Extra Credit Challenge I

Singleton

$$J_{\text{naive-softmax}}(v_c, o, U) = -\log(\hat{y}_o) = -\log\left[\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right]$$

$$= -\left[\log\left(\exp(u_o^T v_c)\right) - \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right]$$

$$= -u_o^T v_c + \log\left[\sum_{w \in Vocab} \exp(u_w^T v_c)\right]$$

$$\Rightarrow \frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}(-u_o^T v_c) + \frac{\partial}{\partial v_c} \log\left[\sum_{w \in Vocab} \exp(u_w^T v_c)\right]$$

Now $\frac{\partial}{\partial v_c}(-u_o^T v_c) = -u_o$, and by the Chain Rule;

$$\frac{\partial}{\partial v_c} \log\left[\sum_{w \in Vocab} \exp(u_w^T v_c)\right] = \left[\sum_{x \in Vocab} \exp(u_x^T v_c)\right]^{-1} \frac{\partial}{\partial v_c}\left[\sum_{w \in Vocab} \exp(u_w^T v_c)\right]$$

$$= \left[\sum_{x \in Vocab} \exp(u_x^T v_c)\right]^{-1} \sum_{w \in Vocab} \exp(u_w^T v_c) u_w$$

$$\Rightarrow \frac{\partial J}{\partial v_c} = -u_o + \frac{\sum_{w \in Vocab} \exp(u_w^T v_c) u_w}{\sum_{x \in Vocab} \exp(u_x^T v_c)} = \left[\sum_{w \in Vocab} \underbrace{\frac{\exp(u_w^T v_c)}{\sum_{x \in Vocab} \exp(u_x^T v_c)}}_{\hat{y}_w} \cdot u_w\right] - u_o$$

$$= \left(\sum_{w \in Vocab} \hat{y}_w u_w\right) - u_o$$

$\Big\{$ $u_o = Uy$ since y is a one-hot vector that has 1 only for word o.

$$= U\hat{y} - Uy$$

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

Extra Credit Challenge II

Singleton

$$J_{naive\text{-}softmax}(v_c, o, U) = -u_o^T v_c + \log\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right]$$

Case I: Suppose $w = o$.

$$\frac{\partial J}{\partial u_w} = \frac{\partial J}{\partial u_o} = \frac{\partial}{\partial u_o}\left(-u_o^T v_c\right) + \frac{\partial}{\partial u_o}\log\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right]$$

$$= -v_c + \left[\sum_{x \in Vocab} \exp(u_x^T v_c)\right]^{-1} \frac{\partial}{\partial u_o}\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right]$$

But $\frac{\partial}{\partial u_o}\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right] = \exp(u_o^T v_c) v_c$ since u_o appears in only one term of the sum. The terms where u_o does not appear vanish when differentiated w.r.t. u_o. Therefore,

$$\frac{\partial J}{\partial u_o} = -v_c + \underbrace{\frac{\exp(u_o^T v_c)}{\sum_{x \in Vocab} \exp(u_x^T v_c)}}_{\hat{y}_o} v_c$$

$$= \hat{y}_o v_c - v_c$$

$$= (\hat{y}_o - 1) v_c.$$

Singleton

Case 2: Suppose $w \neq o$.

$$\frac{\partial J}{\partial u_w} = \frac{\partial}{\partial u_w}\left[-u_o^T v_c\right] + \frac{\partial}{\partial u_w}\log\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right]$$

$$= \quad 0 \quad + \left[\sum_{x \in Vocab} \exp(u_x^T v_c)\right]^{-1} \frac{\partial}{\partial u_w}\left[\sum_{w' \in Vocab} \exp(u_{w'}^T v_c)\right]$$

But ~~under~~ the sum, $\frac{\partial}{\partial u_w} \exp(u_{w'}^T v_c) = 0$ for all values of w' except $w' = w$, in which case it equals $\exp(u_w^T v_c) v_c$. So we have

$$\frac{\partial J}{\partial u_w} = \frac{\exp(u_w^T v_c) v_c}{\sum_{x \in Vocab} \exp(u_x^T v_c)} = \underbrace{\frac{\exp(u_w^T v_c)}{\sum_{x \in Vocab} \exp(u_x^T v_c)}}_{\hat{y}_w} v_c$$

$$= \hat{y}_w v_c.$$

Thus we have $$\frac{\partial J}{\partial u_w} = \begin{cases} (\hat{y}_w - 1) v_c & \text{if } w = o \\ \hat{y}_w v_c & \text{otherwise} \end{cases}$$

or equivalently. $$\frac{\partial J}{\partial u} = v_c (\hat{y} - y)^T$$