a) $-\sum_{w \in vocab} y_w \log(\hat{y}_w) = -y_{w'} \log(\hat{y}_{w'}) = -\log(\hat{y}_{w'})$

* We know that all of $y_w | w \in Vocab$ are zeros except one of them. Thus, we can ommit all words with $y_w = 0$ and keep w' which $y_{w'} = 1$

$$-\log(\hat{y}_{w'}) = -\log(\hat{y}_o) = -\log(P(O=o|C=c)) = J_{naive softmax}$$

* We know that $\hat{y}_o$ is scalar. Thus it can be written instead of $\hat{y}_{w'}$.

b) (i)

$$\frac{\partial J_{naive-softmax}}{\partial v_c} = \frac{-\partial \log(P(O=o|C=c))}{\partial v_c}$$

$$= \frac{-\partial \log\left(\frac{\exp(u_o^T \cdot v_c)}{\sum \exp(u_w^T \cdot v_c)}\right)}{\partial v_c} = -\frac{\partial \log(\exp(u_o^T \cdot v_c))}{\partial v_c}$$

$$+ \frac{\partial \log(\sum \exp(u_w^T \cdot v_c))}{\partial v_c} = -\frac{1}{\exp(u_o^T \cdot v_c)} \cdot \frac{\partial \exp(u_o^T \cdot v_c)}{\partial v_c}$$

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

$$+ \frac{\sum u_w \exp(u_w^T \cdot v_c)}{\sum \exp(u_w^T \cdot v_c)} = -u_o + \sum_{w \in Vocab} u_w \left(\frac{\exp(u_w^T \cdot v_c)}{\sum_{w' \in Vocab} \exp(u_{w'}^T \cdot v_c)}\right)$$

$$= -u_o + \sum_{w \in Vocab} u_w P(O=w|C=c) = -u_o + U \cdot \hat{y}$$

$u_o \in R^{d \times 1}$, $U \in R^{d \times Vocabsize}$, $\hat{y} \in R^{Vocab\_size \times 1}$

Subject:
Date

(ii) $-u_o + U.\hat{y} = 0 \Rightarrow \underset{\text{outside context word}}{u_o} = \underbrace{U.\hat{y}}_{\text{predicted context word}}$

* It means that, gradient will be zero, if the predicted word is equal to outside context word.

(iii)

* The gradient that we found is the error of our model in predicting the correct outside context word. Thus, by subtracting this error from v_c we are moving the vector which is representing v_c to a new position in d-dimential space in order to force our model to predict words with minimum difference with the actual outside context word.

(iv)

L_2 normalization is forcing vectors to have equal length.

This normalization takes away useful information when some features of a specific word is devidable by the same feature of another word.

For example:

$$royalty_{King} = \alpha . royalty_{prince}$$

L_2 normalization makes these two words look equal in the term of royalty.

c) $w=o$:

$$\frac{\partial J_{naive\_softmax}}{\partial u_w} = \frac{-\partial \log\left(\frac{\exp(u_o^T \cdot v_c)}{\Sigma \exp(u_x^T \cdot v_c)}\right)}{\partial u_w}$$

$$= -\frac{\partial \log(\exp(u_o^T \cdot v_c))}{\partial u_w} + \frac{\partial \log(\Sigma \exp(u_x^T \cdot v_c))}{\partial u_w}$$

$$= -\frac{1}{\exp(u_o^T \cdot v_c)} \cdot \frac{\exp(u_o^T \cdot v_c)}{\partial u_w} + \frac{1}{\sum_{x \in Vocab} \exp(u_x^T \cdot v_c)} \cdot \frac{\Sigma \exp(u_x^T \cdot v_c)}{\partial u_w}$$

$$= -v_c + \frac{\exp(u_w^T \cdot v_c)}{\sum_{x \in Vocab} \exp(u_x^T \cdot v_c)} \cdot v_c^T = -v_c + \hat{y}_w \cdot v_c^T$$

$v_c \in R^{d \times 1}$ $\hat{y}_w = R^{1 \times 1}$

$w \neq o$:

$$\frac{\partial J_{naive\_softmax}}{\partial u_w} = \frac{-\partial \log\left(\frac{\exp(u_o^T \cdot v_c)}{\Sigma \exp(u_x^T \cdot v_c)}\right)}{\partial u_w}$$

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

$$= 0 + \frac{1}{\sum_{x \in Vocab} \exp(u_x^T \cdot v_c)} \cdot \frac{\partial \Sigma \exp(u_x^T \cdot v_c)}{\partial u_w} = \frac{\exp(u_w^T \cdot v_c)}{\sum_{x \in Vocab} \exp(u_x^T \cdot v_c)} \cdot v_c^T$$

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

$v_c \in R^{d \times 1}$ $\hat{y}_w \in R^{1 \times 1}$

d)
$$\frac{\partial J_{naive-softmax}}{\partial u_w} \in R^{d \times 1}$$

$$\rightarrow \quad \frac{\partial J_{naive-softmax}}{\partial U} = \left[\frac{\partial J_{naive-softmax}}{\partial u_1}, \frac{\partial J}{\partial u_2}, \dots, \frac{\partial J}{\partial u_{vocab\_len}} \right]$$

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

* This function is not differentiable at $x=0$.

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

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

Scanned with CamScanner

Subject:
Date

g.) (i.)

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

$$= -\frac{1}{\sigma(u_o^T \cdot v_c)} \cdot \frac{\partial \sigma(u_o^T \cdot v_c)}{\partial v_c} - \sum_{s=1}^{K} \frac{1}{\sigma(-u_{w_s}^T \cdot v_c)} \cdot \frac{\partial \sigma(-u_{w_s}^T \cdot v_s)}{\partial v_c}$$

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

$$- \sum_{s=1}^{K} \frac{\sigma(-u_{w_s}^T \cdot v_c)(1-\sigma(-u_{w_s}^T \cdot v_c))}{\sigma(-u_{w_s}^T \cdot v_c)} \cdot \frac{\partial(-u_{w_s}^T \cdot v_c)}{\partial v_c}$$

$$= -u_o(1-\sigma(u_o^T \cdot v_c)) + \sum_{s=0}^{K} u_{w_s}(1-\sigma(u_{w_s}^T \cdot v_c))$$

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

$$= -\frac{\sigma(u_o^T \cdot v_c)(1-\sigma(u_o^T \cdot v_c))}{\sigma(u_o^T \cdot v_c)} \cdot \frac{\partial(u_o^T \cdot v_c)}{\partial u_o} - 0$$

$$= -(1-\sigma(u_o^T \cdot v_c))\, v_c^T$$

$$\frac{\partial J_{neg\text{-}sample}}{\partial u_{w_x}} = -\frac{\partial \log(\sigma(u_o^T \cdot v_c))}{\partial u_{w_x}} - \frac{\partial \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T \cdot v_c))}{\partial u_{w_x}}$$

$$= 0 - \frac{\sigma(-u_{w_x}^T \cdot v_c)(1-\sigma(-u_{w_x}^T \cdot v_c))}{\sigma(-u_{w_x}^T \cdot v_c)} \cdot \frac{\partial(-u_{w_x}^T \cdot v_c)}{\partial u_{w_x}}$$

$$= (1-\sigma(-u_{w_x}^T \cdot v_c)) \cdot v_c^T$$

Subject :
Date

(ii)

$$U^T_{o,\{w_1,\dots,w_k\}} \cdot v_c = \begin{bmatrix} u_o^T \cdot v_c \\ -u_{w_1}^T \cdot v_c \\ \vdots \\ -u_{w_k}^T \cdot v_c \end{bmatrix} \Rightarrow 1 - \sigma(U^T_{o,\{w_1,\dots,w_k\}} \cdot v_c) = \begin{bmatrix} 1-\sigma(u_o^T \cdot v_c) \\ 1-\sigma(u_{w_1}^T \cdot v_c) \\ \vdots \\ 1-\sigma(u_{w_k}^T \cdot v_c) \end{bmatrix}$$

(iii) We don't need to iterate over the vocabulary.

h) All the terms in $\frac{\partial J_{neg\text{-}sample}}{\partial u_{w_x}}$ will be zero, excep the ones with $u_{w_x} = u_{w_s}$

$$\frac{\partial J_{neg\text{-}sample}}{\partial u_{w_x}} = - \frac{\partial \sum_{\substack{1 \le s \le k \\ w_x = w_s}} \log(\sigma(-u_{w_s}^T \cdot v_c))}{\partial u_{w_x}}$$

$$= - \sum_{\substack{1 \le s \le k \\ w_x = w_s}} \frac{\sigma(-u_{w_s}^T \cdot v_c)(1-\sigma(-u_{w_s}^T \cdot v_c))}{\partial u_{w_x}} \cdot \frac{\partial(-u_{w_s}^T \cdot v_c)}{\partial u_{w_x}}$$

$$= \sum_{\substack{1 \le s \le k \\ w_x = w_s}} (1 - \sigma(u_{w_s}^T \cdot v_c))\, v_c^T$$

Subject:
Date

i)

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

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

(iii) $$\frac{\partial J_{skip\text{-}gram}}{\partial u_w} \text{ when } w \neq c = 0$$