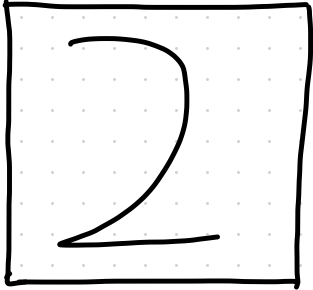


# RECENT SUCCESSES OF NN

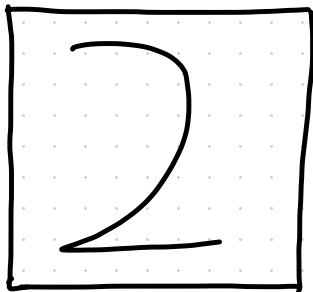
\* state-of-the-art (SOTA) in most fields

# PARADIGM CHANGE

# PARADIGM CHANGE

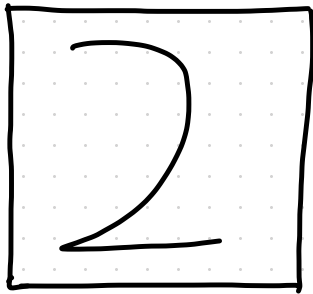


# PARADIGM CHANGE

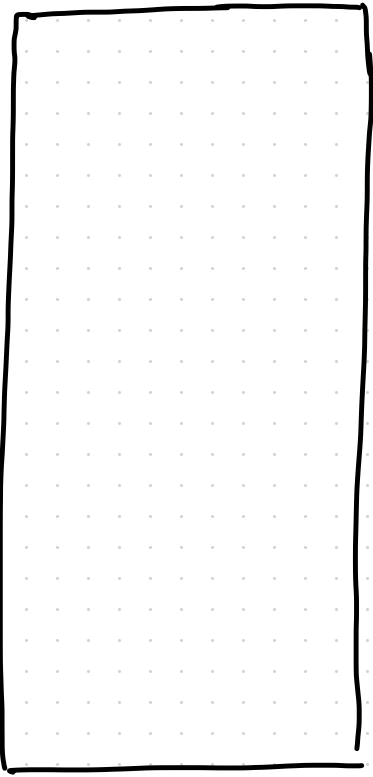


FEATURE  
→  
EXTRACTOR

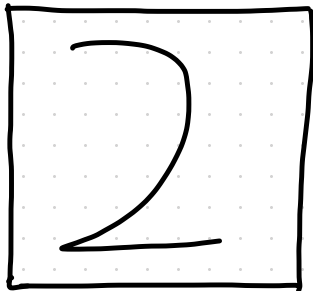
# PARADIGM CHANGE



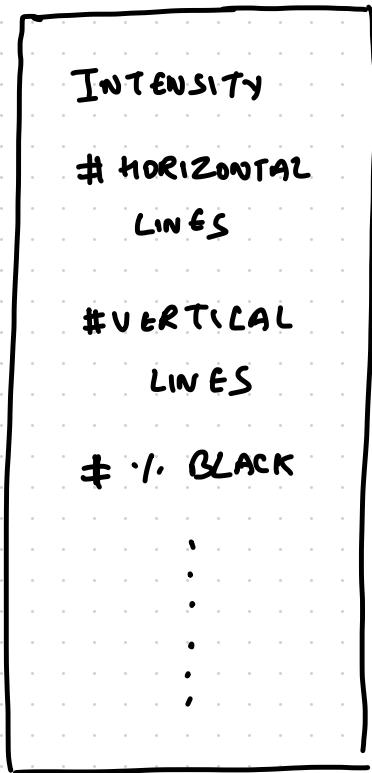
FEATURE  
→  
EXTRACTOR



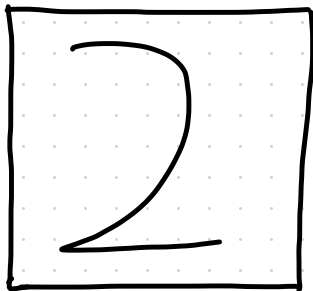
# PARADIGM CHANGE



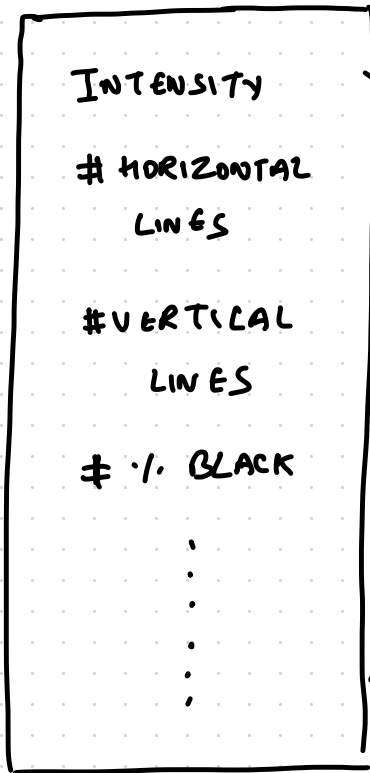
FEATURE  
→  
EXTRACTOR



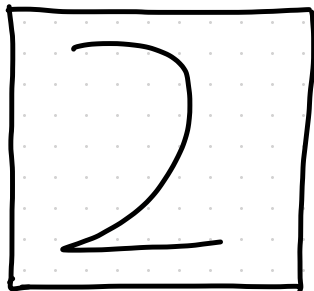
# PARADIGM CHANGE



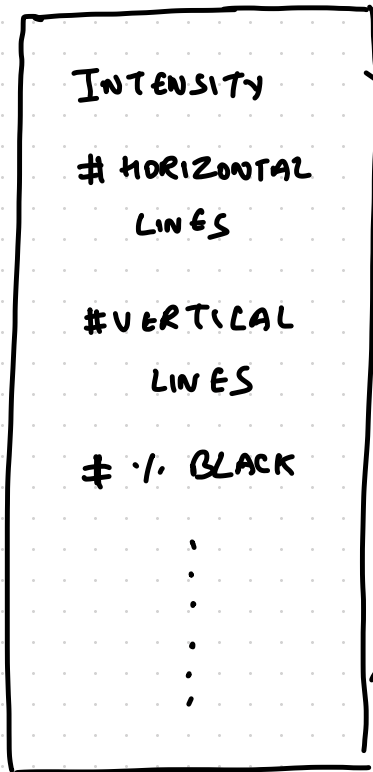
FEATURE  
→  
EXTRACTOR



# PARADIGM CHANGE



FEATURE  
→  
EXTRACTOR

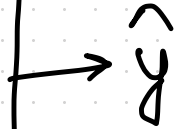
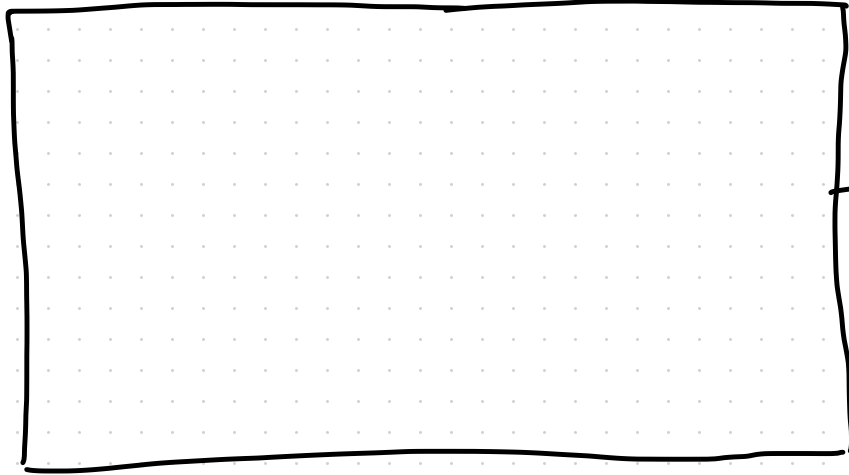
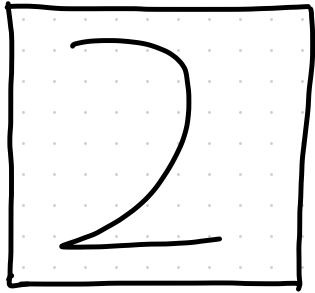


HAND CRAFTED

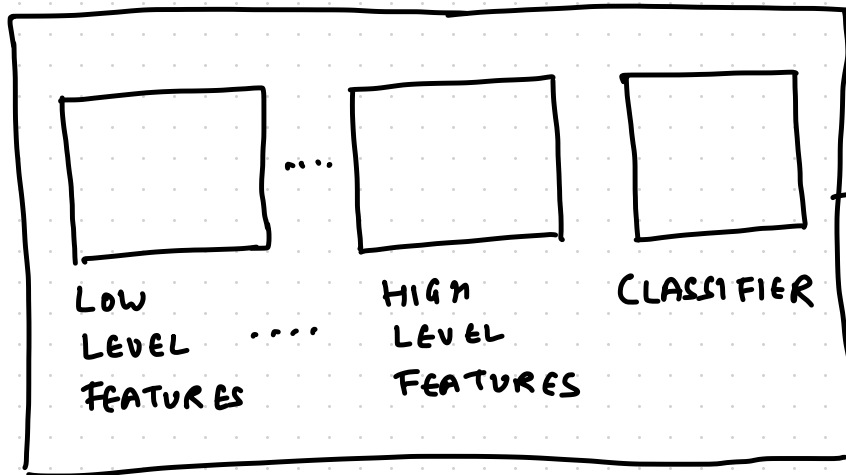
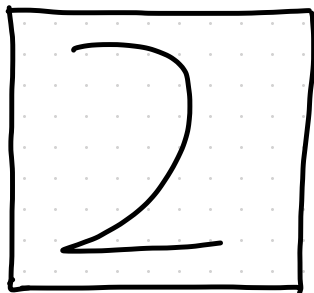




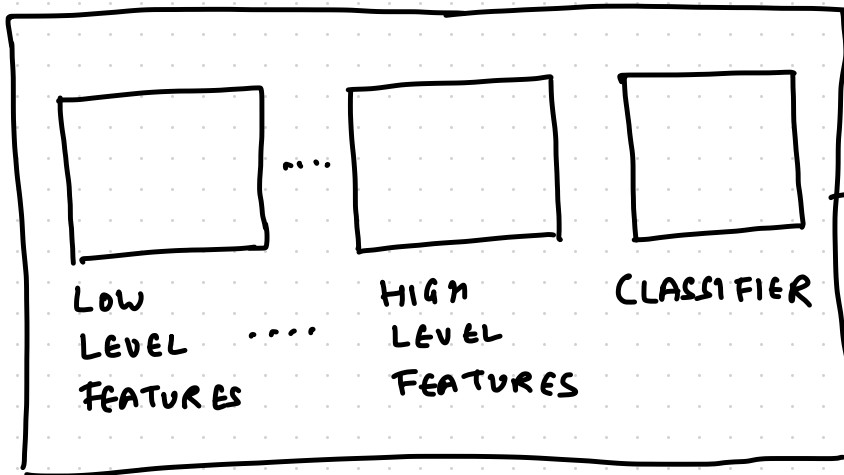
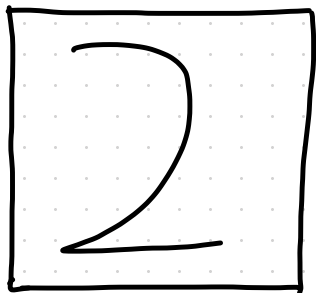
# PARADIGM CHANGE (NNS)



# PARADIGM CHANGE (NNS)



# PARADIGM CHANGE (NNS)

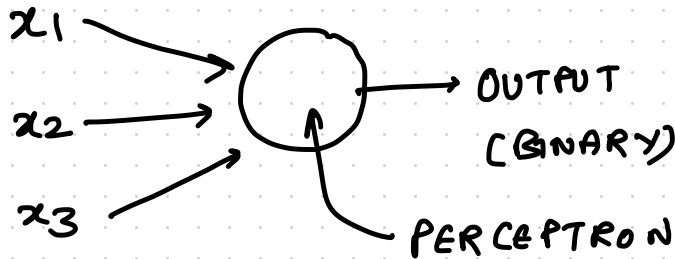


TRAINABLE

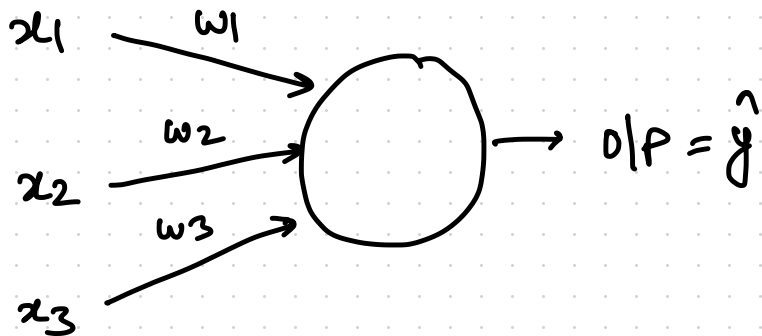
# PERCEPTRON

— ARTIFICIAL NEURON DEVELOPED BY  
ROSENBLATT IN 1960<sup>S</sup> INSPIRED BY  
MCCULLOCH & PITTS

BINARY I/P

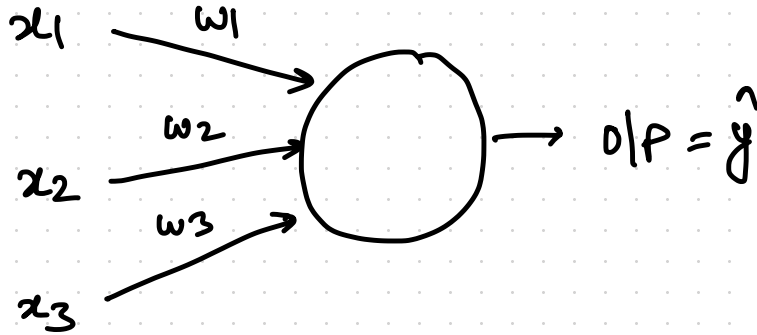


# PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

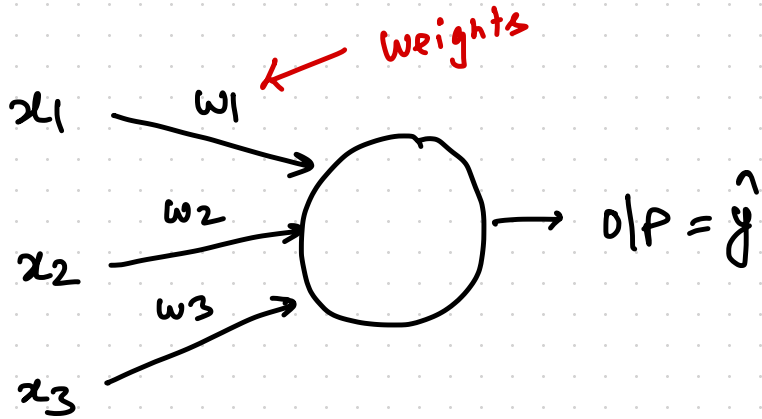
# PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

NEURONS "FIRE" ABOVE  
THRESHOLD

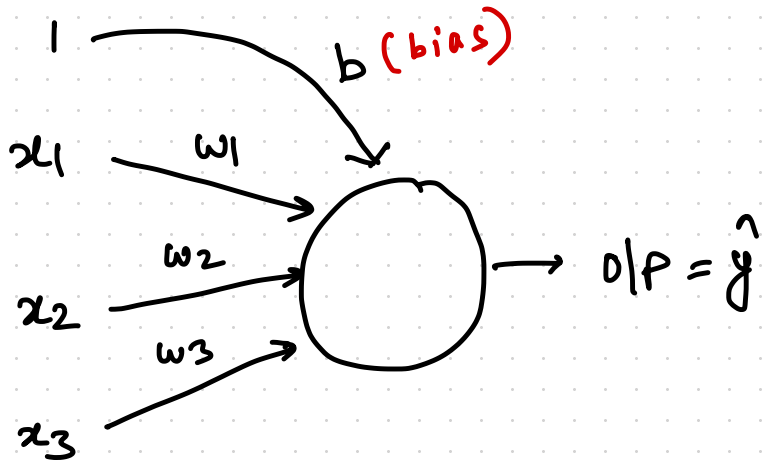
# PERCEPTRON



$$O/P = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i \leq \text{THRESHOLD} \\ 1 & ; \sum w_i x_i > \text{THRESHOLD} \end{cases}$$

NEURONS "FIRE" ABOVE  
THRESHOLD

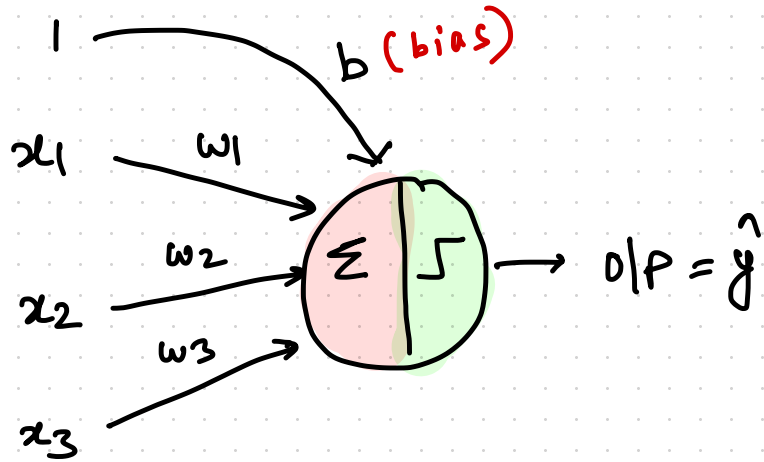
# PERCEPTRON



$$o/p = \hat{y} = \begin{cases} 0 & ; \sum w_i x_i + b \leq 0 \\ 1 & ; \sum w_i x_i + b > 0 \end{cases}$$



# PERCEPTRON

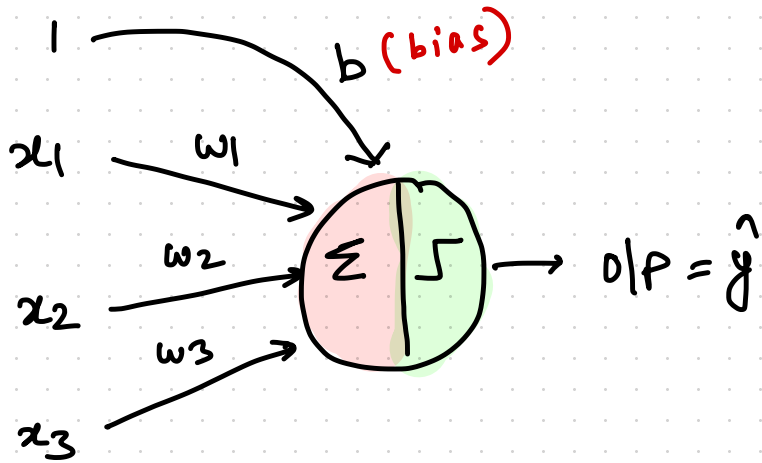


NEURON HAS 2 COMPONENTS

① SUMMATION

② ACTIVATION : STEP FUNCTION

# PERCEPTRON



NEURON HAS 2 COMPONENTS

- ① **SUMMATION**
- ② **ACTIVATION** : STEP FUNCTION

$\text{CSIGN}(\text{STEP})$   
Activation



# LEARNING BINARY GATES

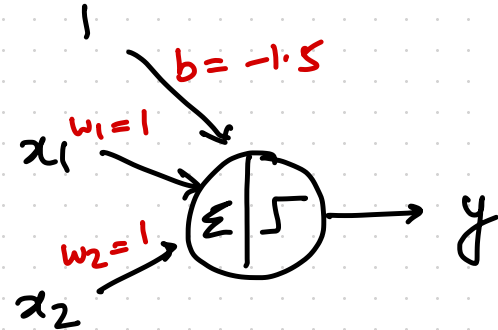
Q) FOR 2 IPs  $x_1$  &  $x_2$  learn  $w_1$ 's and  $b$  for  
BINARY AND

$x_1$	$x_2$	$y$
0	0	0
0	1	0
1	0	0
1	1	1

# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_1$ 's and  $b$  for  
BINARY AND

$x_1$	$x_2$	$y$
0	0	0
0	1	0
1	0	0
1	1	1



# LEARNING BINARY GATES

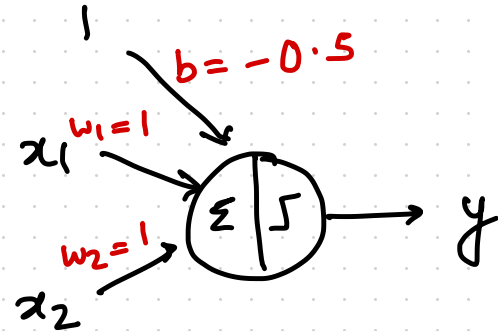
Q) FOR 2 IPs  $x_1$  &  $x_2$  learn  $w_1$ 's and  $b$  for  
BINARY OR

$x_1$	$x_2$	$y$
0	0	0
0	1	1
1	0	1
1	1	1

# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_i$ 's and  $b$  for BINARY OR

$x_1$	$x_2$	$y$
0	0	0
0	1	1
1	0	1
1	1	1



# LEARNING BINARY GATES

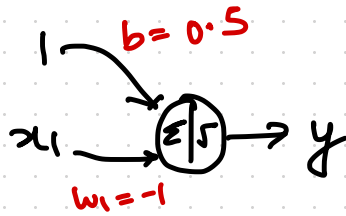
Q) FOR 1 I/Ps  $x_1$  learn  $w_1$  and  $b$  for  
UNARY NOT

$x_1$	$y$
0	1
1	0

# LEARNING BINARY GATES

Q) FOR 1 I/Os  $x_1$  learn  $w_1$ 's and  $b$  for UNARY NOT

$x_1$	$y$
0	1
1	0





# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_1$ 's and  $b$  for NAND

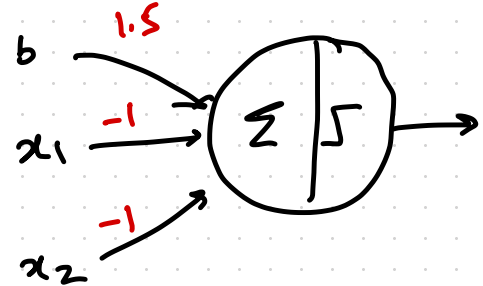
$x_1$	$x_2$	$y$
0	0	1
0	1	1
1	0	1
1	1	0

# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_1, w_2$  and  $b$  for NAND

$x_1$	$x_2$	$y$
0	0	1
0	1	1
1	0	1
1	1	0

APPROACH #1

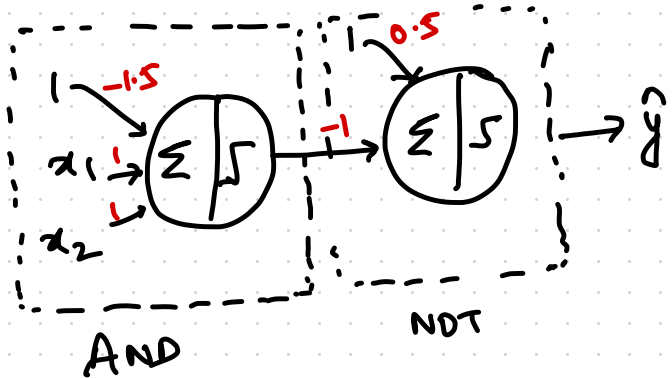
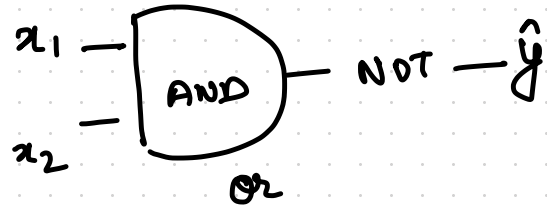


# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_1, w_2$  and  $b$  for NAND

$x_1$	$x_2$	$y$
0	0	1
0	1	1
1	0	1
1	1	0

APPROACH #2



# PERCEPTRON LEARNING ALGORITHM

Input:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $Y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 AUGMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:s] = X$  ;  $X'[s+1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 AUGMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:s] = X$  ;  $x'[1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

S2 INIT  $W \in \mathbb{R}^{D+1}$

# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 ARGUMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:s] = X$  ;  $x'[1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

S2 INIT  $w \in \mathbb{R}^D$

S3 FOR  $I$  IN  $IT$ :

# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 AUGMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:s] = X$  ;  $X'[1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

S2 INIT  $W \in \mathbb{R}^{D+1}$

S3 FOR  $I$  IN  $IT$ :

S3.1 FOR  $n$  in  $N$  :

S3.1.1  $\hat{y} = \text{STEP}(X' \cdot W)$

S3.1.2  $ERR = y - \hat{y}$

S3.1.3  $W \leftarrow W + lr * ERR_n * X'[n]$



# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 AUGMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:s] = X$  ;  $X'[s+1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

S2 INIT  $W \in \mathbb{R}^{D+1}$

S3 FOR  $I$  IN  $IT$ :

S3.1 FOR  $n$  in  $N$ :

S3.1.1  $\hat{y} = \text{STEP}(X'_n \cdot W)$

S3.1.2

$ERR = y - \hat{y}$

S3.1.3

$W \leftarrow W + lr * ERR_n * X'_n$

ERROR IN  $n^{\text{th}}$  sample

$n^{\text{th}}$  sample

# PERCEPTRON LEARNING ALGORITHM

I/P:  $X \rightarrow N \times D$  ;  $y \rightarrow N \times 1$  ;  $lr \rightarrow$  learning rate ;  $it \rightarrow$  #iterations

S1 AUGMENT  $X$  to  $X'$   $\rightarrow$  s.t.  $X'[1:D] = X$  ;  $X'[1] = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

S2 INIT  $W \in \mathbb{R}^{D+1}$

S3 FOR  $I$  IN  $IT$ :

S3.1 FOR  $n$  in  $N$  : ← Analogous to S.G.D

S3.1.1  $\hat{y} = \text{STEP}(x' \cdot W)$

S3.1.2  $ERR = y - \hat{y}$  ← ERROR IN  $n^{\text{th}}$  sample

S3.1.3  $W \leftarrow W + lr * ERR_n * x'[n]$  ←  $n^{\text{th}}$  sample

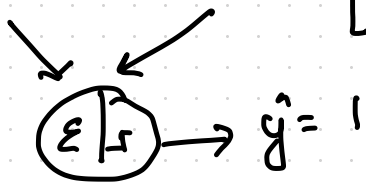
# PERCEPTRON LEARNING ALGORITHM

$$\underline{s31.3} \quad W \leftarrow W + lr * ERR_n * x'[n]$$

IMAGINE

$$W = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$; x'[n] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} ; y = 0$$



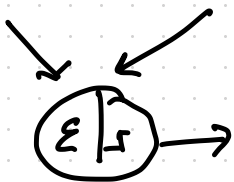
# PERCEPTRON LEARNING ALGORITHM

$$\underline{S31.3} \quad W \leftarrow W + lr * ERR_n * x^i[n]$$

IMAGINE

$$W = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$; x^i[n] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} ; y = 0$$



$$\hat{y} = 1$$

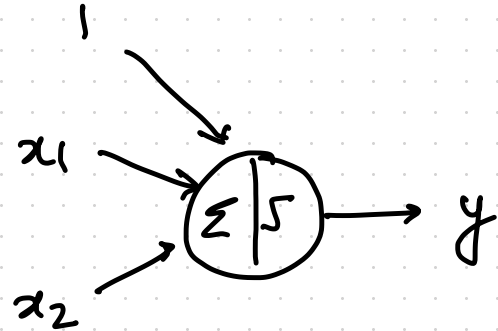
$$ERR_n = 0 - 1 = -1$$

$$W \leftarrow \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 0.01 * -1 * \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$
$$\leftarrow \begin{bmatrix} .99 \\ 1 \\ 1 \end{bmatrix}$$

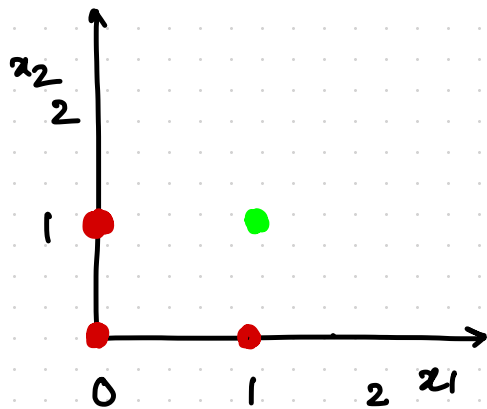
# LEARNING BINARY GATES

Q) FOR 2 I/Ps  $x_1$  &  $x_2$  learn  $w_i$ 's and  $b$  for BINARY XOR

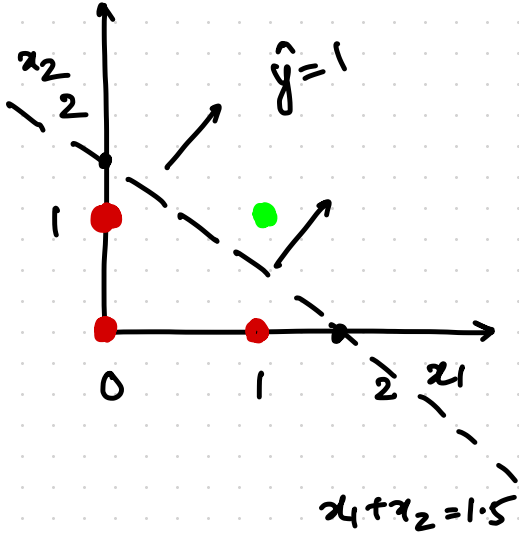
$x_1$	$x_2$	$y$
0	0	0
0	1	1
1	0	1
1	1	0



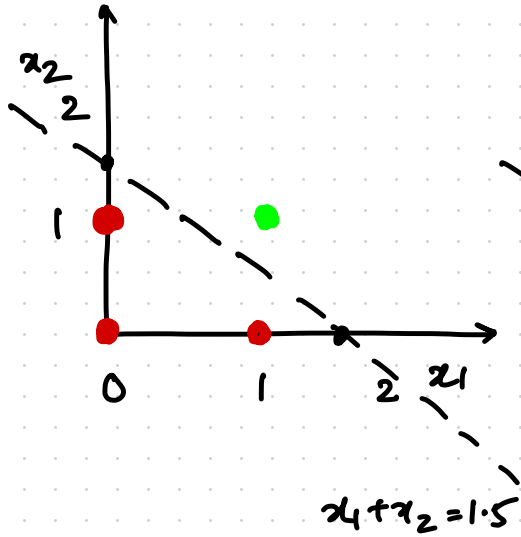
AND



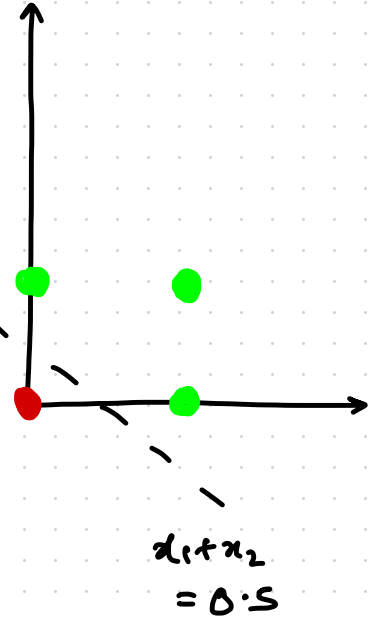
AND



AND

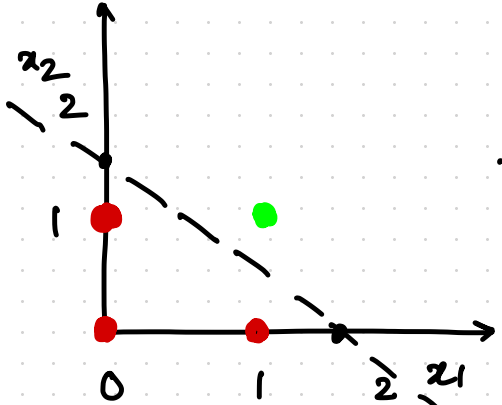


OR



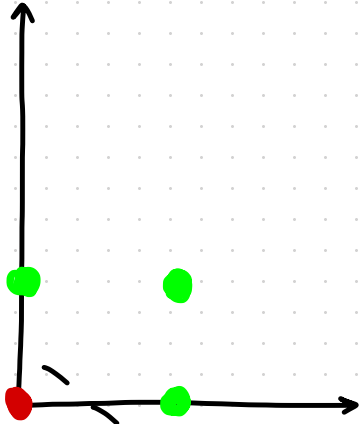


AND



$$x_1 + x_2 = 1.5$$

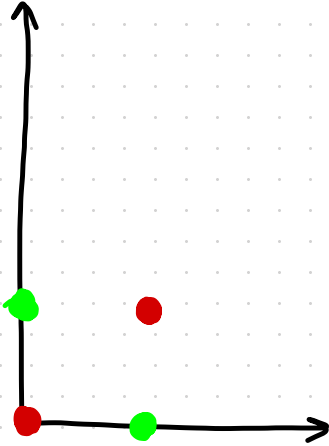
OR



$$x_1 + x_2 = 0.5$$

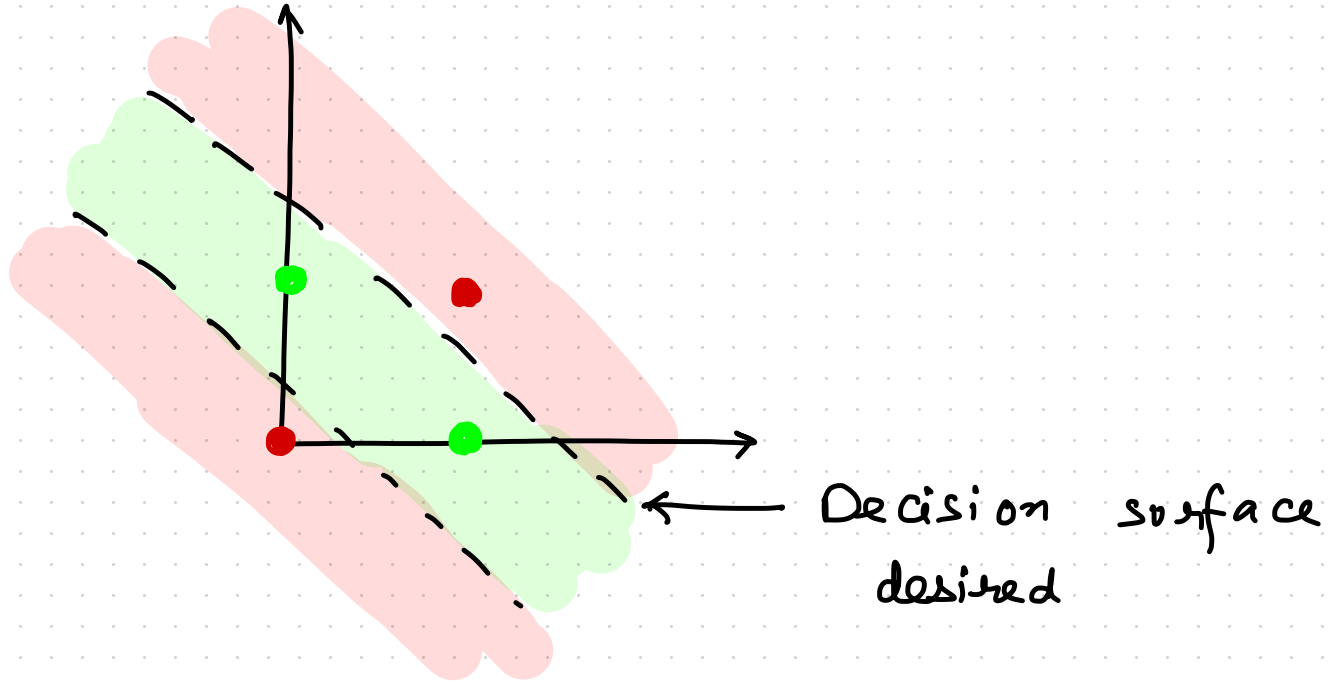
LINEARLY SEPARABLE

XOR

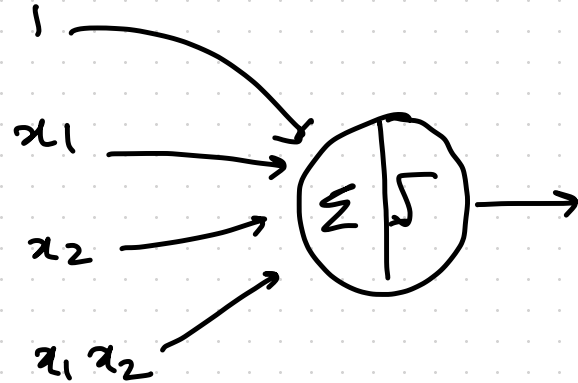
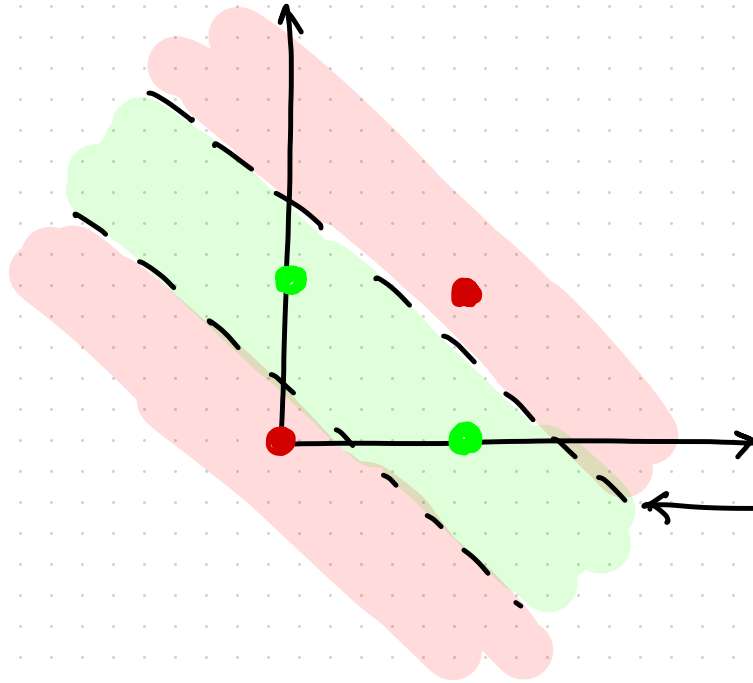


NOT  
SEPARABLE  
LINEARLY

# XDR CLASSIFICATION

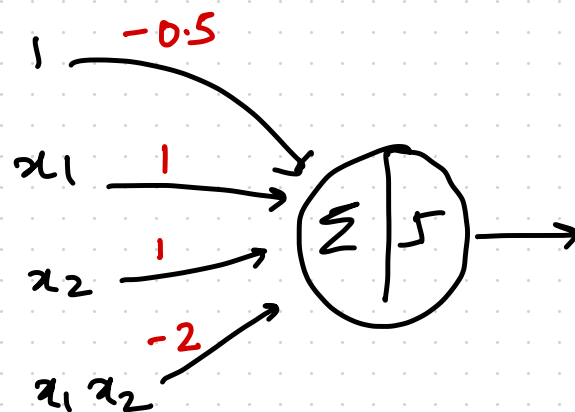
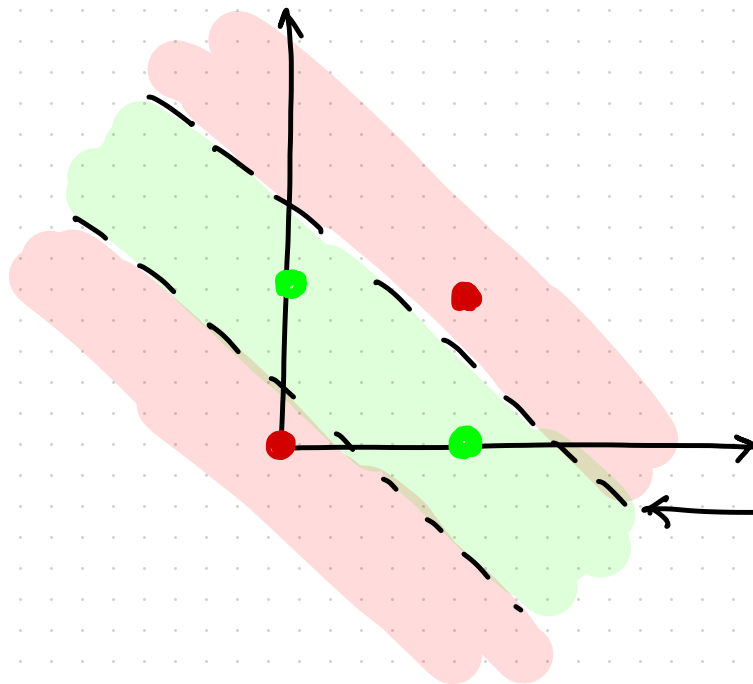


# XDR CLASSIFICATION



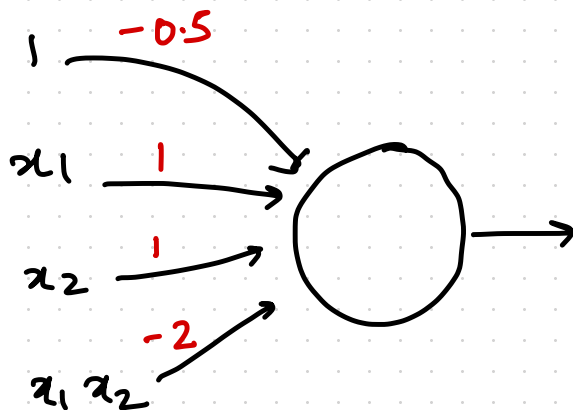
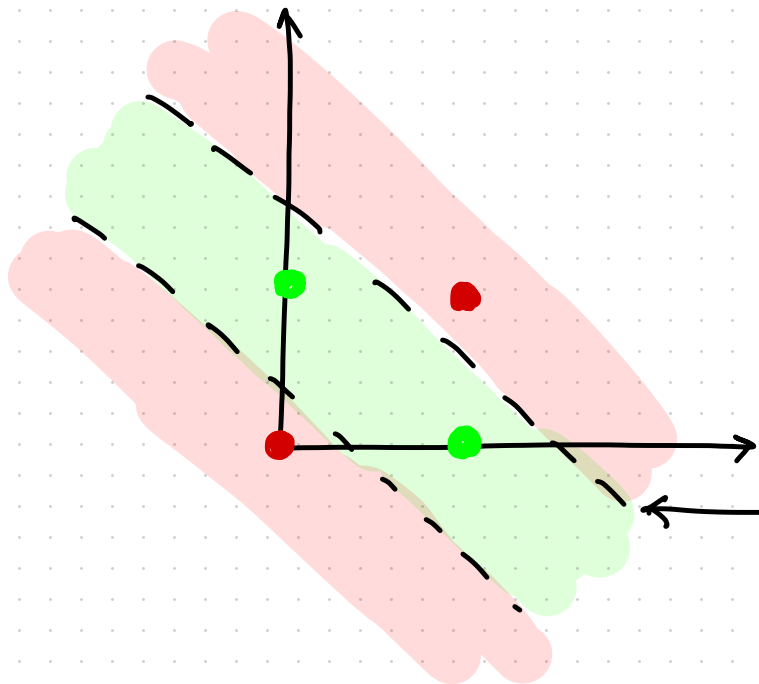
Decision surface desired

# XOR CLASSIFICATION



Decision surface desired

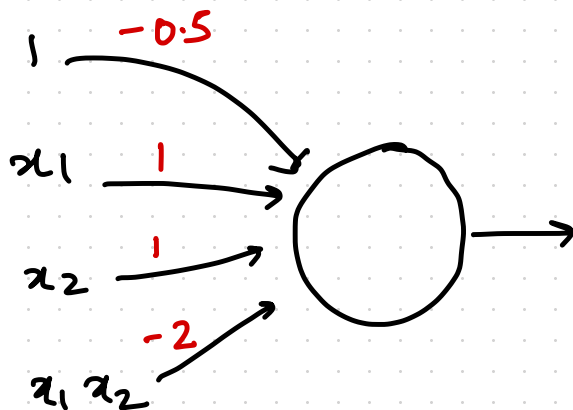
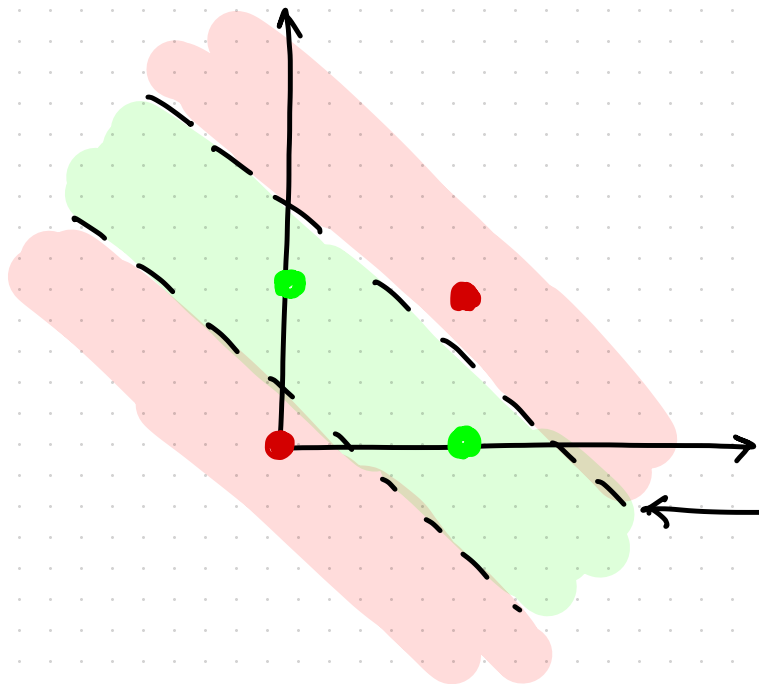
# XDR CLASSIFICATION



FOR  $x_1 = 0; x_2 = 0$ ; we get  
 $\hat{y} = \{-0.5 \leq 0\} = \text{RED CLASS}$

Decision surface  
desired

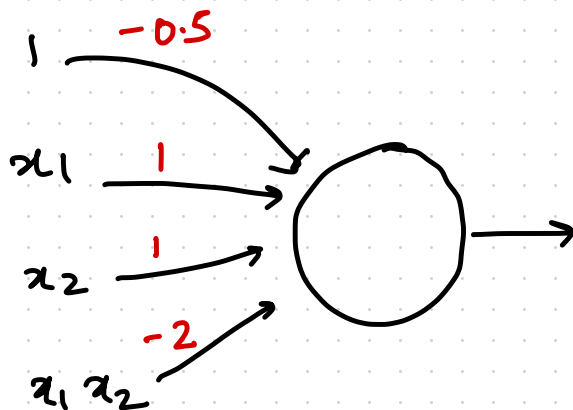
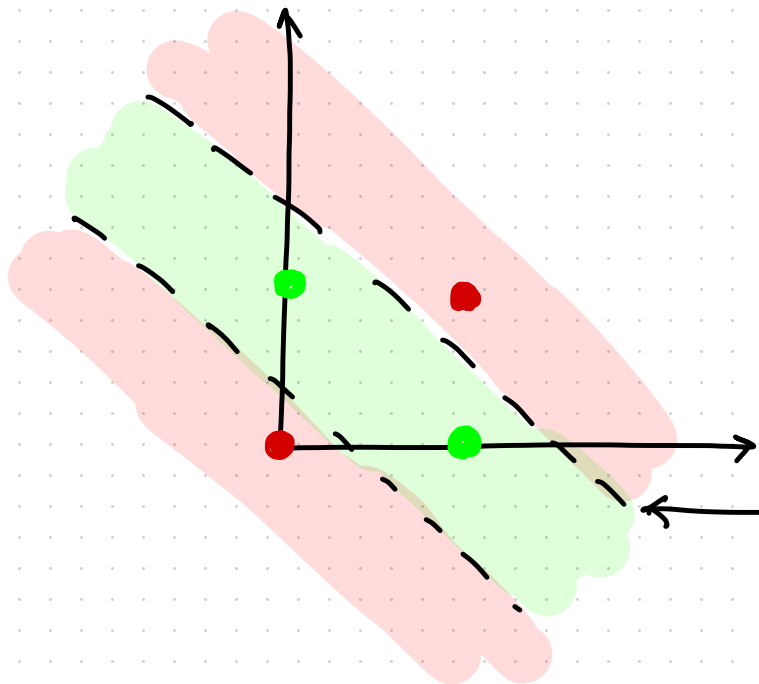
# XDR CLASSIFICATION



FOR  $x_1 = 0 ; x_2 = 1 ;$  we get  
 $\hat{y} = \{-0.5 + 1 \leq 0\} = \text{GREEN CLASS}$

Decision surface desired

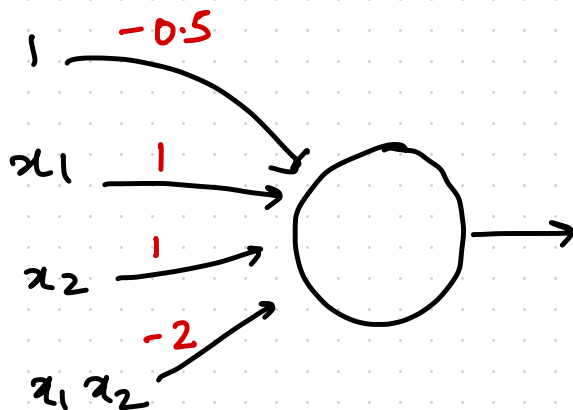
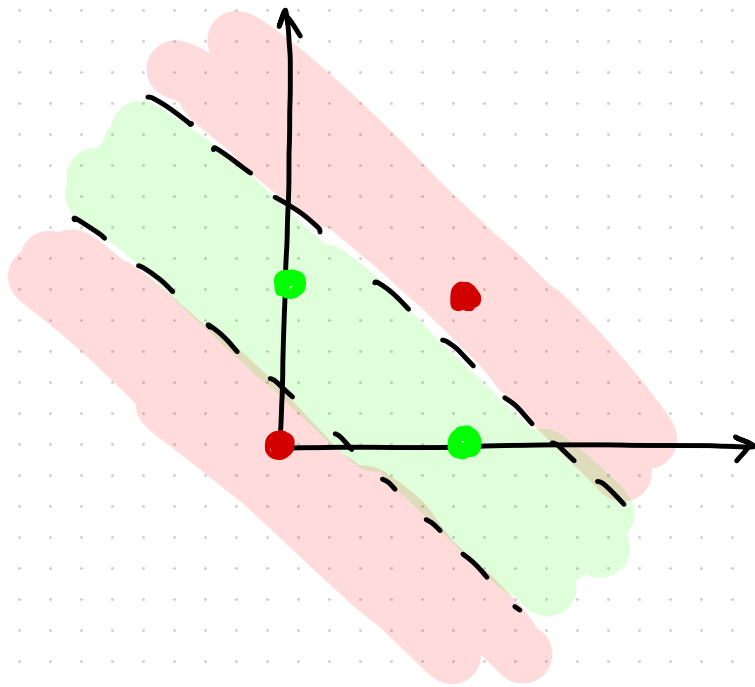
# XDR CLASSIFICATION



FOR  $x_1 = 1$ ;  $x_2 = 0$ ; we get  
 $\hat{y} = \{0.5 \leq 0\} = \text{GREEN CLASS}$

Decision surface  
desired

# XOR CLASSIFICATION

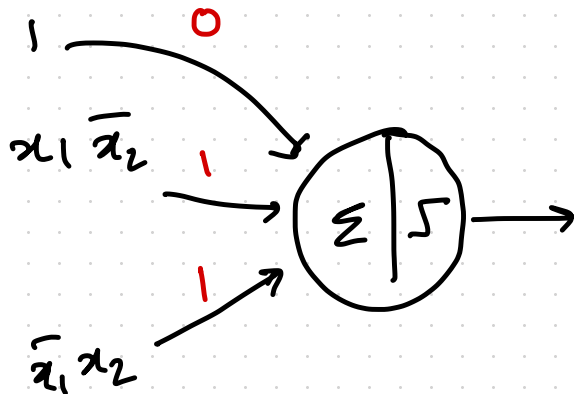
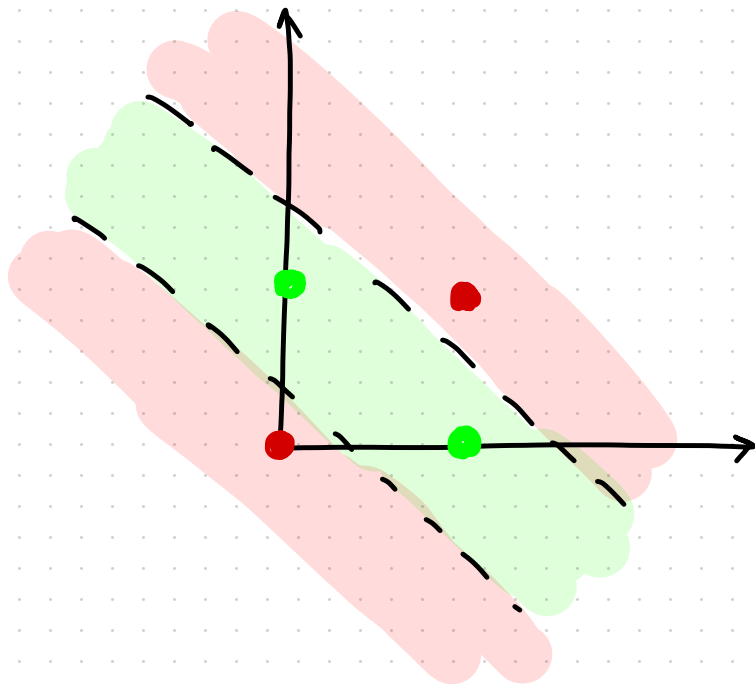


FOR  $x_1 = 1$  ;  $x_2 = 1$  ; we get  
 $\hat{y} = \{-0.5 \leq 0\} = \text{RED CLASS}$

CAN ADD NON-LINEARITY  
BY HAND-CRAFTING  
FEATURES !

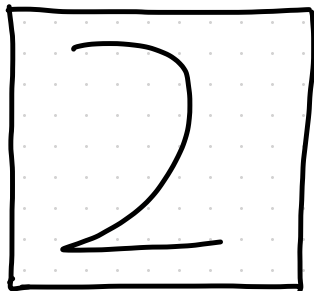


# XOR CLASSIFICATION

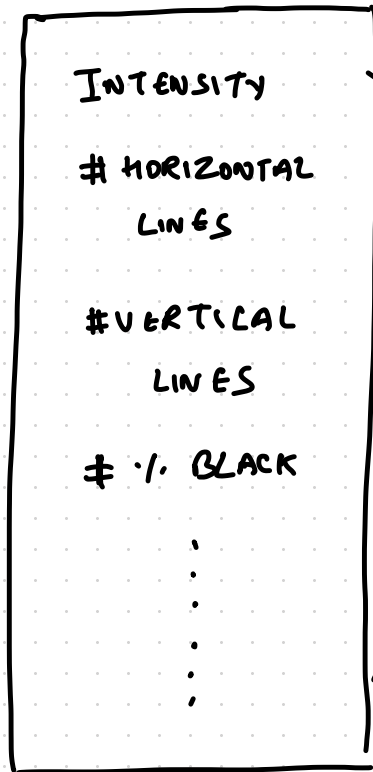


CAN ADD NON-LINEARITY  
BY HAND-CRAFTING  
FEATURES!

# PARADIGM CHANGE



FEATURE  
→  
EXTRACTOR



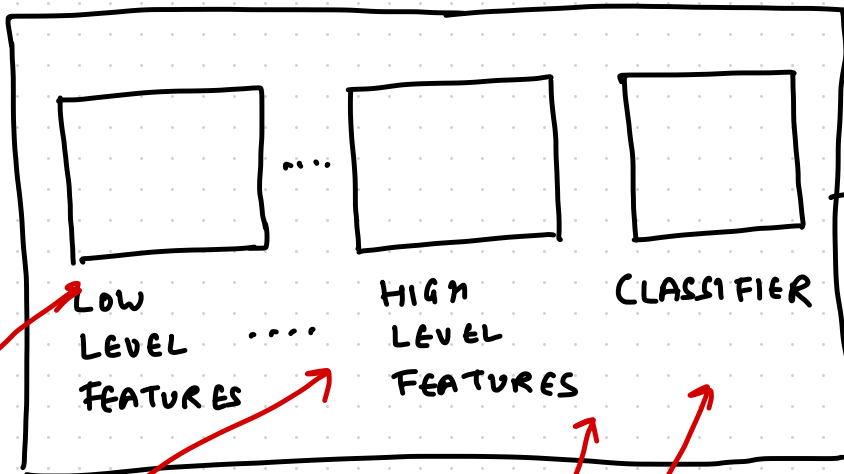
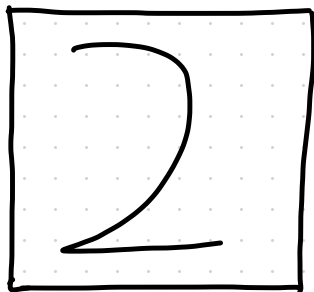
HAND CRAFTED



COULD WE  
DO BETTER?

---

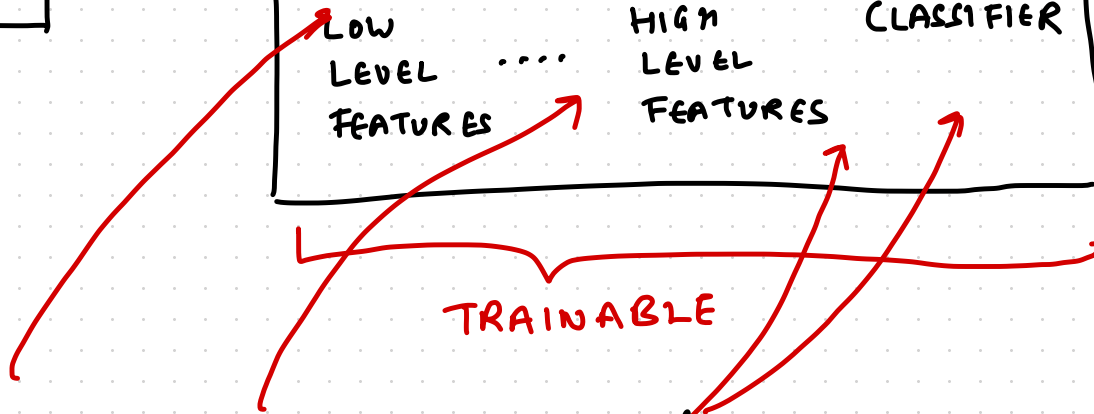
# PARADIGM CHANGE (NNs)



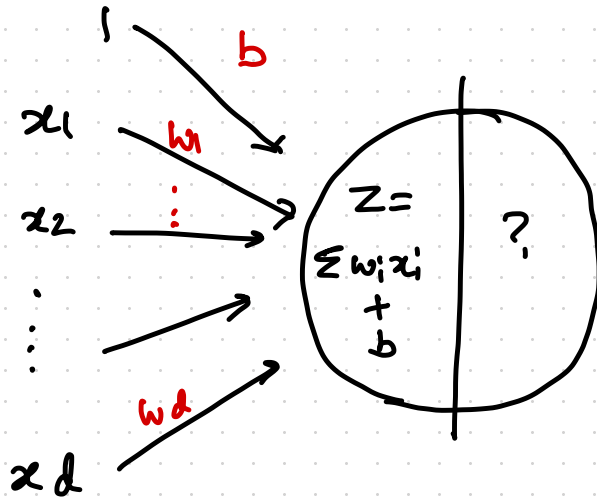
WE NEED

NON-LINEARITY

TRAINABLE

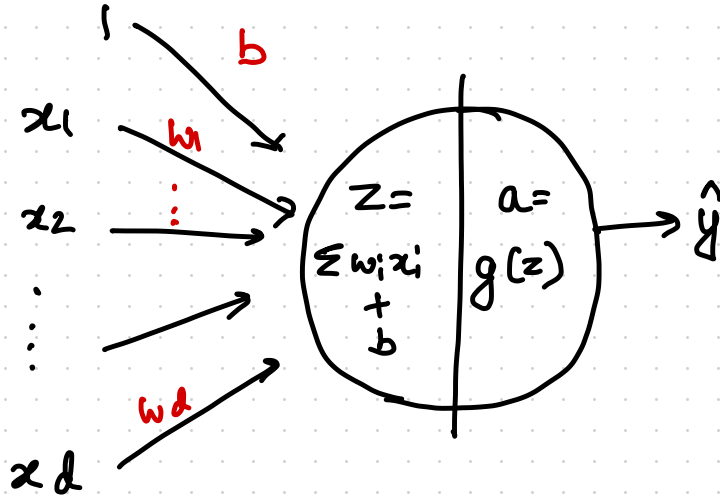


# BACK PROPAGATION SUPPORTED ACTIVATIONS



key idea: use  
activation<sup>n</sup>  
similar to  $\sqrt{\quad}$   
but differentiable

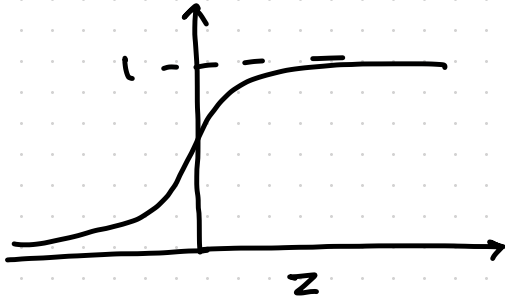
# ADDIN A NON-LINEARITY



$g(z)$ : NON LINEAR  
TRANSFORMATION

# ACTIVATION FUNCTIONS

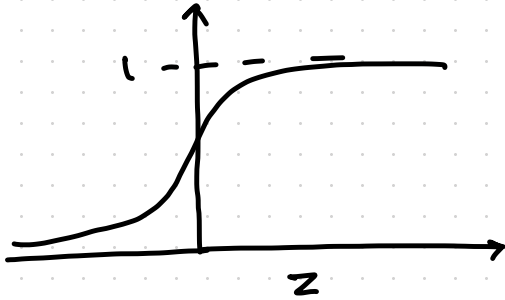
SIGMOID



$$g(z) = \frac{1}{1 + e^{-z}}$$

# ACTIVATION FUNCTIONS

SIGMOID



$$g(z) = \frac{1}{1 + e^{-z}}$$

Q): If we have 1 neuron

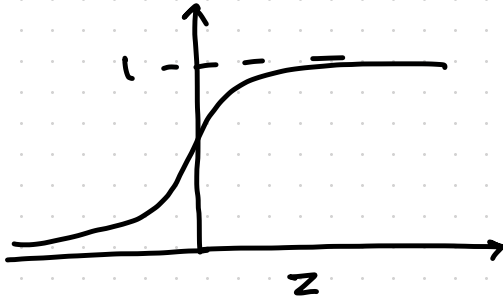
↳

$$g(z) = \frac{1}{1 + e^{-z}}$$

what do we get?

# ACTIVATION FUNCTIONS

SIGMOID



$$g(z) = \frac{1}{1 + e^{-z}}$$

Q): If we have 1 neuron

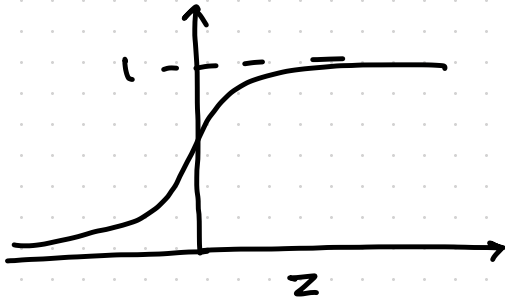
↳  
 $g(z) = \frac{1}{1 + e^{-z}}$  what do we  
get?

Logistic Regression



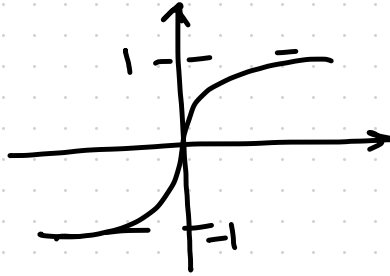
# ACTIVATION FUNCTIONS

SIGMOID



$$g(z) = \frac{1}{1 + e^{-z}}$$

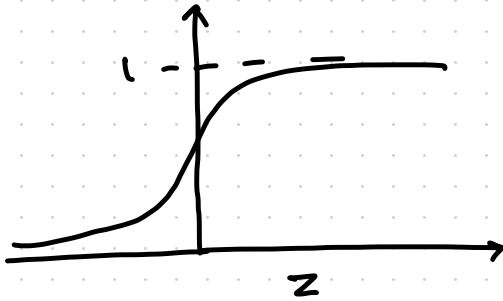
TANH



$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

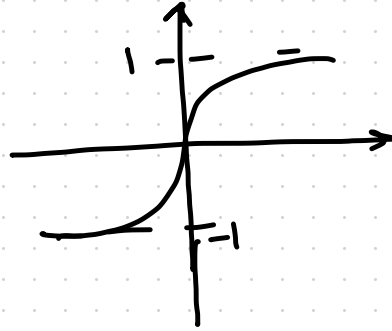
# ACTIVATION FUNCTIONS

SIGMOID



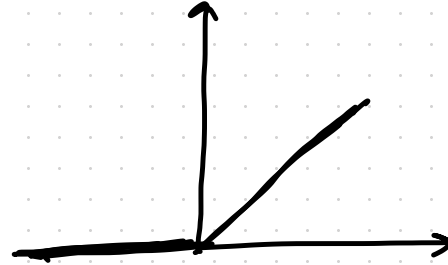
$$g(z) = \frac{1}{1 + e^{-z}}$$

TANH



$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

ReLU



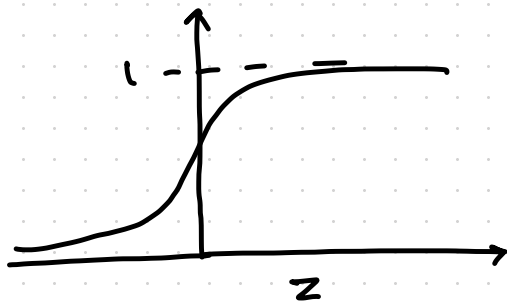
$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & \text{otherwise} \end{cases}$$

or

$$g(z) = \max(0, z)$$

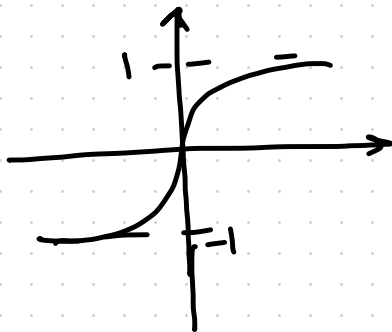
# ACTIVATION FUNCTIONS

SIGMOID



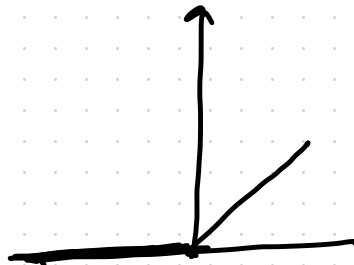
$$g(z) = \frac{1}{1 + e^{-z}}$$

TANH



$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

ReLU

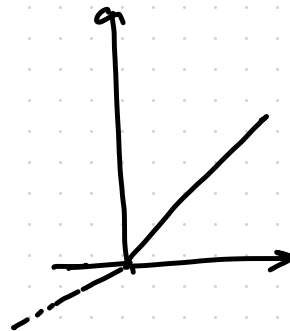


$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & \text{otherwise} \end{cases}$$

or

$$g(z) = \text{MAX}(0, z)$$

Leaky ReLU

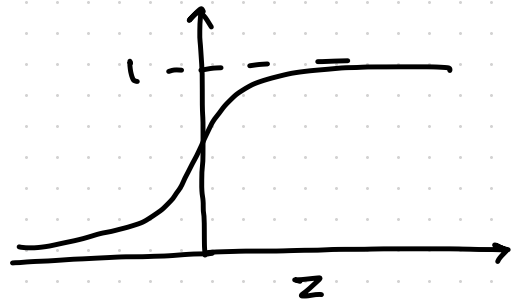


$$g(z) = \text{MAX}(\alpha z, z)$$

$\alpha \rightarrow 0$

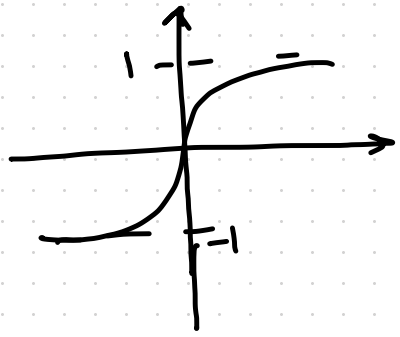
# ACTIVATION FUNCTIONS

SIGMOID



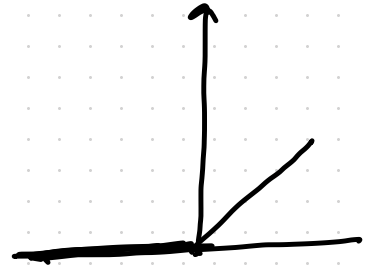
USEFUL FOR  
PROBABILISTIC  
ESTIMATES  
∴ B/w 0 & 1

TANH



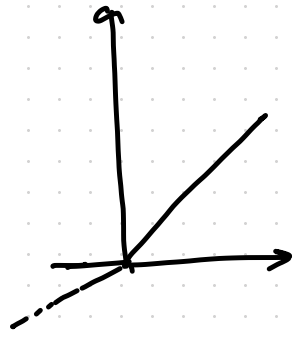
USEFUL IF  
DATA TRANSFORMED  
WITH MEAN 0

ReLU



GAMES  
CHANGER  
(Default)  
Good  
learning  
for  $|z| = \text{high}$ .

Leaky ReLU



Similar  
to  
ReLU  
Learns  
for  $z < 0$   
also

# DESIRABLE ATTRIBUTES OF ACTIVATION FUNCTIONS

1) NON-LINEAR

2) (MOSTLY) SMALL CHANGE IN I/P  $\Rightarrow$  SMALL CHANGE IN O/P

# 1 LAYER PERCEPTRON (NN)

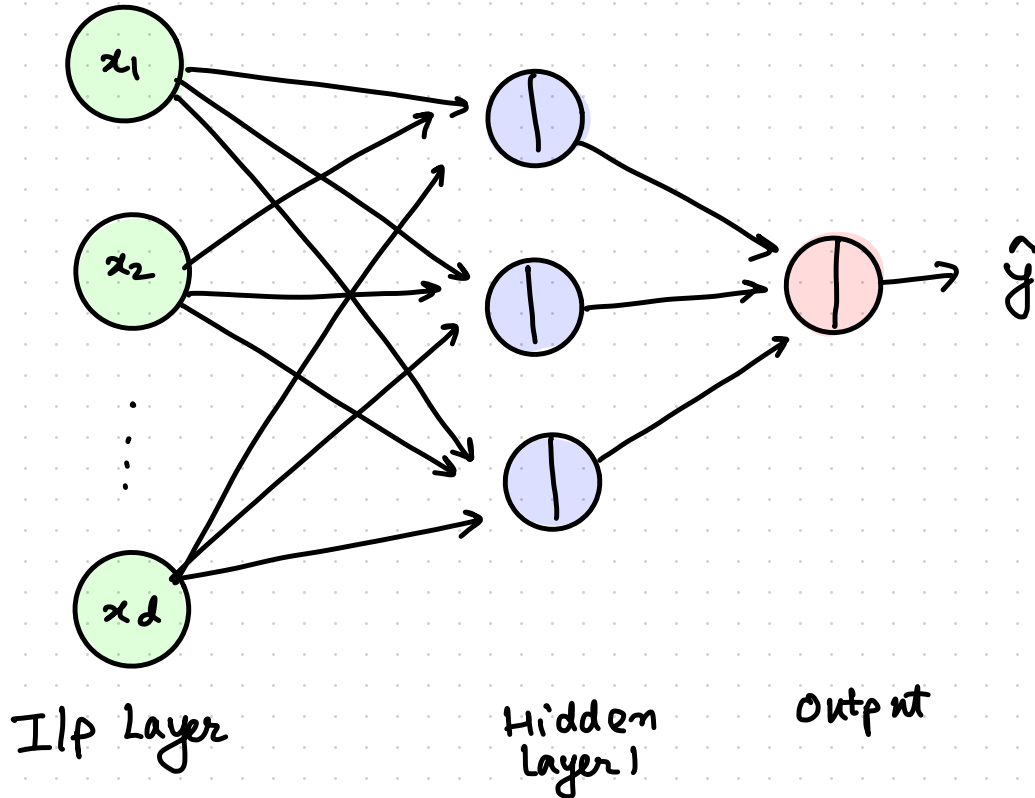


⋮

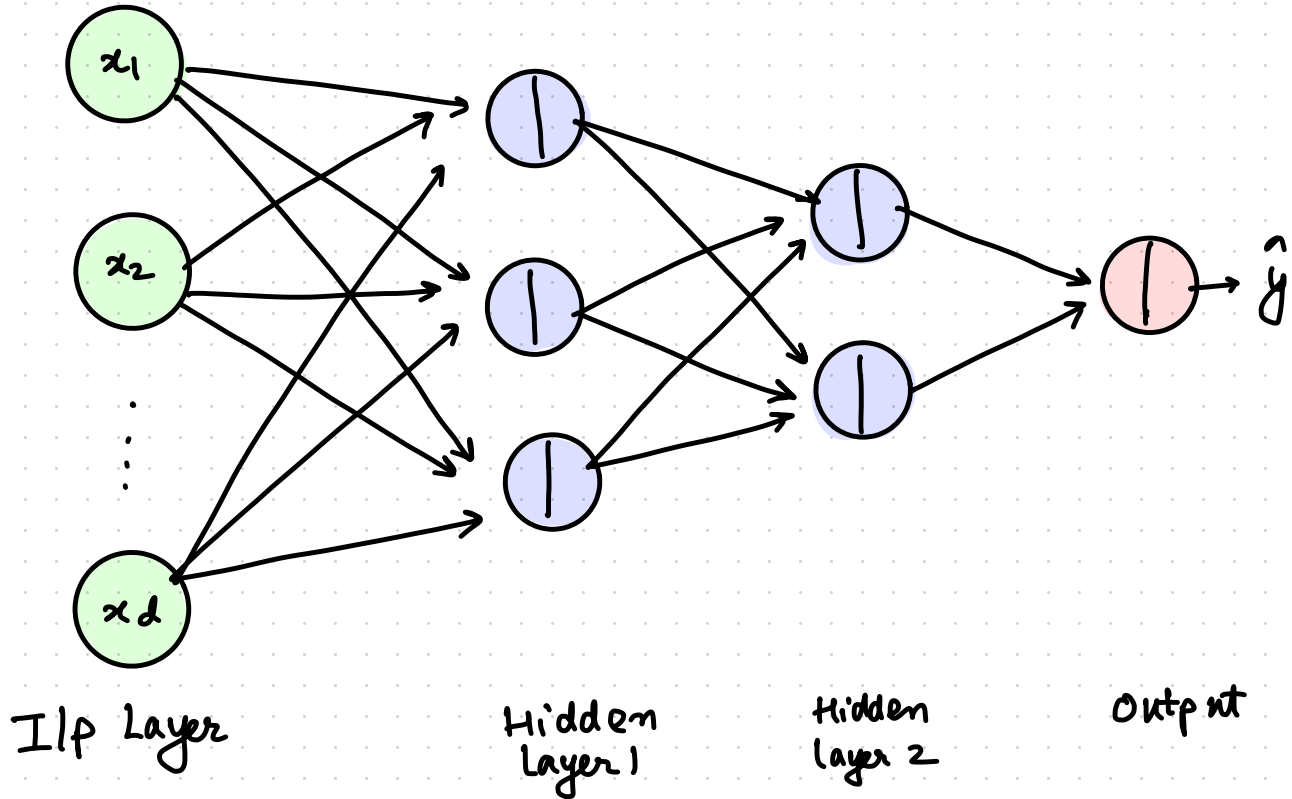


Input Layer

# 1-LAYER PERCEPTRON (NN)

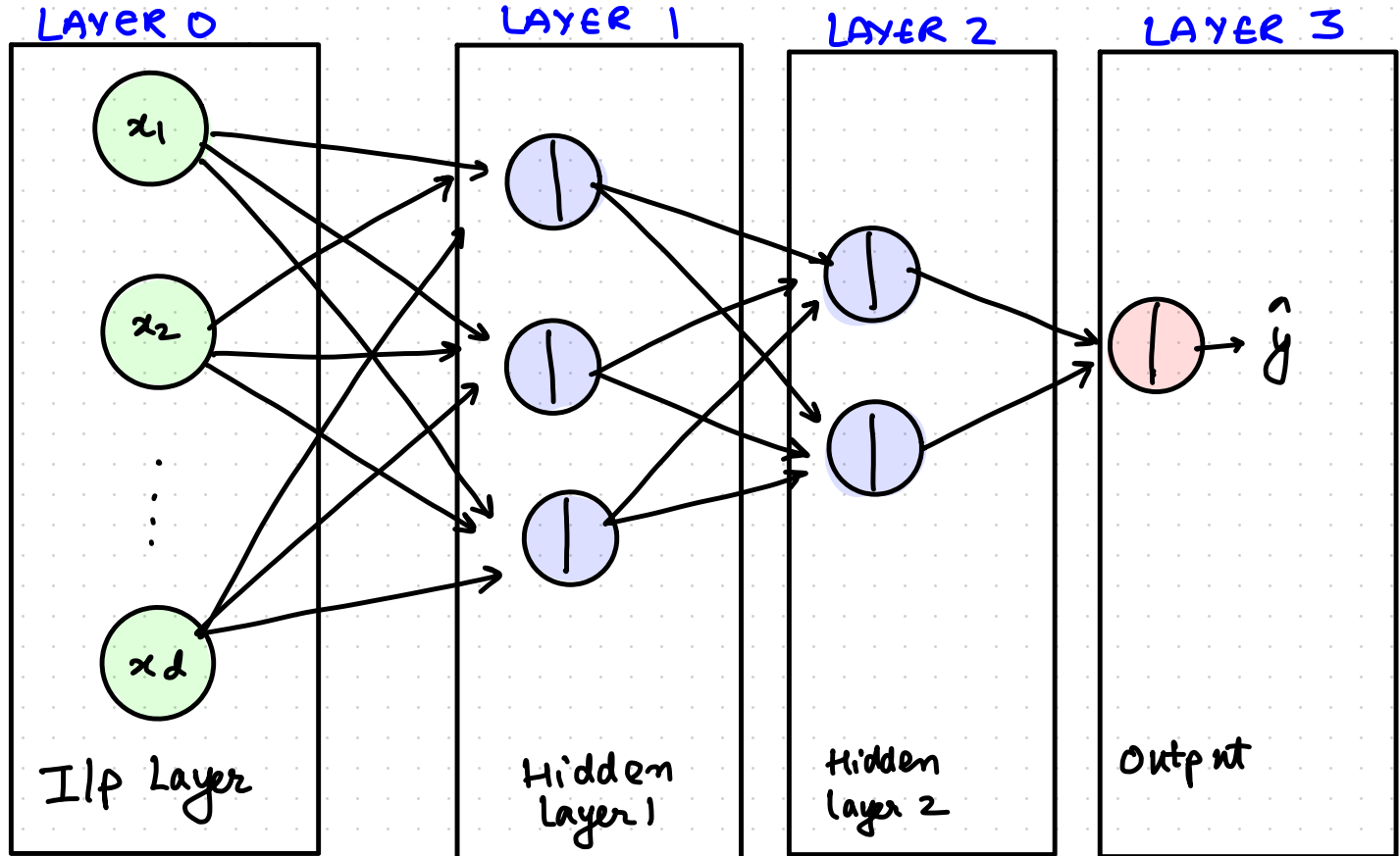


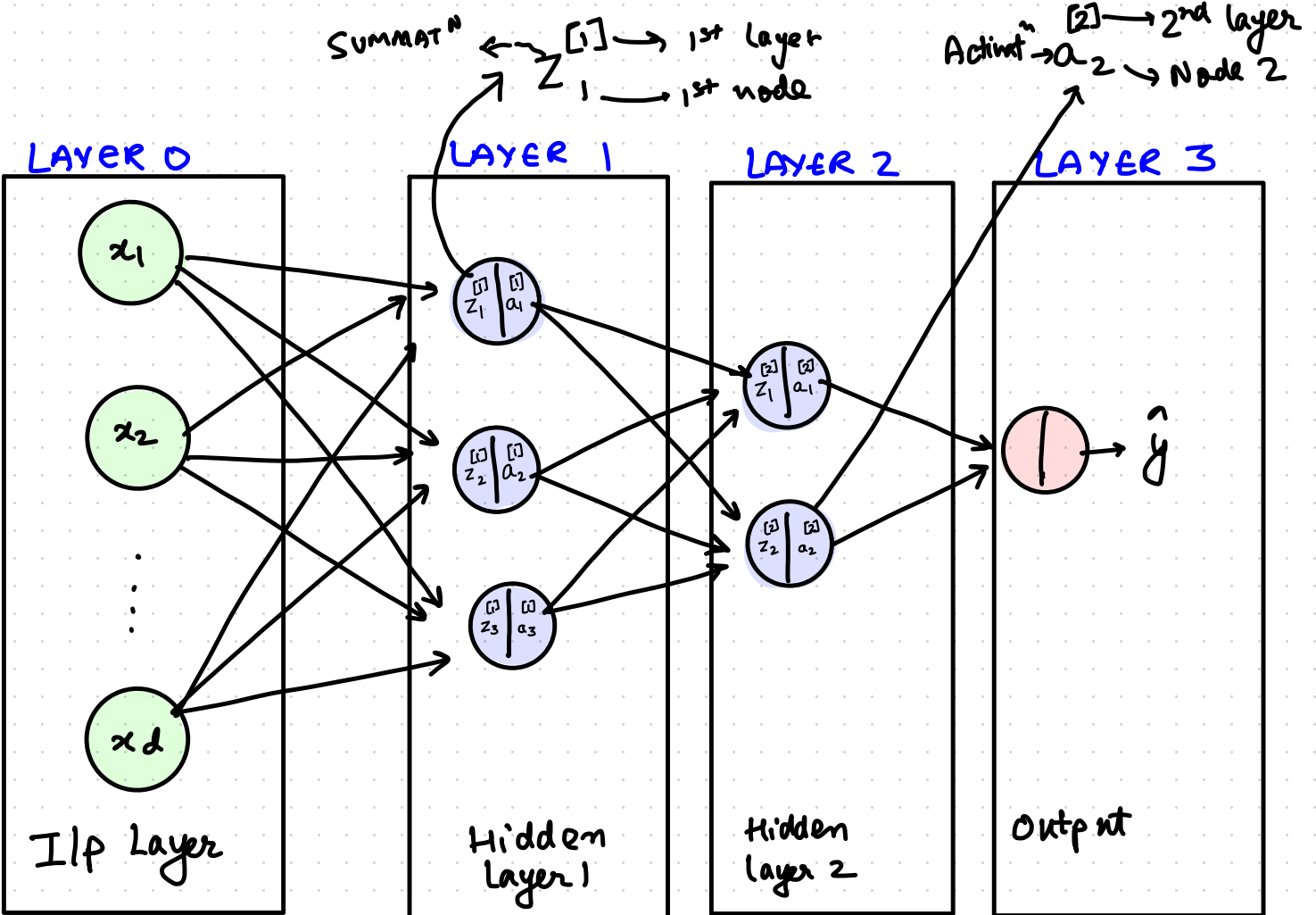
# MULTI-LAYER PERCEPTRON

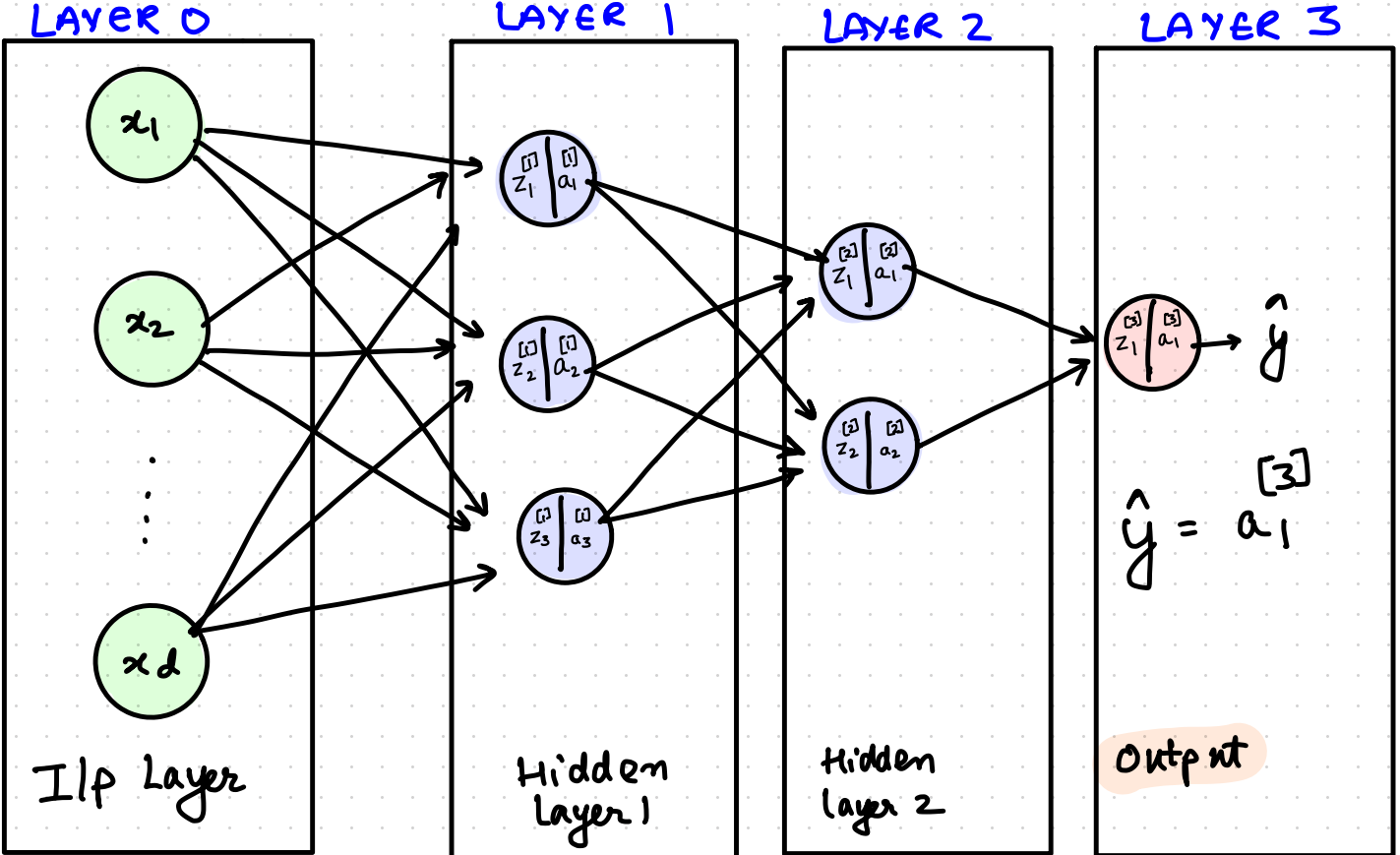


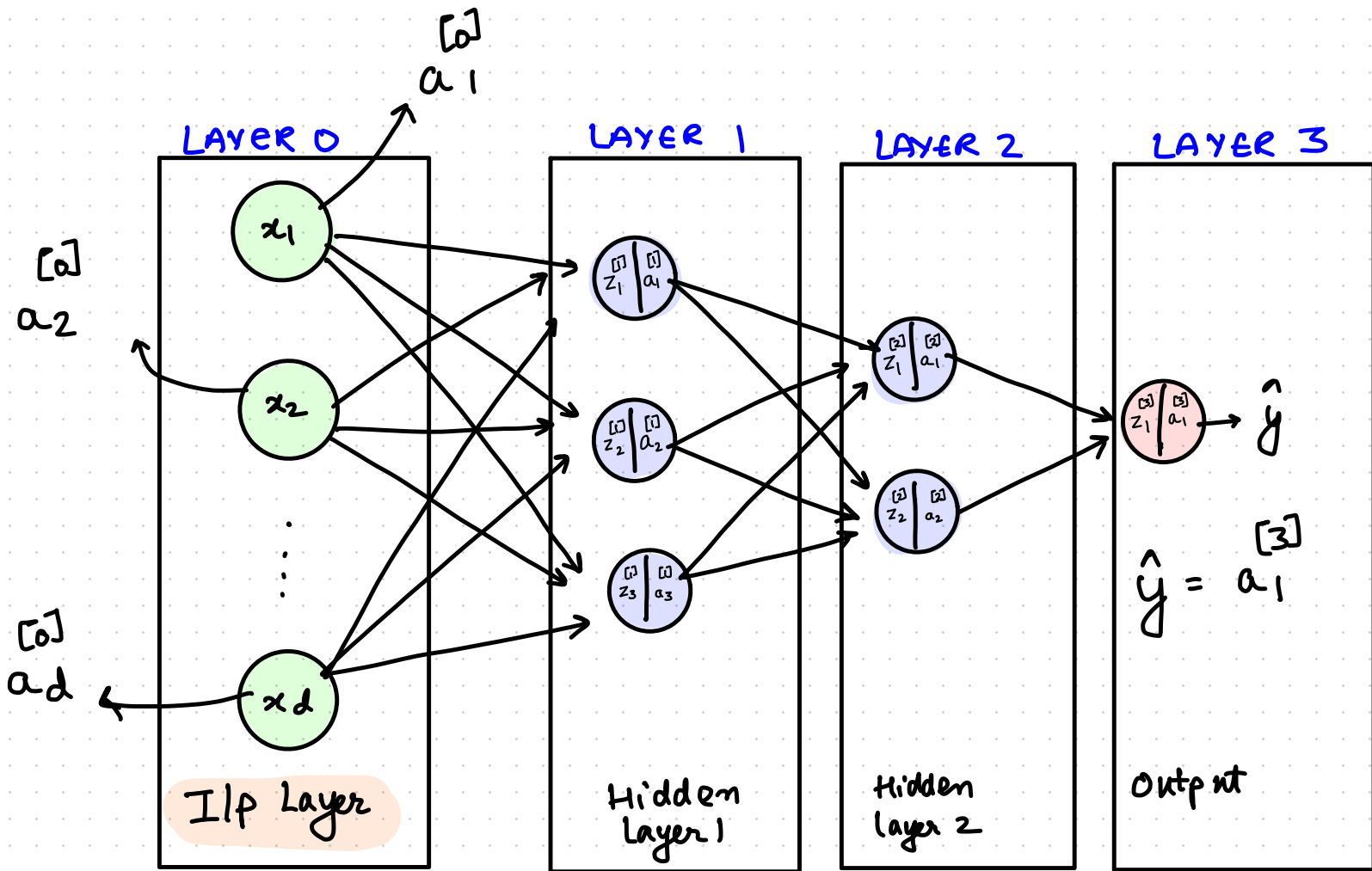


# MULTI-LAYER PERCEPTRON

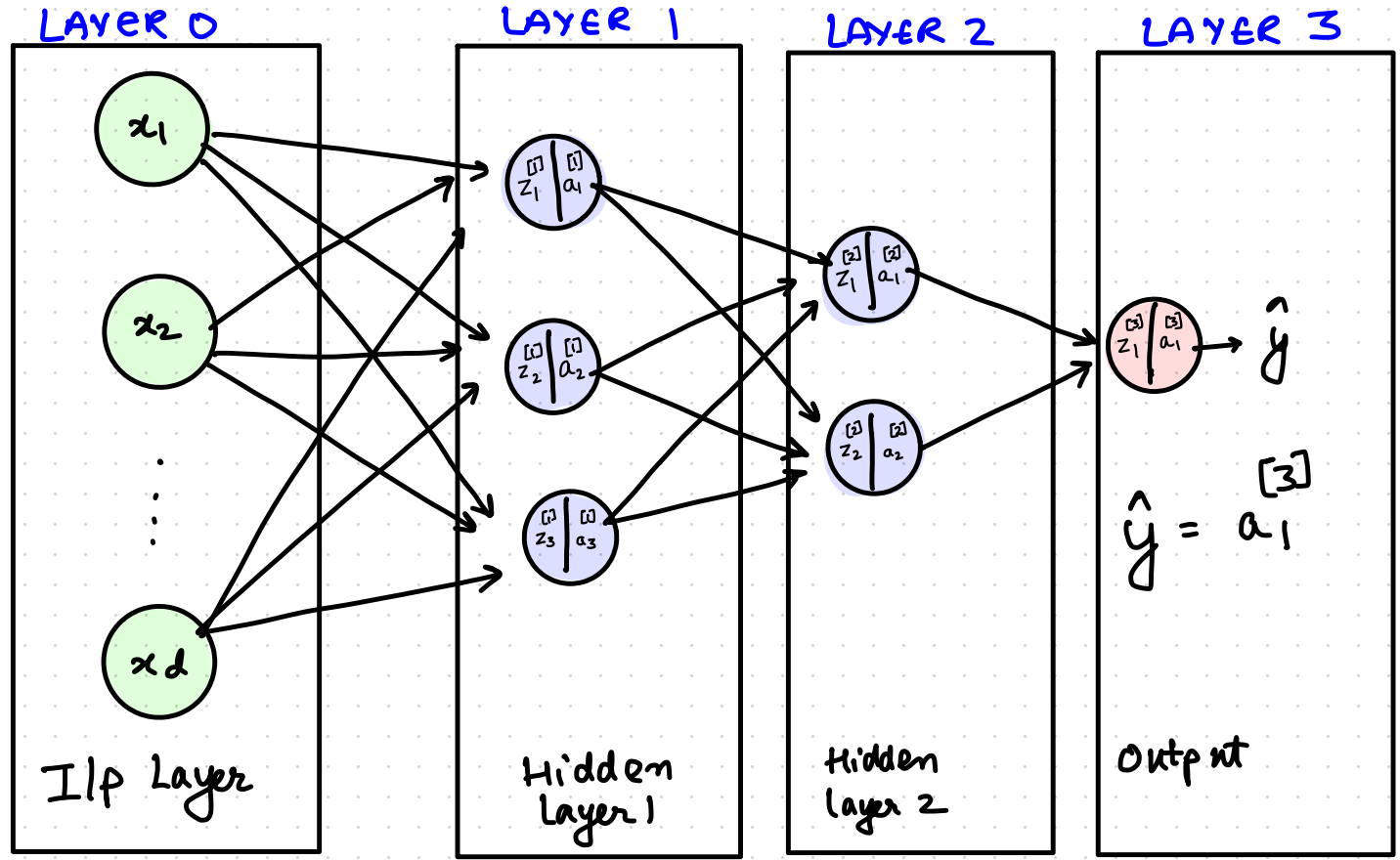




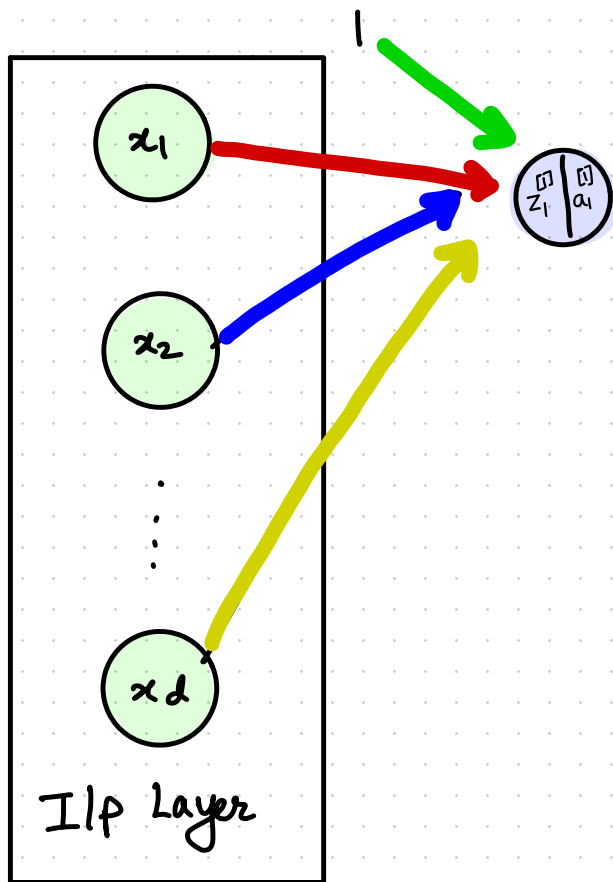




$$\begin{aligned}
 \text{IIP} &= \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \\
 &= \begin{bmatrix} [0] \\ a_1 \\ \vdots \\ [0] \\ a_d \end{bmatrix} \\
 &= a_{d \times 1} \begin{bmatrix} [0] \\ \vdots \\ [0] \end{bmatrix}
 \end{aligned}$$



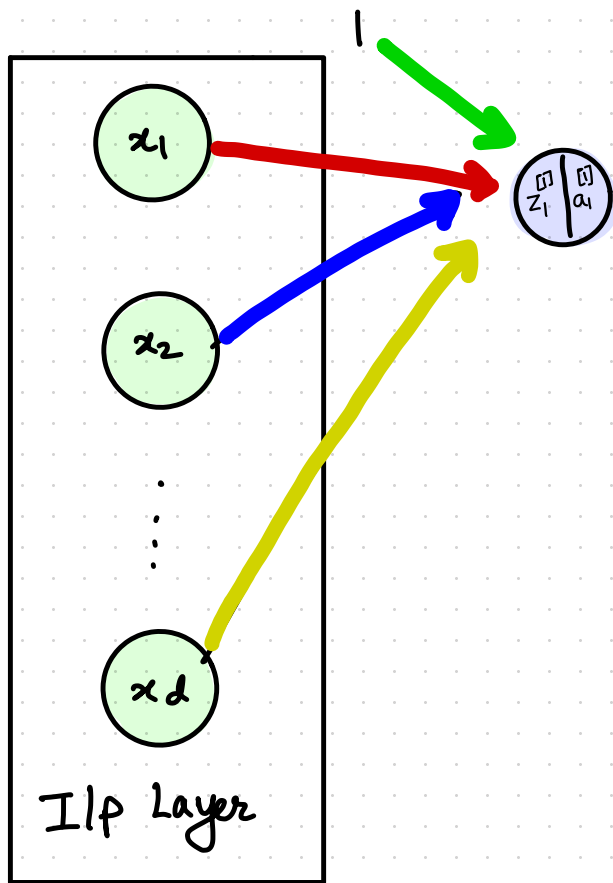
# CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



$$z_1^{[1]} = 1 * b_1^{[1]} + x_1 * w_{1,1}^{[1]} + x_2 * w_{1,2}^{[1]} + \dots + x_d * w_{1,d}^{[1]}$$

← bias layer Node 1

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)

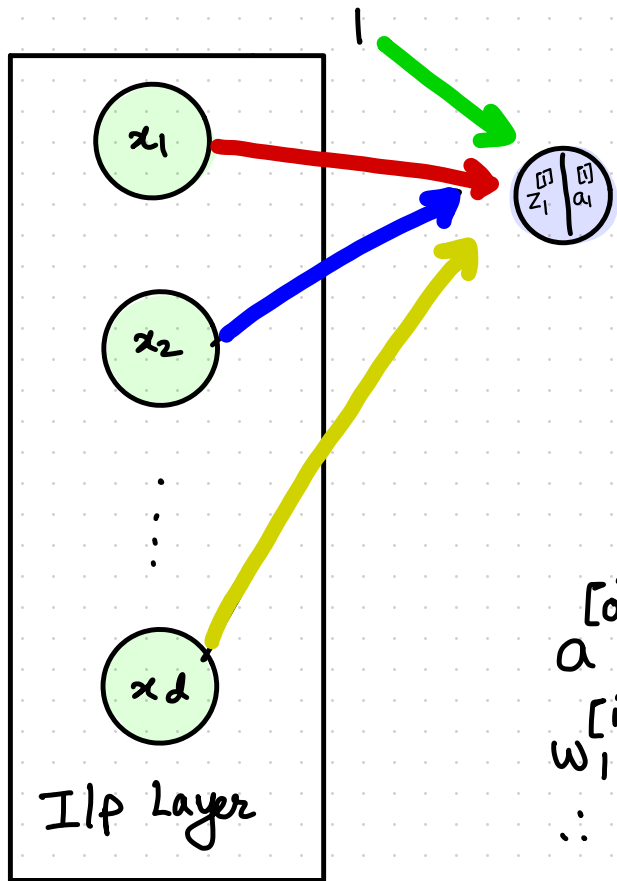


$$z_1^{[1]} = 1 * b_1^{[1]} + x_1 * w_{1,1}^{[1]} + x_2 * w_{1,2}^{[1]} + \dots + x_d * w_{1,d}^{[1]}$$

← bias layer 1 Node 1

$[l]$  ←  $l^{th}$  layer  
 $w_{a,b}$   
 $a^{th}$  node in  $l^{th}$  layer →  $b^{th}$  component of prev. layer activation

# CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



$$z_1^{[1]} = 1 * b_1 + a_1 * w_{1,1}^{[1]} + a_2 * w_{1,2}^{[1]} + \dots + a_d * w_{1,d}^{[1]}$$

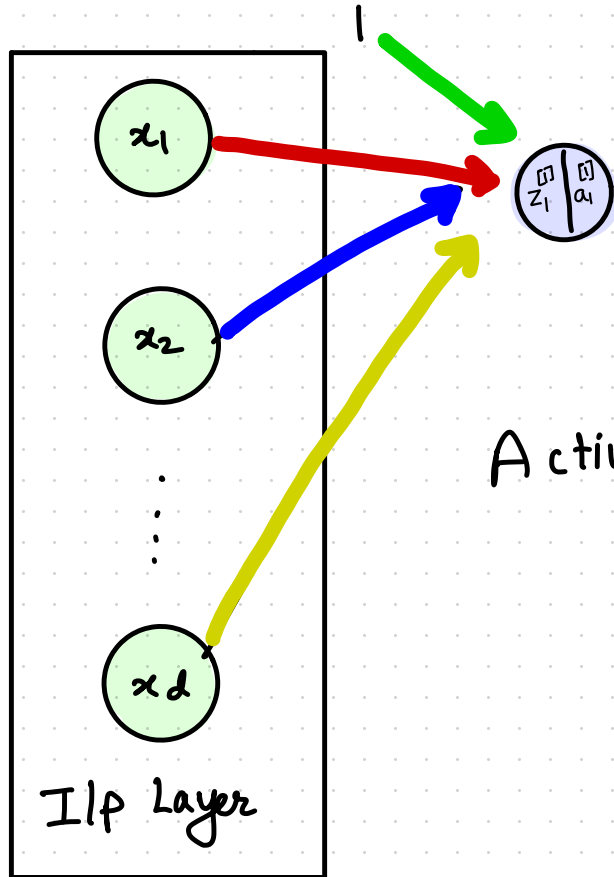
$$a^{[0]} \in \mathbb{R}^D$$

$$w_1^{[1]} \in \mathbb{R}^D$$

$$\therefore z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]}$$



# CONSIDER SINGLE NEURON (LAYER 1, NODE 1)

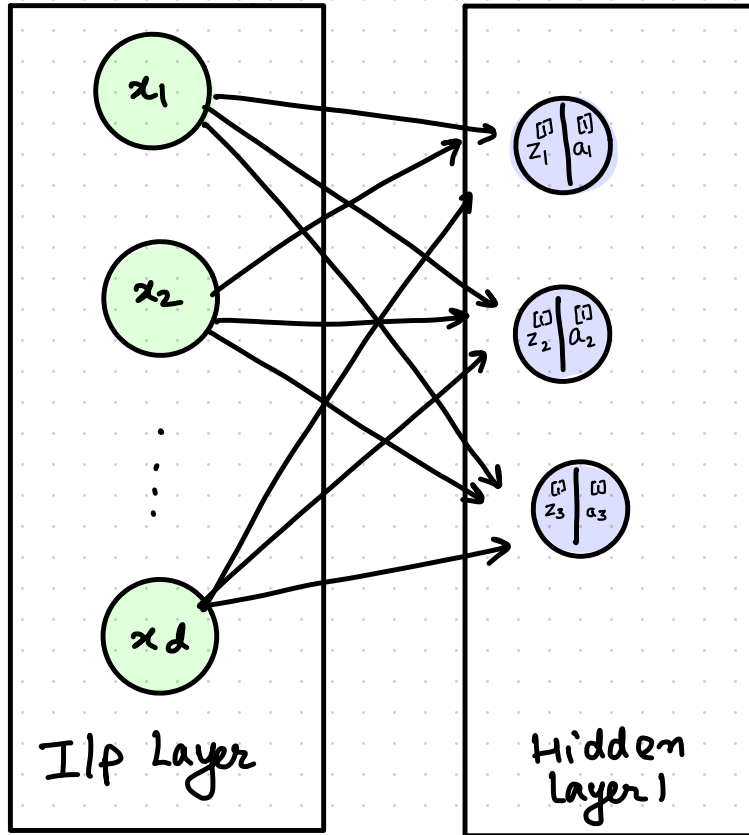


$$z_1^{[1]} = w_1^{[0]T} a^{[0]} + b_1^{[1]}$$

$$\text{Activat}^n = a_1^{[1]} = g(z_1^{[1]})$$

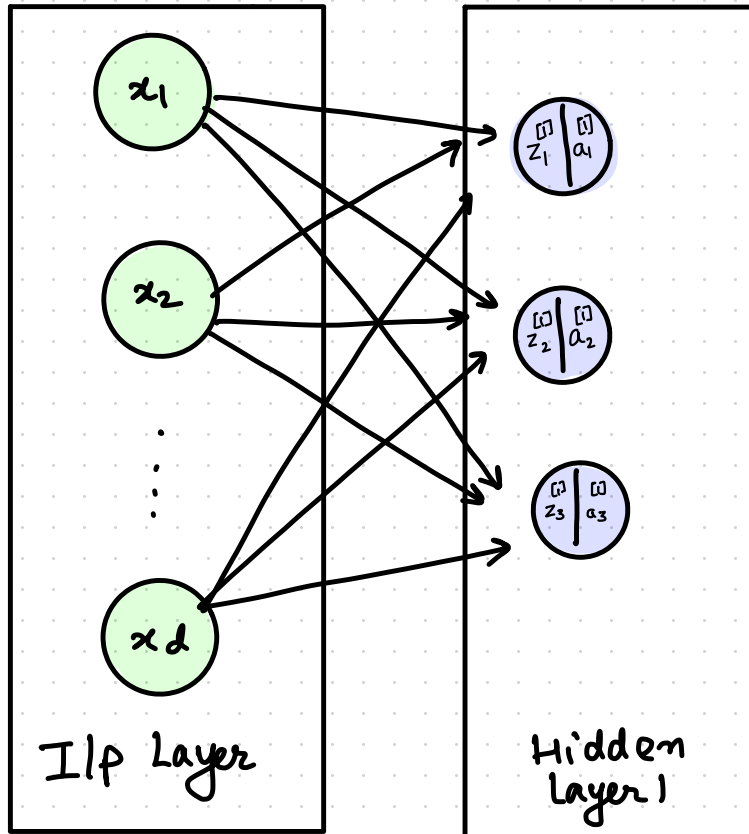
$$a_1^{[1]} \in \mathbb{R}$$

# FORWARD PROPAGATION



$$a_i^{[1]} = g(w_i^{[1]T} a^{[0]} + b_i^{[1]})$$

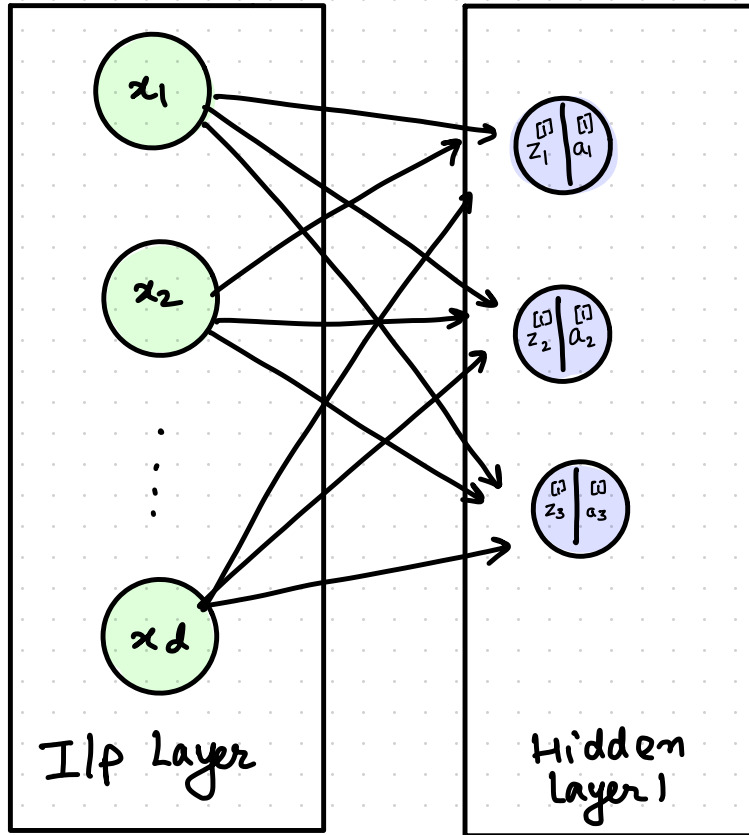
# FORWARD PROPAGATION



$$a_1^{[1]} = g(w_1^{[1]T} a^{[0]} + b_1^{[1]})$$

$$a_2^{[1]} = g(w_2^{[1]T} a^{[0]} + b_2^{[1]})$$

# FORWARD PROPAGATION

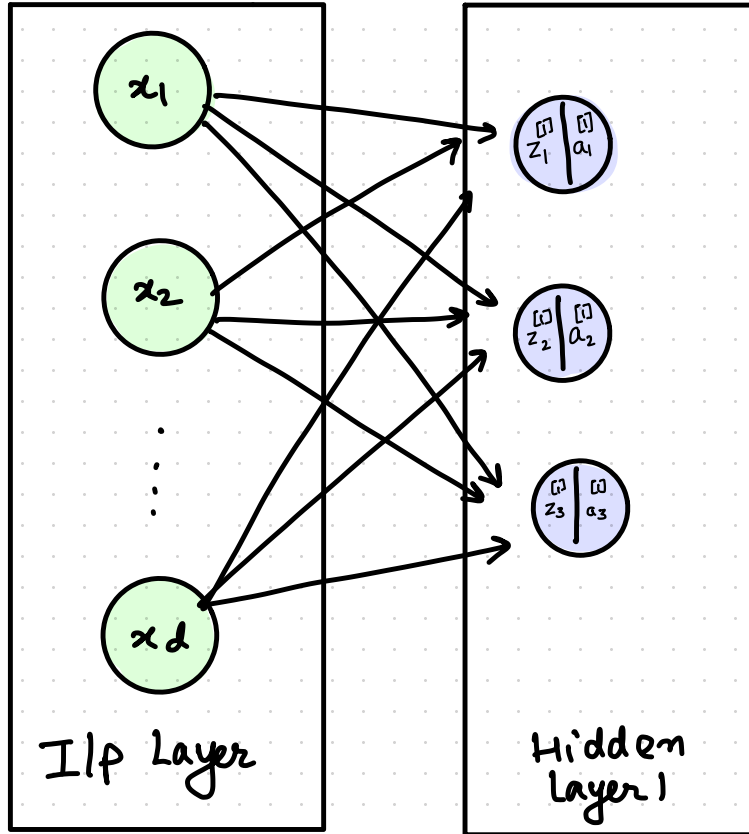


$$a_1^{[1]} = g(w_1^{[1]T} a^{[0]} + b_1^{[1]})$$

$$a_2^{[1]} = g(w_2^{[1]T} a^{[0]} + b_2^{[1]})$$

$$a_3^{[1]} = g(w_3^{[1]T} a^{[0]} + b_3^{[1]})$$

# FORWARD PROPAGATION (VECTORISATION)

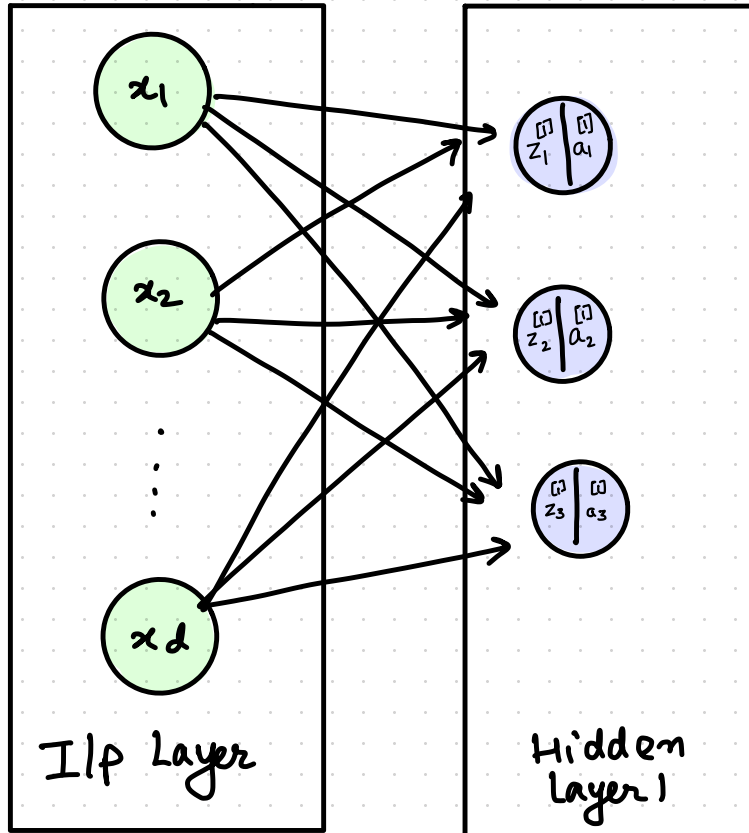


$$z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]}$$

$$z_2^{[1]} = w_2^{[1]T} a^{[0]} + b_2^{[1]}$$

$$z_3^{[1]} = w_3^{[1]T} a^{[0]} + b_3^{[1]}$$

# FORWARD PROPAGATION (VECTORISATION)



$$z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]}$$

$1 \times 1 = 1 \times 3 \quad 3 \times 1 \quad 1 \times 1$

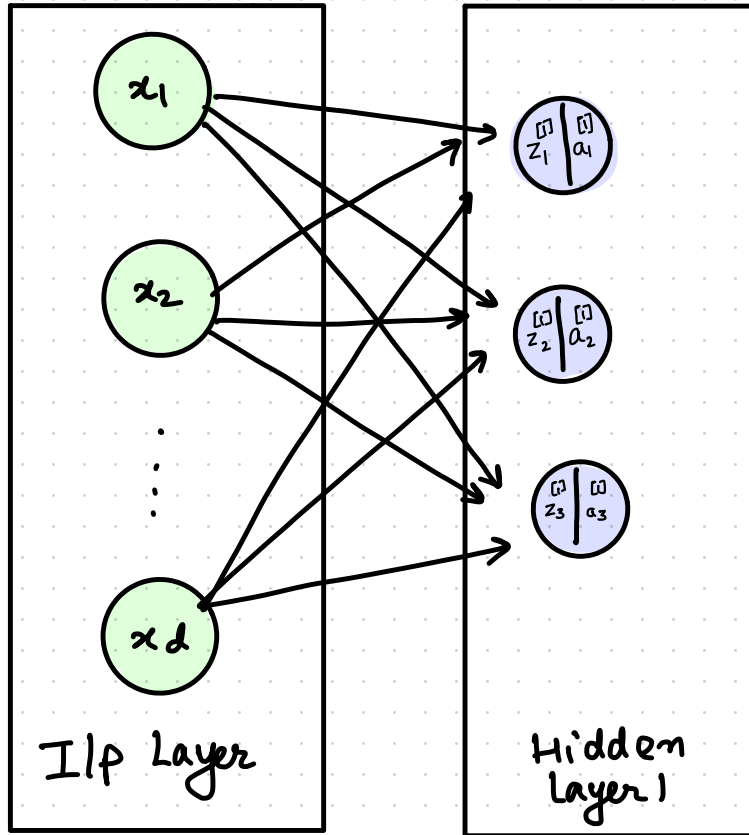
$$z_2^{[1]} = w_2^{[1]T} a^{[0]} + b_2^{[1]}$$

$1 \times 1 = 1 \times 3 \quad 3 \times 1 \quad 1 \times 1$

$$z_3^{[1]} = w_3^{[1]T} a^{[0]} + b_3^{[1]}$$

$1 \times 1 = 1 \times 3 \quad 3 \times 1 \quad 1 \times 1$

# FORWARD PROPAGATION (VECTORISATION)



$$z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]}$$

$1 \times 1$        $1 \times 3$        $3 \times 1$        $1 \times 1$

$$z_2^{[1]} = w_2^{[1]T} a^{[0]} + b_2^{[1]}$$

$1 \times 1$        $1 \times 3$        $3 \times 1$        $1 \times 1$

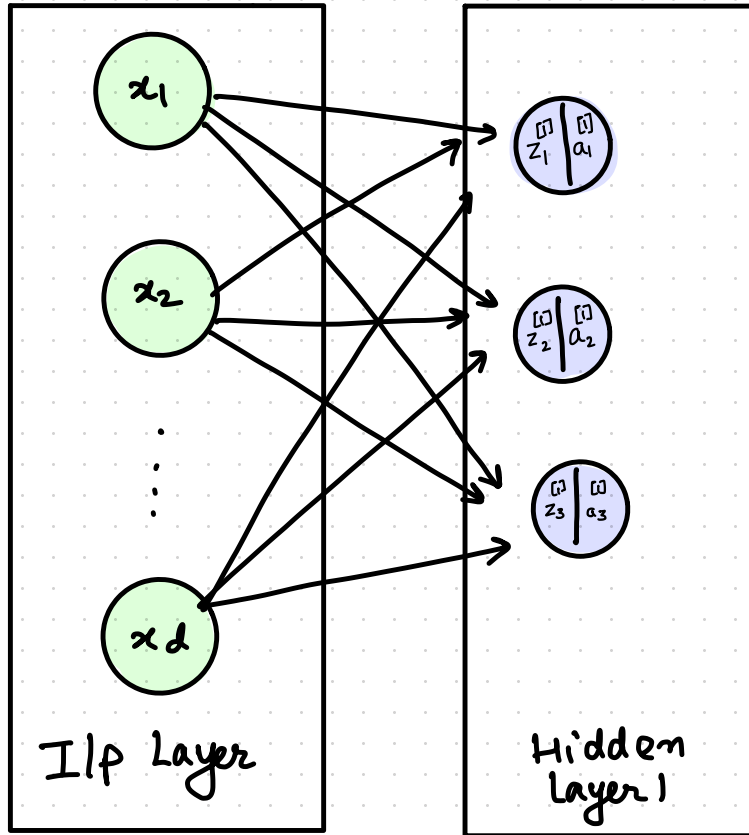
$$z_3^{[1]} = w_3^{[1]T} a^{[0]} + b_3^{[1]}$$

$1 \times 1$        $1 \times 3$        $3 \times 1$        $1 \times 1$

$$z^{[1]} = \begin{bmatrix} - & w_1^{[1]T} \\ - & w_2^{[1]T} \\ - & w_3^{[1]T} \end{bmatrix} a^{[0]} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

$3 \times 1$        $3 \times 3$        $3 \times 1$

# FORWARD PROPAGATION (VECTORISATION)



$$z^{[1]}_{3 \times 1} = \begin{bmatrix} - & w_{11}^{[1]T} \\ - & w_{21}^{[1]T} \\ - & w_{d1}^{[1]T} \end{bmatrix} a^{[0]}_{3 \times 1} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

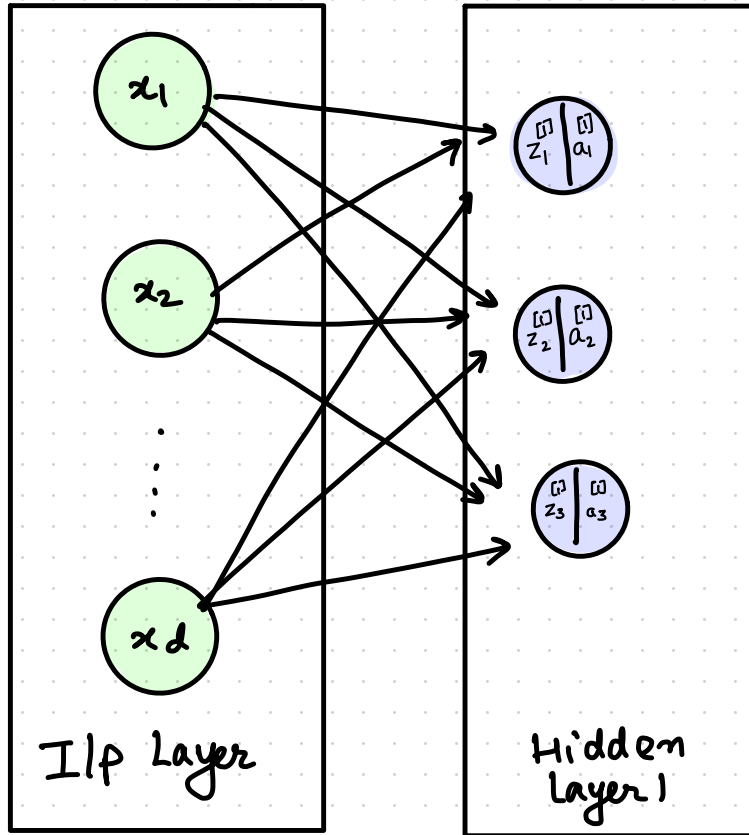
$3 \times 3$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$

↑ capitals for matrices



# FORWARD PROPAGATION (VECTORISATION)



$$z^{[1]}_{3 \times 1} = \begin{bmatrix} - & w_{11}^{[1]T} \\ - & w_{21}^{[1]T} \\ - & w_{31}^{[1]T} \end{bmatrix} a^{[0]}_{3 \times 1} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

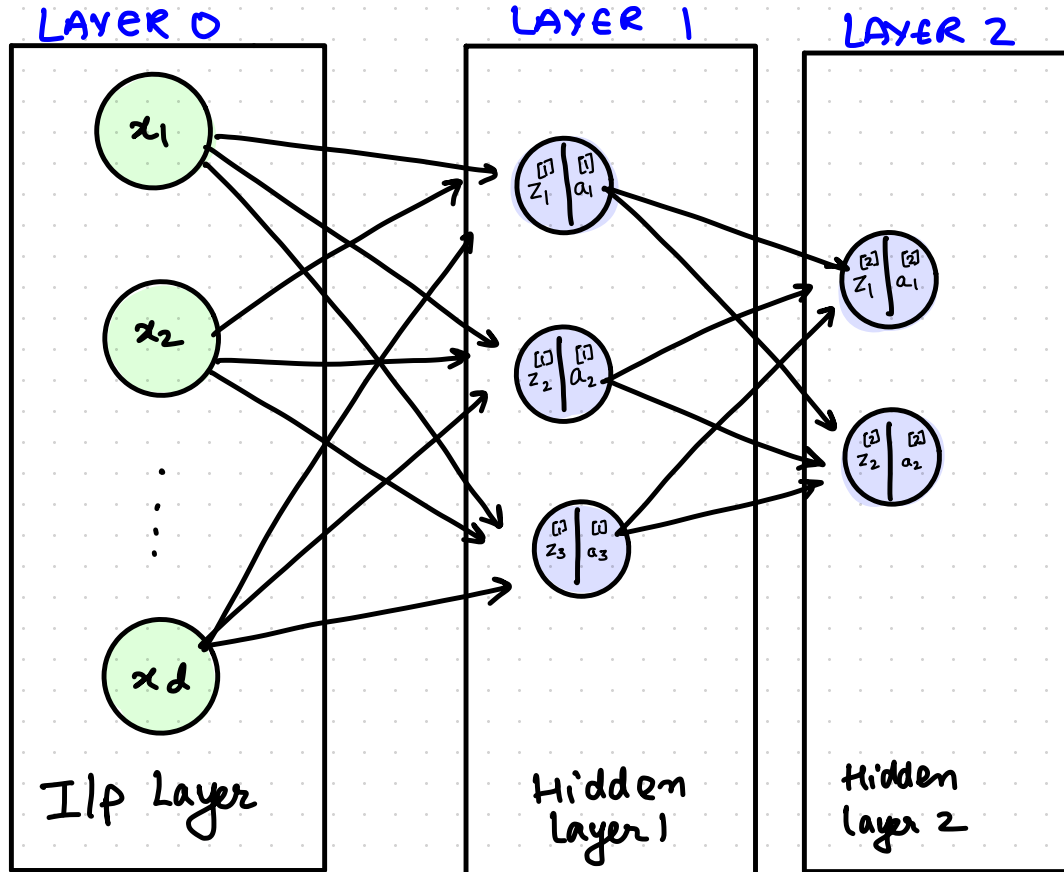
$3 \times 3$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$

↑ capitals for matrices

$$a^{[1]} = g(z^{[1]})$$

# FORWARD PROPAGATION

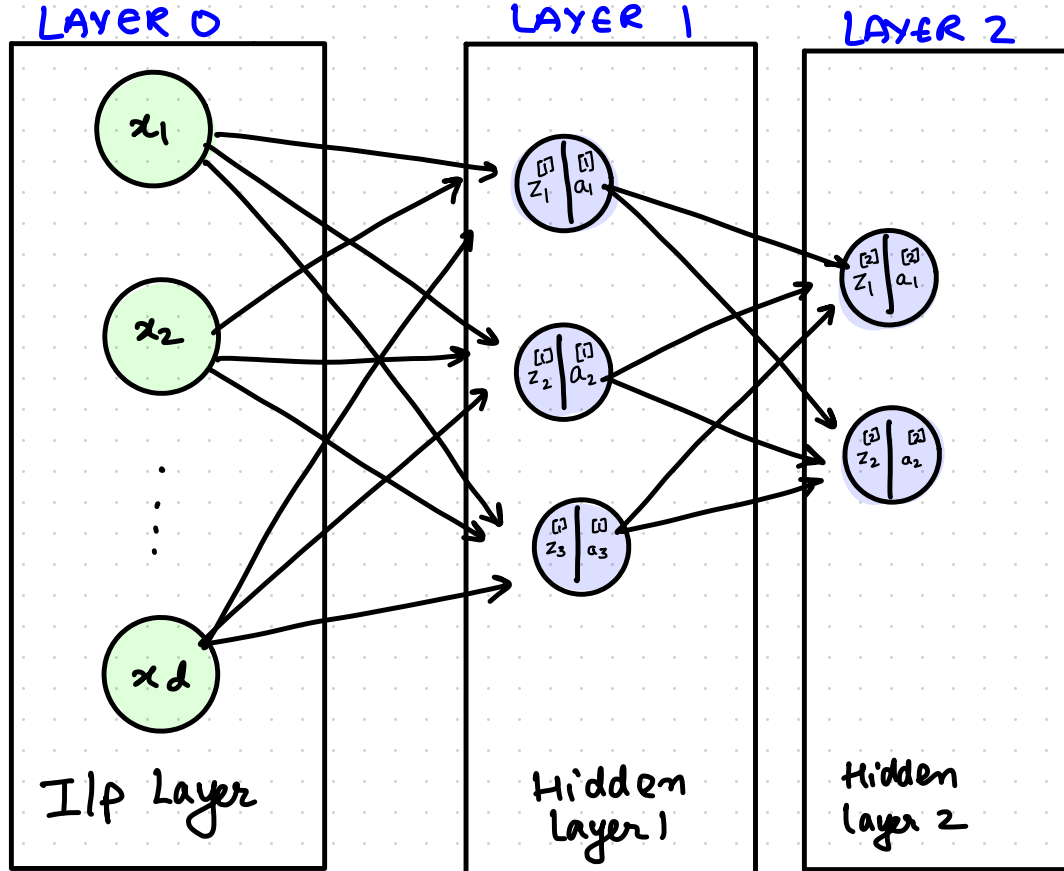


$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

Q. Dim. of  $W^{[2]}$ ?

# FORWARD PROPAGATION



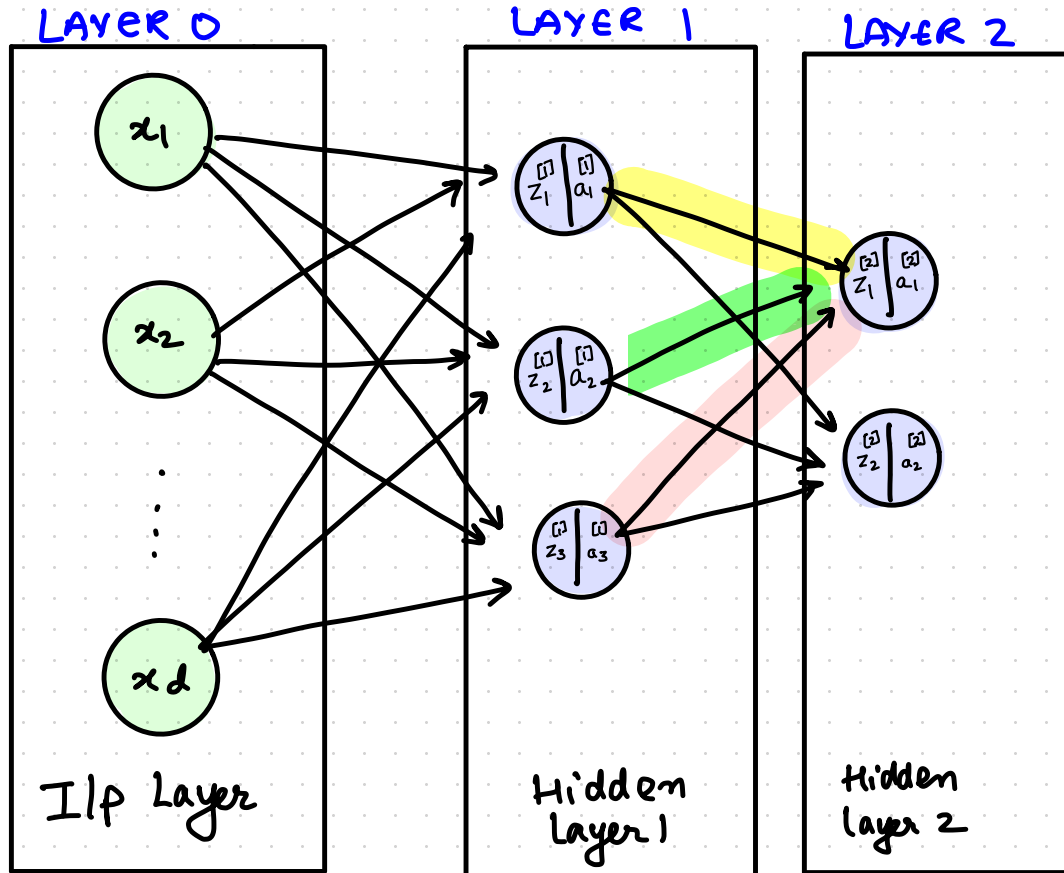
$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

Q. Dim. of  $W^{[2]}$ ?

$$W^{[2]} = \begin{bmatrix} - & W_1^{[2]} & - \\ - & W_2^{[2]} & - \end{bmatrix}$$

# FORWARD PROPAGATION



$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

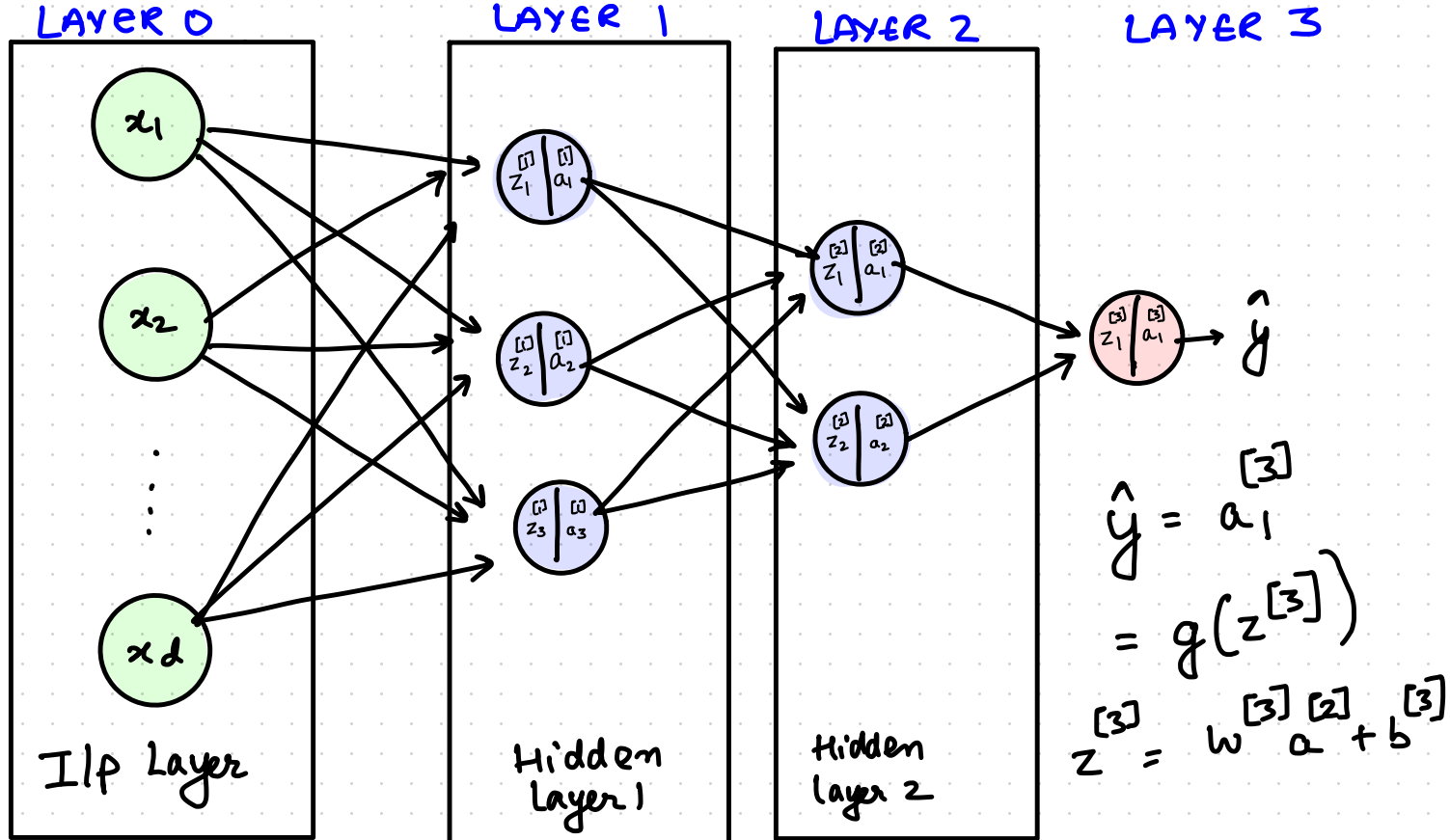
Q. Dim. of  $W^{[2]}$ ?

$$W^{[2]} = \begin{bmatrix} - & W_1^{[2]T} & - \\ - & W_2^{[2]T} & - \end{bmatrix}$$

$$W_1 \in \mathbb{R}^3$$

$$\therefore W^{[2]} \in \mathbb{R}^{2 \times 3}$$

# FORWARD PROPAGATION



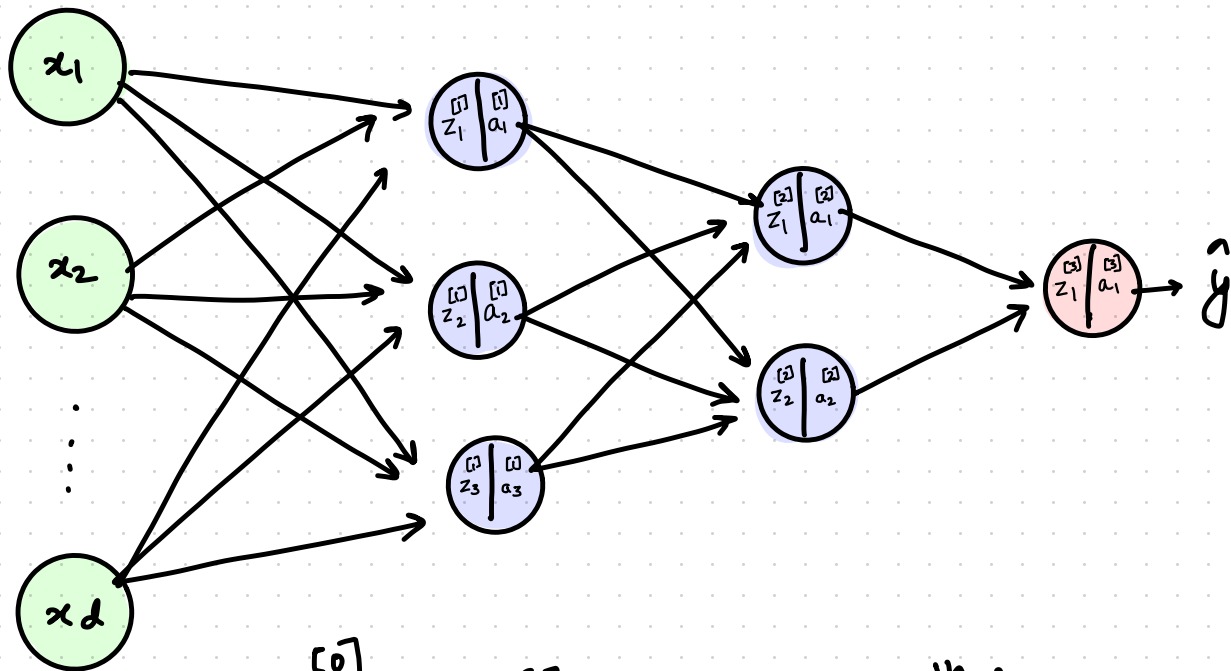
WHAT CAN WE SAY ABOUT SHAPES OF  $a, b, w$

LAYER 0

LAYER 1

LAYER 2

LAYER 3



$a^{[0]} \in \mathbb{R}^d$  or  $\mathbb{R}^{N^{[0]}}$   
 $d \rightarrow \# \text{ i/p features}$

$N^{[0]} = \# \text{ units in } 0^{\text{th}} \text{ layer}$

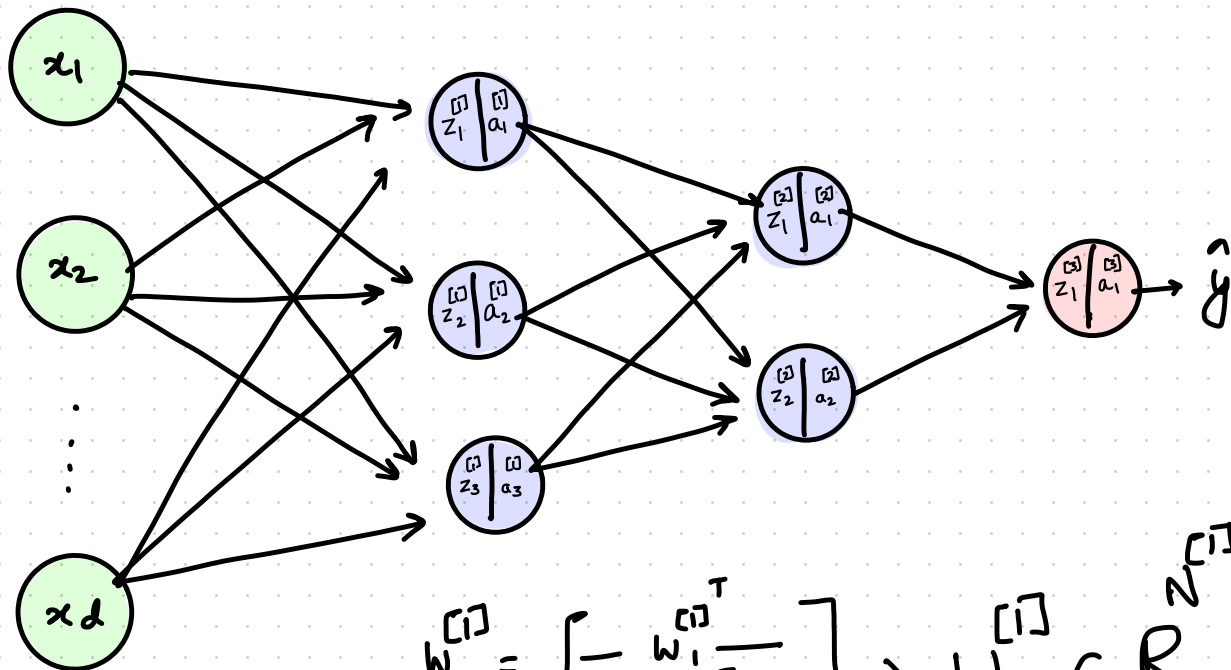
WHAT CAN WE SAY ABOUT SHAPES OF  $a, b, w$

LAYER 0

LAYER 1

LAYER 2

LAYER 3



$a^{[0]} \in \mathbb{R}^d$   
 $d \rightarrow \# \text{ of features}$

$$W^{[1]} = \begin{bmatrix} - & w_{11}^{[1]T} & - \\ - & w_{21}^{[1]T} & - \\ - & \vdots & - \\ - & w_{n1}^{[1]T} & - \end{bmatrix} \Rightarrow W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

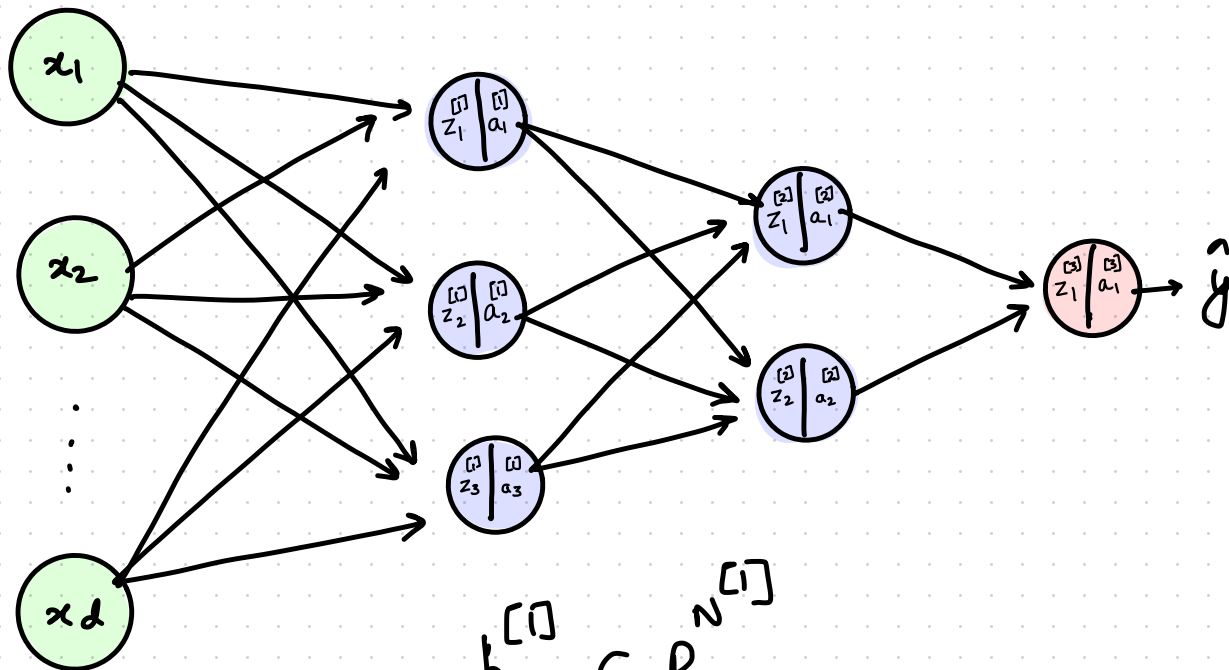
WHAT CAN WE SAY ABOUT SHAPES OF  $a$ ,  $b$ ,  $w$

LAYER 0

LAYER 1

LAYER 2

LAYER 3



$a \in \mathbb{R}^d$   
 $d \rightarrow \# \text{ of features}$

$b \in \mathbb{R}^{N \times 1}$



# SUMMARY OF SHAPES

$N^{[l]}$  : # NODES IN  $l^{\text{th}}$  layer

$a^{[0]}$   $\in \mathbb{R}^{N^{[0]}}$

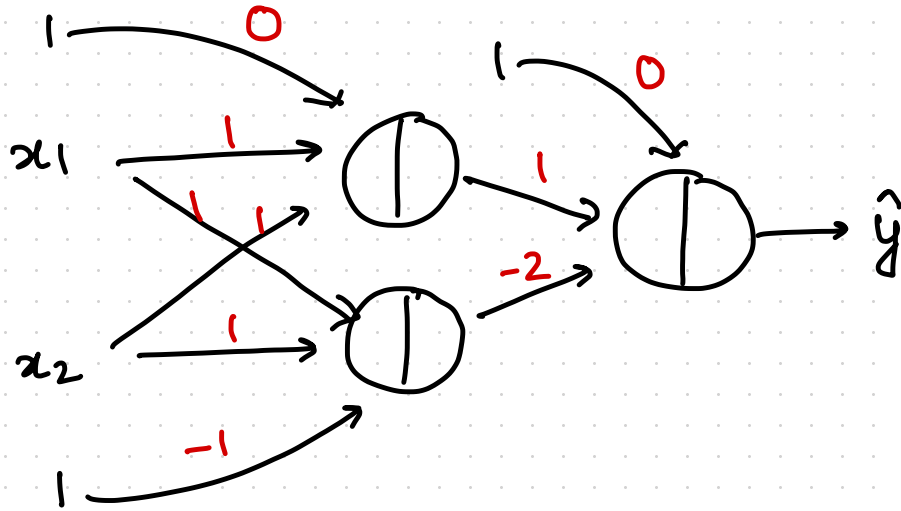
$w^{[l]}$   $\in \mathbb{R}^{N^{[l]} \times N^{[l-1]}}$

$b^{[l]}$   $\in \mathbb{R}^{N^{[l]}}$

$z^{[l]}$   $\in \mathbb{R}^{N^{[l]}}$

$a^{[l]}$   $\in \mathbb{R}^{N^{[l]}}$

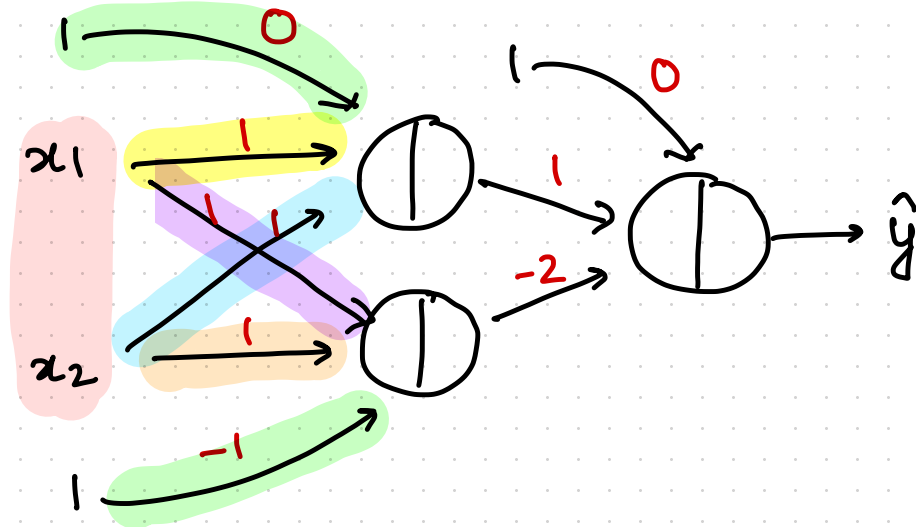
XOR USING "MLP" RELU



CONFIRM IF ABOVE N/W IS CORRECT FOR XOR.

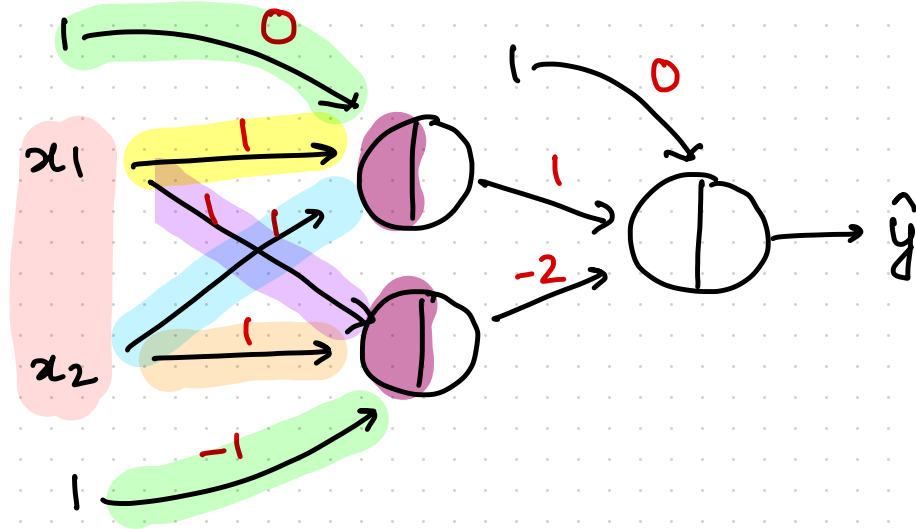
Start with  $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$   $y_{\text{True}} = 0$

# XOR USING "MLP" RELU



$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \quad b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}; \quad W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

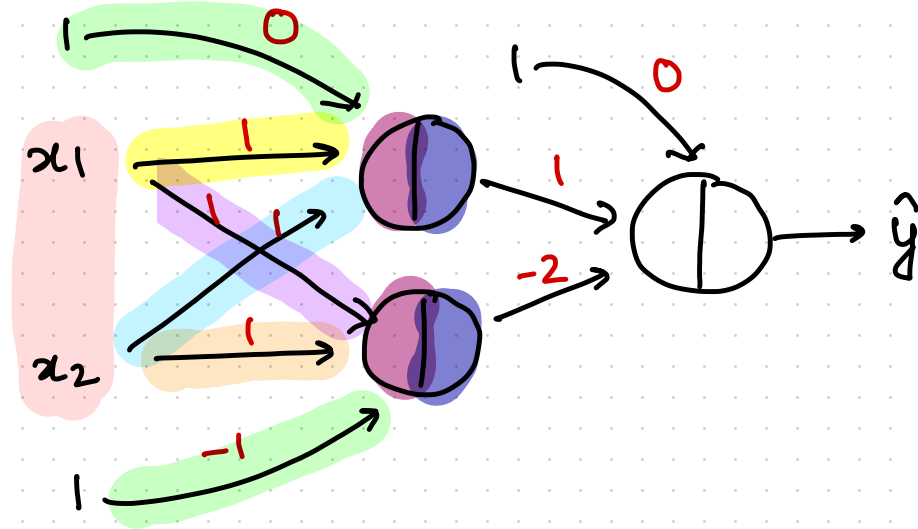
# XOR USING "MLP" RELU



$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \quad b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}; \quad W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

# XOR USING "MLP" RELU

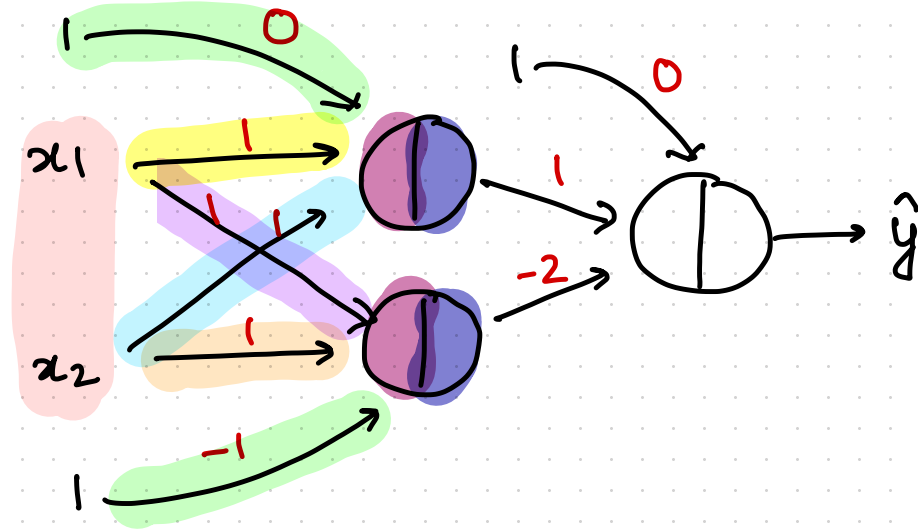


$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \quad b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}; \quad W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$a^{[1]} = \text{RELU} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

# XOR USING "MLP" RELU

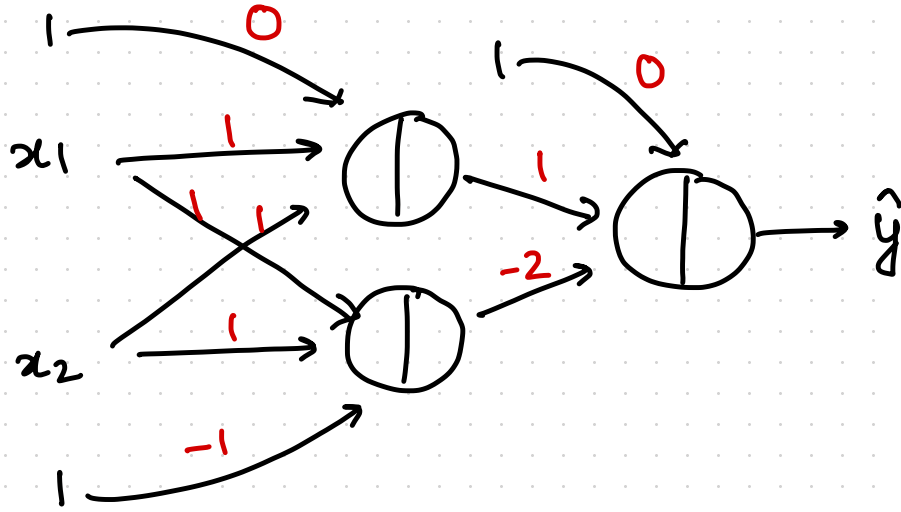


$$a^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} ; W^{[2]} = [1 \ -2] ; b^{[2]} = [0]$$

$$z^{[2]} = [1 \ -2] \begin{bmatrix} 0 \\ 0 \end{bmatrix} + [0] = 0 ; a^{[2]} = \hat{y} = \text{RELU}(0) = 0$$

✓

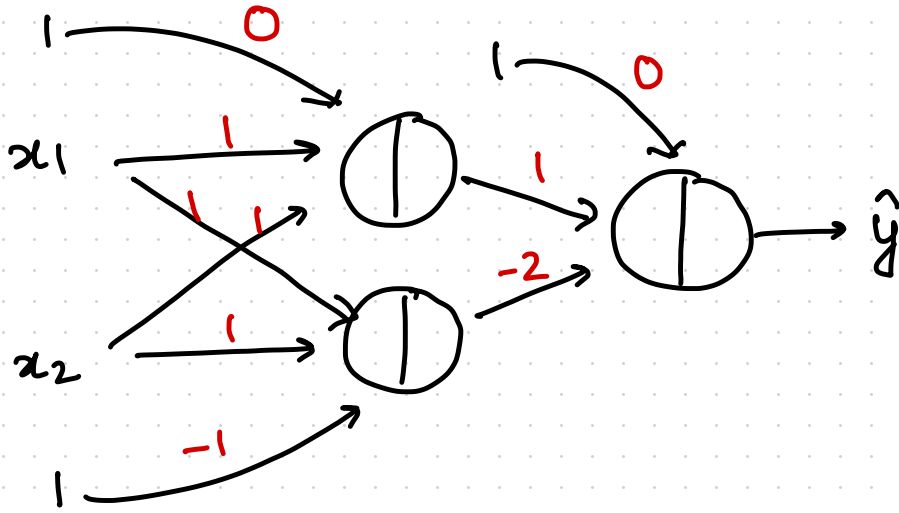
XOR USING "MLP" RELU



CONFIRM IF ABOVE N/W IS CORRECT FOR XOR.

Start with  $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$   $y_{\text{True}} = 1$

# XOR USING "MLP" RELU

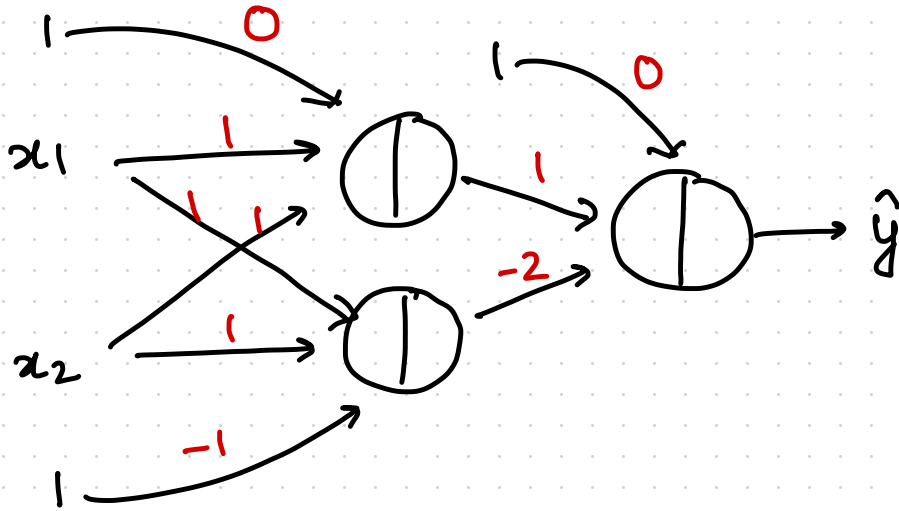


$$z^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$a^{[1]} = \text{RELU}(z^{[1]}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$



# XOR USING "MLP" RELU



$$z^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$a^{[1]} = \text{RELU}(z^{[1]}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$z^{[2]} = [1 \ -2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} + [0] = [1]$$

$$a^{[2]} = \text{RELU}(1) = 1$$

$$= \hat{y}$$

# COMPUTATION FOR N EXAMPLES

$x_{(i)} \in \mathbb{R}^d$  or  $\mathbb{R}^{N^{[0]}}$

$$X = \begin{bmatrix} - x_{(1)}^T - \\ - x_{(2)}^T - \\ - x_{(N)}^T - \end{bmatrix}$$

$$= \begin{bmatrix} - a_{(1)}^{[0]T} - \\ \vdots \\ - a_{(N)}^{[0]T} - \end{bmatrix}$$

$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$   
matrix

# COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in \mathbb{R}^d \text{ or } \mathbb{R}^{N^{[0]}}$$

$$X = \begin{bmatrix} - x_{(1)}^T - \\ - x_{(2)}^T - \\ - x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} - a_{(1)}^{[0]T} - \\ \vdots \\ - a_{(N)}^{[0]T} - \end{bmatrix} = A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$Z^{[l]} \leftarrow \text{Layer}$

$Z^{(i)} \rightarrow \text{Instance / Sample}$

$$= W^{[l]} A^{[0]T} + b^{[l]} \in \mathbb{R}^{N^{[l]}}$$

Independent of 'i'

$$a_{(i)}^{[0]} \in \mathbb{R}^d \equiv \begin{bmatrix} \vdots \end{bmatrix}$$

# COMPUTATION FOR N EXAMPLES

$$x_{(i)} \in \mathbb{R}^D$$

$$X = \begin{bmatrix} - x_{(1)}^T - \\ - x_{(2)}^T - \\ - x_{(N)}^T - \end{bmatrix} = \begin{bmatrix} - a_{(1)}^{[0]T} - \\ \vdots \\ - a_{(N)}^{[0]T} - \end{bmatrix}$$

$$\begin{aligned} z_{(1)}^{[1]} &= \left\{ W^{[1][0]} a_{(1)}^{[0]} + b^{[1]} \right\} \in \mathbb{R}^{N^{[1]}} \\ z_{(2)}^{[1]} &= \left\{ W^{[1][0]} a_{(2)}^{[0]} + b^{[1]} \right\} \in \mathbb{R}^{N^{[1]}} \\ &\vdots \\ z_{(N)}^{[1]} &= \left\{ W^{[1][0]} a_{(N)}^{[0]} + b^{[1]} \right\} \in \mathbb{R}^{N^{[1]}} \end{aligned}$$

# COMPUTATION FOR N EXAMPLES

$$x^{(i)} \in \mathbb{R}^D$$

$$X = \begin{bmatrix} - x^{(1)T} - \\ - x^{(2)T} - \\ - x^{(N)T} - \end{bmatrix} = \begin{bmatrix} - a^{(1)T} - \\ \vdots \\ - a^{(N)T} - \end{bmatrix} = A \in \mathbb{R}^{N \times D}$$

$$\begin{aligned} z^{(1)} &= W a^{(1)} + b^{(1)} \in \mathbb{R}^{N^{(1)}} \\ z^{(2)} &= W a^{(2)} + b^{(1)} \in \mathbb{R}^{N^{(1)}} \\ &\vdots \\ z^{(N)} &= W a^{(N)} + b^{(1)} \in \mathbb{R}^{N^{(1)}} \end{aligned}$$

# COMPUTATION FOR N EXAMPLES

$$x^{(i)} \in \mathbb{R}^D$$

$$X = \begin{bmatrix} - x^{(1)T} - \\ - x^{(2)T} - \\ - x^{(N)T} - \end{bmatrix} = \begin{bmatrix} - a^{(1)T} - \\ \vdots \\ - a^{(N)T} - \end{bmatrix} = A \in \mathbb{R}^{N \times N}^{[0]}$$

$$z^{(1)} = W a^{(1)} + b^{(1)}$$

$$z^{(2)} = W a^{(2)} + b^{(1)}$$

$$\vdots \\ z^{(N)} = W a^{(N)} + b^{(1)}$$

$$\Rightarrow Z = \begin{bmatrix} - z^{(1)T} - \\ - z^{(2)T} - \\ \vdots \\ - z^{(N)T} - \end{bmatrix} \in \mathbb{R}^{N \times N}^{[1]}$$

# COMPUTATION FOR N EXAMPLES

$$x^{(i)} \in \mathbb{R}^D$$

$$X = \begin{bmatrix} - x^{(1)T} - \\ - x^{(2)T} - \\ - x^{(N)T} - \end{bmatrix} = \begin{bmatrix} - a^{[0]T} - \\ \vdots \\ - a^{[0]T} - \end{bmatrix} = A \in \mathbb{R}^{N \times N^{[0]}}$$

$$\begin{aligned} z^{[1]}_{(1)} &= W^{[1][0]} a^{[0]}_{(1)} + b^{[1]} \\ z^{[1]}_{(2)} &= W^{[1][0]} a^{[0]}_{(2)} + b^{[1]} \\ &\vdots \\ z^{[1]}_{(N)} &= W^{[1][0]} a^{[0]}_{(N)} + b^{[1]} \end{aligned}$$

$$\Rightarrow Z = \begin{bmatrix} - z^{[1]}_{(1)} - \\ - z^{[1]}_{(2)} - \\ \vdots \\ - z^{[1]}_{(N)} - \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

# COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} \text{---} & b^{[1]T} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & b^{[1]T} & \text{---} \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} \in \mathbb{R}^{N \times N^{[1]}}$$

*all same entries*



# COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} \text{---} & b^{[1]T} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & b^{[1]T} & \text{---} \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} = A^{[0]} W^{[1]T} + B^{[1]}$$

# COMPUTATION FOR N EXAMPLES

$$A^{[0]} \in \mathbb{R}^{N \times N^{[0]}}$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}}$$

$$B^{[1]} = \begin{bmatrix} \text{---} & b^{[1]T} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & b^{[1]T} & \text{---} \end{bmatrix} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} \in \mathbb{R}^{N \times N^{[1]}}$$

$$Z^{[1]} = A^{[0]} W^{[1]T} + B^{[1]}$$

$$\Rightarrow A^{[1]} = g(Z^{[1]})$$

$$\begin{aligned} Z^{[l]} &= A^{[l-1]} W^{[l]T} + B^{[l]} \\ A^{[l]} &= g(Z^{[l]}) \end{aligned}$$

# XOR ALL EXAMPLES

$$\textcircled{1} X = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}_{4 \times 2} ; \hat{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}_{4 \times 1}$$

# XOR ALL EXAMPLES

$$\textcircled{1} X = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}_{4 \times 2} ; \hat{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}_{4 \times 1}$$

$$W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}_{2 \times 2} ; b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \Rightarrow B^{[1]} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix}_{4 \times 2}$$

# XOR ALL EXAMPLES

$$\textcircled{1} X = A^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}_{4 \times 2}; \hat{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{4 \times 1}$$

$$W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}_{2 \times 2} = W^{[1]T}; b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \Rightarrow B^{[1]} = \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix}_{4 \times 2}$$

$$Z^{[1]} = A^{[0]} W^{[1]T} + B^{[1]} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & -1 \end{bmatrix}$$

# XOR ALL EXAMPLES

$$Z^{[1]} = A^{[0]} W^{[1]} + B^{[1]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

$$A^{[1]} = \text{RELU} \left( \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \right) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}_{4 \times 2}$$

$$W^{[2]} = [1 \ -2]$$

$$b^{[2]} = [0] \Rightarrow B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

# XOR ALL EXAMPLES

$$A^{[1]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}_{4 \times 2}$$

$$W^{[2]} = [1 \ -2]_{1 \times 2}$$

$$b^{[2]} = [0] \Rightarrow B^{[2]} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{4 \times 1}$$

$$Z^{[2]} = A^{[1]} W^{[2]T} + B^{[2]} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

# XOR ALL EXAMPLES

$$z^{[2]} = A^{[1]} w^{[2]T} + B^{[2]} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

$$A^{[2]} = \hat{y} = \text{RELU} \left( \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix} \right) = y_{G.T}$$



## # PARAMETERS

Parameters:  $w^{[l]}$ ;  $b^{[l]}$   $\forall l$

Q: # parameters for XOR example?

# # PARAMETERS

Parameters:  $W^{[l]}$ ;  $b^{[l]}$   $\forall l$

Q: # parameters for XOR example?

$$N^{[0]} = 2; N^{[1]} = 2; N^{[2]} = 1$$

$$W^{[1]} \in \mathbb{R}^{N^{[1]} \times N^{[0]}} = \begin{bmatrix} & \end{bmatrix}_{2 \times 2} = 4 \text{ params} = N^{[1]} * N^{[0]}$$

$$b^{[1]} \in \mathbb{R}^{N^{[1]}} = \begin{bmatrix} & \end{bmatrix}_{2 \times 1} = 2 \text{ params} = N^{[1]}$$

$$W^{[2]} \rightarrow 2 \text{ params} = N^{[2]} * N^{[1]} \quad \& \quad b^{[2]} = N^{[2]} \text{ params}$$

# # PARAMETERS

Parameters:  $w^{[l]}$ ;  $b^{[l]}$   $\forall l$

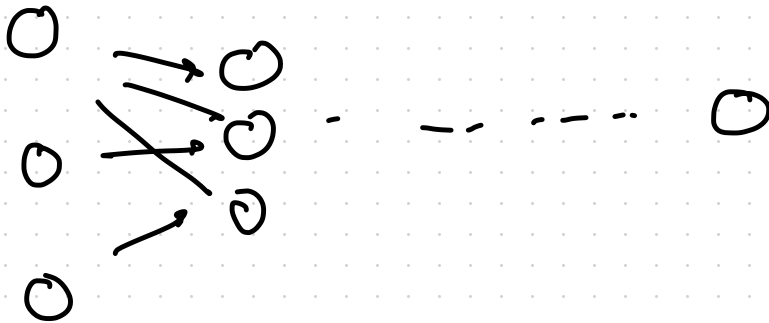
$$\text{PARAMS} = \sum_{l=1}^L N^{[l]} \cdot N^{[l-1]} + N^{[l]}$$

? XOR 4 examples, 9 parameters -!

⇒ NOTEBOOK: XOR demo

# LEARNING PARAMETERS

→ FORWARD PROPAGATION (PREDICT BASED ON CURRENT PARAMS)



← BACKWARD PROPAGATION (CHANGE WEIGHTS TO IMPROVE OBJECTIVE)

# LEARNING PARAMETERS

Assume  $IP: X \in \mathbb{R}^{N \times N^{[O]}}$

$OP: \hat{y}; G.T.: y$

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N L(\hat{y}^{(i)}, y)$$

$$\text{or } \sum_{i=1}^N L(\hat{y}^{(i)}, y)$$

Params:  $w^{[1]}, b^{[1]}, \dots$

# LEARNING PARAMETERS

Assume i/p :  $X \in \mathbb{R}^{N \times N^{[0]}}$

o/p :  $\hat{y}$  ; G.T. :  $y$

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N L(\hat{y}^{(i)}, y) = J(\theta)$$

Params :  $\theta = \{w^{[1]}, b^{[1]}, \dots\}$

## GRADIENT DESCENT

① INIT Params randomly

② Till convergence :

$$w^{[1]} = w^{[1]} - \alpha \frac{\partial J(\theta)}{\partial w^{[1]}}$$

!

→ How TO COMPUTE?

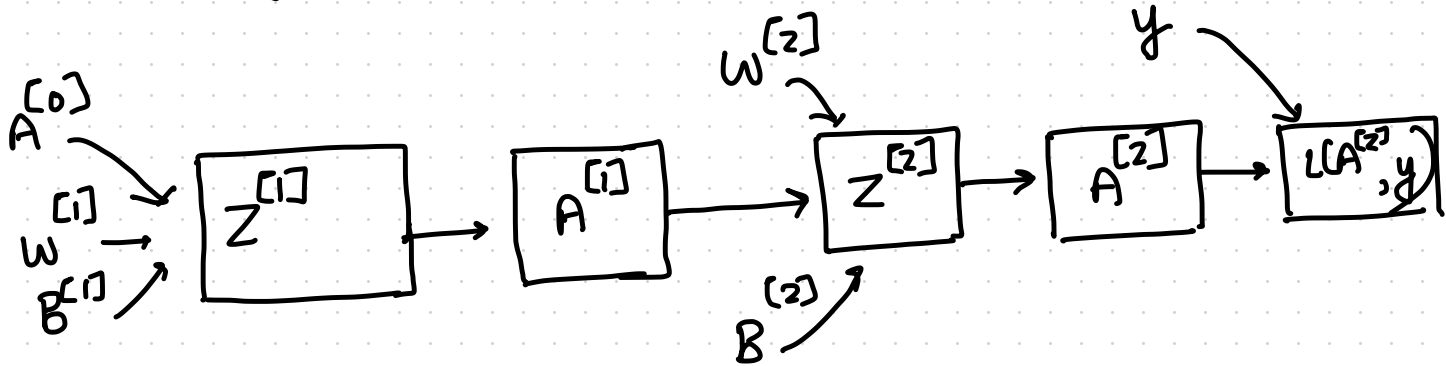
# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$\textcircled{1} z^{[1]} = A^{[0]} w^{[1]T} + B^{[1]}$$

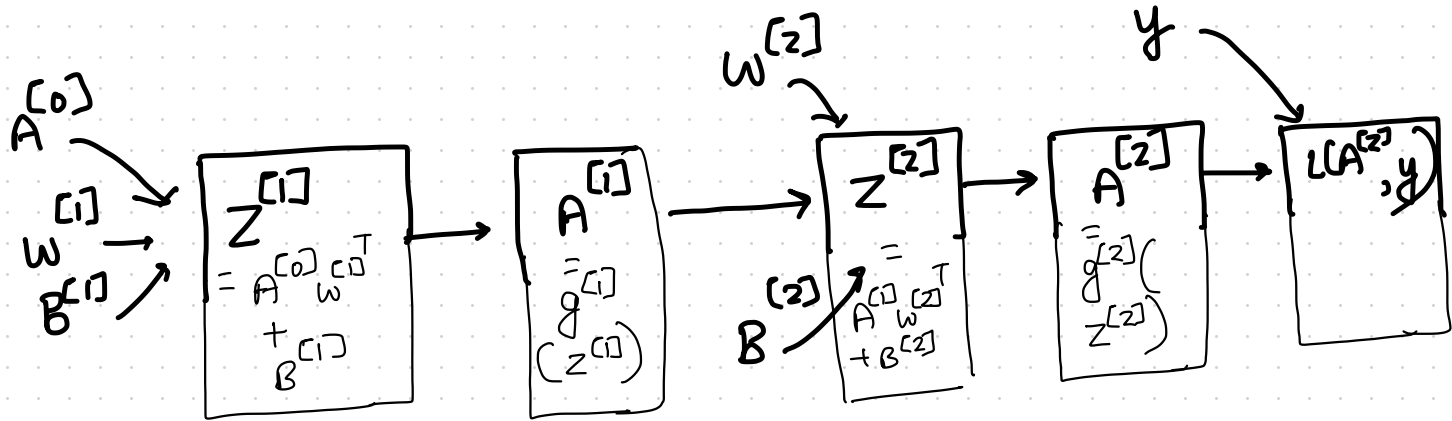
$$\textcircled{2} A^{[1]} = g^{[1]}(z^{[1]})$$

$$\textcircled{3} z^{[2]} = A^{[1]} w^{[2]T} + B^{[2]}$$

$$\textcircled{4} A^{[2]} = g^{[2]}(z^{[2]}) = y$$



# COMPUTATION GRAPH (FOR XOR EXAMPLE)



Let  $g^{[1]} = \text{SIGMOID}$  ; ASSUME CROSS ENTROPY LOSS  
 $g^{[2]} = \text{SIGMOID}$

WHAT IS  $\frac{\partial L(\theta)}{\partial W^{[1]}}$  ? ; WHAT IS  $\frac{\partial L(\theta)}{\partial A^{[2]}}$  ?



# DERIVATIVES OF ACTIVATION FUNCTIONS

RELU

$$g(z) = \begin{cases} z; & z > 0 \\ 0; & z < 0 \\ \text{undefined}; & \text{o/w} \end{cases}$$

↓ Assume  $z \neq 0$

$$g(z) = \begin{cases} z; & z > 0 \\ 0; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z > 0 \\ 0; & \text{o/w} \end{cases}$$

# DERIVATIVES OF ACTIVATION FUNCTIONS

RELU (LEAKY)

$$g(z) = \begin{cases} z; & z \geq 0 \\ \alpha z; & z < 0 \\ \text{undefined;} & \text{o/w} \end{cases}$$

↓

$$g(z) = \begin{cases} z; & z \geq 0 \\ \alpha z; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z \geq 0 \\ \alpha; & z < 0 \end{cases}$$

# DERIVATIVES OF ACTIVATION FUNCTIONS

## SIGMOID

$$g(z) = \frac{1}{1+e^{-z}}$$

$$g'(z) = \frac{-1}{(1+e^{-z})^2} \frac{d}{dz} (1+e^{-z}) = \frac{-1 (e^{-z}) (-1)}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$g'(z) = \frac{1+e^{-z}}{(1+e^{-z})^2} \cdot \frac{-1}{(1+e^{-z})^2} = g(z)(1-g(z))$$

# DERIVATIVES OF ACTIVATION FUNCTIONS

TANH

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{u}{v}$$

$$g'(z) = \frac{v \frac{du}{dz} - u \frac{dv}{dz}}{v^2} = \frac{(e^z + e^{-z})(e^z + e^{-z}) - (e^z - e^{-z})(e^z - e^{-z})}{(e^z + e^{-z})^2}$$
$$= 1 - (g(z))^2$$

# ACTIVATION FUNCTIONS (SUMMARY)

RELU

$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z \geq 0 \\ 0; & z < 0 \end{cases}$$

L-RELU

$$g(z) = \begin{cases} z; & z \geq 0 \\ \alpha z; & z < 0 \end{cases}$$

$$g'(z) = \begin{cases} 1; & z \geq 0 \\ \alpha; & z < 0 \end{cases}$$

SGMOID

$$g(z) = \frac{1}{1 + e^{-z}}$$

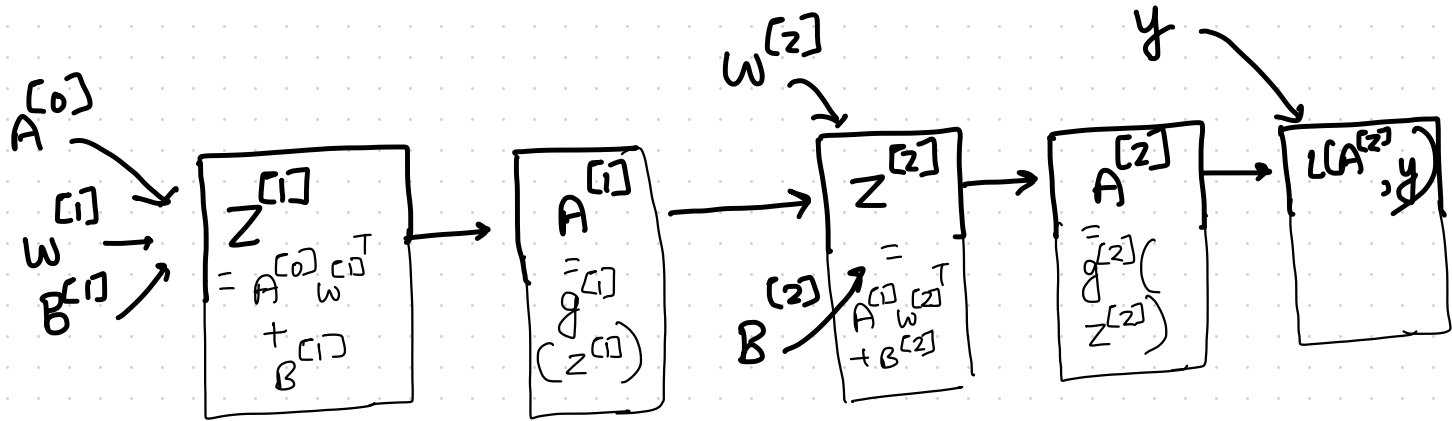
$$g'(z) = g(z) * (1 - g(z))$$

TANH

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g'(z) = 1 - (g(z))^2$$

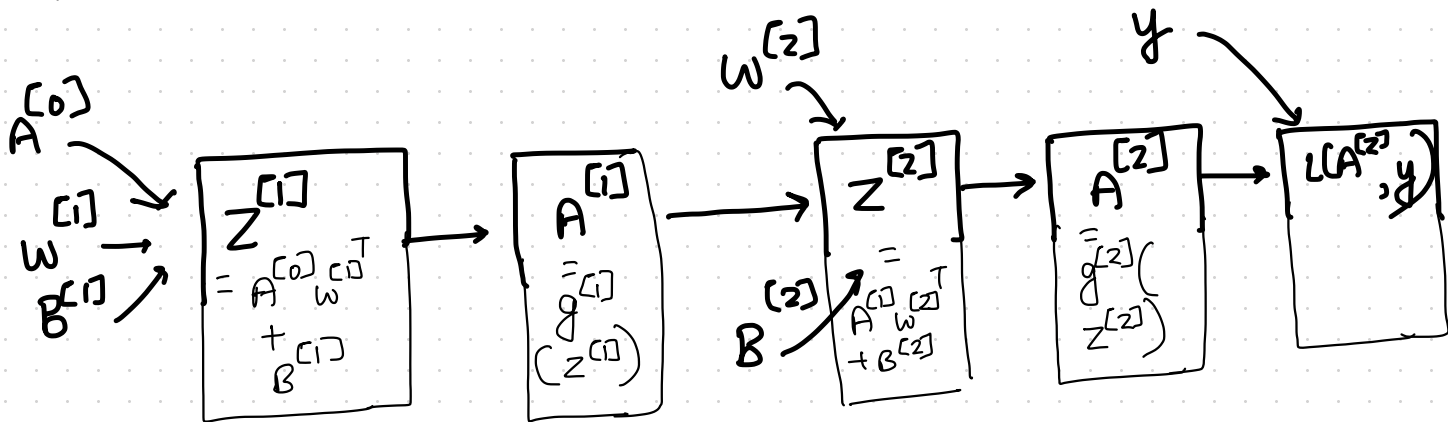
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$L(A^{[2]}, y) = \sum_{i=1}^N -y^{(i)} \log A^{[2]}_{(i)} - (1 - y^{(i)}) \log (1 - A^{[2]}_{(i)})$$

WRITE IN VECTOR FORM

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



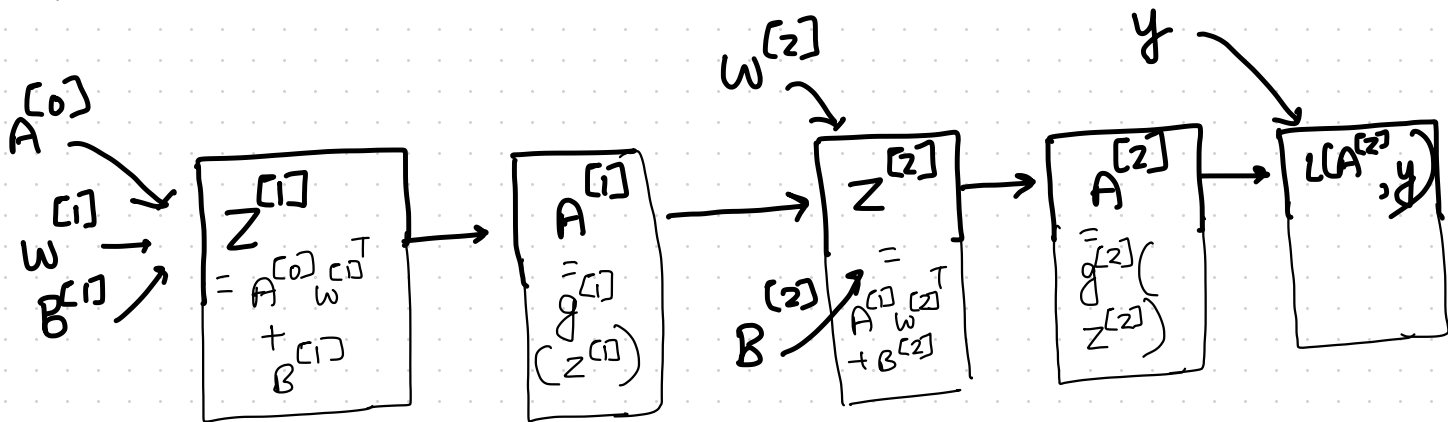
$$L(A^{[2]}, y) = \sum_{i=1}^N -y^{(i)} \log A^{[2]}_{(i)} - (1 - y^{(i)}) \log (1 - A^{[2]}_{(i)})$$

$$= -y^T \log(A^{[2]}) - (1 - y)^T \log(1 - A^{[2]})$$

APPLIED ELEMENT-WISE

$\begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{N \times 1}$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$L(A^{[2]}, y) = \sum_{i=1}^N -y^{(i)} \log A^{[2]}_{(i)} - (1 - y^{(i)}) \log (1 - A^{[2]}_{(i)})$$

$$= -y^T \log(A^{[2]}) - (1 - y)^T \log(1 - A^{[2]})$$

$\begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{N \times 1}$

APPLIED ELEMENT-WISE

$\frac{\partial L(A^{[2]}, y)}{\partial A^{[2]}}$  = ?



$$\text{Let } q = y^T \log(A^{[2]})$$

$$q = y^{(1)} \log A^{(1)[2]} + \dots + y^{(n)} A^{(n)[2]}$$

$$\text{Let } q = y^T \log(A^{[2]})$$

$$q = y^{(1)} \log A^{[2]}_{(1)} + \dots + y^{(N)} \log A^{[2]}_{(N)}$$

$$\frac{\partial q}{\partial A^{[2]}} = \begin{bmatrix} \frac{\partial}{\partial A^{[2]}_{(1)}} (y^{(1)} \log A^{[2]}_{(1)} + \dots) \\ \vdots \\ \frac{\partial}{\partial A^{[2]}_{(N)}} (y^{(1)} \dots) \end{bmatrix} = \begin{bmatrix} \frac{y^{(1)}}{A^{[2]}_{(1)}} \\ \vdots \\ \frac{y^{(N)}}{A^{[2]}_{(N)}} \end{bmatrix}$$

$$\frac{\partial q}{\partial A^{[2]}_{N \times 1}} = y_{N \times 1} \oslash A^{[2]}_{N \times 1} \quad \text{Element-wise division}$$

$$\text{Let } r = (1 - y)^T \log(1 - A^{[2]})$$

$$r = (1 - y^{(1)}) \log(1 - A_{(1)}^{[2]}) + \dots$$

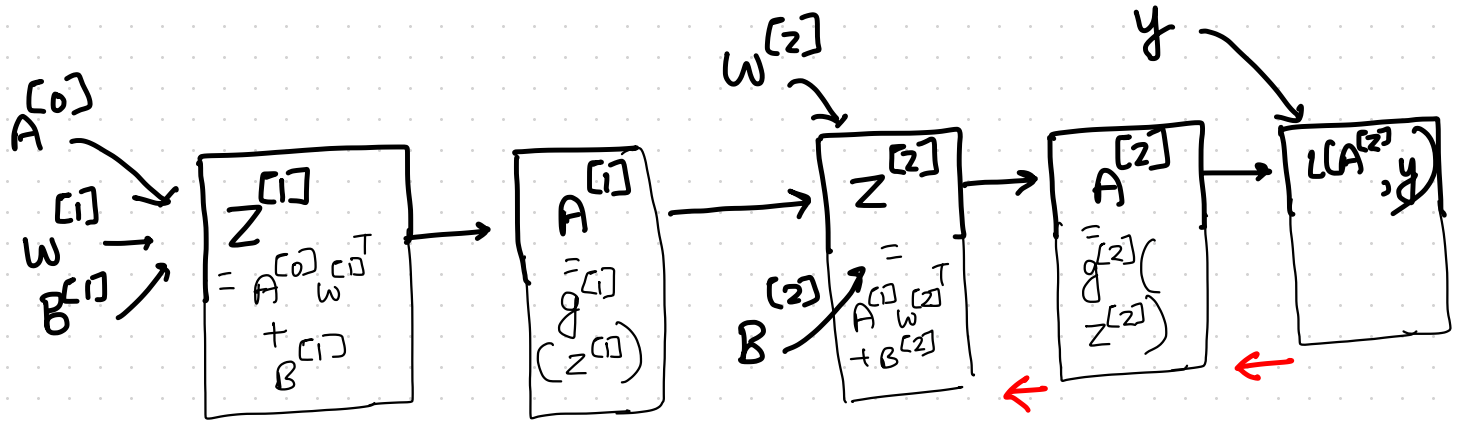
$$\frac{\partial r}{\partial A^{[2]}} = \begin{bmatrix} \frac{\partial}{\partial A_{(1)}^{[2]}} \left\{ (1 - y^{(1)}) \log(1 - A_{(1)}^{[2]}) + \dots \right\} \\ \vdots \\ \frac{\partial}{\partial A_{(N)}^{[2]}} \left\{ (1 - y^{(N)}) \log(1 - A_{(N)}^{[2]}) + \dots \right\} \end{bmatrix} = \begin{bmatrix} \frac{(1 - y^{(1)}) (-1)}{(1 - A_{(1)}^{[2]})} \\ \vdots \\ \frac{(1 - y^{(N)}) (-1)}{(1 - A_{(N)}^{[2]})} \end{bmatrix}$$

$$\frac{\partial r}{\partial A^{[2]}} = -(1 - y) \odot (1 - A^{[2]})$$

$N \times 1$                        $N \times 1$                        $N \times 1$

$$\frac{\partial L(A^{[2]}, y)}{\partial A^{[2]}} = -y \odot A^{[2]} + (1-y) \odot (1-A^{[2]})$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

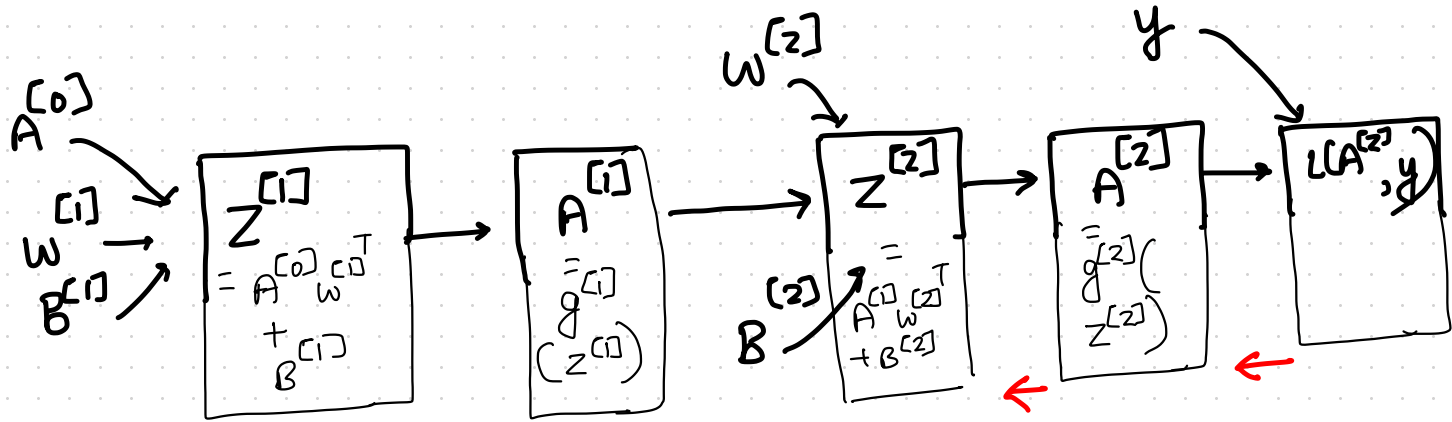


$$\frac{\partial L}{\partial Z^{[2]}} = ?$$

$$\frac{\partial L}{\partial Z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}}$$

(Chain Rule)

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial z^{[2]}} = ?$$

$$\frac{\partial L}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial z^{[2]}} \quad (\text{Chain Rule})$$

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g^{[2]}(z^{[2]}) (1 - g^{[2]}(z^{[2]}))$$

$$\therefore g^{[2]} = \text{SIGMOID}$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} \circ A^{[2]} \circ (1 - A^{[2]})$$

*Nx1*      *Nx1*      *Nx1*

Element-wise multiply

$$= (-y \circ A^{[2]} + (1 - y) \circ (1 - A^{[2]})) \left( A^{[2]} \circ (1 - A^{[2]}) \right)$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} \begin{matrix} A^{[2]} & \circ & (1 - A^{[2]}) \\ \text{N} \times 1 & & \text{N} \times 1 \end{matrix}$$

Element-wise

$$= \left( -y \circ A^{[2]} + (1 - y) \circ (1 - A^{[2]}) \right) \left( A^{[2]} (1 - A^{[2]}) \right)$$

$$= -y \circ A^{[2]} \circ A^{[2]} \circ (1 - A^{[2]}) + (1 - y) \circ (1 - A^{[2]}) \circ A^{[2]} \circ (1 - A^{[2]})$$



# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} A^{[2]} \circ (1 - A^{[2]})$$

*Element-wise*

$$= (-y \circ A^{[2]} + (1-y) \circ (1 - A^{[2]})) \circ (A^{[2]} (1 - A^{[2]}))$$

$$= -y \circ A^{[2]} \circ A^{[2]} \circ (1 - A^{[2]}) + (1-y) \circ (1 - A^{[2]}) \circ A^{[2]} \circ (1 - A^{[2]})$$
$$= -y \circ (1 - A^{[2]}) + (1-y) \circ A^{[2]} = \underbrace{-y}_{N \times 1} + \underbrace{y \circ A^{[2]}}_{N \times 1} + \underbrace{A^{[2]}}_{N \times 1} - \underbrace{y \circ A^{[2]}}_{N \times 1}$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$= \frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial g^{[2]}(z^{[2]})}{\partial z^{[2]}} = \frac{\partial L}{\partial A^{[2]}} g(z^{[2]}) (1 - g(z^{[2]}))$$

$$= \frac{\partial L}{\partial A^{[2]}} \underbrace{A^{[2]}}_{N \times 1} \circ \underbrace{(1 - A^{[2]})}_{N \times 1}$$

Element-wise

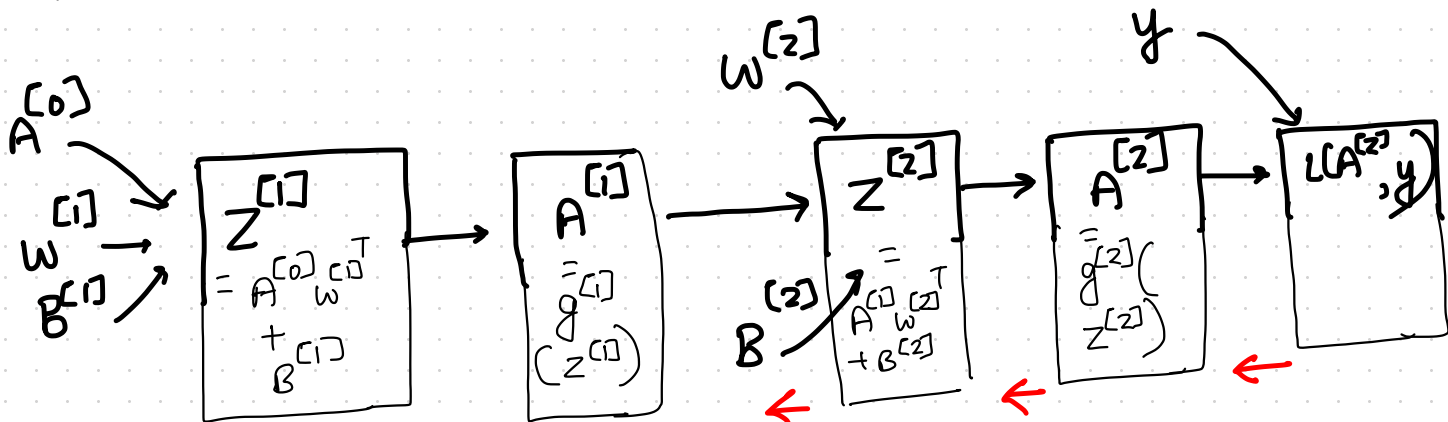
$$= (-y \circ A^{[2]} + (1 - y) \circ (1 - A^{[2]})) \circ (A^{[2]} (1 - A^{[2]}))$$

$$= -y \circ A^{[2]} \circ A^{[2]} \circ (1 - A^{[2]}) + (1 - y) \circ (1 - A^{[2]}) \circ A^{[2]} \circ (1 - A^{[2]})$$

$$= -y \circ (1 - A^{[2]}) + (1 - y) \circ A^{[2]} = \underbrace{-y}_{N \times 1} + \underbrace{y \circ A^{[2]}}_{N \times 1} + \underbrace{A^{[2]}}_{N \times 1} - \underbrace{y \circ A^{[2]}}_{N \times 1}$$

$$\boxed{\frac{\partial L}{\partial z^{[2]}} = A^{[2]} - y}$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial w^{[2]}}$$

$\in \mathbb{R}^{N^{[2]} \times N^{[1]}}$

(chain Rule)

SAME DIMENSION ( $\because L$  is scalar)

ASIDE

\* GRADIENT: VECTOR IN, SCALAR OUT

\* JACOBIAN: VECTOR IN, VECTOR OUT

$$f: \mathbb{R}^N \rightarrow \mathbb{R}^M$$

$$\text{I/P: } \mathbb{R}^N$$

$$\text{O/P: } \mathbb{R}^M$$

$$y = f(x)$$

Derivative of 'f' at 'x' called Jacobian is  $m \times n$  matrix

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & & \frac{\partial y_m}{\partial x_n} \end{pmatrix}_{m \times n}$$

GENERALISED JACOBIAN: TENSOR IN, TENSOR OUT

$$f: \mathbb{R}^{N_1 \times \dots \times N_{D_x}} \rightarrow \mathbb{R}^{M_1 \times \dots \times M_{D_y}}$$

I/P:  $D_x$ -dimensional tensor of shape  $N_1 \times \dots \times N_{D_x}$

O/P:  $D_y$ -dimensional tensor of shape  $M_1 \times \dots \times M_{D_y}$

$$y = f(x)$$

Then;  $\frac{\partial y}{\partial x} = \text{Gen. Jacobian} = (M_1 \times \dots \times M_{D_y}) \times (N_1 \times \dots \times N_{D_x})$

shape

# BACK PROP. WITH TENSORS

$$\text{Let } G = \alpha \beta$$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G: N \times M$$

## BACK PROP. WITH TENSORS

$$\text{Let } G = \alpha \beta$$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G: N \times M$$

→ We are given Loss  $L$  as fun<sup>n</sup> of  $G$

→ we also know  $\frac{\partial L}{\partial G}$

# BACK PROP. WITH TENSORS

$$\text{Let } G = \alpha \beta$$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G: N \times M$$

→ We are given Loss  $L$  as func<sup>n</sup> of  $G$

→ we also know  $\frac{\partial L}{\partial G}$

→ Q:  $\frac{\partial L}{\partial \alpha} = ?$  ;  $\frac{\partial L}{\partial \beta} = ?$



# BACK PROP. WITH TENSORS

$$\text{Let } G = \alpha \beta$$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G: N \times M$$

Let's choose:  $N=1; D=2; M=3$

# BACK PROP. WITH TENSORS

$$\text{Let } G = \alpha \beta$$

$$\alpha: N \times D$$

$$\beta: D \times M$$

$$G: N \times M$$

Let's choose:  $N=1; D=2; M=3$

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

# BACK PROP. WITH TENSORS

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

$$\alpha \beta = \left[ \alpha_{1,1} \beta_{1,1} + \alpha_{1,2} \beta_{2,1} ; \alpha_{1,1} \beta_{1,2} + \alpha_{1,2} \beta_{2,2} ; \alpha_{1,1} \beta_{1,3} + \alpha_{1,2} \beta_{2,3} \right]$$

# BACK PROP. WITH TENSORS

$$G_{1 \times 3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1 \times 2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2 \times 3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

$$\alpha\beta = \left[ \alpha_{1,1} \beta_{1,1} + \alpha_{1,2} \beta_{2,1} ; \alpha_{1,1} \beta_{1,2} + \alpha_{1,2} \beta_{2,2} ; \alpha_{1,1} \beta_{1,3} + \alpha_{1,2} \beta_{2,3} \right]$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,1}} = \left[ \beta_{1,1} ; \beta_{1,2} ; \beta_{1,3} \right]$$

$$\frac{\partial(\alpha\beta)}{\partial \alpha_{1,2}} = \left[ \beta_{2,1} ; \beta_{2,2} ; \beta_{2,3} \right]$$

## BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}] : \text{Shape} = \text{Shape of } \alpha\beta \times \text{Shape of } \alpha_{1,1}$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}] \quad \text{Shape} = (N \times M) \times 1$$

# BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}]$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}]$$

Shape = Shape of  $\alpha\beta$  x Shape of  $\alpha_{1,1}$   
=  $(N \times M) \times 1$

Shape =  $(N \times M) \times 1$

Generalised  
Tensor shape

# BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}]$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}]$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

Generalised  
shape =  
 $1 \times (N \times M)$   
↳ why?

## BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}]$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}]$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

Generalised  
shape =  
 $1 \times (N \times M)$   
 $\hookrightarrow$   
 $\therefore L$  is a  
scalar



## BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}]$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}]$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

# BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}] = \frac{\partial G}{\partial\alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}] = \frac{\partial G}{\partial\alpha_{1,2}}$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}}$$

# BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}] = \frac{\partial G}{\partial\alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}] = \frac{\partial G}{\partial\alpha_{1,2}}$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}}$$

$\xrightarrow{1 \times (N \times M)} \quad (N \times M) \times 1$

# BACK PROP. WITH TENSORS

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,1}} = [\beta_{1,1} ; \beta_{1,2} ; \beta_{1,3}] = \frac{\partial G}{\partial\alpha_{1,1}}$$

$$\frac{\partial(\alpha\beta)}{\partial\alpha_{1,2}} = [\beta_{2,1} ; \beta_{2,2} ; \beta_{2,3}] = \frac{\partial G}{\partial\alpha_{1,2}}$$

We know  $\frac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} \quad \frac{\partial L}{\partial \alpha_{1,2}} \right]$$

$$\frac{\partial L}{\partial \alpha_{1,1}} = \frac{\partial L}{\partial G} \cdot \frac{\partial G}{\partial \alpha_{1,1}} = [dG_{1,1} \beta_{1,1} + dG_{1,2} \beta_{1,2} + dG_{1,3} \beta_{1,3}]$$

$(1 \times (n \times m)) \times ((n \times m) \times 1) \rightarrow$  Dot product of two.

# BACK PROP. WITH TENSORS

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} ; \frac{\partial L}{\partial \alpha_{1,2}} \right] = \left( \frac{\partial L}{\partial G} \right) \beta^T$$

$1 \times (N \times D)$        $1 \times (N \times M)$        $M \times D$

# BACK PROP. WITH TENSORS

$$\frac{\partial L}{\partial \alpha} = \left[ \frac{\partial L}{\partial \alpha_{1,1}} ; \frac{\partial L}{\partial \alpha_{1,2}} \right] = \left( \frac{\partial L}{\partial G} \right) \beta^T$$

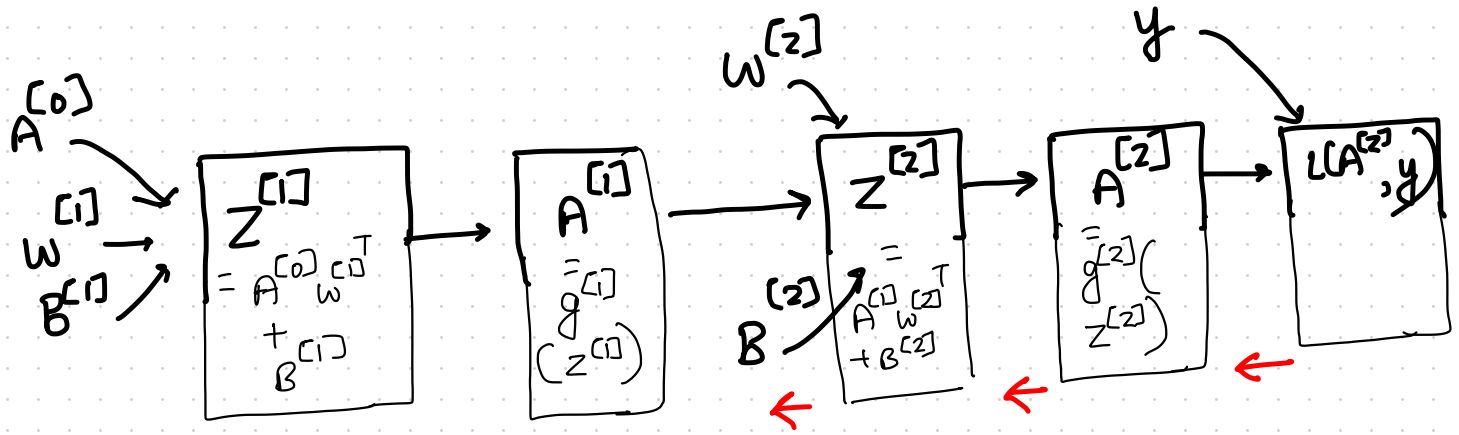
$1 \times (N \times D)$        $1 \times (N \times M)$        $M \times D$

Similarly;

$$\frac{\partial L}{\partial \beta} = \alpha^T \frac{\partial L}{\partial G}$$

$1 \times (D \times M)$        $D \times N$        $1 \times (N \times M)$

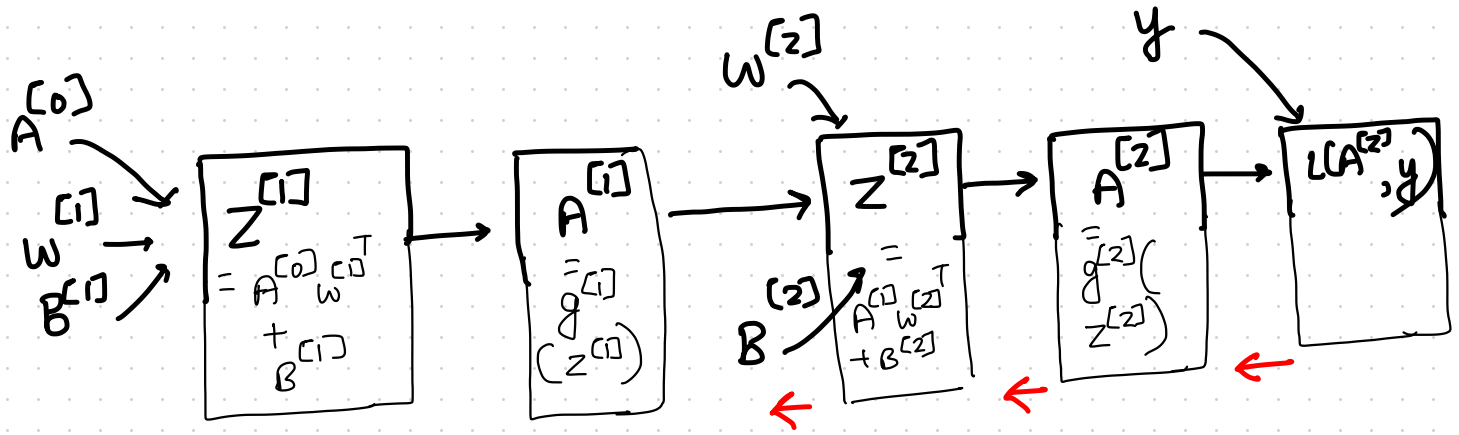
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



Equivalence from Aside

$$\begin{aligned}
 G &\leftrightarrow Z^{[2]} \\
 \alpha &\leftrightarrow A^{[1]} \\
 \beta &\leftrightarrow w^{[2]T}
 \end{aligned}$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



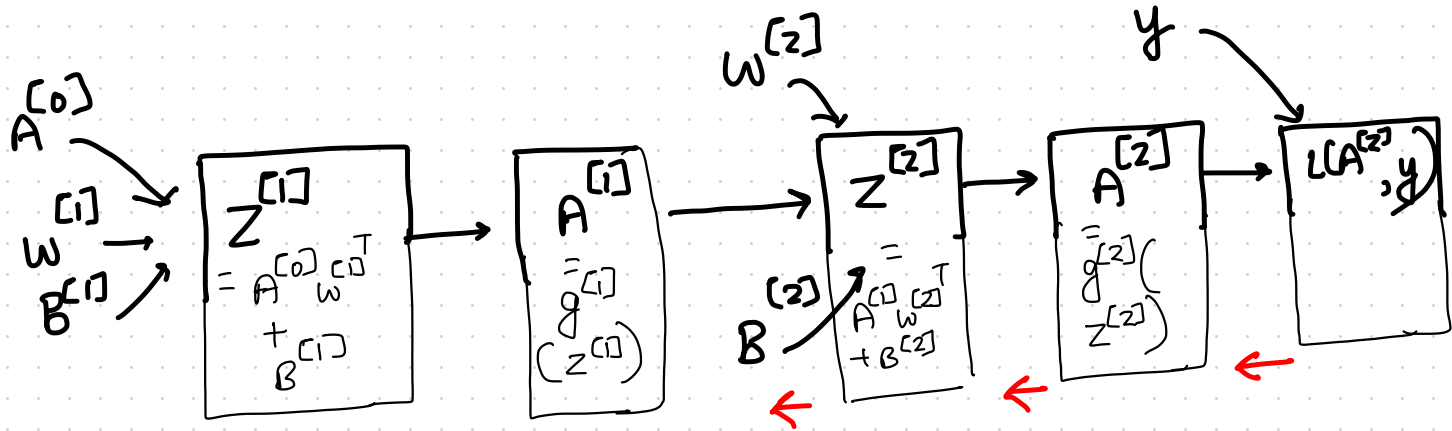
Equivalence from Aside

$$\begin{aligned}
 G &\leftrightarrow z^{[2]} \\
 \alpha &\leftrightarrow A^{[1]} \\
 \beta &\leftrightarrow w^{[2]T}
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial L}{\partial w^{[2]T}} &= A^{[1]T} \frac{\partial L}{\partial z^{[2]}} \\
 &= A^{[1]T} (A^{[2]} - y)
 \end{aligned}$$



# COMPUTATION GRAPH (FOR XOR EXAMPLE)

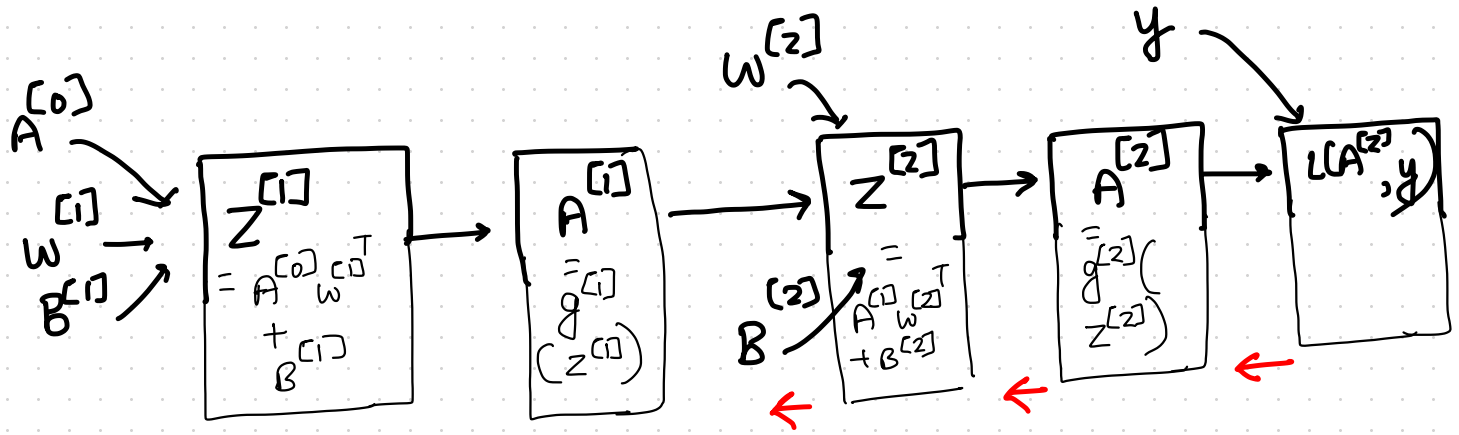


$$\frac{\partial L}{\partial W^{[2]T}} = A^{[1]T} (A^{[2]} - y)$$

$$\Rightarrow \frac{\partial L}{\partial W^{[2]}} = (A^{[2]} - y)^T A^{[1]}$$

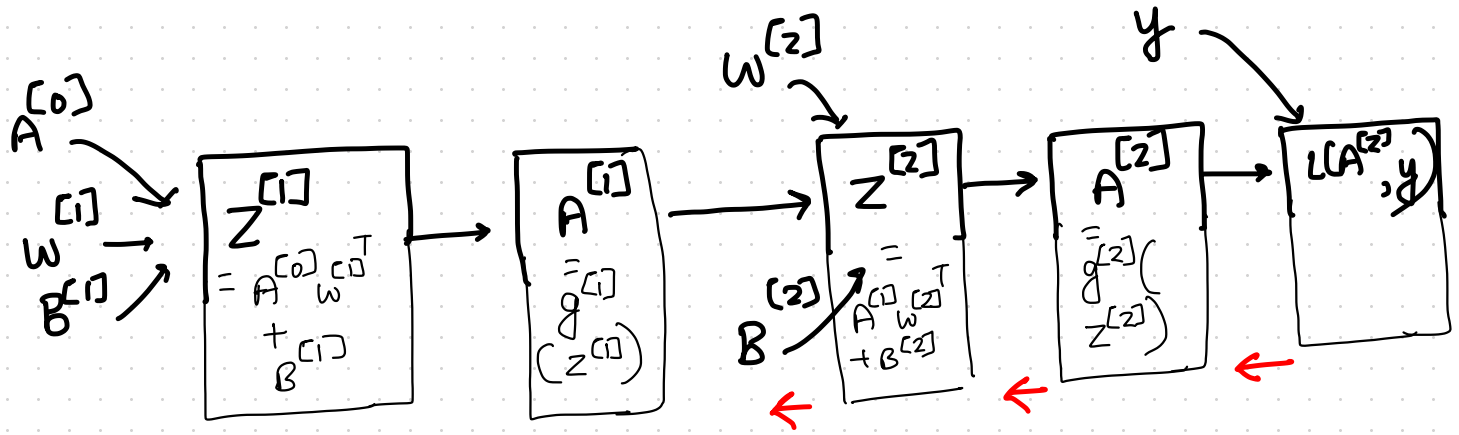
$(N^{[2]} \times N^{[1]})$        $(N^{[2]} \times N)$        $(N \times N^{[1]})$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial B^{[2]}} = ?$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

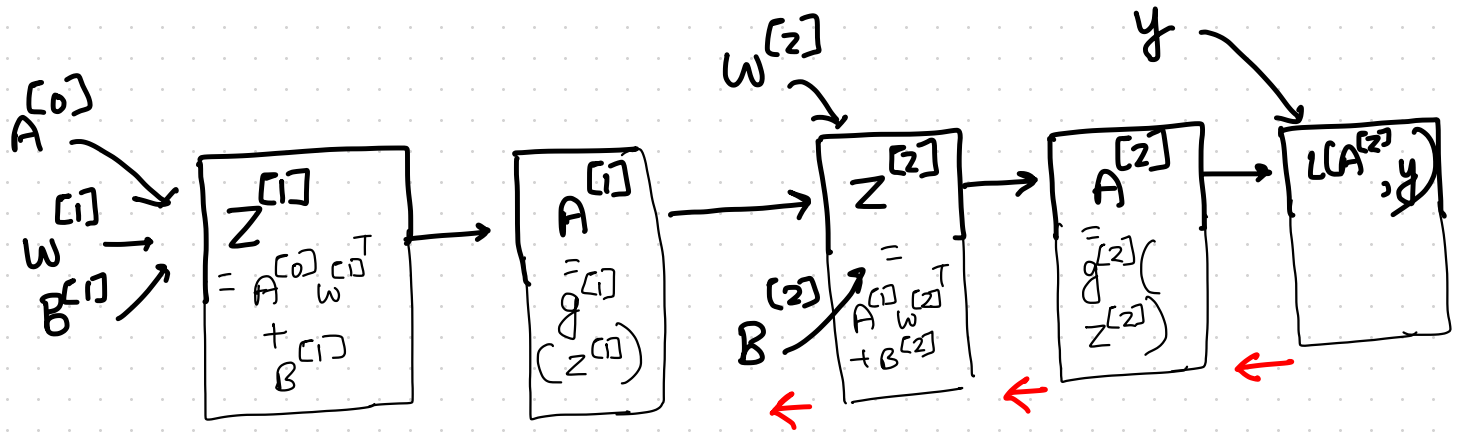


$$\frac{\partial L}{\partial B^{[2]}} = ?$$

$$\text{Let } C = b^{[2]T}$$

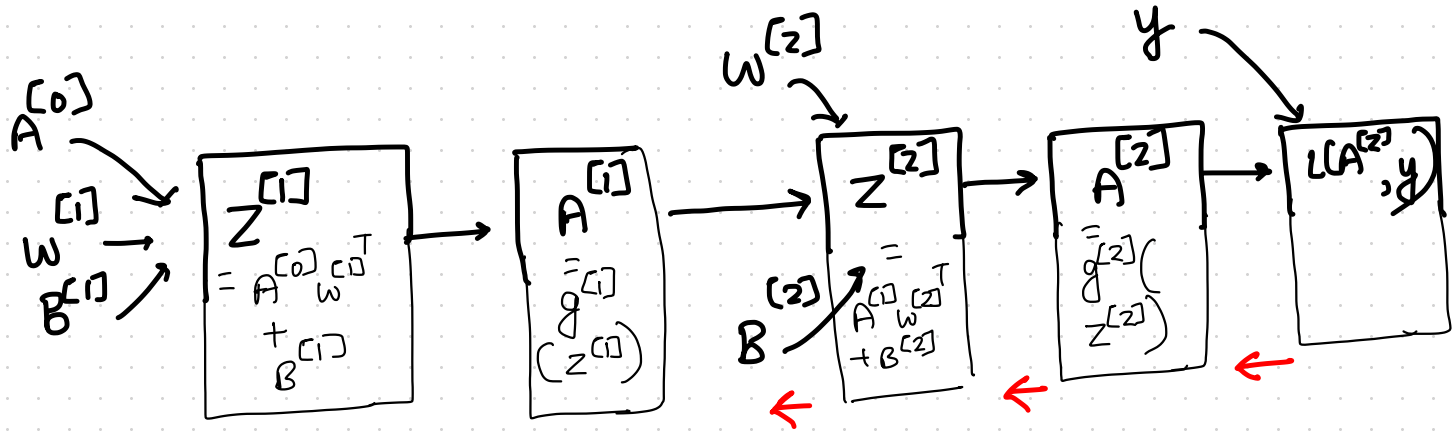
$$B^{[2]} = \begin{bmatrix} -C- \\ -C- \\ \vdots \\ -C- \end{bmatrix}$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial c} = ?$$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

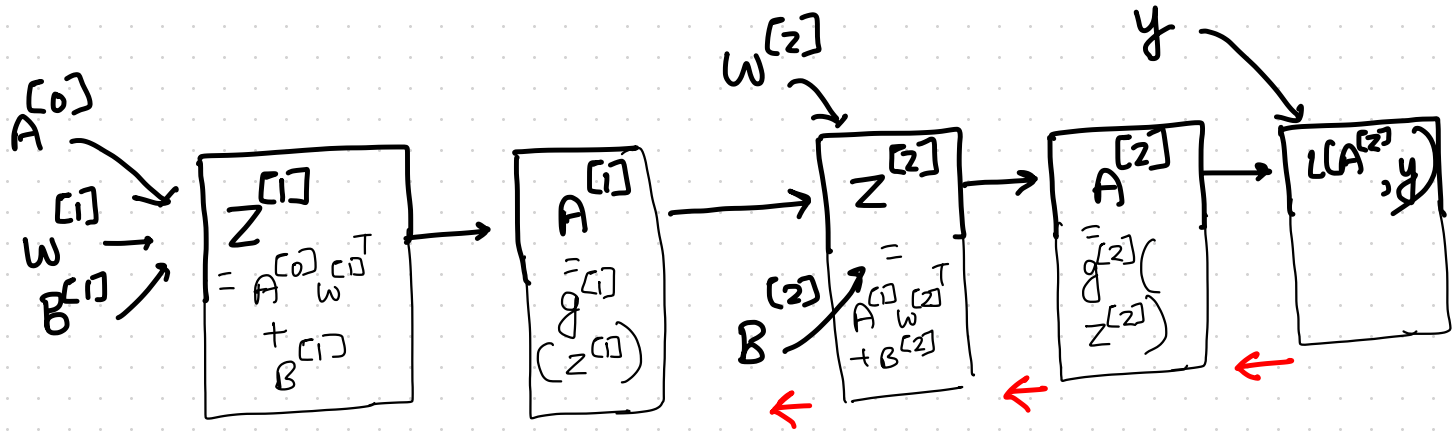


$$\frac{\partial L}{\partial c} = \left[ \frac{\partial L}{\partial c_1} \dots \frac{\partial L}{\partial c_{N^{[2]}}} \right]$$

$(1 \times (1 \times N^{[2]}))$

$$= \left[ \frac{\partial L}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial c_1} \dots \right]$$

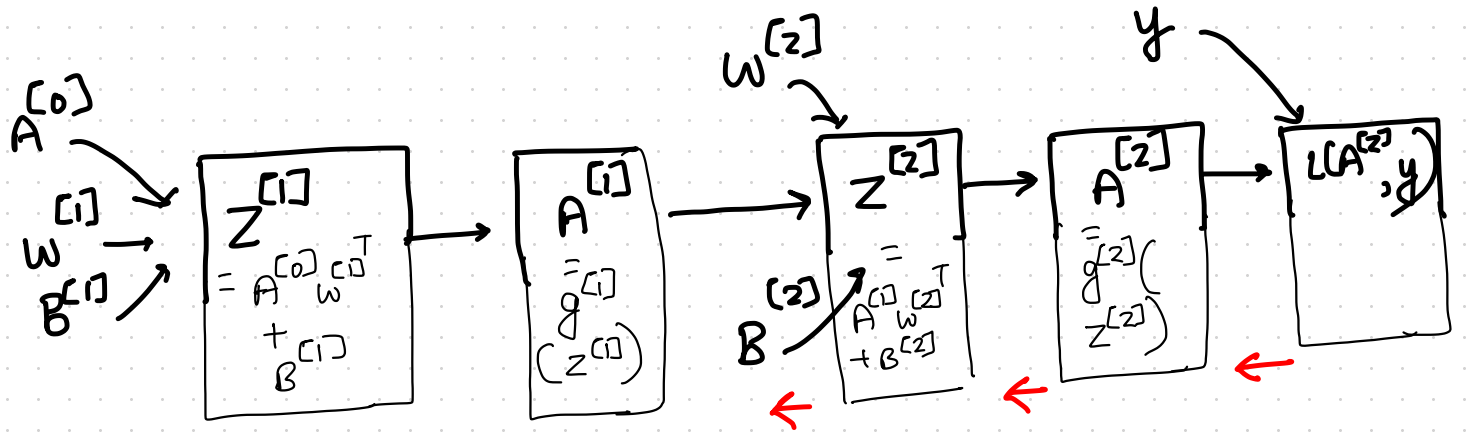
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial c_1} = \frac{\partial L}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial c_1}$$

$$Z^{(2)} = \begin{bmatrix} z_{1,1}^{(2)} & \dots & z_{1,N^{(2)}}^{(2)} \\ \vdots & & \vdots \\ z_{N,1}^{(2)} & \dots & z_{N,N^{(2)}}^{(2)} \end{bmatrix}_{N \times N^{(2)}} = A^{(2)} W^{(2)T} + \begin{bmatrix} c_1 & c_2 & \dots & c_{N^{(2)}} \\ c_1 & c_2 & \dots & c_{N^{(2)}} \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

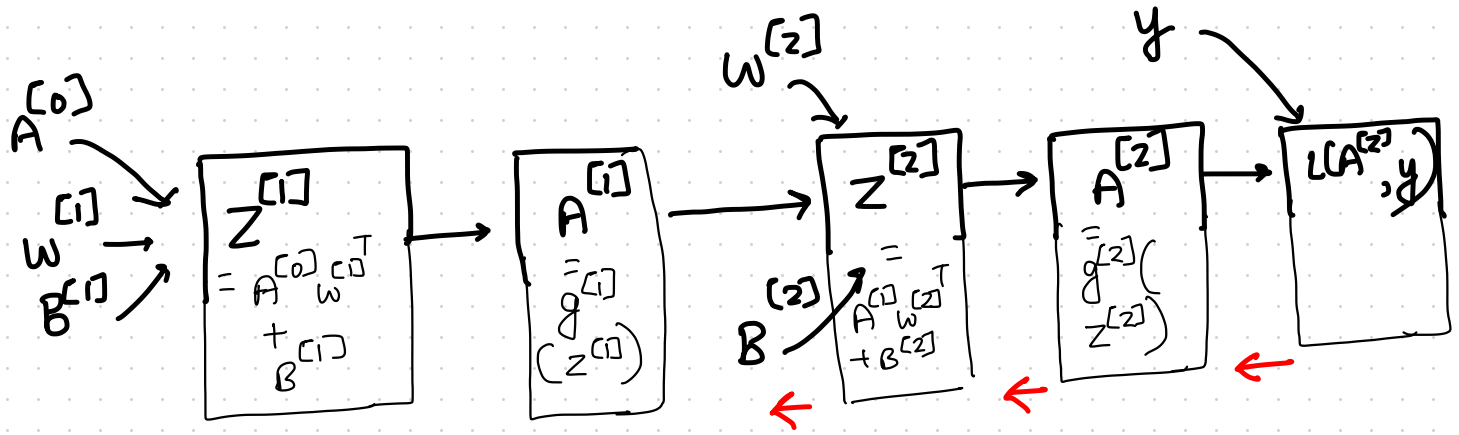
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial z^{[2]}}{\partial c_1} = \begin{bmatrix} \frac{\partial c_1}{\partial c_1} & \frac{\partial c_2}{\partial c_1} & \dots \\ \vdots & \dots & \dots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

(N x N<sup>[2]</sup>) x (1)

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial z}{\partial c_1} = \begin{bmatrix} \frac{\partial c_1}{\partial c_1} & \frac{\partial c_2}{\partial c_1} & \dots \\ \vdots & \vdots & \dots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

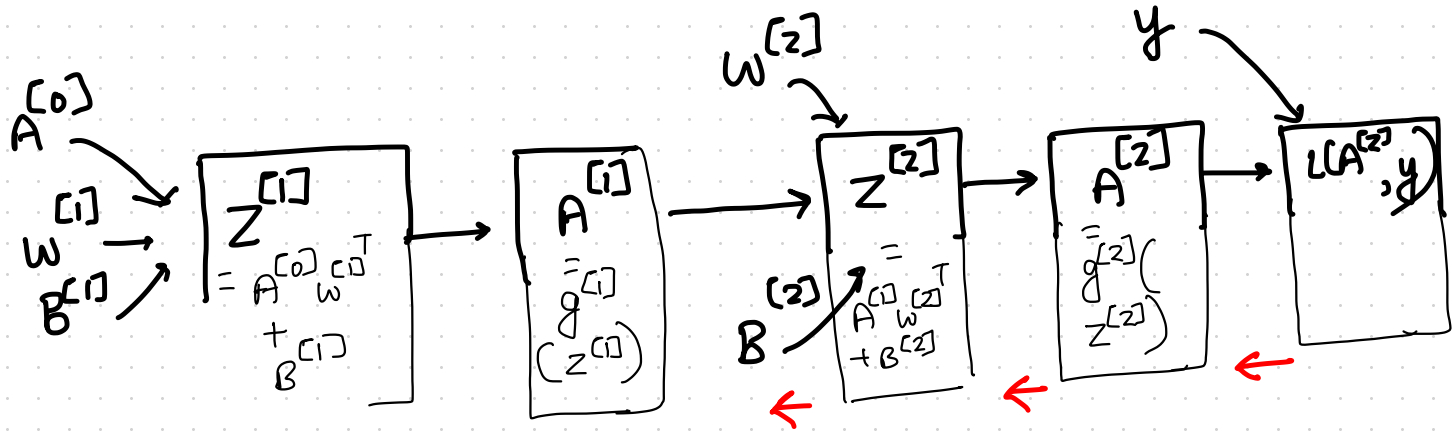
$(N \times N^{[2]}) \times (1)$

$$\frac{\partial L}{\partial c_1} = \frac{\partial L}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial c_1}$$

$1 \times (N \times N^{[2]}) \quad (N \times N^{[2]}) \times 1$



# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\text{Let } \frac{\partial L}{\partial z^{[2]}} = \begin{bmatrix} dz_{1,1}^{[2]} & \dots & dz_{1,N^{[2]}}^{[2]} \\ dz_{N,1}^{[2]} & \dots & dz_{N,N^{[2]}}^{[2]} \end{bmatrix}$$

$$\text{Then } \frac{\partial L}{\partial c_1} = \sum_{i=1}^N dz_{i,1}^{[2]}$$

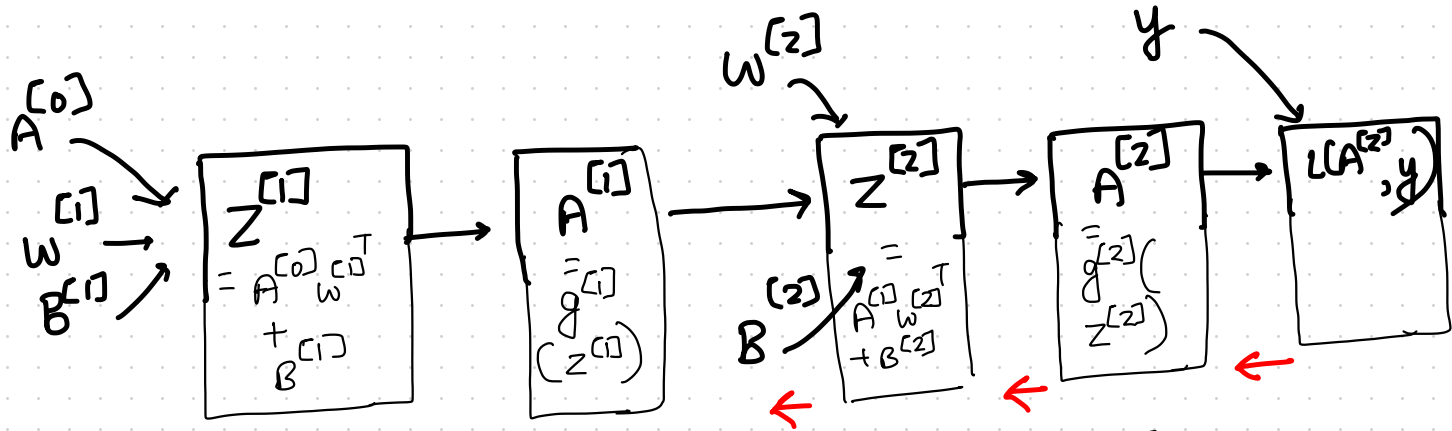
# COMPUTATION GRAPH (FOR XOR EXAMPLE)

$$\frac{\partial L}{\partial c_1} = \sum_{i=1}^N dz_{i,1}^{[2]}$$

$$\Rightarrow \frac{\partial L}{\partial c} = \left[ \sum_{i=1}^N dz_{i,1}^{[2]} \ ; \ \sum_{i=1}^N dz_{i,2}^{[2]} \ \dots \ ; \ \sum_{i=1}^N dz_{i,N}^{[2]} \right]$$

$$\Rightarrow \frac{\partial L}{\partial b} = \left( \frac{\partial L}{\partial c} \right)^T = \begin{bmatrix} \sum_{i=1}^N dz_{i,1}^{[2]} \\ \vdots \\ \sum_{i=1}^N dz_{i,N}^{[2]} \end{bmatrix} \quad N^{[2]} \times 1$$

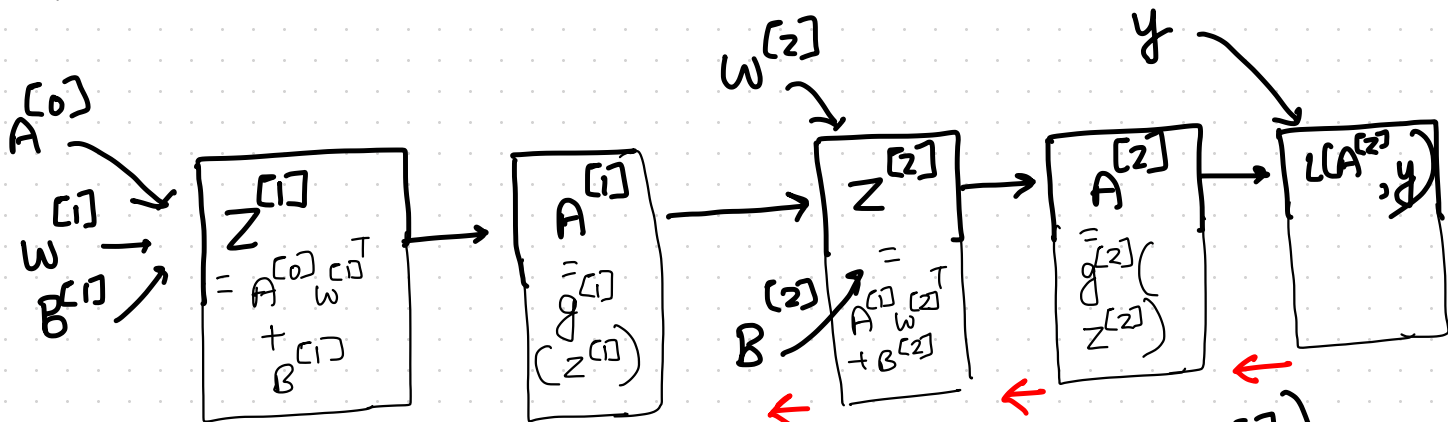
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial A^{[1]}} = \frac{\partial L}{\partial z^{[2]}} \frac{\partial z^{[2]}}{\partial A^{[1]}} = \frac{\partial L}{\partial z^{[2]}} (w^{[2]})^T = \frac{\partial L}{\partial z^{[2]}} w^{[2]}$$



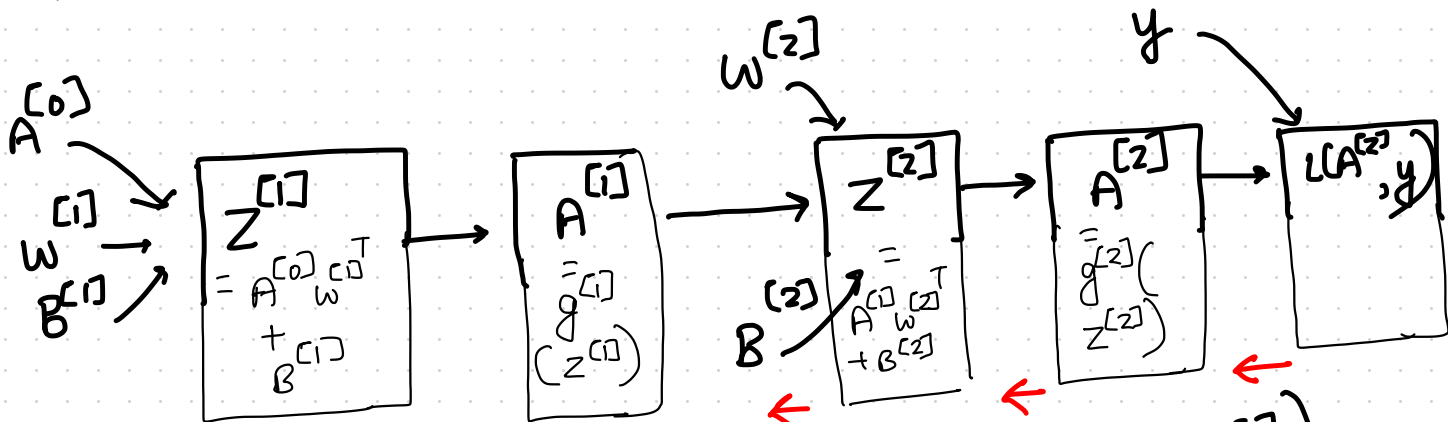
# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial z^{[1]}} = \frac{\partial L}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial z^{[1]}} = \left( \frac{\partial L}{\partial z^{[2]}} w^{[2]} \right) \left( \frac{\partial A^{[1]}}{\partial z^{[1]}} \right)$$

$1 \times (N \times N^{[1]}) \quad \times \quad (N^{[2]} \times N^{[1]}) \quad \times \quad (N \times N^{[1]})$

# COMPUTATION GRAPH (FOR XOR EXAMPLE)

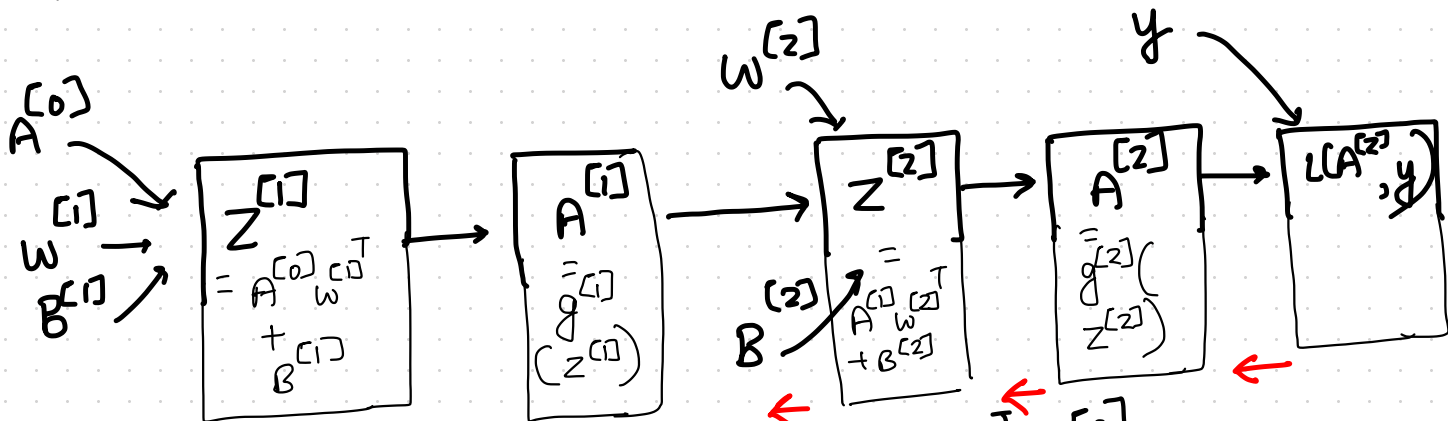


$$\begin{aligned}
 \frac{\partial L}{\partial Z^{[1]}} &= \frac{\partial L}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial Z^{[1]}} = \left( \frac{\partial L}{\partial Z^{[2]}} W^{[2]} \right) \left( \frac{\partial A^{[1]}}{\partial Z^{[1]}} \right) \\
 &= \frac{\partial L}{\partial Z^{[1]}} W^{[2]} \odot g^{[1]'}(Z^{[1]})
 \end{aligned}$$

$1 \times (N \times N^{[2]}) \times (N^{[2]} \times N^{[1]}) \times (N \times N^{[1]}) \times (N \times N^{[1]})$

Element wise or dot product of Jacobian

# COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial W^{[1]T}} = A^{[0]T} \frac{\partial L}{\partial z^{[1]}} \Rightarrow \frac{\partial L}{\partial W^{[1]}} = \left( \frac{\partial L}{\partial z^{[1]}} \right)^T A^{[0]}$$

$$\frac{\partial L}{\partial b^{[1]}} = \begin{bmatrix} \sum_{i=1}^N dz_{i,1}^{[1]} \\ \vdots \\ \sum_{i=1}^N dz_{i,N^{[2]}}^{[1]} \end{bmatrix} \quad N^{[2]} \times 1$$