Homework_4
Shiqiang

(a)

Suppose y_i^t is the only nonzero element of y^t,then

$$J^{(t)}(\theta) = CE\left(\mathbf{y}^{(t)}, \mathbf{y}^{(t)}\right) = -\log \hat{y}_i^t = \log \frac{1}{\hat{y}_i^t}$$

$$PP\left(y^t, \hat{y}^t\right) = \frac{1}{\hat{y}_i^t}$$

Then, $J^{(t)}(\theta) = \log PP\left(y^t, \hat{y}^t\right)$

If the model predictions are completely random, $E\left[\hat{y}_i^t\right] = \frac{1}{|V|}$

the perplexity are $PP\left(y^t, \hat{y}^t\right) = \frac{1}{\hat{y}_i^t} = 10000$

the expected cross-entropies are

$$J^{(t)}(\theta) = \log |V| = \log \frac{1}{\hat{y}_i^t} = \log 10000 = \log_2 10000 \approx 9.21$$

(b)

$As\ known, the\ soft \max\ founction\ is:$

$$\hat{y}_i = \frac{\exp\left(o_i\right)}{\sum_j \exp\left(o_j\right)}$$

$if\quad i = j:$

$$\frac{\partial \hat{y}_j}{\partial o_i} = \frac{\exp\left(o_i\right) \bullet \sum_j \exp\left(o_j\right) - (\exp\left(o_i\right))^{\wedge 2}}{(\sum_j \exp\left(o_j\right))^{\wedge 2}}$$

$$= \frac{\exp\left(o_i\right)}{\sum_j \exp\left(o_j\right)} (1 - \frac{\exp\left(o_i\right)}{\sum_j \exp\left(o_j\right)})$$

$$= \hat{y}_i (1 - \hat{y}_i)$$

$if \quad i \neq j:$

$$\frac{\partial \hat{y}_j}{\partial o_i} = \frac{-\exp\left(o_i\right) \bullet \exp\left(o_j\right)}{(\sum_j \exp\left(o_j\right))^{\wedge 2}}$$

$$= -\frac{\exp\left(o_i\right)}{\sum_j \exp\left(o_j\right)} \bullet \frac{\exp\left(o_j\right)}{\sum_j \exp\left(o_j\right)}$$

$$= -\hat{y}_i \hat{y}_j$$

The loss founction is $J^{(t)}(\theta) = CE\left(\mathbf{y}^{(t)}, \mathbf{y}^{(t)}\right) = -\sum_{j=1}^{|V|} y_j^{(t)} \log \hat{y}_j^{(t)}$

$$\frac{\partial J^{(t)}}{\partial o_i} = \frac{\partial(-\sum_{j=1}^{|V|} y_j^{(t)} \log \hat{y}_j^{(t)})}{\partial o_i}$$

$$= -\sum_{j=1}^{|V|} y_j^{(t)} \frac{1}{\hat{y}_j^{(t)}} \frac{\partial \hat{y}_j^{(t)}}{\partial o_i}$$

$$= -\sum_{i=j} \frac{y_i^{(t)}}{\hat{y}_j^{(t)}} \hat{y}_i^{(t)} (1 - \hat{y}_i^{(t)}) + \sum_{i \neq j} \frac{y_i^{(t)}}{\hat{y}_j^{(t)}} \hat{y}_i^{(t)} \hat{y}_j^{(t)}$$

$$= -\sum_{i=j} y_i^{(t)} (1 - \hat{y}_i^{(t)}) + \sum_{i \neq j} y_i^{(t)} \hat{y}_i^{(t)}$$

$$= \sum_{i \neq j} y_i^{(t)} \hat{y}_i^{(t)} - y_i^{(t)} + y_i^{(t)} \hat{y}_i^{(t)}$$

$$= \hat{y}_i^{(t)} \sum_j y_i^{(t)} - y_i^{(t)}$$

$$= \hat{y}_i^{(t)} - y_i^{(t)}$$

Assume，
$$g^{(t)} = \mathbf{h}^{(t-1)} \mathbf{H} + \mathbf{e}^{(t)} \mathbf{I} + \mathbf{b}_1$$
$$o^{(t)} = \mathbf{h}^{(t)} \mathbf{U} + \mathbf{b}_2$$

Then，
$$\frac{\partial J^{(t)}}{\partial o^{(t)}} = \hat{y}^{(t)} - y^{(t)}$$
$$\frac{\partial J^{(t)}}{\partial g^{(t)}} = \frac{\partial J^{(t)}}{\partial o^{(t)}} \frac{\partial o^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial g^{(t)}} = (\hat{y}^{(t)} - y^{(t)}) \mathbf{U}^T h^{(t)} (1 - h^{(t)})$$

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial o^{(t)}} \frac{\partial o^{(t)}}{\partial b_2} = \hat{y}^{(t)} - y^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial g^{(t)}} \frac{\partial g^{(t)}}{\partial e^{(t)}} \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = (\hat{y}^{(t)} - y^{(t)}) \mathbf{U}^T h^{(t)} (1 - h^{(t)}) \mathbf{I}^T$$

$$\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial g^{(t)}} \frac{\partial g^{(t)}}{\partial I} = (e^{(t)})^T (\hat{y}^{(t)} - y^{(t)}) \mathbf{U}^T h^{(t)} (1 - h^{(t)})$$

$$\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial g^{(t)}} \frac{\partial g^{(t)}}{\partial H} = (h^{(t-1)})^T (\hat{y}^{(t)} - y^{(t)}) \mathbf{U}^T h^{(t)} (1 - h^{(t)})$$

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial g^{(t)}} \frac{\partial g^{(t)}}{\partial h^{(t-1)}} = (\hat{y}^{(t)} - y^{(t)}) \mathbf{U}^T h^{(t)} (1 - h^{(t)}) H^T$$