Question 1

1. (3 points) Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between y and $\hat{y}$; i.e., show that

$$-\sum_{w \in Vocab} y_w \log \hat{y_w} = -\log(\hat{y_o}) \tag{1}$$

Proof. $y_w = 1$ only happens when $w = o$, therefore $left = -1 \times \log(\hat{y_o}) = right$ □

2. (5 points) Compute the partial derivative of $\boldsymbol{J}_{naive-softmax}(v_c; o; U)$ with respect to v_c. Please write your answer in terms of $y, \hat{y}$, and v_c.

Denote $U^T v_c$ as θ, which is the inner products of words in dictionary and center word and is used as prediction probabilities. we have

$$\frac{\partial}{\partial v_c} \boldsymbol{J}_{naive-softmax}(v_c; o; U) = -\sum_{w \in Vocab} \frac{\partial y_w}{\partial v_c} \log \hat{y_w} + \frac{y_w}{\hat{y_w}} \frac{\partial \hat{y_w}}{\partial v_c} \tag{2}$$

$$= -\sum \frac{y_w}{\hat{y_w}} \frac{\partial \hat{y_w}}{\partial \theta} \frac{\partial \theta}{\partial v_c} \tag{3}$$

Note that if $\hat{y_w} = u_o^T v_c$,

$$\frac{\partial \hat{y_w}}{\partial \theta} = \hat{y_w}(1 - \hat{y_w}) \tag{4}$$

else

$$\frac{\partial \hat{y_w}}{\partial \theta} = -\hat{y_w} y_w \tag{5}$$

As we are doing supervised learning, the ground truth is y_o, we have

$$\frac{\partial \boldsymbol{J}}{\partial \theta} = y(\hat{y} - 1) = \hat{y} - y \tag{6}$$

So finally we have the overall partial derivatives

$$\frac{\partial \boldsymbol{J}}{\partial v_c} = \boldsymbol{U}^T (\hat{y} - y) \tag{7}$$

3. (5 points) Compute the partial derivatives of $\boldsymbol{J}_{naive-softmax}(v_c; o; U)$ with respect to each of the 'outside' word vectors, u_w's. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write you answer in terms of $y, \hat{y}$, and v_c.

Similar as before, we have

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{U}} = (\hat{y} - y) v_c^T \tag{8}$$

4. (3 Points) The sigmoid function is given by Equation 4:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \tag{9}$$

Please compute the derivative of $\sigma(x)$ with respect to x, where x is a vector.

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1-\sigma(x)) \tag{10}$$

5. (3 Points) (4 points) Now we shall consider the Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as $w_1, w_2, ..., w_K$ and their outside vectors as $u_1, ..., u_K$. Note that $o \notin \{w_1, ..., w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{neg-sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c)) \tag{11}$$

for a sample $w_1, w_2, ..., w_K$, where $\sigma()$ is the *sigmoid* function.
Please repeat parts (b) and (c), computing the partial derivatives of $\boldsymbol{J}_{neg-sample}$ with respect to v_c, with respect to u_o, and with respect to a negative sample u_k. Please write your answers in terms of the vectors u_o, v_c, and u_k, where $k \in [1, K]$. After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss. Note, you should be able to use your solution to part (d) to help compute the necessary gradients here.

$$\frac{\partial \boldsymbol{J}_{neg-sample}}{\partial v_c} = u_o(\sigma(u_o^T v_c) - 1) + \sum_{k=1}^{K} u_k(\sigma(u_k^T v_c)) \tag{12}$$

$$\frac{\partial \boldsymbol{J}_{neg-sample}}{\partial u_o} = v_c^T(\sigma(u_o^T v_c) - 1) \tag{13}$$

$$\frac{\partial \boldsymbol{J}_{neg-sample}}{\partial u_k} = v_c^T(1 - \sigma(-u_k^T v_c)) = v_c^T \sigma(u_k^T v_c) \tag{14}$$

6. (3 points) Suppose the center word is $c = wt$ and the context window is $[w_{t-m}, ..., w_{t-1}, w_t, w_{t+1}, ..., w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of word2vec, the total loss for the context window is:

$$\boldsymbol{J}_{naive-softmax}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U}) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \boldsymbol{J}(v_c, w_{t+j}, \boldsymbol{U}) \tag{15}$$

Here, $\boldsymbol{J}(v_c, w_{t+j}, \boldsymbol{U})$ represents an arbitrary loss term for the center word $c = wt$ and outside word w_{t+j}. $\boldsymbol{J}(v_c, w_{t+j}, \boldsymbol{U})$ could be $\boldsymbol{J}_{naive-softmax}(v_c, w_{t+j}, \boldsymbol{U})$ or $\boldsymbol{J}_{neg-sample}(v_c, w_{t+j}, \boldsymbol{U})$, depending on your implementation.
Write down three partial derivatives:
(i) $\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{U}$
(ii) $\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{v}_c$
(iii) $\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{v}_w$ when $w \neq c$
(i)

$$\frac{\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \boldsymbol{J}_{naive-softmax}(v_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{U}} \tag{16}$$

(ii)

$$\frac{\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \boldsymbol{J}_{naive-softmax}(v_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{v}_c} \tag{17}$$

(iii)

$$\frac{\partial \boldsymbol{J}_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_w} = 0 \tag{18}$$