(a)

$$-\sum_{w\in Vocab} y_w \log(\hat{y}_w) = -\left[0\cdot\log(\hat{y}_0) + 0\cdot\log(\hat{y}_1) + \cdots + 1\cdot\log(\hat{y}_o) + \cdots + 0\cdot\log(\hat{y}_{Vocab}) \right]$$

$$= -\log(\hat{y}_o) \quad \#$$

$$J_{naive\text{-}softmax}(v_c, o, U) = -\log \frac{\exp(u_o^T v_c)}{\sum_{w\in Vocab} \exp(u_w^T v_c)}$$

$$= \underbrace{-u_o^T v_c}_{①} + \underbrace{\log \sum_{w\in Vocab} \exp(u_w^T v_c)}_{②}$$

(b)

$$\frac{\partial ①}{\partial v_c} = -u_o = -Uy$$

$$\frac{\partial ②}{\partial v_c} = \frac{1}{\sum_{w'\in Vocab} \exp(u_{w'}^T v_c)} \cdot \frac{\partial}{\partial v_c}\left[\sum_{w\in Vocab} \exp(u_w^T v_c)\right]$$

$$= \frac{1}{\sum_{w'\in Vocab} \exp(u_{w'}^T v_c)} \cdot \sum_{w\in Vocab}\left[\exp(u_w^T v_c)\cdot u_w\right]$$

$$= \sum_{w\in Vocab}\left[\frac{\exp u_w^T v_c}{\sum_{w'\in Vocab} \exp(u_{w'}^T v_c)} u_w\right] = \sum_{w\in Vocab} \hat{y}_w \cdot u_w$$

$$= U\hat{y}$$

$$\Rightarrow \frac{\partial J_{naive\text{-}softmax}}{\partial v_c} = U(\hat{y}-y) \quad \#$$

① the gradient is zero when $\hat{y} = y$

② subtracting the gradient is to update v_c to let y be closer t the one-hot vector whose o entry is 1. It is equivalent to increase $P(O=o|C=$

(c)

if $w=o$,

$$\frac{\partial ①}{\partial u_w} = -v_c$$

non-zero only when $w''=w$

$$\frac{\partial ②}{\partial u_w} = \frac{1}{\sum_{w'\in Vocab} \exp(u_{w'}^T v_c)} \cdot \frac{\partial}{\partial u_w}\left(\sum_{w''\in Vocab} \exp(u_{w''}^T v_c)\right)$$

$$= \frac{1}{\sum_{w'\in Vocab} \exp(u_{w'}^T v_c)} \cdot \exp(u_w^T v_c) \cdot v_c$$

$$= \hat{y}_w \cdot v_c$$

$$\Rightarrow \frac{\partial J_{naive\text{-}softmax}}{\partial u_w} = (\hat{y}_w - 1) v_c \overset{w=o}{=} (\hat{y}_o - 1) v_c \quad \#$$

if $w \neq o$

$$\frac{\partial ①}{\partial u_w} = 0$$

$$\frac{\partial ②}{\partial u_w} = \hat{y}_w \cdot v_c$$

$$\Rightarrow \frac{\partial J_{naive\text{-}softmax}}{\partial u_w} = \hat{y}_w \cdot v_c \quad \#$$

(d) $$\frac{\partial J_{naive\text{-}softmax}}{\partial U} = \begin{bmatrix} \frac{\partial J}{\partial u_1} & \frac{\partial J}{\partial u_2} & \cdots & \frac{\partial J}{\partial u_{|Vocab|}} \end{bmatrix} \in R^{d \times |Vocab|} \quad \#$$

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

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

$$= \sigma(x)[1-\sigma(x)] \quad \#$$

$$J_{neg\text{-}sample}(v_c, o, U) = \underbrace{-\log(\sigma(u_o^T v_c))}_{①} \underbrace{- \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))}_{②}$$

$$② \overset{\because\, \sigma(-x) = 1-\sigma(x)}{=} -\sum_{k=1}^{K} \log(1-\sigma(u_k^T v_c))$$

f)

$$\frac{\partial ①}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma'(u_o^T v_c) \cdot u_o$$

$$= [\sigma(u_o^T v_c) - 1] u_o$$

$$\frac{\partial ②}{\partial v_c} = -\sum_{k=1}^{K} \left[\frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma'(-u_k^T v_c) \cdot -u_k \right]$$

$$= \sum_{k=1}^{K} [1 - \sigma(-u_k^T v_c)] u_k \overset{\because\, \sigma(-x)=1-\sigma(x)}{=\!=} \sum_{k=1}^{K} \sigma(u_k^T v_c) u_k$$

$$\Rightarrow \frac{\partial J_{neg\text{-}sample}}{\partial v_c} = [\sigma(u_o^T v_c) - 1] u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c) u_k$$

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

$$\frac{\partial ②}{\partial u_o} = \emptyset \quad \text{because} \quad o \notin \{w_1, w_2, \ldots, w_K\}$$

$$\Rightarrow \frac{\partial J_{neg\text{-}sample}}{\partial u_o} = [\sigma(u_o^T v_c) - 1] v_c \quad \#$$

$$\frac{\partial ①}{\partial u_k} = \emptyset \quad \text{because} \quad o \notin \{w_1, w_2, \ldots, w_K\}$$

$$\frac{\partial ②}{\partial u_k} = \frac{-\partial}{\partial u_k} \sum_{k'=1}^{K} \log(\sigma(-u_{k'}^T v_c)) = \frac{-\partial}{\partial u_k} \log(\sigma(-u_k^T v_c))$$

$$= \frac{-1}{\sigma(-u_k^T v_c)} \cdot \sigma'(-u_k^T v_c) \cdot (-v_c) = [1 - \sigma(-u_k^T v_c)] v_c$$

$$\Rightarrow \frac{\partial J_{neg\text{-}sample}}{\partial u_k} = [1 - \sigma(-u_k^T v_c)] v_c = \sigma(u_k^T v_c) v_c \quad \#$$

(g) When K negative samples may not be distinct

$$\frac{\partial ①}{\partial u_k} = 0 \quad \text{because } o \notin \{w_1, w_2, \dots w_K\}$$

$$\frac{\partial ②}{\partial u_k} = \frac{-\partial}{\partial u_k}\left(\sum_{\substack{1 \le k' \le K \\ w_{k'} = w_k}} \log \sigma(-u_{k'}^T v_c) + \sum_{\substack{1 \le k' \le K \\ w_{k'} \ne w_k}} \log \sigma(-u_{k'}^T v_c) \right)$$

$$= -\sum_{\substack{1 \le k' \le K \\ w_{k'} = w_k}} \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma'(-u_k^T v_c) \cdot (-v_c)$$

$$= \sum_{\substack{1 \le k' \le K \\ w_{k'} = w_k}} \left[1 - \sigma(-u_k^T v_c)\right] v_c = \sum_{\substack{1 \le k' \le K \\ w_{k'} = w_k}} \sigma(u_k^T v_c)\, v_c$$ #

(h)

$$J_{skip\text{-}gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(v_c, w_{t+j}, U)$$

(i) $$\frac{\partial J_{skip\text{-}gram}}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$ #

(ii) $$\frac{\partial J_{skip\text{-}gram}}{\partial v_c} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$ #

(iii) $$\frac{\partial J_{skip\text{-}gram}}{\partial v_w} = \emptyset \quad \text{when } w \ne c$$ #