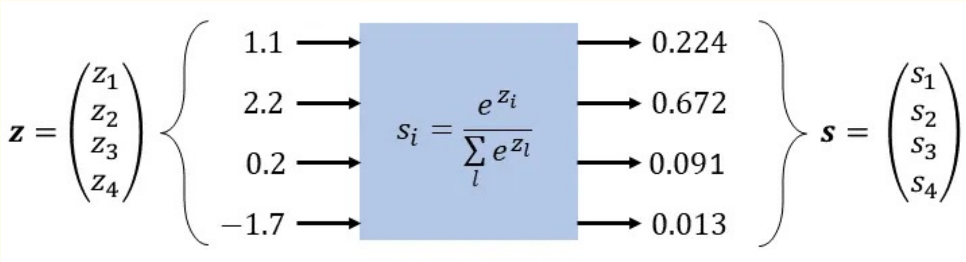
Softmax:



→ derivative of a softmax function is actually a Jacobian matrix.

↳ matrix first order partial derivative.

$$J = \begin{bmatrix} \frac{\partial s_1}{\partial z_1} & \frac{\partial s_1}{\partial z_2} & \cdots & \frac{\partial s_1}{\partial z_n} \\ \vdots & & & \vdots \\ \frac{\partial s_n}{\partial z_1} & \cdots & & \frac{\partial s_n}{\partial z_n} \end{bmatrix}$$

$$s_i = \frac{e^{z_i}}{\sum_{l=1}^{n} e^{z_l}} \quad \forall\ i = 1 \ldots n$$

→ lets calculate the derivative:

→ as output of softmax are all +ve we can take log and then take derivative.

$$\frac{\partial}{\partial z_j}\left(\log(s_i)\right) = \frac{1}{s_i} \cdot \frac{\partial s_i}{\partial z_j}$$

$$\Rightarrow \boxed{\frac{\partial s_i}{\partial z_j} = s_i \cdot \frac{\partial}{\partial z_j}\left(\log(s_i)\right)} \longrightarrow (1)$$

$$\rightarrow \log(s_i) = \log\left(\frac{e^{z_i}}{\sum_{l=1}^{n} e^{z_l}}\right)$$

$$= z_i - \log\left(\sum_{l=1}^{n} e^{z_l}\right)$$

$\rightarrow$ now $\frac{\partial}{\partial z_j}(\log s_i)$

$$= \frac{\partial}{\partial z_j}\left(z_i - \log\left(\sum_{l=1}^{n} e^{z_l}\right)\right)$$

$$\frac{\partial}{\partial z_j}(z_i) = \begin{cases} 1 = \text{if } i=j \\ 0 \quad \text{otherwise (as const)} \end{cases}$$

$\hookrightarrow$ Indicator function

$$= 1(i == j)$$

$$\frac{\partial}{\partial z_j}\left[\log\left(\sum_{l=1}^{n} e^{z_l}\right)\right]$$

$$= \frac{1}{\sum_{l=1}^{n} e^{z_l}} \cdot \frac{\partial}{\partial z_j} \sum_{l=1}^{n} e^{z_l}$$

$$= \frac{1}{\sum_{l=1}^{n} e^{z_l}} \cdot \frac{\partial}{\partial z_j}\left[e^{z_1} + \dots + e^{z_n}\right]$$

$$= \frac{e^{z_j}}{\sum_{l=1}^{n} e^{z_l}} = s_j$$

$$\Rightarrow \left[0 + \dots e^{z_j} + \dots 0\right]$$

$$\frac{\partial}{\partial z_j}(\log s_i) = \mathbb{1}\{i==j\} - s_j \quad \text{—— ②}$$

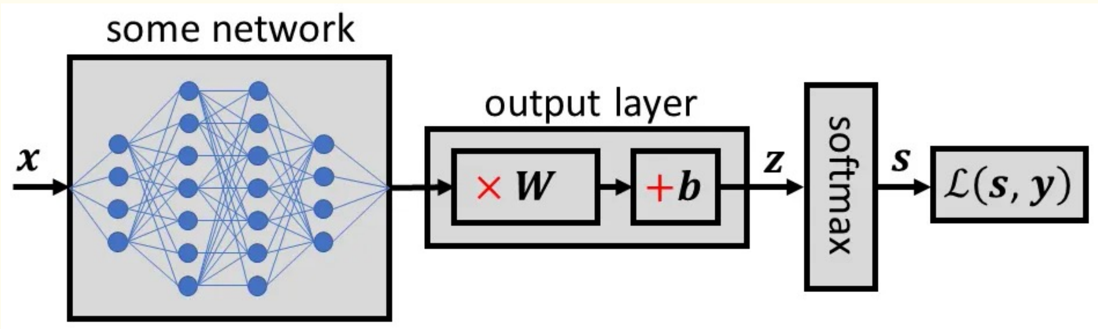
now put ② in ①:

$$\frac{\partial s_i}{\partial z_j} = s_i * (\mathbb{1}(i==j) - s_j) \quad \text{—— ③}$$

$$J_{softmax} = \begin{pmatrix} s_1 \cdot (1 - s_1) & -s_1 \cdot s_2 & -s_1 \cdot s_3 & -s_1 \cdot s_4 \\ -s_2 \cdot s_1 & s_2 \cdot (1 - s_2) & -s_2 \cdot s_3 & -s_2 \cdot s_4 \\ -s_3 \cdot s_1 & -s_3 \cdot s_2 & s_3 \cdot (1 - s_3) & -s_3 \cdot s_4 \\ -s_4 \cdot s_1 & -s_4 \cdot s_2 & -s_4 \cdot s_3 & s_4 \cdot (1 - s_4) \end{pmatrix}$$

Cross-Entropy loss:

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



$$\frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial s_i} \times \frac{\partial s_i}{\partial z_j}$$

$$= \frac{\partial}{\partial s_i}\left(-\sum_{i=1}^{c} y_i \log(s_i)\right) \times s_i \times (\mathbb{1}\{i=j\} - s_j)$$

$$= -\sum_{i=1}^{c} \frac{y_i}{\cancel{s_i}} \times \cancel{s_i} \times (\mathbb{1}\{i=j\} - s_j)$$

$$= -\sum_{i=1}^{c} y_i (\mathbb{1}\{i=j\} - s_j)$$

$$= \sum_{i=1}^{C} y_i s_j - \sum_{i=1}^{C} y_i \times 1\{i=j\}$$

$$= s_j \sum_{i=1}^{C} y_i - \sum_{i=1}^{C} y_i$$

$\hookrightarrow$ 1 only when $i == j$ else 0

$\hookrightarrow$ so $= y_j$

$$= s_j - y_j$$

$$\Rightarrow \boxed{\frac{\partial L}{\partial z_j} = s_j - y_j}$$

- $\sum_{i=1}^{C} y_i = 1$ as y is one-hot encoded value