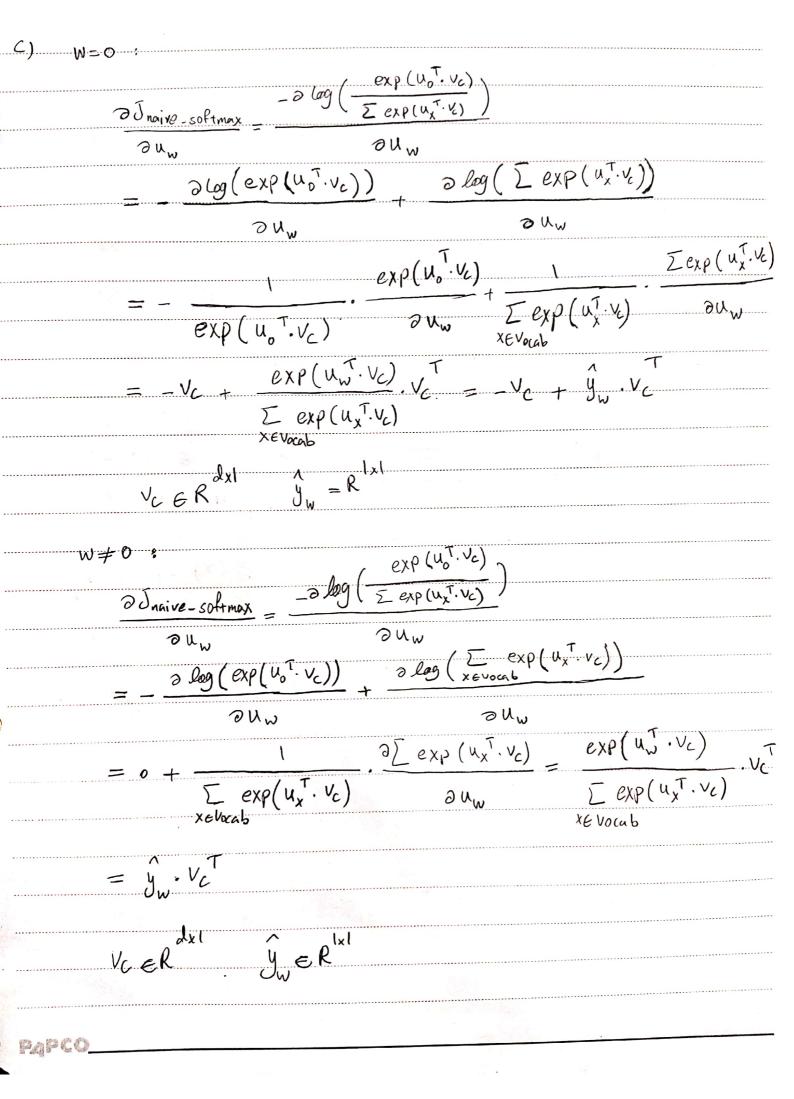
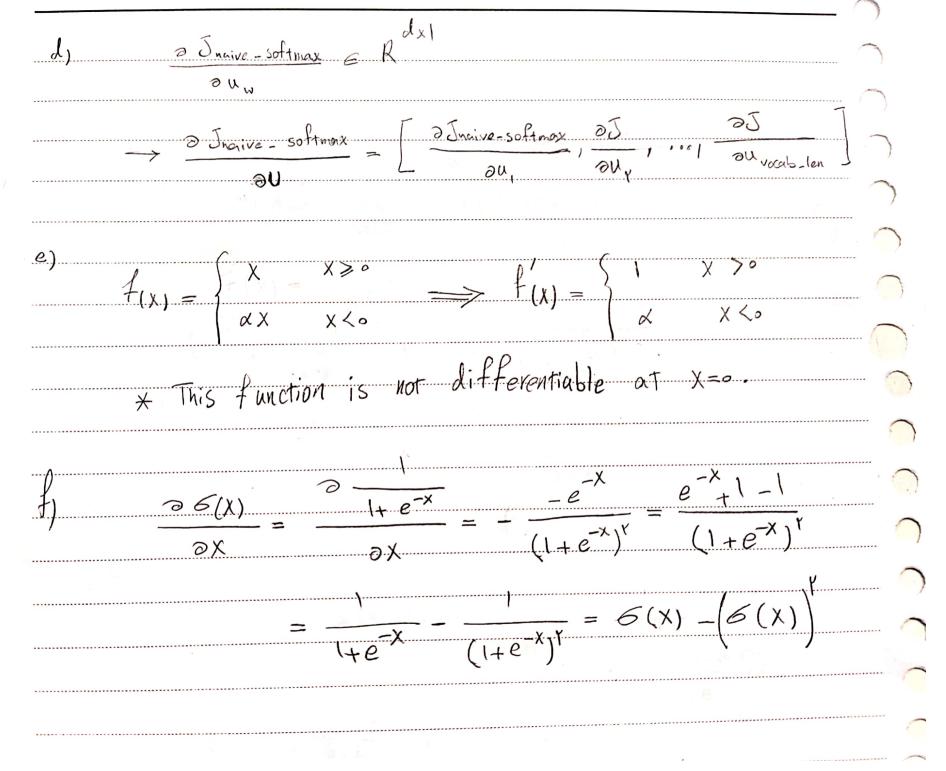
a) $-\sum_{w \in Vocab} y_w \log(y_w) = -y_w \log(y_w) = -\log(y_w)$ * We know that all of $J_w \mid w \in Vocab$ are zeros except one of them. Thus, we can ommit all words with $J_w = 0$ and keep w' which $J_w' = 1$ $-\log\left(\hat{y}_{w'}\right) = -\log\left(\hat{y}_{o}\right) = -\log(P(0=o|C=c)) = \tilde{J}_{naive soft}$ * We know that ŷ is scalar. Thus it can be written instead of ŷw. b) (i) = 5 noive - softmax = - alog(P(O=0|(-c)) $= \frac{\partial V_{c}}{\partial v_{c}} \left(\frac{\exp(u_{o}^{T} \cdot v_{c})}{\sum \exp(u_{w}^{T} \cdot v_{c})} \right) = \frac{\partial \log(\exp(u_{o}^{T} \cdot v_{c}))}{\partial v_{c}}$ $= \frac{\partial \log(\sum \exp(u_{w}^{T} \cdot v_{c}))}{\sum \exp(u_{w}^{T} \cdot v_{c})} = \frac{\partial \log(\exp(u_{o}^{T} \cdot v_{c}))}{\partial v_{c}}$ $= \frac{\partial \log(\sum \exp(u_{w}^{T} \cdot v_{c}))}{\partial v_{c}} = \frac{\exp(u_{o}^{T} \cdot v_{c})}{\partial v_{c}} = \frac{\exp(u_{o}^{T} \cdot v_{c})}{\exp(u_{o}^{T} \cdot v_{c})} \cdot u_{o}$ $= \frac{\partial \log(\sum \exp(u_{w}^{T} \cdot v_{c}))}{\partial v_{c}} = \frac{\exp(u_{o}^{T} \cdot v_{c})}{\exp(u_{o}^{T} \cdot v_{c})} \cdot u_{o}$ $+ \frac{\sum u_w \exp(u_w^T \cdot v_c)}{\sum \exp(u_w^T \cdot v_c)} = u_o + \sum u_w \left(\frac{\exp(u_w^T \cdot v_c)}{\sum \exp(u_w^T \cdot v_c)} \right)$ $= \frac{\sum u_w \exp(u_w^T \cdot v_c)}{\sum \exp(u_w^T \cdot v_c)}$ = $-u_0 + \sum u_W P(D=W|C=c) = -u_0 + U_1 \hat{y}$ $w \in Vocab$ $u_0 \in \mathbb{R}^{d \times 1}$, $u_0 \in \mathbb{R}$ $u_0 \in \mathbb{R}^{d \times 1}$, $u_0 \in \mathbb{R}$

Scanned with CamScanner

(ii)	$u_0 + U_0 \hat{y} = 0 \Rightarrow u_0 = U_0 \hat{y}$
	out side context word > predicted context word
71 =6	* It means that, gradient will be zero, if the predicted word is equal to a outside context word.
(11)	
	* The gradient that we found is the error of
heist	our model in predicting the correct ourside
	context word. Thus, by subtracting this error
	from ve we are moving the vector which is
	representing ve to a new position in d-dimential
	space in order to force our model to predict
	words with minimum difference with the actual
<u> </u>	out side context word.
(iv)	Lz 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 = a. royalty prince
	Le normalization makes these two words look equal
ani	in the term of royalty.
Papco_	

Date





$$\begin{array}{c} g)(i) \\ = J_{\text{neg-sample}} \\ = J_{\text{neg-sample}} \\ = J_{\text{ov}} \\$$

$$(ii) \qquad \qquad U_{\sigma,\{w_1,\dots,w_k\}}^{\mathsf{T}} \mathcal{N}_{c} = \begin{bmatrix} u_{\sigma}^{\mathsf{T}} \mathcal{N}_{c} \\ -u_{\omega_{1}}^{\mathsf{T}} \mathcal{N}_{c} \\ -u_{\omega_{k}}^{\mathsf{T}} \mathcal{N}_{c} \end{bmatrix} \Rightarrow 1 - 5\left(U_{\sigma,\{w_{1},\dots,w_{k}\}}^{\mathsf{T}} \mathcal{N}_{c}\right) = \begin{bmatrix} 1 - 5\left(u_{\sigma}^{\mathsf{T}} \mathcal{N}_{c}\right) \\ 1 - 6\left(u_{\omega_{k}}^{\mathsf{T}} \mathcal{N}_{c}\right) \end{bmatrix}$$

$$= \sum_{1 \le S \le K} \left(1 - 6 \left(u_{w_S}^T \cdot v_L \right) \right) v_L^T$$

$$w_S = w_S$$

Subject:	
Date	

$$\frac{(ii)}{\partial V_{c}} = \frac{\partial J(V_{c}, W_{t+j}, V)}{\partial V_{c}}$$

(iii)
$$\frac{\partial \int_{Skip-gton}}{\partial U_{w}}$$
 when $w \neq c = 0$