This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

We begin with the provided form of the naive softmax loss function,

$$J_{\text{naive-softmax}}(v_c, o, U) = -u_o^T v_c + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right).$$

To arrive at the derivative of the loss function given by equations (3) and (4) in the handout, we take partial derivative with respect to v_c and simplify via the Chain Rule,

$$\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial v_c} &= \frac{\partial}{\partial v_c}\left(-u_o^T v_c + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right) \\
&= -u_o + \frac{\partial}{\partial v_c}\left(\log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right) \\
&= -u_o + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \cdot \frac{\partial}{\partial v_c} \sum_{w \in Vocab} \exp(u_w^T v_c) \\
&= -u_o + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \cdot \sum_{w \in Vocab} \exp(u_w^T v_c) \cdot u_w.
\end{aligned}$$

Now, the Word2Vec algorithm tells us that the conditional probability distribution $\hat{y}_w$ is

$$\hat{y}_w = \frac{\sum_{w \in Vocab} \exp(u_w^T v_c)}{\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)}.$$

Hence, we get

$$\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial v_c} = -u_o + \sum_{w \in Vocab} \hat{y}_w u_w$$

which is the equivalent form in equation (4).

1.b

Similarly, we take the derivative, simplify via the Chain Rule, and use the definition of the conditional probability distribution to arrive at equation (6):

$$
\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial u_w} &= \frac{\partial}{\partial u_w}\left(-u_o^T v_c + \log\left(\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right)\right) \\
&= -v_c + \frac{1}{\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)} \cdot \frac{\partial}{\partial u_w'} \sum_{w' \in Vocab} \exp(u_{w'}^T v_c) \\
&= -v_c + \frac{1}{\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)} \cdot \sum_{w' \in Vocab} \exp(u_{w'}^T v_c) \cdot v_c \\
&= -v_c + \sum_{w \in Vocab} \hat{y}_w v_c.
\end{aligned}
$$

Since y_w is a one-hot vector such that

$$
y_w = \begin{cases} 1, \text{if } w = o \\ 0, \text{otherwise} \end{cases}
$$

we conclude

$$
\begin{aligned}
\frac{\partial J}{\partial u_w} &= -\sum_{w \in Vocab} y_w v_c + \sum_{w \in Vocab} \hat{y}_w v_c \\
&= \begin{cases} (\hat{y}_w - 1)v_c, \text{if } w = 0 \\ \hat{y}_w v_c, \text{otherwise.} \end{cases}
\end{aligned}
$$

Documentation: I did not collaborate with any other students on this assignment.