A2.

1(a) To prove : $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$ with one line.

(vectors: y, $\hat{y}$, scalar: $\hat{y}_o$)

Since y is one hot, $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$

(b)(i) To calculate: $\frac{\partial J}{\partial v_c}(v_c, o, U)$. in terms of $y, \hat{y}, U$.

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

$$\frac{\partial J}{\partial v_c} = -\frac{1}{P(O=o \mid C=c)} \frac{\partial P}{\partial v_c} = -\frac{1}{P}\left(\frac{\partial(\exp(u_o^T v_c))}{\partial v_c} \cdot \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} - \frac{\exp(u_o^T v_c)}{(\sum_{w \in Vocab} \exp(u_w^T v_c))^2} \cdot \frac{\partial(\sum_{w \in Vocab} \exp(u_w^T v_c))}{\partial v_c}\right)$$

$$= -\cancel{\frac{1}{P}} \cdot \frac{1}{\cancel{\sum_{w \in Vocab} \exp(u_w^T v_c)}} \cdot diag(\exp(u_o^T \cdot v_c) \cdot u_o) \cancel{+} - \frac{1}{P} \frac{\exp(u_o^T \cdot v_c)}{(\sum_{w \in Vocab} \exp(u_w^T \cdot v_c))^2} \cdot diag(\sum_{w \in Vocab} \exp(u_w^T v_c) \cdot u_w)$$

$$= -\frac{1}{P} \cdot \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \left(diag(\exp(u_o^T v_c) \cdot u_o) - p \cdot diag(\sum_{w \in Vocab} \exp(u_w^T v_c) \cdot u_w)\right).$$

$$= -diag(u_o) + \frac{1}{\cancel{\sum_{w \in Vocab} \exp(u_w^T v_c)}} \cdot diag(\cancel{\sum_{w \in Vocab} \exp}(u_w^T v_c) \cdot u_w).$$

$$= -Uy^T + U\hat{y}^T.$$ ($u_o = U^T y$: 从 U matrix 中取 o 对应向量)

(ii) iff $y = \hat{y}$.

(iii) $\frac{\partial J}{\partial v_c} = -U(y^T - \hat{y}^T)$. (?)

(c) consider the case where $u_x = \alpha \cdot u_y$ for words $x \neq y$, scalar α.

phrase P_1(with x) before L_2: $u_z + u_x = u_z + \alpha \cdot u_y$

after L_2: $\frac{u_z}{\|u_z\|} + \frac{u_x}{\|u_x\|} = \frac{u_z}{\|u_z\|} + \frac{u_y}{\|u_y\|}$.

replace x to y in P_1: before L_2: $u_z + u_y$

after L_2: $\frac{u_z}{\|u_z\|} + \frac{u_y}{\|u_y\|}$

this takes away the information of α.

if $u_z \cdot u_y < 0$, the value of α can possibly change the result.

(d) To calculate $\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial u_w}$ (two cases: $w=o$ and $w \neq o$).

in terms of $y, \hat{y}, v_c$.

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

case 1: $w=o$.

$$\frac{\partial J}{\partial u_o} = -\frac{1}{P} \cdot \frac{\partial P}{\partial u_o} = -\frac{1}{P}\left(\frac{\partial(\exp(u_o^T v_c))}{\partial u_o} \cdot \frac{1}{\sum_{w \in vocab}(\exp(u_w^T v_c))} - \frac{\exp(u_o^T v_c)}{(\sum_{w \in vocab}(\exp(u_w^T v_c)))^2} \cdot \frac{\partial(\sum_{w \in vocab}(\exp(u_w^T v_c)))}{\partial u_o}\right)$$

$$= -\frac{1}{P}\left(\frac{\exp(u_o^T v_c)}{\sum_{vocab}(\exp(u_w^T v_c))} \cdot \frac{\partial(u_o^T \cdot v_c)}{\partial u_o} - \frac{\exp^2(u_o^T \cdot v_c)}{(\sum_{vocab}(\exp(u_w^T v_c)))^2} \cdot \frac{\partial(u_o^T v_c))}{\partial u_o}\right)$$

$$= -v_c + P(O=o|C=c) v_c = (\hat{y}_o - 1) v_c.$$

case 2: $w \neq o$.

$$\frac{\partial J}{\partial u_w} = -\frac{1}{P} \cdot \frac{\partial P}{\partial u_w} = \frac{1}{P} \cdot \frac{\exp(u_w^T v_c)}{(\sum_{w \in vocab}(\exp(u_w^T v_c)))^2} \cdot \frac{\partial(\sum_{w \in vocab}(\exp(u_w^T v_c)))}{\partial u_w} = \hat{y}_w v_c.$$

(e): $\frac{\partial J}{\partial u} = (\hat{y} - y) v_c$.

(f): Leaky ReLU. $f(x) = \max(\alpha x, x)$. $0 < \alpha < 1$.

$$f'(x) = \begin{cases} 1, & x > 0 \\ \alpha, & x < 0 \end{cases}$$

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

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

(h) Negative sampling loss. K negative samples (words)

$$J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T v_c))$$

(i) $$\frac{\partial J_{neg\text{-}sample}(v_c,o,U)}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)}\frac{e^{u_o^T v_c}}{(e^{u_o^T v_c}+1)^2}\cdot\frac{\partial(u_o^T v_c)}{\partial v_c} - \sum_{s=1}^{k}\left(\frac{1}{\sigma(-u_{w_s}^T v_c)}\cdot\frac{e^{-u_{w_s}^T v_c}}{(e^{-u_{w_s}^T v_c}+1)^2}\cdot\frac{\partial(-u_{w_s}^T v_c)}{\partial v_c}\right)$$

$$= -\frac{1}{e^{u_o^T v_c}+1}\cdot u_o + \sum_{s=1}^{k}\frac{1}{e^{-u_w^T v_c}+1}\cdot u_{w_s}$$

$$= -\left(1-\sigma(u_o^T v_c)\right)u_o + \sum_{s=1}^{k}\left(1-\sigma(-u_w^T v_c)\right)u_{w_s}$$

$$\frac{\partial J_{neg\text{-}sample}(v_c,o,U)}{\partial u_o} = -\frac{e^{u_o^T v_c}+1}{e^{u_o^T v_c}}\cdot\frac{e^{u_o^T v_c}}{(e^{u_o^T v_c}+1)^2}\cdot v_c = -\frac{1}{e^{u_o^T v_c}+1}\cdot v_c = -\left(1-\sigma(u_o^T v_c)\right)v_c.$$

$$\frac{\partial J_{neg\text{-}sample}(v_c,o,U)}{\partial u_{w_s}} = -\frac{e^{-u_{w_s}^T v_c}+1}{e^{-u_{w_s}^T v_c}}\cdot\frac{e^{-u_{w_s}^T v_c}}{(e^{-u_{w_s}^T v_c}+1)^2}\cdot(-v_c) = \frac{v_c}{e^{u_{w_s}^T v_c}+1} = \left(1-\sigma(-u_{w_s}^T v_c)\right)v_c$$

(ii) reuse: $\cancel{\frac{1}{e^{u_o^T v_c}+1}, \frac{1}{e^{-u_{w_s}^T v_c}+1}}$ for $s=[1,\cdots,k]$. $\sigma(u_o^T v_c), \sigma(-u_{w_s}^T v_c)$

first get $u_o, u_{w_1}, \cdots, u_{w_k}$ from $U\cdot M$ ($M_i=0$ if $i\neq o$ and $i \notin w_s$, else $M_i=1$)

(iii) less calculation in matrix ~~multiplication~~ $\longrightarrow$ sigmoid

(i) $J_{neg\text{-}sample}(v_c,o,U) = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{k}\log(\sigma(-u_{w_s}^T v_c))$, w_s might be identical

$$\frac{\partial J_{neg\text{-}sample}(v_c,o,U)}{\partial u_{w_s}} = -\sum_{\substack{i=1\\ w_i=w_s}}^{k}\frac{v_c}{e^{w_s^T v_c}+1} = \sum\left(1-\sigma(-u_{w_s}^T v_c)\right)v_c$$

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

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

(iii) $= 0$