1.

a) Since y is a one-hot vector, the summation over all words is a summation of $w=o$ only

$$-\sum_{w\in Voc} y_w \log(\hat{y}_w) = -[y_1\log(\hat{y}_1), \dots, y_o\log(\hat{y}_o), \dots y_w\log(\hat{y}_w)] = -1\cdot\log(\hat{y}_o)$$
$$= -\log(\hat{y}_o)$$

b) $$\frac{\partial J}{\partial v_c} -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w\in Voc}\exp(u_w^T v_c)}\right) = -\frac{\partial}{\partial v_c}\left(\frac{u_o^T v_c}{\log\sum_w \exp(u_w^T v_o)}\right)$$
$$= \frac{\partial}{\partial v_c}(-u_o^T v_c) + \frac{1}{\sum_w \exp(u_w^T v_o)}\cdot\sum_w \exp(u_w^T v_c)\frac{\partial}{\partial v_c}u_k$$
$$= -u_o + \sum_{x\in Voc}\frac{\exp(u_x^T v_c)}{\sum_w \exp(u_w^T v_j)}\cdot u_x$$
$$= -u_o + \sum_x u_x \frac{\exp(u_x^T v_o)}{\sum_w \exp(u_w^T v_o)}$$
$$= -u_o + \sum_x u_x\, p(x|c)$$
$$= U(\hat{y}-y)$$

c) When $w\neq o$

$$\frac{\partial}{\partial u_w} -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w\in Voc}\exp(u_w^T v_c)}\right) = \frac{\log(\sum_w \exp(u_w^T v_c))}{\partial u_w}$$
$$= \frac{1}{\sum_w \exp(u_w^T v_c)}\cdot\sum_x \exp(u_w^T v_c)\cdot v_c^T$$
$$= \sum_x \frac{\exp(u_w^T v_c)}{\exp(u_w^T v_c)} v_c^T$$
$$= \hat{y} v_c^T$$

$$w=o \Rightarrow -v_c^T + \hat{y}v_c^T$$
$$= (\hat{y}-y)v_c^T$$

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

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

LHS: $\frac{1}{\sigma(u_o^T v_c)} \frac{\partial}{\partial v_c} \sigma(u_o^T v_c)$

$= \frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c)(1 - \sigma(u_o^T v_c))\, u_o$

$= (1 - \sigma(u_o^T v_c))\, u_o$

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

$= \sum_k \frac{1}{\sigma(-u_k^T v_c)} \frac{\partial}{\partial v_c}$

$= \sum_k \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)\,(1 - \sigma(-u_k^T v_c))\,(-u_k)$

$= \sum_k (1 - \sigma(-u_k^T v_c)(-u_k)$

$\frac{\partial J}{\partial v_c} = -(1 - \sigma(u_o^T v_c))\, u_o + \sum_k (1 - \sigma(-u_k^T v_c)\, u_k)$

$\frac{\partial J}{\partial u_o} = -(1 - \sigma(u_o^T v_c))\, v_c^T + 0$

$\frac{\partial J}{\partial u_k} = \log\left(\sigma(-u_k^T v_c)\right)$

$= \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)(1 - \sigma(u_k^T v_c))\, v_c^T$

$= (1 - \sigma(u_k^T v_c))\, v_c^T$

We can use cached activation values to speed up the computation and there is no need to sum over the entire corpus to get probabilities.

f) $\frac{\partial J_{sg}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial U} = \sum_{-m \le j \le m,\ j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$

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

$\frac{\partial J_{sg}}{\partial v_w} = 0 \qquad \text{for } w \neq c$