CS224n Assignment 2 Written

(a) $y_w = 0$ everywhere except $w = o \rightarrow -\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -y_o \log(\hat{y}_o)$.
Since $y_o = 1$, $-y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$

(b) $J = -\log\left(\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}\right)$; $\sum_w$ denotes $\sum_{w \in Vocab}$.

$= -(u_o^T v_c) + \log\left(\sum_w \exp(u_w^T v_c)\right)$

$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \sum_w \exp(u_w^T v_c) \cdot u_w$

$u_o = U_y$ $= -u_o + \sum_w u_w \cdot \left(\frac{\exp(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)}\right)$ $\rightarrow \hat{y}_w$

$= -U_y + \sum_w u_w \cdot \hat{y}_w = -Uy + U\hat{y} = \boxed{U(\hat{y} - y)}$.

(1) Gradient is zero when $y = \hat{y}$

(2) Since we're optimizing v_c to minimize J, we're attempting to maximize $\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}$, which primarily involves increasing the dot product between u_o and v_c, which requires bringing v_c closer to u_o. This is also evident in the intermediate steps of deriving $\frac{\partial J}{\partial v_c}$, where subtracting $-u_o$ implies bringing v_c closer to u_o.

(c) $\frac{\partial J}{\partial u_i} = \begin{cases} i = o: & -v_c + \left(\frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \exp(u_i^T v_c)\right) \cdot v_c = (\hat{y}_i - 1) v_c \\ i \neq o: & \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \exp(u_i^T v_c) \cdot v_c = \hat{y}_i v_c \end{cases}$ ($\hat{y}_i$)

$= (\hat{y}_i - y_i) v_c$

(d) $\frac{\partial J}{\partial U} = \left[v_c(\hat{y}_1 - y_1), v_c(\hat{y}_2 - y_2), \cdots v_c(\hat{y}_{|Vocab|} - y_{|Vocab|})\right]$

$= v_c (\hat{y} - y)^T$

(e) $\frac{df}{dx} = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases}$

(f) $\frac{d\sigma}{dx} = \frac{(e^x+1)\cdot e^x - e^x \cdot e^x}{(e^x+1)^2}$

$= \frac{e^x}{(e^x+1)^2} = \frac{e^x}{e^x+1} \cdot \frac{1}{e^x+1} = \boxed{\sigma(x)\cdot(1-\sigma(x))}$

(g)(i) $\frac{\partial J}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)\cdot(1-\sigma(u_o^T v_c))\cdot u_o$

$-\sum_{s=1}^{K} \frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c)\cdot(1-\sigma(-u_{w_s}^T v_c))\cdot(-u_{w_s})$

$= \boxed{-(1-\sigma(u_o^T v_c))u_o + \sum_{s=1}^{K}(1-\sigma(-u_{w_s}^T v_c))u_{w_s}}$

$\frac{\partial J}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)\cdot(1-\sigma(u_o^T v_c))\cdot v_c$

$= -(1-\sigma(u_o^T v_c))v_c = \boxed{(\sigma(u_o^T v_c)-1)v_c}$

$\frac{\partial J}{\partial u_{w_s}} = -\frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c)\cdot(1-\sigma(-u_{w_s}^T v_c))\cdot(-v_c)$

$= \boxed{(1-\sigma(-u_{w_s}^T v_c))v_c}$

(ii) $1-\sigma(U^T v_c)$

(iii) It only needs to backpropagate to K instead of |Vocab| outside vectors.

(h) $\frac{\partial J}{\partial u_{w_i}} = \frac{\partial}{\partial u_{w_i}}\left(-\log(\sigma(u_o^T v_c))\right) - \frac{\partial}{\partial u_{w_i}} \sum_{\substack{s=1 \\ w_i = w_s}}^{K} \log(\sigma(-u_{w_s}^T v_c))$

$- \frac{\partial}{\partial u_{w_i}} \sum_{\substack{s=1 \\ w_i \neq w_s}}^{K} \log(\sigma(-u_{w_o}^T v_c))$

$= 0 - \sum_{\substack{s=1 \\ w_i = w_s}}^{K} \frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c)\cdot(1-\sigma(-u_{w_s}^T v_c))\cdot(-v_c)$

$- \sum_{\substack{s=1 \\ w_i \neq w_s}}^{K} 0$

$$= \boxed{\sum_{\substack{\delta=1 \\ w_i = w_s}}^{K} \left(1 - \sigma(-u_{w_s}^T v_c)\right) v_c}$$

(i) (i) $\dfrac{\partial J(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial U}$

$$= \boxed{\sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}}$$

(ii) $\dfrac{\partial J(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_c}$

$$= \boxed{\sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}}$$

(iii) $\dfrac{\partial J(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_w}$

$$= \boxed{0}$$