1

a.

$$CE(y,\hat{y}) = -\sum_w y_w \log(\hat{y}_w)$$

$$y_w = [0, \dots, \overset{\overset{o}{\downarrow}}{1}, \dots, 0]$$

$$= -1 \cdot \log(\hat{y}_o) - \sum_{w \neq o} 0 \cdot \log(\hat{y}_w)$$

$$= -\log(\hat{y}_o)$$

b.

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left[-\log(\hat{y}_o)\right] = -\frac{\partial}{\partial v_c}\left[\log(\exp(u_o^T v_c)) - \log\left(\sum_w \exp(u_w^T v_c)\right)\right]$$

$$= -u_o + \frac{\partial}{\partial v_c} \log\left(\sum_w \exp(u_w^T v_c)\right)$$

$$= -u_o + \frac{\sum_w u_w \exp(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)}$$

$$= -u_o + \sum_x \frac{\exp(u_x^T v_c)}{\sum_w \exp(u_w^T v_c)} u_x$$

$$= \boxed{-u_o + \sum_w^V \hat{y}_w u_w}$$

$u_w : d \times 1$ $\quad$ $\frac{\partial}{\partial u_w} : d \times 1$

$v_c : d \times 1$

$$\frac{\partial}{\partial u_w} u_w^T v_c = v_c \qquad \frac{\partial}{\partial v_c} u_w^T v_c = u_w$$

c.

$$\frac{\partial J}{\partial u_w} = -\frac{\partial}{\partial u_w}\left[\underbrace{\log(\exp(u_o^T v_c))}_{\substack{v_c \text{ if } w=o \\ 0 \quad \text{otherwise}}} - \underbrace{\log\left(\sum_i \exp(u_i^T v_c)\right)}_{\frac{\exp(u_w^T v_c)\, v_c}{\sum_i \exp(u_i^T v_c)} = \hat{y}_w v_c}\right]$$

$$= \boxed{\begin{cases} v_c(\hat{y}_w - 1) & \text{if } w = 0 \\ v_c\,\hat{y}_w & \text{otherwise} \end{cases}}$$

d.

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

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

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

$$= \sigma(x) - \sigma^2(x) = \boxed{\sigma(x)(1-\sigma(x))}$$

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

e.

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

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

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

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

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

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

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

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

Looping over $K \ll V$ much faster than over entire vocabulary f.

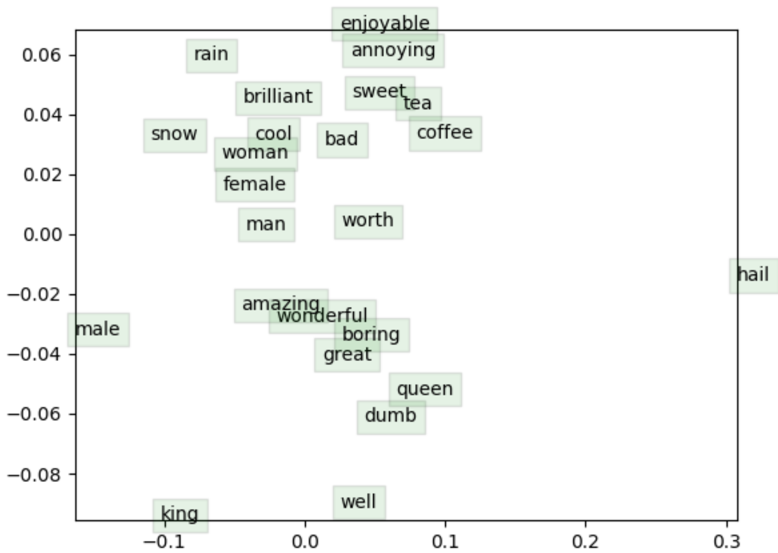
i) $$\frac{\partial J_{skip}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial U} = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

ii) $$\frac{\partial J_{skip}}{\partial v_c} = \sum_j \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

iii) $$\frac{\partial J_{skip}}{\partial v_w} = 0$$
$w \ne c$

2.

c.



- Interchangeable words are grouped [amazing, wonderful]
- Grouped words do not necessarily have similar meaning:
  [boring, great] [enjoyable, annoying]
- Consistent vector from good to bad:
  bad − brilliant ≈ boring − amazing ≈ dumb − great
  ≈ annoying − enjoyable