CS 224n Assignment #2: Written Assignment

Yoshihiro Kumazawa

May 26, 2020

(a)

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

(b)

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

(c)

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

(d)

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

(e)

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

These are computationally less expensive than the naive-softmax loss because its summation ranges over only K numbers, which is usually much smaller than $|Vocab|$.