

Question 1 //

$$h_t = \tanh (w_{xh} x_t + w_{hh} h_{t-1} + b_h)$$

$$z_t = \text{softmax} (w_{hz} h_t + b_z)$$

Let o_t represent the vector before softmax

$$o_t = w_{hz} h_t + b_z$$

$$z_t = \text{softmax}(o_t)$$

Now

$$\frac{\partial E}{\partial w_{hz}} = \frac{\partial E}{\partial o_t} \times \frac{\partial o_t}{\partial w_{hz}} \quad \text{--- (1)}$$

Now $E = \text{cross entropy} = \sum_k y_k \log z_k$

$$\frac{\partial E}{\partial o_t} = - \sum_k y_k \frac{\partial \log p_k}{\partial o_t} = - \sum_k y_k \frac{1}{z_k} \frac{\partial z_k}{\partial o_t}$$

Now softmax dy

$$\frac{\partial E}{\partial o_t} = -y_i (1 - z_i) - \sum_{k \neq i} y_k \frac{1}{z_k} (-z_k \cdot z_i)$$

$$= -y_i + y_i z_i + \sum_{k \neq i} y_k (z_i)$$

$$= z_i \sum_k y_k - y_i$$

$$= z_i - y_i \rightarrow \text{if } y_i \text{ is not intended}$$

Substituting in equation 1

$$\frac{\partial L}{\partial w_{ht}} = \left(z_t \sum_k y_k - y_t \right) \cdot h_t //$$

Now similarly.

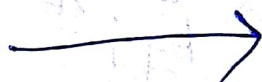
$$\frac{\partial E}{\partial w_{hh}} = \frac{\partial E}{\partial o_t} \times \frac{\partial o_t}{\partial h_t} \times \frac{\partial h_t}{\partial w_{hh}}$$

$$\text{Now } \frac{\partial E}{\partial o_t} = \left(z_t \sum_k y_k - y_t \right)$$

$$\frac{\partial o_t}{\partial h_t} = h_t$$

$$\frac{\partial h_t}{\partial w_{hh}} = 1 - h_t^2 \cdot h_{t-1}$$

$$\frac{\partial E}{\partial w_{hh}} = \left(z_t \sum_k y_k - y_t \right) \cdot h_t \cdot (1 - h_t^2 \cdot h_{t-1})$$



Question 2//

* The tanh function which has been used as an activation function has differentiation value between $(0, 1)$. Thus as we move forward wrt time gradient accumulate becomes lower and lower.

for ex $\frac{\partial E}{\partial W_{HH}}$ has h_t and $(1-h_t^2 \cdot h_{t-1})$ terms and thus $\frac{\partial E}{\partial W_{HH}}$ has very small value and the update would be very small. This ⁰ results ~~in~~ network unable to learn the function and is called problem of vanishing gradients.