Let's learn these derivatives

We have $x_i = U_o^T v_c$

and $y = \text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ we want $\frac{\partial S}{\partial x_i}$

and $S_j = -\log(\hat{y}_j)$

$V \times 1$

$$\frac{\partial S}{\partial x_i} = \frac{\partial S}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial x_i} \qquad \frac{\partial S}{\partial \hat{y}_i} = -\frac{1}{\hat{y}_i}$$

$$\frac{\partial y}{\partial x_i} = \hat{y}_i(\delta_{ij} - \hat{y}_j)$$

$$= \left(-\frac{1}{\hat{y}_i}\right)(\hat{y}_i(\delta_{ij} - \hat{y}_j))$$

$$= \hat{y} - y$$

$N \times 1$ $N \times V$

Now we want $\frac{\partial S}{\partial U_o}$ and $\frac{\partial S}{\partial v_c}$

$$\boxed{\frac{\partial S}{\partial U_o} = \frac{\partial S}{\partial x_i} \frac{\partial x_i}{\partial U_o} = \overset{1 - y_o}{(\hat{y}_o - y_o) v_c}} \qquad \boxed{\frac{\partial S}{\partial U} = (\hat{y} - y) v_c^T}$$

2b
2c
2d

$$\boxed{\underset{N\times 1}{\frac{\partial S}{\partial v_c}} = \sum_{w=1} \underset{V\times 1}{(\hat{y} - y)} \underset{N\times 1}{u_w} = \underset{N\times V}{U^T} \underset{V\times 1}{(\hat{y} - y)}}$$

word vectors are rows in U

when $w \neq o$,

$$\frac{\partial y}{\partial x_o} = \frac{(0)(\Sigma) - (e^{x_o})(e^{x_w})}{\Sigma^2} = -y_o y_w$$

but $\frac{\partial x_o}{\partial u_w} = 0$?

2c) $\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{(0)(1+e^{-x}) - (1)(-e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$

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

2f) $J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum \log(\sigma(-u_k^T v_c))$

$x = u_o^T v_c \quad y_k = -u_k^T v_c$

$z_1 = \sigma(x) \quad z_{2k} = \sigma(y_k)$

$J = -\log z_1 - \sum_k \log(z_{2k})$

$$\frac{\partial J}{v_c} = \underline{\frac{\partial J}{\partial z_1}\frac{\partial z_1}{\partial x}\frac{\partial x}{\partial v_c}} + \sum_k \frac{\partial J}{\partial z_{2k}}\frac{\partial z_{2k}}{\partial y_k}\frac{\partial y_k}{\partial v_c}$$

$\frac{\partial J}{\partial z_1} = -\frac{1}{z_1}$ | $\frac{\partial z_1}{\partial x} = \sigma(x)(1-\sigma(x))$ | $\frac{\partial x}{\partial v_c} = u_o$

$$\frac{\partial J}{\partial z_1}\frac{\partial z_1}{\partial x}\frac{\partial x}{\partial v_c} = \left(-\frac{1}{z_1}\right)(\sigma(x)(1-\sigma(x)))(u_o) = \cancel{-\frac{1}{\sigma(x)}(\sigma(x)(1-\sigma(x)))(u_o)}$$

$$= (\underline{\sigma(u_o^T v_c) - 1})(u_o)$$

$\frac{\partial J}{\partial z_{2k}} = \left(-\frac{1}{z_{2k}}\right)$ | $\frac{\partial z_{2k}}{\partial y_k} = (\sigma(y_k)(1-\sigma(y_k)))$ | $\frac{\partial y_k}{\partial v_c} = -u_k$

$\downarrow$

$$\sum_k (1 - \sigma(\underline{-u_k^T v_c}))\, u_k$$

$$\boxed{\frac{\partial J}{\partial v_c} = (\sigma(u_o^T v_c) - 1)(u_o) + \sum_k (1-\sigma(-u_k^T v_c))\, u_k}$$

2. $$\frac{\partial J}{\partial u_o} = \frac{\partial J}{\partial z_1}\frac{\partial z_1}{\partial x}\frac{\partial x}{\partial u_o} = \left(-\frac{1}{z_1}\right)(z_1(1-z_1))(v_c)$$

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

$$\frac{\partial J}{\partial u_k} = \cancel{\sum_k} \frac{\partial J}{\partial z_{2k}}\frac{\partial z_{2k}}{\partial y_k}\frac{\partial y_k}{\partial u_k} = \cancel{\sum_k}\left(-\frac{1}{z_k}\right)(z_k(1-z_k))(-v_c)$$

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

2g) $J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum \log(\sigma(-u_k^T v_c))$

$\frac{\partial J}{\partial u_k} = N(1-\sigma(-u_k^T v_c))(v_c)$ (N = number of reps)

2h) (i) $\frac{\partial J_{skip\text{-}gram}}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$

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

(iii) $\frac{\partial J_{skip\text{-}gram}}{\partial v_w} = 0$