1 (a) Prove :-

softmax (x) = softmax $(x+c)$

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$\text{softmax}(x+c)_i = \frac{e^{(x+c)_i}}{\sum_j e^{(x+c)_j}} = \frac{e^c e^{x_i}}{e^c \sum_j e^{x_j}}$$

$$= \text{softmax}(x)_i$$

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

$$\frac{d\sigma}{dx} = -\frac{1}{(1+e^{-x})^2} \times -e^{-x}$$

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

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

$$= \sigma(1-\sigma)$$

b) Derive $\frac{\partial CE}{\partial \theta}$ where $CE(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i$

& $\hat{y}_i$ = softmax (θ_i)

Firstly lets derive

$$\frac{\partial \hat{y}}{\partial \theta}$$

$$\hat{y}_k = \frac{e^{\theta_k}}{\sum_j e^{\theta_j}}$$

when $i = k$

$$\frac{\partial \hat{y}_k}{\partial \theta_i} = \frac{e^{\theta_k}}{\sum_j e^{\theta_j}} + e^{\theta_k}\left(-\frac{1}{\left(\sum e^{\theta_j}\right)^2}\right) \cdot e^{\theta_k}$$

$$= \hat{y}_k - \frac{e^{\theta_k}}{\sum_j e^{\theta_j}} \cdot \frac{e^{\theta_k}}{\sum e^{\theta_j}}$$

$$= \hat{y}_k(1 - \hat{y}_k)$$

when $i \neq k$

$$\frac{\partial \hat{y}_k}{\partial \theta_i} = e^{\theta_k} \times \frac{-1}{\left(\sum e^{\theta_j}\right)^2} \times e^{\theta_i}$$

$$= -\frac{e^{\theta_k}}{\sum_j e^{\theta_j}} \cdot \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$$

$$= -\hat{y}_k \cdot \hat{y}_i$$

$$\therefore \frac{\partial \hat{y}_k}{\partial \theta_i} = \hat{y}_k(1 - \hat{y}_k) \quad \text{when } i = k$$

$$= -\hat{y}_k \cdot \hat{y}_i \quad \text{otherwise}$$

Now.

$$\frac{\partial CE}{\partial \theta_i} = -\frac{\partial}{\partial \theta_i}\left(\sum_j y_j \log \hat{y}_j\right)$$

Since, y is a one-hot vector with all but k^{th} dimension, 0.

$$\frac{\partial CE}{\partial \theta_i} = -\frac{\partial}{\partial \theta_i}\left(y_k \log \hat{y}_k\right)$$

$$= -\frac{y_k}{\hat{y}_k}\frac{\partial}{\partial \theta_i}(\hat{y}_k)$$

$$= -\frac{1}{\hat{y}_k}\frac{\partial}{\partial \theta_i}(\hat{y}_k) \qquad (\because y_k = 1)$$

when $i = k$.

$$\frac{\partial CE}{\partial \theta_i} = -\frac{1}{\hat{y}_k}\hat{y}_k(1-\hat{y}_k) = \hat{y}_k - 1$$

when $i \neq k$

$$\frac{\partial CE}{\partial \theta_i} = -\frac{1}{\hat{y}_k} \times -\hat{y}_k \hat{y}_i = \hat{y}_i$$

$$\boxed{\begin{aligned} \frac{\partial CE}{\partial \theta_i} &= \hat{y}_k - 1 && \text{when } i = k \\ &\text{or} \\ &\hat{y}_k - y_k \\ &= \hat{y}_i && \text{otherwise} \end{aligned}}$$

Simplifying above we get

$$\frac{\partial CE}{\partial \theta_i} = \hat{y}_i - y_i$$

(c) Given $h = \text{sigmoid}(xW_1 + b_1)$
$\hat{y} = \text{softmax}(hW_2 + b_2)$

Let $z_1 = xW_1 + b_1$, so that
$h = \text{sigmoid}(z_1)$

& $z_2 = hW_2 + b_2$
so that $\hat{y} = \text{softmax}(z_2)$

$$\frac{\partial J}{\partial x} = \left(\left(\frac{\partial J}{\partial z_2} * \frac{\partial z_2}{\partial h}\right) \circ \frac{\partial h}{\partial z_1}\right) * \frac{\partial z_1}{\partial x}$$

- $\circ$ element-wise multiplication
- $*$ dot-product

From 2(b) $\frac{\partial J}{\partial z_2} = (\hat{y} - y)$

$$\frac{\partial z_2}{\partial h} = \frac{\partial}{\partial h}(\vec{h} \cdot \vec{w}_2 + b_2) = \hat{W}_2^T$$

$$\frac{\partial h}{\partial z_1} = \sigma(z_1)(1 - \sigma(z_1)) = \sigma'$$

$$\frac{\partial z_1}{\partial x} = W_1^T$$

$$\boxed{\therefore \frac{\partial J}{\partial x} = \left(\left((\hat{y} - y) * W_2^T\right) \circ \sigma'\right) * W_1^T}$$

(d) No. of parameters

x: $D_x \times 1$

y: $D_y \times 1$

~~W_1 must be $D_x \times H$ dimensional~~

~~W_2 must be $H \times D_y$ dimensional~~

Adding in the bias unit, hidden ~~No. of inputs to~~ layer receives inputs from $(D_x + 1)$ unit, hence

Dimensions of $W_1 \Rightarrow (D_x+1) \times H$

~~Input~~ Output layer receives inputs from $(H+1)$ units, hence dimensions of $W_2 = (H+1) \times D_y$

$$\boxed{\text{Total no. of parameters} = (D_x+1) \times H + (H+1) \times D_y}$$

word2vec

3) (a) Given predicted word vector v_c corresponding to center word c,

$$\hat{y} = P(O|C) = \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}$$

u_w are "output" word vectors

& u_o is the expected word

To find $\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(\sum y_i \log \hat{y}\right)$

assuming one-hot encoding for y

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log P(O|C)\right)$$

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

$$= -\frac{\partial}{\partial v_c}\left(u_o^T v_c\right) + \left(\frac{1}{\sum e^{u_w^T v_c}}\right)\frac{\partial}{\partial v_c}\left(\sum_x e^{u_x^T v_c}\right)$$

$$= u_o + \frac{1}{\sum e^{u_w^T v_c}}\left(\sum e^{u_x^T v_c} \cdot u_x\right)$$

$$= u_o + \sum_x \left(\frac{e^{u_x^T v_c}}{\sum e^{u_w^T v_c}}\right) \cdot u_x$$

$$= u_o + \sum_x P(x|c) \cdot u_x$$

(b) T

(b) To find $\frac{\partial J}{\partial U_w}$

when $U_w = U_0$

$$\frac{\partial J}{\partial U_0} = -\frac{\partial}{\partial U_0}\left(\log\left(\frac{e^{U_0^T V_c}}{\sum e^{U_w^T V_c}}\right)\right)$$

$$= -\frac{\partial (U_0^T V_c)}{\partial U_0} + \frac{1}{\sum e^{U_w^T V_c}} \frac{\partial}{\partial U_0}\left(\sum_x e^{U_x^T V_c}\right)$$

$$= -V_c + \frac{1}{\sum e^{U_w^T V_c}} e^{U_0^T V_c} \cdot V_c$$

$$= -V_c + P(O|c) \cdot V_c$$

when $U_w \neq U_0$

$$\frac{\partial J}{\partial U_w} = -\frac{\partial}{\partial U_w}\left(\log\left(\frac{e^{U_0^T V_c}}{\sum e^{U_w^T V_c}}\right)\right)$$

$$= -\frac{\partial}{\partial U_w}(U_0^T V_c) + \frac{1}{\sum e^{U_w^T V_c}} \frac{\partial}{\partial U_w}\left(\sum_x e^{U_x^T V_c}\right)$$

$$= 0 + \frac{e^{U_w^T V_c}}{\sum e^{U_w^T V_c}} \cdot V_c$$

$$= P(W|c) \cdot V_c$$

hence

$$\frac{\partial J}{\partial U_w} = V_c\left(P(O|c) - 1\right) \quad \text{when } w = o$$

$$= V_c P(W|c) \quad w \neq o$$

(d) Find $\frac{\partial J}{\partial v_c}$ & $\frac{\partial J}{\partial u_w}$ when negative sampling is applied

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

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

$$= -\frac{1}{\sigma_o} \sigma_o \times (1-\sigma_o) \frac{\partial}{\partial v_c}(u_o^T v_c) - \sum_k \frac{1}{\sigma_{-k}} \times \sigma_{-k}(1-\sigma_{-k}) \frac{\partial}{\partial v_c}(-u_k^T v_c)$$

$$= (\sigma_o - 1) u_o + \sum_k (1-\sigma_{-k}) u_k$$

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

For finding $\frac{\partial J}{\partial u_w}$, consider 3 cases :-

1) when $w = o$ 2) $w = 1 \to k$ 3) $w \neq o$ & $w \neq 1 \to k$

Case I) when $w = o$

$$\frac{\partial J}{\partial u_o} = -\frac{1}{\sigma_o} \times \sigma_o(1-\sigma_o) v_c - 0$$

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

Case II) when $w = 1 \to k$

$$\frac{\partial J}{\partial u_w} = -\frac{1}{\sigma_o} \sigma_o (1-\sigma_o) \cdot 0 - \sum_{k=1}^{K} \frac{1}{\sigma_{-k}} \times \sigma_{-k}(1-\sigma_{-k}) \frac{\partial}{\partial u_w}(-u_k^T v_c)$$

$$= 0 + (1 - \sigma(-u_w^T v_c)) v_c$$

Case III) when $w \neq o$ & $w \neq 1 \to k$

$$\frac{\partial J}{\partial u_w} = 0$$