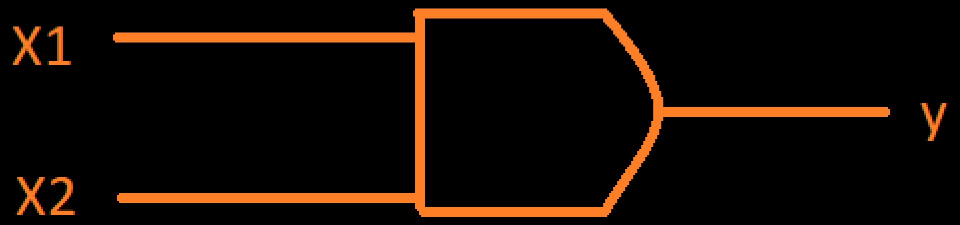
Activation Functions

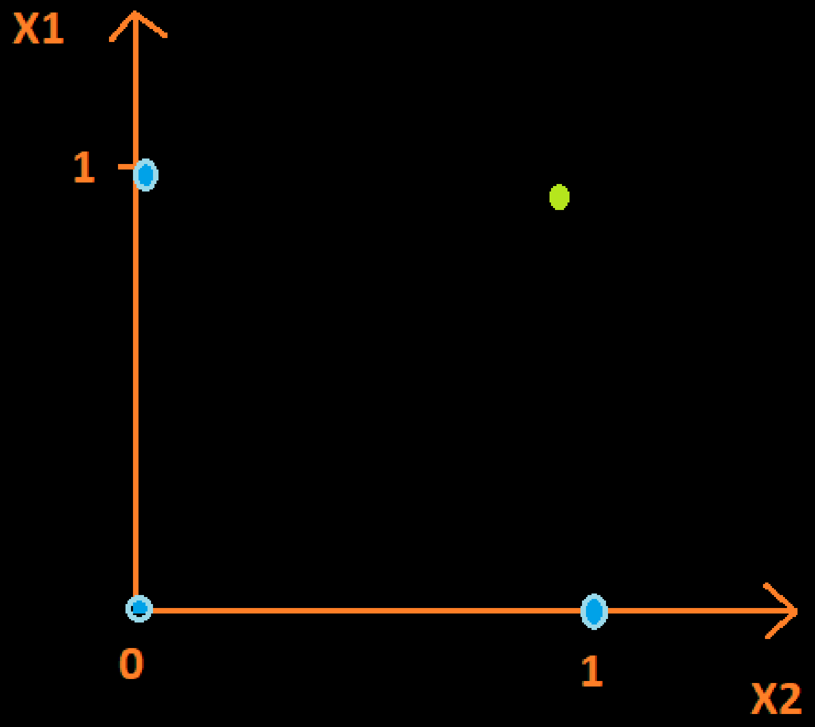
Classification by Perceptron

AND Gate Implementation



X_1	X_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Truth Table

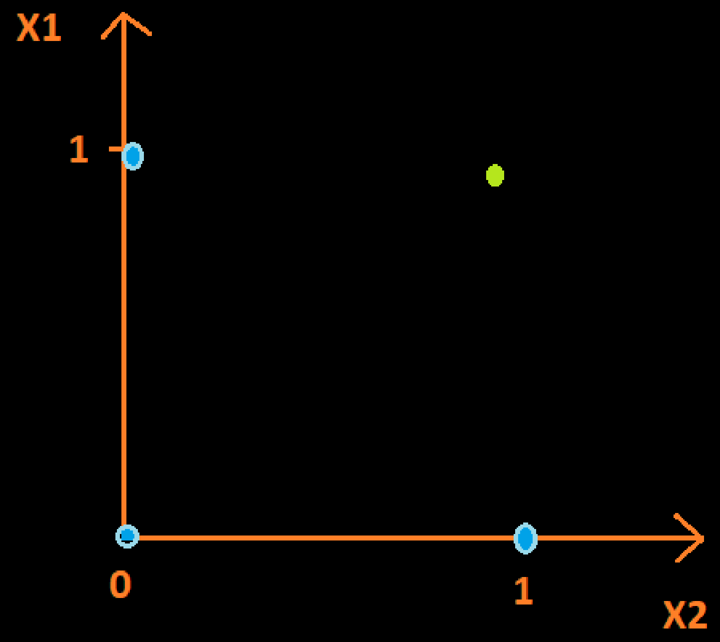


AND Gate in 2D

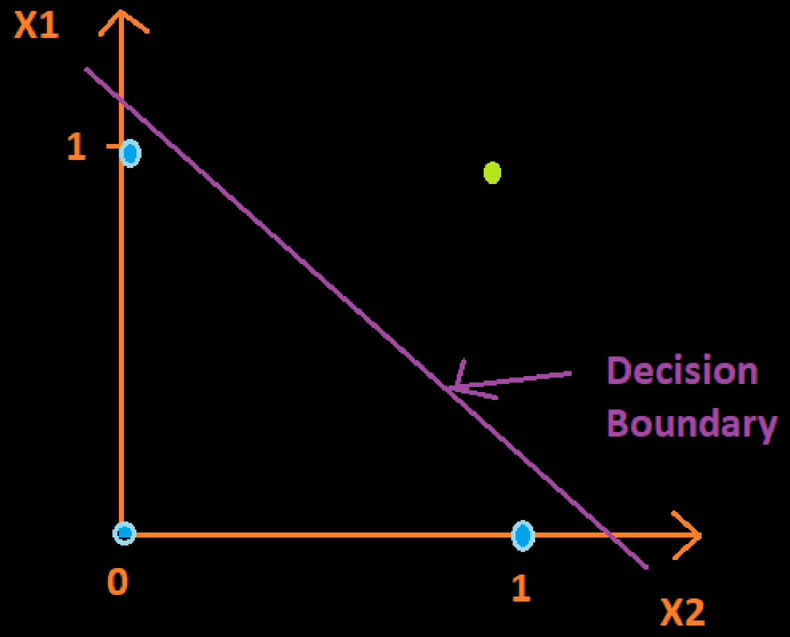
X_1	X_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Truth Table

AND Gate is Linearly separable

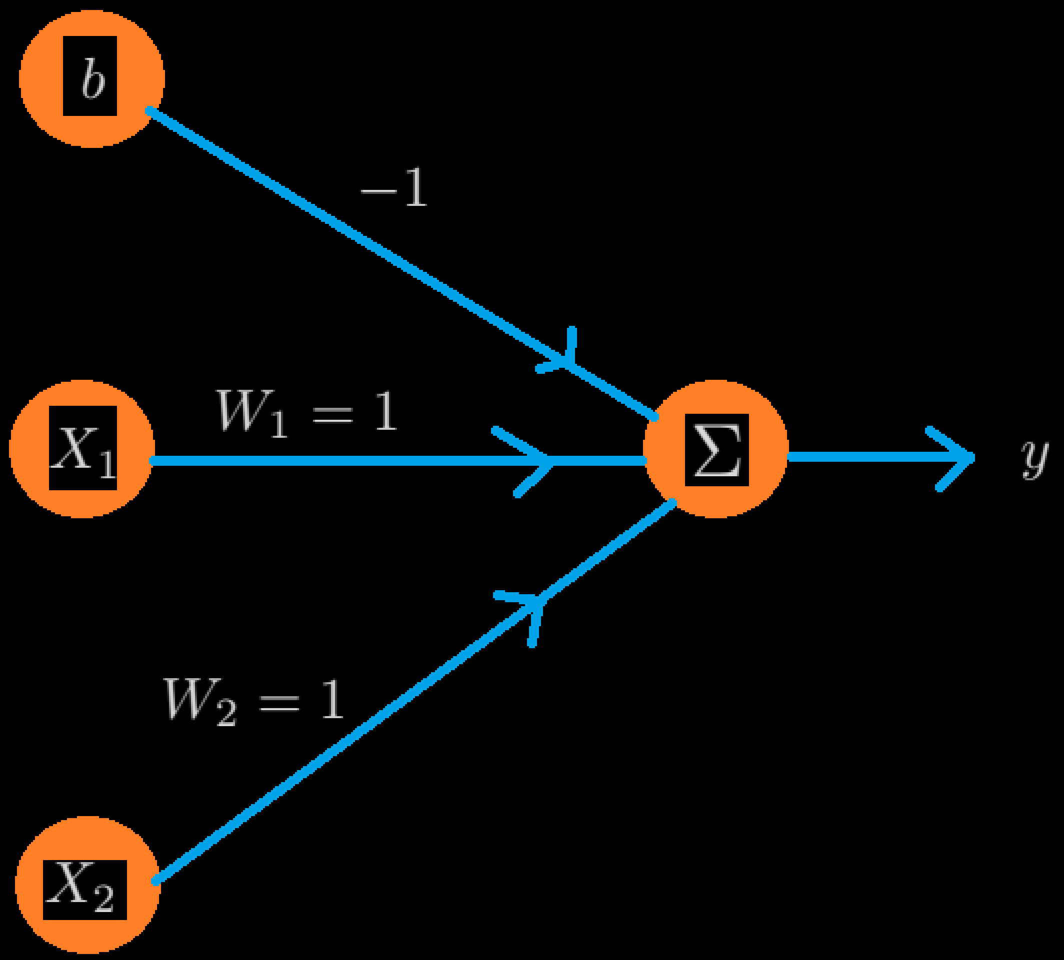


AND Gate in 2D



AND Gate is linearly
Separable

Implementation



$$\boxed{y = W_1X_1 + W_2X_2 + b}$$

The conditions for classification are

$$\boxed{\begin{array}{l} y = 1 \ \text{ if } \ W_1X_1 + W_2X_2 + b > 0 \\ \\ y = 0 \ \text{ if } \ W_1X_1 + W_2X_2 + b \leq 0 \end{array}}$$

Implementation

Case 1 : When the input is [0,0]

$$y = 0 + 0 - 1 = -1 < 0 = 0$$

Case 2 : When the input is [0,1]

$$y = 0 + 1 - 1 = 0$$

Case 3 : When the input is [1,0]

$$y = 1 + 0 - 1 = 0$$

Case 4 : When the input is [1,1]

$$y = 1 + 1 - 1 = 1 > 0 = 1$$

X_1	X_2	y
0	0	0
0	1	0
1	0	0
1	1	1

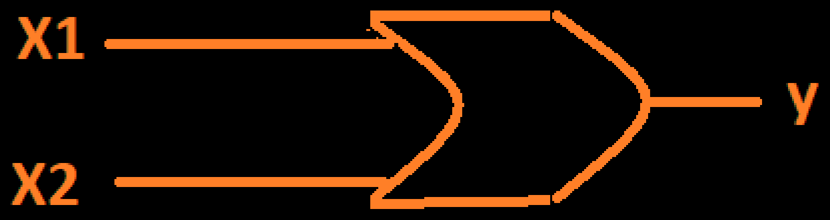
$$\boxed{y = W_1X_1 + W_2X_2 + b}$$

The conditions for classification are

$$y = 1 \text{ if } W_1X_1 + W_2X_2 + b > 0$$

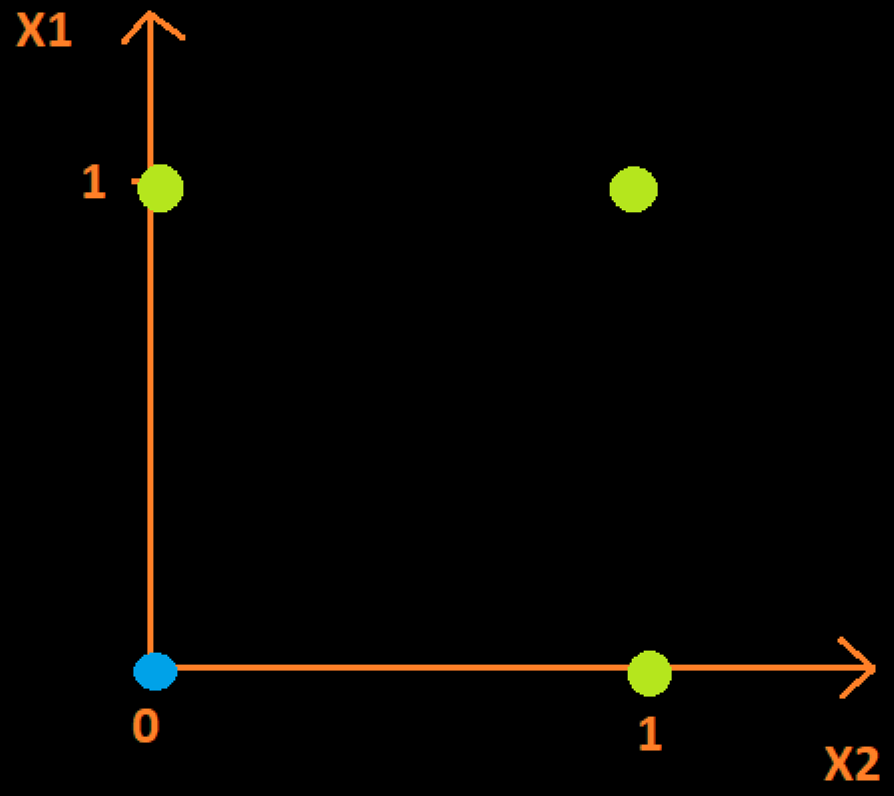
$$y = 0 \text{ if } W_1X_1 + W_2X_2 + b \leq 0$$

OR Gate Implementation



X_1	X_2	y
0	0	0
0	1	1
1	0	1
1	1	1

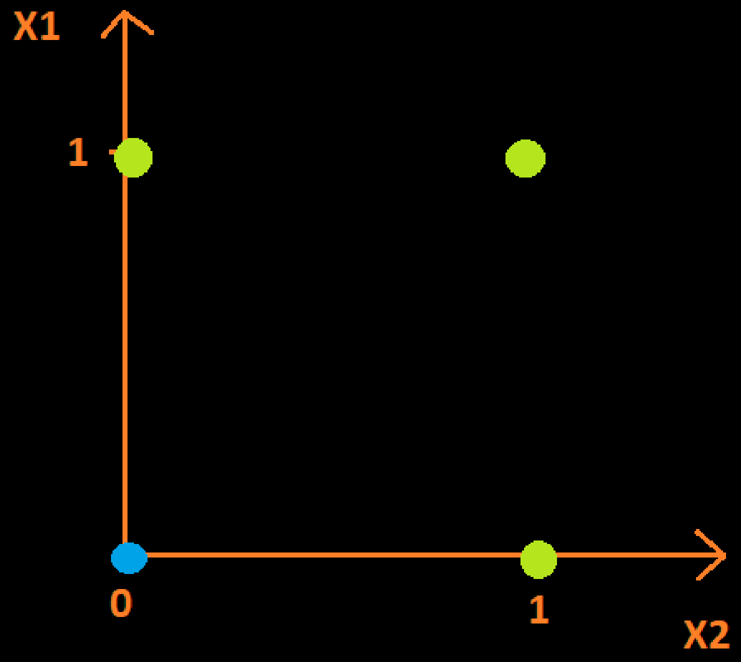
Truth Table

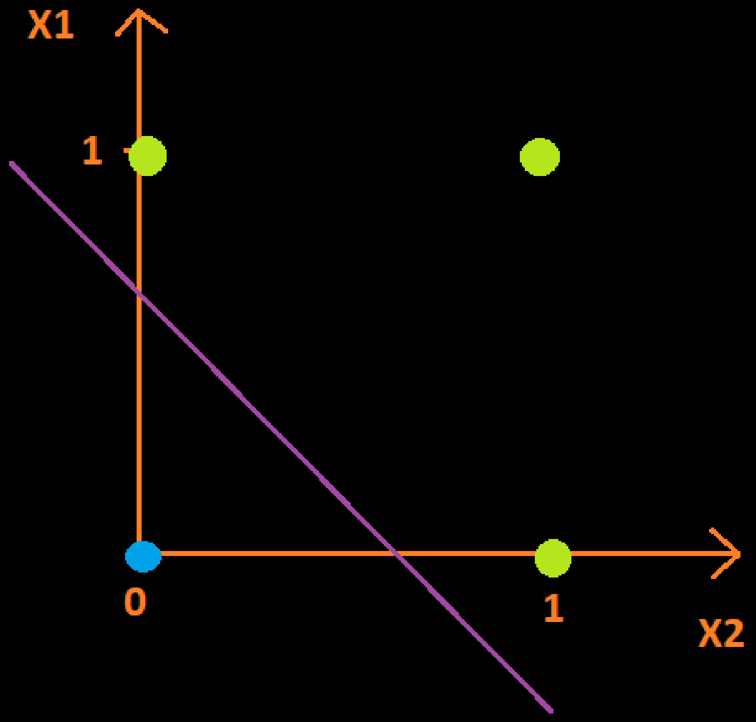


X_1	X_2	y
0	0	0
0	1	1
1	0	1
1	1	1

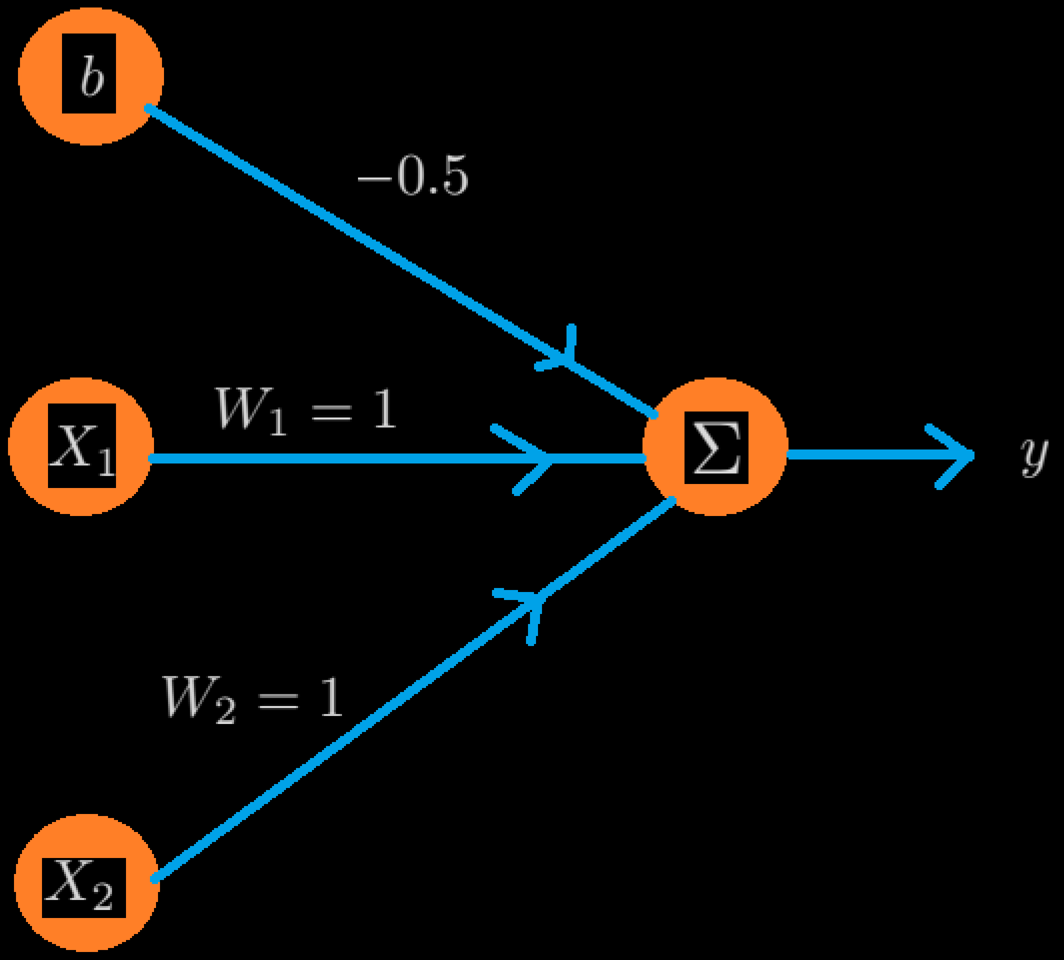
Truth Table

OR Gate is Linearly separable





Implementation



$$\boxed{y = W_1X_1 + W_2X_2 + b}$$

The conditions for classification are

$$\boxed{\begin{aligned} y &= 1 \;\text{ if }\; W_1X_1 + W_2X_2 + b > 0 \\ y &= 0 \;\text{ if }\; W_1X_1 + W_2X_2 + b \leq 0 \end{aligned}}$$

Implementation

Case 1 : When the input is [0,0]

$$y = 0 + 0 - 0.5 = -0.5 < 0 = 0$$

Case 2 : When the input is [0,1]

$$y = 0 + 1 - 0.5 = 0.5 > 0 = 1$$

Case 3 : When the input is [1,0]

$$y = 1 + 0 - 0.5 = 0.5 > 0 = 1$$

Case 4 : When the input is [1,1]

$$y = 1 + 1 - 0.5 = 1.5 > 0 = 1$$

X_1	X_2	y
0	0	0
0	1	1
1	0	1
1	1	1

$$\boxed{y = W_1X_1 + W_2X_2 + b}$$

The conditions for classification are

$$y = 1 \text{ if } W_1X_1 + W_2X_2 + b > 0$$

$$y = 0 \text{ if } W_1X_1 + W_2X_2 + b \leq 0$$

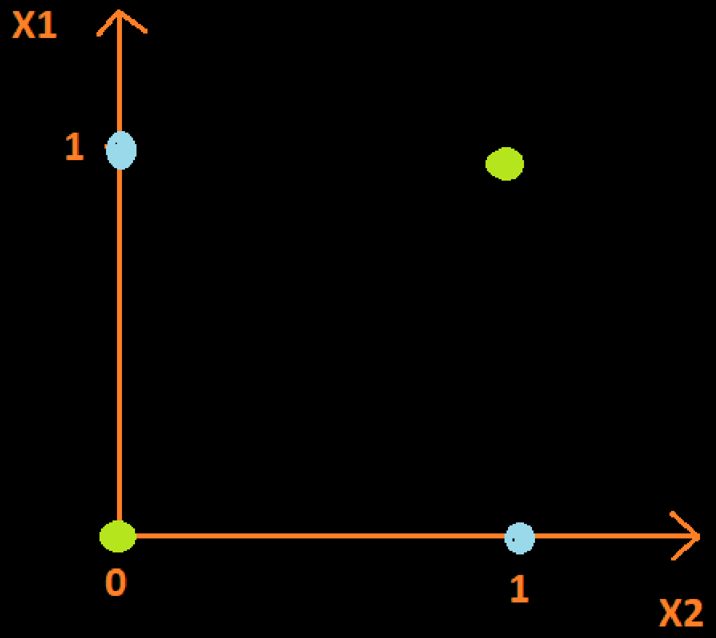
Activation Functions

XOR Gate

X_1	X_2	y
0	0	0
0	1	1
1	0	1
1	1	0

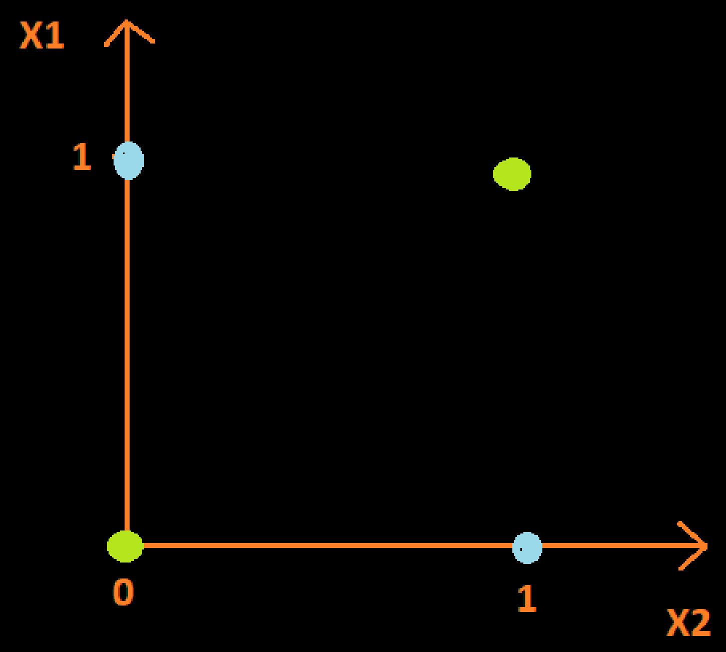
XOR Truth Table

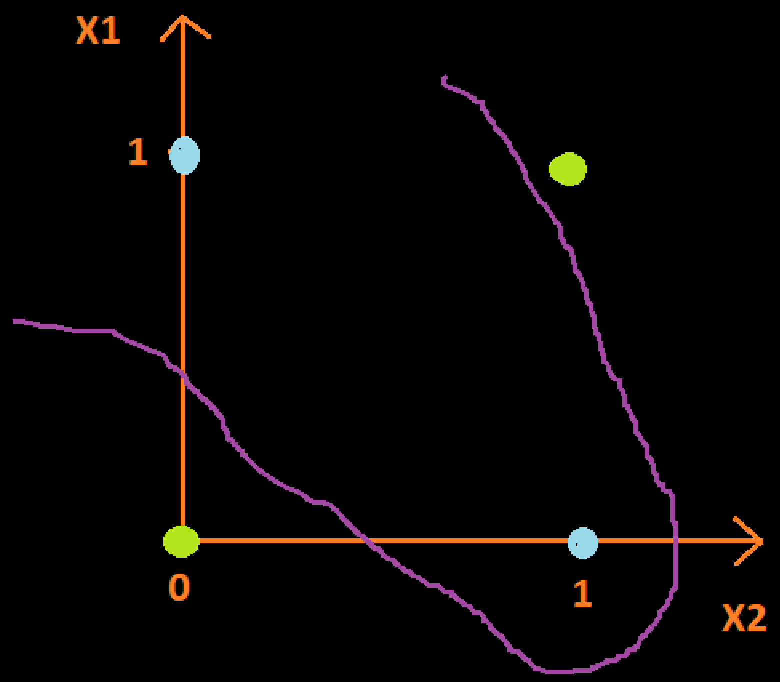
XOR Gate



X_1	X_2	y
0	0	0
0	1	1
1	0	1
1	1	0

XOR Gate



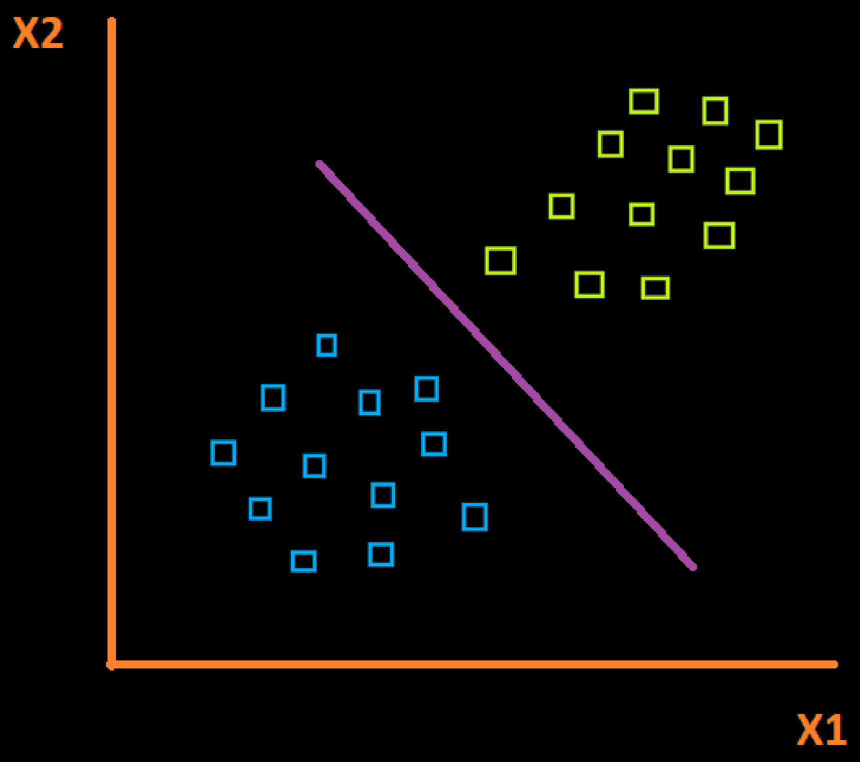


XOR Gate is not linearly Separable

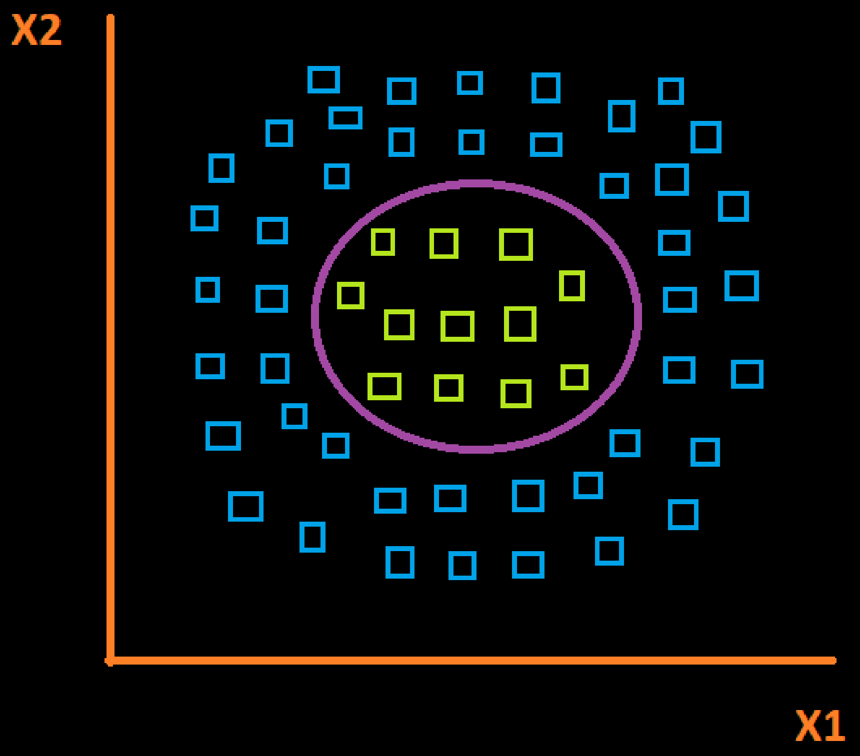
Activation Functions

- Our real world data is non-linear and is therefore linearly unseparable.
- We use activation functions to make our neural network to learn the non-linearity and complexity of data.

Activation Functions

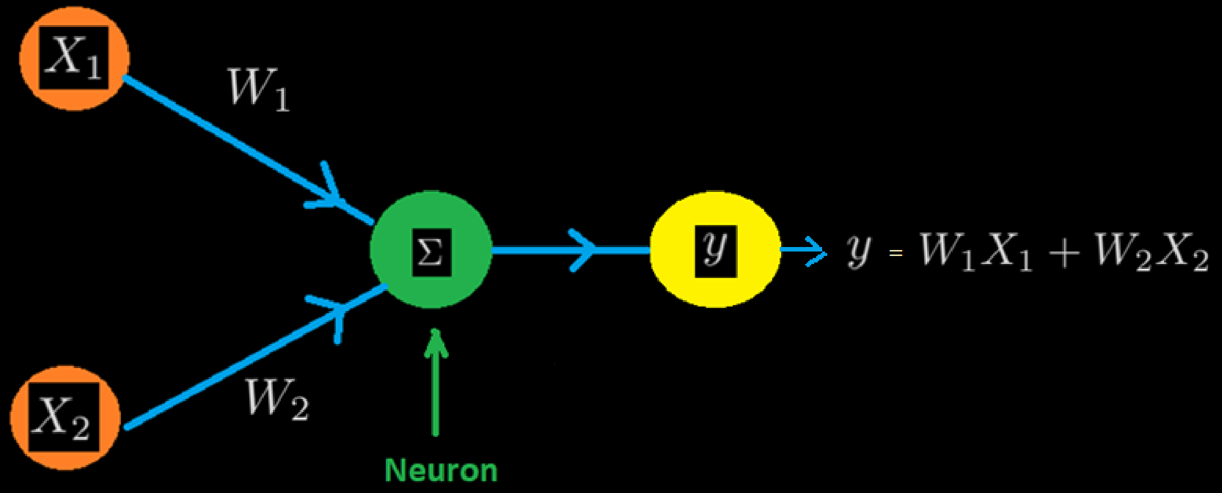


Linearly Separable



Not Linearly Separable

Activation Functions

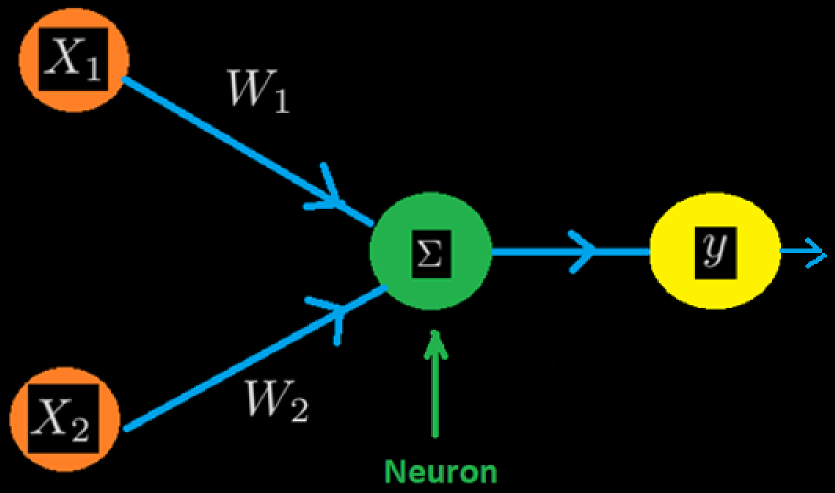


Multiple Linear Regression

$$y = w_0 + w_1X_1 + w_2X_2$$

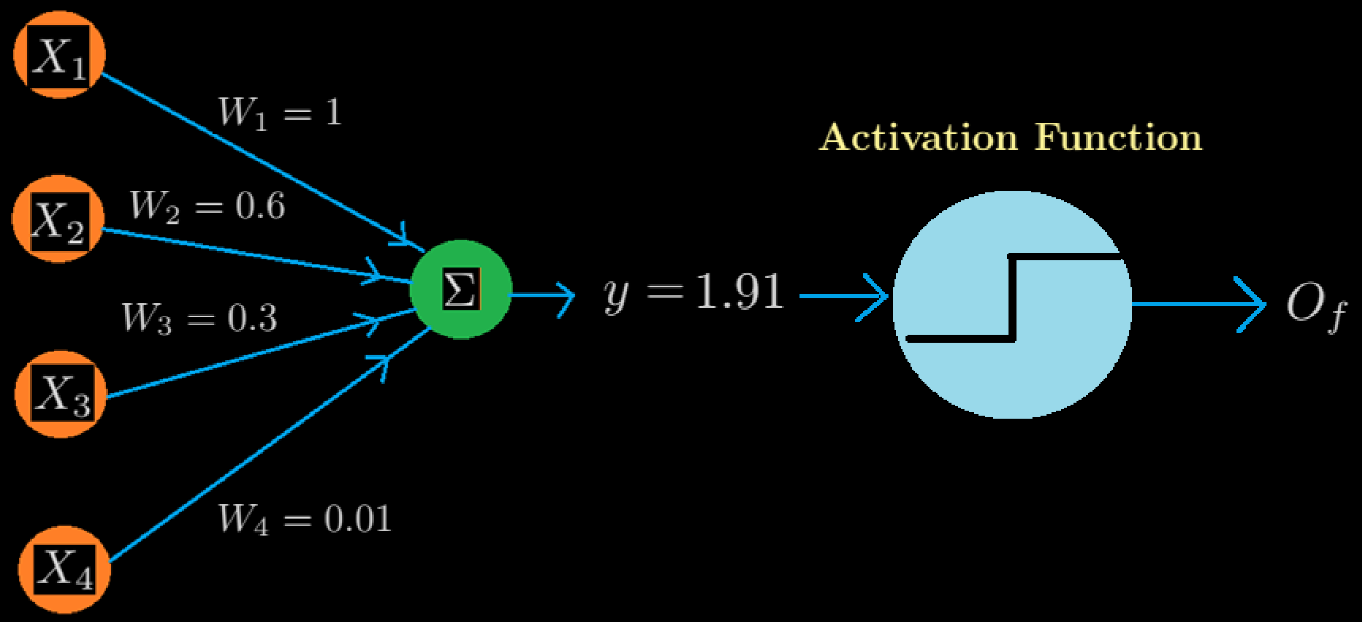
Adding Activation Functions
to
Neural Network

Activation Functions

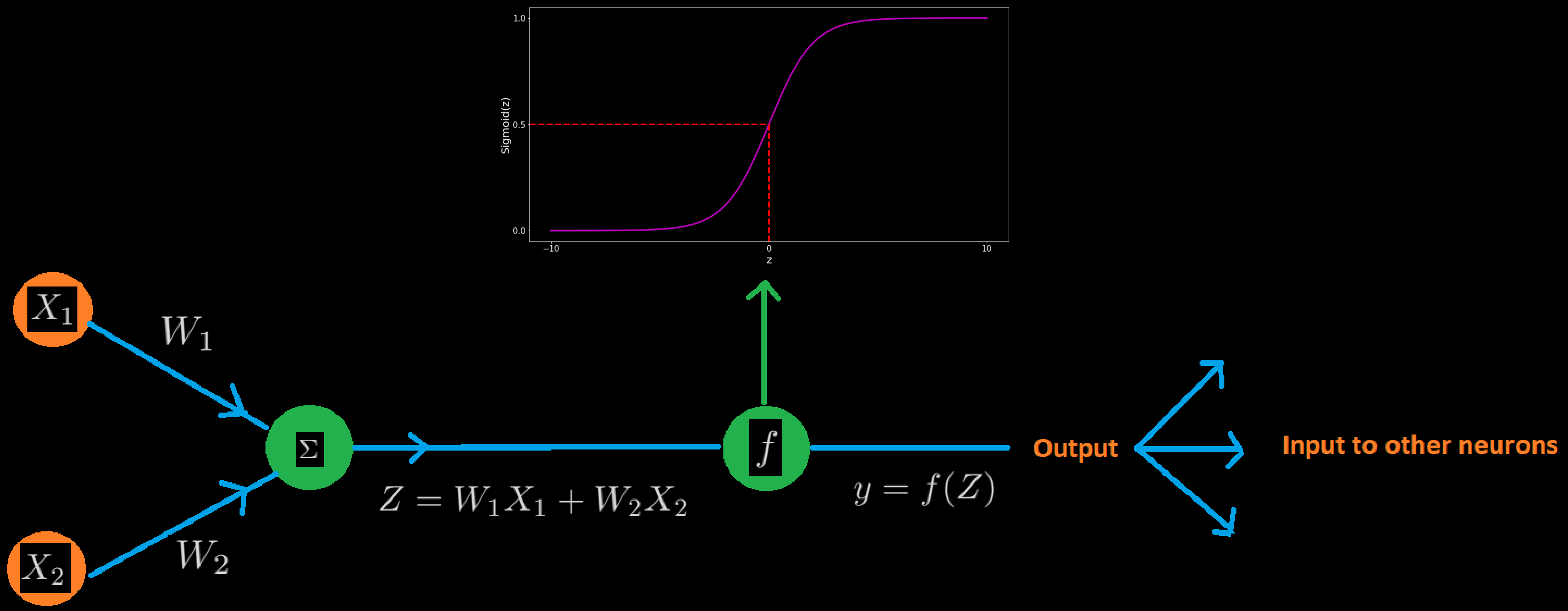


$y = W_1X_1 + W_2X_2$

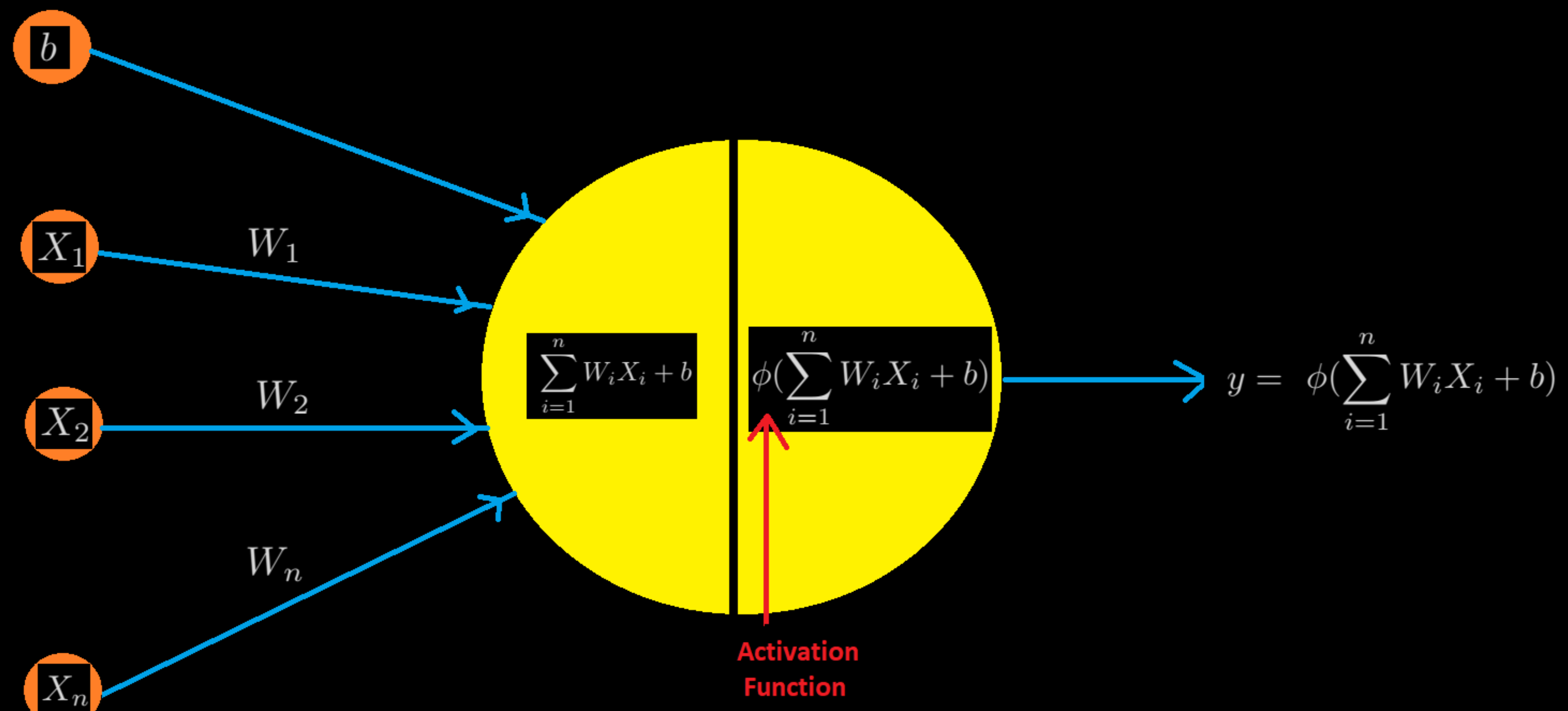
Activation Functions



Activation Functions



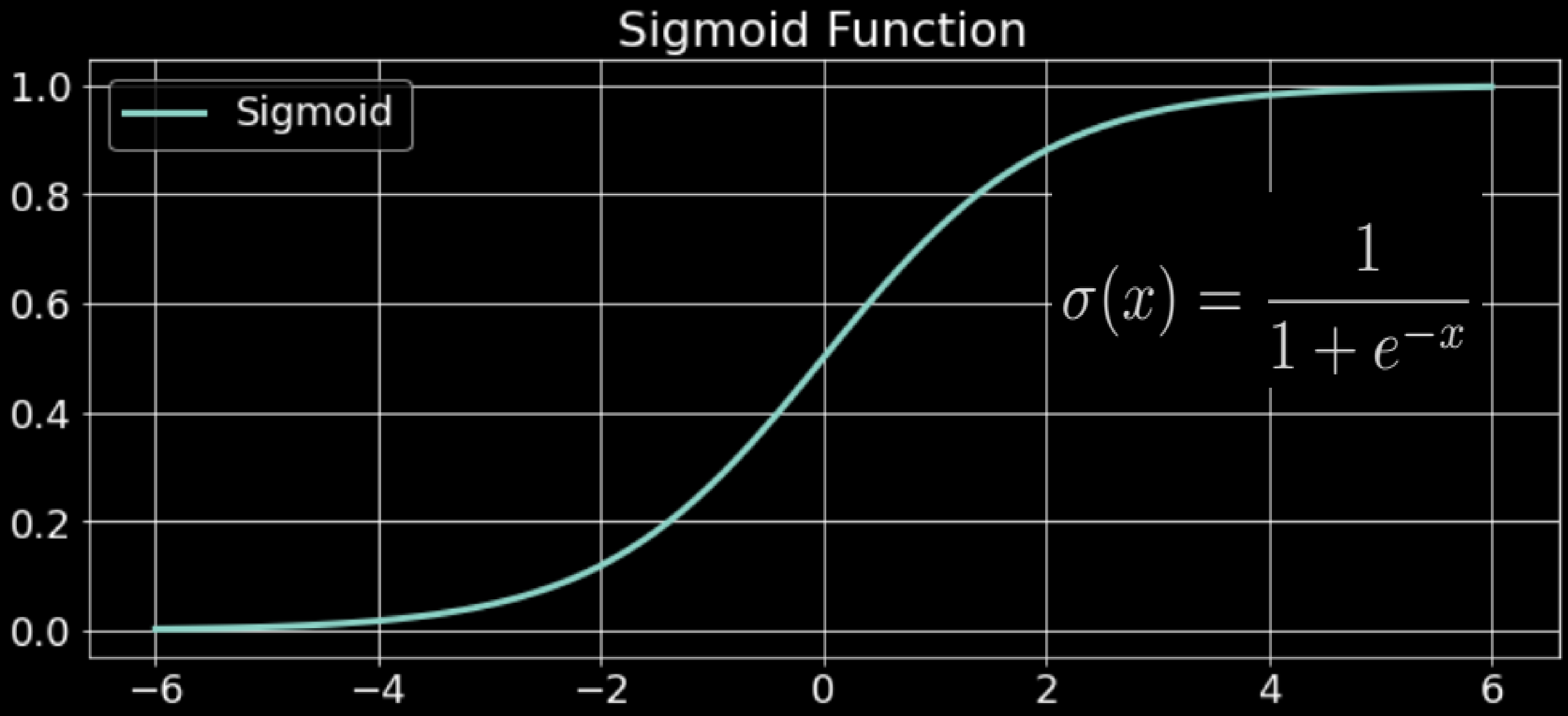
Activation Functions



Internal Visualization of Neuron

Sigmoid Function

Sigmoid Function



Maths of Sigmoid function

$$\boxed{\sigma(x) = \frac{1}{1+e^{-x}}}$$

$$\boxed{\text{for } x = 0}$$

$$\sigma(0) = \frac{1}{1+e^{0}}$$

$$\sigma(0) = \frac{1}{1+1}$$

$$\sigma(0) = \frac{1}{2}$$

$$\boxed{\sigma(0) = 0.5}$$

$$\boxed{\text{for } x = -\infty}$$

$$\sigma(-\infty) = \frac{1}{1+e^{\infty}}$$

$$\sigma(-\infty) = \frac{1}{1+\infty}$$

$$\sigma(-\infty) = \frac{1}{\infty}$$

$$\boxed{\sigma(-\infty) = 0}$$

$$\boxed{\text{for } x = \infty}$$

$$\sigma(\infty) = \frac{1}{1+e^{-\infty}}$$

$$\sigma(\infty) = \frac{1}{1+\frac{1}{e^{\infty}}}$$

$$\sigma(\infty) = \frac{1}{1+\frac{1}{\infty}}$$

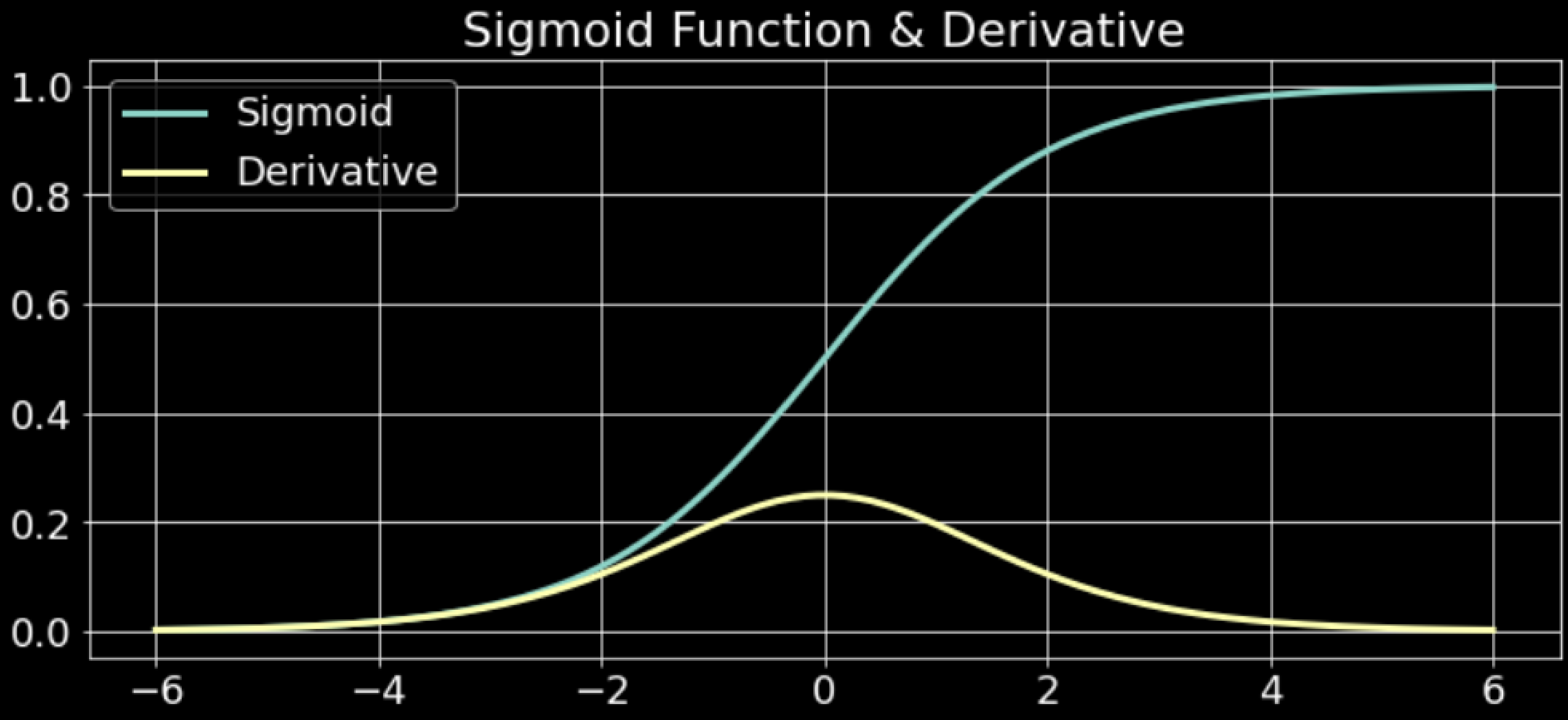
$$\sigma(\infty) = \frac{1}{1+0}$$

$$\boxed{\sigma(\infty) = 1}$$

Properties of Sigmoid function

- It is not good for hidden layers especially when the there are more than two layers in a neural network.
- Its minimum value is zero and maximum value is 1, therefore, it is good for classification layer. it gives us the classification probabilities of classes.
- Sigmoid has the biased average. This means that the average of its input is zero but the average of its output is 0.5

Sigmoid and Its Derivative



Problem With Sigmoid Function

- The magnitude of the derivative of sigmoid is very small.
- Small magnitude can cause vanishing gradient problem in neural networks because of continuous multiplication of gradient terms during learning.

Example

$$0.25 \times 0.25 \times 0.25 = 0.016$$

Exploding Gradient

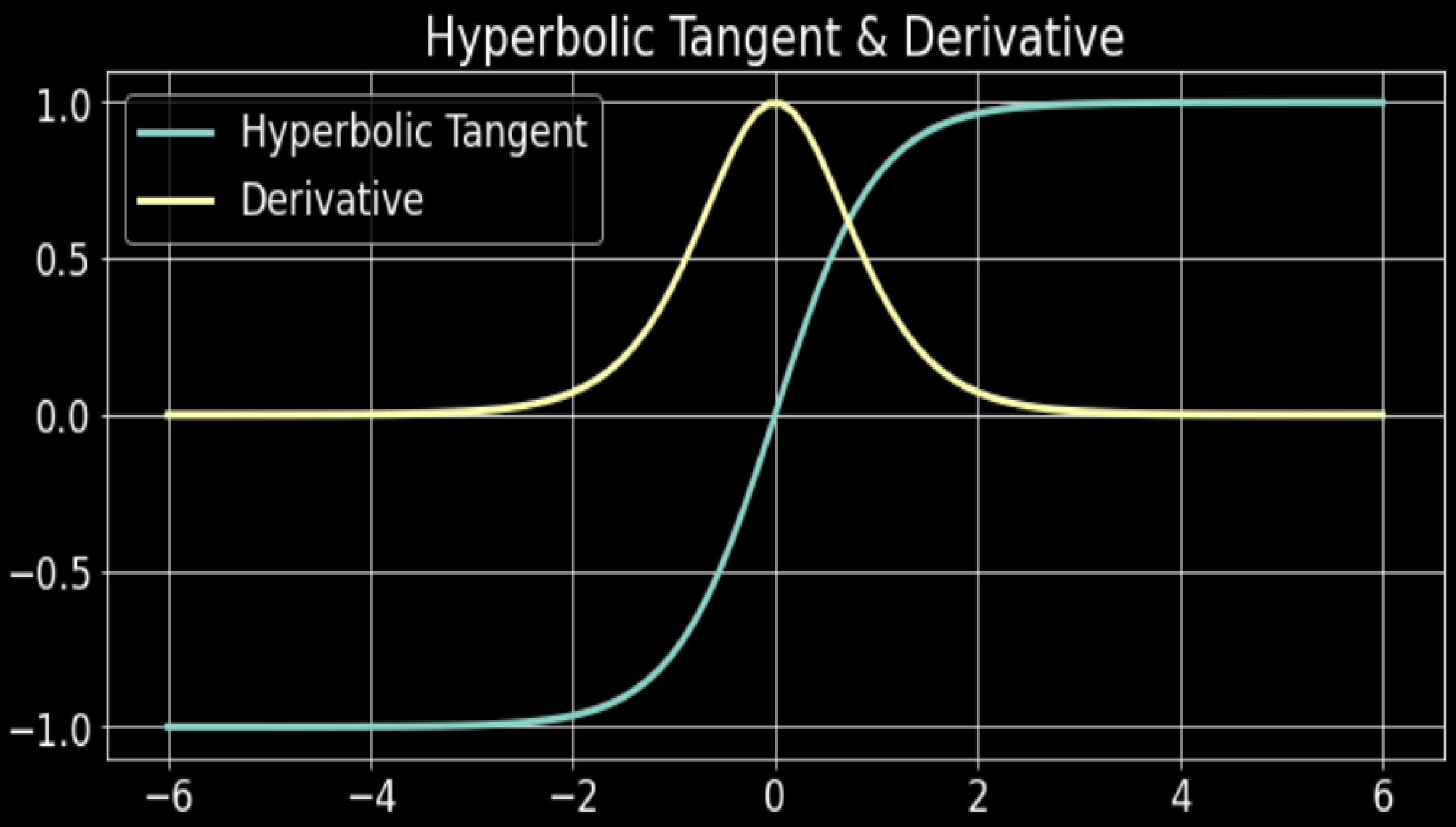
On the otherhand, there are cases when the gradient keep on getting larger and larger during backpropagating from output to input layers. As a result, gradient descent diverges and the values of updated weights become very large. This problem is called the exploding gradient problem.

Example

$1.4 \times 1.4 \times 1.4 = 2.74$

Hyperbolic Tangent Function

Hyperbolic Tangent Function



$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

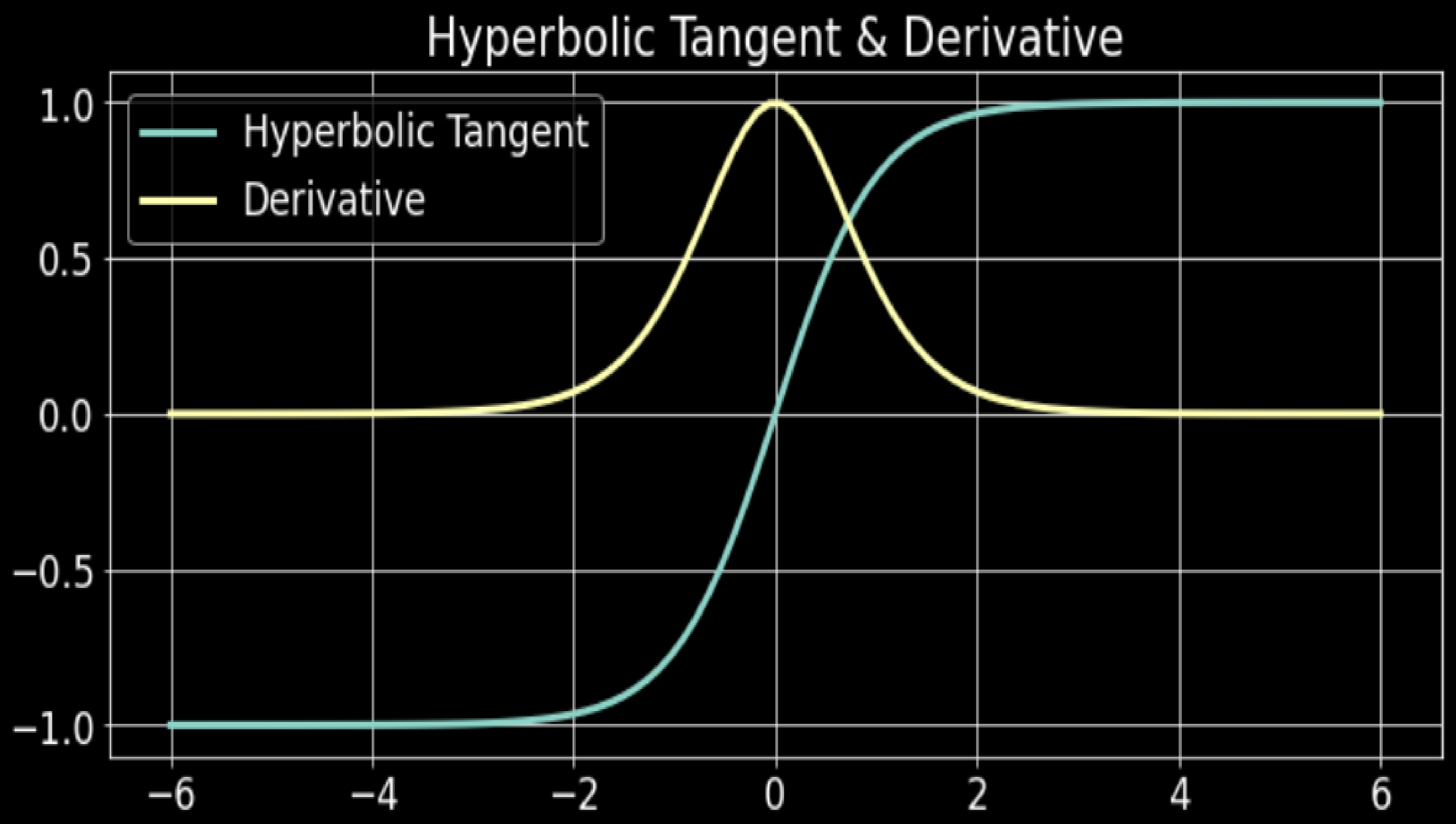
$$\tanh'(x) = 1 - tanh(x)^2$$

Properties of Hyberbolic Tangent Function

- It is somehow okay for hidden layers.
- Its minimum value is -1 and maximum value is 1.
- It doesn't have the biased average. This means that the average of its input and output is zero.
- The advantage of having unbiased average or mean is that when this activation is used with the hidden layers, the mean for the hidden layer comes out be 0 or very close to 0, hence tanh functions helps in centering the data by bringing mean close to 0 which makes learning for the next layer much easier and thus the training of Neural Network becomes faster.

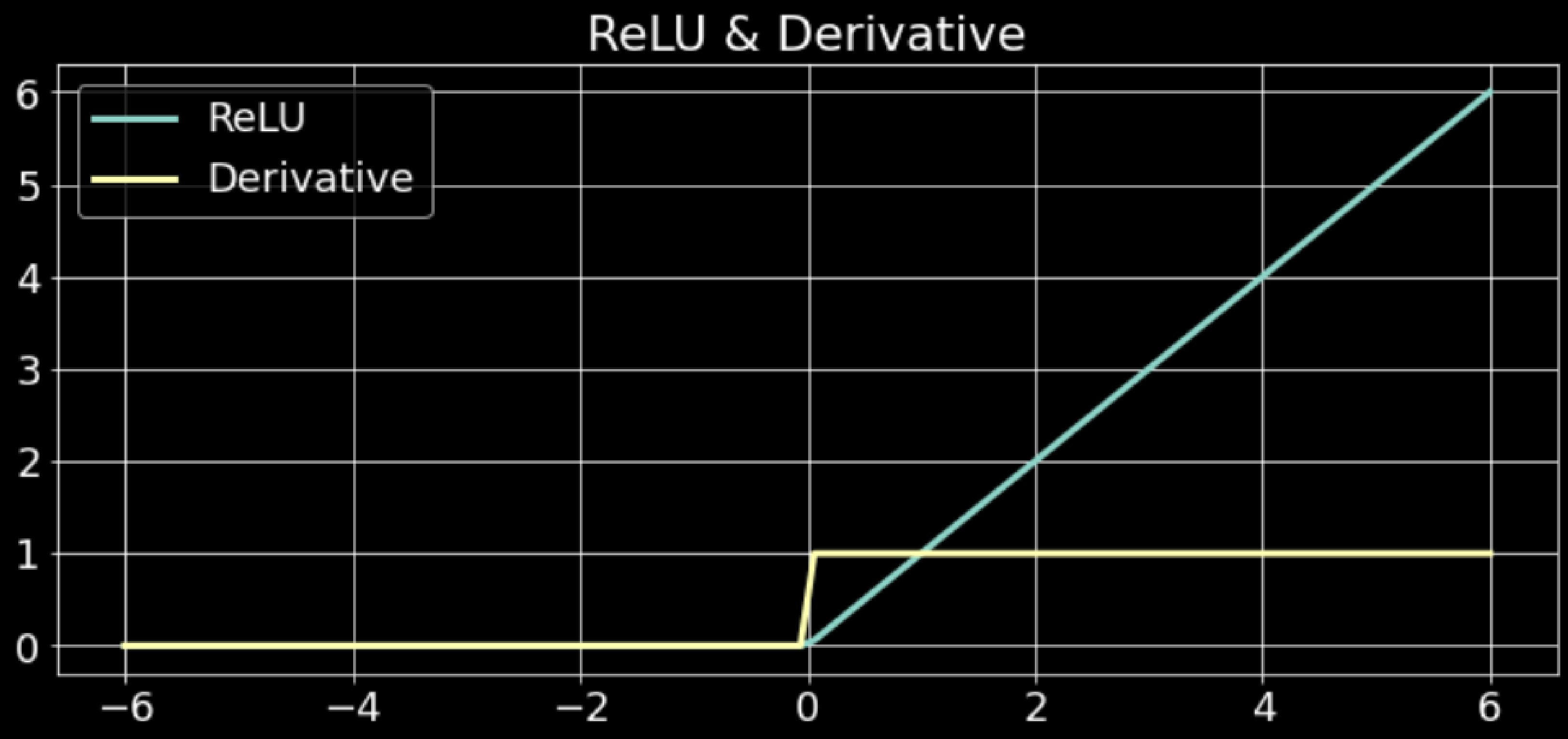
Problem With Hyperbolic Tangent

- Gradient at the tails of -1 and 1 are almost zero.



ReLU and Leaky ReLU

Rectified Linear Unit (ReLU) Function



$$ReLU(x) = max(0, x)$$

0.77	-0.11	0.11	0.33	0.55	-0.11	0.33
-0.11	1.00	-0.11	0.33	-0.11	0.11	-0.11
0.11	-0.11	1.00	-0.33	0.11	-0.11	0.55
0.33	0.33	-0.33	0.55	-0.33	0.33	0.33
0.55	-0.11	0.11	-0.33	1.00	-0.11	0.11
-0.11	0.11	-0.11	0.33	-0.11	1.00	-0.11
0.33	-0.11	0.55	0.33	0.11	-0.11	0.77

Relu Function

0.77	0	0.11	0.33	0.55	0	0.33
0	1.00	0	0.33	0	0.11	0
0.11	0	1.00	0	0.11	0	0.55
0.33	0.33	0	0.55	0	0.33	0.33
0.55	0	0.11	0	1.00	0	0.11
0	0.11	0	0.33	0	1.00	0
0.33	0	0.55	0.33	0.11	0	0.77

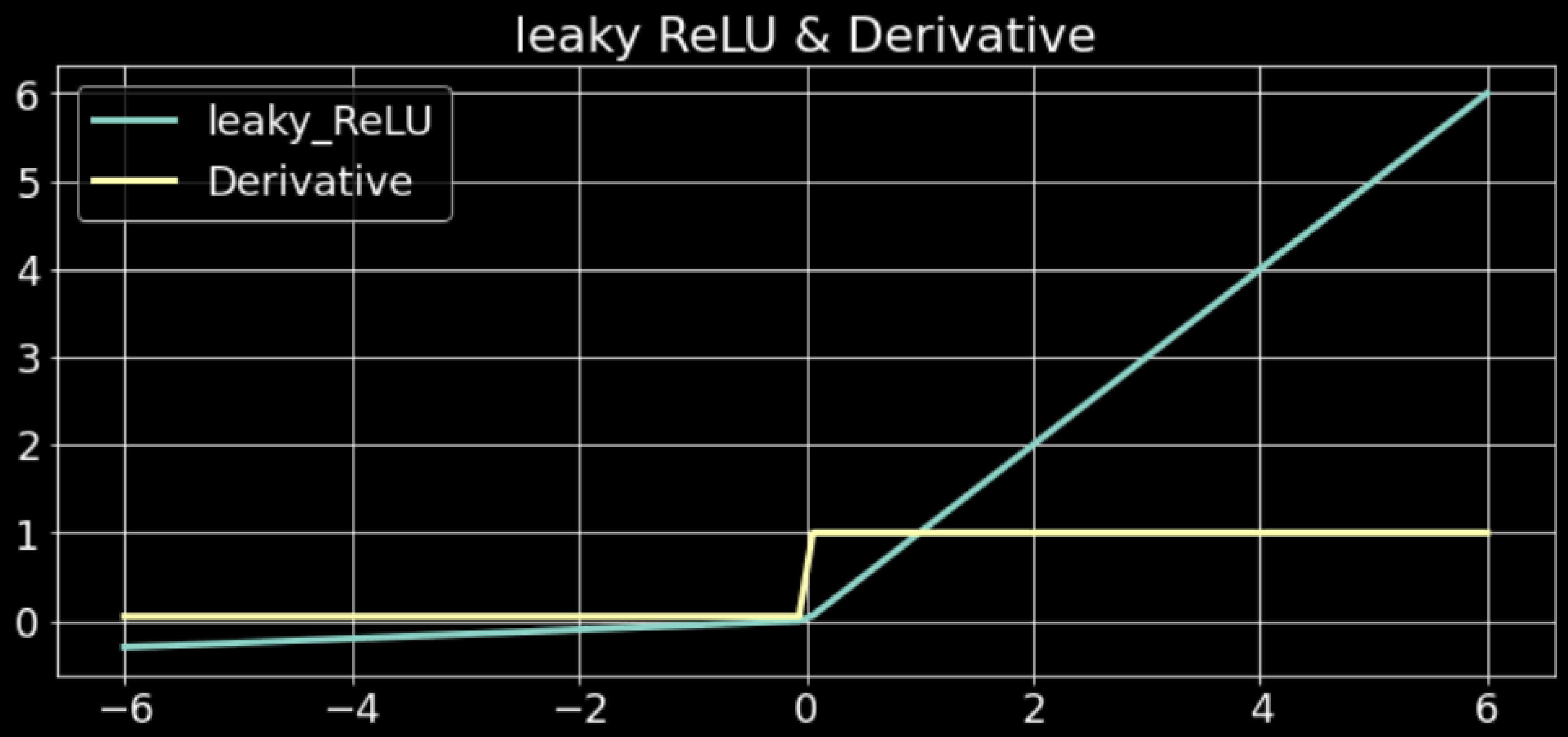
$$ReLU(x) = \begin{cases} x & \text{for } x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$ReLU'(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

Properties of ReLU Function

- It is a piece wise linear function. However, the discontinuity at $x = 0$ makes it Non linear.
- It preserves the positive input and transform negative input to zero.
- It has a linear slope of magnitude one after $x = 0$
- The derivative of the ReLU is 1 in the positive part, and 0 in the negative part.

Leaky ReLU Function



$$ReLU(x) = max(\alpha x, x)$$

where $\alpha = 0.01$

$$ReLU(x) = \begin{cases} x & \text{for } x \geq 0 \\ \alpha x & \text{for } x < 0 \end{cases}$$

$$ReLU'(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ \alpha & \text{for } x < 0 \end{cases}$$

Properties of Leaky ReLU Function

- It is a piece wise linear function. However, the discontinuity at $x = 0$ makes it Non linear.
- It preserves the positive input and transform negative input to a very small fraction.
- It has a linear slope of magnitude one after $x = 0$
- The derivative of the Leaky ReLU is 1 in the positive part, and is a small fraction in the negative part.

Thank you !

Thank you !