Question 1 //

YASH PATEL
GAURAV MAHESHW
ANI

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

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

Let o_t represent intermediate 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 $E = \text{cross entropy} = \sum_k^{n\text{classes}} y_k \log z_k$

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

Now softmax derivative

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

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

$$= z_i \sum_k y_k - y_i$$

$$= z_i - y_i \rightarrow$$ 4 yes one hot encoded

Substituting in equation 1

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

(b) Now similarly.

$$\frac{\partial E}{\partial w_{HH}} = \frac{\partial E}{\partial O_t} \times \frac{\partial O_t}{\partial h_t} \times \frac{\partial h_t}{\partial w_{HH}}$$

Now $$\frac{\partial E}{\partial O_t} = \left(z_i \sum y_k - y_i\right) \circledast$$

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

$$\frac{\partial h_t}{\partial i_{HH}} = 1 - h_t^2 \cdot h_{t-1}$$

$$\frac{\partial E}{\partial w_{HH}} = \left(z_t \sum y_k - y_i\right) \cdot h_t \cdot \left(1 - h_t^2 \cdot h_{t-1}\right)$$

$\longrightarrow$

Question 2//

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

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