

NLP - HW2

Omri Efroni, 204037840
Meitar Shechter, 307938217

December 12th, 2020

1

1.1

Assuming j is the entry where $y_j = 1$:

$$\frac{\partial CE(y, \hat{y})}{\partial \theta_l} = \frac{\partial (-\log(\text{softmax}(\theta)_j))}{\partial \theta_l} = \frac{\partial (-\theta_j + \log(\sum_k e^{\theta_k}))}{\partial \theta_l}$$

For $l = j$:

$$= -1 + \text{softmax}(\theta_l)$$

For $l \neq j$:

$$= \text{softmax}(\theta_l)$$

Or in vector notation:

$$\hat{y} - y$$

1.2

Note $\theta = hW_2 + b_2$:

$$\frac{\partial CE(y, \hat{y})}{\partial x} = \frac{\partial CE(y, \hat{y})}{\partial \theta} \cdot \frac{\partial \theta}{\partial x} \quad (1)$$

Now:

$$\frac{\partial \theta}{\partial x} = \frac{\partial (hW_2 + b_2)}{\partial x} = \frac{\partial \theta}{\partial h} \cdot \frac{\partial h}{\partial x} = \left[\frac{\partial \theta}{\partial h} = W_2^T \right] = W_2^T \cdot \frac{\partial h}{\partial x} \quad (2)$$

Note $\theta^* = xW_1 + b_1$:

$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial \theta^*} \cdot \frac{\partial \theta^*}{\partial x} = \text{diag}(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \quad (3)$$

Plugging this into equation (2):

$$\frac{\partial \theta}{\partial x} = W_2^T \cdot \text{diag}(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \quad (4)$$

And together with subsection (a):

With the same notation as the previous section, for $l = j$:

$$\frac{\partial CE(y, \hat{y})}{\partial x} = (\hat{y} - y) \cdot W_2^T \cdot \text{diag}(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \quad (5)$$

1.3

CODE

1.4

Our perplexity is: "dev perplexity : 115.34486812512428".

2

2.1

Let's note $\theta_1 = h^{(t)}U + b_2$, $\theta_2 = h^{(t-1)}H + e^{(t)}I + b_1$ and $\delta = \frac{\partial CE(y, \hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial \theta_2}$.
Now:

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial CE(y, \hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial b_2} = (\hat{y} - y) \quad (6)$$

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial CE(y, \hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial U} = (h^{(t)})^T \cdot (\hat{y} - y) \quad (7)$$

$$\frac{\partial J^{(t)}}{\partial b_1} \Big|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial b_1} = (\hat{y} - y) \cdot U^T \cdot \text{diag}(h^{(t)} \circ (1 - h^{(t)})) \quad (8)$$

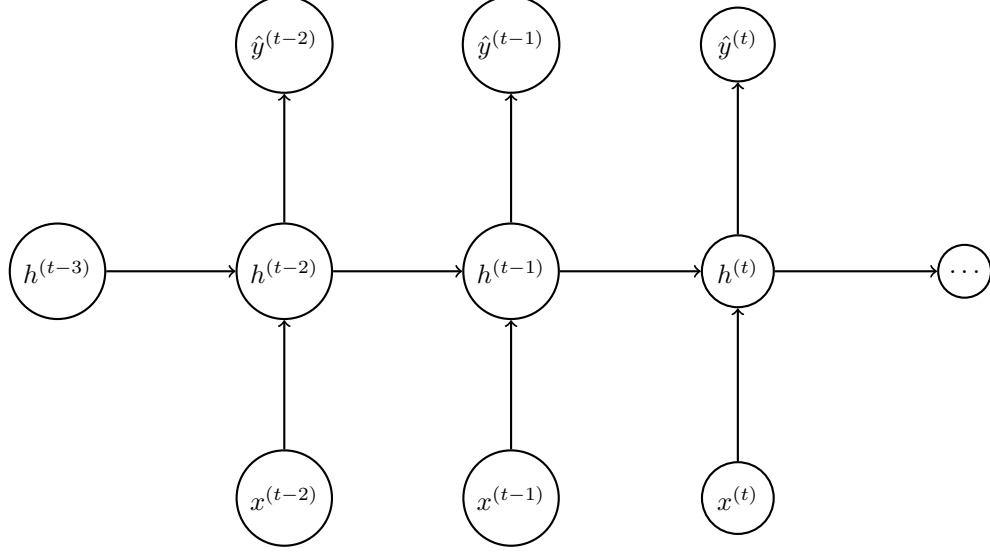
$$\frac{\partial J^{(t)}}{\partial H} \Big|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial H} = (h^{(t-1)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \quad (9)$$

$$\frac{\partial J^{(t)}}{\partial I} \Big|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial I} = (e^{(t)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \quad (10)$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \delta \cdot \frac{\partial \theta_2}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial b_1} \cdot I^T \quad (11)$$

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \delta \cdot \frac{\partial \theta_2}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial b_1} \cdot H \quad (12)$$

2.2



In this subsection our notation is adjusted to the current time step, meaning:
 $\theta_2 = h^{(t-2)}H + e^{(t-1)}I + b_1$

$$\frac{\partial J^{(t)}}{\partial b_1} \Big|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial b_1} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \text{diag}(h^{(t-1)} \circ (1 - h^{(t-1)})) \quad (13)$$

$$\frac{\partial J^{(t)}}{\partial I} \Big|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial I} = (e^{(t-1)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \Big|_{(t-1)} \quad (14)$$

$$\frac{\partial J^{(t)}}{\partial H} \Big|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial H} = (h^{(t-1)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \Big|_{(t-1)} \quad (15)$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial e^{(t-1)}} \cdot \frac{\partial e^{(t-1)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial b_1} \Big|_{(t-1)} \cdot I^T \quad (16)$$

3

Notebook.

4

From logarithms' rules we know: $\log_2(x) = \log_2(e) \cdot \ln(x)$, therefore:

$$2^{-\frac{1}{M} \sum_{i=1}^M \log_2 p(s_i | s_1, \dots, s_{i-1})} = 2^{-\frac{1}{M} \sum_{i=1}^M \ln(p(s_i | s_1, \dots, s_{i-1})) \cdot \log_2(e)} = 2^{-\log_2(e) \frac{1}{M} \sum_{i=1}^M \ln(p(s_i | s_1, \dots, s_{i-1}))} = e^{-\frac{1}{M} \sum_{i=1}^M \ln(p(s_i | s_1, \dots, s_{i-1}))}$$