a). Because y is a hot vectors with 0 if $w \neq o$ and 1 if $w = o$.

Therefore $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -(y_1 \log(\hat{y}_1) + y_2 \log(\hat{y}_2) + \dots + y_o \log(\hat{y}_o) + \dots)$

$= -y_o \log(\hat{y}_o)$

$\because y_o = 1$ when $w = o$

$\therefore = -\log(\hat{y}_o)$

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

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

$$= -\frac{\partial}{\partial v_c}(u_o^T v_c) + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \cdot \sum_{x \in Vocab} u_x \cdot \exp(u_x^T v_c)$$

$$= -u_o + \sum_{x \in Vocab_x} \frac{\exp(u_x^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \cdot u_x$$

$$= -u_o + \sum_{x \in Vocab} P(u_x | v_c) \cdot u_x$$

$$= -u_o + \sum_{x \in Vocab} \hat{y}_x \cdot u_x$$

c). case 1: $w = o$

$$\frac{\partial J}{\partial u_{w=o}} = \frac{-\partial}{\partial u_{w=o}}(\log(\exp(u_o^T v_c)) - \log \sum_{w \in Vocab}((\exp(u_w^T v_c)))$$

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

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

$$= -v_c + \hat{y}_o v_c$$

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

case 2: $w \neq o$

$$\frac{\partial J}{\partial u_{w \neq o}} = \frac{-\partial}{\partial u_{w \neq o}}(\log(\exp(u_o^T v_c)) - \log \sum_{w \in Vocab}((\exp(u_w^T v_c)))$$

$$= 0 - \frac{\exp(u_{w \neq o}^T v_c)}{\sum_{w \in Vocab}(\exp(u_w^T v_c))} \cdot v_c$$

$$= \hat{y}_{w \neq o} v_c$$

d). $$\frac{\partial \sigma(x)}{\partial x} = \frac{1}{(1+e^{-x})^2} \cdot -e^{-x}$$

$$= -\frac{e^{-x}}{(1+e^{-x})^2}$$

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

$$= \sigma(x) \cdot (1-\sigma(x))$$

e). 1. $$\frac{\partial J}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \cdot (\sigma(u_o^T v_c) \cdot (1-\sigma(u_o^T v_c)) \cdot u_o$$
$$+ \sum_{k=1}^{K} \frac{1}{\sigma(u_k^T v_c)} \cdot (\sigma(u_k^T v_c) \cdot (1-\sigma(u_k^T v_c)) \cdot u_k$$
$$= -(1-\sigma(u_o^T v_c)) \cdot u_o + \sum_{k=1}^{K} (1-\sigma(u_k^T v_c)) u_k$$

2. $$\frac{\partial J}{\partial u_o} = -v_c \frac{1}{\sigma(u_o^T v_c)} \cdot (\sigma(u_o^T v_c) \cdot (1-\sigma(u_o^T v_c)) - 0$$
$$= -v_c \cdot (1-\sigma(u_o^T v_c))$$

3. $$\frac{\partial J}{\partial u_k} = \sum_{k=1}^{K} v_c \cdot (1-\sigma(-u_k^T v_c))$$

Instead of calculating all the words in the Vocab, we now only need to calculate k samples.

f). i) $$\frac{\partial J_{skip}(v_c, w_{t-m}, \dots w_{t+m}, U)}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

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

iii) $= 0$