# Backpropagation

## ① Softmax derivative

Softmax derivative

$$\frac{\partial J}{\partial w} = \frac{\partial J}{\partial p} \times \frac{\partial p}{\partial z} \times \frac{\partial z}{\partial w}$$

Cross Entropy w.r.t Softmax

LR unit

$$\text{softmax}_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

What would be derivative of softmax w.r.t inputs

$$\nabla_{\text{softmax}} = \begin{bmatrix} \frac{\partial s_1}{\partial x_1} & \frac{\partial s_1}{\partial x_2} & \cdots & \frac{\partial s_1}{\partial x_n} \\ \frac{\partial s_2}{\partial x_1} & \frac{\partial s_2}{\partial x_2} & \cdots & \frac{\partial s_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial s_n}{\partial x_1} & & \cdots & \frac{\partial s_n}{\partial x_n} \end{bmatrix}$$

Jacobian →

Lets look at $s_i$

$$s_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Lets take log both sides

$$\log s_i = \log \left( \frac{e^{x_i}}{\sum e^{x_j}} \right)$$

$$= \log \left( e^{x_i} \right) - \log \left( \sum e^{x_j} \right)$$

$$\log s_i = x_i - \log \left( \sum e^{x_j} \right)$$

Diffrentiate w.r.t any $x_k$

$$\frac{\partial \log s_i}{\partial x_k} = \frac{\partial x_i}{\partial x_k} - \frac{\partial \log\left(\sum e^{x_i}\right)}{\partial x_k}$$

$$\frac{\partial x_i}{\partial x_k} = \begin{cases} 1 & x_i = x_k \\ 0 & \text{otherwise} \end{cases} = \mathbb{1}(x_i = x_k)$$

$= \delta_{ik}$ where $\mathbb{1}(x_i = x_k)$ (Indicator fun$^n$)

2nd term

$$\frac{\partial \log\left(\sum e^{x_i}\right)}{\partial x_k}$$

$$\left[ \frac{\partial \log g(x)}{\partial x} = \frac{1}{g(x)} \frac{\partial g(x)}{\partial x} \right]$$

$\cdot \backslash$

$$\Rightarrow \frac{1}{\sum e^{x_i}} \left( \frac{\partial \sum e^{x_i}}{\partial x_k} \right)$$

$$= \frac{1}{\sum e^{x_i}} \sum \frac{\partial e^{x_i}}{\partial x_k}$$

$$\delta_{jk} = \begin{cases} 1 & x_i = x_k \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \frac{1}{\sum e^{x_i}} \sum_{j=1}^{C} e^{x_i} \delta_{jk}$$

$$\Rightarrow \frac{e^{x_i}}{\sum e^{x_i}} = S_k$$

So final equation becomes

$$\frac{\partial \log S_i}{\partial x_k} = \delta_{ik} - S_k$$

$\delta_{ik}$ = Indicator func

$$\delta = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases}$$

$$\frac{1}{S_i}\frac{\partial S_i}{\partial x_k} = \delta_{ik} - S_k$$

$$\boxed{\frac{\partial S_i}{\partial x_k} = S_i\left(\delta_{ik} - S_k\right)}$$

So our final Jacobian becomes

$$\nabla_{softmax} = \begin{pmatrix} S_1(1-S_1) & -S_1S_2 & \cdots & -S_1S_n \\ -S_2S_1 & S_2(1-S_2) & \cdots & -S_2S_n \\ \vdots & & & \\ -S_nS_1 & \cdots & \cdots & S_n(1-S_n) \end{pmatrix}$$

## ② Differentiating Cross-Entropy loss

$$\mathcal{L}(s, y) = -\sum_{i=1}^{C} y_i \log s_i$$

$$\frac{\partial L(s, y)}{\partial x_k} = -\sum_{i=1}^{C} y_i \frac{\partial \log s_i}{\partial x_k}$$

$$\Rightarrow -\sum_{i=1}^{C} \left( \frac{y_i}{s_i} \frac{\partial s_i}{\partial x_k} \right)$$

$$= -\sum_{i=1}^{C} \left( \frac{y_i}{s_i} s_i (\delta_{ik} - s_k) \right)$$

$$= -\sum_{i=1}^{C} \left( y_i (\delta_{ik} - s_k) \right)$$

$$= -\sum_{i=1}^{C} \left( y_i \delta_{ik} - y_i s_k \right)$$

$\Rightarrow$

when $i=k$, only then

first term will become $y_k$

$$\frac{\partial L(s,y)}{\partial x_k} = -y_k + \sum_{i=1}^{C} y_k s_k$$

$$\Rightarrow \quad \sum_{i=1}^{C} y_k = 1 \quad \text{(Since its one-hot)}$$

Hence

$$\frac{\partial L(s,y)}{\partial x_k} \Rightarrow -y_k + s_k$$

$\partial x_k$

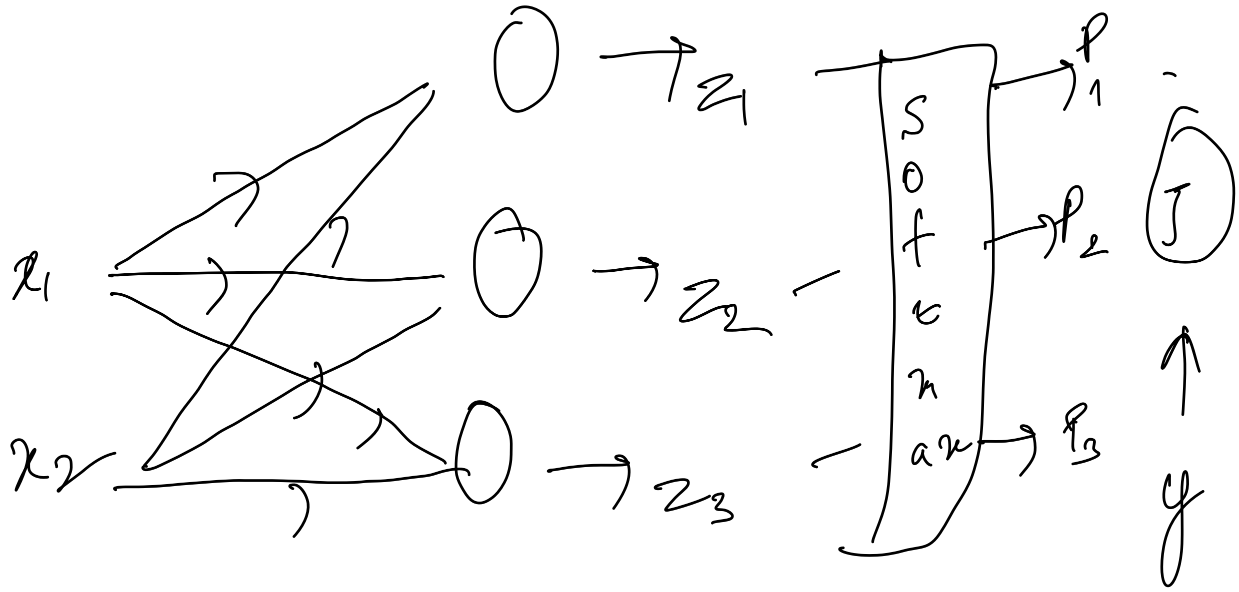
$$\frac{\partial L(s,y)}{\partial x_k} = s_k - y_k$$

In vectorized form this becomes

$$\left[ s_1 - y_1 , s_2 - y_2 \; \ldots \right]$$

$$\boxed{\frac{\partial L(x,y)}{\partial x} = s - y}$$

So now in our original chain equation




$$\frac{\partial J}{\partial w} = \frac{\partial J}{\partial p}\,\frac{\partial p}{\partial z}\,\frac{\partial z}{\partial w}\ ?$$

$$\frac{\partial z}{\partial w} = x$$