a)

$$y_w = \begin{cases} 1, w = o \\ 0, w \not= o \end{cases}$$

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

b)

$$\begin{aligned}
\frac{\partial}{\partial v_c} J_{naive-softmax} &= - \frac{\partial}{\partial v_c} \log P(O = o | C = c) \\
&= - \frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \\
&= - \frac{\partial}{\partial v_c} \log \exp(u_o^T v_c) + \frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} \exp(u_w^T v_c) \\
&= - u_o + \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \frac{\partial}{\partial v_c} \sum_{x=1}^{V} \exp(u_x^T v_c) \\
&= - u_o + \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \sum_{x=1}^{V} \exp(u_x^T v_c) \frac{\partial}{\partial v_c} u_x^T v_c \\
&= - u_o + \frac{1}{\sum_{w=1}^{V} \exp(u_w^T v_c)} \sum_{x=1}^{V} \exp(u_x^T v_c) u_x \\
&= - u_o + \sum_{x=1}^{V} \frac{\exp(u_x^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)} u_x \\
&= - u_o + \sum_{x=1}^{V} P(O = x | C = c) u_x \\
&= - y^T U^T + \hat{y}^T u^T \\
&= U(\hat{y} - y)
\end{aligned}$$

c)

$$
\begin{aligned}
\frac{\partial}{\partial u_w} J_{naive-softmax} &= -\frac{\partial}{\partial u_w} \log \frac{\exp(u_o^T v_c)}{\sum_{m=1}^{V} \exp(u_m^T v_c)} \\
&= -\frac{\partial}{\partial u_w} \log \exp(u_o^T v_c) + \frac{\partial}{\partial u_w} \log \sum_{m=1}^{V} \exp(u_m^T v_c)
\end{aligned}
$$

When $w = o$:

$$
\begin{aligned}
\frac{\partial}{\partial u_o} J_{naive-softmax} &= -v_c + \frac{1}{\sum_{m=1}^{V} \exp(u_m^T)} \sum_{n=1}^{V} \frac{\partial}{\partial u_o} \exp(u_n^T v_c) \\
&= -v_c + \frac{1}{\sum_{m=1}^{V} \exp(u_m^T)} \frac{\partial}{\partial u_o} \exp(u_o^T v_c) \\
&= -v_c + \frac{\exp(u_o^T v_c)}{\sum_{m=1}^{V} \exp(u_m^T)} v_c \\
&= -v_c + P(O = o | C = c) v_c \\
&= (P(O = o | C = c) - 1) v_c
\end{aligned}
$$

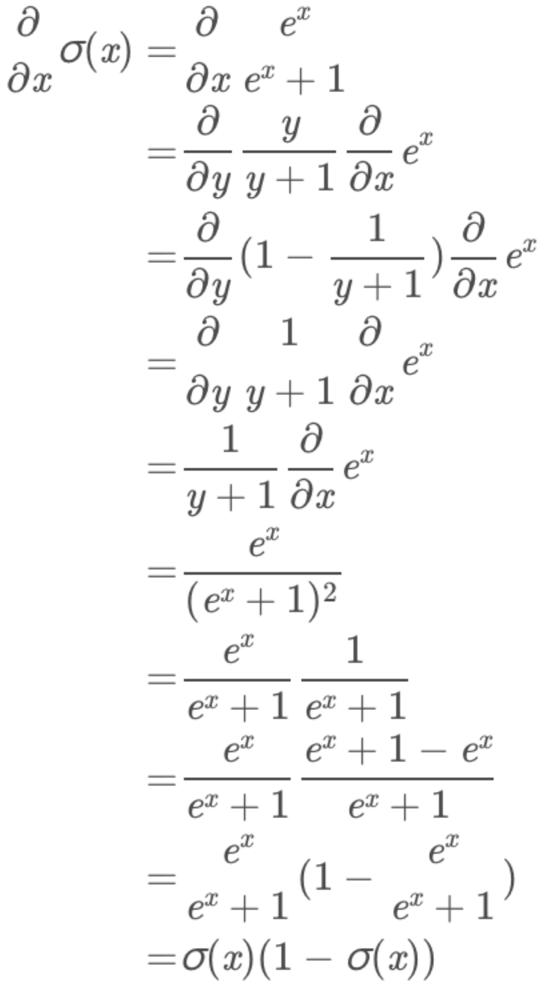
When $w \not= o$:

$$
\begin{aligned}
\frac{\partial}{\partial u_w} J_{naive-softmax} &= \frac{\partial}{\partial u_w} \log \sum_{m=1}^{V} \exp(u_m^T v_c) \\
&= \frac{\exp(u_w^T v_c)}{\sum_{m=1}^{V} \exp(u_m^T)} v_c \\
&= P(O = w | C = c) v_c \\
&= (P(O = o | C = c) - 0) v_c
\end{aligned}
$$

In summary:

$$
\frac{\partial}{\partial u_w} J_{naive-softmax} = (\hat{y}_w - y_w) v_c
$$

d)



$$
\begin{aligned}
\frac{\partial}{\partial x}\sigma(x) &= \frac{\partial}{\partial x}\frac{e^x}{e^x+1} \\
&= \frac{\partial}{\partial y}\frac{y}{y+1}\frac{\partial}{\partial x}e^x \\
&= \frac{\partial}{\partial y}(1-\frac{1}{y+1})\frac{\partial}{\partial x}e^x \\
&= \frac{\partial}{\partial y}\frac{1}{y+1}\frac{\partial}{\partial x}e^x \\
&= \frac{1}{y+1}\frac{\partial}{\partial x}e^x \\
&= \frac{e^x}{(e^x+1)^2} \\
&= \frac{e^x}{e^x+1}\frac{1}{e^x+1} \\
&= \frac{e^x}{e^x+1}\frac{e^x+1-e^x}{e^x+1} \\
&= \frac{e^x}{e^x+1}(1-\frac{e^x}{e^x+1}) \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}
$$

e)

$$\begin{aligned}
\frac{\partial}{\partial v_c} J_{neg-sample} &= -\frac{\partial}{\partial v_c}\log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial v_c}\sum_{k=1}^{K}\log(\sigma(-u_k^T v_c)) \\
&= -\frac{1}{\sigma(u_o^T v_c)}\frac{\partial}{\partial v_c}\sigma(u_o^T v_c) - \sum_{k=1}^{K}\frac{1}{\sigma(-u_k^T v_c)}\frac{\partial}{\partial v_c}\sigma(-u_k^T v_c) \\
&= -\frac{1}{\sigma(u_o^T v_c)}\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))\frac{\partial}{\partial v_c}u_o^T v_c - \sum_{k=1}^{K}\frac{1}{\sigma(-u_k^T v_c)}\sigma(-u_k^T v_c)(1-\sigma(u_k^T v_c))\frac{\partial}{\partial v_c}(-u_k^T v_c) \\
&=(\sigma(u_o^T v_c)-1)u_o - \sum_{k=1}^{K}(\sigma(-u_k^T v_c)-1)u_k
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial u_o} J_{neg-sample} &= -\frac{\partial}{\partial u_o}\log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_o}\sum_{k=1}^{K}\log(\sigma(-u_k^T v_c)) \\
&= -\frac{\partial}{\partial u_o}\log(\sigma(u_o^T v_c)) \\
&= -\frac{1}{\sigma(u_o^T v_c)}\frac{\partial}{\partial u_o}\sigma(u_o^T v_c) \\
&= -\frac{1}{\sigma(u_o^T v_c)}\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))\frac{\partial}{\partial u_o}u_o^T v_c \\
&=(\sigma(u_o^T v_c)-1)v_c
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial u_k} J_{neg-sample} &= -\frac{\partial}{\partial u_k}\log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_k}\sum_{x=1}^{K}\log(\sigma(-u_x^T v_c)) \\
&= -\frac{\partial}{\partial u_k}\sum_{x=1}^{K}\log(\sigma(-u_x^T v_c)) \\
&= -\frac{\partial}{\partial u_k}\log(\sigma(-u_k^T v_c)) \\
&= -\frac{1}{\sigma(-u_k^T v_c)}\frac{\partial}{\partial u_k}\sigma(-u_k^T v_c) \\
&= -\frac{1}{\sigma(-u_k^T v_c)}\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))\frac{\partial}{\partial u_k}(-u_k^T v_c) \\
&=(1-\sigma(-u_k^T v_c))v_c
\end{aligned}$$

f)

$$\frac{\partial}{\partial U} J_{skip-gram}(v_c, w_{t-m}, \ldots w_{t+m}, U) = \sum_{-m \leq j \leq m} \frac{J(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial}{\partial v_c} J_{skip-gram(v_c, w_{t-m,} \ldots w_{t+m}, U)} = \sum_{-m \leq j \leq m} \frac{J(v_c, w_{t+j}, U)}{\partial v_c}$$

$$\frac{\partial}{\partial v_w} J_{skip-gram(v_c, w_{t-m}, \ldots w_{t+m}, U)} =0$$