Question 1: Understanding Word2Vec

Part a)

Since the true empirical distribution y is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else,

$$-\sum_{w\in Vocab} y_w \log\left(\hat{y}_w\right) = -\sum_{w\in Vocab} \mathbb{I}\{y_w = y_o\} \log\left(\hat{y}_w\right) = -\sum_{w\neq o} 0 \log\left(\hat{y}_w\right) - 1 \times \log\left(\hat{y}_o\right) = -\log\left(\hat{y}_o\right)$$

Part b)

As we did in class, we can start with $\frac{\partial}{\partial v_c} \boldsymbol{J}_{\text{naive-softmax}}\left(v_c, o, \boldsymbol{U}\right)$

$$\begin{aligned}
\frac{\partial}{\partial v_c} \boldsymbol{J}_{\text{naive-softmax}}\left(v_c, o, \boldsymbol{U}\right) &= -\frac{\partial}{\partial v_c} log\left[\frac{\exp\left(\boldsymbol{u}_o^{\top} \boldsymbol{v}_c\right)}{\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)}\right] \\
&= -\frac{\partial}{\partial v_c} \boldsymbol{u}_o^{\top} \boldsymbol{v}_c + \frac{\partial}{\partial v_c} log\left[\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)\right] \\
&= -\boldsymbol{u}_o + \left[\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)\right]^{-1} \sum_{j\in \text{ Vocab}} \frac{\partial}{\partial v_c} \exp\left(\boldsymbol{u}_j^{\top} \boldsymbol{v}_c\right) \\
&= -\boldsymbol{u}_o + \sum_{j\in \text{ Vocab}} \frac{\boldsymbol{u}_j^{\top} \boldsymbol{v}_c}{\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)} \boldsymbol{u}_j \\
&= -U\boldsymbol{y} + \boldsymbol{U}\hat{\boldsymbol{y}} = \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}$$

The last line comes from the observations that a one-hot vector times a matrix is basically a column or row lookup table, since it selects and preserves a single row or column depending on multiplication order. Here, each column of U corresponds to a word in the vocabulary, and to select the column corresponding to the outside word, $\boldsymbol{u}_o = Uy$. We perform a similar action for $\hat{y}$, a vector of softmax probabilities.

Part c) First let's take the derivative for $f(x_i) = -log(p(x_i)) = -log\ \frac{e^{x_i}}{\sum_j e^{x_j}}$

$$\begin{aligned}
\frac{\partial}{\partial x_i} f(x_i) &= -\frac{\partial}{\partial x_i} log\ e^{x_i} + \frac{\partial}{\partial x_i} log(\sum_j e^{x_j}) \\
&= -1 + \frac{e^x}{\sum_y e^y} = p(x_i) - 1
\end{aligned}$$

We can then see for a vector of softmax probabilities $\hat{y}$ and a true vector y, we get $\frac{\partial f(x)}{\partial x} = \hat{y} - y$

Now let $x_i = u_i^T v_c$. Then we can note from the chain rule that $\frac{\partial J}{\partial u_w} = \frac{\partial J}{\partial x_w} \frac{\partial x_w}{\partial u_w}$ Applying our finding from above, we see

$$\begin{aligned}
\frac{\partial J}{\partial x_w} &= \hat{y}_w - y_w \\
\frac{\partial x_w}{\partial u_w} &= v_c
\end{aligned}$$

Therefore when $u_w = u_o$ we get

$$\begin{aligned}
\frac{\partial J}{\partial u_w} &= \frac{\partial J}{\partial x_w} \frac{\partial x_w}{\partial u_w} \\
&= (\hat{y}_w - y_w) v_c = (\hat{y}_w - 1) v_c
\end{aligned}$$

When $u_w \neq u_o$ we get

$$\begin{aligned}\frac{\partial J}{\partial u_w} &= \frac{\partial J}{\partial x_w}\frac{\partial x_w}{\partial u_w}\\ &= (\hat{y}_w - y_w)v_c = (\hat{y}_w - 0)v_c = \hat{y}_w v_c\end{aligned}$$

To check our work, we can also calculate the gradients directly. For $u_w = u_o$

$$\begin{aligned}\frac{\partial}{\partial u_o}\boldsymbol{J}_{\text{naive-softmax}}\ (v_c, o, \boldsymbol{U}) &= -\frac{\partial}{\partial u_o}\boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial u_o} log\left[\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)\right]\\ &= -\boldsymbol{v}_c + \left[\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)\right]^{-1} \sum_{j\in\text{ Vocab}} \frac{\partial}{\partial u_o}\exp\left(\boldsymbol{u}_j^\top \boldsymbol{v}_c\right)\\ &= -\boldsymbol{v}_c + \frac{\boldsymbol{u}_o^\top \boldsymbol{v}_c}{\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)}\boldsymbol{v}_c\\ &= -\boldsymbol{v}_c + \hat{\boldsymbol{y}}_o \boldsymbol{v}_c = (\hat{\boldsymbol{y}}_o - 1)\boldsymbol{v}_c\end{aligned}$$

For $u_w \neq u_o$

$$\begin{aligned}\frac{\partial}{\partial u_w}\boldsymbol{J}_{\text{naive-softmax}}\ (v_c, o, \boldsymbol{U}) &= -\frac{\partial}{\partial u_w}\boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial u_w} log\left[\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)\right]\\ &= \left[\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)\right]^{-1} \sum_{j\in\text{ Vocab}} \frac{\partial}{\partial u_w}\exp\left(\boldsymbol{u}_j^\top \boldsymbol{v}_c\right)\\ &= \frac{\boldsymbol{u}_w^\top \boldsymbol{v}_c}{\sum_{w\in\text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)}\boldsymbol{v}_c\\ &= \hat{\boldsymbol{y}}_o \boldsymbol{v}_c\end{aligned}$$

Part d)

Using our result from above:

$$\begin{aligned}\frac{\partial J}{\partial U} &= \begin{bmatrix}\frac{\partial J}{\partial u_1} & \cdots & \frac{\partial J}{\partial u_{|V|}}\end{bmatrix}\\ &= \begin{bmatrix}\hat{y}_1 v_c & \cdots & (\hat{y}_o - 1)v_c & \cdots & \hat{y}_{|V|}v_c\end{bmatrix}\end{aligned}$$

or put another way, let $\left[\frac{\partial J}{\partial U}\right]_k$ denote column k of $\frac{\partial J}{\partial U}$, then $\left[\frac{\partial J}{\partial U}\right]_k = \begin{cases}\hat{y}_k v_c & k \neq o\\ (\hat{y}_k - 1)v_c & k = o\end{cases}$

Part e)

$$\begin{aligned}\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\frac{1}{1+e^{-x}}\\ &= (-1)(1+e^{-x})^{-2}(-1)e^{-x}\\ &= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{(e^{-x}+1)-1}{(1+e^{-x})^2} = \frac{(e^{-x}+1)}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2}\\ &= \frac{1}{(1+e^{-x})} - \frac{1}{(1+e^{-x})^2}\\ &= \frac{1}{(1+e^{-x})}\left[1 - \frac{1}{(1+e^{-x})}\right] = \sigma(x)(1-\sigma(x))\end{aligned}$$

Part f)

$$\begin{aligned}\frac{\partial J}{\partial v_c} &= -\frac{\partial}{\partial v_c}\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K}\frac{\partial}{\partial v_c}\log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right)\\ &= -\left[\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)\right]^{-1}\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c))\boldsymbol{u}_o - \sum_{k=1}^{K}\left[\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)\right]^{-1}\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))(-\boldsymbol{u}_k)\\ &= (\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)-1)\boldsymbol{u}_o + \sum_{k=1}^{K}(1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))(\boldsymbol{u}_k)\end{aligned}$$

For $o \neq k,\ k \in [1, K]$

$$\begin{aligned}\frac{\partial J}{\partial u_o} &= \frac{\partial J}{\partial f}\frac{\partial f}{\partial g}\frac{\partial g}{\partial h}\frac{\partial h}{\partial u_o}\\ &= (\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)-1)\boldsymbol{v}_c\\ \frac{\partial J}{\partial u_k} &= (1-\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))\boldsymbol{v}_c\end{aligned}$$

These gradients should be more efficient to calculate, since we no longer have the denominator from the softmax function to compute, $\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)$. This required a computation across all word vectors, while these gradients either have much smaller sums over K samples or simple vector dot products.

Part g)

For specific word u_i

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

Then for $u_i \neq u_o$

$$\begin{aligned}\frac{\partial J}{\partial u_i} &= -\frac{\partial J}{\partial u_i}\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k\neq i}\frac{\partial J}{\partial u_i}\log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right) - \sum_{j=i}\frac{\partial J}{\partial u_i}\log\left(\sigma\left(-\boldsymbol{u}_j^\top \boldsymbol{v}_c\right)\right)\\ &= -\sum_{j=i}\frac{\partial J}{\partial u_i}\log\left(\sigma\left(-\boldsymbol{u}_j^\top \boldsymbol{v}_c\right)\right)\\ &= \sum_{j=i}(1-\sigma(-\boldsymbol{u}_j^\top \boldsymbol{v}_c))\boldsymbol{v}_c\\ &= n_i(1-\sigma(-\boldsymbol{u}_j^\top \boldsymbol{v}_c))\boldsymbol{v}_c\end{aligned}$$

where n_i is the count of repetitions for word u_i in the sample.

Part h)

i)

$$\partial \boldsymbol{J}_{\text{skip-gram }}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{U} = \sum_{\substack{-m\leq j\leq m\\ j\neq 0}}\frac{\partial}{\partial \boldsymbol{U}}J\left(v_c, w_{t+j}, U\right)$$

ii)

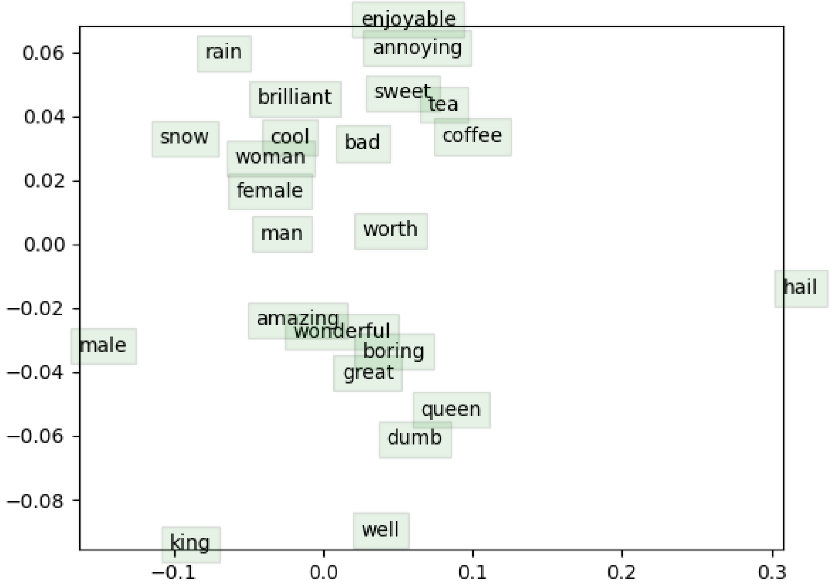
$$\partial \boldsymbol{J}_{\text{skip-gram }}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_{\boldsymbol{c}}=\sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial \boldsymbol{J}\left(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_c$$

iii)

$$\partial \boldsymbol{J}_{\text{skip-gram }}\left(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_w=\sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial \boldsymbol{J}\left(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}\right) / \partial \boldsymbol{v}_m=0$$

Question 2: Implementing Word2Vec

c) Below is the output word vector visualization. The clustering of words appears to make some sense. Words like ”sweet,” ”tea,” and ”coffee” are words associated with beverages while generic adjectives like ”amazing,” ”wonderful,” ”boring,” and ”great” also logically cluster together. The vectors also seem to have some analogy ability, as the vector from male to king is quite similar to that from female to queen. However, some of the outputs are definitely less expected, like how far ”hail” is from other precipitation or how ”enjoyable” and ”annoying” form their own cluster distinct from the adjectives mentioned before.


enjoyable
annoying
rain
brilliant
sweet
tea
snow
cool
bad
coffee
woman
female
man
worth
hail
amazing
wonderful
male
boring
great
queen
dumb
well
king