Q2. Word2Vec

a) Derive the gradients of sigmoid f^n.

$$\nabla \sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$\rightarrow u$
$\rightarrow v.$

using quotient rule.

$$= \frac{(1+e^{-x})(0) - e^{-x}(-1)}{(1+e^{-x})^2}$$

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

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

$$\boxed{\therefore \nabla \sigma(x) = (\sigma(x))^2 \cdot e^{-x}.}$$

b) $$\hat{y} = \frac{e^{u_o^T v_c}}{\sum_{w=1}^{W} e^{u_w^T v_c}}$$ } softmax function

$$J = -\sum_i y_i \log(\hat{y}_i)$$

$$\frac{\partial \hat{y_o}}{\partial v_c} = u_o \hat{y_o} - \sum_{w=1}^{W} u_w \hat{y_w} \quad \text{--- (1)}$$

$$\frac{\partial J}{\partial v_c} = -\sum_i \frac{y_i}{\hat{y_i}} \frac{\partial \hat{y_i}}{\partial v_c} \quad \text{--- (2)}$$

Substitute eq^n (1) in eq^n (2)

$$\frac{\partial J}{\partial v_c} = -\sum_i y_i \left(u_i - \sum_{w=1}^{W} u_w \hat{y}_w\right)$$

c) Derivative of softmax w.r.t u

$$\frac{\partial \hat{y}_o}{\partial u_w} = -v_c \hat{y}_o \quad \text{if } w = o$$

$$\frac{\partial \hat{y}_o}{\partial u_w} = v_c \hat{y}_w \quad \text{if } w \neq o$$

$$\frac{\partial \hat{y}_w}{\partial u_o} = -v_c \hat{y}_w \quad \text{if } w \neq o.$$

derivative of cross entropy:

$$\frac{\partial J}{\partial v_c} = -\sum_i \frac{y_i}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial v_c}$$

substitute softmax derivative in above eq^n:

$$\frac{\partial J}{\partial u_w} = -\sum_i y_i (v_c \hat{y}_i)$$

d) derivative of sampling loss

$$\frac{\partial J}{\partial v_c} = -\sigma(-u_o^T v_c) u_o + \sum_{K=1}^{K} \sigma(u_K^T v_c) u_K.$$

derivative w.r.t u_w:

$$\frac{\partial J}{\partial u_w} = \begin{cases} -\sigma(-u_o^T v_c) v_c & \text{if } w = o \\ \sigma(u_w^T v_c) & \text{if } w \in \{1 \dots k\} \end{cases}$$

e) Gradient w.r.t v_c.

$$\frac{\partial}{\partial v_c} \left(\sum_{-m \leq j < m, j \neq 0} f(w_{c+j}, v_c) \right) = \sum_{-m < j < m, j \neq 0} \frac{\partial f(w_{c+j}, v_c)}{\partial v_c}$$

from part d, we know that.

$$\frac{\partial J}{\partial v_c} = -\sigma(-u_o^T v_c) u_o + \sum_{K=1}^{K} \sigma(u_K^T v_c) u_K.$$

Gradient w.r.t u_K.

$$\frac{\partial}{\partial u_K} \left(\sum_{-m \leq j < m, J \neq 0} f(w_{c+j}, v_c) \right) = \sum_{-m < j < m, J \neq 0} \frac{\partial f(w_{c+J}, v_c)}{\partial u_K}$$

from previous parts.

$$\frac{\partial f(o, v_c)}{\partial u_w} = -\sum_i y_i (v_c \hat{y}_w) \quad \text{if } w = o.$$