное дополнение данных
Data augmentation

i.i.d.

z

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

$$\sigma_t^{*2} = (1-\beta) \sigma_t^2(z) + \beta \sigma_{t-1}^{*2}$$

$$z^* = \frac{z - \mu^*}{\sigma^* + \epsilon}$$

$d \times B$

d

d

$d \times B$

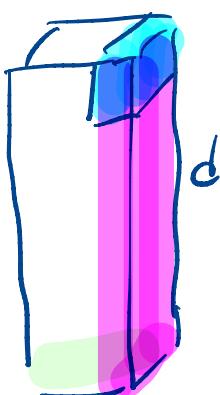
Boro:

$$z \in \mathbb{R}^{d \times B}$$

$$M = d \times B$$

Gano

$$M = 2 \times d + 2 \times d \times B$$



$$E_z = \frac{1}{B \cdot d} \sum_{i=1}^B \sum_{j=1}^d z_{ij} \quad z_{ij} \in \mathbb{R}$$

$$(E_B z)_j = \frac{1}{B} \sum_{i=1}^B z_{ij} \quad \in \mathbb{R}^d$$

$$(E_d z)_i = \frac{1}{d} \sum_{j=1}^d z_{ij} \quad \in \mathbb{R}^B$$

Batch Norm

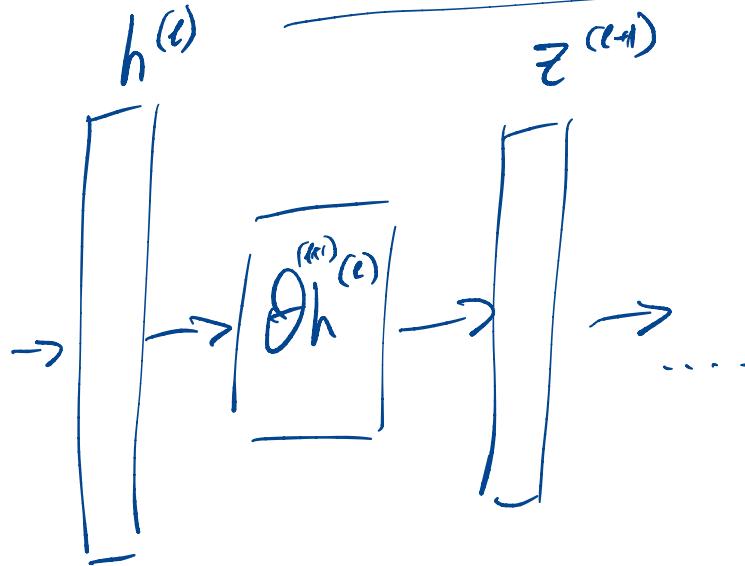
Pros

- ① Может ускорить обучение за счёт стандартизации
- ② Может не сильно отразить следующую инициализацию
- ③ Дает эффект отмены нормализации независимо
- ④ Работает как линейный преобраз.

Cons

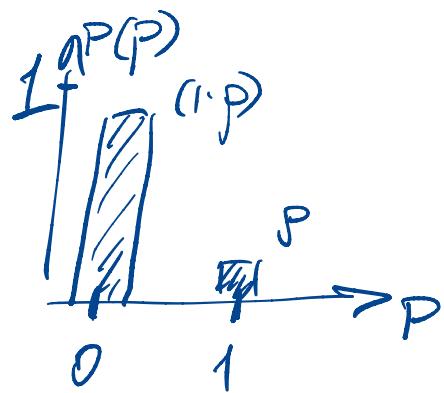
- ① Сложно применять при количествах $B \sim 1-5$
 - ② BN сильно повышает обучение модели
 - ③ BN нельзя применять в случаях, когда данные измерены поставляемые из разных расп-й.
-

Dropout (Удаление)



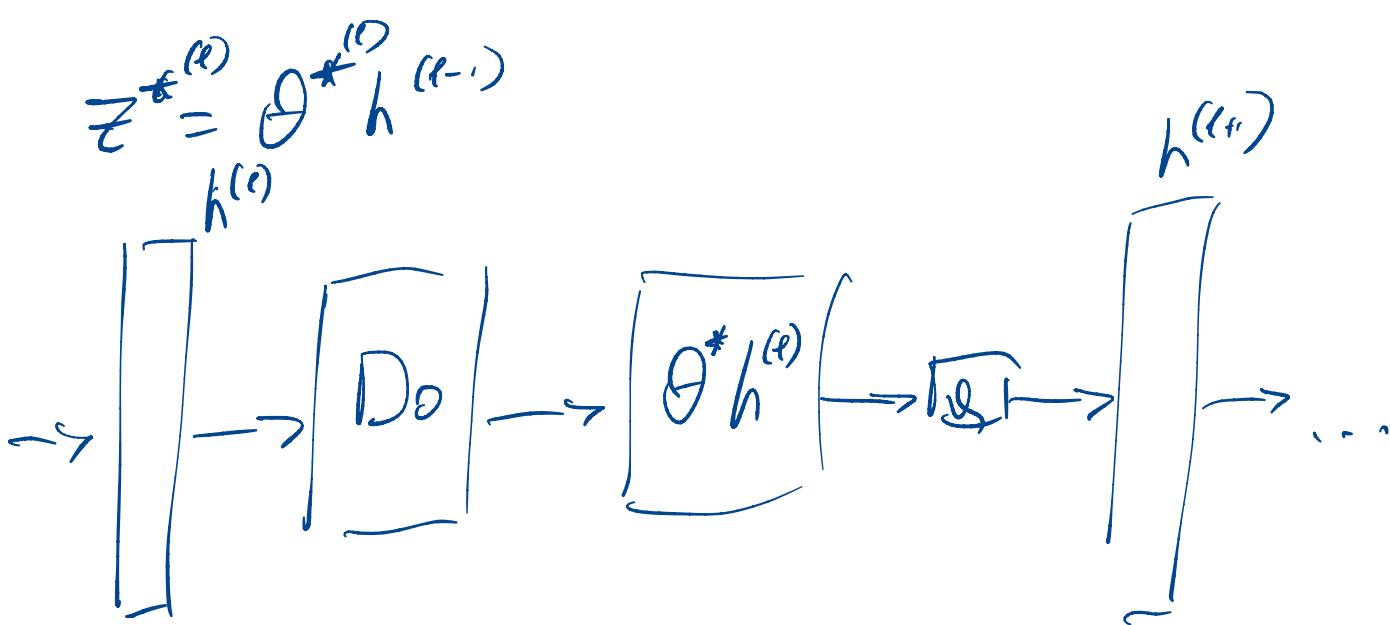
$$P.\text{shape} = \theta.\text{shape}$$

$$P \sim \text{Ber}(p)$$



$$\theta^* = \theta \odot P$$

$$\tilde{z}^* = \theta^* h^{(1)}$$



$$\theta^* = \theta \odot p \quad P \sim \text{Ber}(p)$$

$$z^* = \theta^* \cdot h$$

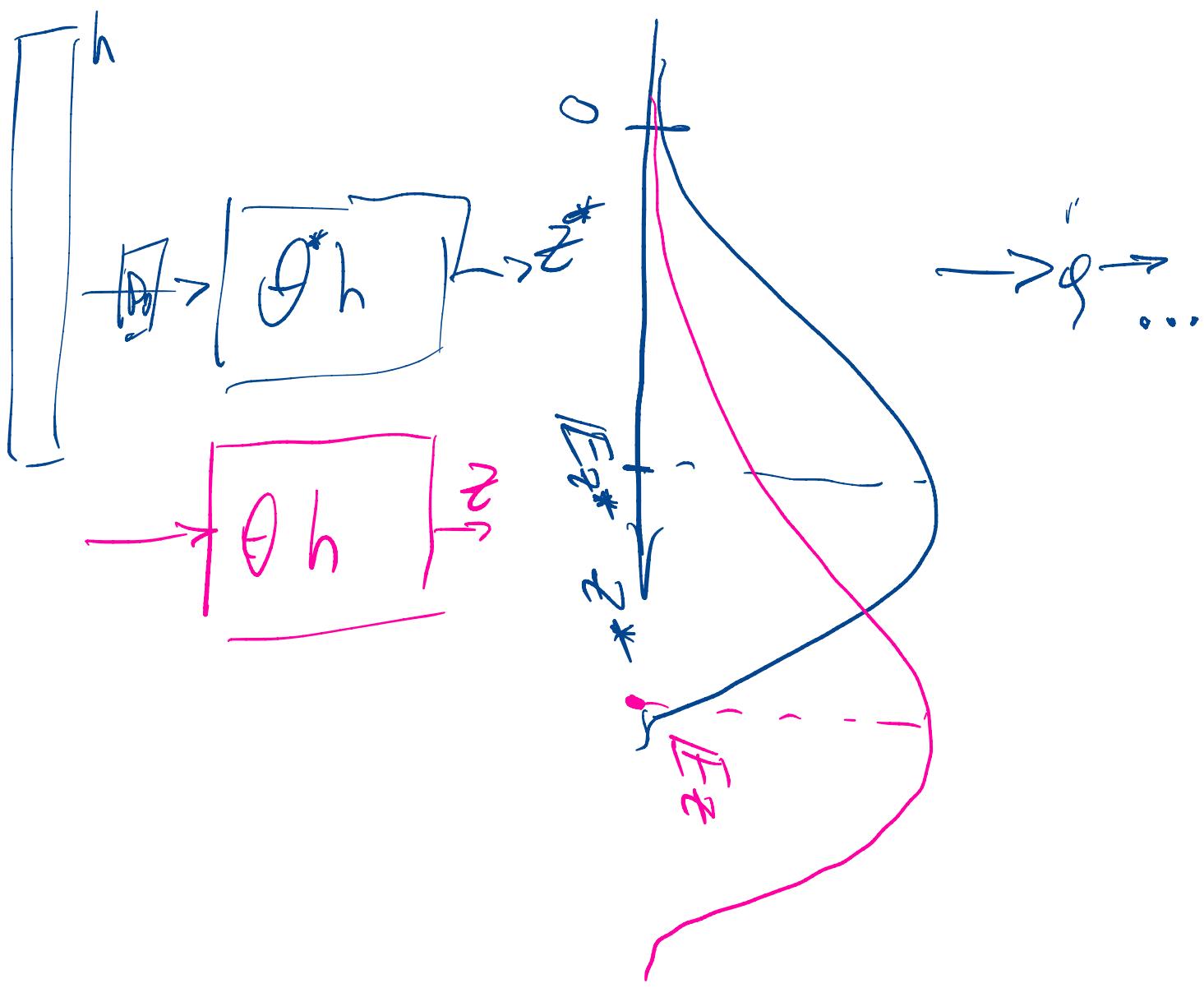
$$E z = \mu$$

$$\begin{aligned} E z^* &= E(\theta^* h) = E(\theta^*) E(h) = \\ &= E(\theta \odot p) E(h) = \cancel{E\theta} \cancel{E p} \cancel{E h} = \\ &= EP \cdot E(\theta h) = EP \mu \end{aligned}$$

$$EP = \frac{1}{N} \sum_{i=1}^N p = \frac{1}{N} \left(\cancel{\theta \cdot (1-p) \cdot N} + 1 \cdot p \cdot N \right) =$$

$$\begin{array}{ccc} \uparrow & (1-p) & = \frac{1}{N} p^N = p \\ \text{Bar chart} & p & \end{array}$$

$$E z^* = p \cdot \mu \quad E z = \mu$$



$$z = \theta h \quad z^* = \theta^* h$$

$$Ez' = \mu \quad \textcircled{\theta h} = Ez^* \\ p^\mu \qquad \qquad \qquad p^\mu$$

$$z = \theta h \quad z^* = \int \theta^* h \\ Ez' = \mu \quad = Ez^* = \int p^\mu$$

Dropout

Объяснение

① $\bar{z}^* = \boxed{P} \Theta h$
 $P \sim \text{Ber}(p)$

Вычисление

$$\bar{z}' = \boxed{P} \Theta h$$

② $\bar{z}^* = \boxed{\frac{1}{p} P} \Theta h$
 $P \sim \text{Ber}(p)$

$$\bar{z}' = \Theta h$$

Propont

Pros

Повышает обобщающую способность

Уменьшает эффект отвлечения рабочего места сотрудника

Может расширяться как вправо, так влево

Cons

① Снижает способность обучение

② Повышает потребление памяти