1 Written: Understanding word2vec (23 points)

a) y is one-hot vector, when w≠o, y_w=0. So $-\sum_{w\in v} y_w \log \hat{y}_\omega = \log \hat{y}_0$

b)$U_o - \sum_{x=1}^{V} P(x|c)U_x = U(\hat{y} - y)$

c)

$$\frac{\partial}{\partial U_w} \log \frac{\exp(U_o^T V_c)}{\sum_{x=1}^{V} \exp(U_x^T V_c)}$$

$$\frac{\partial}{\partial U_w} (\underbrace{\log \exp(U_o^T V_c)}_{V_c / 0} - \log \sum_{x=1}^{V} \exp(U_x^T V_c))$$

$$\frac{\partial}{\partial U_w} \log \sum_{x=1}^{V} \exp(U_x^T V_c)$$

$$= \frac{1}{\sum_{x=1}^{V} \exp(U_x^T V_c)} \cdot \frac{\partial}{\partial U_w} \sum_{x=1}^{V} \exp(U_x^T V_c)$$

$$= \frac{\exp(U_w^T V_c)}{\sum_{x=1}^{V} \exp(U_x^T V_c)} V_c \qquad \therefore \frac{\partial J}{\partial U_w} = \begin{cases} \hat{y}_w V_c, & w \neq o \\ (\hat{y}_w - 1) V_c, & w = o \end{cases}$$

$$= \hat{y}_w V_c$$

d)

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

e)

$$J_{\text{neg-sample}}(v_c, o, U) = -\log(\sigma(u_o^\top v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^\top v_c))$$

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

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

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

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

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

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

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

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

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

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

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

With naïve-softmax loss, we don't need to compute through the whole vocabulary.

f)

(i) $\partial J_{skip-gram}(v_c, w_{t-m}), \ldots. w_{t+m}, U / \partial U = \sum_{-m \le j \le m,\, j \ne 0} \frac{\partial J(v_o, w_{w+j}, U)}{\partial U}$

(ii) $\partial J_{skip-gram}(v_c, w_{t-m}), \ldots. w_{t+m}, U / \partial v_c = \sum_{-m \le j \le m,\, j \ne 0} \frac{\partial J(v_o, w_{w+j}, U)}{\partial v_c}$

(iii) $\partial J_{skip-gram}(v_c, w_{t-m}), \ldots. w_{t+m}, U / \partial v_w (when\, w \ne c) = 0$