$$L_i = -\ln(p_{y_i}) \quad \text{where} \quad p_{y_i} = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}} = 1 - \frac{\sum_{j \neq y_i} e^{s_j}}{e^{s_{y_i}} + \sum_{j \neq y_i} e^{s_j}}$$

$$\frac{\partial}{\partial (s_{y_i})} p_{y_i} = \left(-\sum_{j \neq y_i} e^{s_j}\right) \underbrace{\left(\frac{1}{e^{s_{y_i}} + \sum_{j \neq y_i}(\cdots)}\right)'}$$

$$= (-1)\left(e^{s_{y_i}} + \sum_{j \neq y_i}(\cdots)\right)^{-2}\left(e^{s_{y_i}}\right)$$

$$= \frac{e^{s_{y_i}}\left(\sum_{j \neq y_i} e^{s_j}\right)}{\left(\sum_j e^{s_j}\right)^2} = p_{y_i}(1 - p_{y_i})$$

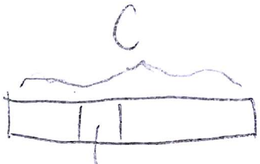
#

$$\frac{\partial L}{\partial (p_{y_i})} = -\frac{1}{p_{y_i}} \quad \#$$

So $$\frac{\partial L}{\partial s_{y_i}} = \left(\frac{\partial L}{\partial p_{y_i}}\right)\left(\frac{\partial (p_{y_i})}{\partial s_{y_i}}\right) = -\frac{1}{p_{y_i}} \; p_{y_i}(1 - p_{y_i}) = \underline{p_{y_i} - 1} \quad \#$$

$$\frac{\partial L}{\partial s_j} = \left(\frac{\partial L}{\partial p_{y_i}}\right)\left(\frac{\partial p_{y_i}}{\partial s_j}\right) = -\frac{1}{p_{y_i}} \left(-p_{y_i}\right)\left(p_j\right) = \underline{p_j} \quad \#$$

$$\frac{\partial L}{\partial s[i]} =$$



p_j if $j \neq y_i$

$p_{y_i} - 1$ if $j = y_i$

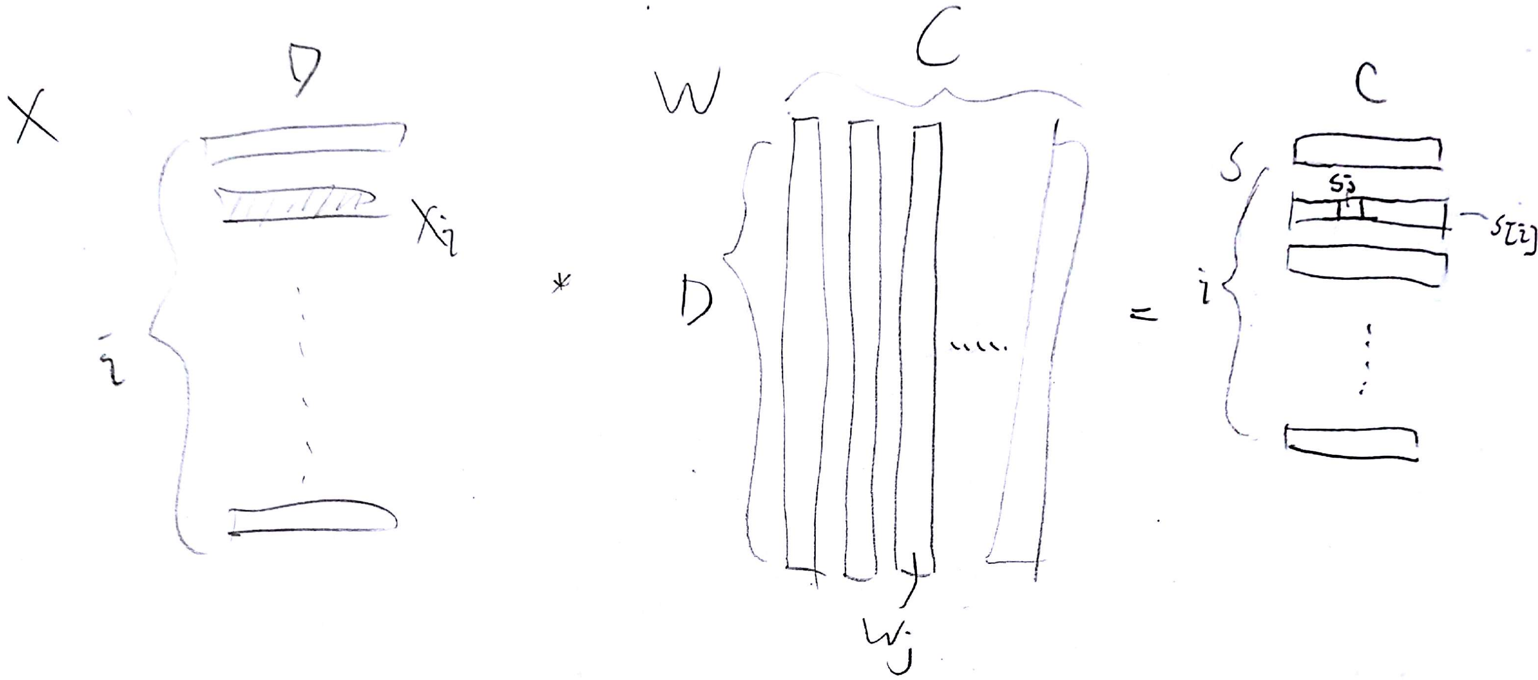
$$\frac{\partial}{\partial s_j}\left(p_{y_i}\right) = \left(e^{s_{y_i}}\right)\left(\frac{1}{e^{s_j}+\sum_{k\neq j} e^{s_k}}\right)'$$

$$= (-1)\left(e^{s_j}\right)\left(\frac{1}{e^{s_j}+\sum_{k\neq j} e^{s_k}}\right)^2$$

$$= -\frac{e^{s_{y_i}}\, e^{s_j}}{\left(\sum_k e^{s_k}\right)^2} = -\left(p_{y_i}\right)\left(p_j\right) \#$$

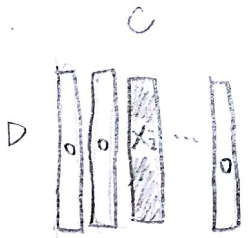
Scanned by CamScanner

$$\frac{\partial S_j}{\partial W_j} = X_i \quad \text{since} \quad S_j = X_i W_j$$

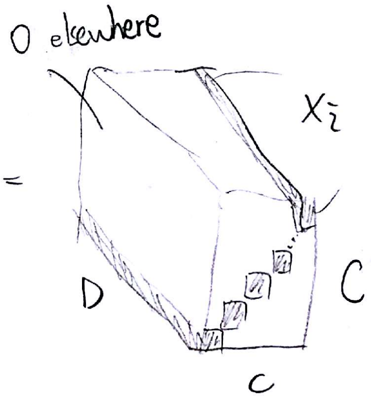


$$\frac{\partial S_j}{\partial W_j} = X_i$$

$$\frac{\partial S_j}{\partial W} =$$



$$\frac{\partial S[i]}{\partial W} =$$



size : $(D \times C \times C)$

$$\frac{\partial S}{\partial W} : i \times (D \times C \times C)$$

For one data point X_i,

$$\frac{\partial L}{\partial W} = \frac{\partial S[i]}{\partial W} \cdot \frac{\partial L}{\partial S[i]}$$