(1) a) $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -y_o \log(\hat{y}_o) = -1 \log(\hat{y}_o) = -\log(\hat{y}_o)$

This happens because the true empirical distribution y is a one-hot vector with a 1 for the true outside word o, and 0 (zero) everywhere else.

b) $\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)\right) =$

$= \frac{\partial}{\partial v_c}\left(-\log(\exp(u_o^T v_c)) + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right) =$

$= \frac{\partial}{\partial v_c}\left(-u_o^T v_c + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right) = -u_o +$

$+ \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \times \frac{\partial}{\partial v_c}\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right) =$

$= -u_o + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \times \left(\sum_{w \in Vocab} \exp(u_w^T v_c)\, u_w\right) =$

$= \sum_{w \in Vocab} \frac{\exp(u_w^T v_c)\, u_w}{\sum_{w \in Vocab} \exp(u_w^T v_c)} - u_o \Leftarrow$ since we know that

$\hat{y}_o = \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}$ and $U y_o = u_o \Leftarrow \left(\sum_{w \in Vocab} \hat{y}_w u_w\right) - U y_o =$

where o is any word in the vocab

$= U\hat{y} - Uy = U(\hat{y} - y)$ [assuming that the y is for the outside word o]

1) The gradient is zero when the model perfectly classifies the context word, meaning that it assigns probability of 1 to it.

2) Subtracting the gradient of the loss function with respect to the v_c vector from the v_c vector moves that vector in the direction of the negative gradient, which is the direction of the steepest descent, and thus maximally minimizes the loss function.

c)

$$\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial u_w} = \frac{\partial}{\partial u_w}\left(-\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}\right)\right) =$$

$(=)$ based on b) $(=) \frac{\partial}{\partial u_w}\left(-u_o^T v_c + \log\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)\right)$

when $w = o$,

$$\frac{\partial}{\partial u_o}\left(-u_o^T v_c + \log\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)\right) = \left(-v_c + \frac{1}{\sum_{w \in vocab} \exp(u_w^T v_c)} \times \right.$$

$$\left. \times \left(\exp(u_o^T v_c) v_c\right)\right) = \frac{\exp(u_o^T v_c) v_c}{\sum_{w \in vocab} \exp(u_w^T v_c)} - v_c =$$

$$= \hat{y}_o v_c - v_c = (\hat{y}_o - 1) v_c$$

when $w \neq o$,

$$\frac{\partial}{\partial u_w}\left(-u_o^T v_c + \log\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)\right) = \frac{1}{\sum_{w \in vocab} \exp(u_w^T v_c)} \times$$

$$\times \left(\exp(u_w^T v_c) v_c\right) = \frac{\exp(u_w^T v_c) v_c}{\sum_{w \in vocab} \exp(u_w^T v_c)} = \hat{y}_w v_c$$

d) In c), we calculated the derivatives of $J_{naive\text{-}softmax}(v_c, o, U)$ with respect to each of the 'outside' word vectors, u_w's.
Since those u_w vectors represent the colums of the U matrix, we can just reuse our answers from c) here.

So, $\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial U} = [\hat{y}_1 v_c \;\; \hat{y}_2 v_c \;\; \dots \;\; (\hat{y}_o - 1) v_c \;\; \dots \;\; \hat{y}_{|vocab|} v_c] \in \mathbb{R}^{d \times |vocab|}$,

$= v_c(\hat{y} - y)^T$

where d is the dimension of the embeddings

e) $f(x) = \max(0, x)$

when $x < 0$, $f'(x) = 0$

when $x > 0$, $f'(x) = 1$

f) $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$, $\sigma'(x) = \frac{e^x(e^x+1) - e^{2x}}{(e^x+1)^2} = \frac{e^x}{(e^x+1)^2} =$

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

g) (i) with respect to v_c

$\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T v_c))\right) =$

$= -\frac{1}{\sigma(u_o^T v_c)} \times \left(1 - \sigma(u_o^T v_c)\right)\sigma(u_o^T v_c) u_o + \sum_{s=1}^{K} \frac{(1-\sigma(-u_{w_s}^T v_c))\sigma(-u_{w_s}^T v_c) u_{w_s}}{\sigma(-u_{w_s}^T v_c)}$

$= -\left(1 - \sigma(u_o^T v_c)\right) u_o + \sum_{s=1}^{K} (1 - \sigma(-u_{w_s}^T v_c)) u_{w_s}$

with respect to u_o

$\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial u_o} = \frac{\partial}{\partial u_o}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T v_c))\right) =$

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

(with respect to u_{w_s})

$$\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial u_{w_s}} = \frac{\partial}{\partial u_{w_s}}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c))\right) =$$

$$= -\frac{1}{\sigma(-u_{w_s}^T v_c)} \times \left((1-\sigma(-u_{w_s}^T v_c))\,\sigma(-u_{w_s}^T v_c)\right) \times (-v_c) =$$

$$= (1-\sigma(-u_{w_s}^T v_c))\, v_c$$

(ii) To minimize duplication, we can reuse the quantities $(1-\sigma(u_o^T v_c))$ and/or $(1-\sigma(-u_{w_s}^T v_c))$ $[s\in[1,\dots,K]]$ We calculated those quantities to find the value of $\partial J/\partial v_c$. Now, we can express $\partial J/\partial v_c$ and $\partial J/\partial U$ in the following way.

$$\partial J/\partial v_c = -U_{o,[w_1,\dots w_K]}\left(1-\sigma(U_{o,[w_1,\dots w_K]}^T v_c)\right) \in R^{d\times 1}$$

$$\partial J/\partial U = -\underbrace{sign(U_{o,[w_1,\dots w_K]})}_{\text{this is done to resolve the discrepancy between the gradients with respect to } u_o \text{ and } u_{w_s},\ w\in[1,\dots K]} \otimes v_c\left(1-\sigma(U_{o,[w_1,\dots w_K]}^T v_c)\right)^T \in R^{d\times|K+1|}$$

sing function; $\otimes$: element-by-element multiplication

(iii) What makes the Negative Sampling loss efficient is that it only involves a fixed K number (much smaller than the vocab size) of words in its computation in addition with the true outside word, unlike soft max loss.

h) $$\frac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial u_{w_s}} = \frac{\partial}{\partial u_{w_s}}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K}\log(\sigma(-u_{w_s}^T v_c))\right)$$

$$-\frac{\partial}{\partial u_{w_s}}\left(\sum_{\substack{a\in[1,K]\\ a\neq s}}\log(\sigma(-u_{w_a}^T v_c)) + \sum_{\substack{a\in[1,K]\\ a=s}}\log(\sigma(-u_{w_a}^T v_c))\right) =$$

$$= \sum_{\substack{a\in[1,K]\\ a=s}}(1-\sigma(-u_{w_a}^T v_c))\, v_c$$

(i) $\frac{\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \dots w_t, \dots w_{t+m}, U)}{\partial U} =$

$$= \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial j(v_c, w_{t+j}, U)}{\partial U}$$

(ii) $\frac{\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \dots, w_t, \dots w_{t+m}, U)}{\partial v_c} =$

$$= \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial j(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii) $\frac{\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \dots, w_t, \dots v_{t+m}, U)}{\partial v_w} =$

$= 0$ (when $w \neq c$)

(2) c) In the graph, we can see that a number of words, which have similar meanings or have closeness in terms of a general topic, ended up being close to each other. Clear examples are (1) woman, female and man; (2) amazing, wonderful and great; (3) tea, coffee, sweet; (4) rain and snow. However, the word "male" did not end up being close to the cluster (1) and the word "hail" was also too far from the words "rain" and "snow", which are expected to be close.