1 Word-Level Neural Bigram Language Model

1.a

In the one-hot vector y, only one value is 1 and the others are zeros.

$$CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_k) = -y_k \log(\hat{y}_k) = -\log(\hat{y}_k)$$

Deriving with respect to the i^{th} item of input vector θ, aka θ_i:

$$\frac{\partial CE(y, \hat{y})}{\partial \theta_i} =$$

$$-\frac{\partial \log \frac{e^{\theta_k}}{\sum_j e^{\theta_j}}}{\partial \theta_i} =$$

$$-\frac{\partial \left(\theta_k - \log \sum_j e^{\theta_j}\right)}{\partial \theta_i} =$$

$$\left(\frac{\partial \log \sum_j e^{\theta_j}}{\partial \theta_i} - \frac{\partial \theta_k}{\partial \theta_i}\right) =$$

$$\begin{cases} (\hat{y}_i - 1) & i = k \\ \hat{y}_i & i \neq k \end{cases}$$

Therefore:

$$\frac{\partial CE(y, \hat{y})}{\partial \theta} = \hat{y} - y$$

1.b

In the one-hot vector y, only one value is 1 and the others are zeros.

$$J = CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_k) = -y_k \log(\hat{y}_k) = -\log(\hat{y}_k)$$

Let's notate:

$$z_1 = xW_1 + b_1 \text{ and } z_2 = hW_2 + b_2$$

Therefore:

$$h = sigmoid(xW_1 + b_1) = sigmoid(z_1)$$
$$\hat{y} = softmax(hW_2 + b_2) = softmax(z_2)$$

So:

$$\frac{\partial CE}{\partial x} =$$
$$\frac{\partial J}{\partial x} =$$
$$\frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial h} \cdot \frac{\partial h}{\partial z_1} \cdot \frac{\partial z_1}{\partial x} =$$
$$(\hat{y} - y) \cdot W_2 \cdot (z_1 - z_1^2) \cdot W_1$$

2 Theoretical Inquiry of a Simple RNN Language Model

2.a

Let's notate:

$$z_1^{(t)} = h^{(t-1)}H + e^{(t)}I + b_1$$
$$z_2^{(t)} = h^{(t)}U + b_2$$

Therefore:

$$\delta_1^{(t)} = \frac{\partial J}{\partial z_1^{(t)}} = \hat{y}^{(t)} - y^{(t)}$$

$$\delta_2^{(t)} = \frac{\partial J}{\partial z_2^{(t)}} = \delta_1^{(t)} \frac{\partial z_2^{(t)}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial z_1^{(t)}} = \delta_1^{(t)} U^T \circ h^{(t)} \circ (1 - h^{(t)})$$

The gradients for the model parameters:

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

$$\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial z_2^{(t)}} \frac{\partial z_2^{(t)}}{\partial b_2} = \delta_1^{(t)}$$

$$\frac{\partial J}{\partial L_{x^{(t)}}} = \frac{\partial J}{\partial z_1^{(t)}} \frac{\partial z_1^{(t)}}{\partial e^{(t)}} \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = \delta_2^{(t)} \ I^T$$

$$\frac{\partial J}{\partial I} = \frac{\partial J}{\partial z_1^{(t)}} \frac{\partial z_1^{(t)}}{\partial I} = (e^{(t)})^T \cdot \delta_2^{(t)}$$

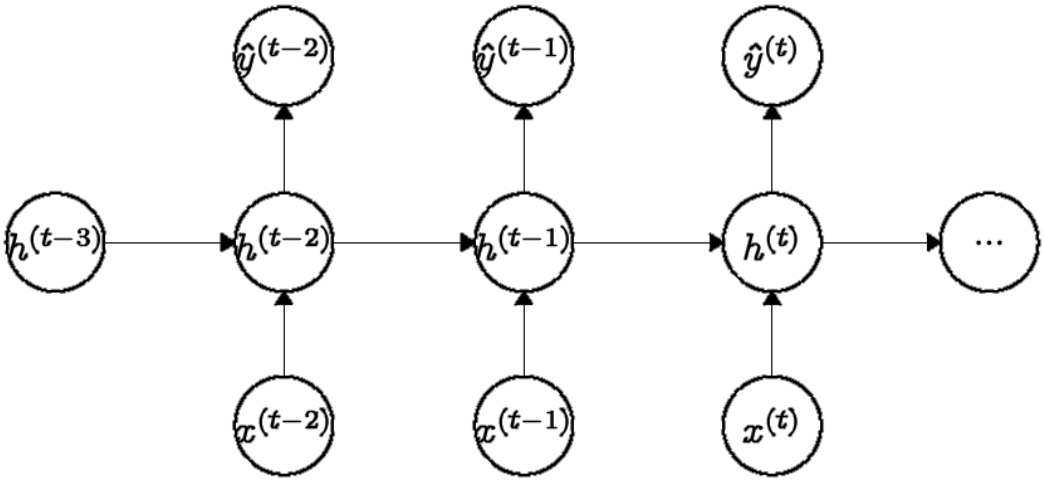
$$\frac{\partial J}{\partial H} = \frac{\partial J}{\partial z_1^{(t)}} \frac{\partial z_1^{(t)}}{\partial H} = (h^{(t-1)})^T \delta_2^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial b_1} = \left(\frac{\partial J^{(t)}}{\partial h^{(t)}} \circ \sigma'(h^{(t-1)} H + e^{(t)} I + b_1) \right)$$

The derivative with respect to the previous hidden layer:

$$\frac{\partial J}{\partial h^{(t-1)}} = \frac{\partial J}{\partial z_1^{(t)}} \ \frac{\partial z_1^{(t)}}{\partial h^{(t-1)}} = \delta_2^{(t)} \ H^T$$

2.b



The "backpropagation-through-time" gradients:

$$\frac{\partial J}{\partial L_x^{(t-1)}} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z_1^{(t-1)}} \frac{\partial z_1^{(t-1)}}{\partial L_x^{(t-1)}} = \delta^{(t-1)} \sigma'(z_1^{(t-1)}) I^T$$

$$\frac{\partial J}{\partial H} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z_1^{(t-1)}} \frac{\partial z_1^{(t-1)}}{\partial H} = (h^{(t-2)})^T \delta^{(t-1)} \sigma'(z_1^{(t-1)})$$

$$\frac{\partial J}{\partial I} = \frac{\partial J}{\partial h^{(t-1)}} \frac{\partial h^{(t-1)}}{\partial z_1^{(t-1)}} \frac{\partial z_1^{(t-1)}}{\partial I} = (e^{(t-1)})^T \delta^{(t-1)} \sigma'(z_1^{(t-1)})$$

$$\frac{\partial J}{\partial b_1}|_{t-1} = \delta^{(t-1)} \circ \sigma'(z_1^{(t-1)})$$

Where:

$$\sigma'(z_1^{(t-1)}) = \frac{\partial h^{(t-1)}}{\partial z_1^{(t-1)}} = diag(h^{(t-1)} \circ (1 - h^{(t-1)}))$$

2.c

The following vectors are of sizes:

$$\hat{y}^{(t)} \in \mathbb{R}^{|V|}, \quad e^{(t)} \in \mathbb{R}^d, \quad h^{(t-1)} \in \mathbb{R}^{D_h}$$

Therefore, computing the following:

$\hat{y}^{(t)}$ *involves* $O(|V| \cdot D_h)$ *computation*

$e^{(t)}$ *can be obtained by going over matrix* L *in* $O(d)$

$h^{(t)}$ *involves* $O(d \cdot D_h + D_h^2)$ *computation, and*

Both forward and backward propagation shall take:

$$O(|V| \cdot D_h + d \cdot D_h + D_h^2)$$

For τ *time steps, it's:*

$$O(\tau(|V| \cdot D_h + d \cdot D_h + D_h^2))$$

Of the three computations, the slowest is $\hat{y}^{(t)}$*, of* $O(|V| \cdot D_h)$ *complexity, especially when* $D_h << |V|$.