1. a) when $w \neq o$, $y_w = 0$, so $-\sum_{w \in Vocab} y_w \log \hat{y}_w = -y_o \log \hat{y}_o = -\log \hat{y}_o$

b) $J = -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} = -u_o^T v_c + \log \sum_{w \in V} \exp(u_w^T v_c)$.

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot \left(\sum_{w \in V} u_w \cdot \exp(u_w^T v_c) \right)$$

$$= -u_o + \sum_{w \in V} \frac{u_w \cdot \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} = -u_o + \sum_{w \in V} u_w \cdot P(o=w \mid c=c)$$

$$= -u_o + \sum_{w \in V} u_w \cdot \hat{y}_w = U(\hat{y} - y)$$

(1) $\frac{\partial J}{\partial v_c} = 0 \Rightarrow \hat{y} = y$.

(2) If zero, then the parameter updating stops.

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

$w = o$: $$\frac{\partial J}{\partial u_w} = -v_c + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot v_c \cdot \exp(u_w^T v_c)$$

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

$$= -v_c + v_c \cdot P(w \mid c) = -v_c + v_c \cdot \hat{y}_w = v_c(\hat{y}_w - 1)$$

$w \neq o$: $\frac{\partial J}{\partial u_w} = v_c \cdot \hat{y}_w$.

d) $$\nabla J = \begin{pmatrix} \frac{\partial J}{\partial u_1} \\ \vdots \\ \frac{\partial J}{\partial u_{|V|}} \end{pmatrix}$$

(e) $$\frac{\partial f(x)}{\partial x} = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases}$$

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

g) $J = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T v_c))$

(i) $\frac{\partial J}{\partial v_c} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c)) \cdot u_o - \sum_{s=1}^{K} (-u_{w_s}) \cdot \frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c))$

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

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

$\frac{\partial J}{\partial u_{w_s}} = -\frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c) \cdot (1-\sigma(-u_{w_s}^T v_c)) \cdot (-v_c)$

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

(ii) we can reuse $\sigma(U^T v_c)$

(iii) Because it takes only K negative samples into account instead of the whole corpus while computing the gradient. it's faster.

(h) $J = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{w_s}^T v_c))$

$$\frac{\partial J}{\partial u_{w_s}} = \sum_{\substack{i=1 \\ w_s = w_i}}^{K} v_c(1-\sigma(-u_{w_i}^T v_c))$$

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

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

$$\frac{\partial J(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_w} = 0, \text{ when } w \ne c.$$

2. see codes.