Calculating gradient

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \neq 0}} \log P(w_{t+j} \mid w_t)$$

$$P(o|c) = \frac{e^{u_o^T v_c}}{\sum_{w=1}^{V} e^{u_w^T v_c}}$$

$$\frac{\partial}{\partial v_c} \log P(o|c) = \frac{\partial}{\partial v_c} \log e^{u_o^T v_c} - \frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} e^{u_w^T v_c}$$

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

$$= u_o - \frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} e^{u_w^T v_c} \qquad (f \cdot g)$$

$$= u_o - \frac{1}{\sum_{w=1}^{V} e^{u_w^T v_c}} \cdot \frac{\partial}{\partial v_c} \sum_{x=1}^{V} e^{u_x^T v_c}$$

$$= u_o - \dots \cdot \sum_{x=1}^{V} \frac{\partial}{\partial v_c} e^{u_x^T v_c}$$

$$= u_o - \dots \cdot \sum_{x=1}^{V} e^{u_x^T v_c} \cdot \frac{\partial}{\partial v_c} u_x^T v_c$$

$$= u_o - \dots \cdot \sum_{x=1}^{V} e^{u_x^T v_c} \cdot u_x$$

$$= u_o - \frac{\sum_{x=1}^{V} e^{\left(u_x^T v_c\right)} \cdot u_x}{\sum_{w=1}^{V} e^{\left(u_w^T v_c\right)}}$$

$$= u_o - \sum_{x=1}^{V} \left(\frac{e^{\left(u_x^T v_c\right)}}{\sum_{w=1}^{V} e^{u_w^T v_c}} \right) u_x$$

softmax

$$= u_0 - \underbrace{\sum_{x=1}^{V} P(x|c)\ u_x}_{\text{expectation}}$$

(average over all context vectors weighted by P over model)

= observed − expected