Question 3

Assuming cross-entropy cost is applied to this prediction and word O is the expected word.

let us denote:

Now, let us first calculate the partial derivative of J w.r.t any :

And now to the partial derivative w.r.t :

1. Derived by using the chain rule:
2. In this case, the new cost function is defined as:

c. (continued):

* Since we’re assuming .
* For :
* For :

It is much more efficient to compute the negative sample cost function because we only have to sum over , instead of summing over the entire vocabulary.

1. Gradients: In the case of skip-gram, we can use the fact that the derivative of a sum is the derivative of every item in the sum. Therefore:

and for :

In both cases, we already know how to calculate the derivatives. All that’s left is to sum them up.

For , in cases where the derivative is simply 0 because the cost function is not a function of for . It is a function of and .

**For CBOW**: In this case, we want to predict the center words given its context:

The reasoning is as follows:  
Since , the derivative of the sum is the derivative of every part of the sum and so:

g. In the following plot we can see a 2D representation of two features in the word vectors, and how the words are spaced in the subspace spanned by those features. We are taking the two features with the highest variance (i.e. most “info”), from the SVD decomposition.