1 Written: Understanding word2vec

(a)

$\boldsymbol{y}_w$ is a one hot vector, so all the zeros in the summation will be dropped, except for where the value of $\boldsymbol{y}_w$ is 1.

(b)

(i)

$$
\begin{aligned}
J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U) &= -\log \frac{\exp \boldsymbol{u}_o^T \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \\
&= -\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c \\
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{v}_c} &= -\boldsymbol{u}_o + \frac{1}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \frac{\partial \sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c}{\partial \boldsymbol{v}_c} \\
&= -\boldsymbol{u}_o + \frac{1}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \sum_{w \in \text{Vocab}} \boldsymbol{u}_w \exp \boldsymbol{u}_w^T \boldsymbol{v}_c \\
&= -U\boldsymbol{y} + \sum_{w \in \text{Vocab}} \frac{\exp \boldsymbol{u}_w^T \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \boldsymbol{u}_w \\
&= -U\boldsymbol{y} + \sum_{w \in \text{Vocab}} \hat{\boldsymbol{y}}_w \boldsymbol{u}_w \\
&= -U\boldsymbol{y} + U\hat{\boldsymbol{y}} \\
&= U(\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}
$$

(ii)

- $\hat{\boldsymbol{y}} = \boldsymbol{y}$
- $(\hat{\boldsymbol{y}} - \boldsymbol{y})$ is in the null space of U

(c)

- where $w = o$ we have:

$$\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} &= \frac{\partial}{\partial \boldsymbol{u}_w}[-\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c] \\
&= -\boldsymbol{v}_c^T + \frac{1}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \frac{\partial}{\partial \boldsymbol{u}_w}[\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c] \\
&= -\boldsymbol{v}_c^T + \frac{\exp \boldsymbol{u}_w^T \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \boldsymbol{v}_c^T \\
&= -\boldsymbol{v}_c^T + \hat{\boldsymbol{y}}_w \boldsymbol{v}_c^T \\
&= (\hat{\boldsymbol{y}}_w - 1) \boldsymbol{v}_c^T
\end{aligned}$$

- where $w \neq o$ we have:

$$\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} &= \frac{\partial}{\partial \boldsymbol{u}_w}[-\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c] \\
&= \frac{1}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \frac{\partial}{\partial \boldsymbol{u}_w}[\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c] \\
&= \frac{\exp \boldsymbol{u}_w^T \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp \boldsymbol{u}_w^T \boldsymbol{v}_c} \boldsymbol{v}_c^T \\
&= \hat{\boldsymbol{y}}_w \boldsymbol{v}_c^T \\
&= (\hat{\boldsymbol{y}}_w - 0) \boldsymbol{v}_c^T
\end{aligned}$$

For it to be same shape as $\boldsymbol{u}_w$, we need to transpose $\boldsymbol{v}_c^T$, so the partial derivatives are:

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} = \begin{cases} (\hat{\boldsymbol{y}}_w - 1)\boldsymbol{v}_c & \text{if } w = o\,, \\ (\hat{\boldsymbol{y}}_w - 0)\boldsymbol{v}_c & \text{if } w \neq o\,. \end{cases}$$

or, for the general case:

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

(d)

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial U} = \boldsymbol{v}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^T$$

(e)

$$\frac{\partial \max \alpha x, x}{\partial x} = \begin{cases} \alpha & \text{if } x < 0\,, \\ 1 & \text{if } x > 0\,. \end{cases}$$

(f)

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

(g)

(i)

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

(ii)

$$\sigma(U_{o,\{w_1,\ldots,w_K\}}^T \boldsymbol{v}_c) - \mathbf{1}$$

(h)

$$
\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(\boldsymbol{v}_c, o, U)}{\partial - \boldsymbol{u}_{w_s}} &= \frac{\partial}{\partial - \boldsymbol{u}_{w_s}} - \sum_{\substack{1 \leq t \leq K \\ \boldsymbol{w}_t = \boldsymbol{w}_s}} \log \sigma(-\boldsymbol{u}_{w_t}^T \boldsymbol{v}_c) \\
&= - \sum_{\substack{1 \leq t \leq K \\ \boldsymbol{w}_t = \boldsymbol{w}_s}} \frac{-\sigma(-\boldsymbol{u}_{w_t}^T \boldsymbol{v}_c)(1 - \sigma(-\boldsymbol{u}_{w_t}^T \boldsymbol{v}_c))\boldsymbol{v}_c}{\sigma(-\boldsymbol{u}_{w_t}^T \boldsymbol{v}_c)} \\
&= \sum_{\substack{1 \leq t \leq K \\ \boldsymbol{w}_t = \boldsymbol{w}_s}} (\sigma(-\boldsymbol{u}_{w_t}^T \boldsymbol{v}_c) - 1)\boldsymbol{v}_c
\end{aligned}
$$

(i)

(i)

$$
\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, U)}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t+j}, U)}{\partial U}
$$

(ii)

$$
\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, U)}{\partial \boldsymbol{v}_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t+j}, U)}{\partial \boldsymbol{v}_c}
$$

(iii)

$$
\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, U)}{\partial \boldsymbol{v}_w} = 0
$$