**(b)**

$$J_{naive\text{-}softmax}(\vec{v}_c, o, U) = -\log P(O=o \mid C=c)$$

$$= -\log \frac{\exp(\vec{u}_o^T \vec{v}_c)}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)}$$

$$\frac{\partial}{\partial \vec{v}_c} J_{naive\text{-}softmax}$$

$$= \frac{\partial}{\partial \vec{v}_c}\left[-\log \exp(\vec{u}_o^T \vec{v}_c) + \log \sum_w \exp(\vec{u}_w^T \vec{v}_c)\right]$$

$$= -\frac{\partial}{\partial \vec{v}_c} \vec{u}_o^T \vec{v}_c + \frac{1}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \cdot \frac{\partial}{\partial \vec{v}_c} \sum_w \exp(\vec{u}_w^T \vec{v}_c)$$

$$= -\vec{u}_o + \frac{1}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \sum_w \frac{\partial}{\partial \vec{v}_c} \exp(\vec{u}_w^T \vec{v}_c)$$

$$= -\vec{u}_o + \frac{1}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \sum_w \exp(\vec{u}_w^T \vec{v}_c) \cdot \vec{u}_w$$

$$= -\vec{u}_o + \sum_w \frac{\exp(\vec{u}_w^T \vec{v}_o)\, \vec{u}_w}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)}$$

$y$: true distribution

$$= -\vec{u}_o + \underbrace{\sum_w P(O=w \mid C=c)}_{\text{predicted distribution.}} \vec{u}_w$$

$\hat{y}$: predicted distribution

$$= -U \cdot y + U\hat{y} = U(\hat{y} - y)$$

(c) $J_{naive\text{-}softmax}(\vec{v}_c, o, U) = -\log \dfrac{\exp(\vec{u}_o^T \vec{v}_c)}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)}$

① $w \neq o$

$$\frac{\partial}{\partial \vec{u}_w} J = \cancel{-\frac{\partial}{\partial \vec{u}_w} \vec{u}_o^T \vec{v}_c} + \frac{\partial}{\partial \vec{u}_w} \log \sum_w \exp(\vec{u}_w^T \vec{v}_c)$$

$$= 0 + \frac{1}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \sum_w \frac{\partial}{\partial \vec{u}_w} \exp(\vec{u}_w^T \vec{v}_c)$$

$$= \frac{1}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \frac{\partial}{\partial \vec{u}_w} \exp(\vec{u}_w \vec{v}_c)$$

$$= \frac{\exp(\vec{u}_w \vec{v}_c)}{\sum_w \exp(\vec{u}_w^T \vec{v}_c)} \cdot \vec{v}_c$$

$$= P(O=w \mid C=c) \cdot \vec{v}_c$$

$$= \hat{y}_w \cdot \vec{v}_c$$

② $w = o$

$$\frac{\partial}{\partial \vec{u}_o} J = -\vec{v}_c + \hat{y}_o \cdot \vec{v}_c$$

$$= -y_o \vec{v}_c + \hat{y}_o \vec{v}_c$$

$$= (\hat{y}_o - y_o) \vec{v}_c$$

**(d)**

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

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

$$= \sigma(x)\left(1-\sigma(x)\right) = \sigma(x)\sigma(-x)$$

重要性质：$\sigma(x)+\sigma(-x)=1$

**(e)**

①

$$\frac{\partial}{\partial \vec{v}_c} J = -\frac{\partial}{\partial \vec{v}_o}\log\left(\sigma(\vec{u}_o^T\vec{v}_c)\right) - \frac{\partial}{\partial \vec{v}_c}\sum_{k=1}^{K}\log\left(\sigma(-\vec{u}_k^T\vec{v}_c)\right)$$

$$= -\frac{1}{\sigma(\vec{u}_o^T\vec{v}_c)}\cdot\frac{\partial}{\partial \vec{v}_c}\sigma(\vec{u}_o^T\vec{v}_c) - \sum_{k=1}^{K}\frac{1}{\sigma(-\vec{u}_k^T\vec{v}_c)}\cdot\frac{\partial}{\partial \vec{v}_c}\sigma(-\vec{u}_k^T\vec{v}_c)$$

$$= \left(\sigma(\vec{u}_o^T\vec{v}_c)-1\right)\cdot\frac{\partial}{\partial \vec{v}_c}\vec{u}_o^T\vec{v}_c + \sum_{k=1}^{K}\left(\sigma(-\vec{u}_k^T\vec{v}_c)-1\right)\frac{\partial}{\partial \vec{v}_c}-\vec{u}_k^T\vec{v}_c$$

$$= \left(\sigma(\vec{u}_o^T\vec{v}_c)-1\right)\cdot\vec{u}_o^T + \sum_{k=1}^{K}\vec{u}_k\left(1-\sigma(-\vec{u}_k^T\vec{v}_c)\right)$$

$$= -\sigma(-\vec{u}_o^T\vec{v}_c)\,\vec{u}_o + \sum_{k=1}^{K}\sigma(\vec{u}_k^T\vec{v}_c)\,\vec{u}_k$$

②

$$\frac{\partial}{\partial \vec{u}_o} J_{neg\text{-}sample} = -\frac{\partial}{\partial \vec{u}_o}\log\left(\sigma(\vec{u}_o^T\vec{v}_c)\right) - \cancel{\sum_{k=1}^{K}\frac{\partial}{\partial \vec{u}_o}\log\left(\sigma(-\vec{u}_k^T\vec{v}_c)\right)}$$

$$= -\frac{1}{\sigma(\vec{u}_o^T \vec{v}_c)} \frac{\partial}{\partial \vec{u}_o} \sigma(\vec{u}_o^T \vec{v}_c)$$

$$= -\sigma(-\vec{u}_o^T \vec{v}_c) \vec{v}_c$$

(3) $\frac{\partial}{\partial \vec{u}_k} J_{neg\text{-}sample}$

$$= \cancel{-\frac{\partial}{\partial \vec{u}_k} \log(\sigma(\vec{u}_o^T \vec{v}_c))} - \sum_{k=1}^{K} \frac{\partial}{\partial \vec{u}_k} \log(\sigma(-\vec{u}_k^T \vec{v}_c))$$

$$= -\frac{\partial}{\partial \vec{u}_k} \log(\sigma(-\vec{u}_k^T \vec{v}_c))$$

$$= \sigma(\vec{u}_k^T \vec{v}_c) \vec{v}_c$$

Only need to compute for at most $K$ words, but the naive softmax has to go through the whole vocabulary, computing for $|V|$ words!

(f)

(i) $\frac{\partial}{\partial U} J_{skip\text{-}gram}(\vec{v}_c, w_{t-m}, \cdots, w_{t+m}, U)$

$$= \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial}{\partial U} J(\vec{v}_c, w_{t+j}, U)$$

(ii) $\frac{\partial}{\partial \vec{v}_c} J_{skip\text{-}gram}(\vec{v}_c, w_{t-m}, \cdots, w_{t+m}, U)$

$$= \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial}{\partial \vec{v}_c} J(\vec{v}_c, w_{t+j}, U)$$

(iii) $\frac{\partial}{\partial \vec{v}_w} J_{skip\text{-}gram}(\vec{v}_c, w_{t-m}, \cdots, w_{t+m}, U)$
$w \ne c$

$$= 0$$