Lecture 2

Minimize error rate

$$\frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i] \to \min_a$$

$y_i \in \{1, \ldots, c\}$ – classes

$$p = \begin{pmatrix} p(\hat{y}_i = 1) \\ p(\hat{y}_i = 2) \\ \vdots \\ p(\hat{y}_i = c) \end{pmatrix}$$

$$z_i = \sigma(W_i z_{i-1} + b_i)$$

$$z_n = W_n z_{n-1} + b_n$$

$$z_n \in \mathbb{R}^c$$

$$p = \{p_i\}_{i=1}^{c} \qquad p_i \geq 0, \quad \sum_{i=1}^{c} p_i = 1$$

Softmax

$$z = (z_1, \ldots, z_c)$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{c} \exp(z_j)} \geq 0$$

p – probabilities $\quad \forall i,j \quad z_i > z_j \Rightarrow p_i > p_j$

z – logits

$$\hat{y} \to 1, \dots n \qquad \hat{y} \to \begin{pmatrix} p_1 \\ \vdots \\ p_y \\ \vdots \\ p_c \end{pmatrix} \to 1$$

$$y \to 1, \dots n$$

$$y \to \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} y$$

y - target $\in \{1, \dots, c\}$

$$-\sum_{k=1}^{c} [k = y] \log p_k \to \min_{\{W, b\}}$$

$$\|$$

$$-\log p_y \to \min \sim p_y \to \max$$

NLL Negative log-likelihood

$$\{(x_i, y_i)\}_{i=1}^{\ell}$$

$$-\sum_{i=1}^{\ell} \sum_{k=1}^{c} [k = y_i] \log p_k^{(i)} \to \min_{\{W, b\}}$$

$$-\sum_{i=1}^{\ell} \log p_{y_i}^{(i)} \to \min \sim \prod_{i=1}^{\ell} p_{y_i}^{(i)} \to \max$$

$$\text{NLLLoss}(\text{Softmax}(z), y) \to \min$$

$$\log p_i = \log \text{softmax}(z_i) = \log \frac{\exp(z_i)}{\sum_{k=1}^{c} \exp z_k} =$$

$$= \underset{\substack{\| \\ z_i}}{\log \exp(z_i)} - \underbrace{\log\left(\sum_{k=1}^{c} \exp z_k\right)}_{\text{logsumexp}(z)}$$

$$\log \frac{\exp(z_i - \max_j z_j)}{\sum_{k=1}^{c} \exp(z_k - \max_j z_j)}$$

$$\text{Cross entropy loss}(\underset{\substack{\uparrow \\ \text{logits} \in \mathbb{R}^c}}{z}, y)$$

$$CE\left(\{p_i\}_{i=1}^{c}, \{q_i\}_{i=1}^{c}\right) = -\sum_{i=1}^{c} q_i \log p_i$$

$$\hat{y} = \begin{pmatrix} p_1 \\ \vdots \\ p_c \end{pmatrix} = p \qquad y = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}_y = q$$

$$NLL = CE = -\sum_{i=1}^{c} q_i \log p_i \to \min$$

Dropout

$z_k \in \mathbb{R}^{d_k}$ $\qquad d_k$ — hidden size

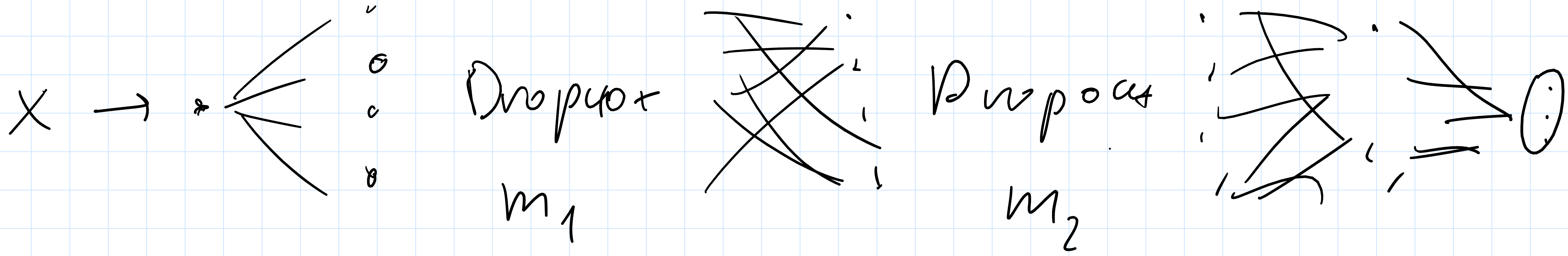
$z_k = \sigma(W_k z_{k-1} + b_k)$

$m \in \{0,1\}^{d_k}$ $\qquad m_i \sim \mathrm{Bernoulli}(1-p)$

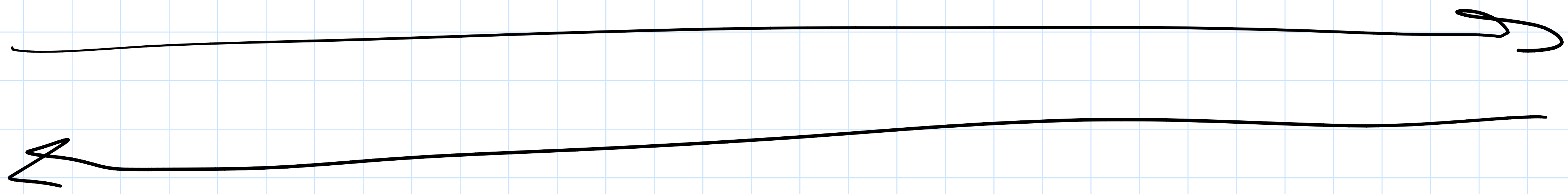
$p \in [0{,}1 ; 0{,}25]$

$y = m \odot z_k$

element wise product



$m_1 \neq m_2$



X_{next} $\qquad m_{1_{next}}$ $\qquad m_{2_{next}}$

$\mathbb{E}[y_k] = \mathbb{E}[m \odot z_k] = \mathbb{E}[m] \odot z_k =$

$y_k \in \mathbb{R}^{d_k}$ $\qquad = (1-p) \cdot z_k$

$\langle w_{k+1}, z_k \rangle \leftarrow 1-p$

Train mode:

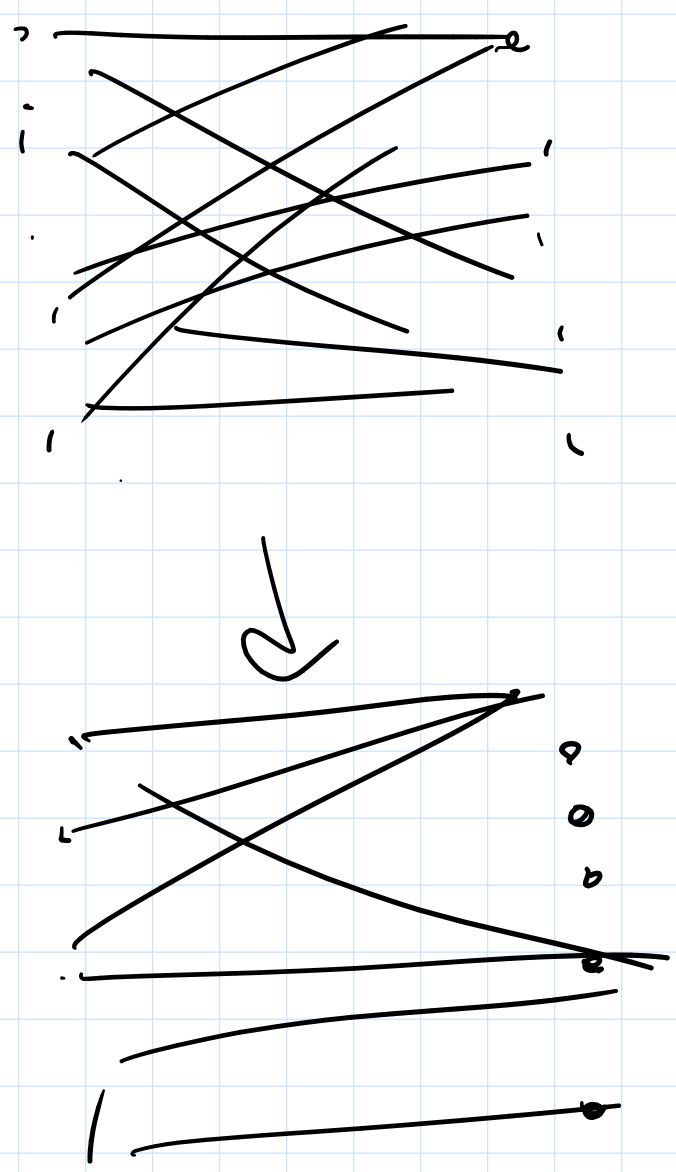
$$m_i \sim \text{Bernoulli}(1-p)$$

$$y_k = m \odot z_k \frac{1}{1-p} \qquad \mathbb{E}[y_k] = z_k$$

Eval mode:

$$y = z$$

$$m = \begin{pmatrix} 1 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \leftarrow m_1 \\ \\ \\ \leftarrow m_i \\ \\ \\ \leftarrow m_{d_k} \end{matrix}$$



$$m = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Batch Normalization

$(-\infty, +\infty)$

Train mode

$z \in \mathbb{R}^{B \times d_k}$ B – batch size

$\{(x_i, y_i)\}$

$z_i \in \mathbb{R}^{d_k}$

$$W z + b$$
$$\uparrow$$
$$\mathbb{R}^d$$

$x_1 \quad y_1 \rightarrow$

$x_2 \quad y_2 \rightarrow$

$$\left(x_1, x_2 \dots x_B \right)$$

$$\Downarrow$$

$$W \overset{\overset{B}{\longleftrightarrow}}{Z} \updownarrow^{d} + \left(b, \dots, b \right)$$

$$\mu = \frac{1}{B} \sum_{i=1}^{B} z_i \qquad \sigma^2 = \frac{1}{B} \sum_{i=1}^{B} (z_i - \mu)^2$$

$$\mu, \sigma \in \mathbb{R}^{d_k}$$

$$\hat{Z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \qquad \boxed{\text{element-wise}}$$

$\leftarrow$ small

$$\frac{1}{B} \sum \hat{z}_i = 0$$

$$\frac{1}{B} \sum \hat{z}_i^2 \approx 1$$

$$y_i = \hat{z}_i \odot w + b \qquad w, b \in \mathbb{R}^{d_k}$$

$$\underset{\alpha}{w} \qquad \underset{\beta}{b}$$

learnable

$$\text{running\_mean} := \text{running\_mean}_{prev} (1 - m) + \mu \cdot m$$

$$\text{running\_var} := \text{running\_var}_{prev} (1 - m) + \sigma^2 m \frac{B}{B-1}$$

$m \approx 0{,}1$

Eval mode:

$$\hat{z}_i = \frac{z_i - \text{running\_mean}}{\sqrt{\text{running\_var} + \varepsilon}}$$

$\varepsilon = \text{const}$

$$y_i = \hat{z}_i \odot w + b$$