

Derive Gradient ~~Descent~~ For $\boxed{\rightarrow \text{softmax} \rightarrow \text{cross entropy} \rightarrow}$

1. $Z \rightarrow \boxed{\text{softmax function}} \rightarrow \text{softmax_out}$

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix} \rightarrow \text{softmax_out}_i = \frac{e^{z_i}}{\sum_{j=0}^k e^{z_j}} \rightarrow \begin{bmatrix} \text{softmax_out}_0 \\ \text{softmax_out}_1 \\ \text{softmax_out}_2 \\ \vdots \\ \text{softmax_out}_k \end{bmatrix}$$

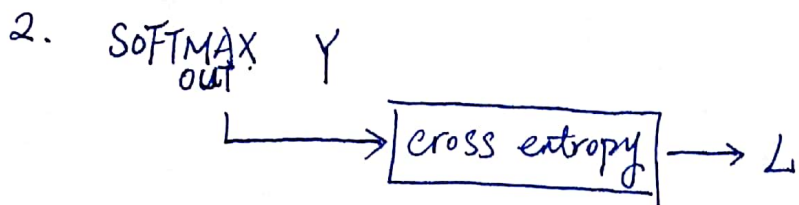
$$\frac{\partial \text{softmax_out}_n}{\partial z_m} = \frac{\partial}{\partial z_m} e^{z_n} (e^{z_0} + e^{z_1} + \dots + e^{z_k})^{-1}$$

let $e^{z_0} + e^{z_1} + \dots + e^{z_k} = S$

$$= \begin{cases} e^{z_n} (S)^{-1} + e^{z_n} \cdot (-1) (S)^{-2} e^{z_n}, & \text{when } m=n \\ (-1) e^{z_n} (S)^{-2} e^{z_m}, & \text{when } m \neq n \end{cases}$$

$$= \begin{cases} e^{z_n} (S)^{-1} (1 - e^{z_n} (S)^{-1}), & \text{when } m=n \\ -(e^{z_n} S^{-1}) (e^{z_m} S^{-1}), & \text{when } m \neq n \end{cases}$$

$$= \begin{cases} \text{softmax_out}_n (1 - \text{softmax_out}_n), & \text{when } m=n \\ -(\text{softmax_out}_n) (\text{softmax_out}_m), & \text{when } m \neq n \end{cases}$$



$$\begin{bmatrix} \text{softmax_out}_0 \\ \text{softmax_out}_1 \\ \vdots \\ \text{softmax_out}_k \end{bmatrix} \quad \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix}$$

$$\rightarrow \left[- \sum_{i=0}^k y_i \log(\text{softmax_out}_i) \right] \rightarrow L$$

$$\frac{\partial L}{\partial \text{softmax_out}_m} = -(y_m) \left(\frac{1}{\text{softmax_out}_m} \right)$$

3. Combine 1. and 2.

$$\begin{aligned} \frac{\partial L}{\partial z_m} &= \frac{\partial L}{\partial \text{SOFTMAX_OUT}} * \frac{\partial \text{SOFTMAX_OUT}}{\partial z_m} \\ &= \sum_{n=0}^k \frac{\partial L}{\partial \text{softmax_out}_n} * \frac{\partial \text{softmax_out}_n}{\partial z_m} \end{aligned}$$

assume $y_j = 1$ for $j = l$ else $y_j = 0$ recall that Y is a one hot vector

$$= -y_l \left(\frac{1}{\text{softmax_out}_l} \right) * \frac{\partial \text{softmax_out}_l}{\partial z_m}$$

$$= \begin{cases} -y_m \left(\frac{1}{\text{softmax_out}_m} \right) (\text{softmax_out}_m) (1 - \text{softmax_out}_m) & \text{when } l=m \\ + y_l \left(\frac{1}{\text{softmax_out}_l} \right) (\text{softmax_out}_l) (\text{softmax_out}_m) & \text{when } l \neq m \end{cases}$$

$$= \text{softmax_out}_m - y_m$$