Assignment: 2

a) Show that, naive-softmax loss is same as cross entropy loss between y and $\hat{y}$.

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

Given, true empirical distribution,

$$\left.\begin{matrix} w = o \\ w \neq o \end{matrix}\right\} \begin{matrix} y_w = 1 \\ y_w = 0 \end{matrix}$$

Cross entropy $= -\sum_{w \in V} y_w \log(\hat{y}_w)$

$$= -(y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + \cdots + y_o \log \hat{y}_o + \cdots + y_{|V|} \log \hat{y}_{|V|})$$

$$= -(0 + 0 + y_o \log \hat{y}_o + \cdots 0) = -1(\log \hat{y}_o)$$

$\because (w = o, y_w = 1)$

$$= \boxed{-\log(\hat{y}_o)} \Rightarrow \text{Proved}$$

b) → Derivative of $J_{naive\text{-}softmax}(v_c, o, U)$. $= ?$ $w \in V$

→ Write terms $y, \hat{y}, U$.

→ Present in vectorize form.

$$J(v_c, o, U) = -\log(\hat{y}_o)$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left[-\left(\log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w \in V} \exp(u_w^T \cdot v_c)}\right)\right]$$

// applying log property

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

$$= \frac{\partial}{\partial v_c}\left\{-u_o^T v_c + \log \sum_{w \in V} \exp(u_w^T v_c)\right\}$$ // log & exp are inverses

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

$$\because \frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$$

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

Since, $\frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$ is

the probability distribution over words. $(\hat{y}_w)$-

$$\Rightarrow \frac{\partial J}{\partial v_c} = -u_o^T + \sum_{w \in V} \hat{y}_w u_w$$

Since, $y_w = \begin{cases} 1 & \text{if } w = o \\ 0 & \text{if } w \neq o \end{cases}$, we can write

$$\boxed{u_o = y_w u_w}$$

$$\Rightarrow \frac{\delta J}{\delta v_c} = -\left(\sum_{w \in V} y_w u_w + \hat{y_w} u_w\right)$$

$$= \sum_{w \in V} u_w (-y_w + \hat{y_w})$$

$$= \sum_{w \in V} u_w (\hat{y_w} - y)$$

After vectorizing $\Rightarrow \frac{\delta J}{\delta v_c} = \boxed{U(\hat{y} - y)}$

(c) Compute derivative w.r.t u_w. Consider both cases when $w = o$ (outside word) and $w \neq o$ (all other words).

$$\frac{\delta J}{\delta u_w} = \frac{\delta}{\delta w_w}(-\log(y_o))$$

$$= \frac{\delta}{\delta w_w}\left(-\log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T \cdot v_c)}\right)$$

$$= \frac{\delta}{\delta u_w}\left[-\log(\exp(u_o^T v_c)) + \log \sum_{w \in V} \exp(u_w^T \cdot v_c)\right]$$

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

$$= -\frac{\partial u_o^T (v_c)}{\partial w_w} + \frac{\partial}{\partial u_w}\left[\log \sum_{w \in V} \exp(u_w^T \cdot v_c)\right].$$

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

$\Rightarrow$ (derivative of sum = Sum of derivatives)

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

Since, $\frac{\partial u_o^T}{\partial u_w} = \begin{cases} 1 & \text{if } w = o \\ 0 & \text{if } w \neq o \end{cases}$ and equivalent to y_w.

$$\Rightarrow \sum_{w \in V} -y_w v_c + \frac{1}{\sum_{w \in V} \exp(u_w^T \cdot v_c)} \sum_{w \in V} \exp(u_w^T \cdot v_c) \cdot v_c.$$

Since, $\frac{\exp(u_w^T \cdot v_c)}{\sum_{w \in V} \exp(u_w^T \cdot v_c)}$ is conditional probability, denoted as $\hat{y}_w$.

$$\Rightarrow \frac{\partial J}{\partial u_w} = \sum_{w \in V} \left(-y_w v_c + \hat{y}_w v_c\right)$$

$$= \boxed{\sum_{w \in V} v_c \left(\hat{y_w} - y_w\right)}$$

case1: $w = 0$. (context word is the outside word)

$$\frac{\partial J}{\partial u_{w=0}} = v_c \left(\hat{y}_{w=0} - y_{w=0}\right)$$

Since, $y_{w=0} = 1$

$$\Rightarrow \boxed{\frac{\partial J}{\partial u_{w=0}} = v_c \left(\hat{y}_{w=0} - 1\right)}$$

case2: $w \neq 0$ Since $y_{w \neq 0} = 0$

$$\Rightarrow \boxed{\frac{\partial J}{\partial u_{w \neq 0}} = v_c \hat{y}_{w \neq 0}}$$

$$\frac{\partial J}{\partial w_w} = \begin{cases} v_c \left(\hat{y_w} - 1\right) & \text{if } w = 0 \\ v_c \left(\hat{y}_w\right) & \text{otherwise} \end{cases}$$

(d) Compute partial derivative of $J_{naive\text{-}softmax}$ with respect to U.

Derivative of scalar y by a matrix A is given as

$$\frac{\partial y}{\partial A_{m\times n}} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y}{\partial A_{m1}} & \frac{\partial y}{\partial A_{m2}} & \cdots & \frac{\partial y}{\partial A_{mn}} \end{bmatrix}$$

Given, ww = Outside word w_o,

$\Rightarrow$ Derivative of y w.r.t 'U' is -

$$\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial U} = \begin{bmatrix} \frac{\partial J}{\partial u_1} & \frac{\partial J}{\partial u_2} & \cdots & \frac{\partial J}{\partial u_{|V|}} \end{bmatrix}$$

where,

$$\frac{\partial J}{\partial u_w} = \begin{cases} v_c(\hat{y} - 1) & \text{if } w = o \\ v_c \hat{y} & \text{otherwise} \end{cases}$$

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

$$\sigma = \frac{1}{1+e^{-x}} \quad \text{(given)}.$$

$$\frac{\partial \sigma}{\partial x} = \frac{\partial}{\partial x}\left[\frac{1}{1+e^{-x}}\right] = \frac{\partial}{\partial x}(1+e^{-x})^{-1}$$

$$= -1\,(1+e^{-x})^{-2}\,[-e^{-x}].$$

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

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

$$= \frac{1}{1+e^{-x}}\left[\left(\frac{e^{-x}+1}{1+e^{-x}}\right) - \frac{1}{1+e^{-x}}\right]$$

$$= \boxed{\sigma(x)\,[1-\sigma(x)]}$$

(b) Consider negative sampling loss: for a center word c, outside word o,

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

Compute derivatives of $J_{\text{neg-sample}}$ w.r.t v_c, u_o and u_k. Tell why this function is efficient.

① Derivative of $J_{\text{neg-sample}}$ w.r.t v_c:

$$\frac{\partial}{\partial v_c} = \frac{\partial}{\partial v_c}\left[-\log\left(\sigma\left(u_o^T v_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-u_o^k v_c\right)\right)\right]$$

$$= \frac{\partial}{\partial v_c}\left[-\log\left(\sigma\left(u_o^T v_c\right)\right) - \sum_{k=1}^{K} \frac{\partial}{\partial v_c} \log\left(\sigma\left(-u_o^k v_c\right)\right)\right]$$

$$= \frac{-1}{\sigma\left(u_o^T v_c\right)} \frac{\partial}{\partial v_c}\left(\sigma\left(u_o^T v_c\right)\right) - \sum_{k=1}^{K}\left[\frac{1}{\sigma\left(-u_o^k v_c\right)} \frac{\partial}{\partial v_c}\left(\sigma\left(-u_o^k v_c\right)\right)\right]$$

$$\boxed{\frac{\partial}{\partial x}\left(\sigma(x)\right) = \sigma(x)\left(1-\sigma(x)\right)} \quad \text{(Part e)}$$

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

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

$$\Rightarrow \boxed{u_o\left(1-\sigma\left(u_o^T v_c\right)\right) - \sum_{k=1}^{K} u_k\left(1-\sigma\left(-u_k^T v_c\right)\right)}$$

2) Derivative of $J_{neg\text{-}sample}$ w.r.t u_o:

$$\frac{\partial J}{\partial u_o} = \frac{\partial}{\partial u_o}\left[-\log\left(\sigma(u_o^T v_c)\right) - \sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)\right]$$

$\Rightarrow$ Since, $a \notin \{w_1, \dots, w_K\}$

$$\Rightarrow \frac{\partial}{\partial u_o}\left[\sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)\right] = 0$$

$$\Rightarrow \frac{\partial}{\partial u_o}\left(-\log\left(\sigma(u_o^T v_c)\right) - 0\right)$$

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

$$\Rightarrow -v_c\left(1 - \sigma(u_o^T v_c)\right)$$

$$\Rightarrow \boxed{v_c\left(\sigma(u_o^T v_c) - 1\right)}$$

3) Compute Derivative of $J_{neg\text{-}sample}$ w.r.t u_k:

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

Since, $o \notin \{w_1, w_2, \dots w_K\}$.

$$\Rightarrow \frac{\partial J}{\partial u_k}\left(-\log\left(\sigma(u_o^T v_c)\right) = 0\right.$$

$$\Rightarrow \frac{\partial J}{\partial u_k} = 0 - \sum_{k=1}^{K} \frac{\partial}{\partial u_k} \left(\log \left(\sigma(-u_k^T v_c) \right) \right)$$ (A)

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

$\Rightarrow$ all the derivative terms where $w \neq k$, becomes 0, while $\boxed{w = k}$, only the term u_w remains

$$\Rightarrow \frac{\partial J}{\partial u_k} = -v_c \left(1 - \sigma(-u_k^T v_c)\right)$$

$$= \boxed{v_c \left(\sigma(-u_k^T v_c) - 1\right)}$$

Negative sampling loss is much more memory and compute efficient since it requires only K sampled words. $O(K)$ - However, Naive-Softmax uses all the words inside Vocabulary $O(|V|)$ and normalize probabilities

(g) compute derivative of Jneg-sample w.r.t u_k. where K samples are drawn from Vocabulary

Since, K words are sampled from vocabulary which can't be distinct. Let's break sum into 2 sums

↳ ① Sum of all sampled words $w_i = w_k$

② Sum of all sampled words $w_i \neq w_k$

⇒ Consider Eq ① from last part

$$\frac{\partial J}{\partial u_k} = -\frac{\partial}{\partial u_k}\left[\sum_{k=1}^{K} \left(\sigma(-u_k^T v_c)\right)\right]$$

$$= \frac{-\partial}{\partial u_k}\left[\sum_{i\in\{1,\dots,K\}: w_i = w_k} \log\left(\sigma(-u_i^T v_c)\right) + \sum_{i\in\{1,\dots,K\}; w_i \neq w_k} \log\sigma(-u_i^T v_c)\right]$$

Now, $\frac{\partial}{\partial u_k} \log\left(\sigma(-u_i^T; w_i \neq w_k\ v_c)\right) = 0.$

$$\Rightarrow \frac{-\partial}{\partial u_k} \sum_{i\in\{1,\dots,K\}; w_i = w_k} \log\left(\sigma(-u_i^T v_c)\right)$$

$$\Rightarrow \frac{-\partial}{\partial u_k} \sum_{i\in\{1,\dots,K\}; w_i = w_k} \log\left(\sigma(-u_k^T v_c)\right) \Rightarrow -\sum_{i\in\{1,\dots K\}; w_i = w_k} \frac{\partial J}{\partial u_k}\left(\log(\sigma(-u_k^T w)\right.$$

$$\Rightarrow -\sum_{\substack{i\in\{1,\dots K\} \\ w_i = w_j}} \frac{1}{\log(\sigma(-u_k^T v_c))} \cdot \sigma(-u_k^T v_c) \cdot -v_c)\left(1-\sigma(-w_k^T v_c)\right)$$

$$\Rightarrow \sum_{k=1 ; w_i = w_u}^{K} \cdot v_c \left(1 - \sigma(-w_k^T v_c)\right)$$

$$\Rightarrow \boxed{\sum_{i \in \{1,\dots K\} ; w_i = w_u} -v_c \left(\sigma(-u_k^T v_c) - 1\right)} \quad \frac{\partial \bar{v}}{\partial u_u}$$

when samples are drawn from vocabulary.

(h)

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

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

- $$\frac{\partial J_{\text{skip-gram}}(v_c, w_{t+m'}, \dots, w_{t+m})}{\partial v_w ; w \neq c} = 0$$