RECENT SUCCESSES OF NN

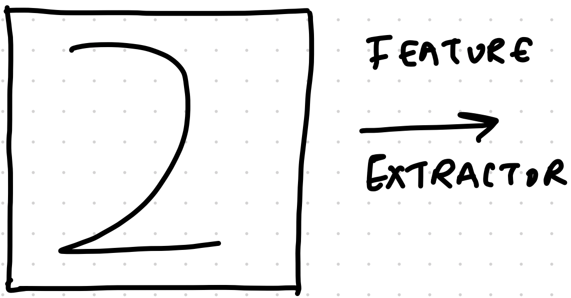
* State-of-the-art (SOTA) in most fields

PARADIGM CHANGE

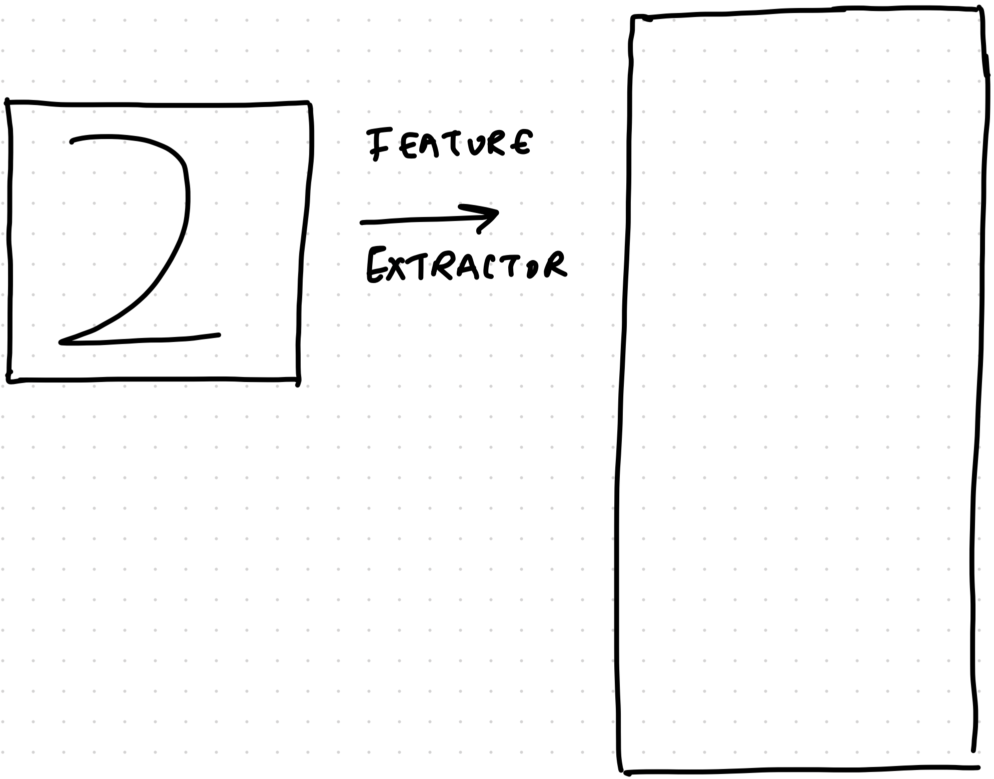
PARADIGM CHANGE

$\boxed{2}$

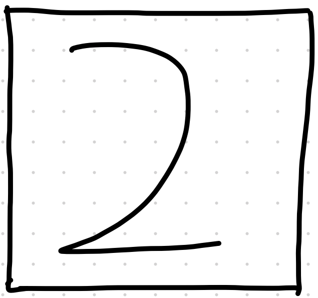
PARADIGM CHANGE



PARADIGM CHANGE



PARADIGM CHANGE



FEATURE
$\longrightarrow$
EXTRACTOR

INTENSITY

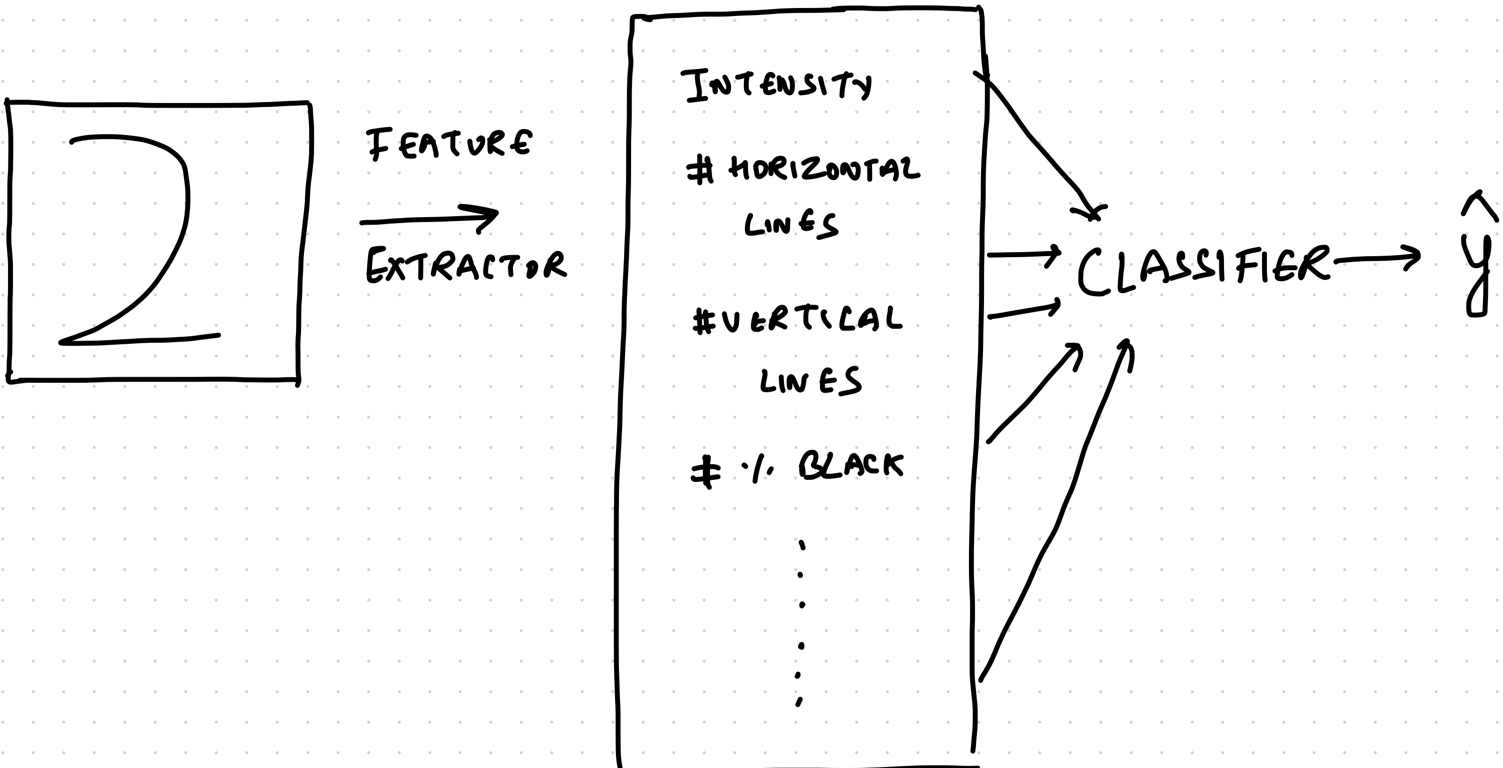
HORIZONTAL LINES

#VERTICAL LINES

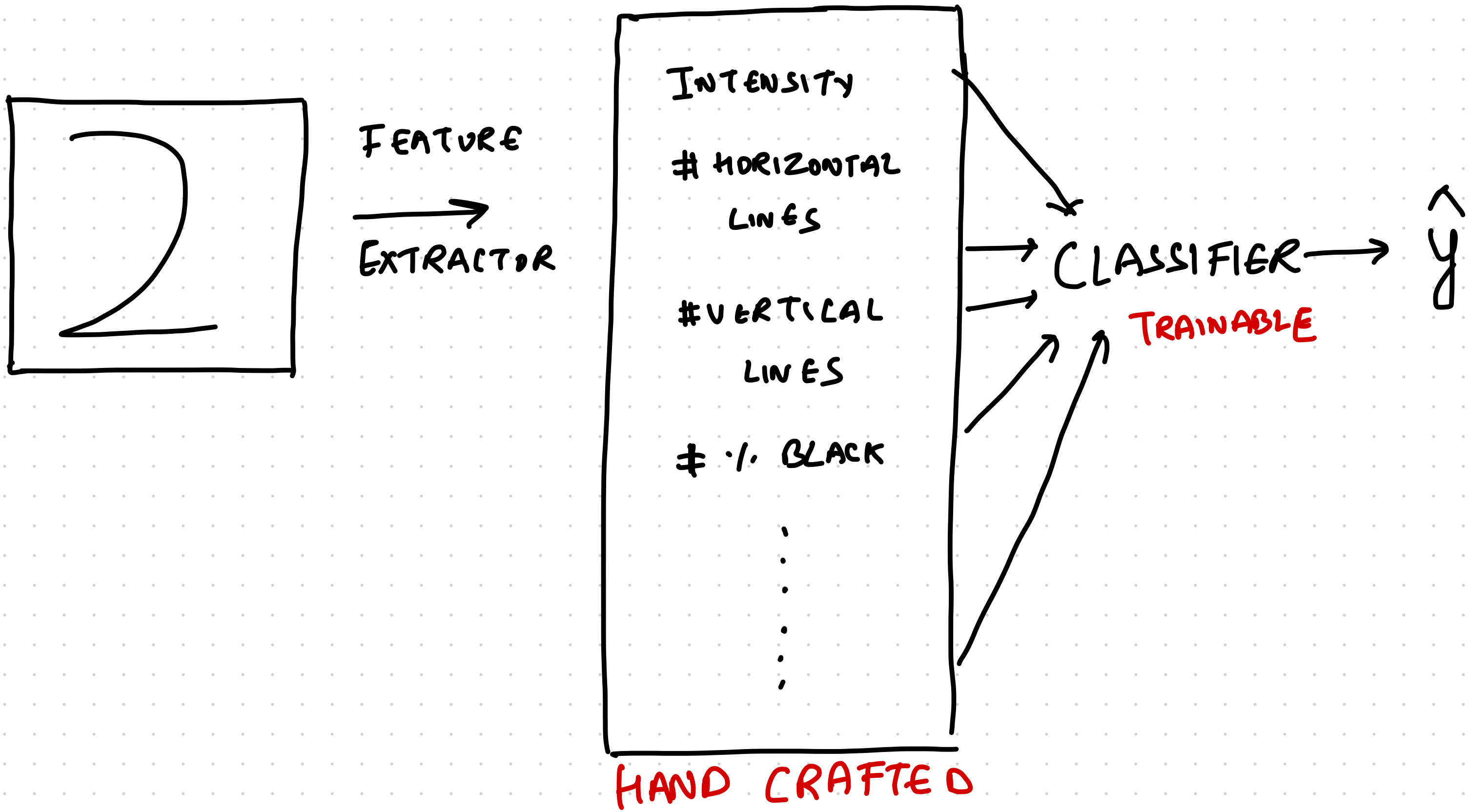
% BLACK

⋮

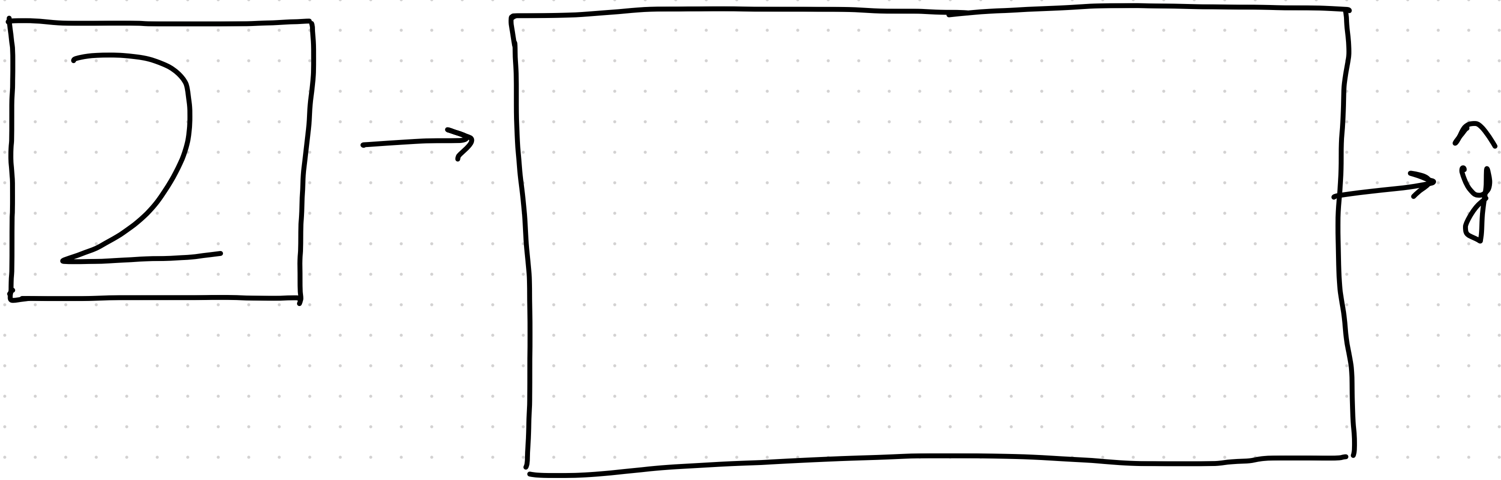
PARADIGM CHANGE



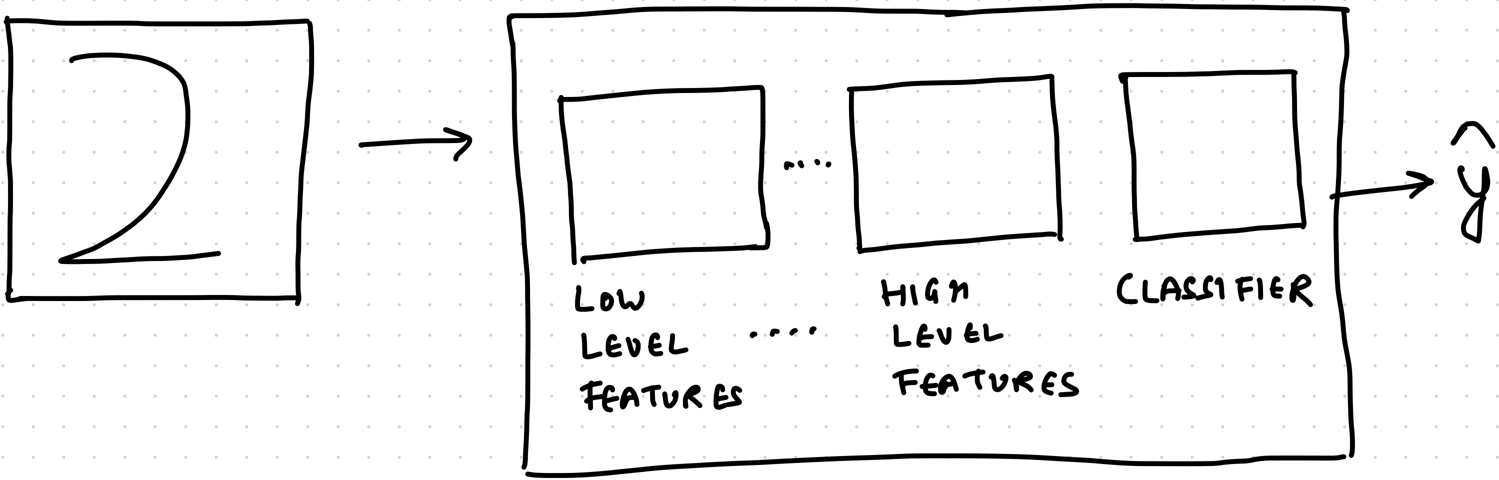
PARADIGM CHANGE



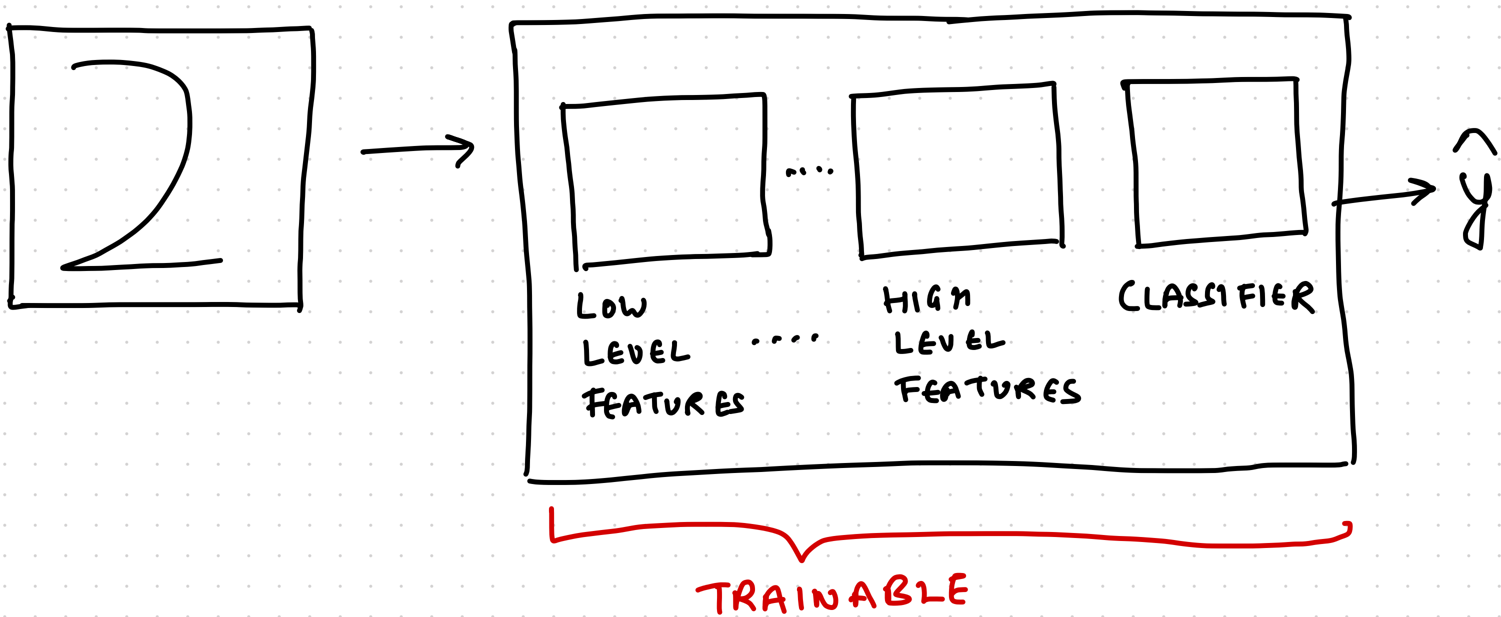
PARADIGM CHANGE (NNS)



PARADIGM CHANGE (NNS)



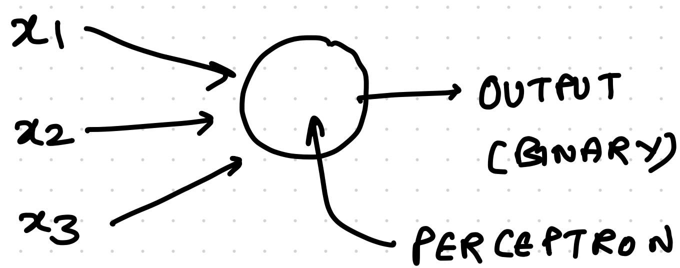
PARADIGM CHANGE (NNS)



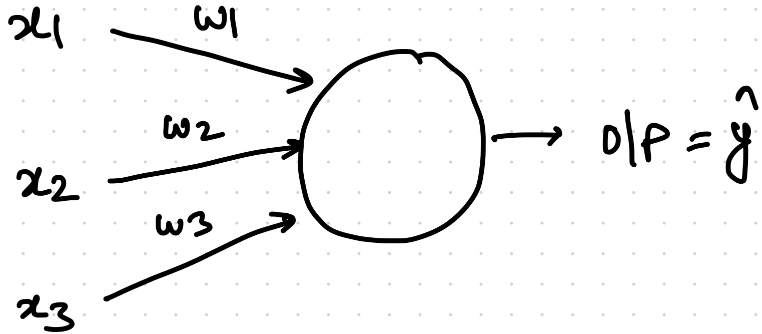
PERCEPTRON

- ARTIFICIAL NEURON DEVELOPED BY ROSENBLATT IN 1960S INSPIRED BY MC CULLOCH & PITTS

BINARY I/P

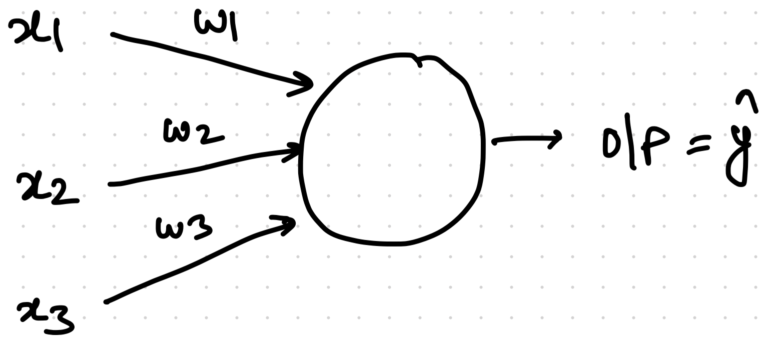


PERCEPTRON



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

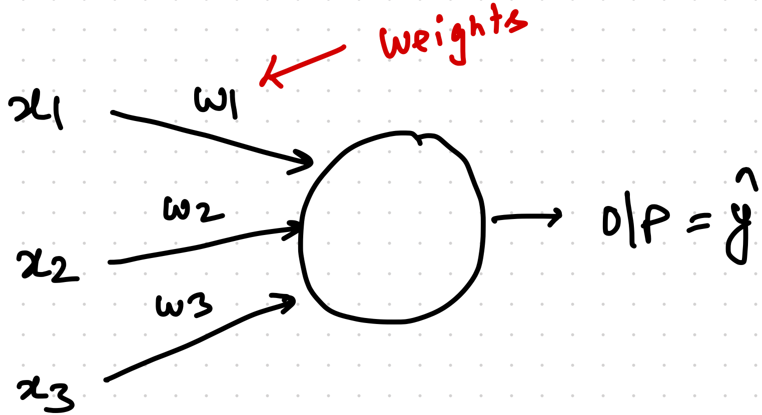
PERCEPTRON



$$O|P = \hat{y} = \begin{cases} 0 \; ; & \Sigma\, w_i x_i \le \text{THRESHOLD} \\ 1 \; ; & \Sigma\, w_i x_i > \text{THRESHOLD} \end{cases}$$

NEURONS "FIRE" ABOVE THRESHOLD

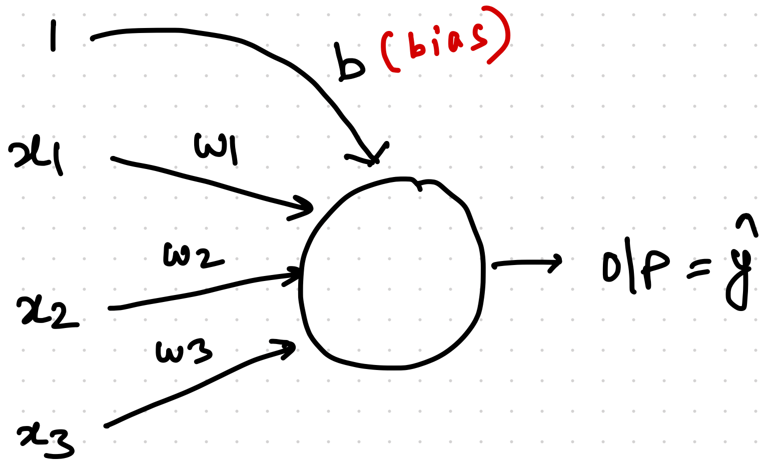
PERCEPTRON



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

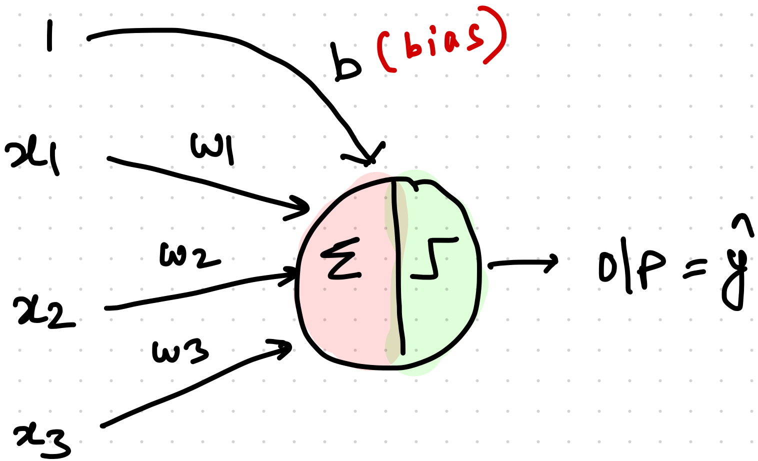
NEURONS "FIRE" ABOVE THRESHOLD

PERCEPTRON



$$O|P = \hat{y} = \begin{cases} 0 \, ; & \sum w_i x_i + b \leq 0 \\ 1 \, ; & \sum w_i x_i + b > 0 \end{cases}$$

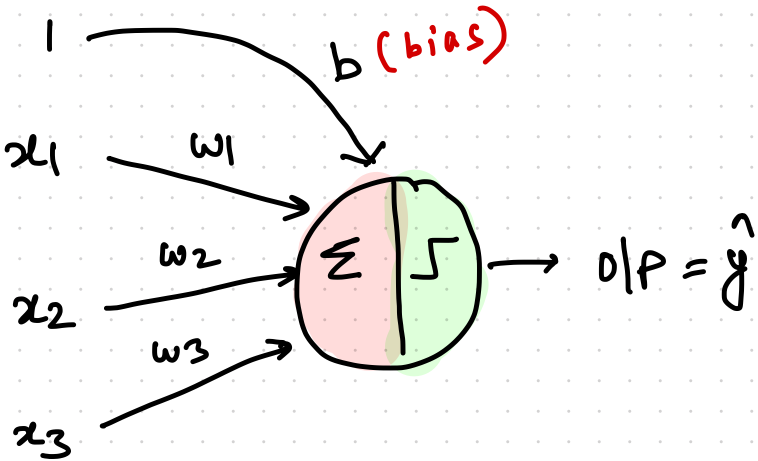
PERCEPTRON



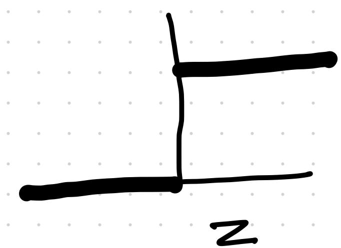
NEURON HAS 2 COMPONENTS

1. SUMMATION
2. ACTIVATION : STEP FUNCTION

PERCEPTRON



(SIGN/STEP)
Activation



NEURON HAS 2 COMPONENTS

1. SUMMATION
2. ACTIVATION : STEP FUNCTION

LEARNING BINARY GATES

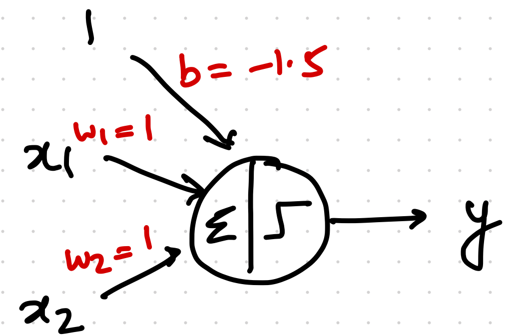
Q) FOR 2 I/ps x_1 & x_2 learn w_i's and b for BINARY AND

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_i'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

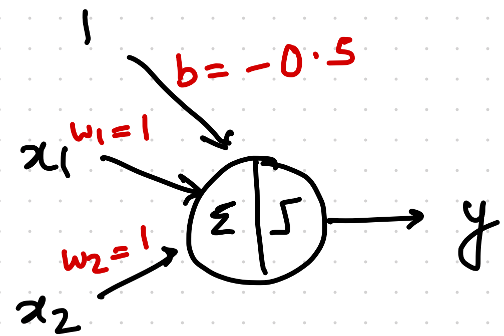
Q) FOR 2 I/PS x_1 & x_2 learn w_i's and b for BINARY OR

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

LEARNING BINARY GATES

Q) FOR 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

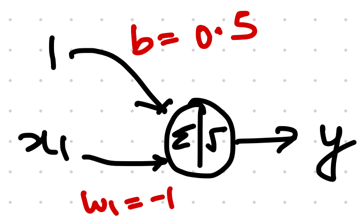
Q) FOR 1 I/ps x_1 learn w_i's and b for UNARY NOT

x_1	y
0	1
1	0

LEARNING BINARY GATES

Q) FOR 1 I/ps x_1 learn w_i'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_i's and b for NAND

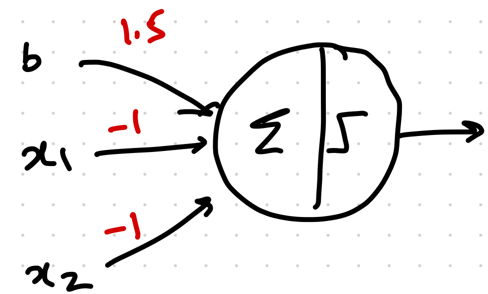
x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0

LEARNING BINARY GATES

Q) FOR 2 I/ps x_1 & x_2 learn w_i's and b for NAND

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

APPROACH #1

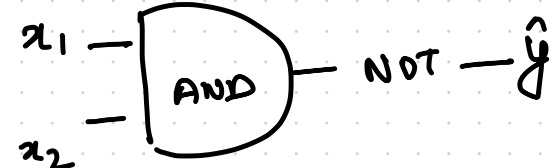


LEARNING BINARY GATES

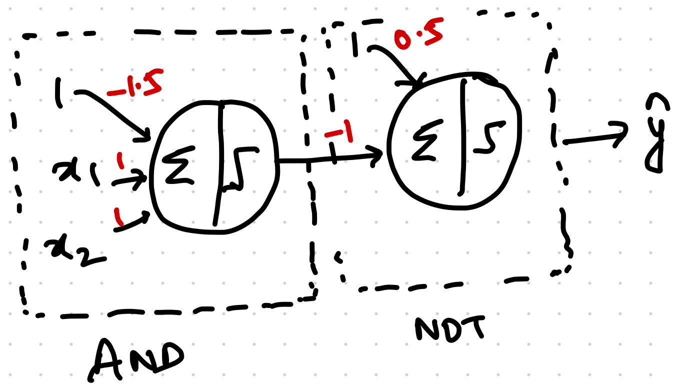
Q) FOR 2 I/ps x_1 & x_2 learn w_i's and b for NAND

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

APPROACH #2



or



PERCEPTRON LEARNING ALGORITHM

I/p: $X \hookrightarrow N \times D$; $y \hookrightarrow N \times 1$; $lr \hookrightarrow$ learning rate ; $it \hookrightarrow$ #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' → s.t. $x'[1:s] = X$; $x'[1] = \begin{bmatrix} 1 \\ \vdots \\ \vdots \end{bmatrix}$

PERCEPTRON LEARNING ALGORITHM

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

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

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

PERCEPTRON LEARNING ALGORITHM

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

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

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

S3 FOR I IN IT:

PERCEPTRON LEARNING ALGORITHM

I/p: X ; y ; lr ; it

$X \to N \times D$; $y \to N \times 1$; $lr \to$ learning rate ; $it \to$ #iterations

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

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

S3 FOR I IN IT:

S3.1 FOR d in D:

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

PERCEPTRON LEARNING ALGORITHM

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

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

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

S3 FOR I IN IT:

S3.1 FOR d in D:

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

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

PERCEPTRON LEARNING ALGORITHM

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

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

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

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

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

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

ERROR IN n^{th} sample

n^{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' → s.t. $X'[1:s] = X$; $X'[1] = \begin{bmatrix} 1 \\ \vdots \end{bmatrix}$

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 $ERR = y - \hat{y}$

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

ERROR IN n^{th} sample ← Analogous to Gradient

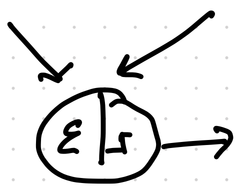
$X'[n]$ → n^{th} sample

PERCEPTRON LEARNING ALGORITHM

$$\boxed{\underline{S31.3} \quad w \leftarrow w + lr * ERR_n * x'[n]}$$

IMAGINE

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



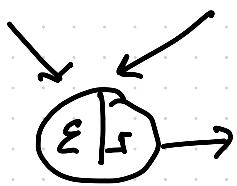
$$\hat{y} = 1$$

PERCEPTRON LEARNING ALGORITHM

$$\underline{S3.1.3} \qquad w \leftarrow w + lr * ERR_n * x'[n]$$

IMAGINE

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



$$\hat{y} = 1$$

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

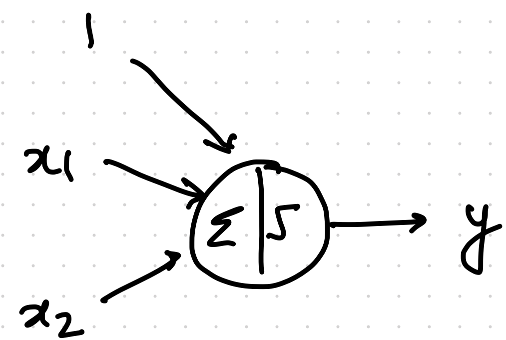
$$w \leftarrow \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 0.01 * -1 * \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\leftarrow \begin{bmatrix} .99 \\ 1 \\ 1 \end{bmatrix}$$

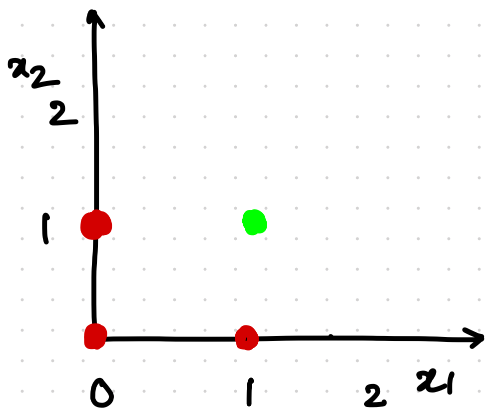
LEARNING BINARY GATES

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

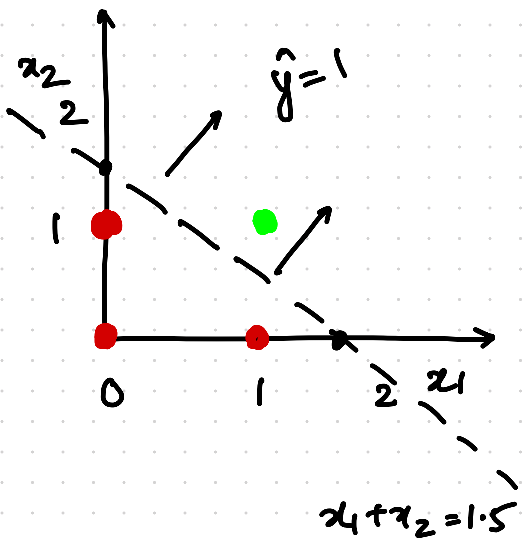
x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0



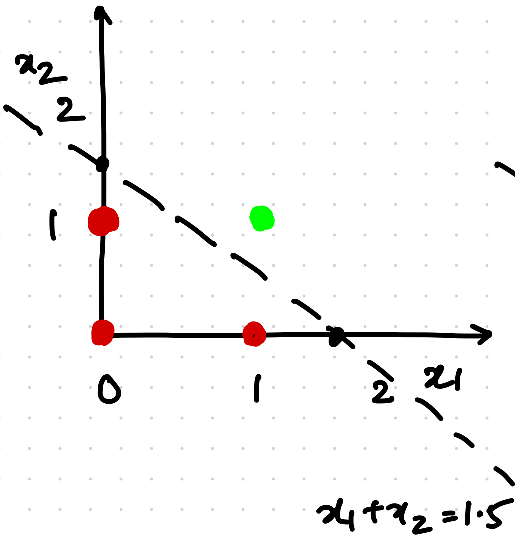
AND



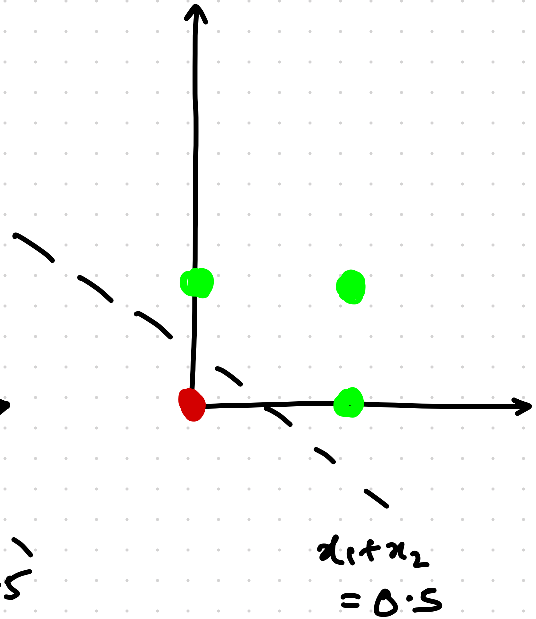
AND

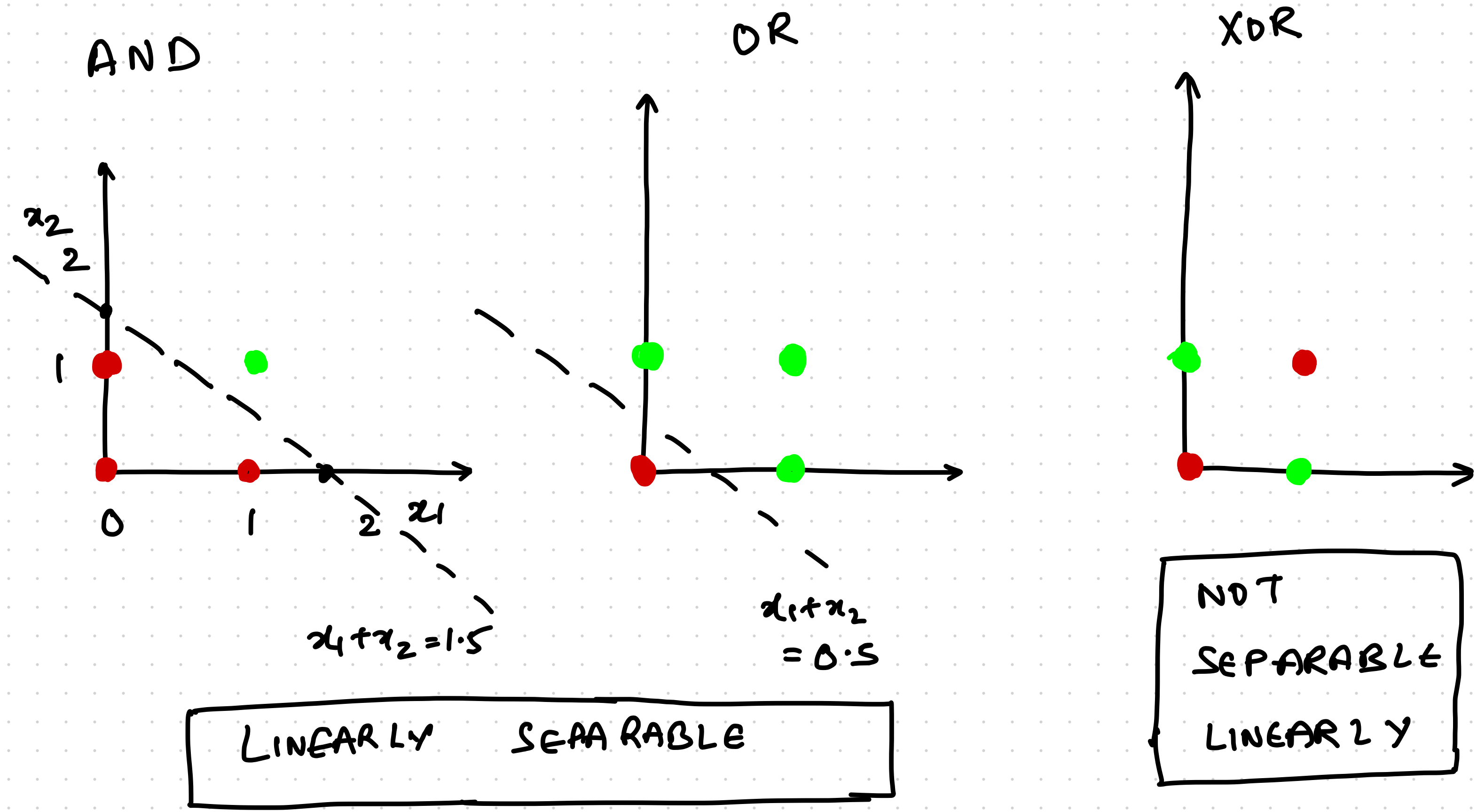


AND

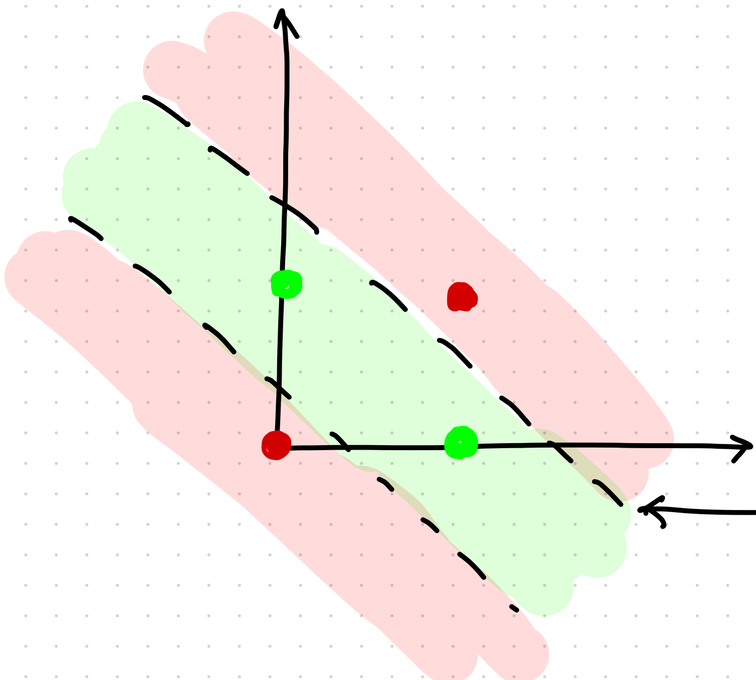


OR



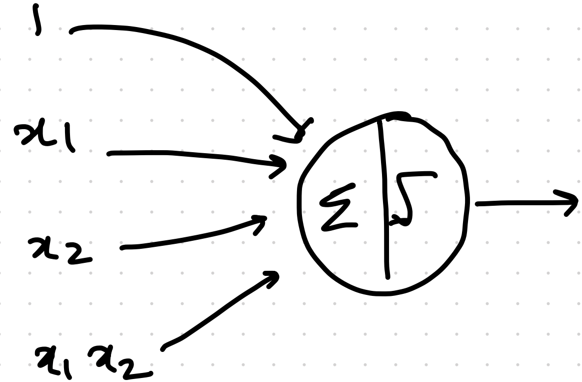

AND
OR
XOR
x_2
2
1
0
1
2
x_1
$x_1+x_2=1.5$
x_1+x_2
$=0.5$
LINEARLY SEPARABLE
NOT
SEPARABLE
LINEARLY

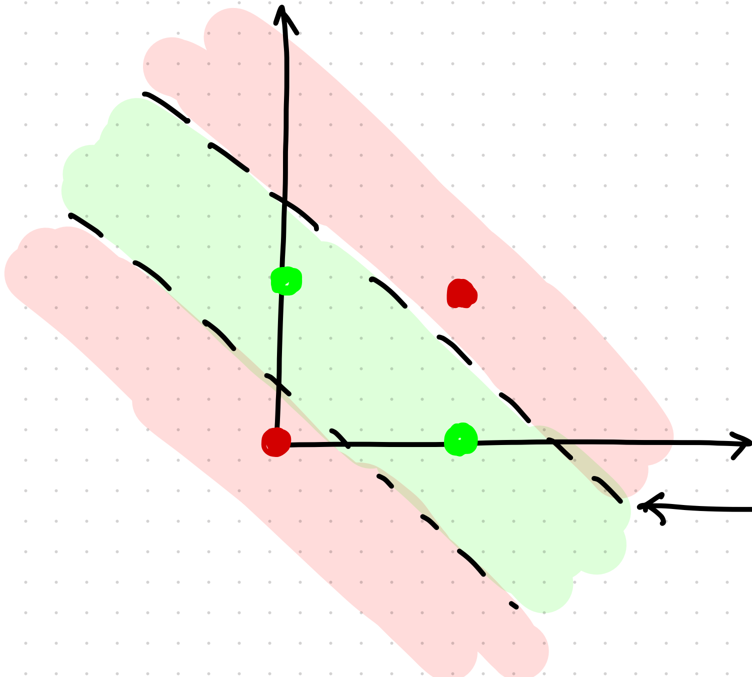
XOR CLASSIFICATION



Decision surface desired

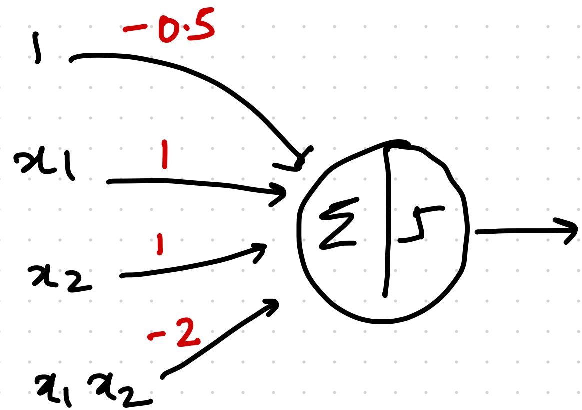
XOR CLASSIFICATION

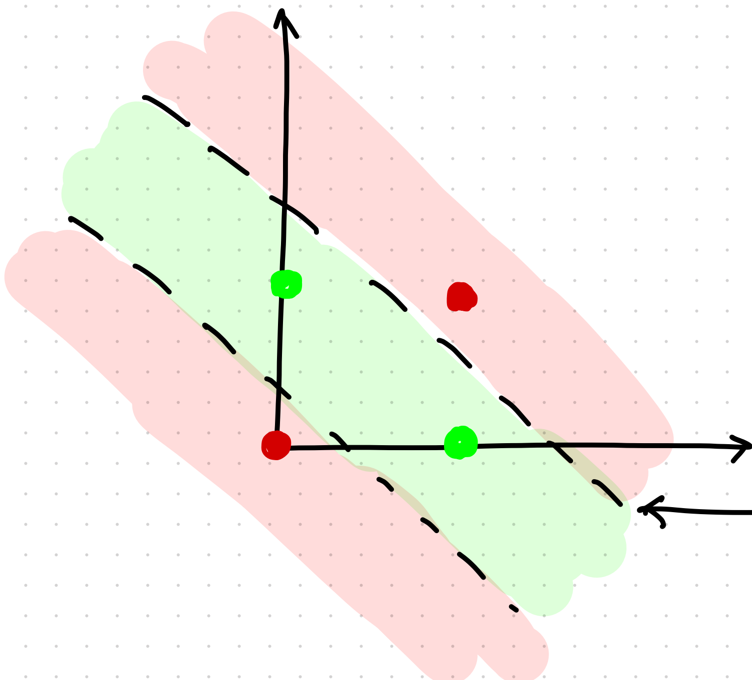




Decision surface desired

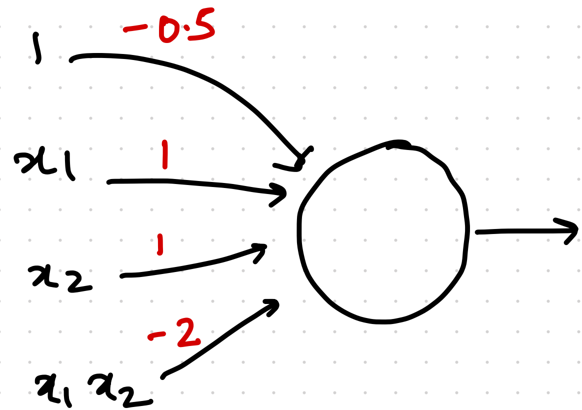
XOR CLASSIFICATION

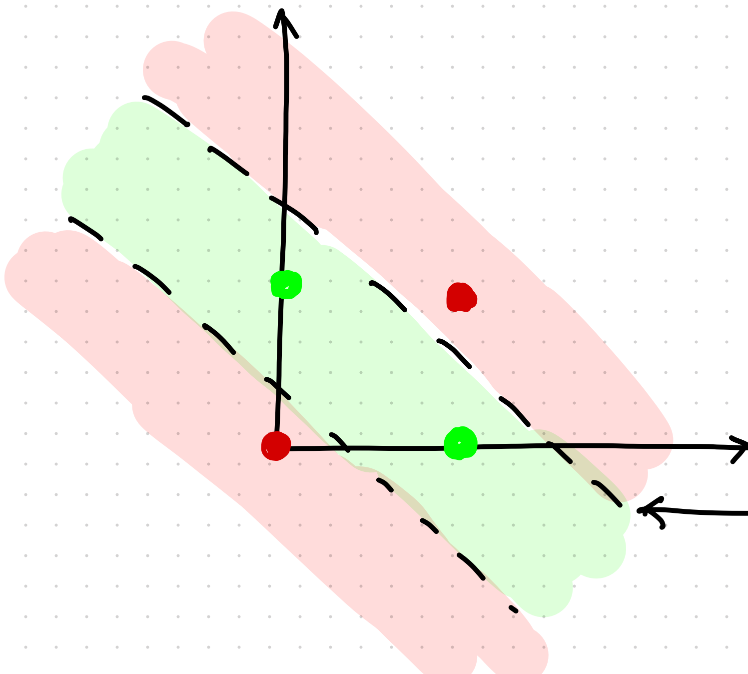




Decision surface desired

XOR CLASSIFICATION



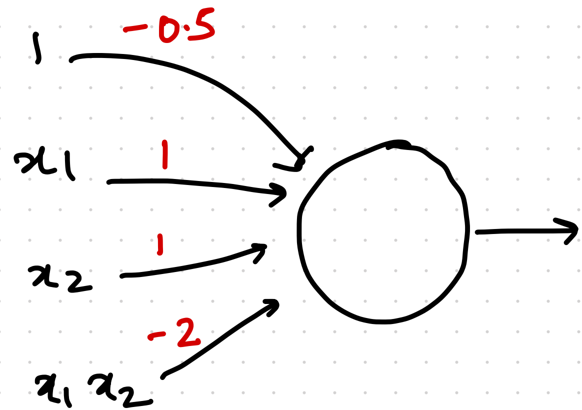


FOR $x_1 = 0$; $x_2 = 0$; we get

$\hat{y} = \{-0.5 \leq 0\}$ = RED CLASS

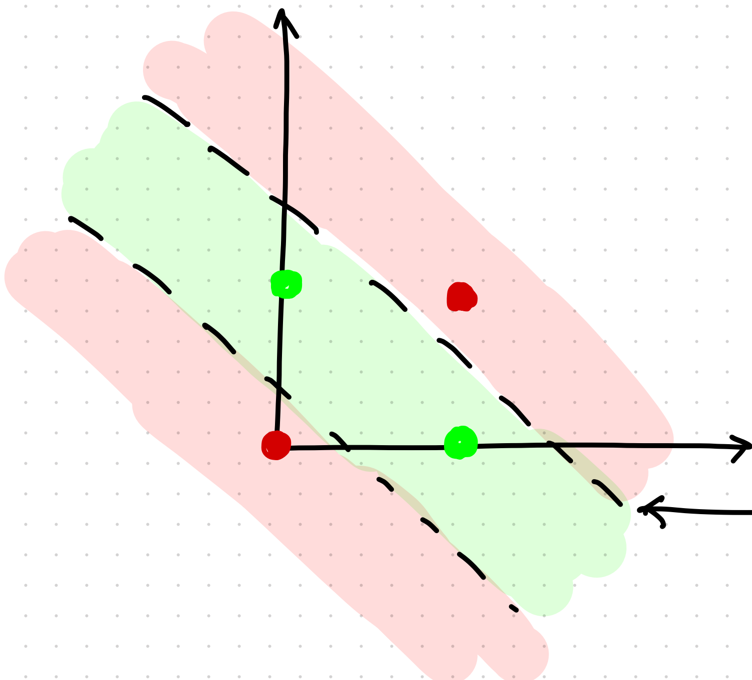
Decision surface desired

XOR CLASSIFICATION



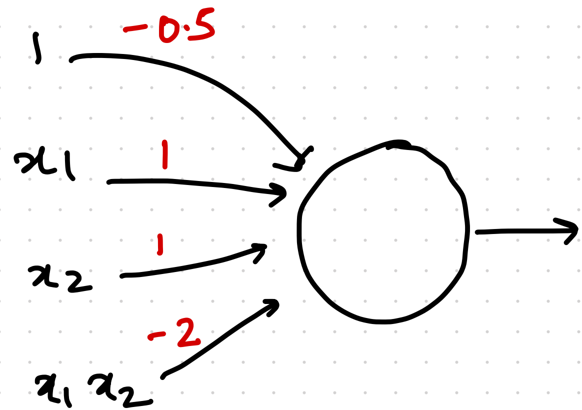
FOR $x_1 = 0 ; x_2 = 1$; we get

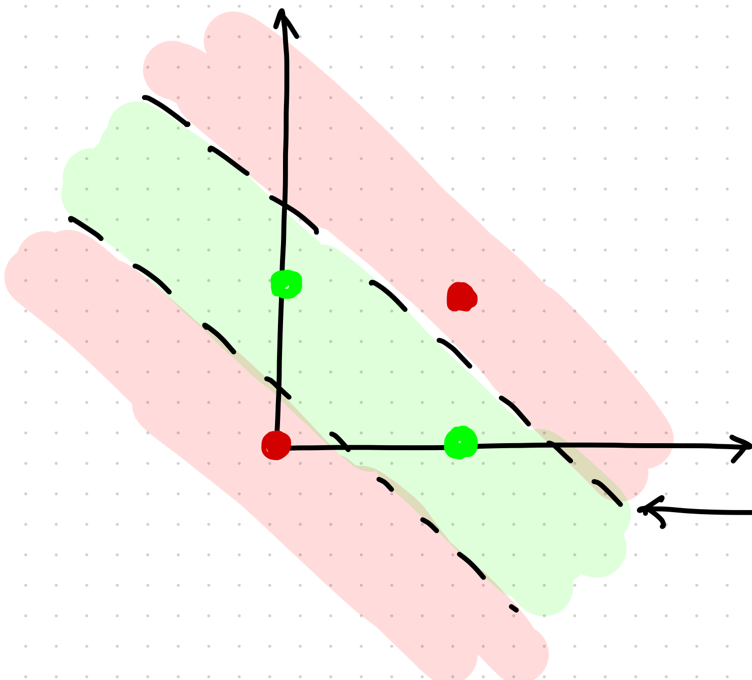
$\hat{y} = \{-0.5 + 1 \leq 0\} =$ GREEN CLASS



Decision surface desired

XOR CLASSIFICATION



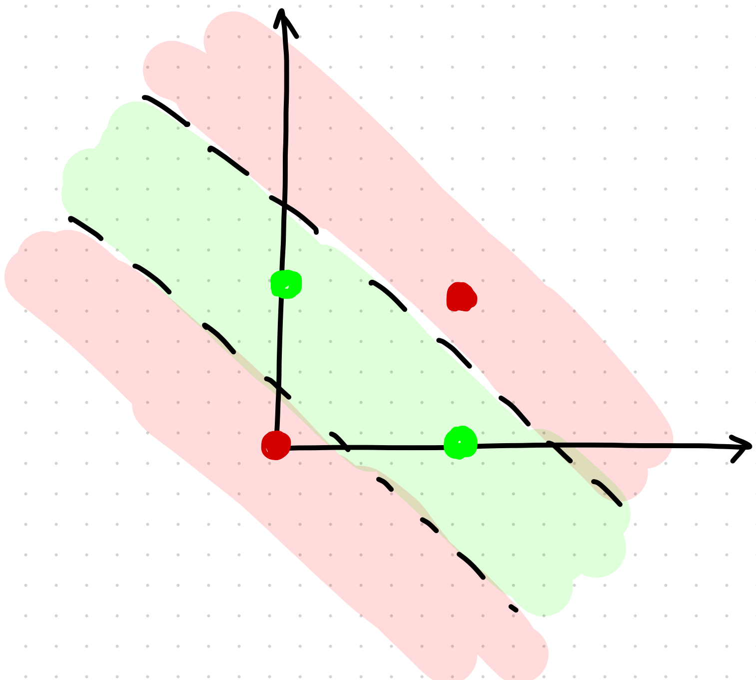


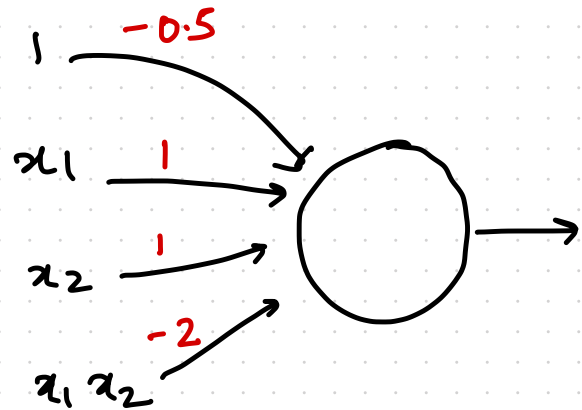
FOR $x_1 = 1$; $x_2 = 0$; we get

$$\hat{y} = \{0.5 \le 0\} = \text{GREEN CLASS}$$

Decision surface desired

XOR CLASSIFICATION



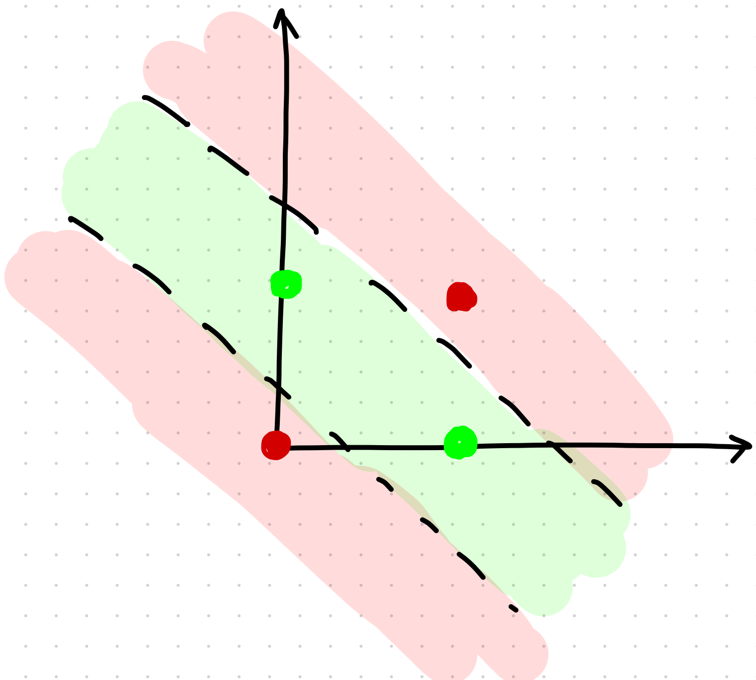


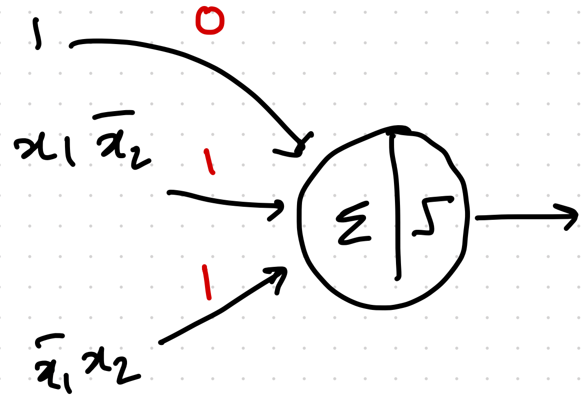
FOR $x_1 = 1$; $x_2 = 1$; we get

$\hat{y} = \{-0.5 \le 0\} =$ RED CLASS

CAN ADD NON-LINEARITY BY HAND-CRAFTING FEATURES !

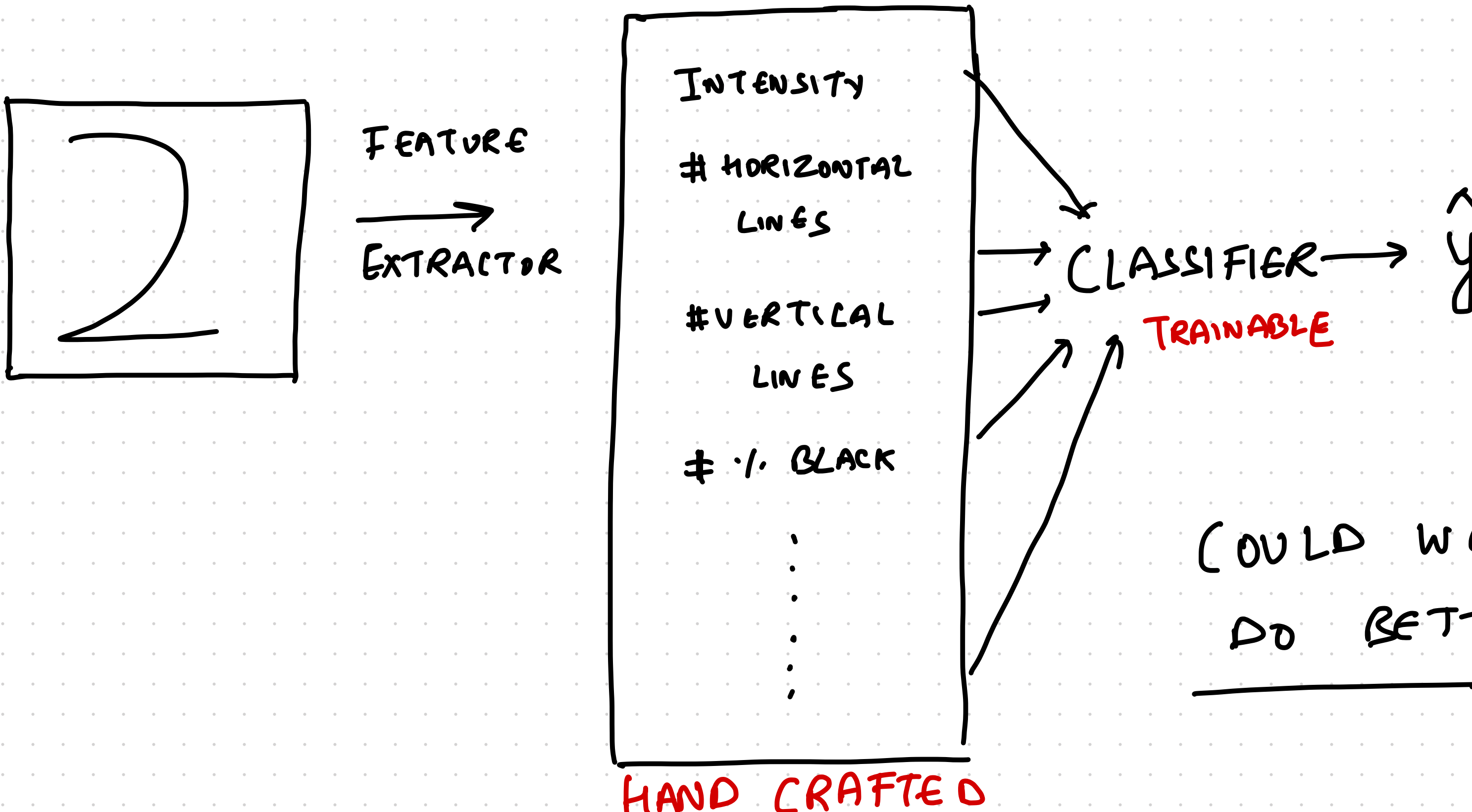
XOR CLASSIFICATION





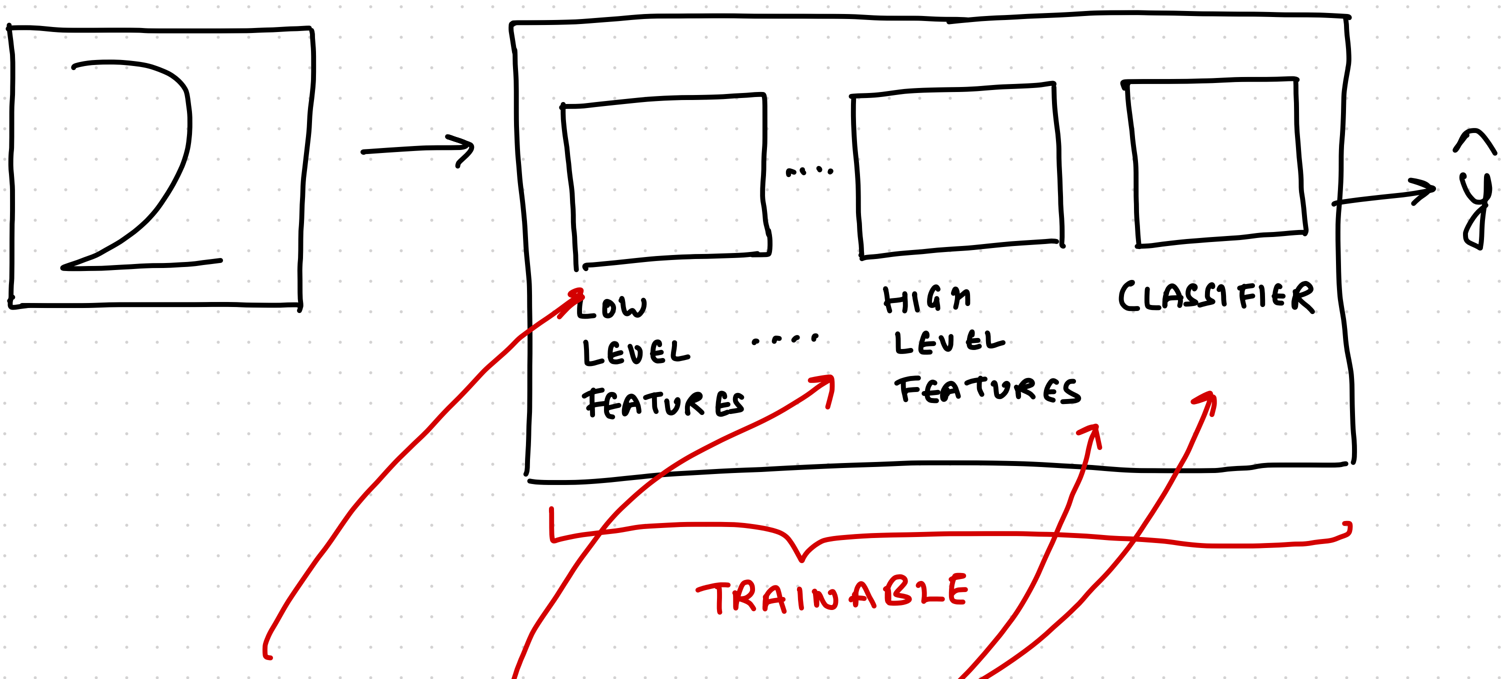
CAN ADD NON-LINEARITY BY HAND-CRAFTING FEATURES!

PARADIGM CHANGE



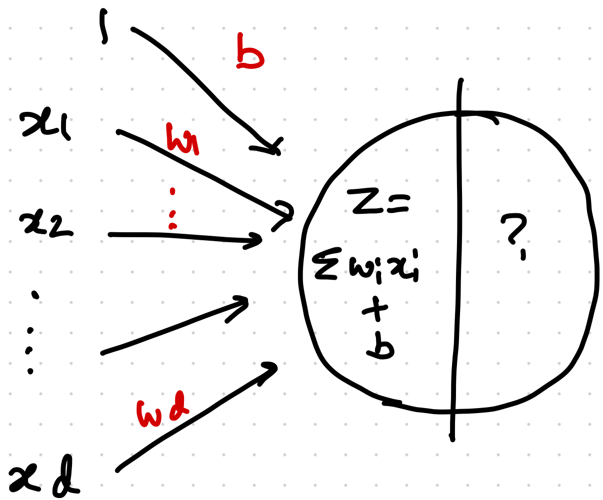
COULD WE DO BETTER?

PARADIGM CHANGE (NNS)



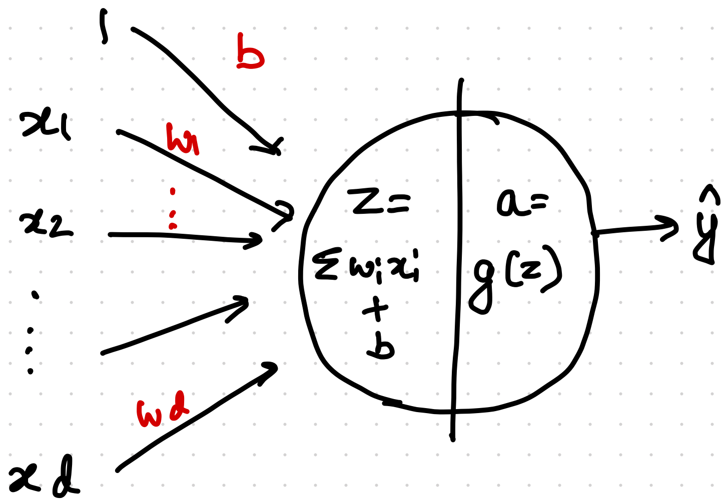
WE NEED NON-LINEARITY

BACK PROPAGATION SUPPORTED ACTIVATIONS



key idea: use activat^n similar to  but differentiable

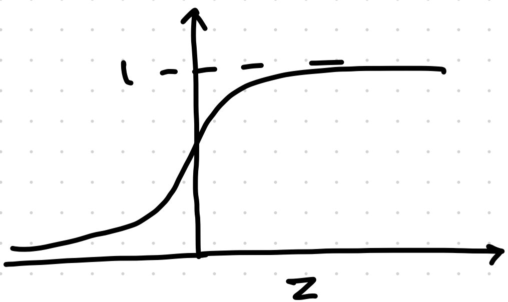
ADDING A NON-LINEARITY



$g(z)$: NON LINEAR TRANSFORMATION

ACTIVATION FUNCTIONS

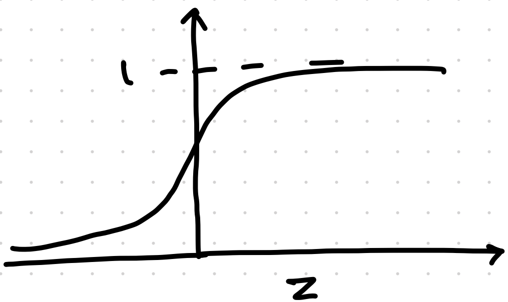
SIGMOID



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

ACTIVATION FUNCTIONS

SIGMOID



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

Q): If we have 1 neuron

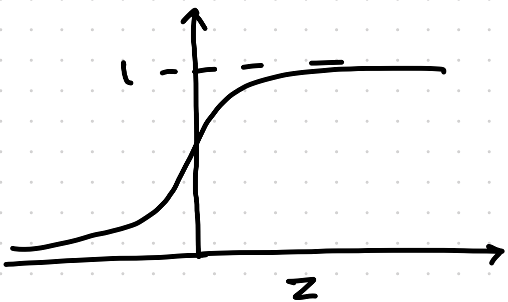
&

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

get?

ACTIVATION FUNCTIONS

SIGMOID



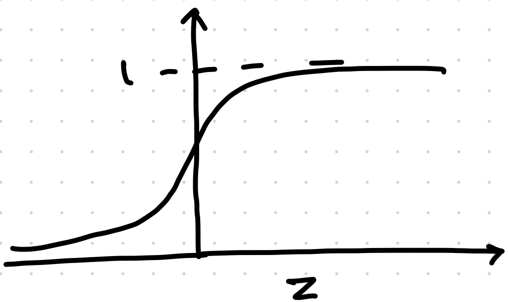
$$g(z) = \frac{1}{1+e^{-z}}$$

Q): If we have 1 neuron & $g(z) = \frac{1}{1+e^{-z}}$ what do we get?

Logistic Regression

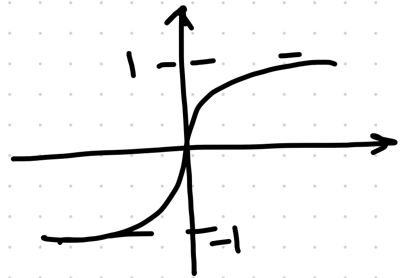
ACTIVATION FUNCTIONS

SIGMOID



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

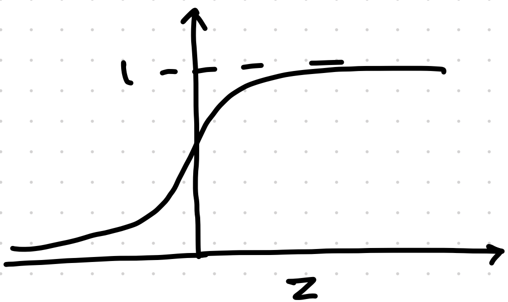
TANH



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

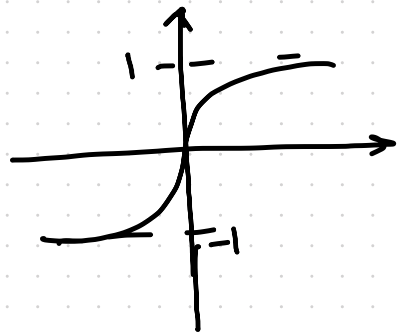
ACTIVATION FUNCTIONS

SIGMOID



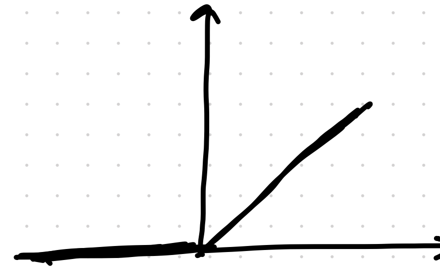
$$g(z) = \frac{1}{1+e^{-z}}$$

TANH



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

ReLU



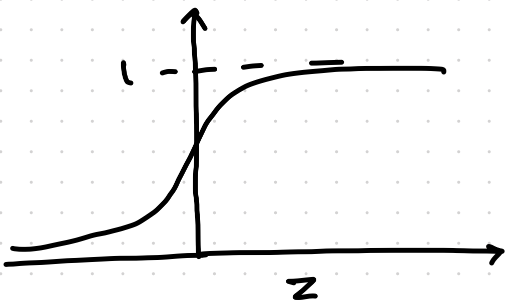
$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & o/w \end{cases}$$

or

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

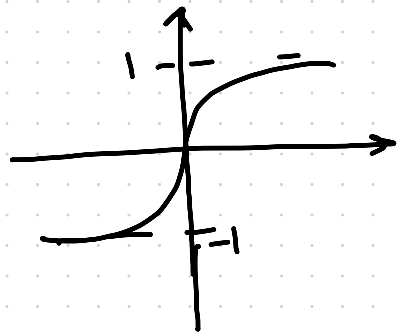
ACTIVATION FUNCTIONS

SIGMOID



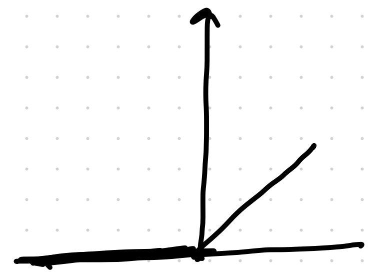
$$g(z) = \frac{1}{1+e^{-z}}$$

TANH



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

ReLU

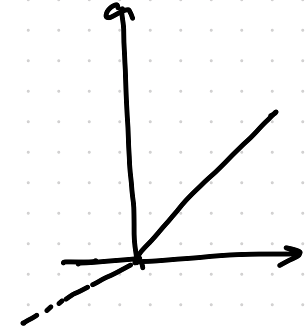


$$g(z) = \begin{cases} z; & z \geq 0 \\ 0; & o/w \end{cases}$$

or

$$g(z) = MAX(0, z)$$

Leaky ReLU

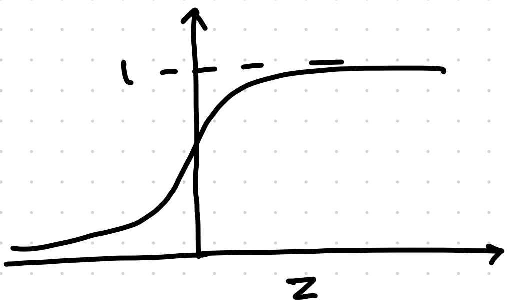


$$g(z) = MAX(\alpha z, z)$$

$$\alpha \to 0$$

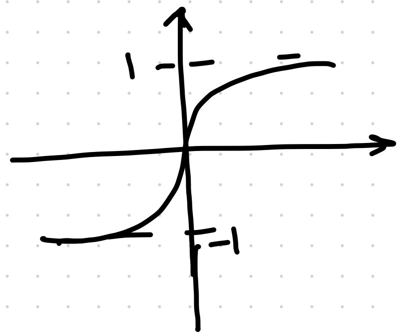
ACTIVATION FUNCTIONS

SIGMOID



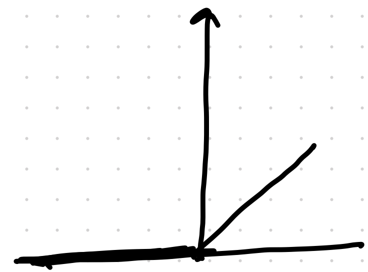
USEFUL FOR PROBABILISTIC ESTIMAES ∵ B/W 0 & 1

TANH



USEFUL IF DATA TRANSFORMED WITH MEAN 0

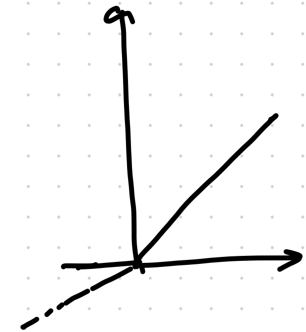
ReLU



GAME CHANGER (Default)

Good learning for $|z|$ = high.

Leaky ReLU



Similar to ReLU

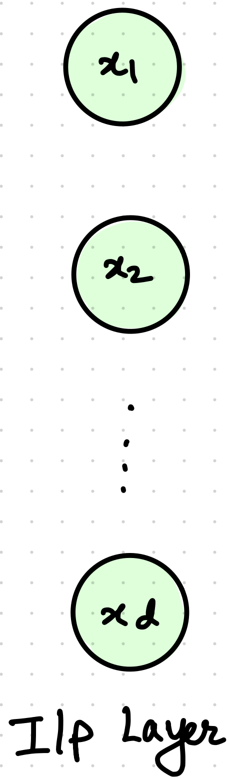
Learns for $z < 0$ also

DESIRABLE ATTRIBUTES OF ACTIVATION FUNCTIONS

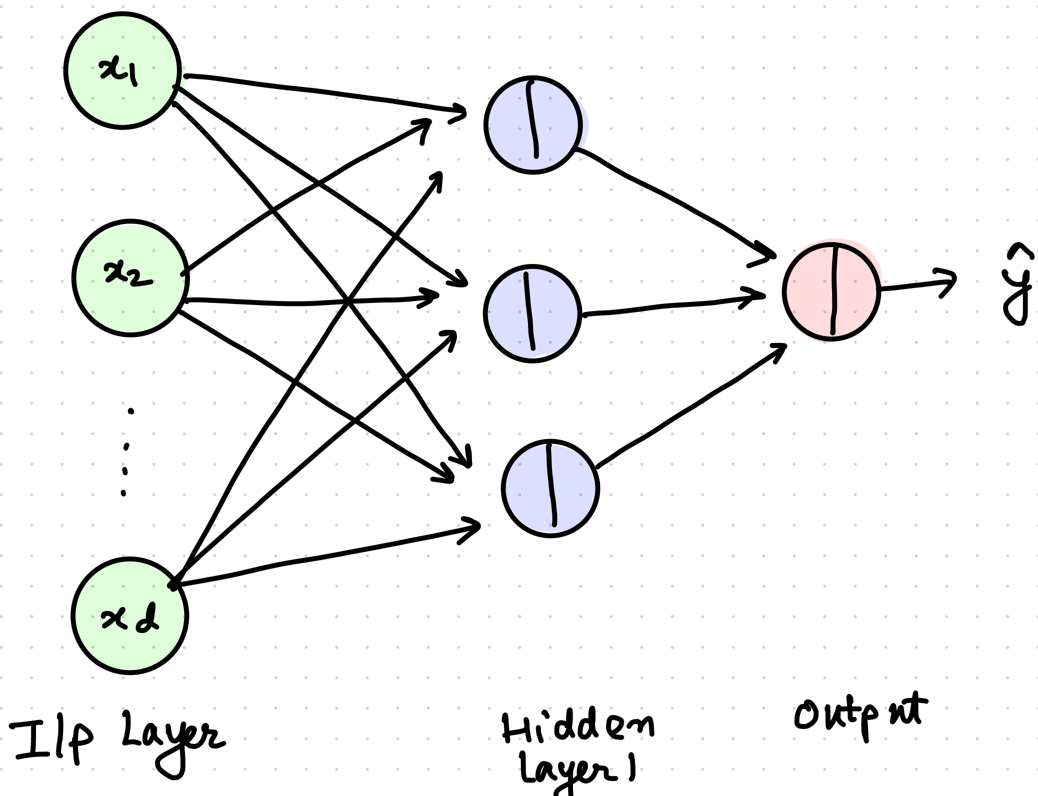
1) NON-LINEAR

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

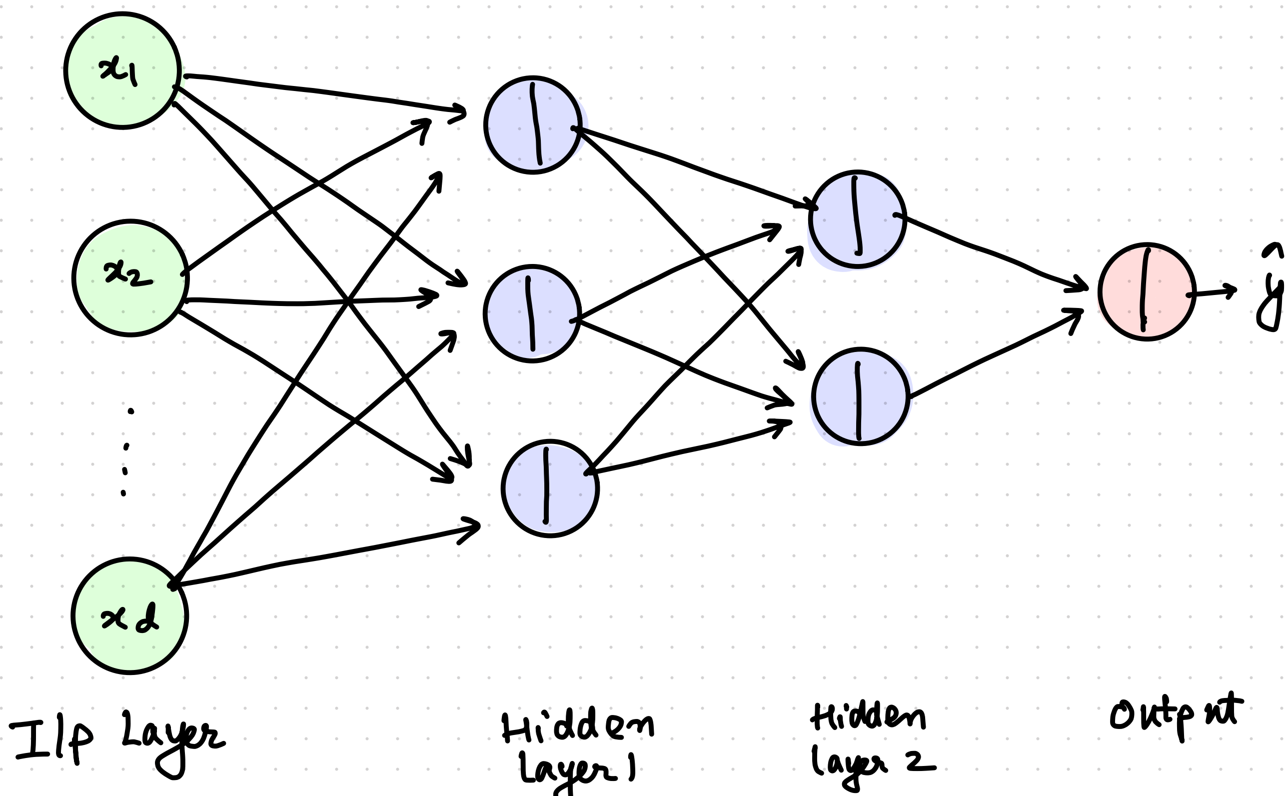
1 LAYER PERCEPTRON (NN)



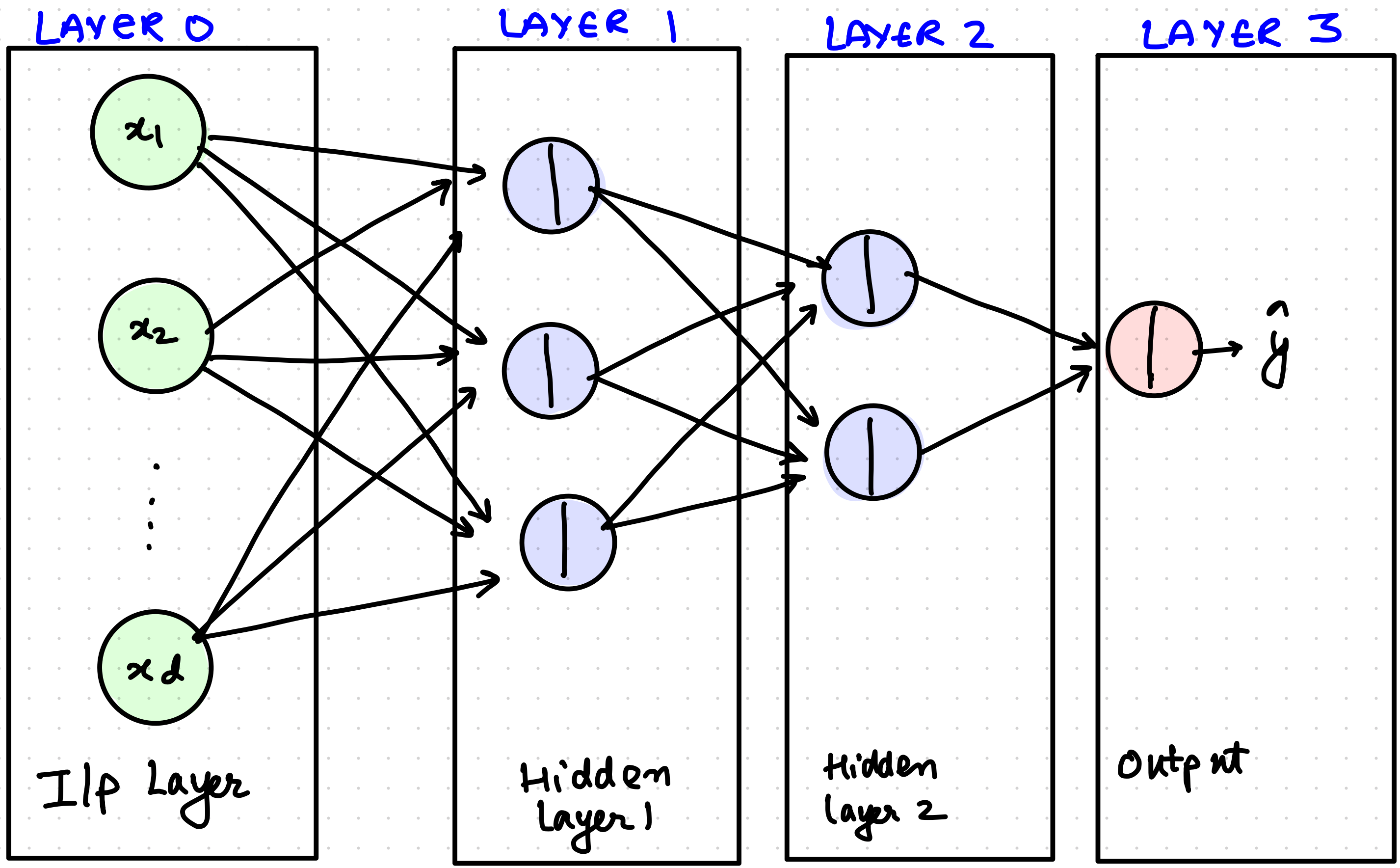
1- LAYER PERCEPTRON (NN)

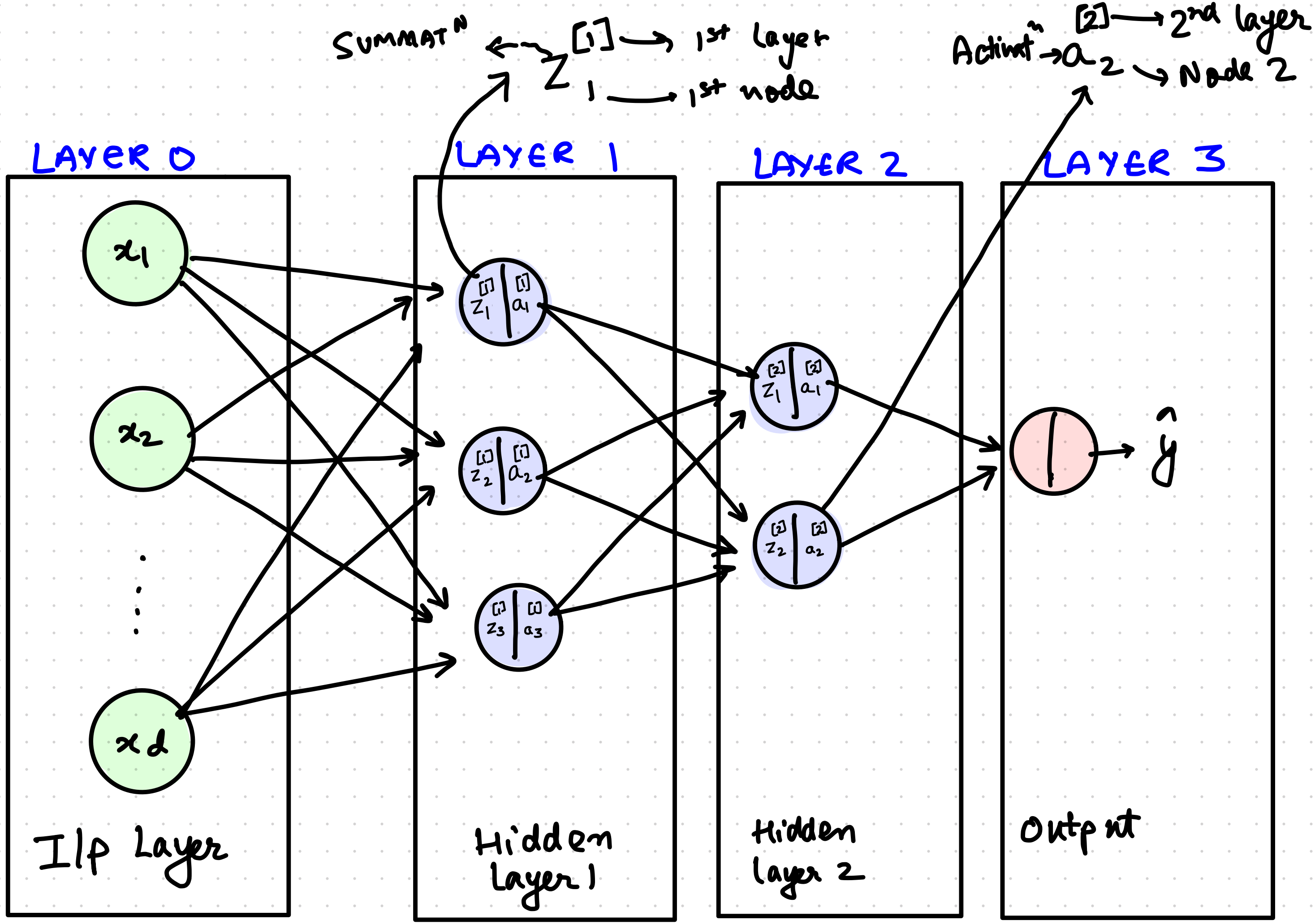


MULTI - LAYER PERCEPTRON

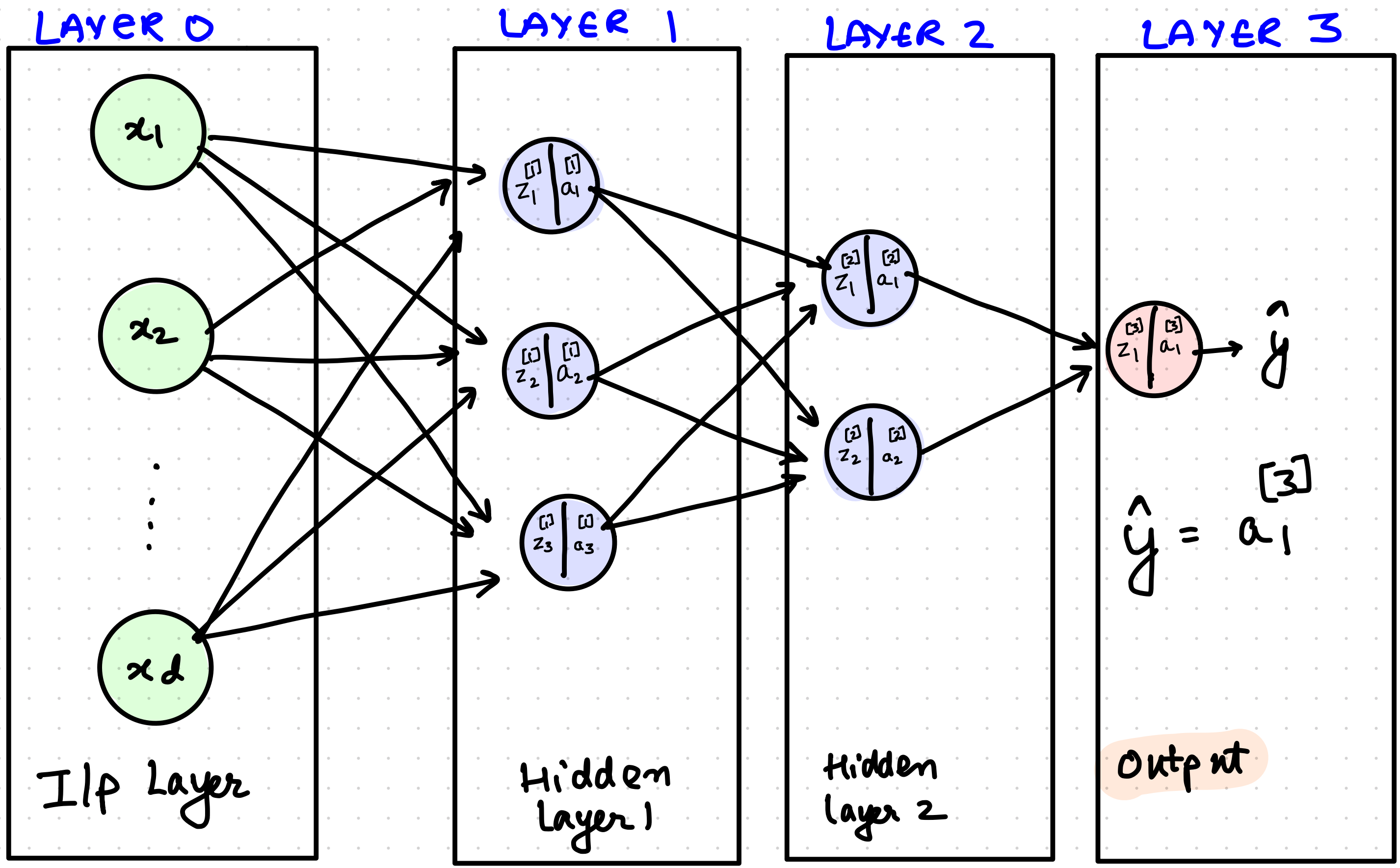


MULTI - LAYER PERCEPTRON

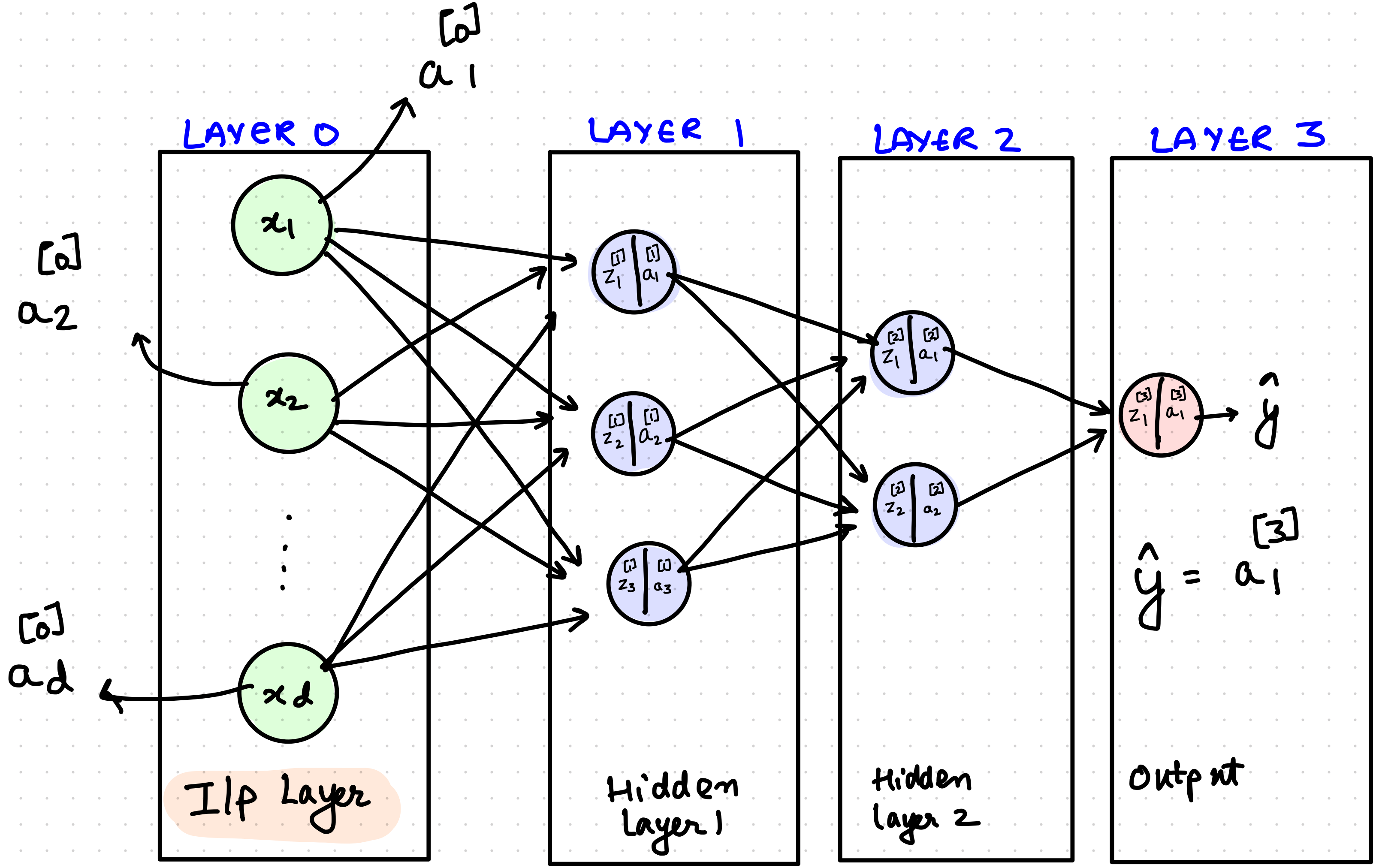


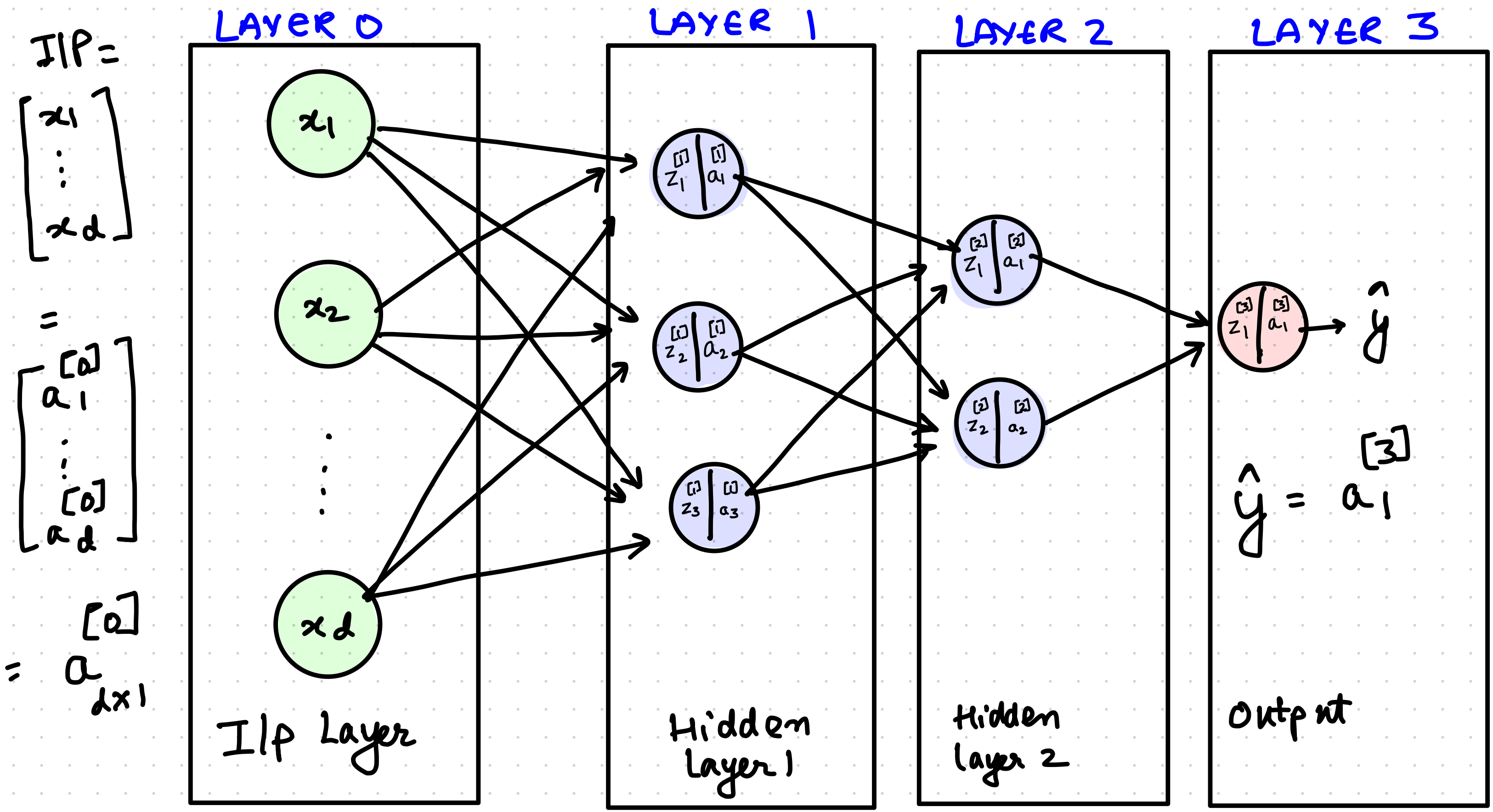


SUMMATN
[1] → 1st Layer
1 → 1st node
Activatn → a
[2] → 2nd layer
2 → Node 2
LAYER 0
LAYER 1
LAYER 2
LAYER 3
x1
x2
xd
I/p Layer
Hidden Layer 1
Hidden layer 2
Output
ŷ



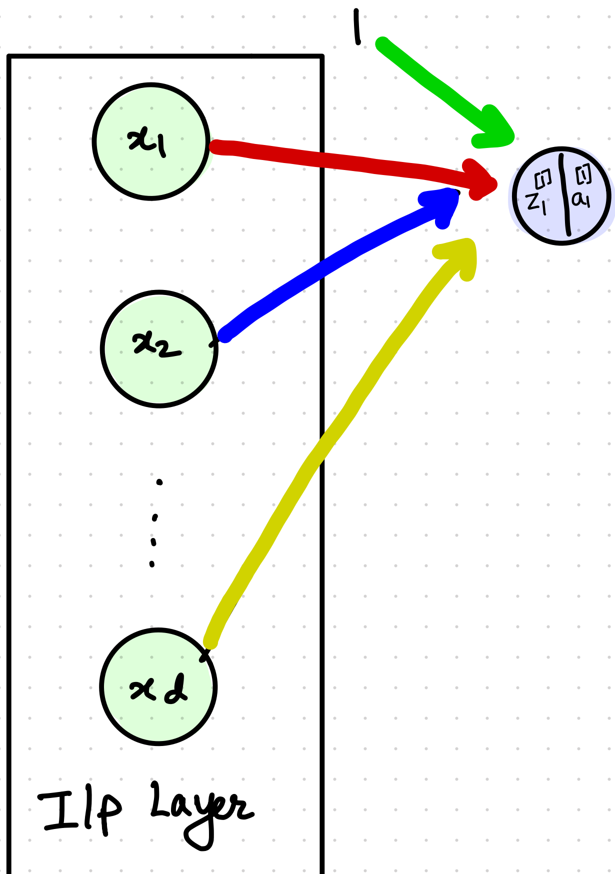
LAYER 0
LAYER 1
LAYER 2
LAYER 3
x_1
x_2
x_d
$z_1^{[1]}$ $a_1^{[1]}$
$z_2^{[1]}$ $a_2^{[1]}$
$z_3^{[1]}$ $a_3^{[1]}$
$z_1^{[2]}$ $a_1^{[2]}$
$z_2^{[2]}$ $a_2^{[2]}$
$z_1^{[3]}$ $a_1^{[3]}$
$\hat{y}$
$\hat{y} = a_1^{[3]}$
I/p Layer
Hidden layer 1
Hidden layer 2
Output


LAYER 0
LAYER 1
LAYER 2
LAYER 3
$a_1^{[0]}$
$a_2^{[0]}$
$a_d^{[0]}$
x_1
x_2
x_d
$z_1^{[1]}$ $a_1^{[1]}$
$z_2^{[1]}$ $a_2^{[1]}$
$z_3^{[1]}$ $a_3^{[1]}$
$z_1^{[2]}$ $a_1^{[2]}$
$z_2^{[2]}$ $a_2^{[2]}$
$z_1^{[3]}$ $a_1^{[3]}$
$\hat{y}$
$\hat{y} = a_1^{[3]}$
I/p Layer
Hidden Layer 1
Hidden layer 2
Output



LAYER 0
LAYER 1
LAYER 2
LAYER 3
I/P =
$\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} a_1^{[0]} \\ \vdots \\ a_d^{[0]} \end{bmatrix} = a^{[0]}_{d\times 1}$
x_1
x_2
x_d
$z_1^{[1]}$ $a_1^{[1]}$
$z_2^{[1]}$ $a_2^{[1]}$
$z_3^{[1]}$ $a_3^{[1]}$
$z_1^{[2]}$ $a_1^{[2]}$
$z_2^{[2]}$ $a_2^{[2]}$
$z_1^{[3]}$ $a_1^{[3]}$
$\hat{y}$
$\hat{y} = a_1^{[3]}$
I/P Layer
Hidden Layer 1
Hidden layer 2
Output

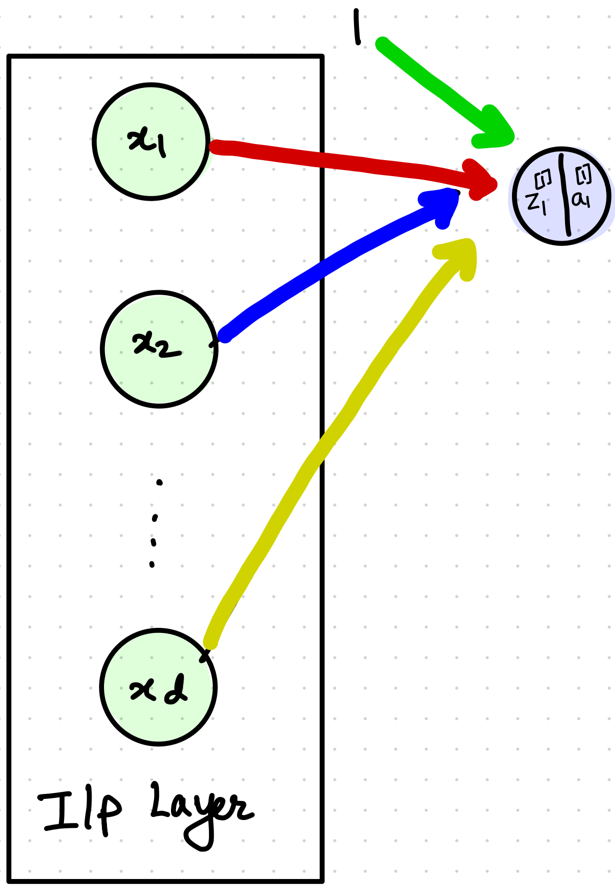
CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



$$Z_1^{[1]} = 1 * b_1^{[1]} + x_1 * w_{1,1}^{[1]} + x_2 * w_{1,2}^{[1]} + \dots + x_d * w_{1,d}^{[1]}$$

bias layer 1
Node 1

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



bias (layer 1, Node 1)

$$Z_1^{[1]} = 1 * b_1^{[1]} + x_1 * w_{1,1}^{[1]} + x_2 * w_{1,2}^{[1]} + \dots + x_d * w_{1,d}^{[1]}$$

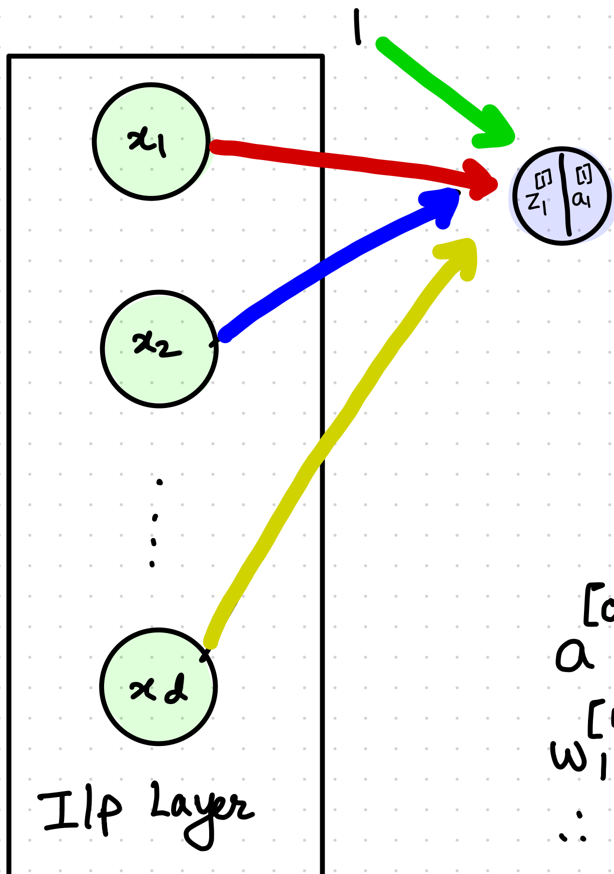
$$w_{a,b}^{[l]}$$

[l] ← l^{th} layer

a → a^{th} node in l^{th} layer

b → b^{th} component of prev. layer activat^n

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)



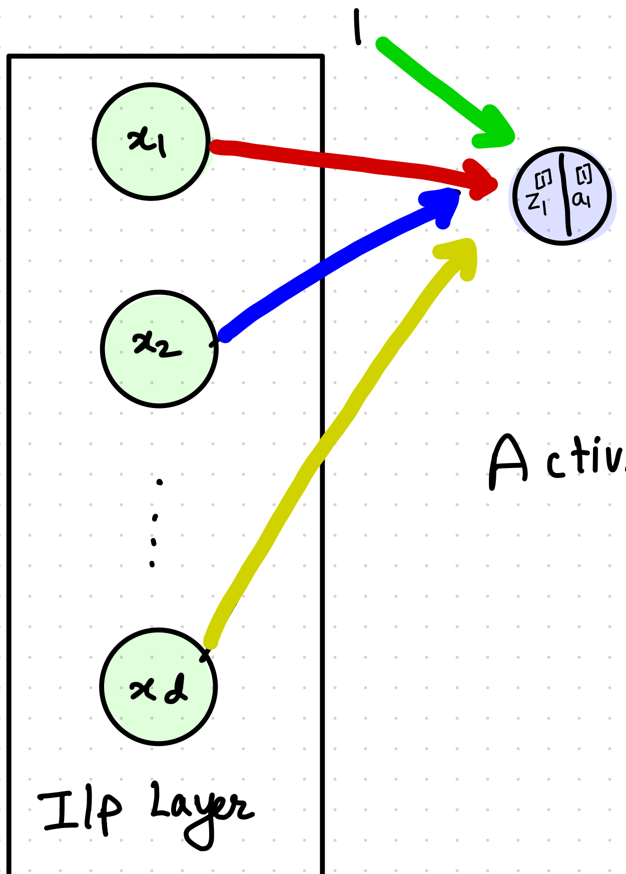
$$Z_1^{[1]} = 1 * b_1^{[1]} + a_1^{[0]} * w_{1,1}^{[1]} + a_2^{[0]} * w_{1,2}^{[1]} + \cdots + a_d^{[0]} * w_{1,d}^{[1]}$$

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

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

$$\therefore Z_1^{[1]} = w_1^{[1]^T} a^{[0]} + b_1^{[1]}$$

CONSIDER SINGLE NEURON (LAYER 1, NODE 1)

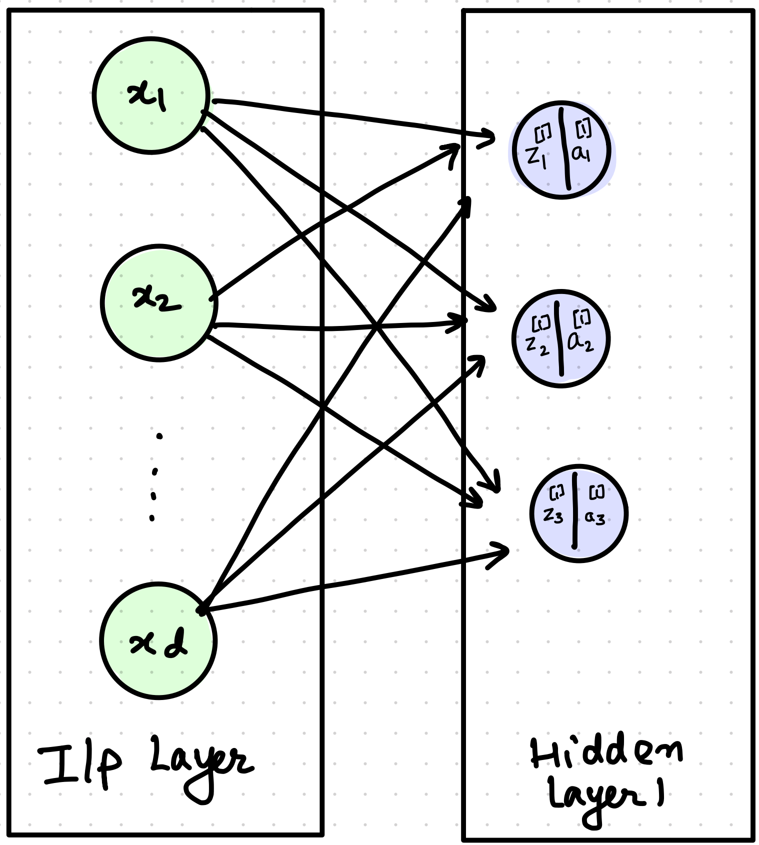


$$z_1^{[1]} = w_1^{[1]T} a^{[0]} + b_1^{[1]}$$

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

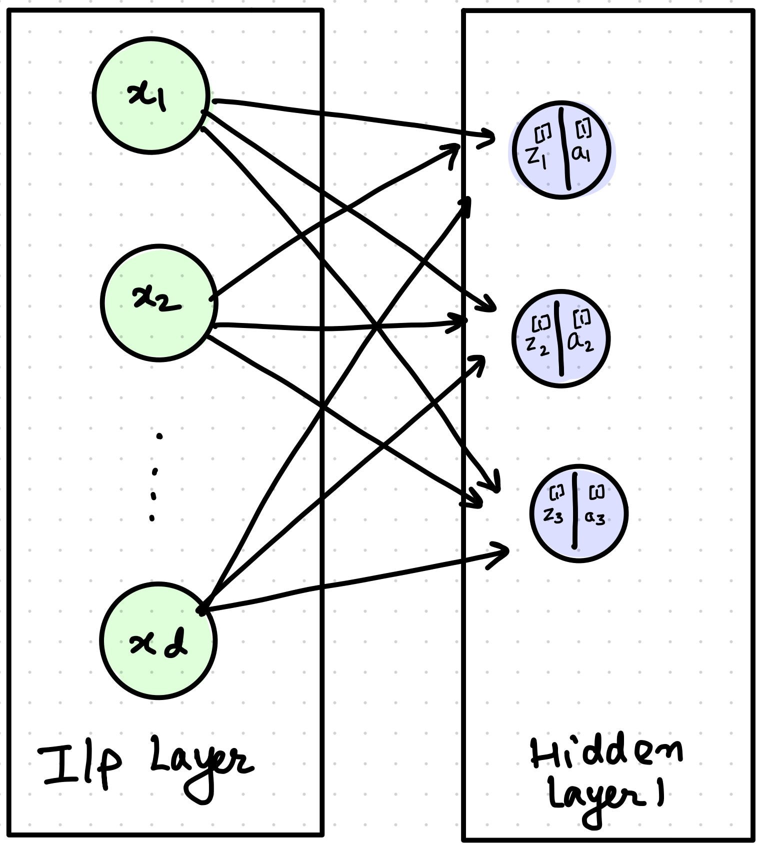
$$a_1^{[1]} \in R$$

FORWARD PROPAGATION



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

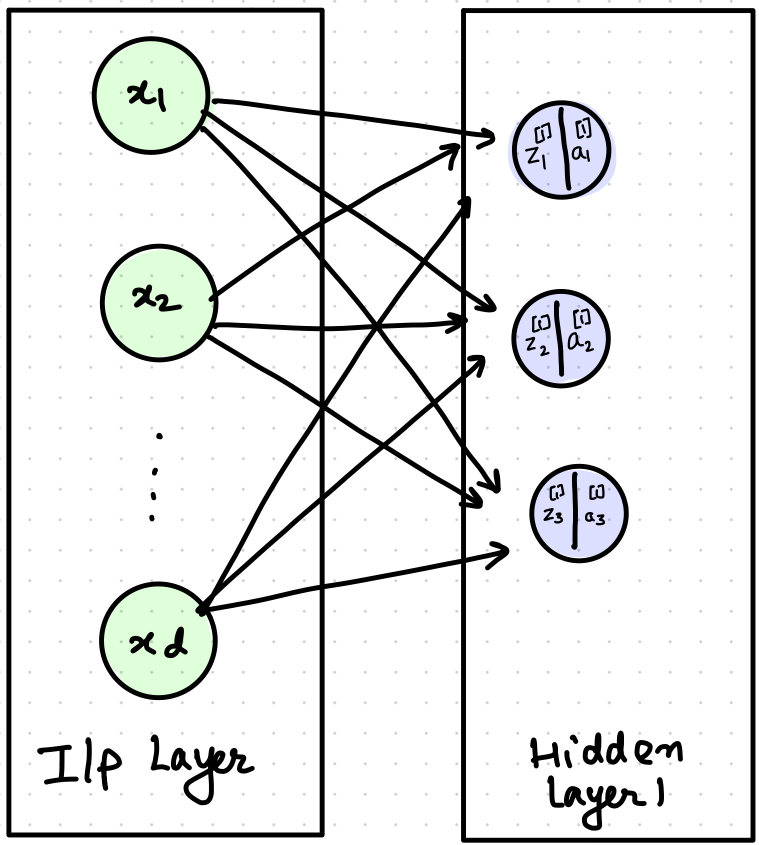
FORWARD PROPAGATION



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

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

FORWARD PROPAGATION

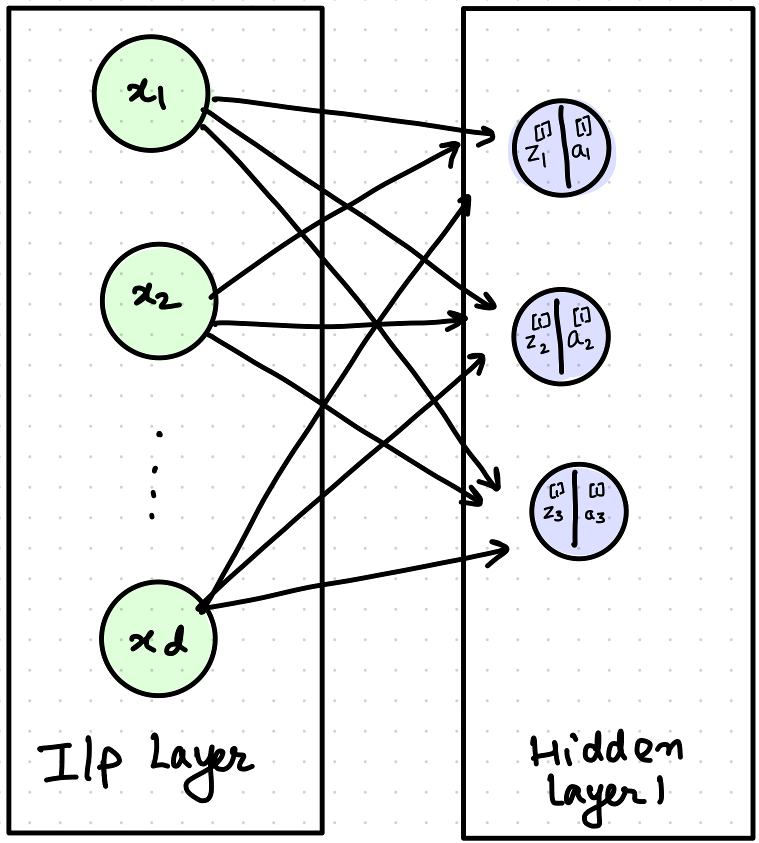


$$a_1^{[1]} = g\left({w_1^{[1]}}^T a^{[0]} + b_1^{[1]}\right)$$

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

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

FORWARD PROPAGATION (VECTORISATION)

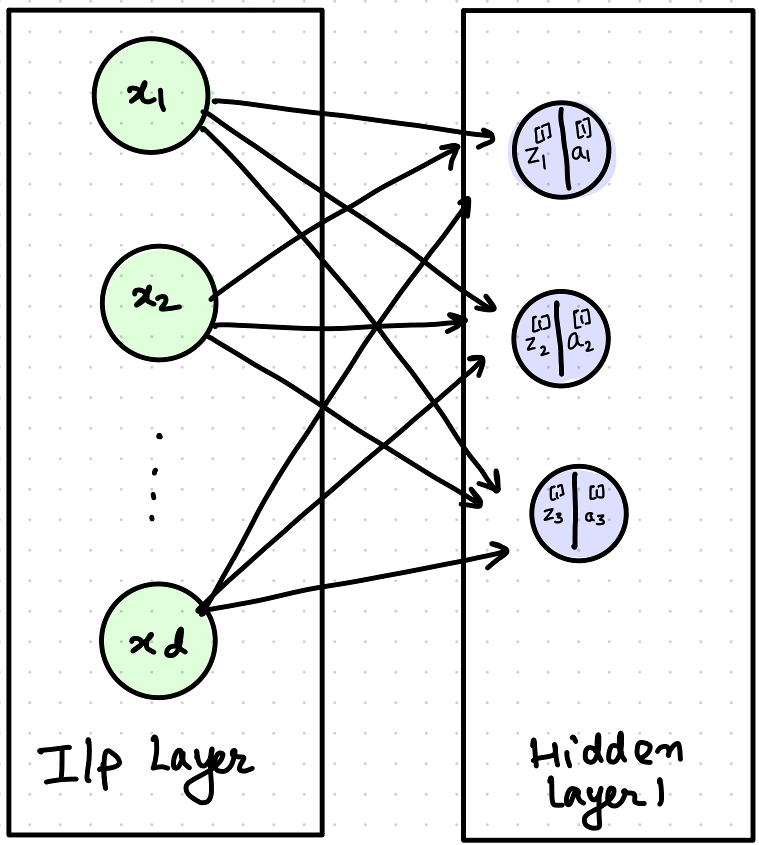


$$z_1^{[1]} = w_1^{[1]^T} a^{[0]} + b_1^{[1]}$$

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

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

FORWARD PROPAGATION (VECTORISATION)

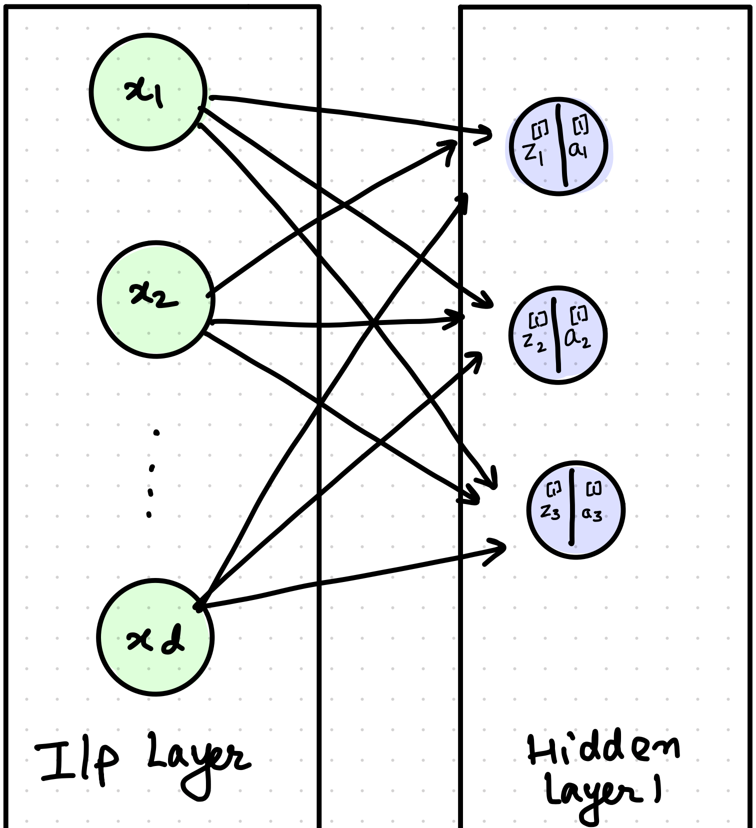


$$\underset{1\times1}{z_1^{[1]}} = \underset{1\times3}{{w_1^{[1]}}^T} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_1^{[1]}}$$

$$\underset{1\times1}{z_2^{[1]}} = \underset{1\times3}{{w_2^{[1]}}^T} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_2^{[1]}}$$

$$\underset{1\times1}{z_3^{[1]}} = \underset{1\times3}{{w_3^{[1]}}^T} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_3^{[1]}}$$

FORWARD PROPAGATION (VECTORISATION)



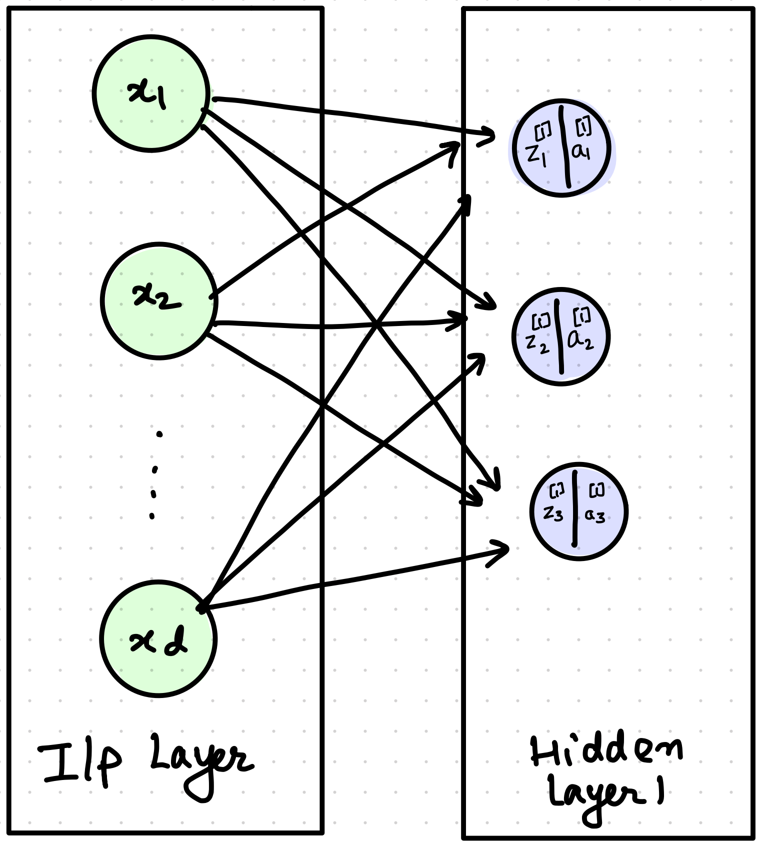
$$\underset{1\times1}{z_1^{[1]}} = \underset{1\times3}{w_1^{[1]T}} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_1^{[1]}}$$

$$\underset{1\times1}{z_2^{[1]}} = \underset{1\times3}{w_2^{[1]T}} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_2^{[1]}}$$

$$\underset{1\times1}{z_3^{[1]}} = \underset{1\times3}{w_3^{[1]T}} \underset{3\times1}{a^{[0]}} + \underset{1\times1}{b_3^{[1]}}$$

$$\underset{3\times1}{z^{[1]}} = \underset{3\times3}{\begin{bmatrix} — & w_1^{[1]T} & — \\ — & w_2^{[1]T} & — \\ — & w_3^{[1]T} & — \end{bmatrix}} \underset{3\times1}{a^{[0]}} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

FORWARD PROPAGATION (VECTORISATION)

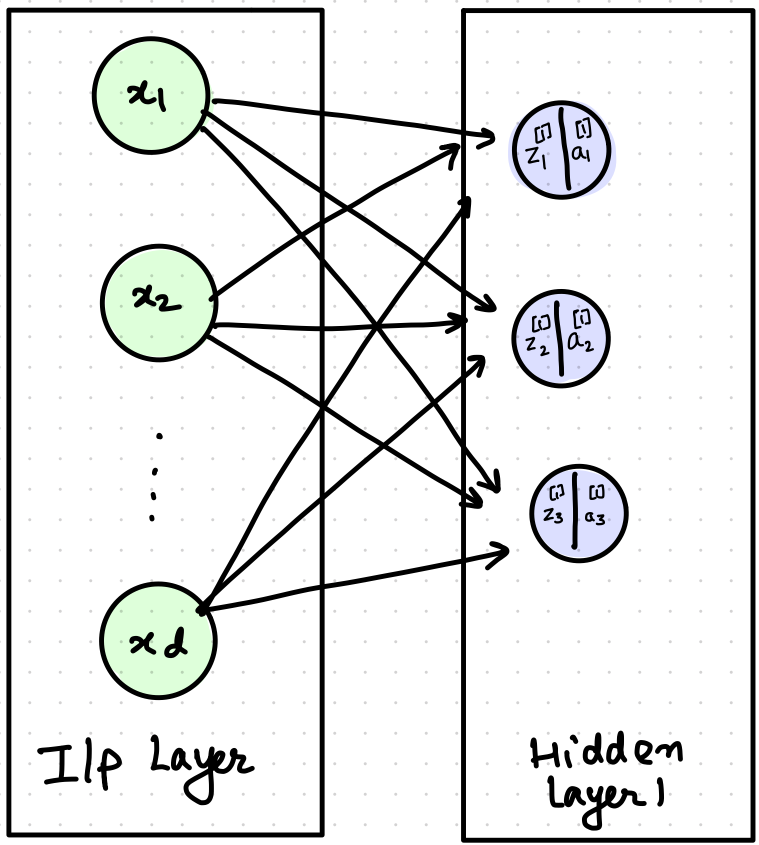


$$\underset{3\times1}{z^{[1]}} = \underset{3\times3}{\begin{bmatrix} — & w_1^{[1]T} & — \\ — & w_2^{[1]T} & — \\ — & w_3^{[1]T} & — \end{bmatrix}} \underset{3\times1}{a^{[0]}} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

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

capitals for matrices

FORWARD PROPAGATION (VECTORISATION)



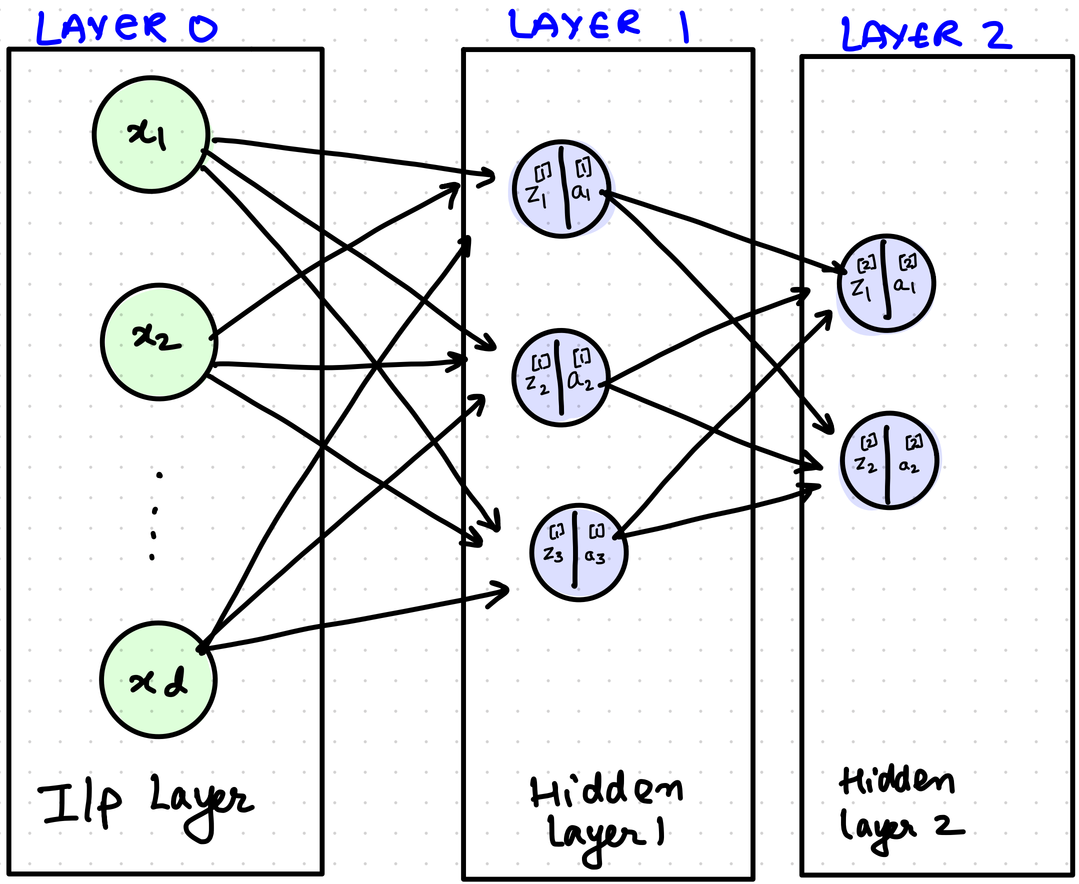
$$\underset{3\times1}{z^{[1]}} = \underset{3\times3}{\begin{bmatrix} — w_1^{[1]T} — \\ — w_2^{[1]T} \\ — w_3^{[1]T} \end{bmatrix}} \underset{3\times1}{a^{[0]}} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

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

capitals for matrices

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

FORWARD PROPAGATION

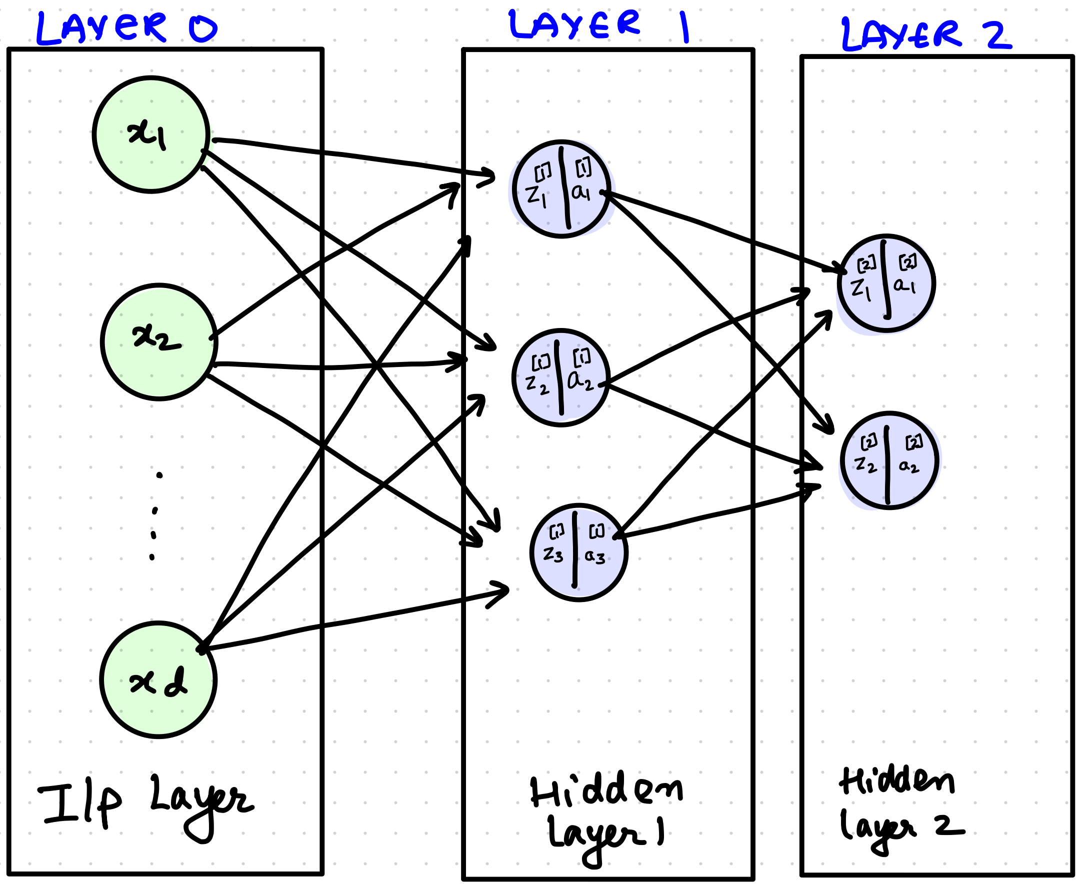


$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

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

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

FORWARD PROPAGATION



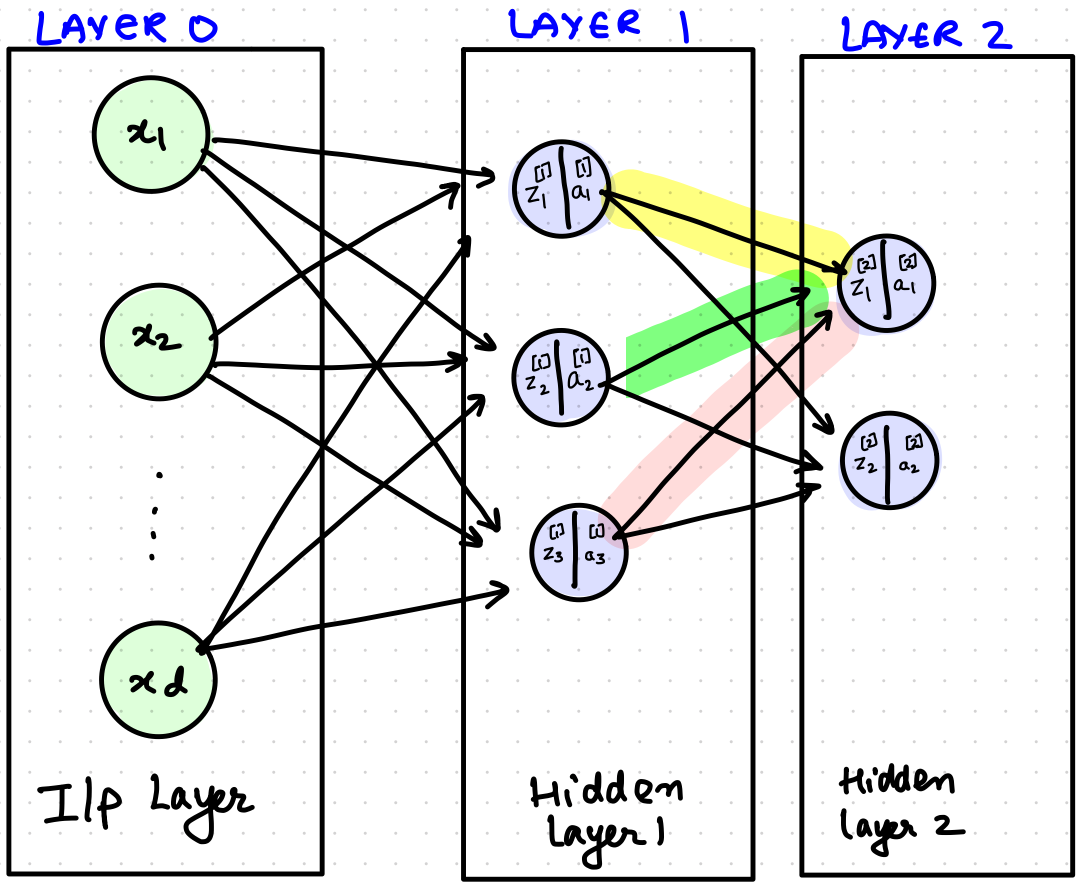
$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

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

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

$$W^{[2]} = \begin{bmatrix} - & w_1^{[2]} & - \\ - & w_2^{[2]} & - \end{bmatrix}$$

FORWARD PROPAGATION



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

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

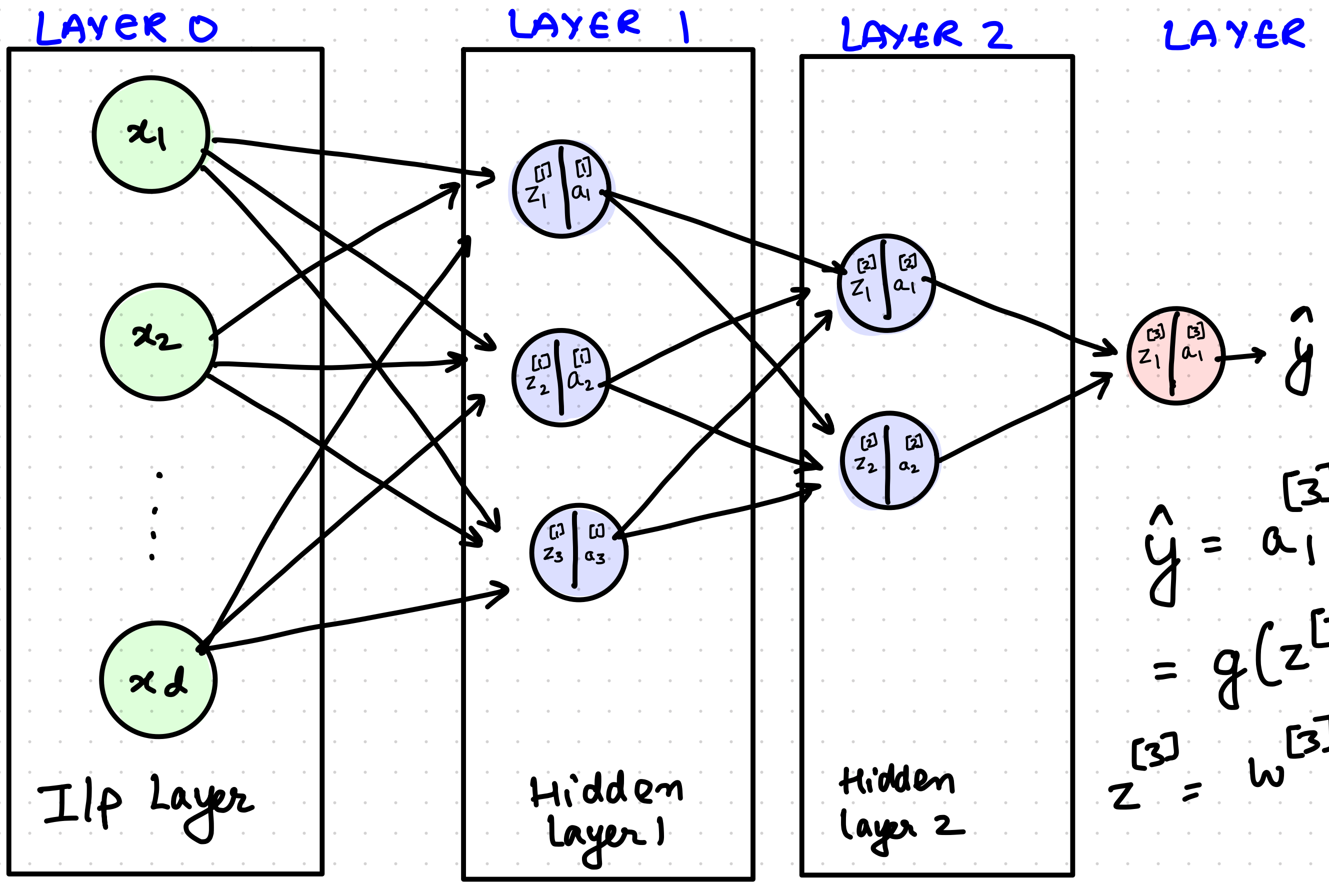
Q. Dim. of $W^{[2]}$?

$$W^{[2]} = \begin{bmatrix} - \; {w_1^{[2]}}^T \; - \\ - \; {w_2^{[2]}}^T \; - \end{bmatrix}$$

$$w_1^{[2]} \in R^3$$

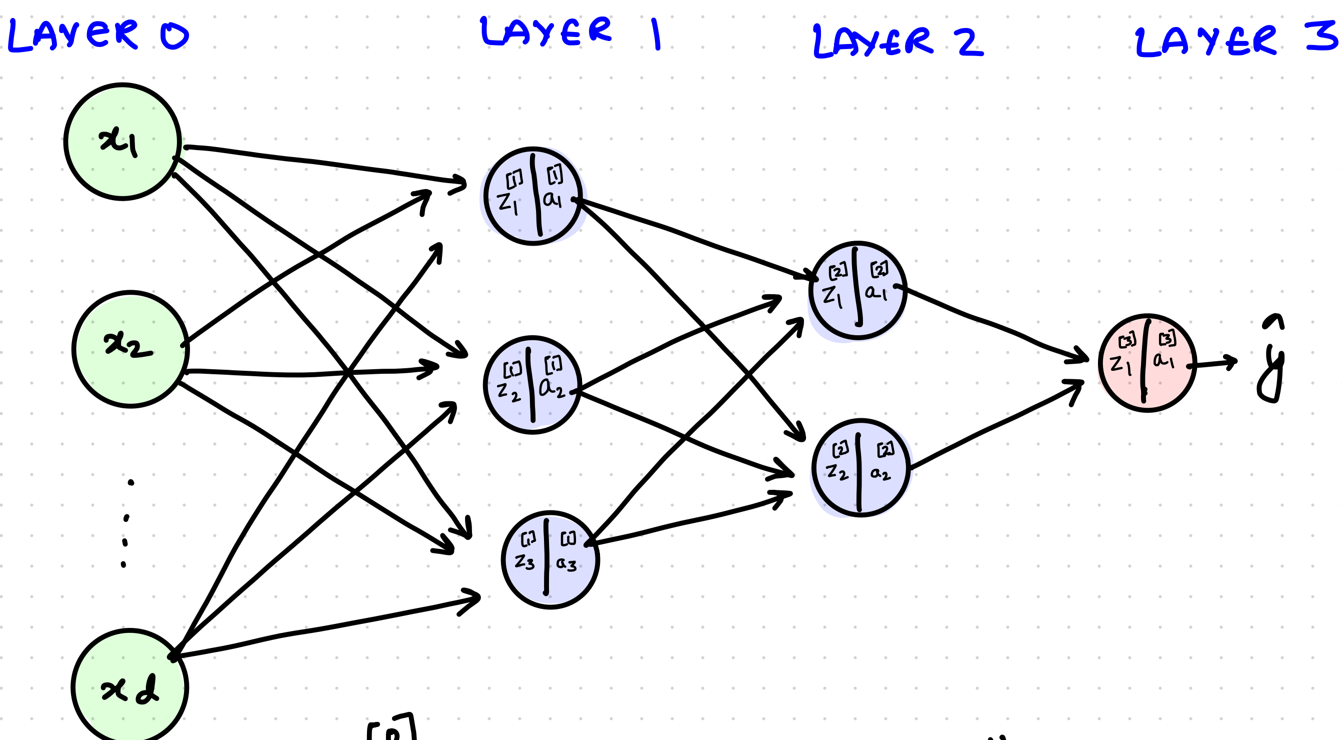
$$\therefore \; W^{[2]} \in R^{2\times 3}$$

FORWARD PROPAGATION



$$\hat{y} = a_1^{[3]}$$
$$= g\left(z^{[3]}\right)$$
$$z^{[3]} = w^{[3]} a^{[2]} + b^{[3]}$$

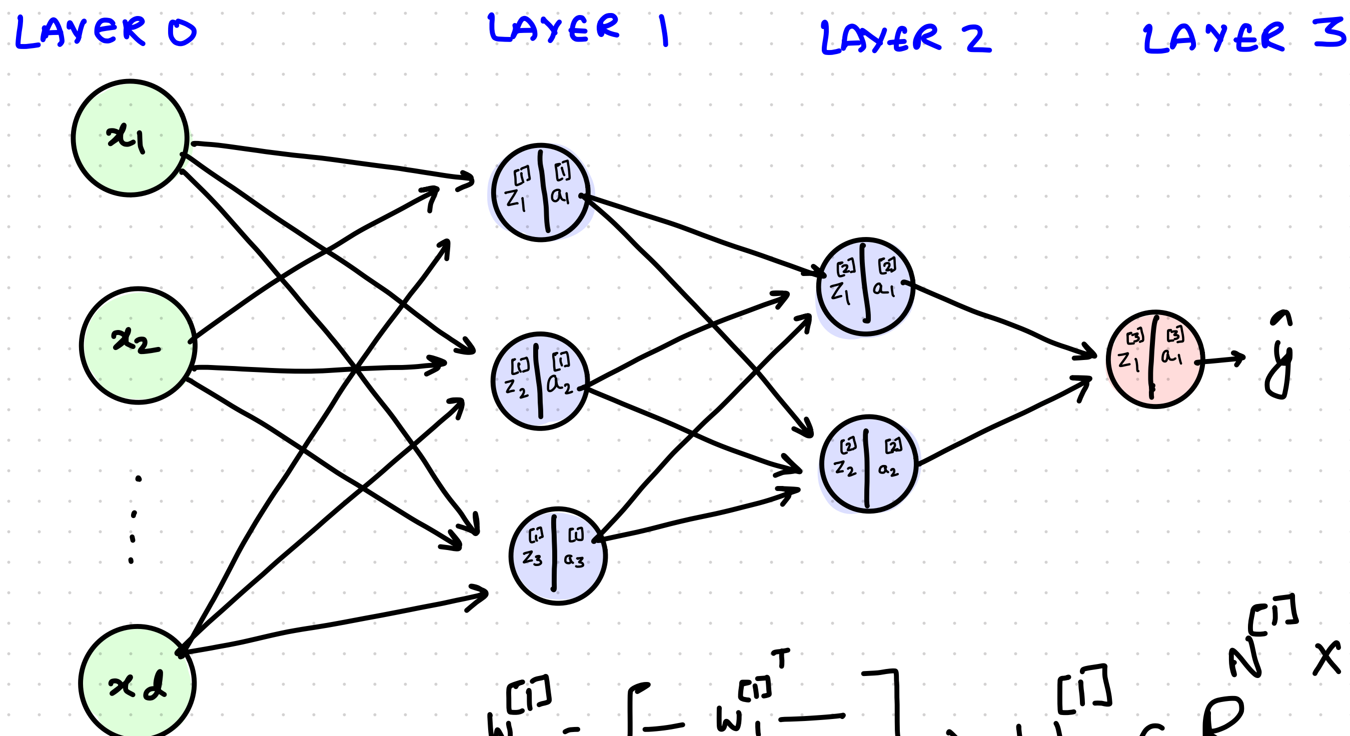
WHAT CAN WE SAY ABOUT SHAPES OF a, b, W



$a^{[0]} \in R^d$ or $R^{N^{[0]}}$ $\quad N^{[0]}$ = # units in 0^{th} Layer

$d \rightarrow$ # i/p features

WHAT CAN WE SAY ABOUT SHAPES OF a, b, W

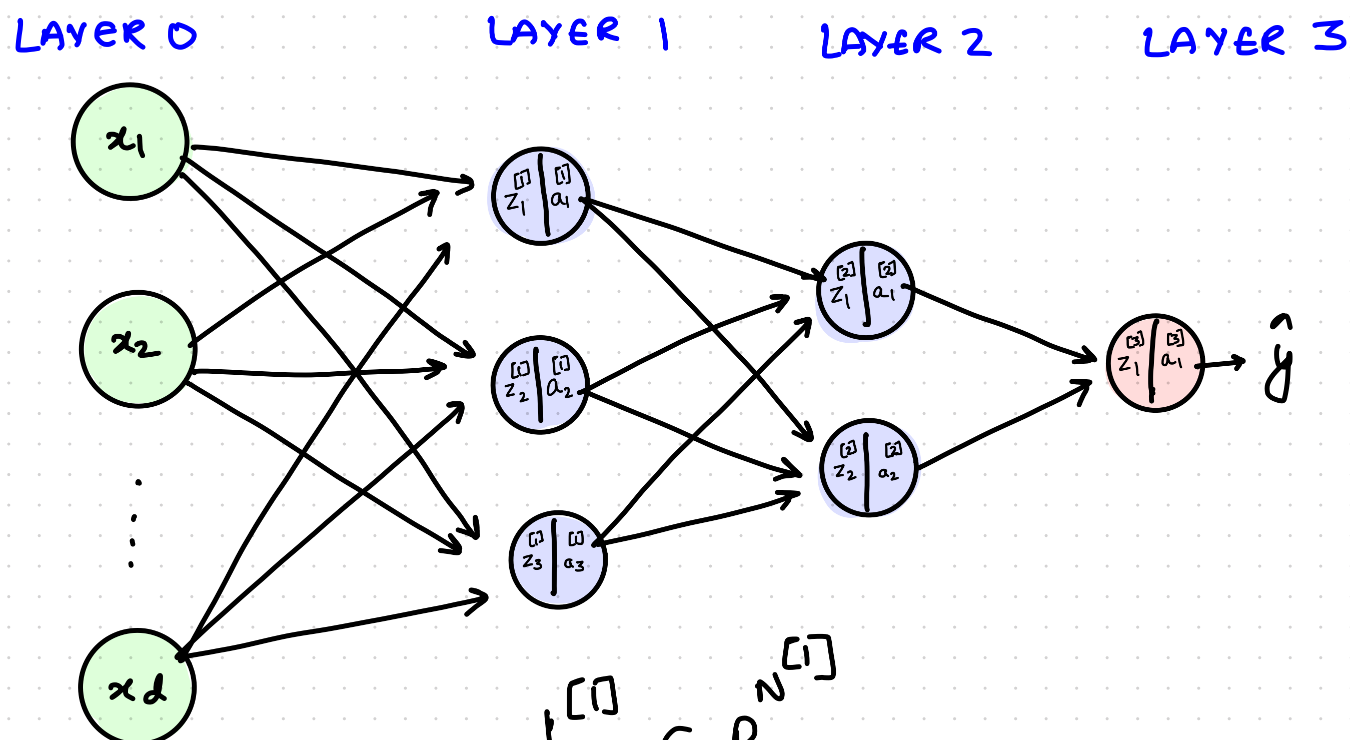


$a^{[0]} \in R^{d}$

$d \rightarrow$ # i/p features

$$W^{[1]} = \begin{bmatrix} - \; w_1^{[1]T} \; - \\ - \; w_2^{[1]T} \; - \\ \vdots \\ - \; w_{n^{[1]}}^{[1]T} \; - \end{bmatrix} \Rightarrow W^{[1]} \in R^{n^{[1]} \times n^{[0]}}$$

WHAT CAN WE SAY ABOUT SHAPES OF a, b, W



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

$a^{[0]} \in R^{d}$

$d \rightarrow$ # i/p features

SUMMARY OF SHAPES

$N^{[l]}$: # NODES IN l^{th} Layer

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

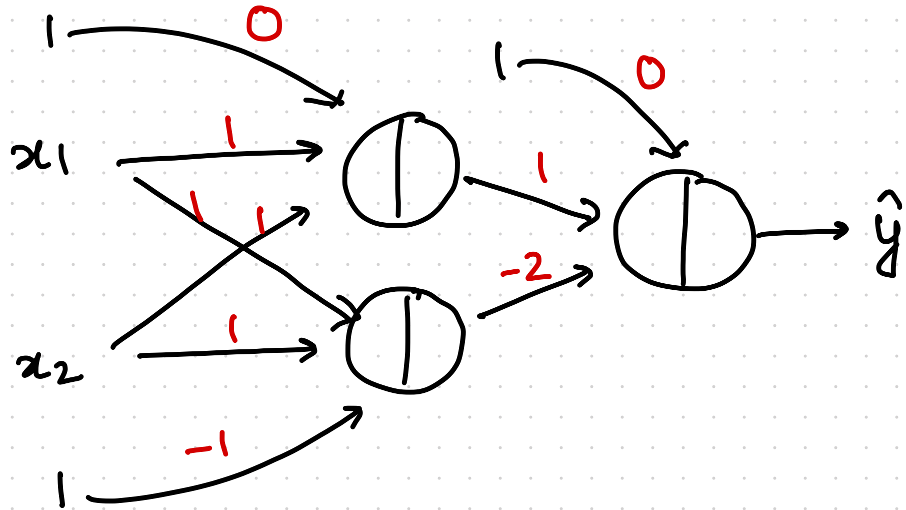
$W^{[l]} \in R^{N^{[l]} \times N^{[l-1]}}$

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

$Z^{[l]} \in R^{N^{[l]}}$

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

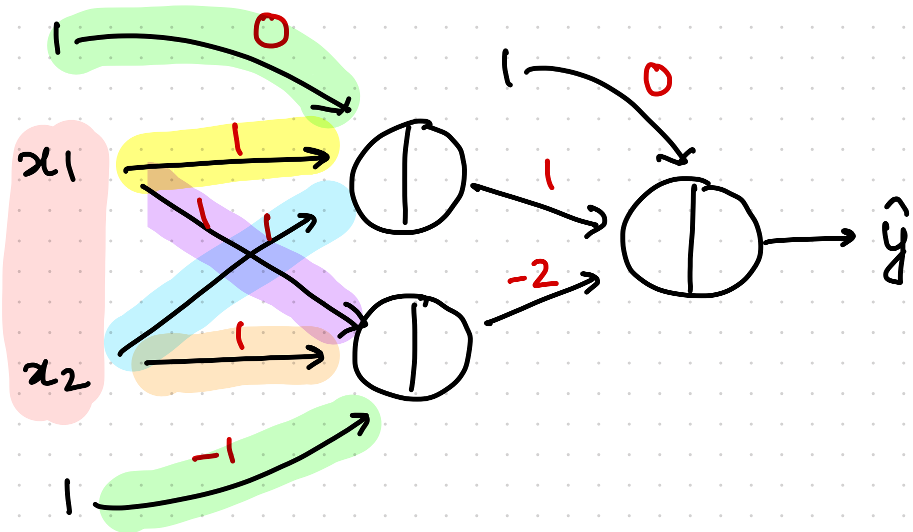
XOR USING "MLP" RELU



CONFIRM IF ABOVE N/W IS CORRECT FOR XOR.

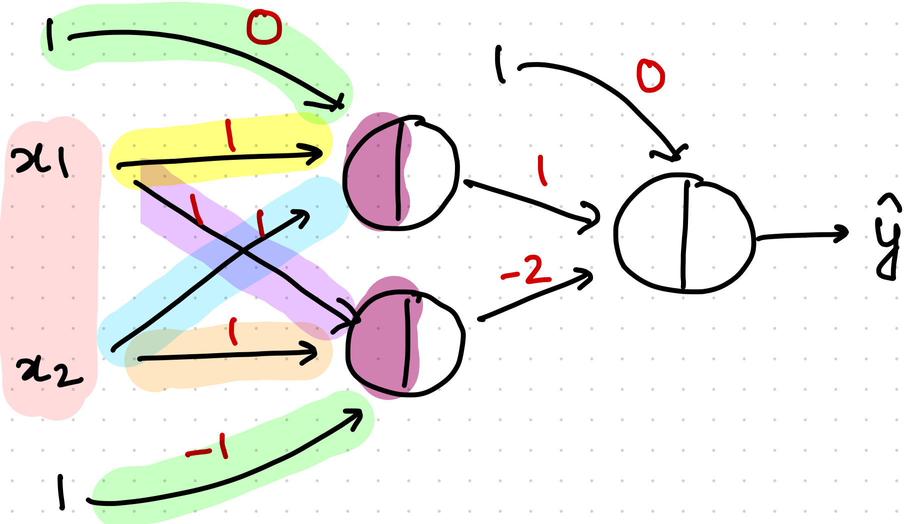
Start with $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $y_{TRUE} = 0$

XOR USING "MLP" RELU



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

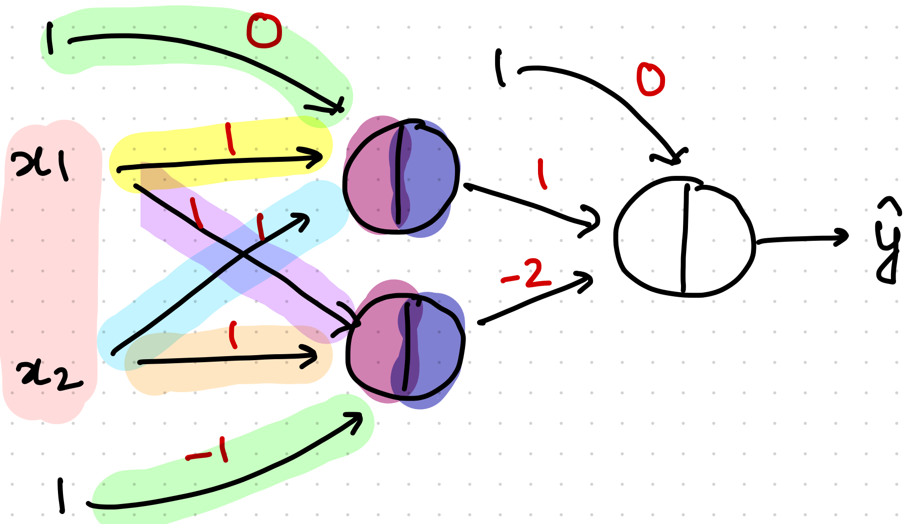
XOR USING "MLP" RELU



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

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

XOR USING "MLP" RELU

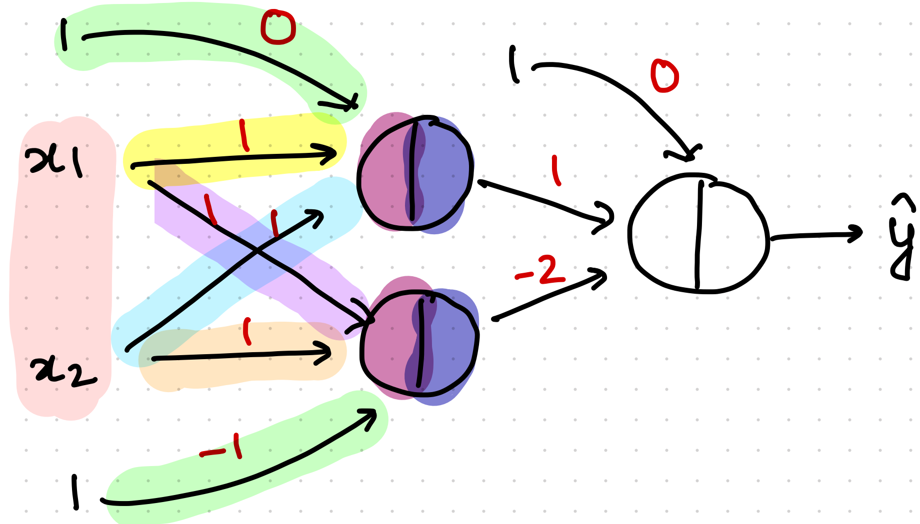


$$a^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; ; \; b^{[1]} = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \; ; \; W^{[1]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

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

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

XOR USING "MLP" RELU

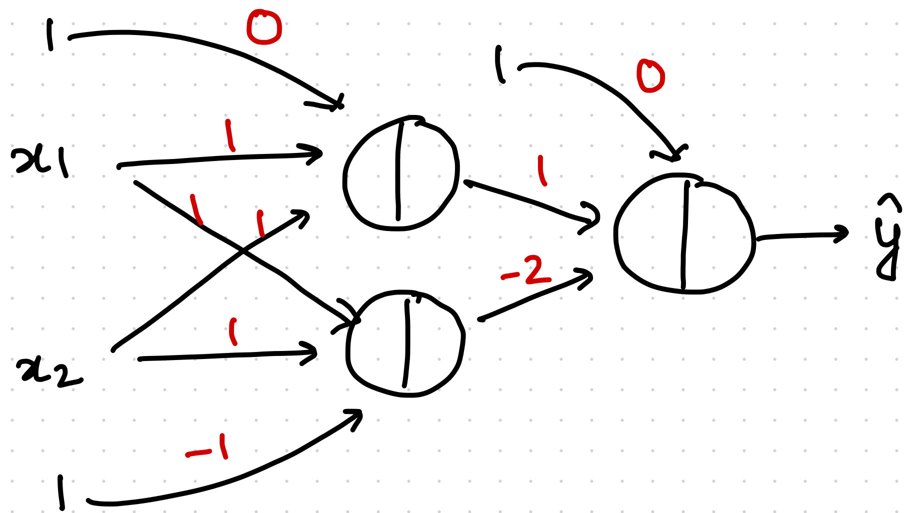


$$a^{[1]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad ; \quad W^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix} \quad ; \quad b^{[2]} = \begin{bmatrix} 0 \end{bmatrix}$$

$$Z^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} = 0 \quad ; \quad a^{[2]} = \hat{y} = \mathrm{RELU}([0]) = 0$$

✓

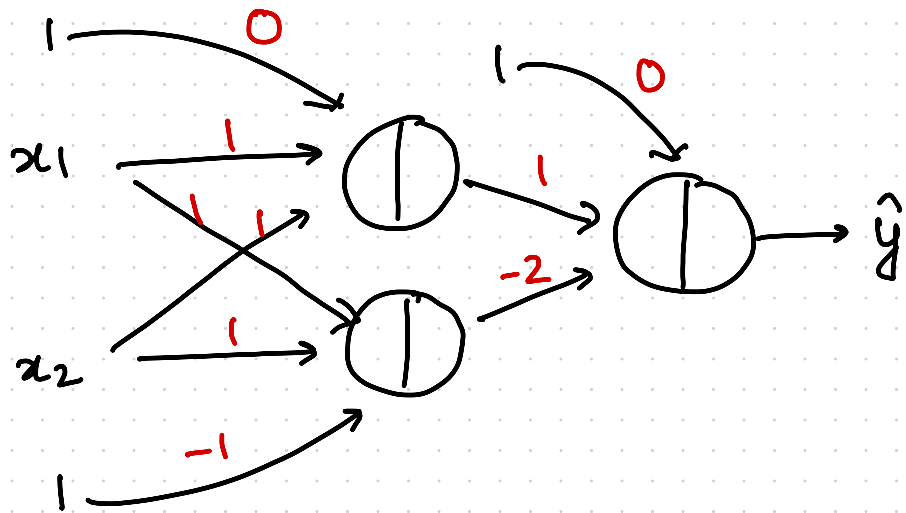
XOR USING "MLP" RELU



CONFIRM IF ABOVE N/W IS CORRECT FOR XOR.

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

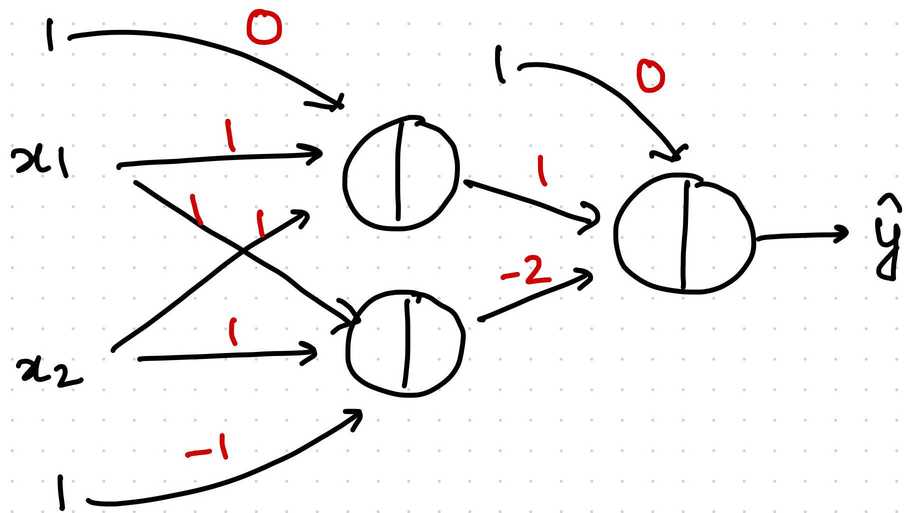
XOR USING "MLP" RELU



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

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

XOR USING "MLP" RELU



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

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

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

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

$$= \hat{y}$$

COMPUTATION FOR N EXAMPLES

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

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

matrix

COMPUTATION FOR N EXAMPLES

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

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

$Z^{[1]}$ ← Layer

(i) → Instance/ Sample

$$Z^{[1]}_{(i)} = W^{[1]} A^{[0]T}_{(i)} + b^{[1]} \in R^{N^{[1]}}$$

Independent of i

$$a^{[0]}_{(i)} \in R^d = \begin{bmatrix} \vdots \end{bmatrix}$$

COMPUTATION FOR N EXAMPLES

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

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

$$z_{(1)}^{[1]} = \left\{ W^{[1]} a_{(1)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

$$z_{(2)}^{[1]} = \left\{ W^{[1]} a_{(2)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

$$\vdots$$

$$z_{(N)}^{[1]} = \left\{ W^{[1]} a_{(N)}^{[0]} + b^{[1]} \right\} \in R^{N^{[1]}}$$

COMPUTATION FOR N EXAMPLES

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

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

$$z_{(1)}^{[1]} = W^{[1]} a_{(1)}^{[0]} + b^{[1]} \in R^{n^{[1]}}$$

$$z_{(2)}^{[1]} = W^{[1]} a_{(2)}^{[0]} + b^{[1]} \in R^{n^{[1]}}$$

$$\vdots$$

$$z_{(N)}^{[1]} = W^{[1]} a_{(N)}^{[0]} + b^{[1]} \in R^{n^{[1]}}$$

COMPUTATION FOR N EXAMPLES

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

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

$$z_{(1)}^{[1]} = W^{[1]} a_{(1)}^{[0]} + b^{[1]}$$

$$z_{(2)}^{[1]} = W^{[1]} a_{(2)}^{[0]} + b^{[1]}$$

$$\vdots$$

$$z_{(N)}^{[1]} = W^{[1]} a_{(N)}^{[0]} + b^{[1]}$$

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

COMPUTATION FOR N EXAMPLES

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

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

$$z_{(1)}^{[1]} = W^{[1]} a_{(1)}^{[0]} + b^{[1]}$$
$$z_{(2)}^{[1]} = W^{[1]} a_{(2)}^{[0]} + b^{[1]}$$
$$\vdots$$
$$z_{(N)}^{[1]} = W^{[1]} a_{(N)}^{[0]} + b^{[1]}$$

$$\Rightarrow Z^{[1]} = \begin{array}{c|c} & \overleftrightarrow{\;\; N^{[1]} \;\;} \\ \hline \updownarrow N & \begin{bmatrix} — z_{(1)}^{[1]T} — \\ — z_{(2)}^{[1]T} — \\ \vdots \\ — z_{(N)}^{[1]T} — \end{bmatrix} \end{array} \in R^{N \times N^{[1]}}$$

COMPUTATION FOR N EXAMPLES

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

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

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

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

all same entries

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

COMPUTATION FOR N EXAMPLES

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

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

$$b^{[1]} \in R^{n^{[1]}}$$

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

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

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

COMPUTATION FOR N EXAMPLES

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

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

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

$b^{[1]} \in R^{n^{[1]}}$

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

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

$$Z^{[1]} = A^{[0]} W^{[1]^T} + B^{[1]} \Rightarrow A^{[1]} = g\left(Z^{[1]}\right)$$

XOR ALL EXAMPLES

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

XOR ALL EXAMPLES

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

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

XOR ALL EXAMPLES

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

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

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

XOR ALL EXAMPLES

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

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

$$W^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix}$$

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

XOR ALL EXAMPLES

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

$$W^{[2]} = \begin{bmatrix} 1 & -2 \end{bmatrix}_{1\times 2}$$

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

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

XOR ALL EXAMPLES

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

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

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?

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

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

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

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

PARAMETERS

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

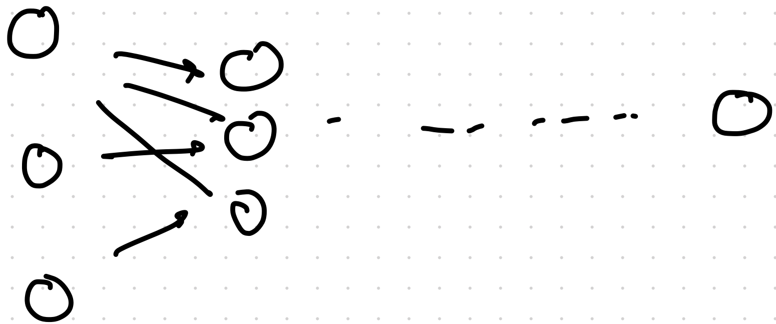
$$\boxed{\text{PARAMS} = \sum_{l=1}^{L} N^{[l]} \cdot N^{[l-1]} + N^{[l]}}$$

? XOR 4 examples, 9 parameters !!

$\Rightarrow$ NOTEBOOK: XOR demo

LEARNING PARAMETERS

FORWARD PROPAGATION (PREDICT BASED ON CURRENT PARAMS)



BACKWARD PROPAGATION (CHANGE WEIGHTS TO IMPROVE OBJECTIVE)

LEARNING PARAMETERS

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

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

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

Params : $W^{[1]}, b^{[1]}, \ldots\ldots$

LEARNING PARAMETERS

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

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

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

Params : $\theta = \{W^{[1]}, b^{[1]}, \ldots\ldots\}$

GRADIENT DESCENT

① INIT params randomly

② Till convergence :

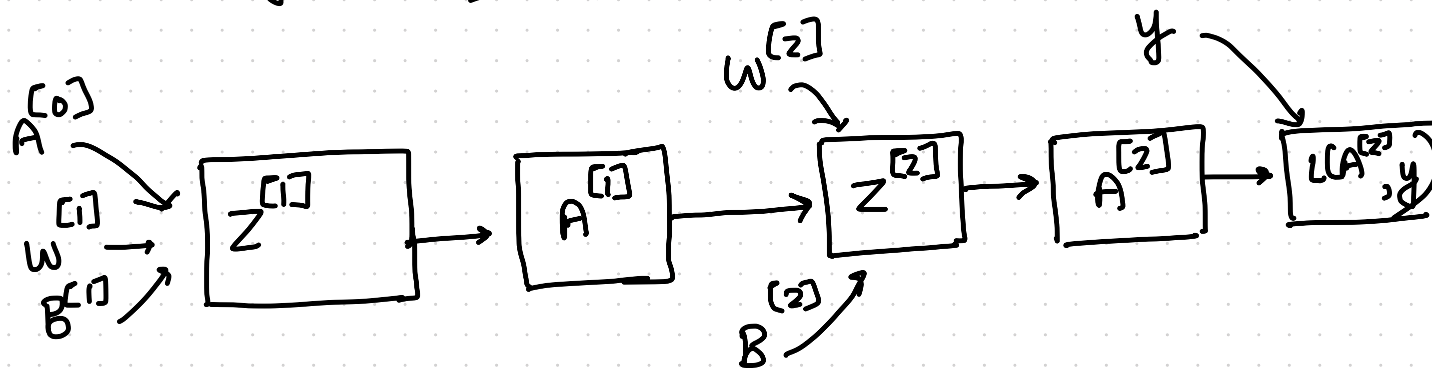
$$W^{[1]} = W^{[1]} - \alpha \boxed{\frac{\partial J(\theta)}{\partial W^{[1]}}} \ldots$$

$\vdots$

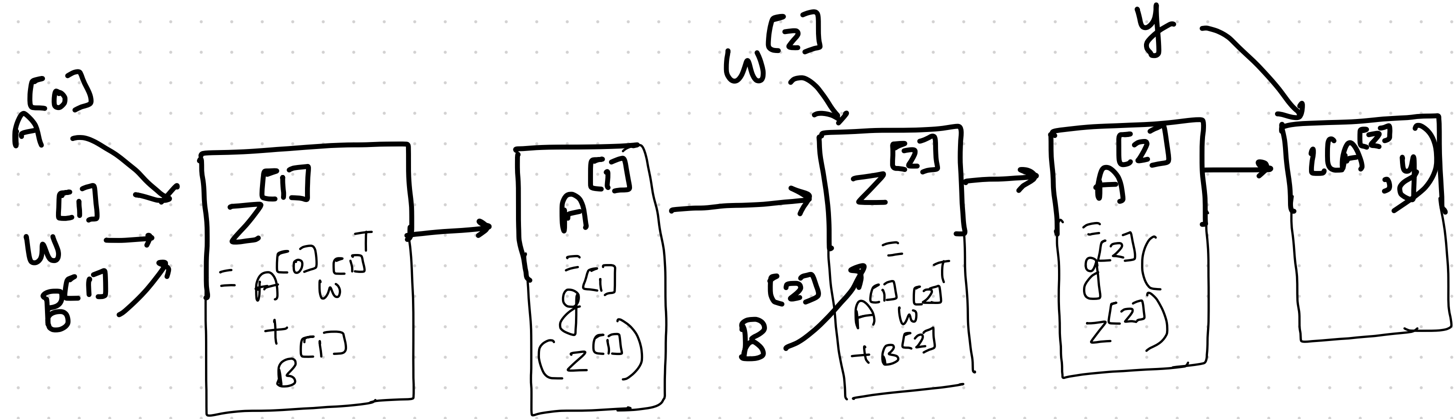
HOW TO COMPUTE?

COMPUTATION GRAPH (FOR XOR EXAMPLE)

1. $Z^{[1]} = A^{[0]} W^{[1]^T} + B^{[1]}$
2. $A^{[1]} = g^{[1]}(Z^{[1]})$
3. $Z^{[2]} = A^{[1]} W^{[2]^T} + B^{[2]}$
4. $A^{[2]} = g^{[2]}(Z^{[2]}) = \hat{y}$



COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

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

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

DERIVATIVES OF ACTIVATION FUNCTIONS

RELU

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

$\downarrow$ Assume $z \neq 0$

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

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

DERIVATIVES OF ACTIVATION FUNCTIONS

RELU (LEAKY)

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

$$\downarrow$$

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

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

DERIVATIVES OF ACTIVATION FUNCTIONS

SIGMOID

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

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

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

DERIVATIVES OF ACTIVATION FUNCTIONS

TANH

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

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

$$= 1 - \left(g(z)\right)^2$$

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

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

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

SIGMOID

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

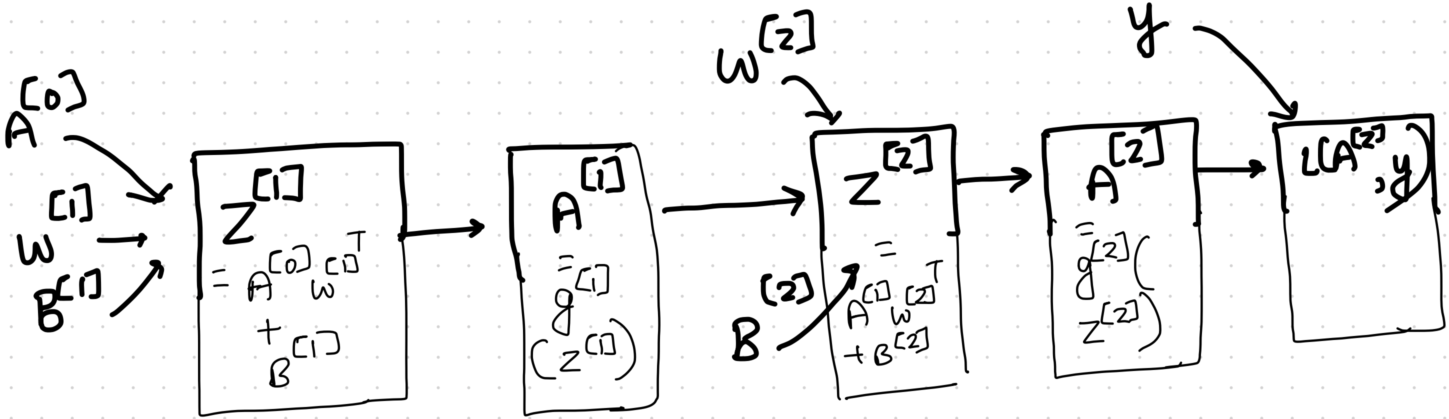
$$g'(z) = g(z) * (1 - g(z))$$

TANH

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

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

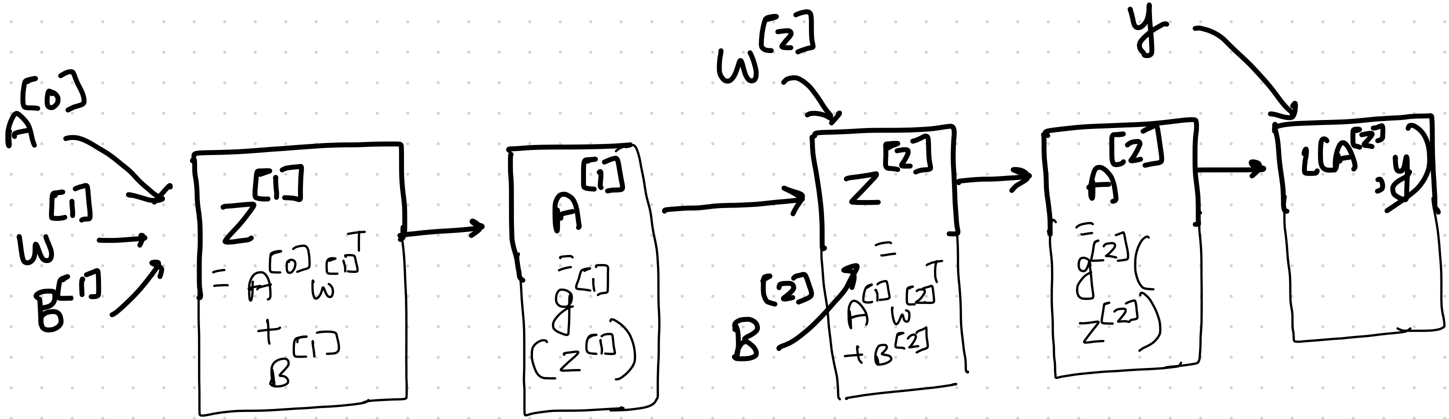
COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

WRITE IN VECTOR FORM

COMPUTATION GRAPH (FOR XOR EXAMPLE)



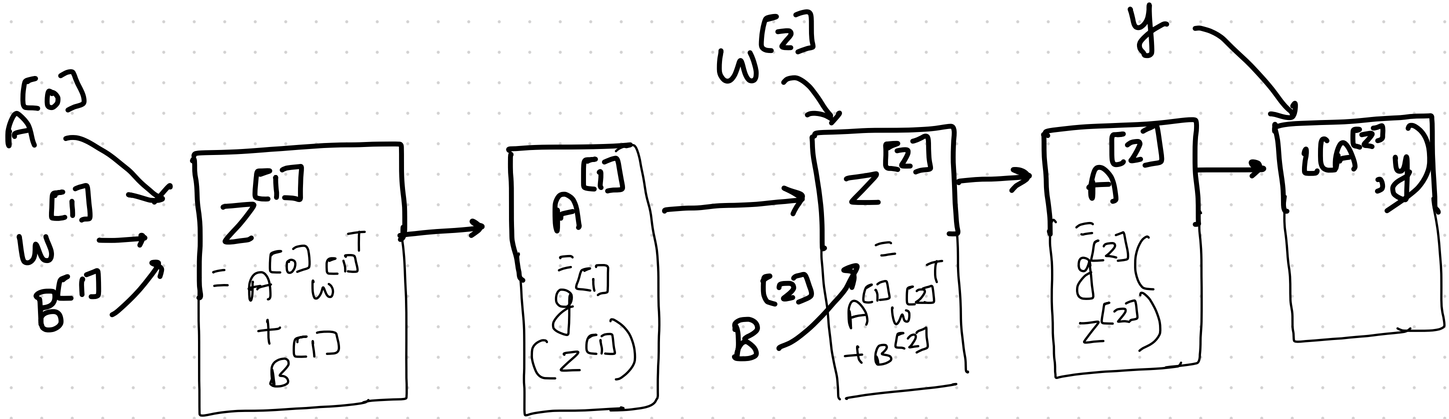
$$L(A^{[2]}, y) = \sum_{i=1}^{N} -y_{(i)} \log A^{[2]}_{(i)} - (1 - y_{(i)}) \log (1 - A^{[2]}_{[i]})$$

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

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

APPLIED ELEMENT-WISE

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

$$= -\underbrace{y^T}_{1\times N} \log \underbrace{(A^{[2]})}_{N\times 1} - \underbrace{(1 - y)^T}_{1\times N} \log \underbrace{(1 - A^{[2]})}_{N\times 1} \qquad \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}_{N\times 1}$$

APPLIED ELEMENT-WISE

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

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

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

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

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

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

$$\frac{\partial q}{\partial A^{[2]}}_{N \times 1} = y_{N \times 1} \oslash A^{[2]}_{N \times 1}$$

Element-wise division

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

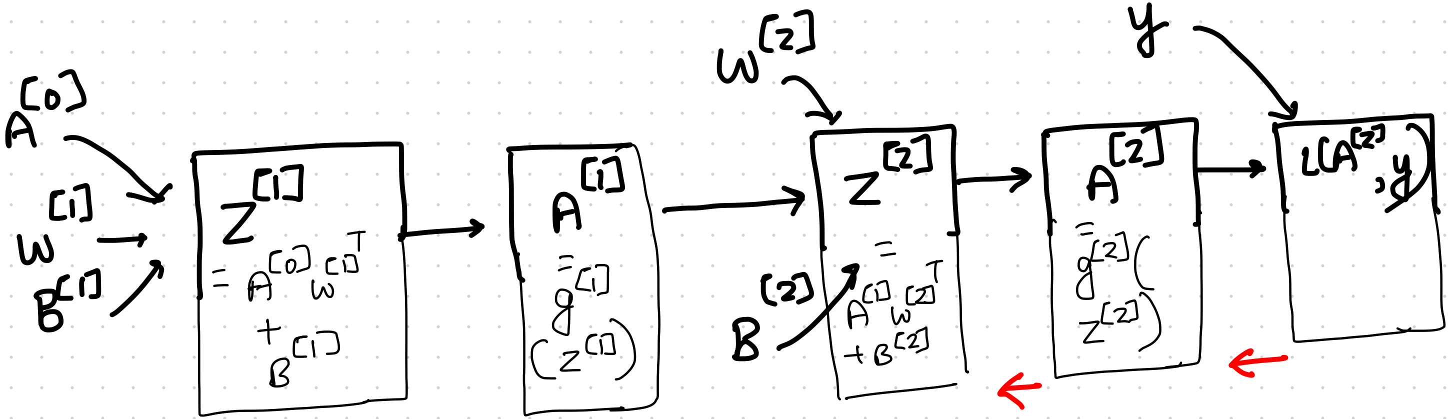
$$r = \left(1 - y(1)\right) \log \left(1 - A^{[2]}_{(1)}\right) + \cdots\cdots$$

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

$$\underset{N \times 1}{\frac{\partial r}{\partial A^{[2]}}} = -\underset{N \times 1}{(1-y)} \oslash \underset{N \times 1}{(1 - A^{[2]})}$$

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

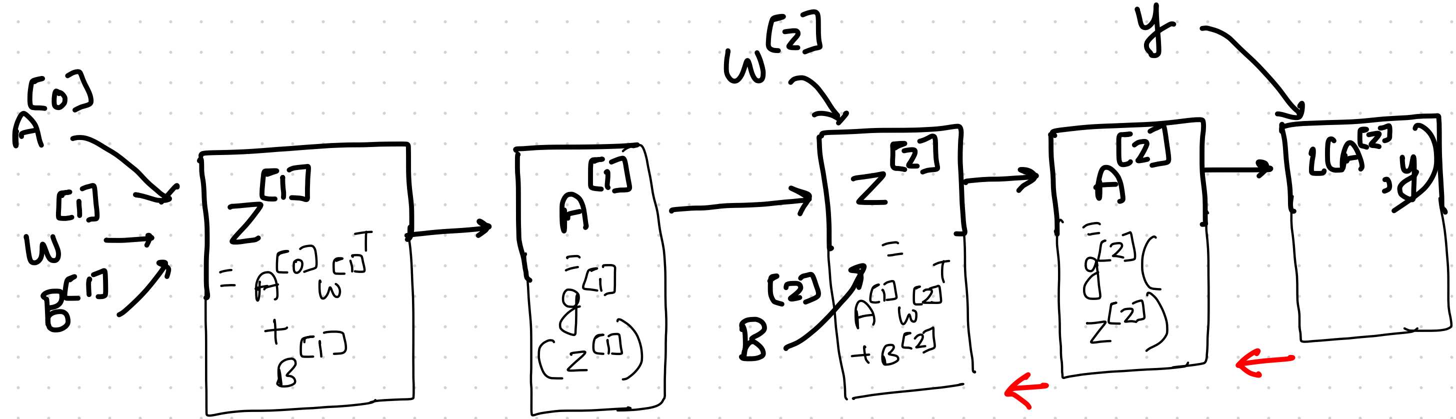


$$\frac{\partial L}{\partial Z^{[2]}} = ?$$

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

(Chain Rule)

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

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

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

$\because g^{[2]} = $ SIGMOID

COMPUTATION GRAPH (FOR XOR EXAMPLE)

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

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

Element-wise multiply

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

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

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

Element-wise

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

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

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

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

Element-wise

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

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

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

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

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

Element-wise

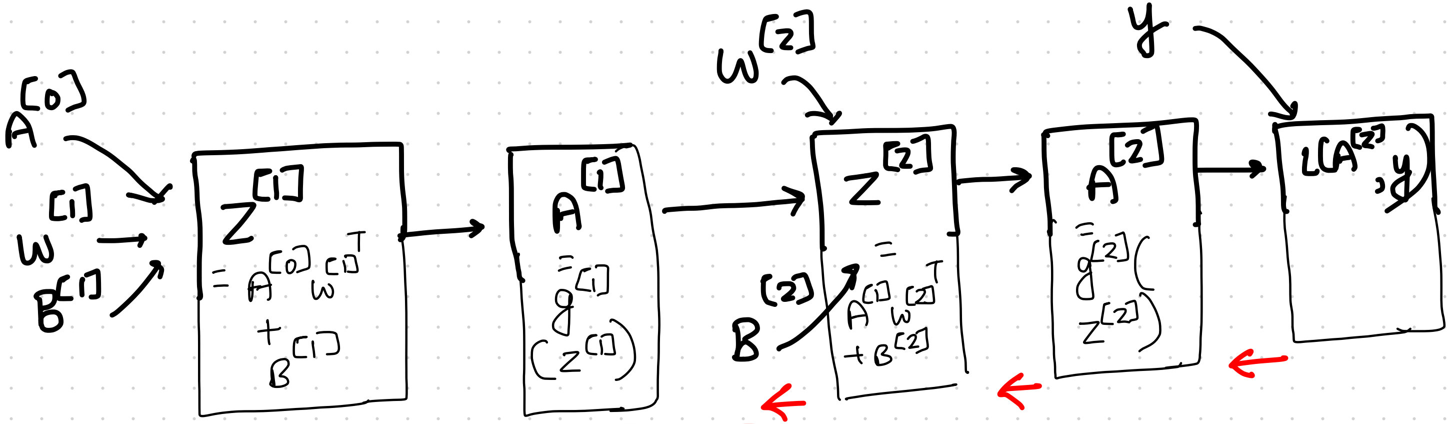
$$= \left(-y \oslash A^{[2]} + (1-y) \oslash (1-A^{[2]})\right)\left(A^{[2]}(1-A^{[2]})\right)$$

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

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

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

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

$f: R^N \longrightarrow R^M$

I/p: R^N

O/p: R^M

$y = f(x)$

Derivative of 'f' at 'x' called Jacobian is $M \times N$ matrix

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & & \frac{\partial y_m}{\partial x_N} \end{pmatrix}_{M \times N}$$

GENERALISED JACOBIAN : TENSOR IN, TENSOR OUT

$$f : R^{N_1 \times \cdots \times N_{D_x}} \longrightarrow R^{M_1 \times \cdots \times M_{D_y}}$$

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

o/p: D_y- dimensional tensor of shape $M_1 \times \cdots M_{D_y}$

$$y = f(x)$$

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

shape

BACK PROP. WITH TENSORS

Let $G = \alpha \beta$

α : $N \times D$

β : $D \times M$

G : $N \times M$

BACK PROP. WITH TENSORS

Let $G = \alpha\beta$

$\alpha : N \times D$

$\beta : D \times M$

$G : N \times M$

→ we are given Loss L as funcⁿ of G

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

BACK PROP. WITH TENSORS

Let $G = \alpha \beta$

$\alpha : N \times D$

$\beta : D \times M$

$G : N \times M$

→ We are given Loss L as funcⁿ 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

Let $G = \alpha\beta$

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

α : $N \times D$

β : $D \times M$

G : $N \times M$

BACK PROP. WITH TENSORS

Let $G = \alpha \beta$

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

$\beta : D \times M$

$G : N \times M$

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

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

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

BACK PROP. WITH TENSORS

$$G_{1\times3} = \begin{pmatrix} G_{1,1} & G_{1,2} & G_{1,3} \end{pmatrix}$$

$$\alpha_{1\times2} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} \end{pmatrix}$$

$$\beta_{2\times3} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix}$$

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

BACK PROP. WITH TENSORS

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

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

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

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

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

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

BACK PROP. WITH TENSORS

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

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

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

BACK PROP. WITH TENSORS

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

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

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

$\underline{\text{Shape}}$ = $(N\times M)\times 1$

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

$\therefore$ 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 $\dfrac{\partial L}{\partial G} = [dG_{1,1} \quad dG_{1,2} \quad dG_{1,3}]$

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

$$\frac{\partial L}{\partial \alpha_{1,1}} = \underbrace{\frac{\partial L}{\partial G}}_{1\times(N\times M)} \cdot \underbrace{\frac{\partial G}{\partial \alpha_{1,1}}}_{(N\times M)\times 1}$$

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

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

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

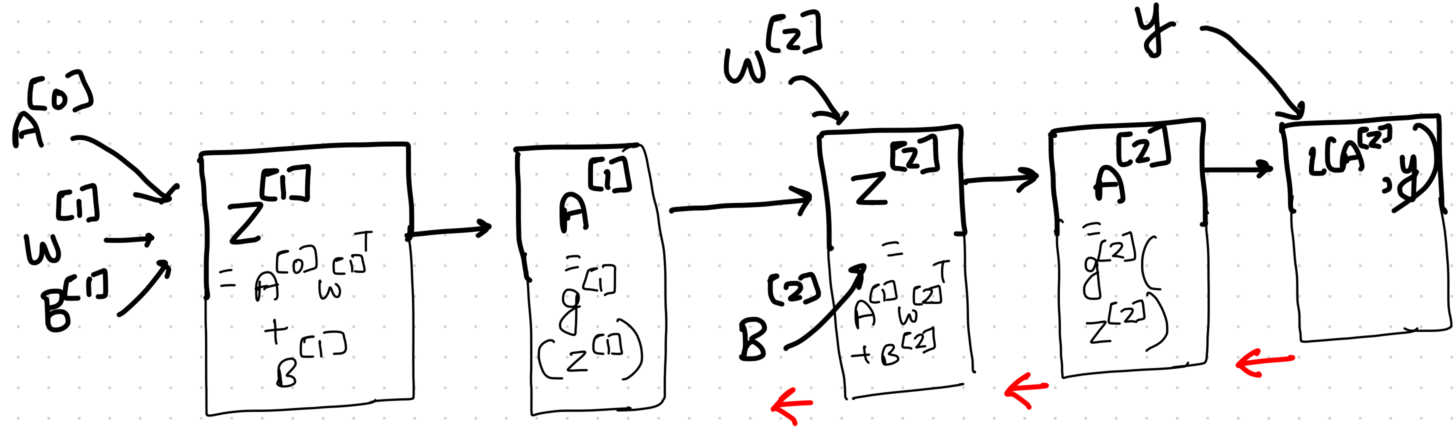
BACK PROP. WITH TENSORS

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

Similarly;

$$\underset{1\times(D\times M)}{\frac{\partial L}{\partial \beta}} = \underset{D\times N}{\alpha^T} \quad \underset{1\times(N\times M)}{\frac{\partial L}{\partial G}}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



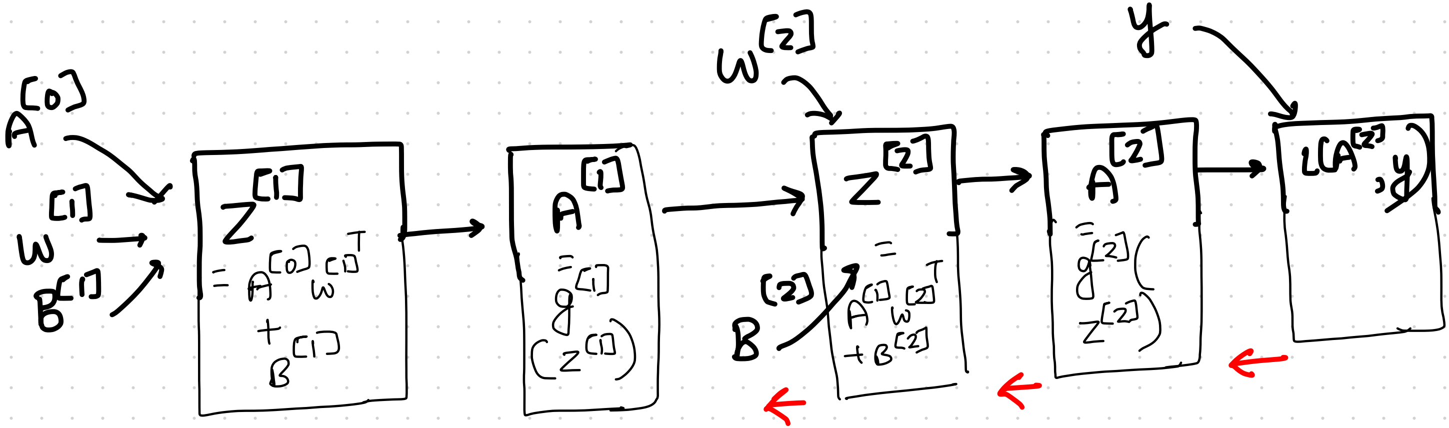
Equivalence from Aside

$G \longleftrightarrow Z^{[2]}$

$\alpha \longleftrightarrow A^{[1]}$

$\beta \longleftrightarrow W^{[2]^T}$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



Equivalence from <u>Aside</u>

$G \longleftrightarrow Z^{[2]}$

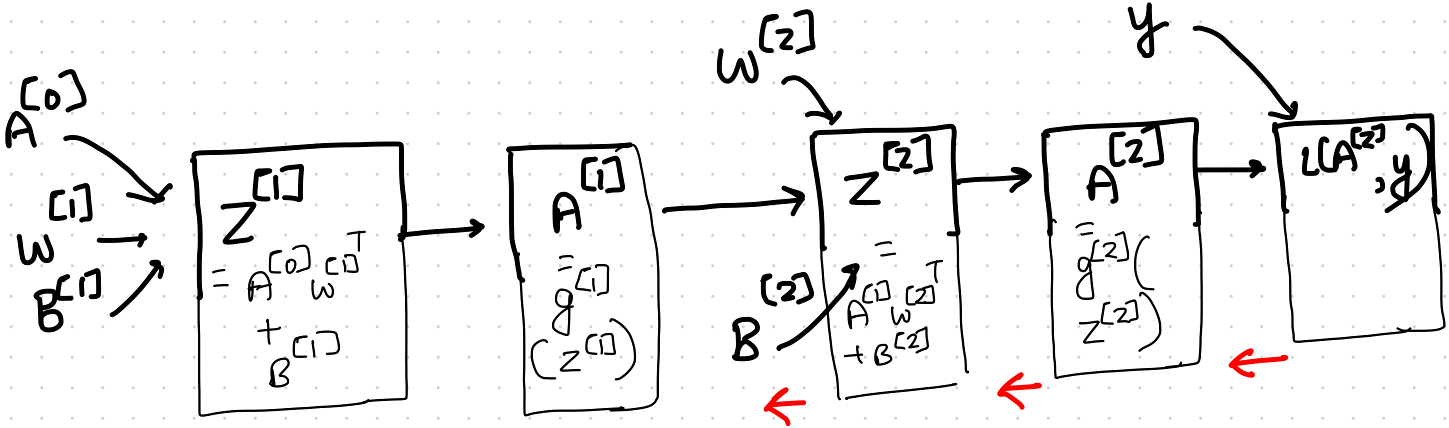
$\alpha \longleftrightarrow A^{[1]}$

$\beta \longleftrightarrow W^{[2]^T}$

$$\frac{\partial L}{\partial W^{[2]^T}} = A^{[1]^T} \frac{\partial L}{\partial Z^{[2]}}$$

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

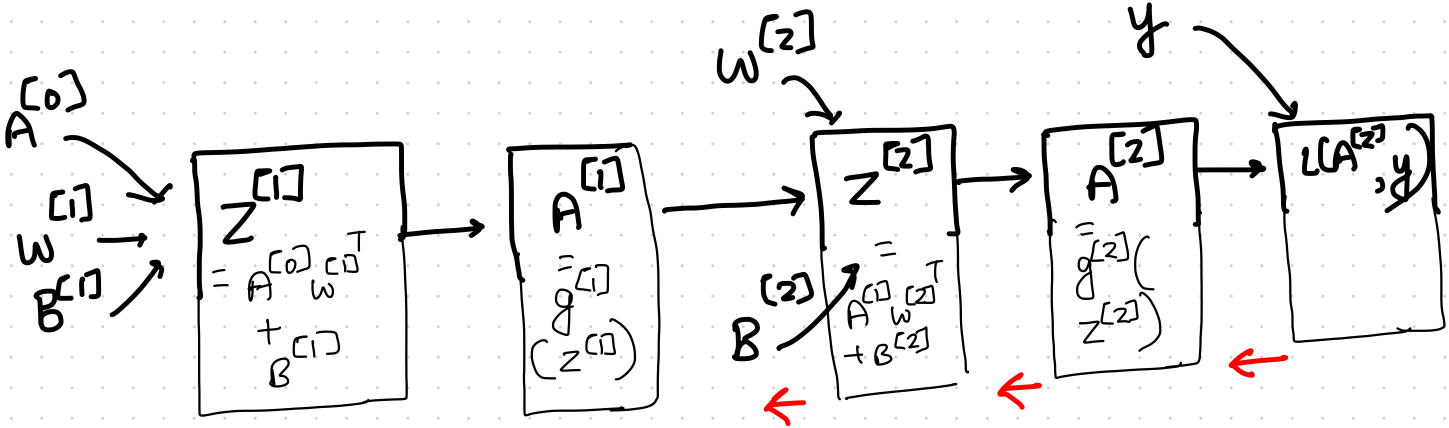
COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

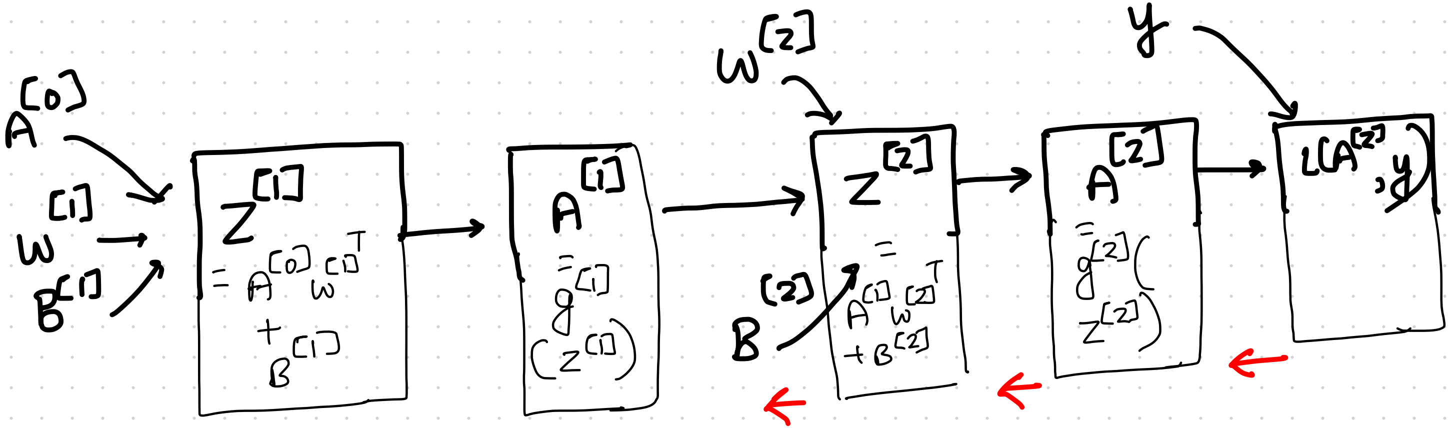
$$\boxed{\Rightarrow \underbrace{\frac{\partial L}{\partial W^{[2]}}}_{1\times(N^{[2]}\times N^{[1]})} = \underbrace{\left(A^{[2]} - y\right)^T}_{N^{[2]}\times N} \underbrace{A^{[1]}}_{N\times N^{[1]}}}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

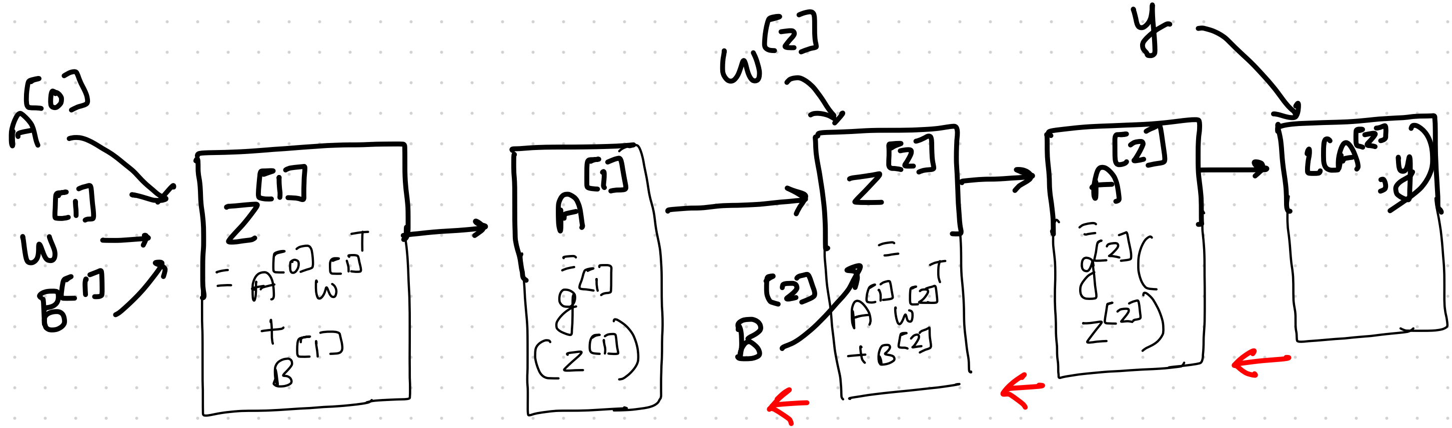


$$\frac{\partial L}{\partial B^{[2]}} = ?$$

Let $c = b^{[2]^T}$

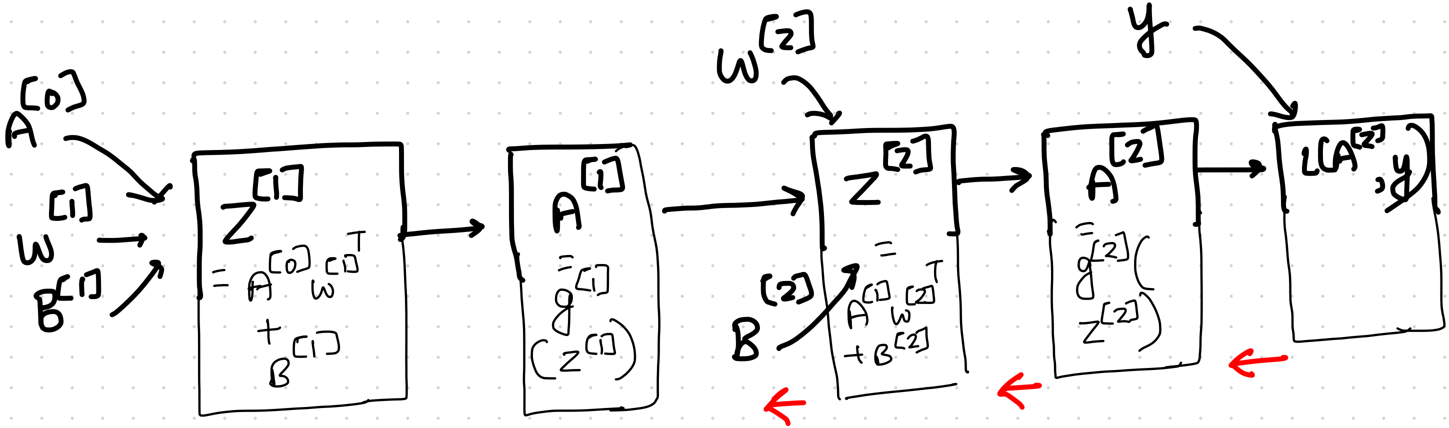
$$B^{[2]} = \begin{bmatrix} - & c & - \\ - & c & - \\ & \vdots & \\ - & c & - \end{bmatrix}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

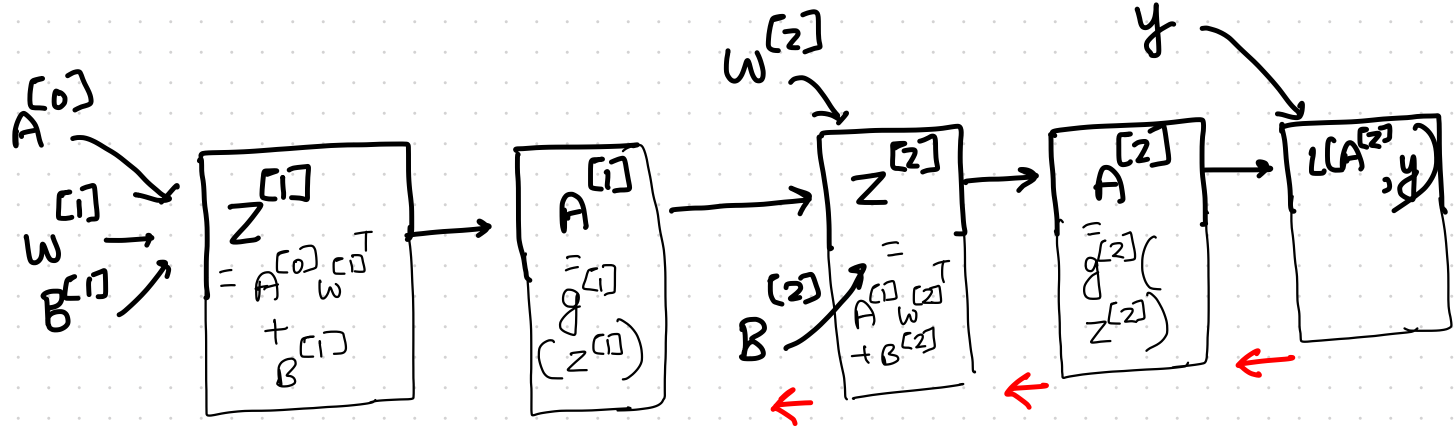
COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\frac{\partial L}{\partial C}_{1\times(1\times N^{[2]})} = \left[\frac{\partial L}{\partial c_1} \cdots\cdots \frac{\partial L}{\partial c_{N^{[2]}}}\right]$$

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

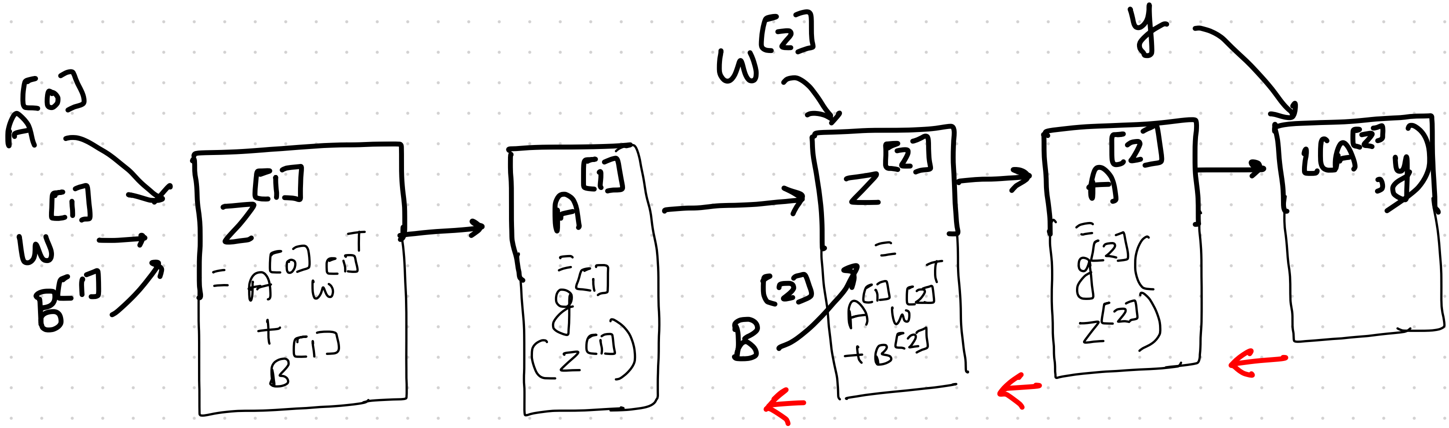
COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

$$Z^{[2]} = \begin{bmatrix} Z^{[2]}_{1,1} & \cdots & Z^{[2]}_{1,N^{[2]}} \\ \vdots & & \\ Z^{[2]}_{N,1} & \cdots & Z^{[2]}_{N,N^{[2]}} \end{bmatrix}_{N \times N^{[2]}} = A^{[2]} W^{[2]^T} + \begin{bmatrix} c_1 & c_2 \cdots & c_{N^{[2]}} \\ c_1 & c_2 \cdots & c_{N^{[2]}} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

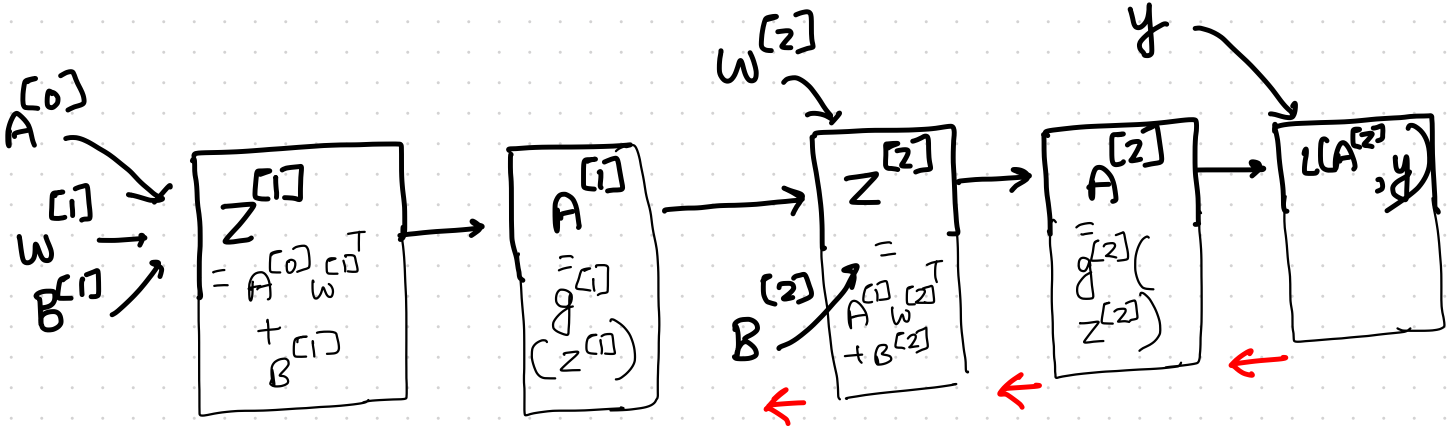
COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)

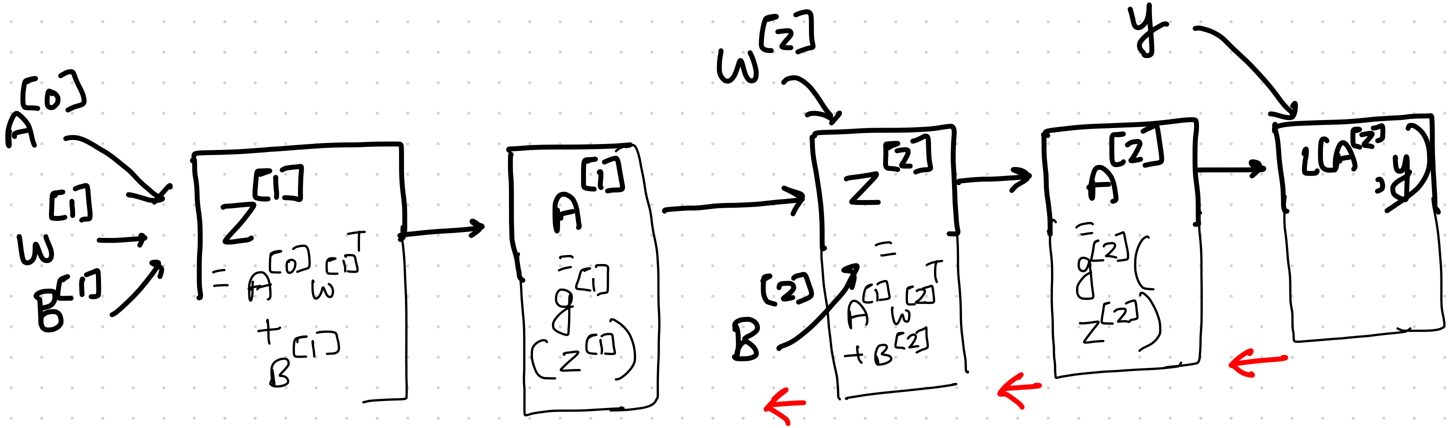


$$\frac{\partial Z}{\partial c_1} = \begin{bmatrix} \frac{\partial c_1}{\partial c_1} & \frac{\partial c_2}{\partial c_1} & \cdots \\ \vdots & & \cdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & \cdots & & & \\ 1 & & & & \\ 1 & \cdots & & \cdots & \end{bmatrix}$$

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

$$\frac{\partial L}{\partial c_1} = \underbrace{\frac{\partial L}{\partial z^{[2]}}}_{1 \times (N \times N^{[2]})} \underbrace{\frac{\partial z^{[2]}}{\partial c_1}}_{(N \times N^{[2]}) \times 1}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



Let $\dfrac{\partial L}{\partial Z^{[2]}} = \begin{bmatrix} dz^{[2]}_{1,1} & \cdots & dz^{[2]}_{1,N^{[2]}} \\ \\ dz^{[2]}_{N,1} & \cdots & dz^{[2]}_{N,N^{[2]}} \end{bmatrix}$

Then $\boxed{\dfrac{\partial L}{\partial c_1} = \sum_{i=1}^{N} dz^{[2]}_{i,1}}$

COMPUTATION GRAPH (FOR XOR EXAMPLE)

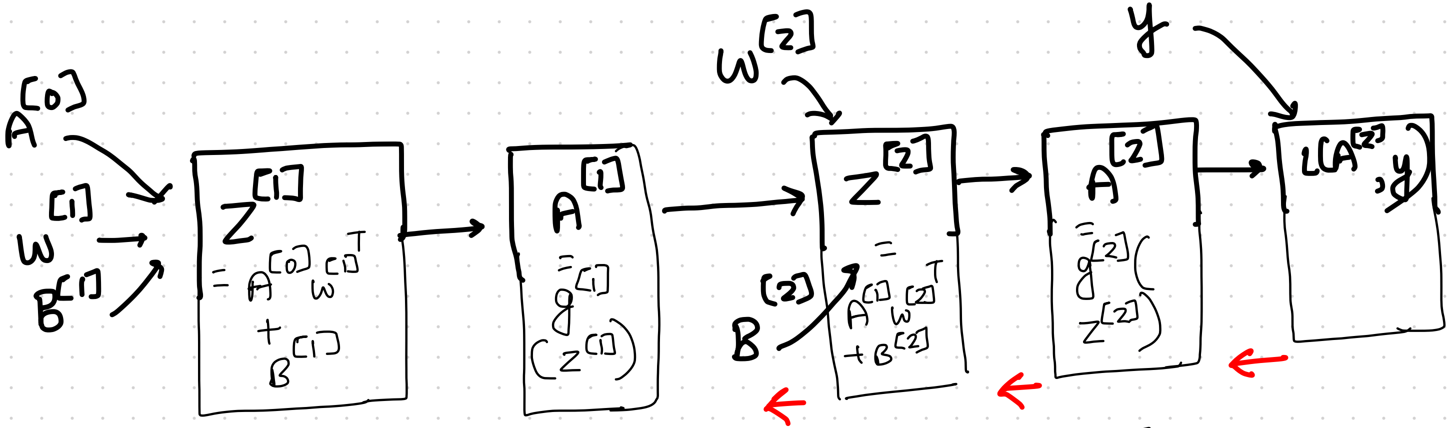
$$\frac{\partial L}{\partial c_1} = \sum_{i=1}^{N} dz^{[2]}_{i,1}$$

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

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

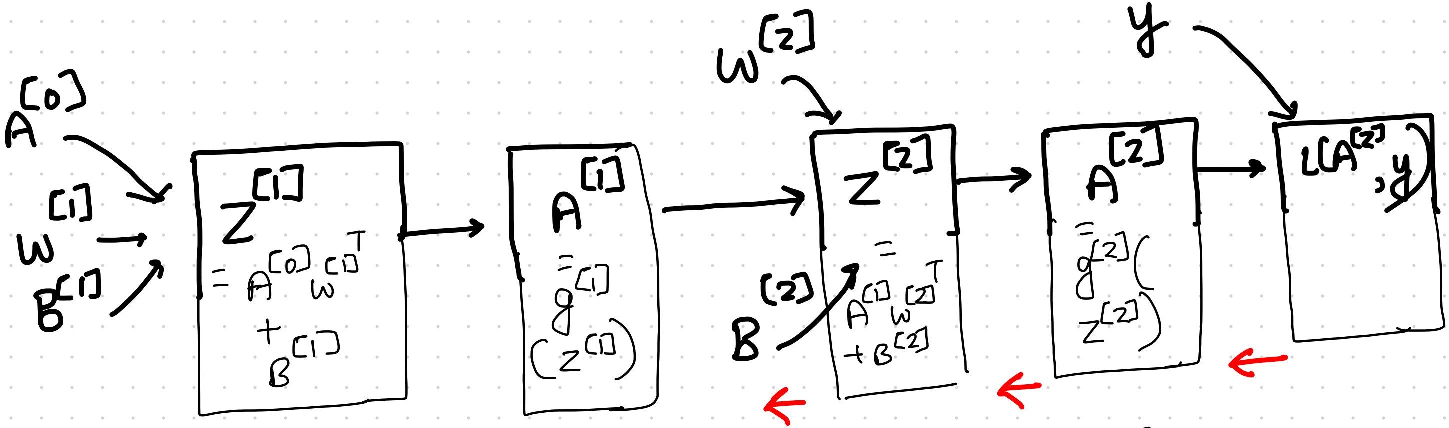
$N^{[2]} \times 1$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



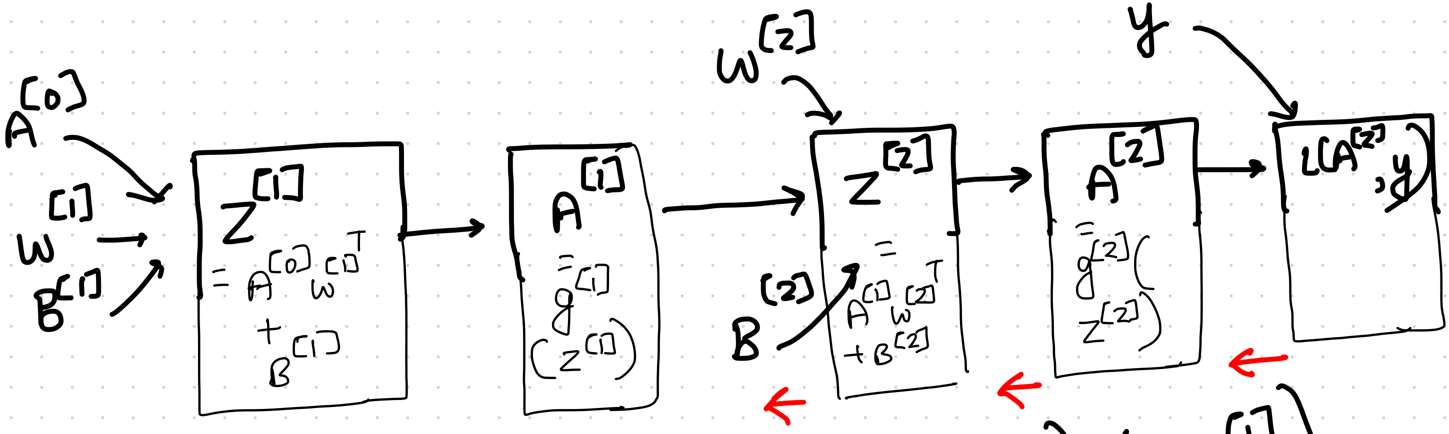
$$\frac{\partial L}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \left(W^{[2]^T}\right)^T = \frac{\partial L}{\partial Z^{[2]}} W^{[2]}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



$$\underbrace{\frac{\partial L}{\partial A^{[1]}}}_{1\times(N\times N^{[1]})} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial A^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \left(W^{[2]T}\right)^T = \underbrace{\frac{\partial L}{\partial Z^{[2]}}}_{1\times(N\times N^{[2]})} \underbrace{W^{[2]}}_{N^{[2]}\times N^{[1]}}$$

COMPUTATION GRAPH (FOR XOR EXAMPLE)



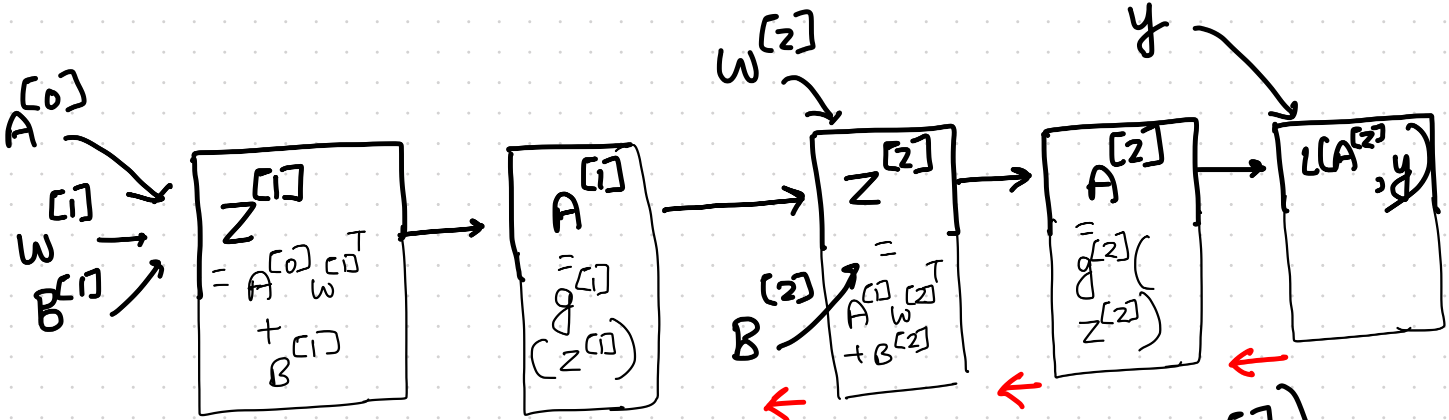
$$\frac{\partial L}{\partial Z^{[1]}} = \frac{\partial L}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial Z^{[1]}} = \left(\frac{\partial L}{\partial Z^{[2]}} W^{[2]}\right)\left(\frac{\partial A^{[1]}}{\partial Z^{[1]}}\right)$$

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

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

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

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

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

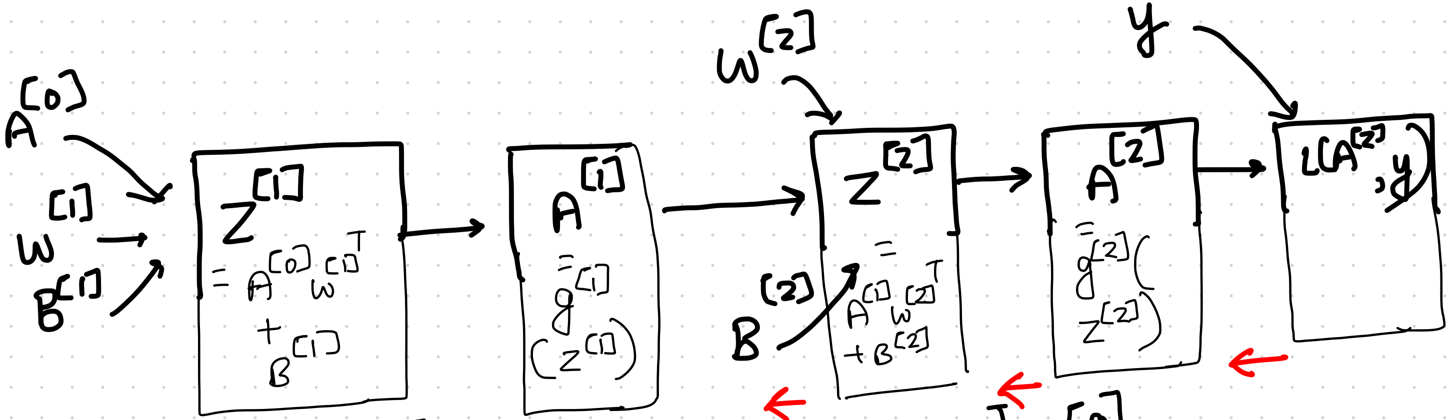
$1 \times (N \times N^{[2]}) \times (N^{[2]} \times N^{[1]})$

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

$$= \frac{\partial L}{\partial Z^{[1]}} W^{[2]} \odot g^{[1]'}(Z^{[1]})$$

Element wise or dot product of Jacobian

COMPUTATION GRAPH (FOR XOR EXAMPLE)



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

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