a. naive-softmax loss가 cross entropy loss와 같아지는 이유

$$-\sum_{w \in Vocab} y_w \log \hat{y}_w = -\log \hat{y}_o$$

y는 실제 확률 분포, $\hat{y}$는 모델에서 구한 확률 분포

y는 context word o에 해당하는 element만 1인 one-hot vector ∴ 참

b, c b = naive softmax loss 식을 v_c에 대해 편미분

$$\frac{\partial J}{\partial v_c} = -u_o + \sum_{x \in Vocab} P(x|c)\, u_x$$

실제 분포와 가중치가 있는 확률분포의 차이값 계산

$$\frac{\partial J}{\partial v_c} = U(\hat{y} - y)$$

c = naive softmax loss 식을 u에 대해 편미분 ($w = o$인 경우)

$$\frac{\partial J}{\partial u_o} = (\hat{y}_o - y_o)\, v_c$$

↓

확률분포의 element끼리 더하고 뺀 값 (scalar 값)

($w \neq o$인 경우)

$$\frac{\partial J}{\partial u_w} = \hat{y}_w v_c$$

$$\frac{\partial J}{\partial U} = (\hat{y} - y)\, v_c^T \quad / \text{전체 } U \text{에 대해 편미분}$$

d Sigmoid 편미분

$$\frac{\partial \sigma}{\partial x} = \sigma(1-\sigma)$$

e negative sample에 대한 loss의 편미분 식, neg sample의 loss

↓

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

K : negative samples

o : neg sample ×

<각각 미분한 결과>

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

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

$$\frac{\partial J}{\partial u_k} = \sum_{x=1}^{K} (1-\sigma(-u_x^T v_c)) \frac{\partial u_x^T v_c}{\partial u_K}$$