1. a) $J_{\text{naive-softmax}} = -\log P(O=o \mid C=c) = -\log \hat{y}_o = -\sum_w y_w \log \hat{y}_w$

Since $y_w = \begin{cases} 1 & \text{if } w=o \\ 0 & \text{otherwise} \end{cases}$.

b) $\frac{\partial J_{\text{naive-softmax}}}{\partial v_c} = -\frac{\partial}{\partial v_c} \log \hat{y}_o = -\frac{\partial}{\partial v_c} \log \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}$

$= -\frac{\partial}{\partial v_c} \left(u_o^T v_c - \log \sum_w e^{u_w^T v_c} \right)$

$= -u_o + \frac{1}{\sum_w e^{u_w^T v_c}} \sum_w e^{u_w^T v_c} u_w$

$= -u_o + \sum_w \hat{y}_w u_w = -U y + U \hat{y}$

$= U(\hat{y} - y)$ where columns of U are outside word vectors

c) $\frac{\partial J}{\partial u_w} = -\frac{\partial}{\partial u_w} \left(u_o^T v_c - \log \sum_w e^{u_w^T v_c} \right)$

$= -1(w=o) v_c + \frac{1}{\sum_x e^{u_x^T v_c}} e^{u_w^T v_c} v_c$

$= -1(w=o) v_c + \hat{y}_w v_c$

$= (\hat{y}_w - 1(w=o)) v_c$

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

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

$= \sigma(x)\left(1 - \sigma(x)\right)$

e) $\frac{\partial J_{neg\text{-}sample}}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c)\left(1-\sigma(u_o^T v_c)\right) u_o$

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

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

$\frac{\partial J_{neg\text{-}sample}}{\partial u_o} = -\left(\sigma(u_o^T v_c) - 1\right) v_c$

$\frac{\partial J_{neg\text{-}sample}}{\partial u_k} = -\left(\sigma(-u_k^T v_c) - 1\right) v_c$

Only need to calculate $\sigma(u_w^T v_c)$ for $k << |Vocab|$ word vectors, so $J_{neg\text{-}sample}$ is much more efficient to calculate than $J_{naive\text{-}softmax}$