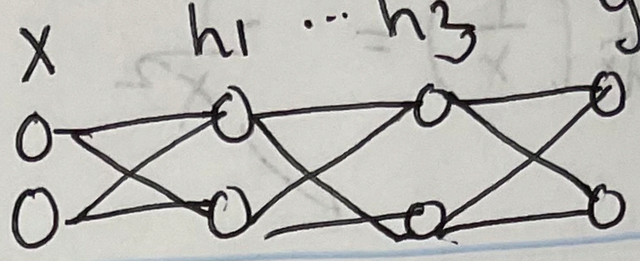


→ 3 layers

layer

Q.1.) f → k=3^ fully connected MLP

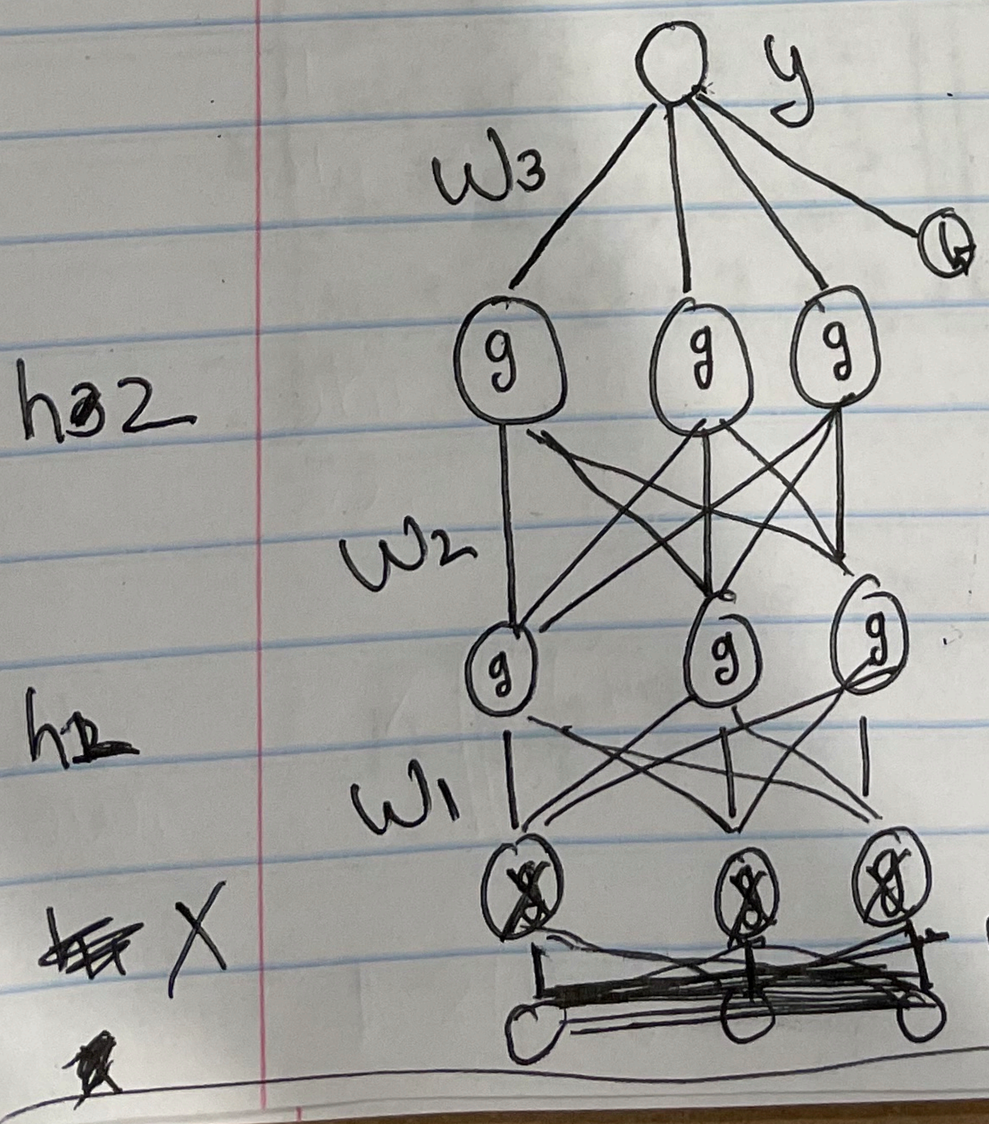
g is a generic 1-Lipschitz activation

X is a mx1 column vector

$$h_1 = g(W_1^T X + b_1) \qquad h_3 = g(W_3^T X + b_3)$$
$$h_2 = g(W_2^T X + b_2)$$
$$\hat{y} = W^T h_{n-1} + b_n \qquad \therefore \hat{y} = W^T h_2 + b_3$$

$$L(y, \hat{y}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

Find: $\frac{\partial L}{\partial W_3}, \frac{\partial L}{\partial W_2}, \frac{\partial L}{\partial W_1}, \frac{\partial L}{\partial b_3}, \frac{\partial L}{\partial b_2}, \frac{\partial L}{\partial b_1}$



$$\therefore \frac{\partial L}{\partial y} = 2(y_n - \hat{y}_n)$$

$$\frac{\partial y}{\partial W_3} = h_3$$

$$\therefore \frac{\partial L}{\partial W_3} = 2(y_n - \hat{y}_n) h_2$$

Prep

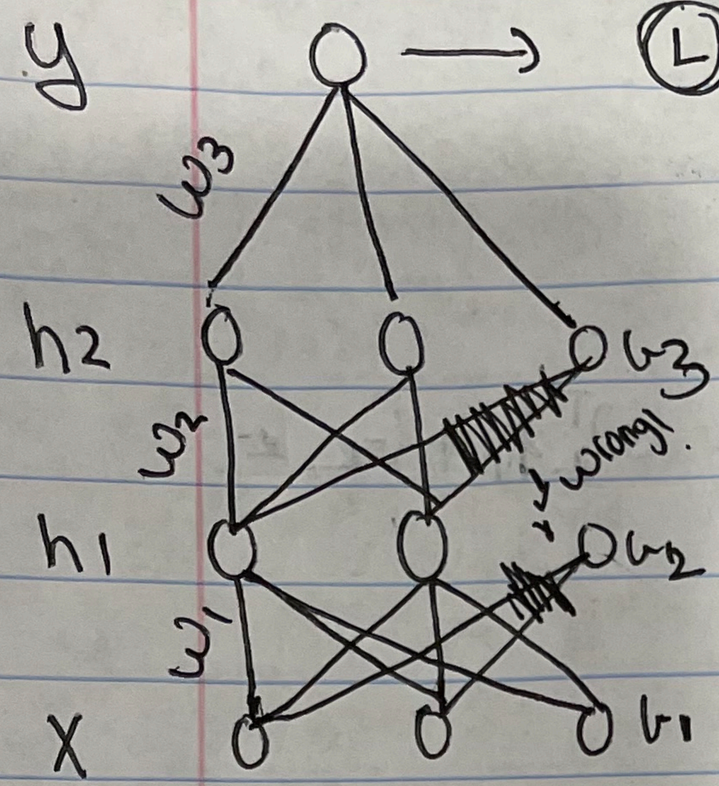
$$\frac{d}{dx}(a-x)^2 = 2(x-a)$$

a^2+x^2-2ax

$2x-2a = 2(x-a)$

let

$z_i = w_i^T h_{i-1} + b_i$



(L) $y = w_3^T h_2 + b_3$

$h_2 = g(w^T h_1 + b_2)$

$h_1 = g(w^T x + b_1)$

$L = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightarrow MSE$

$\frac{\partial L}{\partial \hat{y}} = 2(\hat{y} - y_i)$

$\frac{\partial y}{\partial w_3} = h_2 \quad \therefore \frac{\partial L}{\partial w_3} = 2(\hat{y} - y_i)h_2$

$\frac{\partial y}{\partial b_3} = 1 \quad \therefore \frac{\partial L}{\partial b_3} = 2(\hat{y} - y)$

$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h_2} \frac{\partial h_2}{\partial z_2} \quad \therefore \frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial h_2} \frac{\partial h_2}{\partial z_2} \frac{\partial z_2}{\partial w_2}$

$g = 1$ Lipshitz $\quad \frac{\partial z_2}{\partial w_2} = h_1; \frac{\partial h}{\partial z_2} = \frac{\partial g(z_2)}{\partial (z_2)}; \frac{\partial y}{\partial h_2} = w_3 \quad \frac{\partial L}{\partial y} = 2(\hat{y} - y)$

$\therefore \frac{\partial L}{\partial w_2} = h_1 \left(\frac{\partial g(z_2)}{\partial (z_2)} \cdot w_3 \cdot 2(\hat{y} - y) \right)^T$

$L = (\hat{y} - y)^2$ $\quad y = W_3^T h_2 + b_3 \quad h_2 = g(z_2) \quad z_2 = W_2^T h_1 + b_2$

$h_1 = g(z_1) \quad z_1 = W_1^T x + b_1$

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial y} \frac{dy}{dh_2} \frac{\partial h_2}{\partial z_2} \frac{\partial z_2}{\partial b_2}$$

$\frac{\partial z_2}{\partial b_2} = 1$ since $z_2 = W_2^T h_1 + b_2$

$\therefore \frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial h_2} W_3$

$\therefore \frac{\partial L}{\partial b_2} = \frac{\partial (g z_2)}{\partial z_2} \cdot W_3 \cdot 2(\hat{y} - y)$

Graidents W.R.T W_1, b_1

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial y} \frac{dy}{h_2} \frac{\partial h_2}{\partial z_2} \frac{\partial z_2}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial W_1}$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial y} \frac{dy}{\partial h_2} \frac{\partial h}{\partial z_2} \frac{\partial z_2}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial b_1}$$

$\frac{\partial h_1}{\partial z_1} = \frac{\partial g(z_1)}{\partial z_1} \quad \frac{\partial z_1}{\partial W_1} = X \quad \frac{\partial z_1}{\partial b_1} = 1$

$\therefore \frac{\partial L}{\partial W_1} = X \left(\left(2(\hat{y} - y) \cdot W_3 \cdot \frac{\partial (g z_2)}{\partial z_2} \right) \cdot W_2 \cdot \frac{\partial g(z_1)}{\partial (z_1)} \right)^T$

$\frac{\partial L}{\partial b_1} = 1 \cdot 2(\hat{y} - y) \cdot W_3 \cdot \frac{\partial g(z_2)}{\partial z_2} \cdot W_2 \cdot \frac{\partial g(z_1)}{\partial z_1}$

$$z_n = w^T x$$

let ~~$z_n = \sigma(w^T x)$~~ $\hat{y}_n = \sigma(z_n)$

(3.3) $L(D:w) = \sum -y_n \log(w^T x)) - (1 - y_n)(\log(1 - \sigma(w^T x))$

$$\therefore \frac{\partial L}{\partial w} = \frac{\partial L}{\partial z_n} \frac{\partial z_n}{\partial w} \qquad \frac{\partial z_n}{\partial w} = x_n$$

$$\frac{\partial L}{\partial z_n} = \left(-\frac{y_n}{\hat{y}_n} + \frac{1 - y_n}{1 - \hat{y}_n}\right) \hat{y}_n (1 - \hat{y}_n) \quad \text{(chain rule)}$$

$$= (\hat{y}_n - y_n)$$

~~$\frac{\partial z_n}{\partial w} =$~~

① $$\therefore \frac{\partial L}{\partial w} = \sum_{n=1}^{N} \frac{\partial L}{\partial z_n} \frac{\partial z_n}{\partial w} = (\hat{y}_n - y_n) x_n$$

② Hessian = second derivatives

$$H = \frac{\partial^2 L}{\partial w \partial w^T} \qquad \therefore H_{ij} = \sum_{n=1}^{N} \frac{\partial^2 L}{\partial w_i \partial w_j} = \sum_{n=1}^{N} \hat{y}_n (1 - y_n) x_{ni} x_{nj}$$

$\therefore$ The hessian $= X^T R X$ for $R =$ the diagonal matrix with entries $R_{nn} = \hat{y}_n (1 - \hat{y}_n)$

The diagonal approximation is just

$H_{ii} = \sum_{n=1}^{N} \hat{y}_n(1-\hat{y}_n)x_{ni}^2$ and for H_{ij} where $i=j:=0$