Problem 3

Skip-gram will predict the surrounding words u_w based on the center word v_c.

a.) Derive the gradient wrt v_c

Problem gives the following expression

$$\hat{y}_o = p(o \mid c) = \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}$$

but I don't like the notation; it will make doing Einstein notation difficult. Note that o is used as the word itself, as an index to u specifying the output word vector corresponding to the word, and the location of the 1-value in the one-hot vector y corresponding to the word. To make it easier to do einstein, I'm going to recast the vectors (u and v) in terms of the matrices from which they come (U and V). v_c is the c'th column of the matrix V and u_o is the o'th row of U. His notation also reveals that he has switched from row-vector convention (in the previous problem) to column-vector convention which is EXTREMELY annoying that he didn't just pick one and stick with it. Anyways, because we're working with the matrices now, the u vector is effectively already transposed and we don't need to worry about the transpose here. As written below, it's the inner product of the o'th row of U and the c'th column of V. Also dropping the explicit some over w, it is now implied. Thus...

$$\hat{y}_o = \frac{\exp(U_{oj} V_{jc})}{\exp(U_{wk} V_{kc})}$$

The cost function is

$$J = CE(y, \hat{y}).$$

We want to find the gradient wrt the elements in V. Because I've already used the indices j and k as dummies in the definition of $\hat{y}$, I cannot use them again. So let us take the derivative wrt V_{ic}

$$\frac{\partial J}{\partial V_{ic}} = \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial V_{ic}}$$

Only the latter term in the chain rule is new. We calculated the first term in problem 2.

$$\begin{aligned}
\frac{\partial \hat{y}_a}{\partial V_{ic}} &= \frac{\partial}{\partial V_{ic}}\left[\frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})}\right] \\
&= \left[\frac{\frac{\partial}{\partial V_{ic}}\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})} - \frac{\exp(U_{aj}V_{jc})}{(\exp(U_{wk}V_{kc}))^2}\frac{\partial}{\partial V_{ic}}\exp(U_{zh}V_{hc})\right] \\
&= \left[\frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})}\frac{\partial}{\partial V_{ic}}(U_{am}V_{mc}) - \frac{\exp(U_{aj}V_{jc})}{(\exp(U_{wk}V_{kc}))^2}\exp(U_{zh}V_{hc})\frac{\partial}{\partial V_{ic}}(U_{zn}V_{nc})\right] \\
&= \left[\frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})}U_{am}\frac{\partial V_{mc}}{\partial V_{ic}} - \frac{\exp(U_{aj}V_{jc})}{(\exp(U_{wk}V_{kc}))^2}\exp(U_{zh}V_{hc})U_{zn}\frac{\partial V_{nc}}{\partial V_{ic}}\right] \\
&= \left[\hat{y}_a U_{am}\frac{\partial V_{mc}}{\partial V_{ic}} - \hat{y}_a\hat{y}_z U_{zn}\frac{\partial V_{nc}}{\partial V_{ic}}\right] \\
&= \left[\hat{y}_a U_{am}\delta_{mi} - \hat{y}_a\hat{y}_z U_{zn}\delta_{ni}\right] \\
&= \left[\hat{y}_a U_{ai} - \hat{y}_a\hat{y}_z U_{zi}\right]
\end{aligned}$$

In the first line above, notice that the a'th element of $\hat{y}$ corresponds to using the a'th row of U in the numerator exponential. Earlier we were using the index o here, but it doesn't matter what index you use as long as you're consistent. In fact, you had better to NOT use o here, because it's already been defined as indexing on our target words and we need the indices in the derivatives to be general. The second line just distributes the derivative through. Notice the appearance of the indices z and h. The z distinguishes that this is now a different sum than the one that is implied by w elsewhere in the equation. Similarly, h is used to index over this new inner product.

In the third line, new indices m and n appear to distinguish from the inner products over j and h — inner products with respect to which we have already taken derivatives. This is subtle and one of the hardest aspects to grasp of using einstein notation — knowing when and why to use new indices — so don't worry if you don't get it at first. The high level reasoning for introducing m is that the derivative has already moved past the inner product over j, so it would be incorrect to have new terms appear with j in them because they would mess up the inner product over j.

If you survived line 3 above, then you're home clear. Now plug everything back in to find

$$\begin{aligned}\frac{\partial J}{\partial V_{ic}} &= \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial V_{ic}} \\ &= -\frac{y_a}{\hat{y}_a}\left[\hat{y}_a U_{ai} - \hat{y}_a \hat{y}_z U_{zi}\right] \\ &= -y_a\left[U_{ai} - \hat{y}_z U_{zi}\right] \\ &= -y_a U_{ai} + y_a \hat{y}_z U_{zi}\end{aligned}$$

In the second term, there are no repeated a indices, so that's a sum over the one-hot vector which is just 1. Converting to matrix notation

$$\begin{aligned}\frac{\partial J}{\partial V_{:c}} &= \hat{y}^T U - y^T U \\ &= (\hat{y} - y)^T U\end{aligned}$$

Recall that $\hat{y}$ is conditioned on c even though it lacks a formal index.

b.) Derive the gradient wrt u_w

What a slog! Okay, that's enough complaining. Similar to part a. Let's not use the index w because it already appears in the definition of $\hat{y}$, instead let's take the derivative of the zth row of U

$$\frac{\partial J}{\partial U_{zi}} = \frac{\partial J}{\partial \hat{y}_a}\frac{\partial \hat{y}_a}{\partial U_{zi}}$$

Only need to compute the latter term

$$\begin{aligned}
\frac{\partial \hat{y}_a}{\partial U_{zi}} &= \frac{\partial}{\partial U_{zi}}\left[\frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})}\right] \\
&= \frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})} V_{mc} \frac{\partial U_{am}}{\partial U_{zi}} - \frac{\exp(U_{aj}V_{jc})}{(\exp(U_{wk}V_{kc}))^2} \frac{\partial}{\partial U_{zi}} \exp(U_{ph}V_{hc}) \\
&= \frac{\exp(U_{aj}V_{jc})}{\exp(U_{wk}V_{kc})} V_{mc}\delta_{az}\delta_{mi} - \frac{\exp(U_{aj}V_{jc})}{(\exp(U_{wk}V_{kc}))^2} \exp(U_{ph}V_{hc}) V_{nc}\delta_{pz}\delta_{ni} \\
&= \hat{y}_a V_{ic}\delta_{az} - \hat{y}_a \hat{y}_p V_{nc}\delta_{pz}\delta_{ni} \\
&= \hat{y}_a V_{ic}\delta_{az} - \hat{y}_a \hat{y}_z V_{ic}
\end{aligned}$$

Since the above calculation was substantially similar to part a.), I took the liberty of skipping many steps. Now plug everything in

$$\begin{aligned}
\frac{\partial J}{\partial U_{zi}} &= \frac{\partial J}{\partial \hat{y}_a} \frac{\partial \hat{y}_a}{\partial U_{zi}} \\
&= -\frac{y_a}{\hat{y}_a}\left[\hat{y}_a V_{ic}\delta_{az} - \hat{y}_a \hat{y}_z V_{ic}\right] \\
&= -y_a\left[V_{ic}\delta_{az} - \hat{y}_z V_{ic}\right] \\
&= -y_a V_{ic}\delta_{az} + y_a \hat{y}_z V_{ic} \\
&= -y_z V_{ic} + \hat{y}_z V_{ic} \\
&= (\hat{y}_z - y_z) V_{ic}
\end{aligned}$$

Then in matrix notation

$$\frac{\partial J}{\partial U_{z:}} = [\hat{y} - y]_z V_{:c}$$

or

$$\frac{\partial J}{\partial u_z} = [\hat{y} - y]_z v_c$$

or

$$\begin{aligned}
\frac{\partial J}{\partial U} &= (\hat{y} - y) v_c^T \\
&= v_c (\hat{y} - y)^T
\end{aligned}$$

c.) Now do for negative sampling cost function

Here's the cost function, for minimization, written in Einsein notation using my convention of dealing with U and V matrices directly

$$J(o,c) = -\log(\sigma(U_{oj}V_{jc})) - \log(\sigma(-U_{kj}V_{jc})).$$

Note: o and c are NOT like the other indices! They are parameters / arguments to the function. They do not imply summations. So in the coming calculations you will see terms with multiple o and k subscripts, but keep in mind that this does not mean there is a sum over them.

Derivative wrt V

$$\begin{aligned}
\frac{\partial J}{\partial V_{ic}} &= -\frac{1}{\sigma(U_{oj}V_{jc})}\sigma'(U_{on}V_{nc})\frac{\partial}{\partial V_{ic}}[U_{om}V_{mc}] \\
&\qquad - \frac{1}{\sigma(-U_{kj}V_{jc})}\sigma'(-U_{kn}V_{nc})\frac{\partial}{\partial V_{ic}}[-U_{km}V_{mc}] \\
&= -\frac{1}{\sigma(U_{oj}V_{jc})}\sigma'(U_{on}V_{nc})U_{om}\delta_{mi} \\
&\qquad + \frac{1}{\sigma(-U_{kj}V_{jc})}\sigma'(-U_{kn}V_{nc})U_{km}\delta_{mi} \\
&= -(1-\sigma(U_{on}V_{nc}))U_{om}\delta_{mi} \\
&\qquad + (1-\sigma(-U_{kn}V_{nc}))U_{km}\delta_{mi} \\
&= -(1-\sigma(u_o^T v_c))U_{om}\delta_{mi} + (1-\sigma(-u_k^T v_c))U_{km}\delta_{mi} \\
&= -(1-\sigma(u_o^T v_c))U_{oi} + (1-\sigma(-u_k^T v_c))U_{ki}
\end{aligned}$$

Now convert to matrix notation

$$\frac{\partial J}{\partial v_c} = -(1-\sigma(u_o^T v_c))u_o + \sum_k (1-\sigma(-u_k^T v_c))u_k$$

Derivative wrt U

$$\begin{aligned}\frac{\partial J}{\partial U_{zi}} &= -\frac{1}{\sigma(U_{oj}V_{jc})}\sigma'(U_{on}V_{nc})\frac{\partial}{\partial U_{zi}}[U_{om}V_{mc}] \\ &\qquad - \frac{1}{\sigma(-U_{kj}V_{jc})}\sigma'(-U_{kn}V_{nc})\frac{\partial}{\partial U_{zi}}[-U_{km}V_{mc}] \\ &= -\frac{1}{\sigma(U_{oj}V_{jc})}\sigma'(U_{on}V_{nc})V_{mc}\delta_{mi}\delta_{oz} \\ &\qquad + \frac{1}{\sigma(-U_{kj}V_{jc})}\sigma'(-U_{kn}V_{nc})V_{mc}\delta_{mi}\delta_{kz} \\ &= -(1-\sigma(U_{zn}V_{nc}))V_{mc}\delta_{mi}\delta_{oz} \\ &\qquad + (1-\sigma(-U_{kn}V_{nc}))V_{mc}\delta_{mi}\delta_{kz} \\ &= -(1-\sigma(U_{zn}V_{nc}))V_{ic}\delta_{oz} \\ &\qquad + (1-\sigma(-U_{kn}V_{nc}))V_{ic}\delta_{kz}\end{aligned}$$

This one is a little tricky because you cannot evaluate the δ's that depend on z until a z has been chosen. I didn't see it at first either, but notice that they are mutually exclusive; if $\delta_{oz} = 1$ then $\delta_{kz} = 0$ and vice versa. In other words, z is a parameter in the equation, not an index over which you can sum. Convert to matrix notation

$$\frac{\partial J}{\partial u_z} = -(1-\sigma(u_z^T v_c))v_c\delta_{oz} + (1-\sigma(-u_k v_c))v_c\delta_{kz}$$

Which reveals that if u corresponds to your expected word o, then

$$\frac{\partial J}{\partial u_o} = -(1-\sigma(u_o^T v_c))v_c.$$

If u is not the expected word, but is one of the k words you're negatively sampling, then

$$\frac{\partial J}{\partial u_k} = (1-\sigma(-u_k v_c))v_c.$$

If u is any other word, then the derivative is zero. This will be the case for the vast majority of the words in your lexicon.

d.) Gradients for all word vectors in skip-gram and CBOW

The cost function, where F is either the cross-entropy or negative sampling cost functions from above.

$$J(c-m,\dots,c+m) = \sum_{-m\le j\le m, j\ne 0} F(w_{c+j}, v_c)$$

Skip–gram softmax

First using the cross-entropy,

$$\begin{aligned}\frac{\partial J}{\partial v_c} &= \sum_j \frac{\partial}{\partial v_c} F(w_{c+j}, v_c) \\ &= \sum_j U^T \left[\hat{y}(c+j) - y(c+j)\right]\end{aligned}$$

and, for an individual output vector,

$$\begin{aligned}\frac{\partial J}{\partial u_k} &= \sum_j \frac{\partial}{\partial u_k} F(w_{c+j}, v_c) \\ &= \sum_j \left[\hat{y}(c+j) - y(c+j)\right]_k v_c\end{aligned}$$

or for all of them at once

$$\frac{\partial J}{\partial U} = \sum_j \left[\hat{y}(c+j) - y(c+j)\right] \otimes v_c$$

NOTE: You can code this up and compare against the solutions code so you know that these expressions are correct; however, the default SGD parameters in the code are tuned for the negative sampling cost function. This cost function won't converge unless you tweak them. I don't care to spend the time on that.

Skip–gram negative sampling

$$\begin{aligned}\frac{\partial J}{\partial v_c} &= \sum_j \frac{\partial}{\partial v_c} F(w_{c+j}, v_c) \\ &= \sum_j \frac{\partial}{\partial v_c}\left[-\log(\sigma(u_{c+j}^T v_c)) - \sum_p \log(\sigma(-u_p^T v_c))\right] \\ &= \sum_j \left[-(1-\sigma(u_{c+j}^T v_c))u_{c+j} + \sum_p (1-\sigma(-u_p^T v_c))u_p\right]\end{aligned}$$

Gotta be careful with this one; the sum over p cannot be easily combined with the sum over j. For every j, a new batch of negative samples must be generated for the sum over p. Won't be an issue though the way the code has modularized these functions.

CBOW Gradients

Diminishing returns. I'd rather move forward than work on this.