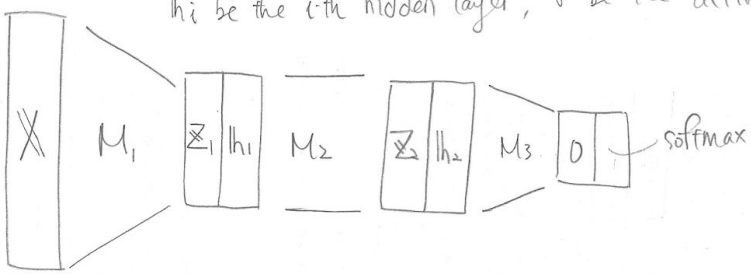
1. Explain why “non-linearity” is required property for any activation function deployed in a deep neural network.

Let X be the input vector, M_i be the learnable parameters
h_i be the i-th hidden layer, σ be the activation function



$Z_1 = M_1 X$, $h_1 = \sigma(Z_1)$

$Z_2 = M_2 h_1$, $h_2 = \sigma(Z_2)$

Assume σ to be a linear function : $f(x) = cx$

$h_1 = cZ_1$, $h_2 = cZ_2 = cM_2 h_1 = cM_2 cM_1 X = c^2 M_2 M_1 X$

$\therefore$ No non-linearity (regardless # of hidden layers) $= M'X$
if σ is linear ✱

2. Derive $\partial L/\partial W_{x2,h1}$ using backward propagation

$$\frac{\partial L}{\partial W_{x_2,h_1}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_{x_2,h_1}} = \underbrace{\frac{\partial L}{\partial z_1}} \cdot x_2$$

$\sigma'(z_1) \Downarrow$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} = \left(\underbrace{\frac{\partial L}{\partial z_3}} \cdot \underbrace{\frac{\partial z_3}{\partial h_1}}_{W_{h_1,h_3}} + \frac{\partial L}{\partial z_4} \cdot \underbrace{\frac{\partial z_4}{\partial h_1}}_{W_{h_1,h_4}} \right) \cdot \boxed{\frac{\partial h_1}{\partial z_1}}$$

$\sigma'(z_3) \Downarrow$

$$\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial h_3} \cdot \frac{\partial h_3}{\partial z_3} = \left(\underbrace{\frac{\partial L}{\partial o_1}} \cdot \underbrace{\frac{\partial o_1}{\partial h_3}}_{W_{h_3,o_1}} + \frac{\partial L}{\partial o_2} \cdot \underbrace{\frac{\partial o_2}{\partial h_3}}_{W_{h_3,o_2}} \right) \cdot \boxed{\frac{\partial h_3}{\partial z_3}}$$

$$\frac{\partial L}{\partial o_1} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial o_1} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial o_1} = -\frac{y_1}{\hat{y}_1} \left[\frac{e^{o_1}}{e^{o_1}+e^{o_2}} - \left(\frac{e^{o_1}}{e^{o_1}+e^{o_2}} \right)^2 \right] = y_1 \cdot \hat{y}_1 - y_1$$

($\hat{y}_1$ and $\hat{y}_1^2$)

$$= \hat{y}_1 - 1$$

#

(assuming $y_1 = 1$, which is ground truth, and $y_2 = 0$)

!!! true label