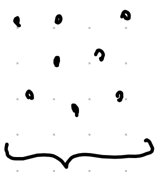
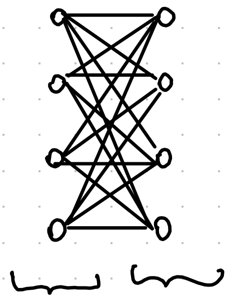
GRADIENT OF SOFTMAX + CROSS-ENTROPY

Let us consider the last layer of classification and the cost.



N size of dataset



$\longrightarrow L = \text{cross-entropy}$

K K

logits outputs

- $x \in \mathbb{R}^{N \times K}$ inputs
- $\tilde{y} \in \mathbb{R}^{N \times K}$ true targets (one-hot encoded)

$$L = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} \tilde{y}_{ik} \log \sigma_k(x_{i\cdot})$$

Recall that

$$\sigma_k(x_{i\cdot}) = \frac{e^{x_{ik}}}{\sum_{h=0}^{K-1} e^{x_{ih}}}$$

We want to compute

$$\frac{\partial L}{\partial x} \in \mathbb{R}^{K \times N}$$

$$\frac{\partial}{\partial x_{jl}} \Big(\sigma_k(x_{i,}) \Big) =$$ First of all:

0 if $j \neq i$

Next, we distinguish two cases:

<u>Case 1:</u> $k = l$

$$\frac{\partial}{\partial x_{ik}} \left(\frac{e^{x_{ik}}}{\sum_{h=0}^{K-1} e^{x_{ih}}} \right) = \frac{e^{x_{ik}}}{\sum_{h=0}^{K-1} e^{x_{ih}}} - \frac{\left(e^{x_{ik}}\right)^2}{\left(\sum_{h=0}^{K-1} e^{x_{ih}}\right)^2} =$$

$$= \sigma_k(x_{i\cdot}) - \sigma_k(x_{i\cdot})^2 =$$

$$= \sigma_k(x_{i\cdot}) \left(1 - \sigma_k(x_{i\cdot})\right)$$

Case 2: $k \neq l$

$$\frac{\partial}{\partial x_{il}} \left(\frac{e^{x_{ik}}}{\sum_{h=0}^{K-1} e^{x_{ih}}} \right) = - \frac{e^{x_{ik}}}{\left(\sum_{h=0}^{K-1} e^{x_{ih}}\right)^2} e^{x_{il}} =$$

$$= - \sigma_k(x_{i\cdot}) \sigma_l(x_{i\cdot})$$

In a compact way:

$$\frac{\partial}{\partial x_{jl}} \left(\sigma_k(x_{i\cdot})\right) = Id_{ij} \, \sigma_k(x_{i\cdot}) \left(Id_{kl} - \sigma_l(x_{i\cdot})\right)$$

We are ready to compute the differential of the loss:

$$\frac{\partial L}{\partial x_{jl}} = \frac{\partial}{\partial x_{jl}} \left(- \frac{1}{N} \sum_{i=0}^{N-1} \sum_{K=0}^{K-1} \tilde{y}_{ik} \log \sigma_k(x_{i\cdot}) \right) =$$

$$= - \frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} \tilde{y}_{ik} \frac{1}{\sigma_k(x_{i\cdot})} Id_{ij} \, \sigma_k(x_{i\cdot}) \left(Id_{kl} - \sigma_l(x_{i\cdot})\right)$$

$$= - \frac{1}{N} \sum_{k=0}^{K-1} \tilde{y}_{jk} \left(Id_{kl} - \sigma_l(x_{j\cdot})\right) =$$

$$= - \frac{1}{N} \tilde{y}_{jl} + \frac{1}{N} \sum_{k=0}^{K-1} \tilde{y}_{jk} \sigma_l(x_{j\cdot}) =$$

$$= -\frac{1}{N}\tilde{y}_{jl} + \frac{1}{N}\sigma_l(x_{j\cdot})$$

$$= \frac{1}{N}\left(\sigma_l(x_{j\cdot}) - \tilde{y}_{jl}\right)$$

Setting $q \in \mathbb{R}^{N \times K}$, $q_{jl} = \sigma_l(x_{j\cdot})$,

$$\frac{\partial L}{\partial x} = \frac{1}{N}(q - \tilde{y})^T \in \mathbb{R}^{K \times N}.$$