$$P(O=o|C=c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \quad (1)$$

$$J_{naive\text{-}softmax}(v_c, o, U) = -\log P(O=o|C=c) \quad (2)$$

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

sol: $\because y_w = 1$ iff $w=o$ for $\vec{y}$, else $y_w = 0$

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

(b) Compute $\frac{\partial J_{naive\text{-}softmax}}{\partial v_c}$

sol: $\vec{y} = [0, \dots, 1, \dots, 0]^T$ where $y_o = 1$, $y_{w \neq o} = 0$ $U_{|V| \times n}$

$\vec{\hat{y}} = [P(O=w_1|C=c), \dots, P(O=w_{|V|}|C=c)]^T$ $|V| \times 1$

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}\right) = \frac{\partial}{\partial v_c}\left(-u_o^T v_c + \log\left(\sum_w \exp(u_w^T v_c)\right)\right)$$

$$= -u_o + \frac{1}{\sum_j \exp(u_j^T v_c)} \sum_w \exp(u_w^T v_c)\, u_w$$

$$= -u_o + \sum_w \frac{\exp(u_w^T v_c)}{\sum_j \exp(u_j^T v_c)} u_w = -u_o + \sum_w \hat{y}_w u_w$$

$$= -\sum_w y_w u_w + \sum_w \hat{y}_w u_w = -U^T \vec{y} + U^T \vec{\hat{y}}$$

$$= U^T(\vec{\hat{y}} - \vec{y}) = -U^T(\vec{y} - \vec{\hat{y}})$$

* update stops if $|\vec{y} - \vec{\hat{y}}| \approx 0$

#

(c) Compute $\frac{\partial J_{naive\text{-}softmax}}{\partial u_w}$ for $w=0$ and $w\neq 0$.

Sol: $J = -u_o^T v_c + \log\left(\sum_w \exp(u_w^T v_c)\right)$

$$\frac{\partial J}{\partial u_o} = -v_c + \frac{1}{\sum_w \exp(u_w^T v_c)} \exp(u_o^T v_c) v_c$$

$$= -v_c + \hat{y}_o \cdot v_c = -v_c + (\vec{\hat{y}}^T \vec{y}) v_c$$

$$= -(1 - \vec{\hat{y}}^T \cdot \vec{y}) v_c \quad \text{update stops if } \langle \vec{\hat{y}}, \vec{y} \rangle \approx 1$$

$$\Updownarrow$$

$$\hat{y}_o \approx 1$$

$$\frac{\partial J}{\partial u_w}\Big|_{w\neq o} = \frac{1}{\sum_w \exp(u_w^T v_c)} \exp(u_w^T v_c) v_c$$

$$= \hat{y}_w \cdot v_c \quad \text{update stops if } \hat{y}_w \approx 0. \quad \#$$

(d) $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$ (4)

Compute $\frac{d\vec{\sigma}(x)}{d\vec{x}}$. if scalar, $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$

Sol: result should a $n \times n$ Jaccobi matrix where $n = len(\vec{x})$.

$x = (x_1, \cdots, x_n)$ $\sigma(x) = (\sigma(x_1), \cdots, \sigma(x_n)) = \left(\frac{1}{1+e^{x_1}}, \cdots, \frac{1}{1+e^{x_n}}\right)$

$\frac{\partial \sigma(x_i)}{\partial \sigma(x_j)} = \sigma(x_i)(1-\sigma(x_i))$ if $i=j$ else 0

$$\left[\frac{d\sigma(\vec{x})}{d\vec{x}}\right]_{ij} = \frac{\partial \sigma(\vec{x})|_i}{\partial \vec{x}|_j} = \begin{cases} \sigma(x_i)(1-\sigma(x_i)) & i=j \\ 0 & i\neq j \end{cases}$$

$$\downarrow$$

$$\begin{bmatrix} \sigma(x_1)(1-\sigma(x_1)) & & 0 \\ & \ddots & \\ 0 & & \sigma(x_n)(1-\sigma(x_n)) \end{bmatrix} = diag\left[\sigma(\vec{x}) \cdot (1-\sigma(\vec{x}))\right]$$

$\cdot$: element wise multiply

(e) $J_{\text{neg-sample}}(v_c, o, U) = -\log \sigma(u_o^T v_c) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

Compute $\partial J / \partial v_c$, $\partial J / \partial u_o$ and $\partial J / \partial u_k$ $(k = 1, \dots K)$

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

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

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

$$= -u_o + \sum_{w=0,1}^{K} \sigma(u_w^T v_c) u_w$$

$$= -u_o + \sum_{w \in \{o, w_1, \dots w_K\}} P(O = w | C = c) u_w$$

update stops when $P(O = o | C = c) \approx 1$ and $P(O = w_k | C = c) \approx 0$ $\forall k \in \{1, \dots K\}$ #

$$\frac{\partial J}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c)(1 - \sigma(u_o^T v_c)) v_c - 0$$

$$= -(1 - \sigma(u_o^T v_c)) v_c = -(1 - P(O = o | C = c)) v_c$$

update stops when $P(O = o | C = c) \approx 1$ #

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

$$= (1 - \sigma(-u_k^T v_c)) v_c = \sigma(u_k^T v_c) v_c$$

$$= P(O = w_k | C = c) v_c$$

update stops when $P(O = w_k | C = c) \approx 0$ #

(f) $c = w_t$

context window $[w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m}]$

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

Compute

(i) $\partial J_{\text{skip-gram}} / \partial U$

(ii) $\partial J_{sg} / \partial v_c$

(iii) $\partial J_{sg} / \partial v_w \quad w \neq c$

in terms of $\partial J(v_c, w_{t+j}, U)/\partial U$ and $\partial J(v_c, w_{t+j}, U)/\partial v_c$

Sol:

$$\frac{\partial J_{sg}}{\partial U} = \frac{\partial}{\partial U}\left(\sum_{\substack{-m \le j \le m \\ j \neq 0}} J(v_c, w_{t+j}, U)\right)$$

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

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

$$\frac{\partial J_{sg}}{\partial v_{w \mid w \neq c}} = 0$$

#