CS224n - A2 Written Part — Sara Kodeiri 96521443

(a) $y \rightarrow$ true distribution $\quad \hat{y} \rightarrow$ predicted distribution

$$-\sum_{w \in V} y_w \log(\hat{y}_w) = -\left(y_1 \log(\hat{y}_1) + y_2 \log(\hat{y}_2) + \cdots + y_V \log(\hat{y}_V)\right) = -\log(\hat{y}_o)$$

all these values are zero except when $w = o$. In that case, $y_o = 1$

(b) $J_{nsm} = -\log\left(e^{u_o^T v_c} \times \left(\sum_{w \in V} e^{u_w^T v_c}\right)\right) = \log\left(\sum_{w \in V} e^{u_w^T v_c}\right) - \log\left(e^{u_o^T v_c}\right)$

(J_A) (J_B)

$J = J_A - J_B$

$$\Rightarrow \frac{\partial J}{\partial v_c} = \frac{\partial J_A}{\partial v_c} - \frac{\partial J_B}{\partial v_c} \qquad \frac{\partial J_A}{\partial v_c} = \frac{\partial (u_o^T v_c)}{\partial v_c} = u_o$$

$$\frac{\partial J_B}{\partial v_c} = \frac{\partial \left(\log \sum_{w \in V} e^{u_w^T v_c}\right)}{\partial v_c} = \frac{1}{\sum_{w \in V} e^{u_w^T v_c}} \sum_{s \in V} u_s e^{u_s^T v_c} \rightarrow \hat{y}_s$$

chain rule

$$\Rightarrow \frac{\partial J_B}{\partial v_c} = \sum_{s \in V} \hat{y}_s u_s$$

$$\Rightarrow \frac{\partial J_{nsm}}{\partial v_c} = \sum_{w \in V} \hat{y}_w u_w - u_o$$

(c.1) $w = 0$

$$\frac{\partial J_{nsm}}{\partial u_w} = \frac{\partial (-\log(\hat{y}_o))}{\partial u_w} = -\frac{\partial}{\partial u_w}\left(\log \frac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}\right)$$

$$= \frac{-\partial}{\partial u_w}(u_o^T v_c) + \frac{\partial}{\partial u_w}\left(\log\left(\sum_{w \in V} e^{u_w^T v_c}\right)\right) \quad \Leftarrow \text{generalized}$$

now we use the fact that $w = 0$ ⇒

$$-v_c + \frac{1}{\sum_{w \in V} e^{u_w^T v_c}}\left(e^{u_w^T v_c} v_c\right)$$

$$= -v_c + v_c \hat{y}_w = v_c(\hat{y}_o - 1)$$

(c.2) $w \neq 0$

$$0 + v_c \hat{y}_w = v_c \hat{y}_w$$

(d) $$\frac{\partial J_{nsm}}{\partial U} = \left[\frac{\partial J(v_c, o, U)}{\partial u_1}, \frac{\partial J}{\partial u_2}, \cdots, \frac{\partial J}{\partial u_V} \right]$$

(e) $$\sigma'(x) = \left(\frac{e^x}{e^x+1}\right)' = \frac{-e^{-x}}{-(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})} \times \frac{1}{(1+e^{-x})}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \Rightarrow e^{-x} = \frac{1}{\sigma(x)} - 1 \qquad e^{-x}\,\sigma(x) \times \sigma(x)$$

$$\Rightarrow \sigma'(x) = \left(\frac{1}{\sigma(x)} - 1\right) \times \sigma(x) \times \sigma(x) = \sigma(x)\left(1 - \sigma(x)\right)$$

(f) $$\frac{\partial J_{ns}}{\partial v_c} = -\frac{\partial}{\partial v_c}\log\left(\sigma(u_o^T v_c)\right) - \frac{\partial}{\partial v_c}\left(\sum_{k=1}^{K}\log\left(\sigma(-u_k^T v_c)\right)\right) \quad (2)$$

$$\frac{-1}{\sigma(u_o^T v_c)}\left(u_o\,\sigma(u_o^T v_c)\left(1-\sigma(u_o^T v_c)\right)\right) = u_o\left(\sigma(u_o^T v_c) - 1\right)$$

(2): $$\sum_{k=1}^{K} \frac{-u_k\,\sigma(A)\left(1-\sigma(A)\right)}{\sigma(A)} = \sum_{k=1}^{K} -u_k\left(1-\sigma(-u_k^T v_c)\right)$$

$$\frac{\partial J_{ns}}{\partial v_c} = u_o\left(\sigma(u_o^T v_c) - 1\right) + \sum_{k=1}^{K} u_k\left(1-\sigma(-u_k^T v_c)\right)$$

- $\frac{\partial J_{ns}}{\partial u_o}$ ⇒ the Σ doesn't have any u_o in it ! ↑ just like previous part. The first part is
  ↳ $v_c\left(\sigma(u_o^T v_c) - 1\right)$

- $\frac{\partial J_{ns}}{\partial u_K}$ ⇒ First part has no u_K. In the second part, since there's only one match, Σ gets out of the way.

$$\frac{\partial J_{ns}}{\partial u_K} = v_c\left(1 - \sigma(-u_K^T v_c)\right)$$

- The naive softmax considers the whole vocabulary for computation but the negative-sampling loss function only needs k samples and $K \ll V$ & is more efficient.

(g) $J_{ns} = -\log(\sigma(u_o^T v_c)) - \sum_{k \in A} \log(\sigma(-u_k^T v_c)) - \sum_{k \in B} \log(\sigma(-u_k^T v_c))$

↳ equal to u_k ↳ unequal to u_k

$$\frac{\partial J_{ns}}{\partial u_k} = 0 - \sum_{k \in A} \log(\sigma(-u_k^T v_c)) - 0$$

$$= n v_c (1 - \sigma(-u_k^T v_c))$$

(h) (i) $\left[\frac{\partial J}{\partial w_{t-m}}, \frac{\partial J}{\partial w_{t-m+1}}, \ldots, \frac{\partial J}{\partial w_{t+m}}\right]$

(ii) $\sum_{-m \leq j \leq m} \frac{\partial J_{sg}(v_c, w_{t+j}, U)}{\partial v_c}$

(iii) Zero