1. Softmax

(a) we want to prove : $\text{softmax}(X+c) = \text{softmax}(X)$

we know that: $\text{Softmax}(X)_i = \dfrac{e^{x_i}}{\sum_{j=1}^{Dim(X)} e^{x_j}}$

so: $\text{Softmax}(X+c)_i = \dfrac{e^{x_i+c}}{\sum_{j=1}^{Dim(X)} e^{x_j+c}} = \dfrac{e^{x_i} \times e^{c}}{\sum_{j=1}^{Dim(X)} e^{x_j} \times \underbrace{e^{c}}_{\text{constant}}} = \dfrac{e^{x_i} \times \cancel{e^{c}}}{\cancel{e^{c}} \sum_{j=1}^{Dim(X)} e^{x_j}}$

$\Rightarrow \text{softmax}(X+c) = \text{softmax}(X)$

---

2. Neural Network Basics:

(a) we want to derive the gradients of the sigmoid function and show that it can be rewritten as a function of the function value :

$$\sigma(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1} \xrightarrow[\text{using chain rule}]{\text{take derivative}} \sigma'(x) = -(1+e^{-x})^{-2} \times -e^{-x}$$

$$\Rightarrow \sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \underbrace{\frac{e^{-x}}{1+e^{-x}}}_{1-\sigma(x)} \times \underbrace{\frac{1}{1+e^{-x}}}_{\sigma(x)} = \sigma(x)(1-\sigma(x))$$

(b) we want to derive the gradient w.r.t the inputs of a softmax function when cross entropy loss is used for evaluation :

$$\begin{cases} \hat{y} = \text{Softmax}(\theta) \\ CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i) \end{cases}$$

$\dfrac{\partial CE}{\partial \theta} = -\sum_i y_i \dfrac{\partial}{\partial \theta} \log(\hat{y}_i) \Rightarrow$ we know that y is a one-hot label vector. we assume that only the K-th dimension of y is one , so:

$\dfrac{\partial CE}{\partial \theta} = -\cancel{y_K}^{\,1} \dfrac{\partial}{\partial \theta} \log(\hat{y}_K) = -\dfrac{\partial}{\partial \theta} \log\left(\dfrac{e^{\theta_K}}{\sum_j e^{\theta_j}}\right) \Rightarrow$ θ is a vector. so we expand this derivative

Gradient w.r.t θ_i $\Rightarrow$ $\frac{\partial}{\partial \theta_i}\left[\log \frac{e^{\theta_K}}{\sum_j e^{\theta_j}}\right] = \frac{\partial}{\partial \theta_i}\left[\log e^{\theta_K} - \log \sum_j e^{\theta_j}\right] = \frac{\partial}{\partial \theta_i}\left[\theta_K - \log \sum_j e^{\theta_j}\right]$

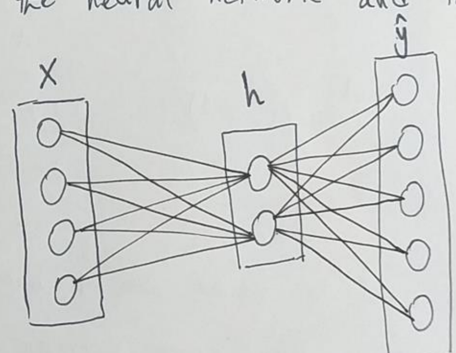
Now there is two options $\Rightarrow$

$$\begin{cases} \text{if } i=K \Rightarrow \frac{\partial}{\partial \theta_K}\left[\theta_K - \log \sum_j e^{\theta_j}\right] = 1 - \frac{e^{\theta_K}}{\sum_j e^{\theta_j}} = 1-\hat{y}_K = \underbrace{1}_{y_i \text{ when } i=K}-\hat{y}_i = y_i - \hat{y}_i \\ \text{if } i\neq K \Rightarrow \frac{\partial}{\partial \theta_i}\left[\theta_K - \log \sum_j e^{\theta_j}\right] = 0 - \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} = \underbrace{0}_{y_i \text{ when } i\neq K} - \hat{y}_i = y_i - \hat{y}_i \end{cases}$$

so anyway we can rewritten the gradient like: $\frac{\partial}{\partial \theta}\left[\log \frac{e^{\theta_K}}{\sum_j e^{\theta_j}}\right] = y - \hat{y}$

finally $\Rightarrow$ $\frac{\partial CE}{\partial \theta} = -\frac{\partial}{\partial \theta}\left[\log \frac{e^{\theta_K}}{\sum_j e^{\theta_j}}\right] = \hat{y} - y$

(c) we want to move backward through the neural network and take all derivatives step by step.



let's assume $\begin{cases} z^2 = hw^2 + b^2 \\ z^1 = xw^1 + b^1 \end{cases}$

$h = \text{sigmoid}(xw^1 + b^1)$

$\hat{y} = \text{softmax}(hw^2 + b^2)$

from part (b) we know that:

$$\frac{\partial CE}{\partial z^2} = (\hat{y} - y) = dz^2$$

$$\underbrace{\frac{\partial CE}{\partial w^2}}_{H\times D_y} = \frac{\partial z^2}{\partial w^2} \times \frac{\partial CE}{\partial z^2} = \underbrace{\frac{\partial z^2}{\partial w^2}}_{\text{it should be } H\times 1} \times \underbrace{(\hat{y} - y)}_{1\times D_y}$$

let's compute this term

③

According to the neural network architecture we can conclude

$$\begin{cases} h_1 = \dfrac{\partial z_1^2}{\partial w_{11}^2} = \dfrac{\partial z_2^2}{\partial w_{12}^2} = \dfrac{\partial z_3^2}{\partial w_{13}^2} = \cdots = \dfrac{\partial z_{D_y}^2}{\partial w_{1D_y}^2} \\ h_2 = \dfrac{\partial z_1^2}{\partial w_{21}^2} = \dfrac{\partial z_2^2}{\partial w_{22}^2} = \dfrac{\partial z_3^2}{\partial w_{23}^2} = \cdots = \dfrac{\partial z_{D_y}^2}{\partial w_{2D_y}^2} \\ \vdots \\ h_H = \dfrac{\partial z_1^2}{\partial w_{H1}^2} = \dfrac{\partial z_2^2}{\partial w_{H2}^2} = \dfrac{\partial z_3^2}{\partial w_{H3}^2} = \cdots = \dfrac{\partial z_{D_y}^2}{\partial w_{HD_y}^2} \end{cases}$$

$$\frac{\partial CE}{\partial w^2} = \begin{bmatrix} \frac{\partial CE}{\partial w_{11}^2} & \frac{\partial CE}{\partial w_{12}^2} & \cdots & \frac{\partial CE}{\partial w_{1D_y}^2} \\ \frac{\partial CE}{\partial w_{21}^2} & \frac{\partial CE}{\partial w_{22}^2} & \cdots & \frac{\partial CE}{\partial w_{2D_y}^2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial CE}{\partial w_{H1}^2} & \frac{\partial CE}{\partial w_{H2}^2} & \cdots & \frac{\partial CE}{\partial w_{HD_y}^2} \end{bmatrix}_{H\times D_y} = \begin{bmatrix} \frac{\partial z_1^2}{\partial w_{11}^2}\times\frac{\partial CE}{\partial z_1^2} & \frac{\partial z_2^2}{\partial w_{12}^2}\times\frac{\partial CE}{\partial z_2^2} & \cdots & \frac{\partial z_{D_y}^2}{\partial w_{1D_y}^2}\times\frac{\partial CE}{\partial z_{D_y}^2} \\ \frac{\partial z_1^2}{\partial w_{21}^2}\times\frac{\partial CE}{\partial z_1^2} & \frac{\partial z_2^2}{\partial w_{22}^2}\times\frac{\partial CE}{\partial z_2^2} & \cdots & \frac{\partial z_{D_y}^2}{\partial w_{2D_y}^2}\times\frac{\partial CE}{\partial z_{D_y}^2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial z_1^2}{\partial w_{H1}^2}\times\frac{\partial CE}{\partial z_1^2} & \frac{\partial z_2^2}{\partial w_{H2}^2}\times\frac{\partial CE}{\partial z_2^2} & \cdots & \frac{\partial z_{D_y}^2}{\partial w_{HD_y}^2}\times\frac{\partial CE}{\partial z_{D_y}^2} \end{bmatrix}$$

$$\Rightarrow \frac{\partial CE}{\partial w^2} = \begin{bmatrix} h_1\times(\hat{y}_1-y_1) & h_1(\hat{y}_2-y_2) & \cdots & h_1(\hat{y}_{D_y}-y_{D_y}) \\ h_2\times(\hat{y}_1-y_1) & h_2\times(\hat{y}_2-y_2) & \cdots & h_2\times(\hat{y}_{D_y}-y_{D_y}) \\ \vdots & \vdots & & \vdots \\ h_H\times(\hat{y}_1-y_1) & h_H\times(\hat{y}_2-y_2) & \cdots & h_H\times(\hat{y}_{D_y}-y_{D_y}) \end{bmatrix}_{H\times D_y} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_H \end{bmatrix}_{H\times 1} \begin{bmatrix} \hat{y}_1-y_1 & \hat{y}_2-y_2 & \cdots & \hat{y}_{D_y}-y_{D_y} \end{bmatrix}_{1\times D_y}$$

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_H \end{bmatrix}_{H\times 1} = h^T = \frac{\partial z^2}{\partial w^2}$$

Finally

$$\Rightarrow \frac{\partial CE}{\partial w^2} = \frac{\partial z^2}{\partial w^2}\times(\hat{y}-y) = h^T(\hat{y}-y) = h^T dz^2 = dw^2$$

Gradient w.r.t b^2

$$\Rightarrow \frac{\partial CE}{\partial b^2} = \begin{bmatrix} \frac{\partial CE}{\partial b_1^2} & \frac{\partial CE}{\partial b_2^2} & \cdots & \frac{\partial CE}{\partial b_{D_y}^2} \end{bmatrix}$$

$$\frac{\partial CE}{\partial b^2} = \begin{bmatrix} \underbrace{\frac{\partial z_1^2}{\partial b_1^2}}_{1}\times\frac{\partial CE}{\partial z_1^2} & \underbrace{\frac{\partial z_2^2}{\partial b_2^2}}_{1}\times\frac{\partial CE}{\partial z_2^2} & \cdots & \underbrace{\frac{\partial z_{D_y}^2}{\partial b_{D_y}^2}}_{1}\times\frac{\partial CE}{\partial z_{D_y}^2} \end{bmatrix}$$

$$\frac{\partial CE}{\partial b^2} = \begin{bmatrix} \frac{\partial CE}{\partial z_1^2} & \frac{\partial CE}{\partial z_2^2} & \cdots & \frac{\partial CE}{\partial z_{D_y}^2} \end{bmatrix} = \frac{\partial CE}{\partial z^2} = (\hat{y}-y) = db^2$$

Gradient w.r.t h $\Rightarrow$ $\underbrace{\frac{\partial CE}{\partial h}}_{Dim=1\times H} = \underbrace{\frac{\partial CE}{\partial Z^2}}_{1\times D_y} \times \underbrace{\frac{\partial Z^2}{\partial h}}_{\text{should be}: D_y\times H} = dZ^2 \times \frac{\partial Z^2}{\partial h}$ $\rightarrow$ let's compute this term

we know that $\rightarrow$
$$\begin{cases} Z_1^2 = h_1 w_{11}^2 + h_2 w_{21}^2 + \cdots + h_H w_{H1}^2 + b_1^2 \\ Z_2^2 = h_1 w_{12}^2 + h_2 w_{22}^2 + \cdots + h_H w_{H2}^2 + b_2^2 \\ \vdots \\ Z_{D_y}^2 = h_1 w_{1D_y}^2 + h_2 w_{2D_y}^2 + \cdots + h_H w_{HD_y}^2 + b_{D_y}^2 \end{cases} \Rightarrow w_{ij}^2 = \frac{\partial Z_j^2}{\partial h_i}$$

so:
$$\frac{\partial Z^2}{\partial h} = \underbrace{\begin{bmatrix} \frac{\partial Z_1^2}{\partial h_1} & \frac{\partial Z_1^2}{\partial h_2} & \cdots & \frac{\partial Z_1^2}{\partial h_H} \\ \frac{\partial Z_2^2}{\partial h_1} & \frac{\partial Z_2^2}{\partial h_2} & \cdots & \frac{\partial Z_2^2}{\partial h_H} \\ \vdots & \vdots & & \vdots \\ \frac{\partial Z_{D_y}^2}{\partial h_1} & \frac{\partial Z_{D_y}^2}{\partial h_2} & \cdots & \frac{\partial Z_{D_y}^2}{\partial h_H} \end{bmatrix}}_{D_y\times H} = \underbrace{\begin{bmatrix} w_{11}^2 & w_{21}^2 & \cdots & w_{H1}^2 \\ w_{12}^2 & w_{22}^2 & \cdots & w_{H2}^2 \\ \vdots & & & \\ w_{1D_y}^2 & w_{2D_y}^2 & \cdots & w_{HD_y}^2 \end{bmatrix}}_{D_y\times H}$$

$$\Rightarrow \frac{\partial Z^2}{\partial h} = w^{2T}$$

finally: $\frac{\partial CE}{\partial h} = \left(\frac{\partial CE}{\partial Z^2}\right)\times\left(\frac{\partial Z^2}{\partial h}\right) = (\hat{y}-y)w^{2T} = dh$

Gradient w.r.t Z^1 $\Rightarrow$ $\frac{\partial CE}{\partial Z^1} = \frac{\partial CE}{\partial h} \times \frac{\partial h}{\partial Z^1} = dh \times \frac{\partial h}{\partial Z^1}$ $\rightarrow$ let's compute this term

we know that $\Rightarrow$ $h = \underbrace{\text{Sigmoid}(Z^1)}_{\sigma(Z^1)} \Rightarrow \frac{\partial h}{\partial Z^1} = \sigma'(Z^1)$

$\Rightarrow$ So: $\underbrace{\frac{\partial CE}{\partial Z^1}}_{Dim=1\times H} = \frac{\partial CE}{\partial h} \times \frac{\partial h}{\partial Z^1} = \underbrace{(\hat{y}-y)w^{2T}}_{1\times H} \circ \underbrace{\sigma'(Z^1)}_{1\times H} = dZ^1$

Gradient w.r.t W^1 $\Rightarrow$ $\underbrace{\frac{\partial CE}{\partial W^1}}_{Dim = D_x \times H} = \underbrace{\frac{\partial Z^1}{\partial W^1}}_{\text{Should be: } D_x \times 1} \times \underbrace{\frac{\partial CE}{\partial Z^1}}_{1 \times H} = \frac{\partial Z^1}{\partial W^1} \times dZ^1$ let's compute this term

According to the neural network, we can conclude

$$\begin{cases} x_1 = \frac{\partial Z_1^1}{\partial W_{11}^1} = \frac{\partial Z_2^1}{\partial W_{12}^1} = \cdots = \frac{\partial Z_H^1}{\partial W_{1H}^1} \\ x_2 = \frac{\partial Z_1^1}{\partial W_{21}^1} = \frac{\partial Z_2^1}{\partial W_{22}^1} = \cdots = \frac{\partial Z_H^1}{\partial W_{2H}^1} \\ \vdots \\ x_{D_x} = \frac{\partial Z_1^1}{\partial W_{D_x1}^1} = \frac{\partial Z_2^1}{\partial W_{D_x2}^1} = \cdots = \frac{\partial Z_H^1}{\partial W_{D_xH}^1} \end{cases}$$

$$\frac{\partial CE}{\partial W^1} = \begin{bmatrix} \frac{\partial CE}{\partial W_{11}^1} & \frac{\partial CE}{\partial W_{12}^1} & \cdots & \frac{\partial CE}{\partial W_{1H}^1} \\ \frac{\partial CE}{\partial W_{21}^1} & \frac{\partial CE}{\partial W_{22}^1} & \cdots & \frac{\partial CE}{\partial W_{2H}^1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial CE}{\partial W_{D_x1}^1} & \frac{\partial CE}{\partial W_{D_x2}^1} & \cdots & \frac{\partial CE}{\partial W_{D_xH}^1} \end{bmatrix}_{D_x H} = \begin{bmatrix} \frac{\partial Z_1^1}{\partial W_{11}^1} \times \frac{\partial CE}{\partial Z_1^1} & \frac{\partial Z_2^1}{\partial W_{12}^1} \times \frac{\partial CE}{\partial Z_2^1} & \cdots & \frac{\partial Z_H^1}{\partial W_{1H}^1} \times \frac{\partial CE}{\partial Z_H^1} \\ \frac{\partial Z_1^1}{\partial W_{21}^1} \times \frac{\partial CE}{\partial Z_1^1} & \frac{\partial Z_2^1}{\partial W_{22}^1} \times \frac{\partial CE}{\partial Z_2^1} & \cdots & \frac{\partial Z_H^1}{\partial W_{2H}^1} \times \frac{\partial CE}{\partial Z_H^1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial Z_1^1}{\partial W_{D_x1}^1} \times \frac{\partial CE}{\partial Z_1^1} & \frac{\partial Z_2^1}{\partial W_{D_x2}^1} \times \frac{\partial CE}{\partial Z_2^1} & \cdots & \frac{\partial Z_H^1}{\partial W_{D_xH}^1} \times \frac{\partial CE}{\partial Z_H^1} \end{bmatrix}$$

$$\Rightarrow \frac{\partial CE}{\partial W^1} = \begin{bmatrix} x_1 \frac{\partial CE}{\partial Z_1^1} & x_1 \frac{\partial CE}{\partial Z_2^1} & \cdots & x_1 \frac{\partial CE}{\partial Z_H^1} \\ x_2 \frac{\partial CE}{\partial Z_1^1} & x_2 \frac{\partial CE}{\partial Z_2^1} & \cdots & x_2 \frac{\partial CE}{\partial Z_H^1} \\ \vdots & & & \\ x_{D_x} \frac{\partial CE}{\partial Z_1^1} & x_{D_x} \frac{\partial CE}{\partial Z_2^1} & \cdots & x_{D_x} \frac{\partial CE}{\partial Z_H^1} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{D_x} \end{bmatrix}_{D_x \times 1} \begin{bmatrix} \frac{\partial CE}{\partial Z_1^1} & \frac{\partial CE}{\partial Z_2^1} & \cdots & \frac{\partial CE}{\partial Z_H^1} \end{bmatrix}_{1 \times H}$$

$$= X^T = \frac{\partial Z^1}{\partial W^1}$$

$$\Rightarrow \frac{\partial CE}{\partial W^1} = X^T \frac{\partial CE}{\partial Z^1} = X^T (\hat{y} - y) W^{2T} \circ \sigma'(Z^1) = dW^1$$

Gradient w.r.t b^1 $\rightarrow$ $\frac{\partial CE}{\partial b^1} = \left[\frac{\partial CE}{\partial b^1_1} \quad \frac{\partial CE}{\partial b^1_2} \quad \cdots \quad \frac{\partial CE}{\partial b^1_H} \right]$

$$\frac{\partial CE}{\partial b^1} = \left[\cancelto{1}{\frac{\partial z^1_1}{\partial b^1_1}} \times \frac{\partial CE}{\partial z^1_1} \quad \cancelto{1}{\frac{\partial z^1_2}{\partial b^1_2}} \times \frac{\partial CE}{\partial z^1_2} \quad \cdots \quad \cancelto{1}{\frac{\partial z^1_H}{\partial b^1_H}} \times \frac{\partial CE}{\partial z^1_H} \right]$$

$$\frac{\partial CE}{\partial b^1} = \left[\frac{\partial CE}{\partial z^1_1} \quad \frac{\partial CE}{\partial z^1_2} \quad \cdots \quad \frac{\partial CE}{\partial z^1_H} \right] = \frac{\partial CE}{\partial z^1} = (\hat{y} - y) w^{2T} \circ \sigma'(z^1) = db^1$$

And finally we can compute $\frac{\partial CE}{\partial X} = \frac{\partial CE}{\partial z^1} \times \frac{\partial z^1}{\partial X}$ $\rightarrow$ similar to computation of $\frac{\partial z^2}{\partial h}$, it is equal to: w^{1T}

$$\Rightarrow \quad \frac{\partial CE}{\partial X} = (\hat{y} - y) w^{2T} \circ \sigma'(z^1) \, w^{1T}$$

(d) How many parameters are there in this neural network?? let's suppose:

- D_x : input vector dimension
- D_y : output vector dimension
- H : number of hidden units

$$\# \text{ parameters} = \# \text{ Parameter in } w^2, b^2, w^1, b^1 =$$

$$= H \times D_y + 1 \times D_y + D_x \times H + 1 \times H$$

$$= (H+1) D_y + (D_x + 1) H$$

---

3. Word2Vec :

(a) with following assumptions, we want to derive the gradients of cost function with respect to V_c :

- V_c : predicted word vector corresponding to the center word c for skipgram
- u_0 : output word vector corresponding to the expected word 0

(b) Derive the gradient w.r.t all output vectors:

$$\frac{\partial J}{\partial u_i} = \frac{\partial}{\partial u_i}\left[-u_0^T v_c + \log\left[\sum_{w=1}^{W} \exp(u_w^T v_c)\right]\right]$$

(1) if $i \neq 0 \Rightarrow$ $$\frac{\partial J}{\partial u_i} = \left[-(0) + \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \times v_c \exp(u_i^T v_c)\right]$$

$$\Rightarrow \frac{\partial J}{\partial u_i} = \left[-(0) + v_c \times \underbrace{\frac{\exp(u_i^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)}}\right] = [-\overset{v_c y_i}{\overset{\shortparallel}{(0)}} + v_c \hat{y}_i]$$

$$\Rightarrow \frac{\partial J}{\partial u_i} = v_c(\hat{y}_i - y_i)$$

(2) if $i = 0 \Rightarrow$ $$\frac{\partial J}{\partial u_0} = [-\overset{v_c y_0}{\overset{\shortparallel}{v_c}} + v_c \hat{y}_0] = v_c(\hat{y}_0 - y_0) = v_c(\hat{y}_i - y_i)$$

$\Rightarrow$ so $\Rightarrow$ $\frac{\partial J}{\partial u_w} = v_c(\hat{y}_w - y_w)$ $\Rightarrow$ or more generally $\Rightarrow$ $\frac{\partial J}{\partial U} = v_c(\hat{y} - y)^T$

any word in vocabulary

$\text{Dim} = N \times W = (N \times 1) \times (1 \times W)$

(c) Repeat part (a) and (b) assuming we are using the negative sampling loss function (Note that $0 \notin \{1, 2, \ldots K\}$):

$$J_{negative\text{-}sample}(0, v_c, U) = -\log(\sigma(u_0^T v_c)) - \sum_{K=1}^{K} \log(\sigma(-u_K^T v_c))$$

$$\frac{\partial J}{\partial v_c} = -\frac{\partial}{\partial v_c} \log(\sigma(u_0^T v_c)) - \sum_{K=1}^{K} \frac{\partial}{\partial v_c} \log(\sigma(-u_K^T v_c))$$

$$\frac{\partial J}{\partial v_c} = -\frac{\cancel{\sigma(u_0^T v_c)}(1 - \sigma(u_0^T v_c))}{\cancel{\sigma(u_0^T v_c)}} u_0 - \sum_{K=1}^{K} \frac{-\cancel{\sigma(-u_K^T v_c)}(1 - \sigma(-u_K^T v_c))}{\cancel{\sigma(-u_K^T v_c)}} u_K$$

$$\frac{\partial J}{\partial v_c} = (\sigma(u_0^T v_c) - 1) u_0 - \sum_{K=1}^{K} (\sigma(-u_K^T v_c) - 1) u_K$$

9

Gradient w.r.t u_0 $\Rightarrow$ $\frac{\partial J}{\partial u_0} = -\frac{\partial}{\partial u_0} \log\left(\sigma(u_0^T v_c)\right) = -\frac{\cancel{\sigma(u_0^T v_c)}\left(1-\sigma(u_0^T v_c)\right)}{\cancel{\sigma(u_0^T v_c)}} \times v_c$

$$\frac{\partial J}{\partial u_0} = v_c\left(\sigma(u_0^T v_c) - 1\right)$$

Gradient w.r.t u_K $\Rightarrow$ $\frac{\partial J}{\partial u_K} = -\frac{\partial}{\partial u_K} \sum_{K>1}^{K} \log\left(\sigma(-u_K^T v_c)\right)$

$$\frac{\partial J}{\partial u_K} = -\cancel{\sum_{K=1}^{K}} \frac{\cancel{\sigma(-u_K^T v_c)}\left(1-\sigma(-u_K^T v_c)\right)}{\cancel{\sigma(-u_K^T v_c)}} \times -v_c$$

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

(d) for skip-gram, the cost for a context centered around c is:

$$J_{\text{skip-gram}}(\text{word}_{c-m \dots c+m}) = \sum_{-m \le j \le m,\, j \ne 0} F(\omega_{c+j}, v_c)$$

Then the gradients for the cost of one context window are:

$$\begin{cases} \dfrac{\partial J_{\text{skip-gram}}(\text{word}_{c-m \dots c+m})}{\partial U} = \displaystyle\sum_{-m \le j \le m,\, j \ne 0} \dfrac{\partial F(\omega_{c+j}, v_c)}{\partial U} \\[2ex] \dfrac{\partial J_{\text{skip-gram}}(\text{word}_{c-m \dots c+m})}{\partial v_c} = \displaystyle\sum_{-m \le j \le m,\, j \ne 0} \dfrac{\partial F(\omega_{c+j}, v_c)}{\partial v_c} \\[2ex] \dfrac{\partial J_{\text{skipgram}}(\text{word}_{c-m \dots c+m})}{\partial v_j} = 0 \quad \text{for all the other } j \end{cases}$$

for CBOW instead of using v_c as the predicted vector, we use $\hat{v}$ (10)

defind below:

$$\hat{v}_c = \sum_{-m \leq j \leq m, j \neq 0} v_{c+j}$$

then the CBOW cost is:

$$J_{CBOW}(word_{c-m \dots c+m}) = F(w_c, \hat{v}_c)$$

for CBOW then we have:

$$\begin{cases} \dfrac{\partial J_{CBOW}(word_{c-m \dots c+m})}{\partial U} = \dfrac{\partial F(w_c, \hat{v}_c)}{\partial U} \\[2ex] \dfrac{\partial J_{CBOW}(word_{c-m \dots c+m})}{\partial v_j} = \dfrac{\partial F(w_c, \hat{v}_c)}{\partial \hat{v}_c} \quad \text{for all } j \in \{c-m \dots c+m\} \\[2ex] \dfrac{\partial J_{CBOW}(word_{c-m \dots c+m})}{\partial v_j} = 0 \quad \text{for all } j \notin \{c-m \dots c+m\} \end{cases}$$