(a) Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between $y$ and $\hat{y}$.

Show that $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

↳ Cross entropy loss between the true probability distribution $p$ and another distribution $q$ is $-\sum_i p_i \log(q_i)$

With given center word $c$, $y$ is true emprical distribution and $\hat{y}$ is the predicted distribution. (one hot vector with a 1 for the true outside word $o$, and 0 elsewhere)

$$-\sum_{w \in V} y_w \log(\hat{y}_w) = -\sum_{w \neq o, w \in V} 0 * \log(\hat{y}_w) - 1 * \log(\hat{y}_o) = -\log(\hat{y}_o)$$

(b) Compute the partial derirative of $J_{naive\text{-}softmax}(v_c, o, U)$ with respect to $v_c$. Please write your answer in terms of $y$, $\hat{y}$ and $U$.

$$\frac{\partial J(v_c, o, U)}{\partial v_c} = -\frac{\partial \log(P(O=o|C=c))}{\partial v_c}$$

$$= \frac{-\partial \log(\exp(u_o^T v_c))}{\partial v_c} + \frac{\partial \log\left(\sum_{w=1}^{V} \exp(u_w^T v_c)\right)}{\partial v_c}$$

$$= -u_o + \sum_{w=1}^{V} \frac{\exp(u_w^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \cdot u_w$$

$$= -u_o + \underbrace{\sum_{w=1}^{V} P(O=w|C=c)}_{\hat{y}} u_w$$

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

(c) Compute the partial derivatives of $J_{naive\text{-}softmax}(v_c, o, U)$ with respect to each of the 'outside' word vectors, $u_w$'s. There will be two cases: when $w=o$, the true 'outside' word vector, and $w \neq o$ for all other words.
Please write your answer in terms of $y$, $\hat{y}$, and $v_c$.

$$\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial u_w} = -\frac{\partial \log(\exp(u_o^T v_c))}{\partial u_w} + \frac{\partial \log\left(\sum_{w=1}^{V} \exp(u_w^T v_c)\right)}{\partial u_w}$$

Case 1: $w=o$, $\frac{\partial \log(\exp(u_o^T v_c))}{\partial u_o} + \frac{\partial \log\left(\sum_{w=1}^{V} \exp(u_w^T v_c)\right)}{\partial u_o}$

$$= -v_c + \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \cdot \frac{\partial \exp(u_o^T v_c)}{\partial u_o}$$

$$= -v_c + \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} v_c = \hat{y}_o v_c - v_c$$

$$= (P(O=o \mid C=c) - 1) v_c$$

Case 2: $w != o$

$$0 + \frac{\partial}{\partial u_w} \log \sum_{w \in V} \exp(u_w^T v_c)$$

$$= \frac{\exp(u_w^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \cdot v_c = \hat{y}_w v_c = P(O=w \mid C=c) v_c$$

$$\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial U} = (\hat{y} - y)^T v_c$$

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Korea

(d) The sigmoid function is given by Equation 4:

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

Please compute the derivative of $\sigma(x)$ with respect to $x$, where $x$ is a vector.

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

(e) consider Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that $K$ negative samples are drawn from the vocabulary refers to them as $w_1, w_2, \ldots w_K$, outside vectors as $u_1, \ldots u_K$, $o \notin \{w_1, \ldots w_K\}$

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

Repeat parts (b) and (c), computing partial derivative $J_{neg\text{-}sample}$ with respect to $v_c$, with respect to $u_o$, with respect to a negative sample $u_k$.

$$\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial v_c} = \frac{\partial\left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right)}{\partial v_c}$$

$$= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} \cdot \frac{\partial u_o^T v_c}{\partial v_c} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^T v_c))}{\partial v_c}$$

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

$$= -(1-\sigma(u_o^T v_c))u_o + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c))u_k$$

e. $\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial u_o} = \frac{-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))}{\partial u_o}$ 0

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

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

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

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

f.

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

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

iii) $\frac{\partial}{\partial v_w} J_{skip\text{-}gram}(v_c, w_{t-m}, \cdots w_{t+m}, U) = 0$