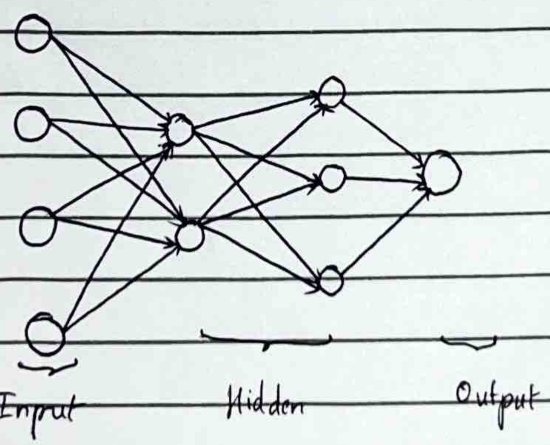
Assignment 4
WiDS

Problem 1.

Q1) For ~~d~~ building a neural network to recognize handwritten digits if we use a single output neuron with a linear activation function as it outputs the value as it is, we have no way of checking the probability and allowing the computer to decide.
whereas, with 10 output neurons with a softmax activation function which converts the output to a number between 0 and 1, we get an understanding of how probable each digit is. One-hot encoded targets clearly help us see which digit has been identified as only that is given the value '1' and the rest are '0'.

Problem 2.

(a)



(b) As there are 2 hidden layers,

$$g_1(x) = \text{ReLu}(W_1x + b_1) \quad \text{or} \quad \max(0, w_1x + b_1)$$

$$g_2(x) = \max(0, w_2(w_1x + b_1) + b_2)$$

$$f(x) = W_3 g_2(x) + b_3$$

$$= W_3(\max(0, w_2(w_1x + b_1) + b_2)) + b_3$$

(c) $X = \begin{bmatrix} 3 \\ -2 \\ 1 \\ 0 \end{bmatrix}_{4\times1}$

$W_1 = \begin{bmatrix} -1 & 1 & 0 & 2 \\ 0 & 1 & 1 & -1 \end{bmatrix}_{2\times4}$ $b_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}_{2\times1}$

$\cancel{g_1(x)}$ ∴ output $= \begin{bmatrix} -5 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ -1 \end{bmatrix}$

$g_1(x) = \text{ReLu(output)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$W_2 = \begin{bmatrix} 1 & 2 \\ 1 & 6 \\ -1 & 1 \end{bmatrix}$ $b_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$

output $= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$

$g_2(x)$ ReLU(output) $= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$w_3 = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}_{1\times3}$ $b_3 = [1]$

Output $f(x)$ $= 1 + 1 = \underline{\underline{2}}$.

(d) Parameters $= 4\times2 + 2\times3 + 3\times1$ (weights) $+ 2 + 3 + 1$ (biases)

$= 17 + 6 = \underline{\underline{23}}$.

Problem 3.

(a) $f(z) = \max(0, z)$

$$f(z) = \begin{cases} z & z > 0 \\ 0 & z \leq 0 \end{cases}$$

$$f'(z) = \begin{cases} 1 & z > 0 \\ 0 & z < 0 \end{cases}$$ (undefined at 0 ~~as LHL ≠ RHL~~)

(b) As the pre-activation sum z is negative, after passing through the ReLU activation function ($f(z)$) it will amount to 0

The gradient of the loss w.r.t w_i will also then be 0

$$\left(\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial f(z)} \cdot \underbrace{\frac{\partial f(z)}{\partial z}}_{0} \underbrace{\frac{\partial z}{\partial w_i}}_{x_i} \right)$$

(c) No, the weights of this neuron will never change during subsequent steps as $\frac{\partial L}{\partial w_i}$ will always be zero. ~~and~~

Problem 4.

(a) $J_{total}(\omega) = J_{data}(\omega) + \frac{\lambda \omega^2}{2}$

Gradient:

$$\frac{\partial J_{total}}{\partial \omega} = \frac{\partial J_{data}}{\partial \omega} + \lambda \omega$$

Gradient descent update:

$$\omega_{new} = \omega_{old} - \eta \frac{\partial J_{total}}{\partial \omega}$$

$$\therefore \omega_{new} = \omega_{old} - \eta\left(\frac{\partial J_{data}}{\partial \omega} + \lambda \omega_{old} \right)$$

Date ______
Page ______

(b) $$W_{new} = W_{old} - \eta \frac{\partial J_{data}}{\partial w} - n\lambda W_{old}.$$

$$W_{new} = \underbrace{(1 - n\lambda)}_{\downarrow} w_{old} - \eta \frac{\partial J_{data}}{\partial w}$$

$$0 < (1 - n\lambda) < 1$$

(c) L2 regularization is often referred to as "Weight Decay" as ~~it~~ each weight gets multiplied by a factor slightly less than 1 causing it to decay.