Assignment 4

Fudi(Fred) Wang

November 2019

Part 1

(a) Since $\boldsymbol{y}^{(t)}$ is a one hot vector, without loss of generation, let the index such that $y_j^{(t)} = 1$ be i, we have:

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

Hence, minimizing the mean cross-entropy loss will also minimize the mean perplexity across the training set.
Now, since model predictions were completely random (chosen uniformly from the vocabulary), and the sum of the $\hat{y}_i^{(t)}$s is equal to one, we have:

$$E(\hat{y}_i^{(t)}) = \frac{1}{\mid V \mid} \qquad \text{for i} \in (1, \mid V \mid)$$

Therefore,

$$E(PP^{(t)}) = \mid V \mid$$

So the corresponding cross-entropy loss when $\mid V \mid = 10000$ is just

$$E(J^{(t)}(\theta)) = log10000 = 4$$

Of course, if log is base 2, then we get the number 13.29.

(b) First, let $\boldsymbol{x}^{(t)} = \boldsymbol{h}^{(t)}\boldsymbol{U} + \boldsymbol{b}_2$.
The softmax function is defined as:

$$softmax(x_i) = \frac{x_i}{\sum_j x_j} = \frac{1}{1 + e^{-x_i}\sum_{j \neq i} e^{x_j}}$$

where the second equality holds by dividing both the numerator and the denominator by e^{x_i} Note that the derivative of the softmax function can be discussed as two cases:

$$\frac{\partial softmax(x_i)}{\partial x_i} = \frac{e^{-x_i}\sum_{j \neq i} e^{e^{x_j}}}{(1 + e^{-x_i}\sum_{j \neq i} e^{x_j})^2} = softmax(x_i)(1 - softmax(x_i))$$

and

$$\frac{\partial softmax(x_i)}{\partial x_j} = -\frac{e^{x_i}e^{x_j}}{(\sum_j e^{x_j})^2} = -softmax(x_i)softmax(x_j)$$

And recall that from previous assignments, $\sigma^{'}(x) = \sigma(x)(1-\sigma(x))$. Hence we have the following partial derivatives:

$$\frac{\partial J^{(t)}}{\partial \boldsymbol{b_2}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \frac{\partial \hat{y}^{(t)}}{\partial \boldsymbol{x}^{(t)}} \frac{\partial \boldsymbol{x}^{(t)}}{\partial \boldsymbol{b_2}} = -\frac{1}{\hat{y}^{(t)}} \hat{y}^{(t)}(1-\hat{y}^{(t)}) = \boldsymbol{\hat{y}}^{(t)} - \boldsymbol{y}^{(t)}$$

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

$$\frac{\partial J^{(t)}}{\partial \boldsymbol{I}}|_{(t)} = (\boldsymbol{e}^{(t)})^T (\boldsymbol{\hat{y}}^{(t)} - \boldsymbol{y}^{(t)}) \boldsymbol{U}^T \boldsymbol{h}^{(t)} (1 - \boldsymbol{h}^{(t)})$$

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

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