$$P(O=o \mid C=c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} = \hat{y}_o \in \mathbb{R}^{1\times 1}$$

$u_o \in \mathbb{R}^{n\times 1}$, $v_c \in \mathbb{R}^{n\times 1}$

$$U^T = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{Vocab} \end{bmatrix}^T \in \mathbb{R}^{n\times Vocab}$$

(b)-(i)

$$J_{naive\text{-}softmax}(v_c, o, U) = -\log P(O=o \mid C=c) = -\sum_{w \in Vocab} y_w \cdot \log \hat{y}_w = -\log(\hat{y})$$

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

$U \in \mathbb{R}^{Vocab \times n}$

$$\underbrace{\frac{\partial J}{\partial v_c}}_{\in \mathbb{R}^{1\times n}} = -u_o^T + \frac{\sum_w u_w^T \exp(u_w^T v_c)}{\sum_w \exp(u_o^T v_c)} = \exp(\underbrace{U}_{Vocab\times n})\underbrace{v_c}_{n\times 1}$$

$$= -u_o^T + \sum_w \underbrace{P(O=w \mid C=c)}_{1\times 1} \cdot \underbrace{u_w^T}_{1\times n}$$

$$= -u_o^T + \sum_w \underbrace{\hat{y}_w}_{1\times 1} \cdot \underbrace{u_w^T}_{1\times n}$$

$$= \sum_w \underbrace{(\hat{y}_w - y_w)}_{1\times 1} \cdot \underbrace{u_w^T}_{1\times n} = (U(\hat{y}-y))^T \in \mathbb{R}^{1\times n}$$

$v_c \in \mathbb{R}^{1\times n}$, $u_o \in \mathbb{R}^{1\times n}$ 일 때

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

$$\frac{\partial J}{\partial v_c} = \underbrace{(\hat{y}-y)}_{1\times Vocab}\underbrace{U}_{Vocab\times n} = -u_o + \frac{\sum_w u_w \exp(v_c u_w^T)}{\sum_w \exp(v_c u_w^T)}$$

(b)-(ii) $\frac{\partial J}{\partial v_c} = (U(\hat{y}-y))^T = \vec{0}$, The gradient is zero when the predicted distribution equals the true distribution ($\hat{y}=y$)

(b)-(iii)

$$v_c \leftarrow v_c - \eta \cdot \frac{\partial J}{\partial v_c} \iff v_c \leftarrow v_c - \eta \cdot (U\hat{y} - Uy)^T$$

(d)

i) $w=o$

$$\underbrace{\frac{\partial J}{\partial w}}_{1\times n} = -v_c^T + \frac{v_c^T \cdot \exp(u_o^T v_c)}{\sum_i \exp(u_i^T v_c)} = \left(-1 + \frac{\exp(u_o^T v_c)}{\sum_i \exp(u_i^T v_c)}\right) v_c^T = \underbrace{(\hat{y}_o - y_o)}_{1\times 1} \underbrace{v_c^T}_{1\times n}$$

ii) $w \neq o$

$$\underbrace{\frac{\partial J}{\partial w}}_{1\times n} = \frac{v_c^T \exp(u_w^T v_c)}{\sum_i \exp(u_i^T v_c)} = \underbrace{\hat{y}_w}_{1\times 1} \cdot \underbrace{v_c^T}_{1\times n}$$

$\hat{y} \in \mathbb{R}^{Vocab\times 1}$

(e)

$$\underbrace{\frac{\partial J}{\partial U}}_{Vocab \times n} = \begin{bmatrix} \frac{\partial J}{\partial u_1} \\ \vdots \\ \frac{\partial J}{\partial u_{Vocab}} \end{bmatrix} = \begin{bmatrix} \underbrace{\frac{\partial J}{\partial u_1}}_{1\times n} & \cdots & \frac{\partial J}{\partial u_{Vocab}} \end{bmatrix}^T$$

$\hat{y} \in \mathbb{R}^{1\times Vocab}$

(f) $$\frac{df}{dx} = \begin{cases} 1 & , x > 0 \\ \alpha & , x < 0 \end{cases}$$

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

(h)-(i) Let $x = u_o^T v_c$, $y = -u_{w_s}^T v_c$, $f = \log \Rightarrow J = -f(\sigma(x)) - \sum_{s=1}^{K} f(\sigma(y))$

$u_o, v_c \in \mathbb{R}^{1\times d}$

① $$\underbrace{\frac{\partial J}{\partial v_c}}_{\in \mathbb{R}^{1\times d}} = -\frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial x}\cdot\frac{\partial x}{\partial v_c} - \sum_{s=1}^{K} \frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial y}\cdot\frac{\partial y}{\partial v_c}$$

$$= -\frac{1}{\cancel{\sigma(x)}}\cdot\cancel{\sigma(x)}\cdot(1-\sigma(x))\cdot u_o^T + \sum_{s=1}^{K}\frac{1}{\cancel{\sigma(y)}}\cdot\cancel{\sigma(y)}\cdot(1-\sigma(y))\cdot(+u_{w_s}^T)$$

$$= (\sigma(u_o^T v_c)-1)u_o^T + \sum_{s=1}^{K}(1-\sigma(-u_{w_s}^T v_c))\cdot u_{w_s}^T$$

$(\sigma(u_o v_c^T)-1)u_o + \sum_{s=1}^{K}(1-\sigma(-u_{w_s} v_c^T))u_{w_s}$

② $$\frac{\partial J}{\partial u_o} = -\frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial x}\cdot\frac{\partial x}{\partial u_o} - \sum_{s=1}^{K}\underbrace{\frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial y}}\cdot\boxed{\frac{\partial y}{\partial u_o}}$$

$$= -\frac{1}{\cancel{\sigma(x)}}\cdot\cancel{\sigma(x)}\cdot(1-\sigma(x))\cdot v_c^T - \sum_{s=1}^{K}\frac{1}{\sigma(y)}\cdot\sigma(y)(1-\sigma(y))\cdot 0$$

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

$$③\ \frac{\partial J}{\partial u_{w_s}} = -\frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial x}\cdot\frac{\partial x}{\partial u_{w_s}} - \sum_{s=1}^{K}\frac{\partial f}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial y}\cdot\frac{\partial y}{\partial u_{w_s}}$$

$$u_{w_s} \leftarrow u_{w_s} - \eta\cdot\frac{\partial J}{\partial u_{w_s}}$$

$$= -\frac{1}{\cancel{\sigma(x)}}\cdot\cancel{\sigma(x)}\cdot(1-\sigma(x))\cdot 0 + \frac{1}{\cancel{\sigma(y)}}\cdot\cancel{\sigma(y)}\cdot(1-\sigma(y))\cdot(+v_c^T)$$

$$= \left(1-\sigma(-u_{w_s}^T v_c)\right)\cdot v_c^T$$

(h)-(ii) $U_{o,\{w_1,\dots,w_K\}} = [u_o, -w_1, \dots, -w_K]$ $\quad\therefore\ \sigma\left(U_{o,\{w_1,\dots,w_K\}}^T v_c\right)$

(i) $$J_{neg\text{-}sample} = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c))$$

$$= -\log(\sigma(u_o^T v_c)) - \left\{\sum_{i:w_i=w_s}\log(\sigma(-u_{w_i}^T v_c)) + \sum_{j:w_j\neq w_s}\log(\sigma(-u_{w_j}^T v_c))\right\}$$

$$\frac{\partial J}{\partial u_{w_s}} = -\sum_{i:w_i=w_s}\frac{\partial}{\partial u_{w_s}}\left\{\log(\sigma(-u_{w_i}^T v_c))\right\} = -\sum_{i:w_i=w_s}\sigma(-u_{w_s}^T v_c)\,v_c$$

$$J_{neg\text{-}sample} = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c))$$

$u_o \in \mathbb{R}^{1\times d},\ v_c \in \mathbb{R}^{1\times d}$

$$= -\log(\sigma(u_o^T v_c)) - \left\{\sum_{i:w_i=w_s}\log(\sigma(-u_{w_i}^T v_c)) + \sum_{j:w_j\neq w_s}\log(\sigma(-u_{w_j}^T v_c))\right\}$$

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

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

(j)-(iii) $$\frac{\partial J_{skip\text{-}gram}(v_c, w_{t-m},\dots,w_{t+m}, U)}{\partial v_w} = \sum_{\substack{-m\le j\le m,\, j\neq 0\\ w_{t+j}=w}}\frac{J(v_c, w_{t+j}, U)}{\partial v_w}$$