Assignment 2 - written

1(a) Given

$$-\sum_{w \in Vocab} y_w \; log(\hat{y}_w) = -log(\hat{y}_o)$$

Show that $J_{naive-softmax}(v_c, o, U) = -logP(O = o|C = c)$

$y_i =$ one hot vector for true outside word, “o”. $\therefore y_i = 1$ if $i == o$, else $y_i = 0$

$$-\sum_{w \in Vocab} y_w \; log(\hat{y}_w) = -[y_1 \; log(\hat{y}_1) + \cdots \; y_o \; log(\hat{y}_o) + \cdots \; y_w \; log(\hat{y}_w)]$$

$$= -y_o \; log(\hat{y}_o) \; \texttt{substitute} \; y_o = 1$$

$$= -log(\hat{y}_o) \in \mathbb{R}^{1 \times 1} \; \texttt{(a scalar)}$$

1(b) Find

$\dfrac{\partial}{\partial v_c} J_{naive-softmax}(v_c, o, U)$ `in terms of` $y, \hat{y},$ `and` U

Given

$\hat{y} = P(O = o|C = c)$, $J = CE(y, \hat{y})$, $\hat{y} = softmax(\theta) \in \mathbb{R}^n$

$\hat{y}$ is the probability of an outside word given a center word. Softmax converts θ scores to probabilities.

$\dfrac{\partial J}{\partial \theta} = \hat{y} - y$ `(Eq. 1) See identity #7 [1]`

$U =$ matrix of columns for outside vectors $U \in \mathbb{R}^{n \times |V|}$

$|V| =$ vocabulary size

$V =$ input matrix $V \in \mathbb{R}^{n \times |V|}$

$n =$ length of embedding

$v_c =$ column vector for center word $\in \mathbb{R}^n$

$\theta = Uv_c =$ `score vector` $\in \mathbb{R}^{n \times 1}$ `(Eq. 2) , and` $\dfrac{\partial \theta}{\partial v_c} = U$ `(Eq. 3) See identity #2 [1]`

$$\frac{\partial J}{\partial \theta} \frac{\partial \theta}{\partial v_c} = (\hat{y} - y)U \in \mathbb{R}^n$$

1(c) Find

$\frac{\partial}{\partial U_w} J_{naive-softmax}(v_c, o, U)$ in terms of $y, \hat{y},$ and v_c

Given

$o =$ outside word

$\theta = Uv_c$ where $u_w \in w = o$ (Eq. 4)

$\theta = Uv_c$ where $u_w \in w \neq o$ (Eq. 5)

$\frac{\partial J}{\partial W_{ij}} = \delta^\top x^\top$ Where $\delta = \frac{\partial J}{\partial z}$ and $z = Wx$ (Eq. 6)

See identity #6 [1] (matrix times column vector with respect to the matrix)

Substitute parameters: $x = v_c$, $W = U$, and $z = \theta = Uv_c$

$$\frac{\partial J}{\partial U} = \left(\frac{\partial J}{\partial \theta}\right)^\top \left(\frac{\partial \theta}{\partial U}\right)^\top = (\hat{y} - y)^\top v_c^\top \in \mathbb{R}^{|V| \times n} \text{ same dimension as } U^\top$$

is $\hat{y}$ a row vector of length V?

This result appears to calculate gradients for both $w = o$ and $w \neq o$. My thinking is that the resulting matrix would show sparse (zero) entries where $w \neq o$

1(d)

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

Solve $\frac{\partial \sigma}{\partial x}$ in terms of $\sigma(x)$

Quotient Rule: $\frac{d}{dxf} = \frac{(denominator * \frac{d}{dx}numerator) - (numerator * \frac{d}{dx}denominator)}{denominator^2}$

$$\frac{d}{dx}numerator = \frac{d}{dx}1 = 0$$

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

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

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

substitute $\sigma(x) \Rightarrow \sigma(x)(1 - \sigma(x))$ (Eq. 7)

1(e)

Given $J_{neg-sample}(v_c, o, U) = -log(\sigma(u_o^\top v_c)) - \sum_{k=1}^{k} log(\sigma(-u_k^\top v_c))$

Find $\frac{\partial J}{\partial v_c}, \frac{\partial J}{\partial u_o}, \frac{\partial J}{\partial u_k}$

$log(x) = log_e(x) = ln(x) \Rightarrow \frac{d}{dx} log(x) = \frac{1}{x}$ (Eq. 8)

Chain Rule: $\frac{\partial}{\partial v_c} \underbrace{-log}_{f}(\underbrace{\sigma(\underbrace{u_o^\top v_c}_{g_2})}_{g_1}) - \frac{\partial}{\partial v_c} \sum_{k=1}^{k} \underbrace{log}_{f}(\underbrace{\sigma(\underbrace{-u_k^\top v_c}_{g_2})}_{g_1})$ Use Eq. 7 & Eq. 8

$$f'g_1' =$$

$$-\frac{1}{\sigma(u_o^\top v_c)} \cdot \sigma(u_o^\top v_c)(1 - \sigma(u_o^\top v_c)) - \sum_{k=1}^{k} \frac{1}{\sigma(-u_k^\top v_c)} \cdot \sigma(-u_k^\top v_c)(1 - \sigma(-u_k^\top v_c))$$

$$\frac{\partial J}{\partial v_c} = f'g_1'g_2' = (\sigma(u_o^\top) - 1)u_o - \sum_{k=1}^{k}(1 - \sigma(-u_k^T v_c))(-u_k)$$

$$\frac{\partial J}{\partial v_c} = (\sigma(u_o^\top v_c) - 1)u_o + \sum_{k=1}^{k}(1 - \sigma(-u_k^\top v_c))u_k$$

Chain Rule: $\frac{\partial}{\partial u_o} \underbrace{-log}_{f}(\underbrace{\sigma(\underbrace{u_o^\top v_c}_{g_2})}_{g_1}) - \frac{\partial}{\partial u_o} \underbrace{\sum_{k=1}^{k} log(\sigma(-u_k^\top v_c))}_{zero: u_k \neq u_o}$

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

Chain Rule: $\frac{\partial}{\partial u_k} \underbrace{-log(\sigma(u_o^\top v_c))}_{zero: u_k \neq u_o} - \frac{\partial}{\partial u_k} \sum_{k=1}^{k} \underbrace{log}_{f}(\underbrace{\sigma(\underbrace{-u_k^\top v_c}_{g_2})}_{g_1})$

$$\frac{\partial J}{\partial u_k} = \sum_{k=1}^{k} -\frac{1}{\sigma(-u_k^\top v_c)} \cdot \sigma(-u_k^\top v_c)(1 - \sigma(-u_k^\top v_c))(-v_c)$$

$$\frac{\partial J}{\partial u_k} = \sum_{k=1}^{k}(1 - \sigma(-u_k^\top v_c))v_c$$

Neg-sample should be more efficient than naive-softmax because neg-sample calculates zero for a lot of terms where $u_k \neq u_o$

1(f)

Given: center word, $c = w_t$

Context window $= [w_{t-m} \cdots w_{t+m}]$

$m =$ context window size

$J_{skipgram}(v_c, w_{t-m} \ldots w_{t+m}, U) = \sum_{-m \leq j \leq m, j \neq 0} J(v_c, w_{t+j}, U)$

`Find:` $\partial J_{skipgram}(v_c, w_{t-m} \cdots w_{t+m}, U)$ `for` $\frac{\partial J}{\partial U}, \frac{\partial J}{\partial v_c}, \frac{\partial J}{\partial v_w}$ `where` $w \neq c$

`Use` $_{neg-sample}(v_c, w_{t+j}, U)$

$$\frac{\partial J}{\partial U} = \sum_{-m \leq j \leq m, j \neq 0} \frac{J(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J}{\partial v_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{J(v_c, w_{t+j}, U)}{\partial v_c}$$

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

Only center word vectors v_c contribute to loss calculations.

[1] gradient_notes.pdf http://web.stanford.edu/class/cs224n/readings/gradient-notes.pdf