CS 224n HW #2

1a) $J_{\text{naive-softmax}}(v_c, o, U) = -\log P(O=o \mid C=c) \triangleq J_1$

$J_{\text{cross-entropy}} = \sum_{i \in \text{Vocab}} -y_i \log(\hat{y}_i) \triangleq J_2$

$\cancel{dJ = \frac{1}{P(O=o \mid C=c)} \partial P(O=o \mid C=c)}$

$y_i = 1$ when $O = o$
$y_i = 0$ otherwise

$\Rightarrow -\sum y_i (\log(\hat{y}_i))$
$= -\log(\hat{y}_o)$

$\hat{y} \triangleq P(O=o \mid C=c)$

$\Downarrow$

$J_1 = J_2$

b) $\partial J_1 / \partial v_c = \dfrac{\partial -\log(P(O=o \mid C=c))}{\partial v_c}$

$= \dfrac{\partial -\log\left[(e^{u_o^T v_c}) / (\sum_{w \in \text{Vocab}} e^{u_w^T v_c})\right]}{\partial v_c}$

$= \partial(-\log(e^{u_o^T v_c}) / \partial v_c) - \partial(-\log(\sum_{w \in V} e^{u_w^T v_c}))$

$= \underbrace{\dfrac{\partial - u_o^T v_c}{\partial v_c}}_{\triangleq T_1} + \underbrace{\dfrac{\partial \log(\sum_w e^{u_w^T v_c})}{\partial v_c}}_{\triangleq T_2}$

1c) $\frac{\partial J}{\partial u_w} = \frac{\partial \log \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}}{\partial u_w}$

~~when w=0~~

$= \frac{\partial \cdot \log e^{u_o^T v_c}}{\partial u_w} = \frac{\partial u_o^T v_c}{\partial u_w} + \frac{\partial + \log \sum e^{u_w^T v_c}}{\partial u_w} \triangleq T_2$

$\triangleq T_1$

= Case: when $w = 0$

$T_1 = \frac{\partial u_o^T v_c}{\partial u_o} = -v_c$

$T_2 = \frac{\partial \log \sum e^{u_w^T v_c}}{\partial u_o} = \frac{\partial \log \left(e^{u_o^T v_c} + \sum_{x \in W, x \neq 0} e^{u_x^T v_c}\right)}{\partial u_o}$

$= \cancel{\partial \log e} \frac{1}{\sum_w e^{u_x^T v_c}} e^{u_o^T v_c} v_c$

$T_1 + T_2 = -v_c + \frac{1}{\sum_w e^{u_w^T v_c}} e^{u_o^T v_c} v_c$

$= \cancel{\frac{v_c \sum e^{u_w^T v_c} + e^{u_o^T v_c} v_c}{\sum e^{u_w^T v_c}}}$

$= -(v_c - v_c\, p(O=o \mid c=c)) = -v_c(1 - p(O=o \mid C=c))$

$= -v_c(1-\hat{y}) \, y$

$=$

1c) $\frac{\partial J}{\partial u_w} = \frac{\partial \log \frac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}}{\partial u_w}$

~~when w=0~~

$= \underbrace{\frac{\partial \cdot \log e^{u_o^T v_c}}{\partial u_w} = \frac{\partial u_o^T v_c}{\partial u_w}}_{\triangleq T_1} + \frac{\partial + \log \sum e^{u_w^T v_c}}{\partial u_w} \triangleq T_2$

Case: when $w = 0$

$T_1 = \frac{\partial u_o^T v_c}{\partial u_o} = -v_c$

$T_2 = \frac{\partial \log \sum e^{u_w^T v_c}}{\partial u_o} = \frac{\partial \log \left(e^{u_o^T v_c} + \sum_{x \in W, x \neq 0} e^{u_x^T v_c}\right)}{\partial u_o}$

$= \cancel{\partial \log e} \frac{1}{\sum_w e^{u_x^T v_c}} e^{u_o^T v_c} v_c$

$T_1 + T_2 = -v_c + \frac{1}{\sum_w e^{u_w^T v_c}} e^{u_o^T v_c} v_c$

$= \cancel{\frac{v_c \sum e^{u_w^T v_c} + e^{u_o^T v_c} v_c}{\sum e^{u_w^T v_c}}}$

$= -(v_c - v_c\, p(O=o \mid c=c)) = -v_c(1 - p(O=o \mid c=c))$

$= -v_c(1 - \hat{y}) \, y$

$= \cancel{\Theta}$

$=$

1c, cont'd) cases when $w \neq o$

$$T_1 = \frac{\partial u_o^T v_c}{\partial u_w} = 0$$

$$T_2 = \frac{\partial \log \sum e^{u_w^T v_c}}{\partial u_w}$$

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

$$T_1 + T_2 = v_c \frac{e^{u_w^T v_c}}{\sum_w e^{u_w^T v_c}} = v_c \hat{y} (1-y)$$

1a) $$\frac{\partial \sigma(x)}{\partial x} = \partial\left(\frac{1}{1+e^{-x}}\right)/\partial x = \frac{(1+e^{-x})(0) + 1(1+e^{-x})(-e^{-x})}{(1+e^{-x})^2}$$

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

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

$$T_1 \frac{\partial T_1}{\partial v_c} = -u_o^T \qquad T_2 \frac{\partial T_2}{\partial v_c} = \frac{1}{\sum_w (e^{u_w^T v_c})} \frac{\partial e^{u_o^T v_c}}{\partial v_c}$$

$\boxed{1c, \text{cont'd}}$

$$\frac{\partial \sum_w e^{u_w^T v_c}}{\partial v_c} = \sum_w \frac{\partial e^{u_w^T v_c}}{\partial v_c}$$

$$= \sum_{w \neq o} \cancel{\frac{\partial e^{u_w^T v_c}}{\partial v_c}} + \sum_{w=o} \frac{\partial e^{u_o^T v_c}}{\partial v_c}$$

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

$$\partial J = T_1 + T_2$$

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

$$= -u_o^T + u_o^T P(O=o|C=c)$$

$$= u_o^T (P(O=o|C=c) - 1)$$

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

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

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

1e) $\dfrac{\partial J_{neg\text{-}sample}}{\partial u_k} = \cancel{\dfrac{\partial \left(\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c)) \right.}{\partial u_k}}$

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

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

1d) $\frac{\partial \sigma(x)}{\partial x} = \partial\left(1/(1+e^{-x})\right)/\partial x = \frac{\cancel{(1+e^{-x})(0)} + e^{-x}(+1)}{(1+e^{-x})^2}$

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

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

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

1e) $J_{neg\text{-}sample} = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

$$\frac{\partial J}{\partial v_c} = \underbrace{\frac{-\log(\sigma(u_o^T v_c))}{\partial v_c}}_{T_1} - \underbrace{\sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))}_{T_2}$$

$$T_1 = -\frac{1}{\cancel{\sigma(u_o^T v_c)}}(\cancel{\sigma(u_o^T v_c)})(1-\sigma(u_o^T v_c))\frac{\partial u_o^T v_c}{\partial v_c}$$

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

$$T_2 = \sum_{k=1}^{K} \frac{1}{\cancel{\sigma(-u_k^T v_c)}}(\cancel{\sigma(-u_k^T v_c)})(1-\sigma(-u_k^T v_c))\frac{\partial -u_k^T v_c}{\partial v_c}$$

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

$$T_1 + T_2 = -u_o(1-\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \sigma(-u_k^T v_c)\, u_k$$

(1e, cont'd)

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

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

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

efficiency advantage of sigmoid vs naive-softmax: don't have to compute $\sum_w e^{u_w^T v_c}$ for all w

---

1f) i) $$\frac{\partial J_{skip\text{-}gram}}{\partial U} = \sum_{\substack{-m \le j \\ j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

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

iii) $$\frac{\partial J_{skip\text{-}gram}}{\partial v_w} = \sum_j \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_w} = 0 \; ?$$