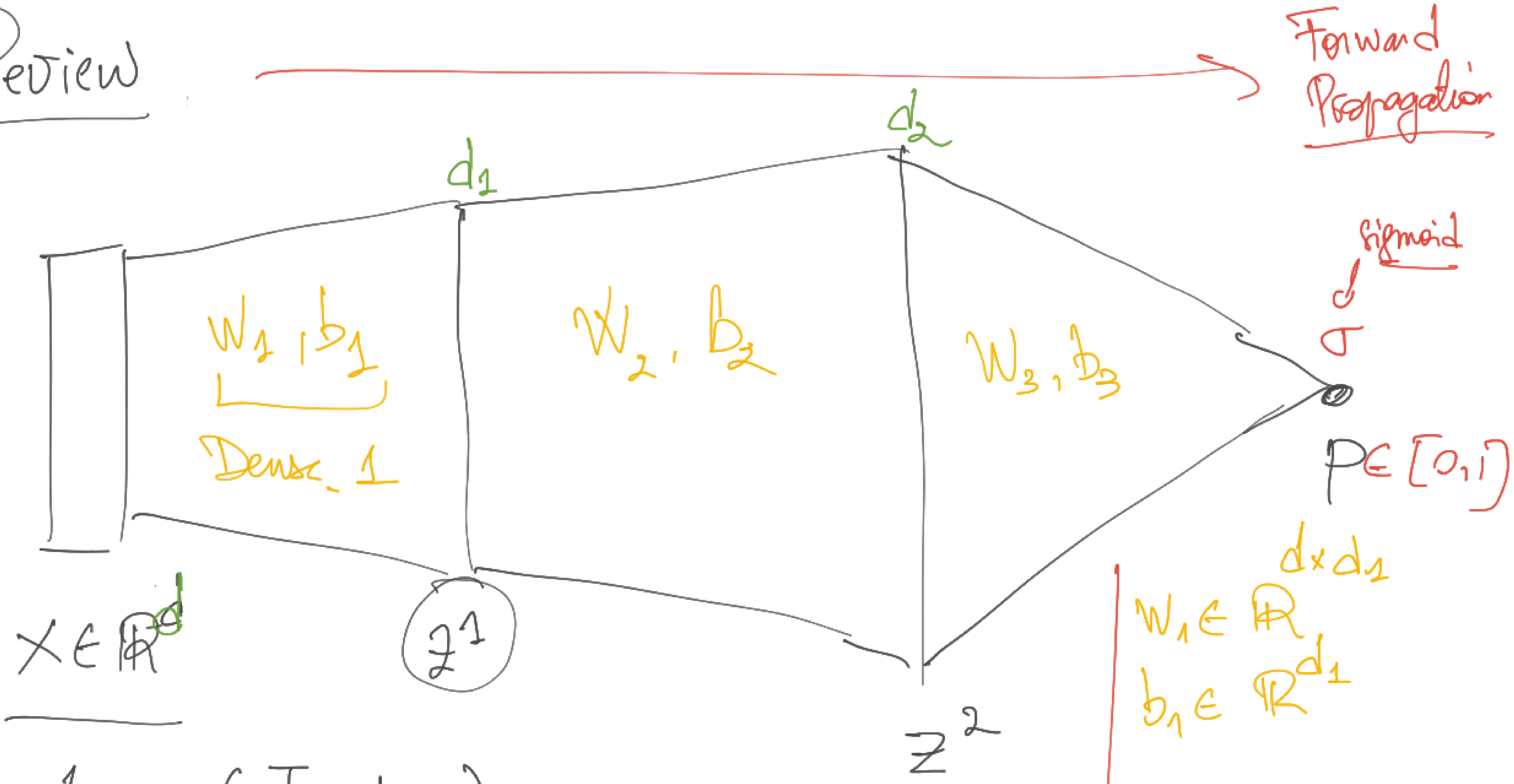
Review



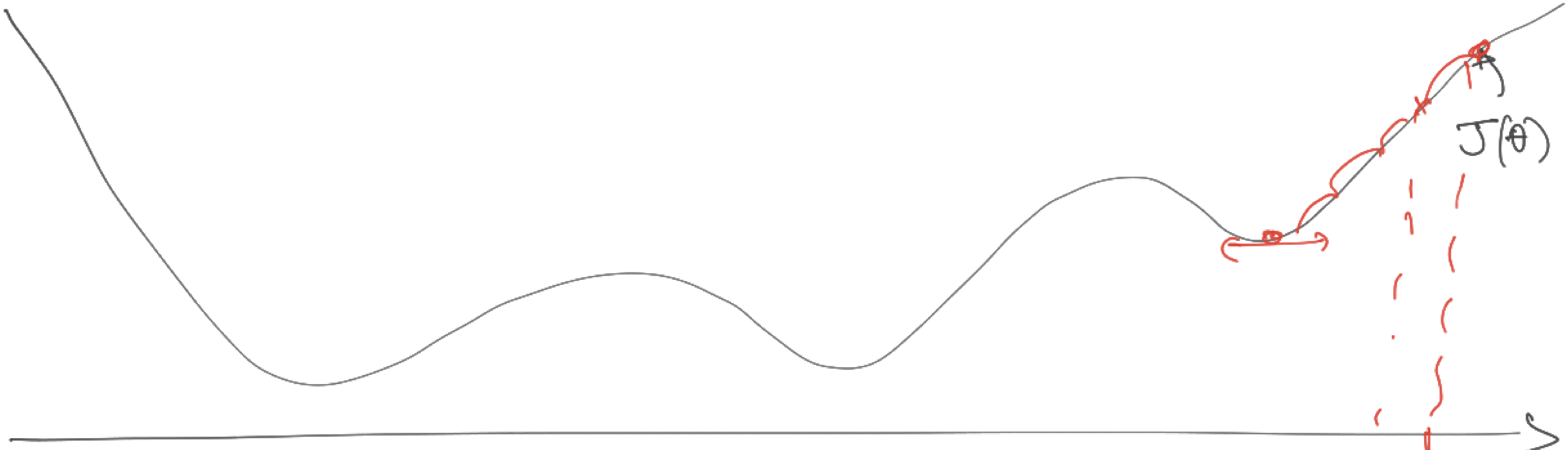
$W_1 \in \mathbb{R}^{d \times d_1}$

$b_1 \in \mathbb{R}^{d_1}$

$$z^1 = \sigma(W_1^T x + b_1)$$

$$z^2 = \sigma(W_2^T z^1 + b_2) \quad ; \quad p = \sigma(W_3^T z^2 + b_3)$$

Parameters : $\theta = \{ W_i, b_i \; ; \; 1 \le i \le 3 \} = \begin{bmatrix} \\ \\ \end{bmatrix}$ all parameters



$$\theta_{k+1} \leftarrow \theta_k - \eta \nabla_\theta J$$

$\eta \sim 10^{-3}$

$$\begin{bmatrix} \frac{\partial J}{\partial \theta^1} \\ \frac{\partial J}{\partial \theta^2} \\ \vdots \end{bmatrix}$$

parameters

Loss function: $J(\theta) = -\frac{\log(\mathcal{L}(\theta))}{N}$

e.g. Binary Classification pb:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left\{\underbrace{y_i}_{}\log \underbrace{p_i}_{} + (1-y_i)\log(1-p_i)\right\}$$

Dataset : $\{\underbrace{x_i}_{d}, \underline{y_i}\}_{1\le i\le N}$

$$x_i \xrightarrow{f_\theta} p_i \qquad p_i = \overset{NN}{\boxed{f_\theta}}(x_i)$$

e) Multiclass classification pb : (with K categories)

Dataset : $\{x_i, y_i\}_{1 \le i \le N}$; $x_i \in \mathbb{R}^d$; $y_i \in \{1, \dots, K\}$

$\hookrightarrow \hat{y}_i = [0 \dots 1 0 \dots 0]_K$

$$x_i \xrightarrow{f_\theta} \underline{\underline{p_i}} = \begin{bmatrix} p_i^1 \\ \vdots \\ p_i^K \end{bmatrix}$$

$$J(\theta) = {}_{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{y}_i^k \log(p_i^k)$$

Dataset : Approach in Programming Session 4.

doc_1 : "Neural Networks are fun" $\Rightarrow [\ \rule{3cm}{0.4pt}\ ]_V$

doc_2 : Attention models are awesome $\Rightarrow [\ \rule{3cm}{0.4pt}\ ]_V$

$\vdots$

$\underline{\underline{\text{doc}_N}}$: This course is boring $\Rightarrow [\ \rule{3cm}{0.4pt}\ ]_V$

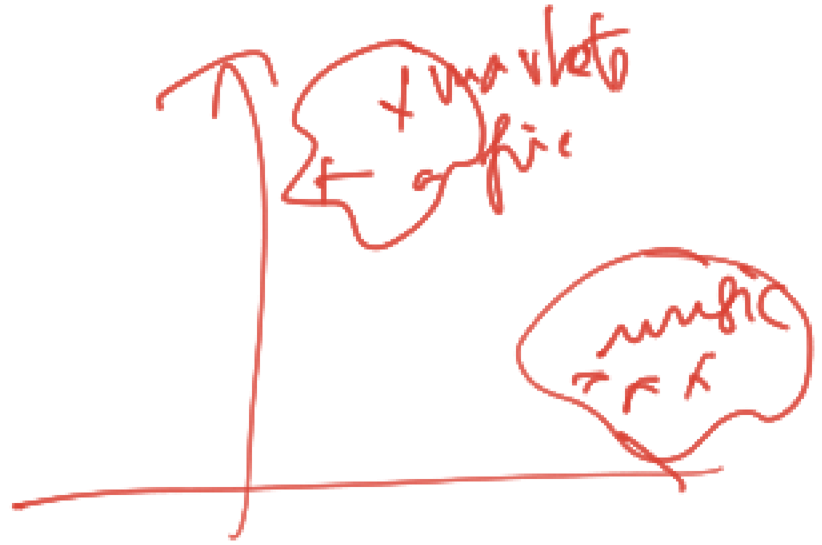
---

data processed is of shape (N, V)

Recap

GloVe approach

$D \sim 50-300$



$$W = \begin{bmatrix} — & w_1 & — \\ & \vdots & \\ & \vdots & \\ — & w_V & — \end{bmatrix} \in \mathbb{R}^{V \times D}$$

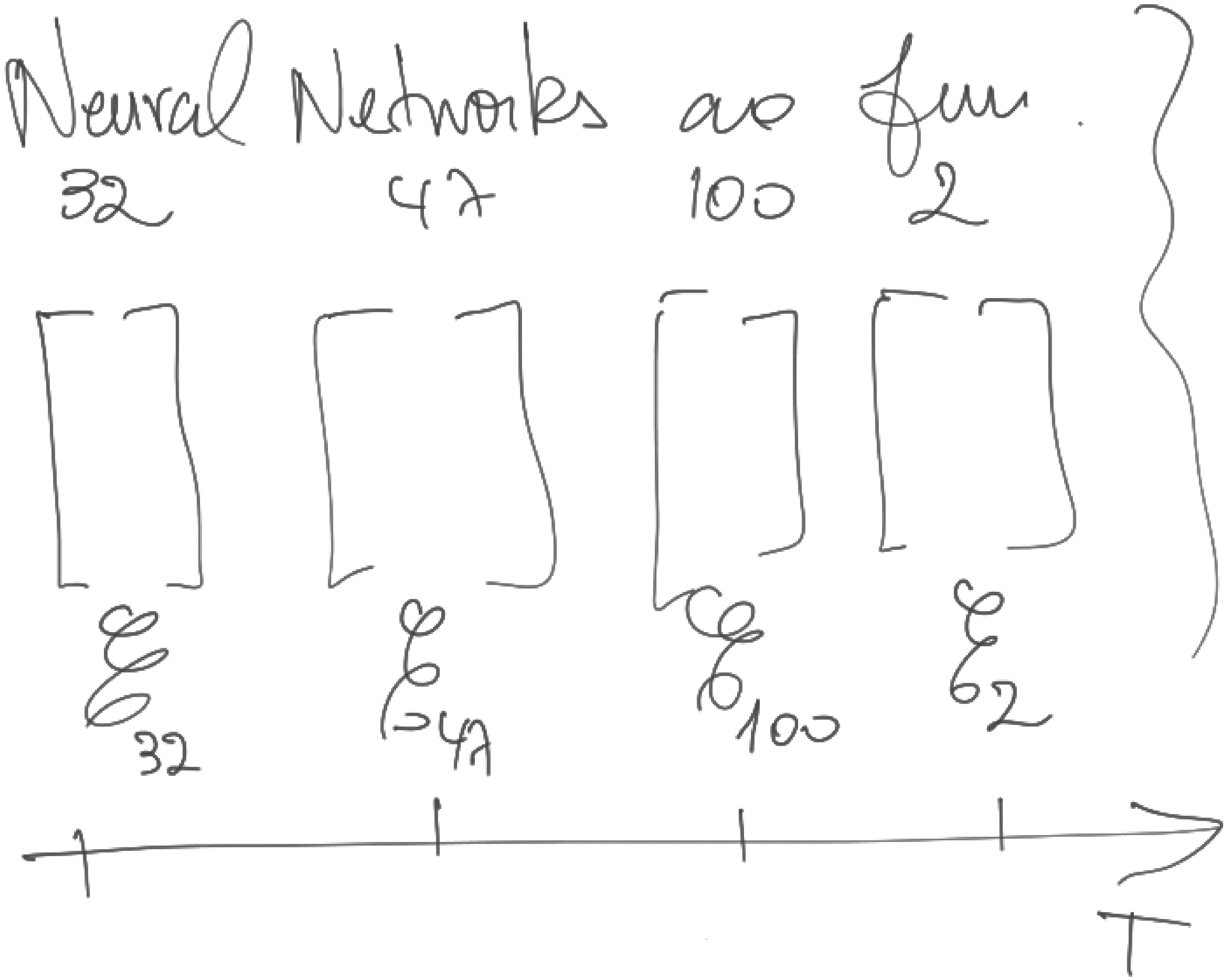
} D-dim representation of each word (as a center word)

(~ users)

$$\tilde{W} = \begin{bmatrix} — & \tilde{w}_1 & — \\ & \vdots & \\ — & \tilde{w}_V & — \end{bmatrix} \in \mathbb{R}^{V \times D}$$

} D-dim representati of each word (as a context word)

(~ movies)



Embedding Matrix (GloVe)

$$E = \begin{array}{c|c} \begin{bmatrix} \leftarrow & e_1 & \text{---} \\ \text{---} & e_2 & \text{---} \\ & \vdots & \\ \text{---} & e_V & \text{---} \end{bmatrix} & \begin{matrix} 1 \\ \vdots \\ \vdots \\ V \end{matrix} \end{array}$$

$\in \mathbb{R}^{V \times D}$

$\text{doc}_1 \longrightarrow [e_{32}, e_{47}, e_{100}, e_2] \in \underline{\underline{\mathbb{R}^{T \times D}}}$

$[\ ]_D \quad [\ ]_D \quad [\ ]_D \quad [\ ]_D$

$$\left.\begin{array}{l} \text{doc}_1 \longrightarrow (T, D) \\ \text{doc}_2 \longrightarrow (T, D) \\ \vdots \\ \text{doc}_N \longrightarrow (T, D) \end{array}\right\} \begin{array}{c}\text{Dataset} \\ \text{processed} \\ \underline{\underline{(N, T, D)}}\end{array}$$