Exercise 1. (a) Softmax is invariant to constant offsets in the input, that is, for any input vector x and any constant c,

$$
\begin{aligned}
softmax(x_i + c) &= \frac{e^{(x_i+c)}}{\sum_j e^{(x_j+c)}} \\
&= \frac{e^x * e^c}{\sum_j e^{x_j} e^c} \\
&= \frac{e^x * e^c}{e^c \sum_j e^{x_j}} \\
&= \frac{e^{x_i}}{\sum_j e^{x_j}} \\
&= softmax(x)_i
\end{aligned}
$$

Exercise 2. (a) Derive the gradients of the sigmoid function and show that it can be rewritten as a function of the function value

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

Exercise 2. (b) Derive the gradient with regard to the inputs of a softmax function when cross entropy loss is used for evaluation

$$\hat{y} - y$$

Exercise 2. (c) Derive the gradients with respect to the inputs x to an one-hidden-layer neural network

$$z_1 = xW_1 + b_1$$

$$z_2 = hW_2 + b_2$$

$$\delta_1 = \frac{\partial CE}{\partial x} = \hat{y} - y$$

$$\delta_2 = \frac{\partial CE}{\partial h} = \delta_1 \frac{\partial z_2}{\partial h} = \delta_1 W_2^T$$

$$\delta_3 = \frac{\partial CE}{\partial z_1} = \delta_2 \frac{\partial h}{\partial z_1}$$

$$\frac{\partial CE}{\partial x} = \delta_3 \frac{\partial z_1}{\partial x} = \delta_3 W_1^T$$

Exercise 2. (d) Parameters in this neural network is (Dx+1)H+(H+1)Dy

Exercise 3. (a) Derive the gradients with respect to vc:

$$\frac{\partial J}{\partial v_c} = U^T \left(\hat{y} - y\right)$$

Exercise 3. (b) As in the previous part, derive gradients for the output word vectors uw

$$\frac{\partial J}{\partial U} = v_c \left(\hat{y} - y\right)^T$$

Exercise 3. (c) Describe with one sentence why this cost function is much more efficient to compute than the softmax-CE loss

$$\frac{\partial J}{\partial v_c} = \left(\sigma\left(u_o^T v_c\right) - 1\right) u_o - \sum_{k=1}^{K} \left(\sigma\left(-u_k^T v_c\right) - 1\right) u_k$$

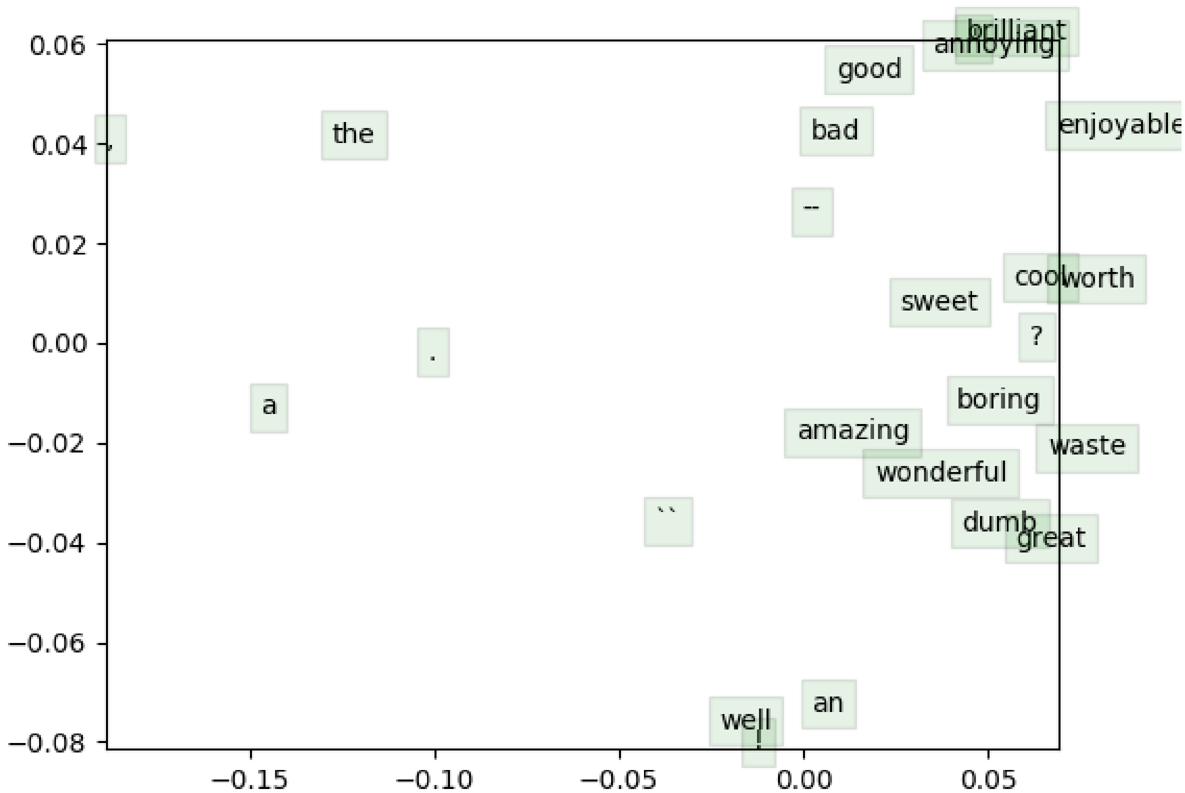
$$\frac{\partial J}{\partial u_o} = \left(\sigma\left(u_o^T v_c\right) - 1\right) v_c$$

$$\frac{\partial J}{\partial u_k} = -\left(\sigma\left(-u_k^T v_c\right) - 1\right) v_c$$

Exercise 3. (d) Derive gradients for all of the word vectors for skip-gram and CBOW given the previous parts and given a set of context words

$$\frac{\partial J_{skip\text{-}gram}\left(Word_{c-m,\dots c+m}\right)}{\partial U} = \sum_{-m<=j<=m, j\neq 0} \frac{\partial F\left(w_{c+j}, v_c\right)}{\partial U}$$

$$\frac{\partial J_{skip\text{-}gram}\left(Word_{c-m,\dots c+m}\right)}{\partial v_c} = \sum_{-m<=j<=m, j\neq 0} \frac{\partial F\left(w_{c+j}, v_c\right)}{\partial v_c}$$



$$\frac{\partial J_{skip\_gram}\left(Word_{c-m,\ldots c+m}\right)}{\partial v_j} = 0, j \neq c$$

$$\frac{\partial J_{CBOM}\left(Word_{c-m,\ldots c+m}\right)}{\partial U} = \sum_{-m<=j<=m, j\neq 0} \frac{\partial F\left(w_c, \hat{u}\right)}{\partial U}$$

$$\frac{\partial J_{CBOW}\left(Word_{c-m,\ldots c+m}\right)}{\partial u_j} = \sum_{-m<=j<=m, j\neq 0} \frac{\partial F\left(w_c, \hat{u}\right)}{\partial \hat{u}}, j \in (c-m, \ldots c+m)$$

$$\frac{\partial J_{skip\_gram}\left(Word_{c-m,\ldots c+m}\right)}{\partial u_j} = 0, j \notin (c-m, \ldots c+m)$$

Exercise 3. (g)