Solution to Problem (a):

Since $y_o = 1$ and $y_w = 0$ for any $w \in Vocab \setminus \{o\}$, The equation (3) is obvious. ■

Solution to Problem (b):

According to the equations (1) and (2), note that

$$\begin{aligned}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\log\left(\frac{\exp(\boldsymbol{u}_o^T\boldsymbol{v}_c)}{\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\right)\\ &= -\boldsymbol{u}_o^T\boldsymbol{v}_c + \log\left(\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)\right).\end{aligned}$$

Thus, the desired partial derivative is,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{v}_c}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\boldsymbol{u}_o + \sum_{x\in Vocab}\left(\frac{\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}{\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\boldsymbol{u}_x\right)\\ &= -\boldsymbol{u}_o + \sum_{x\in Vocab}(\hat{y}_x\boldsymbol{u}_x)\\ &= \sum_{x\in Vocab}((\hat{y}_x - y_x)\boldsymbol{u}_x)\\ &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}).\end{aligned}$$

Note that this solution strictly follows “the shape convention”. (Transposed answer could be possible in some case.) ■

Solution to Problem (c):

We may take advantage of the first equation in the previous solution. First, when $w = o$,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{u}_o}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -\boldsymbol{v}_c + \frac{\exp(\boldsymbol{u}_o^T\boldsymbol{v}_c)}{\sum_{w\in Vocab}\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}\boldsymbol{v}_c\\ &= -\boldsymbol{v}_c + \hat{y}_o\boldsymbol{v}_c.\end{aligned}$$

On the other hand, when $w \neq o$,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{u}_w}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= \frac{\exp(\boldsymbol{u}_w^T\boldsymbol{v}_c)}{\sum_{x\in Vocab}\exp(\boldsymbol{u}_x^T\boldsymbol{v}_c)}\boldsymbol{v}_c\\ &= \hat{y}_w\boldsymbol{v}_c.\end{aligned}$$

In short, for any $w \in Vocab$,

$$\frac{\partial}{\partial\boldsymbol{u}_w}\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = (\hat{y}_w - y_w)\boldsymbol{v}_c.$$

■

Solution to Problem (d):

$$\frac{\partial}{\partial \boldsymbol{U}} \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = \begin{bmatrix} \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_{|Vocab|}} \end{bmatrix}.$$

■

Solution to Problem (e):

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

■

Solution to Problem (f):

Note that $\sigma(-x) = 1 - \sigma(x)$ and $\frac{d}{dx}\log(\sigma(x)) = \frac{\sigma(x)(1-\sigma(x))}{\sigma(x)} = 1 - \sigma(x)$.

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= -(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))\boldsymbol{u}_o - \sum_{k=1}^{K} (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))(-\boldsymbol{u}_k) \\ &= (\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) - 1)\boldsymbol{u}_o + \sum_{k=1}^{K} \sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c)\boldsymbol{u}_k, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{u}_o} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= (\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) - 1)\boldsymbol{v}_c, \\ \frac{\partial}{\partial \boldsymbol{u}_k} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= \sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c)\boldsymbol{v}_c. \end{aligned}$$

The negative sampling loss function is much more efficient than the naive-softmax loss, because it uses only a portion of vocabulary, so the computational burden gets significantly decreased. ■

Solution to Problem (g):

(i) $\frac{\partial}{\partial \boldsymbol{U}} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U}) = \sum_{-m \le j \le m;\, j \ne 0} \frac{\partial}{\partial \boldsymbol{U}} \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$,

(ii) $\frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U}) = \sum_{-m \le j \le m;\, j \ne 0} \frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$, and

(iii) $\frac{\partial}{\partial \boldsymbol{v}_w} \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, \boldsymbol{U}) = \boldsymbol{0}$, when $w \ne c$.

■