This document proves all the mathematical details needed for skip-gram word2vec model introduced by Mikolov et al. in 2013 which is accessible at arXiv:1310.4546. Also, some of the notations are inspired by the Stanford CS224N class notes lectured by Prof. Christopher D. Manning.

1. Naïve Softmax

1.1 Loss function

For a center word w_c and its context window $\{w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m}\}$ with size m, the predicted probability of observing such window is:

$$\hat{Y} = P(w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m} | w_c)$$

Assuming naïve conditional independence (a.k.a. bi-gram independence):

$$\hat{Y} = P(w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m} | w_c) = \prod_{\substack{j=-m \\ j\neq 0}}^{m} P(w_{c+j}|w_c)$$

The true probability for observing such window is 1. So,

$$Y = \begin{cases} 1 & (w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m} | w_c) \\ 0 & (otherwise|w_c) \end{cases}$$

Thus, the entropy between the two distribution is defined as:

$$H(Y, \hat{Y}) = -1 \times \log(\hat{Y}) - 0 \times \log(1 - \hat{Y}) = -\log(\hat{Y})$$

We can re-write the cross entropy formula:

$$H(Y, \hat{Y}) = -\log(\hat{Y}) = -\log\left(\prod_{\substack{j=-m \\ j\neq 0}}^{m} P(w_{c+j}|w_c)\right) = -\sum_{\substack{j=-m \\ j\neq 0}}^{m} \log\left(P\left(w_{c+j}|w_c\right)\right)$$

Note that in Skip-gram's Naïve Softmax implementation, we define:

$$\hat{y}_{c+j} = P(w_{c+j}|w_c) = P(u_{c+j}|v_c) = \frac{\exp(u_{c+j}^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \quad \forall -m \leq j \leq m, j \neq 0$$

where u and v are d dimensional, i.e., shape $(d, 1)$. They also represent the columns of matrices U and Veach with $(d, |V|)$ shape where they hold the entire context and center word vectors for the entire vocab according to their orders, respectively. Note that we denote the size of the vocabulary by $|V|$.

For a center word w_c and a context word $w_{c+j} \in \{w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m}\}$, we can define the predicted and true probability distributions as well as the cross-entropy as below:

$\hat{y} = [\hat{y}_1, \dots, \hat{y}_{c+j}, \dots, \hat{y}_{|V|}]$: a $(|V|, 1)$ vector. We can also use the vectorized Softmax notation as below:

$\hat{y} = softmax(Uv_c)$

y_{c+j} = one-hot vector of size $(|V|, 1)$ with 1 at position $c + j$

$$H(y_{c+j}, \hat{y}) = -\sum_{w \in Vocab} y_{c+j_w} \log(\hat{y}_w) = \underbrace{-\sum_{\substack{w \in Vocab \\ w \neq c+j}} y_{c+j_w} \log(\hat{y}_w)}_{zero} - \underbrace{y_{c+j_{c+j}}}_{one} \log(\hat{y}_{c+j})$$

$$= -\log(\hat{y}_{c+j}) = -\log\left(\frac{\exp(u_{c+j}^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)$$

where y_{c+j_w} and $\hat{y}_w$ denote the elements of vectors y_{c+j} and $\hat{y}$, respectively.

Note that we can re-write the cross-entropy for a center word's context as below:

$$H(Y, \hat{Y}) = -\sum_{\substack{j=-m \\ j \neq 0}}^{m} \log(\hat{y}_{c+j}) = \sum_{\substack{j=-m \\ j \neq 0}}^{m} H(y_{c+j}, \hat{y})$$

We set the skip-gram loss function to the Naïve Softmax-based cross-entropy:

$$J(v_c, w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m}, U) = H(Y, \hat{Y}) = \sum_{\substack{j=-m \\ j \neq 0}}^{m} H(y_{c+j}, \hat{y})$$

$$= \sum_{\substack{j=-m \\ j \neq 0}}^{m} J_{naive-softmax}(v_c, w_{c+j}, U)$$

Or, simply:

$$J(v_c, w_{c-m}, \dots, w_{c+m}, U) = -\sum_{\substack{j=-m \\ j \neq 0}}^{m} \log(\hat{y}_{c+j})$$

1.2 Gradients:

$$(i)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_c = \sum_{\substack{-m\le j\le m \\ j\ne 0}} \partial J_{naive-softmax}\big(v_c, w_{c+j}, U\big)/\partial v_c$$

$$(ii)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial U = \sum_{\substack{-m\le j\le m \\ j\ne 0}} \partial J_{naive-softmax}(v_c, w_{c+j}, U)/\partial U$$

$$(iii)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_{w\ne c} = \sum_{\substack{-m\le j\le m \\ j\ne 0}} \partial J_{naive-softmax}\big(v_c, w_{c+j}, U\big)/\partial v_w = 0$$

1.2.1 Gradients with respect to center word

$$\begin{aligned}
\frac{\partial J_{naive-softmax}\big(v_c, w_{c+j}, U\big)}{\partial v_c} &= -\frac{\partial \log\big(\hat{y}_{c+j}\big)}{\partial v_c} = -\frac{\partial}{\partial v_c} log\left(\frac{\exp\big(u_{c+j}^T v_c\big)}{\sum_{w\in Vocab}\exp\,(u_w^T v_c)}\right) \\
&= -\frac{\partial}{\partial v_c}\left(u_{c+j}^T v_c - \log\left(\sum_{w\in Vocab}\exp(u_w^T v_c)\right)\right) \\
&= -u_{c+j} + \frac{\frac{\partial}{\partial v_c}\big(\sum_{x\in Vocab}\exp(u_x^T v_c)\big)}{\sum_{w\in Vocab}\exp\,(u_w^T v_c)} \\
&= -u_{c+j} + \sum_{x\in Vocab} u_x \frac{\exp(u_x^T v_c)}{\sum_{w\in Vocab}\exp\,(u_w^T v_c)} = -u_{c+j} + \sum_{x\in Vocab} u_x \hat{y}_x \\
&= -u_{c+j} + U\hat{y}
\end{aligned}$$

$$\partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_c = \sum_{\substack{-m\le j\le m \\ j\ne 0}} [-u_{c+j} + U\hat{y}] = 2m\left(U\hat{y} - \frac{1}{2m}\sum_{\substack{-m\le j\le m \\ j\ne 0}} u_{c+j}\right)$$

$$\partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_{w\ne c} = 0$$

We can now form the gradient matrix $\partial J/\partial V$:

$$\partial J/\partial V = [0 \quad \dots \quad 0 \quad \partial J/\partial v_c \quad 0 \quad \dots \quad 0]_{(d,|V|)}$$

1.2.2 Gradients with respect to the context word and other words

$$\begin{aligned}
\frac{\partial J_{naive-softmax}\big(v_c, w_{c+j}, U\big)}{\partial u_{c+j}} &= -\frac{\partial}{\partial u_{c+j}}\left(u_{c+j}^T v_c - \log\left(\sum_{w\in Vocab}\exp(u_w^T v_c)\right)\right) \\
&= -v_c + \frac{\frac{\partial}{\partial u_{c+j}}\big(\sum_{x\in Vocab}\exp(u_x^T v_c)\big)}{\sum_{w\in Vocab}\exp\,(u_w^T v_c)} = -v_c + v_c\frac{\exp\big(u_{c+j}^T v_c\big)}{\sum_{w\in Vocab}\exp\,(u_w^T v_c)} \\
&= v_c\big(\hat{y}_{c+j} - 1\big)
\end{aligned}$$

$$\frac{\partial J_{naive-softmax}(v_c, w_{c+j}, U)}{\partial u_{k \neq c+j}} = -\frac{\partial}{\partial u_k}\left(u_{c+j}^T v_c - \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right)$$

$$= 0 + \frac{\frac{\partial}{\partial u_k}\left(\sum_{x \in Vocab} \exp(u_x^T v_c)\right)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} = v_c \frac{\exp(u_k^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} = v_c \hat{y}_k$$

Pulling all gradients together:

$$\partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial U = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} [v_c \hat{y}_1, \dots, \underbrace{v_c(\hat{y}_{c+j} - 1)}_{c+j^{th}\ element}, \dots, v_c \hat{y}_{|V|}]$$

$$= \left[2m v_c \hat{y}_1, \dots, \underbrace{\left((2m-1) v_c \hat{y}_{c-m} + v_c(\hat{y}_{c-m} - 1)\right), \dots, \left((2m-1) v_c \hat{y}_{c+m} + v_c(\hat{y}_{c+m} - 1)\right)}_{2m\ context\ words}, \dots, 2m v_c \hat{y}_{|V|}\right]$$

$$= \left[2m v_c \hat{y}_1, \dots, \underbrace{v_c(2m\hat{y}_{c-m} - 1), \dots, v_c(2m\hat{y}_{c+m} - 1)}_{2m\ context\ words}, \dots, 2m v_c \hat{y}_{|V|}\right]$$

2. Negative Sampling

2.1 Loss function

For a center word w_c and its context window $\{w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m}\}$ with size m, we define the negative sample sets as $A_{c-m}, \dots, A_{c-1}, A_{c+1}, \dots, A_{c+m}$ where:

$A_{c+j} = \{$ K randomly chosen indices from 1 to $|V|$ excl. $c + j\}$ for $-m \leq j \leq m, j \neq 0$

In this method, we are again interested in the probability of observing the $2m$ context words given the corresponding center word. However, we use an auxiliary binary classification, treating the training context words as positive examples and samples from a noise distribution $P_n(w)$as negative examples. We will use the unigram distribution of the training data as the noise distribution. Assume that noise samples are K times more frequent than data samples.

Given a center word w_c, we define the following random variable with Bernoulli distribution :

$$D = \begin{cases} 1, & w \sim P(w|w_c)\ context\ word\ pdf \\ 0, & w \sim P_n(w) \end{cases}$$

$$\hat{y}_w = P(D = 1|w_c, w) = \frac{P(w|D = 1, w_c)P(D = 1|w_c)}{P(w|D = 1, w_c)P(D = 1|w_c) + P(w|D = 0, w_c)P(D = 0|w_c)}$$

$$= \frac{1}{1 + \frac{P(w|D = 0, w_c)P(D = 0|w_c)}{P(w|D = 1, w_c)P(D = 1|w_c)}} = \frac{1}{1 + K\frac{P(w|\ w_c)}{P_n(w)}} =$$

$$= \sigma\left(\log\left(\frac{P(w|\ w_c)}{KP_n(w)}\right)\right) \overset{Negative\ Sampling}{\triangleq} \sigma(u_w^T v_c)$$

Therefore, if we can find $\hat{Y}$, and because we know $P_n(w)$ and K, we can find $P(w|\, w_c)$ which is the probability of observing the context word w given the corresponding center word w_c. Note that mixture model probability:

$$\begin{aligned} P_{mixture}(w|\, w_c) &= P(w|\, w_c, D=1)\, P(D=1|w_c) + P(w|\, w_c, D=0)\, P(D=0|w_c) \\ &= P(w|\, w_c)\, 1/(K+1) + P_n(w)\, K/(K+1) \end{aligned}$$

We want to minimize the binary classification loss function:

$$\begin{aligned} -\mathbb{E}_{P_{mixture}}[\log P(D|w, w_c)] &= -(\mathbb{E}_P[\log(P(D=1|w_c, w))] + K\mathbb{E}_n[\log(P(D=0|w_c, w))]) \\ &= -\left(\log(\hat{y}_w) + \sum_{k\in A_w} \log(1-\hat{y}_k)\right) \\ &= -\left(\log(\sigma(u_w^T v_c)) + \sum_{k\in A_w} \log(1-\sigma(u_k^T v_c))\right) = J_{neg-sample}(v_c, w, U) \end{aligned}$$

And the loss function for all 2m context words is (assuming bigram independence):

$$J(v_c, w_{c-m}, \dots, w_{c+m}, U) = -\sum_{\substack{j=-m \\ j\neq 0}}^{m} \left[\log(\hat{y}_{c+j}) + \sum_{k\in A_{c+j}} \log(1-\hat{y}_k)\right]$$

2.2 Gradients:

$$(i)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_c = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \partial J_{neg-sample}(v_c, w_{c+j}, U)/\partial v_c$$

$$(ii)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial U = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \partial J_{neg-sample}(v_c, w_{c+j}, U)/\partial U$$

$$(iii)\ \partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_{w\neq c} = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \partial J_{neg-sample}(v_c, w_{c+j}, U)/\partial v_w = 0$$

2.2.1 Gradients with respect to center word

$$\frac{\partial J_{neg-sample}(v_c, w_{c+j}, U)}{\partial v_c} = -\frac{\partial}{\partial v_c}\left(\log\left(\sigma\left(u_{c+j}^T v_c\right)\right) + \sum_{k \in A_{c+j}} \log\left(\sigma(-u_k^T v_c)\right)\right)$$

$$= -\left(\frac{\partial \log\left(\sigma\left(u_{c+j}^T v_c\right)\right)}{\partial v_c} + \sum_{k \in A_{c+j}} \frac{\partial \log\left(\sigma(-u_k^T v_c)\right)}{\partial v_c}\right)$$

$$= -\left(\frac{\partial \sigma\left(u_{c+j}^T v_c\right)/\partial v_c}{\sigma\left(u_{c+j}^T v_c\right)} + \sum_{k \in A_{c+j}} \frac{\partial \sigma(-u_k^T v_c)/\partial v_c}{\sigma(-u_k^T v_c)}\right)$$

$$= -\left(\frac{u_o \sigma\left(u_{c+j}^T v_c\right)\left(1 - \sigma\left(u_{c+j}^T v_c\right)\right)}{\sigma\left(u_{c+j}^T v_c\right)} + \sum_{k \in A_{c+j}} \frac{-u_k \sigma(-u_k^T v_c)\left(1 - \sigma(-u_k^T v_c)\right)}{\sigma(-u_k^T v_c)}\right)$$

$$= -\left(u_{c+j}\left(1 - \sigma\left(u_{c+j}^T v_c\right)\right) - \sum_{k \in A_{c+j}} u_k\left(1 - \sigma(-u_k^T v_c)\right)\right)$$

We know that:

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

So,

$$\frac{\partial J_{neg-sample}(v_c, w_{c+j}, U)}{\partial v_c} = -u_{c+j}\sigma\left(-u_{c+j}^T v_c\right) + \sum_{k \in A_{c+j}} u_k \sigma(u_k^T v_c)$$

$$= u_{c+j}(\hat{y}_{c+j} - 1) + \sum_{k \in A_{c+j}} u_k \hat{y}_k$$

$$\partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_c = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \left[u_{c+j}(\hat{y}_{c+j} - 1) + \sum_{k \in A_{c+j}} u_k \hat{y}_k\right]$$

$$\partial J(v_c, w_{c-m}, \dots, w_{c+m}, U)/\partial v_{w \ne c} = 0$$

We can now form the gradient matrix $\partial J/\partial V$:

$$\partial J/\partial V = [0 \quad \dots \quad 0 \quad \partial J/\partial v_c \quad 0 \quad \dots \quad 0]_{(d,|V|)}$$

2.2.2 Gradients with respect to the context word and other words

$$\frac{\partial J_{neg-sample}\left(v_c, w_{c+j}, U\right)}{\partial u_{c+j}} = -\frac{\partial}{\partial u_{c+j}}\left(\log\left(\sigma\left(u_{c+j}^T v_c\right)\right) + \sum_{k\in A_{c+j}} \log\left(\sigma\left(-u_k^T v_c\right)\right)\right)$$

$$= -\left(\frac{\partial \log\left(\sigma\left(u_{c+j}^T v_c\right)\right)}{\partial u_{c+j}} + \underbrace{\sum_{k\in A_{c+j}} \frac{\partial \log\left(\sigma\left(-u_k^T v_c\right)\right)}{\partial u_{c+j}}}_{zero}\right) = -\frac{\partial\sigma\left(u_{c+j}^T v_c\right)/\partial u_{c+j}}{\sigma\left(u_{c+j}^T v_c\right)}$$

$$= -\frac{v_c \sigma\left(u_{c+j}^T v_c\right)\left(1-\sigma\left(u_{c+j}^T v_c\right)\right)}{\sigma\left(u_{c+j}^T v_c\right)} = -v_c\left(1-\sigma\left(u_{c+j}^T v_c\right)\right) = -v_c\sigma\left(-u_{c+j}^T v_c\right)$$

$$= v_c\left(\hat{y}_{c+j} - 1\right)$$

$$\frac{\partial J_{neg-sample}\left(v_c, w_{c+j}, U\right)}{\partial u_{k\in A_{c+j}}} = -\frac{\partial}{\partial u_k}\left(\log\left(\sigma\left(u_{c+j}^T v_c\right)\right) + \sum_{x\in A_{c+j}} \log\left(\sigma\left(-u_x^T v_c\right)\right)\right)$$

$$= -\left(\underbrace{\frac{\partial \log\left(\sigma\left(u_{c+j}^T v_c\right)\right)}{\partial u_k}}_{zero} + \sum_{x\in A_{c+j}} \frac{\partial \log\left(\sigma\left(-u_x^T v_c\right)\right)}{\partial u_k}\right) = -n_k \frac{\partial\sigma\left(-u_k^T v_c\right)/\partial u_k}{\sigma\left(-u_k^T v_c\right)}$$

$$= -n_k \frac{-v_c\sigma\left(-u_k^T v_c\right)\left(1-\sigma\left(-u_k^T v_c\right)\right)}{\sigma\left(-u_k^T v_c\right)} = n_k v_c \sigma\left(u_k^T v_c\right) = n_k v_c \hat{y}_k$$

Where n_k is the number of time word w_k is negatively sampled and appeared in A_{c+j}.

$$\frac{\partial J_{neg-sample}\left(v_c, w_{c+j}, U\right)}{\partial u_{i\notin A_{c+j}\,\&\, i\neq c+j}} = 0$$

Pulling all together:

$$\partial J/\partial U = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \left[\partial J/\partial u_1 \quad \dots \quad \partial J/\partial u_{|V|}\right]_{(d,|V|)}$$

where :

$$\partial J/\partial u_i = \begin{cases} v_c\left(\hat{y}_i - 1\right) & i = c+j \\ n_k v_c \hat{y}_i & i \in A_{c+j} \\ 0 & otherwise \end{cases}$$

END.