This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

U is the outside word matrix composed of u_o outside vectors for each outside word.
V is the center word matrix composed of v_c center vectors for each center word.
Both U and V contains a vector for each word w in the vocabulary.
U : (embedsize X numtokens)
V : (embedsize X numtokens)

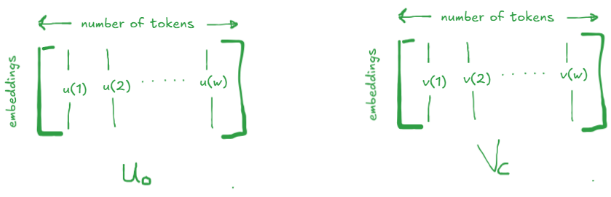

Figure 1: U V matrix

u_o is the column o of U

v_c is the column c of V

Cross-entropy "loss" between the true probability distribution y and predicted distribution $\hat{y}$ is represented as:
$-\sum_i ylog(\hat{y})$ ... (1)
were y and $\hat{y}$ are vectors of length equal to the number of words in the vocabulary.

y : (num-tokens X 1)
$\hat{y}$: (num-tokens X 1)

Probability of an outside word o given a center word c is the softmax function:

$\hat{y}_o = P(O = oC = c) = \frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)}$ ... (2)

$J_{native-softmax}(v_c, o, U) = -\sum_{w \epsilon Vocab} y_w log(\hat{y_o})$ ... (3)

For a given center word v_c there is only one outside word u_o which is the k^{th} word in the vocabulary, where $y_k = 1$. Hence the $y_i = 0$ where i is not equal to k.

$J_{native-softmax}(v_c, o, U) = -y_k log(y_o\hat{(}k))$ ... (4)

$J_{native-softmax}(v_c, o, U) = -\log(\hat{y_o})$ ... (5)

Substituting (2) to (5)

$J_{native-softmax}(v_c, o, U) = -\log(\frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)})$ .... (6)

$J_{native-softmax}(v_c, o, U) = -\log(exp(u_o^T v_c)) + log(\sum_w exp(u_w^T v_c))$ .... (7)

$J_{native-softmax}(v_c, o, U) = -u_o^T v_c + log(\sum_w exp(u_w^T v_c))$ .... (7)

Partial derivative of $J_{native-softmax}$ with respect to v_c

$\frac{\partial J}{\partial v_c} = -u_o + \frac{\partial (log(\sum_w exp(u_w^T v_c)))}{\partial v_c}$... (8)

Derivative of a nested functions (by chain rule):

$\frac{df(g(x))}{dx} = \frac{d(f(g(x))}{d(g(x))} . \frac{d(g(x))}{dx}$

$f(g(v_c)) = log(g(v_c))$
$g(v_c) = \sum_w exp(u_w^T v_c)$
$\frac{d(f(g(v_c))}{d(g(v_c))} = \frac{1}{g(v_c)}$

$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_w exp(u_w^T v_c)} . \frac{\partial (\sum_w exp(u_w^T v_c))}{\partial v_c}$ ... (9)

$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_w exp(u_w^T v_c)} . \sum_w exp(u_w^T v_c) \frac{\partial ((u_w^T v_c))}{\partial v_c}$ ... (10)

$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_w exp(u_w^T v_c)} . \sum_w exp(u_w^T v_c) u_w^T$ ... (11)

$\frac{\partial J}{\partial v_c} = -u_o + \sum_w \frac{exp(u_w^T v_c)}{\sum_w exp(u_w^T v_c)} . u_w^T$ ... (12)

Substituting (2)

$\frac{\partial J}{\partial v_c} = -u_o + \sum_w^V \hat{y}_w . u_w^T$ ... (12)

$\frac{\partial J}{\partial v_c} = -u_o + \sum_w^V \hat{y}_w . u_w$ ... (13)

V is the number of words in vocabulary.

1.b

$J_{native-softmax}(v_c, o, U) = -u_o^T v_c + log(\sum_w exp(u_w^T v_c))$ .... (1)

$\frac{\partial J}{\partial u_w} = -\frac{\partial u_o^T v_c}{\partial u_w} + \frac{\partial log(\sum_w exp(u_w^T v_c)}{\partial u_w}$ ... (2)

$\frac{\partial J}{\partial u_w} = -\frac{\partial u_o^T v_c}{\partial u_w} + \frac{1}{\sum_w exp(u_w^T v_c)} \frac{\partial(\sum_w exp(u_w^T v_c)}{\partial u_w}$ ... (3)

$\frac{\partial J}{\partial u_w} = -\frac{\partial u_o^T v_c}{\partial u_w} + \frac{\sum_w exp(u_w^T v_c)}{\sum_w exp(u_w^T v_c)} \frac{\partial(u_w^T v_c)}{\partial u_w}$ ... (4)

$\frac{\partial J}{\partial u_w} = -\frac{\partial u_o^T v_c}{\partial u_w} + \frac{exp(u_w^T v_c)}{\sum_w exp(u_w^T v_c)} v_c$ ... (5)

$\frac{\partial J}{\partial u_w} = -\frac{\partial u_o^T v_c}{\partial u_w} + \hat{y}_w v_c$ ... (6)

When w = o

$\frac{\partial J}{\partial u_w} = -v_c + \hat{y}_w v_c = (\hat{y}_w - 1) v_c$ ... (7)

When w = otherwise

$\frac{\partial J}{\partial u_w} = \hat{y}_w v_c$ ... (8)