Homework_4

Lv Xinpeng

a) According to the cross-entropy loss and perplexity information, we can get:

$$PP^{(t)}(y^{(t)},\ \hat{y}^{(t)}) = 2^{J^{(t)}(\theta)}$$

So, that minimizing the (arithmetic) mean cross-entropy loss will also minimize the (geometric) mean perplexity across the training set.
For |V| = 10000, cross-entropy loss = $\log_2$|V|
The result is approximately equal to 13.285.

b) According to the forward propagation:

$$\begin{aligned} \boldsymbol{e}^{(t)} &= \boldsymbol{x}^{(t)}\boldsymbol{L} \\ \boldsymbol{h}^{(t)} &= \text{sigmoid}\left(\boldsymbol{h}^{(t-1)}\boldsymbol{H} + \boldsymbol{e}^{(t)}\boldsymbol{I} + \boldsymbol{b}_1\right) \\ \hat{\boldsymbol{y}}^{(t)} &= \text{softmax}\left(\boldsymbol{h}^{(t)}\boldsymbol{U} + \boldsymbol{b}_2\right) \end{aligned}$$

Assume:

$$g^{(t)} = h^{(t-1)}H + e^{(t)}I + b_1$$

$$k^{(t)} = h^{(t)}U + b_2$$

And here is the cross-entropy function:

$$J^{(t)}(\theta) = CE(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)}) = -\sum_{j=1}^{|V|} y_j^{(t)} \log \hat{y}_j^{(t)}$$

So, we can get:

$$\begin{aligned} \frac{\partial J^{(t)}}{\partial b_2} &= \frac{\partial J^{(t)}}{\partial k^{(t)}} \bullet \frac{\partial k^{(t)}}{\partial b_2} = \hat{y}^{(t)} - y^{(t)} \\ \frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} &= \frac{\partial J^{(t)}}{\partial g^{(t)}} \bullet \frac{\partial g^{(t)}}{\partial e^{(t)}} \bullet \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = (\hat{y}^{(t)} - y^{(t)})U^T h^{(t)}(1-h^{(t)})I^T \\ \frac{\partial J^{(t)}}{\partial I_t} &= \frac{\partial J^{(t)}}{\partial g^{(t)}} \bullet \frac{\partial g^{(t)}}{\partial I_t} = e^{(t)T}(\hat{y}^{(t)} - y^{(t)})U^T h^{(t)}(1-h^{(t)}) \\ \frac{\partial J^{(t)}}{\partial H_t} &= \frac{\partial J^{(t)}}{\partial g^{(t)}} \bullet \frac{\partial g^{(t)}}{\partial H_t} = h^{(t-1)T}(\hat{y}^{(t)} - y^{(t)})U^T h^{(t)}(1-h^{(t)}) \\ \frac{\partial J^{(t)}}{\partial h^{(t-1)}} &= \frac{\partial J^{(t)}}{\partial g^{(t)}} \bullet \frac{\partial g^{(t)}}{\partial h^{(t-1)}} = (\hat{y}^{(t)} - y^{(t)})U^T h^{(t)}(1-h^{(t)})H^T \end{aligned}$$