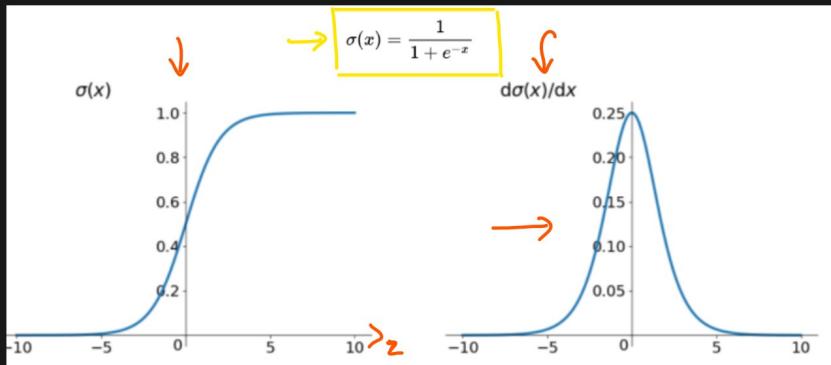


Activation Functions

① Sigmoid Activation function $[0 \text{ to } 1]$ $z = \sum_{i=1}^n w_i^T x + b$

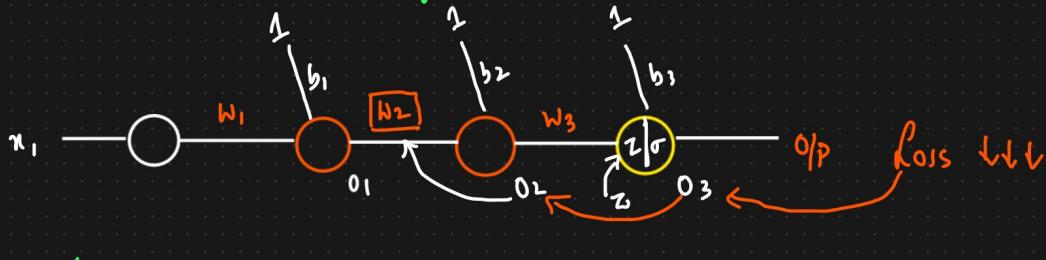


$$\sigma(z) \Rightarrow 0 \text{ to } 1$$

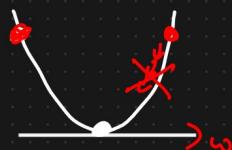
$$\phi(z) \Rightarrow$$

$$\frac{\partial \sigma(z)}{\partial z} = 0 \text{ to } 0.25$$

Forward Propagation



Backward Propagation.



$$w_{2\text{new}} = w_{2\text{old}} - \eta \left[\frac{\partial h}{\partial w_{2\text{old}}} \right] \quad \boxed{0.001} \quad \boxed{0.0001} \quad \Rightarrow w_{2\text{new}} \approx w_{2\text{old}}$$

$$\frac{\partial h}{\partial w_{2\text{old}}} = \frac{\partial h}{\partial o_3} \neq \boxed{\frac{\partial o_3}{\partial o_2}} * \frac{\partial o_2}{\partial w_2}$$

$$0.20 \downarrow * 0.01 \times \det z = (o_2 * w_3) + b_3$$

$$\frac{\partial o_3}{\partial o_2} = \frac{\partial (\sigma(z))}{\partial z} * \frac{\partial z}{\partial o_2} \quad [0 \text{ to } 1]$$

$$= [0 - 0.25] * \frac{\partial [(o_2 * w_3) + b_3]}{\partial o_2}$$

$$= [0 - 0.25] * w_3 \Rightarrow \text{Small value} \Rightarrow$$

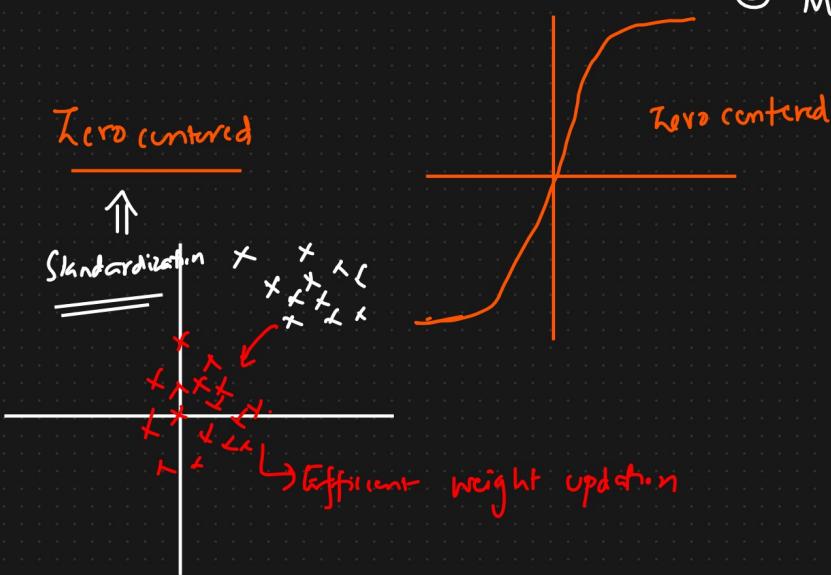
Advantages

- ① Binary Classification Suitable.
- ② clear prediction i.e. very close 1 or 0

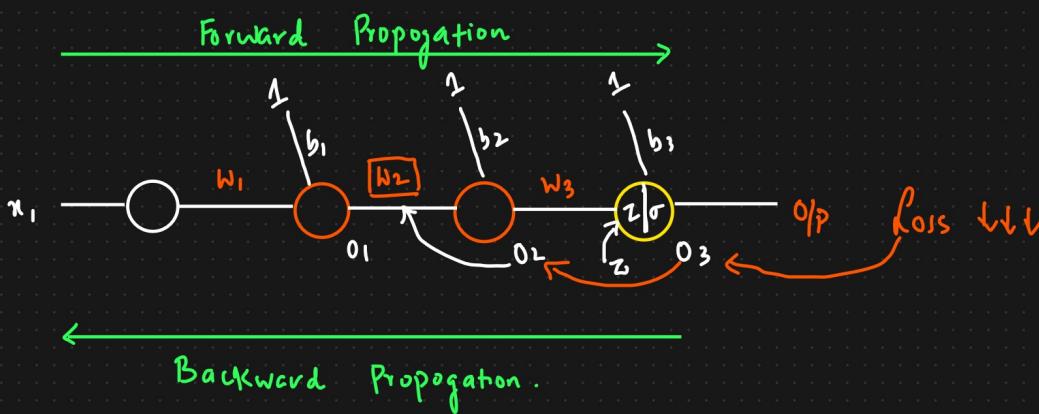
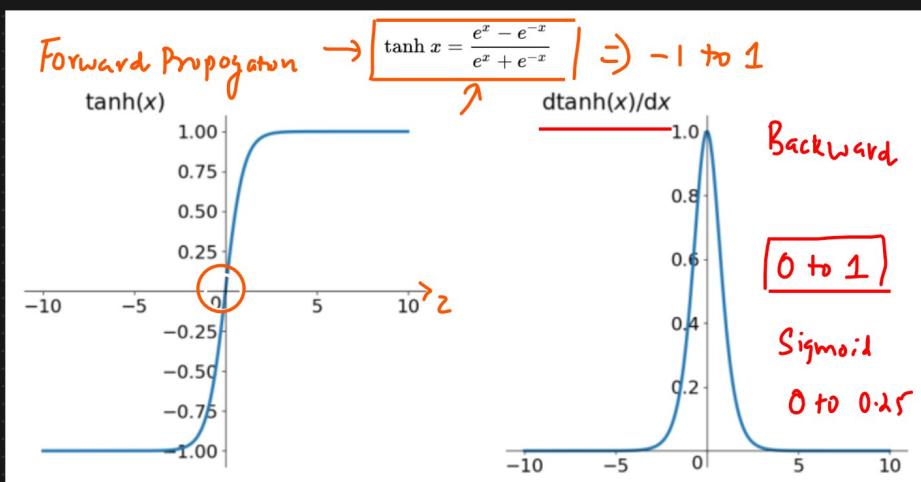
Disadvantages

- ① Prone to vanishing Gradient Problem.
- ② Function output is not zero centered \Rightarrow Efficient weight update
- ③ Mathematical operation are relatively time consuming

Zero centered



② Tanh Activation Function



$$\frac{\partial h}{\partial w_{2012}} = \frac{\partial h}{\partial o_3} \neq \boxed{\frac{\partial o_3}{\partial o_2}} * \frac{\partial o_2}{\partial w_2}$$

↓↓↓

$$0 \cdot 20_{11} * 0 \cdot 01 \times \text{det } z = (o_2 * w_3) + b_3$$

$$\begin{aligned}\frac{\partial o_3}{\partial o_2} &= \boxed{\frac{\partial (\tanh(z))}{\partial z}} * \frac{\partial z}{\partial o_2} && [0 \text{ to } 1] \\ &= [0 - 1] * \frac{\partial [(o_2 * w_3) + b_3]}{\partial o_2} \\ &= [0 - 1] * w_3 \Rightarrow \text{Small value} \Rightarrow\end{aligned}$$

Advantages

① Zero Centric \Rightarrow weight updation is efficient

Disadvantages

① Prone to Vanishing Gradient Problem

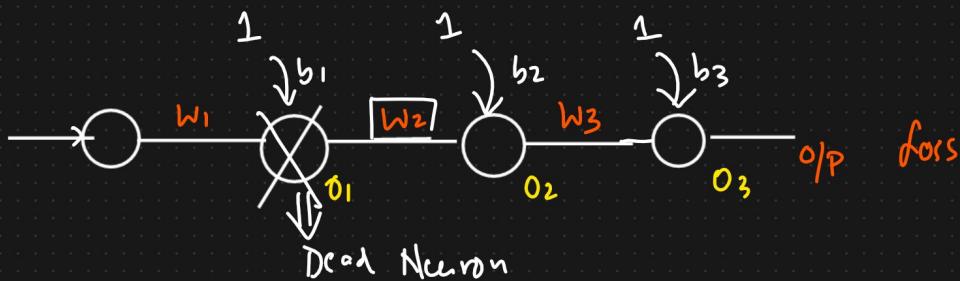
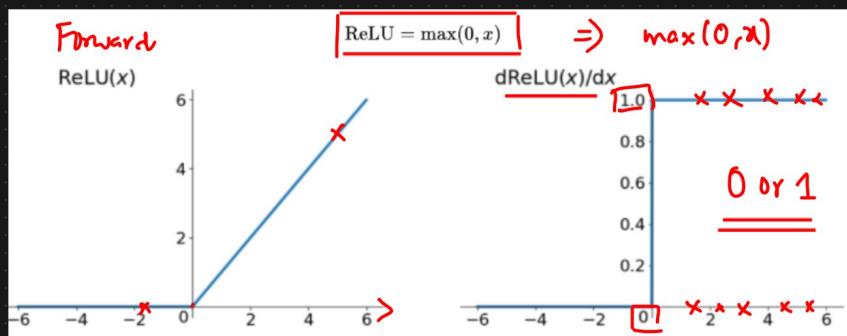
② Train Complexity

[Rectified Linear Unit].

③ ReLU Activation Function

Tanh $\Rightarrow 0 \text{ to } 1$

Sigmoid $\Rightarrow 0 \text{ to } 0.25$



$$\frac{\partial h}{\partial w_{201d}} = \frac{\partial h}{\partial o_3} * \boxed{\frac{\partial o_3}{\partial o_2}} * \frac{\partial o_2}{\partial w_2}$$

$\downarrow \downarrow \downarrow \downarrow$

$$0.20_{11} * 0.01 * \text{let } z = (o_2 * w_3) + b_3$$

$$\frac{\partial o_3}{\partial o_2} = \boxed{\frac{\partial (\text{relu}(z))}{\partial z}} * \frac{\partial z}{\partial o_2}$$

$[0 \rightarrow 1]$

$$= [0 \text{ or } 1] * \frac{\partial [(o_2 * w_3) + b_3]}{\partial o_2}$$

$$= \begin{bmatrix} \text{-ve} & \text{+ve} \end{bmatrix} * w_3 \Rightarrow \text{Small value} \Rightarrow$$

If ReLU output is $\boxed{1} \Rightarrow$ Weight updation will happen

If ReLU output is $\boxed{0} \Rightarrow$ Dead Neuron

$$w_{2\text{new}} = w_{201d} - \eta \boxed{\frac{\partial h}{\partial w_{201d}}} \Rightarrow 0$$

If Derivative of $\text{ReLU}(z)$ is 0

$$\boxed{w_{2\text{new}} \approx w_{201d}} \Rightarrow \text{Dead Neuron}$$

If $z = \text{+ve}$ $\frac{\partial \text{ReLU}(z)}{\partial z} = 1$

If $z = \text{-ve}$ $\frac{\partial \text{ReLU}(z)}{\partial z} = 0$

Advantages

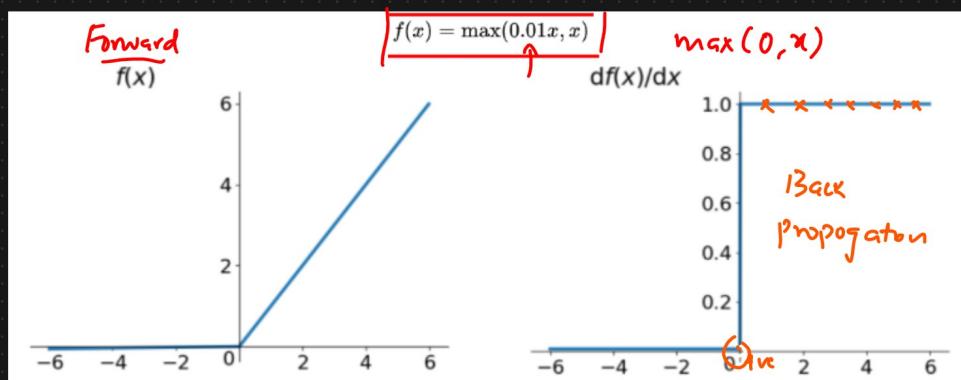
- ① Solving Vanishing Gradient Problem
- ② $\text{Max}(0, x) \rightarrow$ Calculation is Superfast. The ReLU function has a linear relationship.
- ③ It is much faster than Sigmoid or Tanh.

Disadvantages

- ① Dead Neuron
- ② ReLU function O/P $(0, x) \Rightarrow 0$ or zero number
↓
It is not zero centric

④ Leaky ReLU And Parametric ReLU

$\max(\lambda x, x)$ → hyperparameter $\lambda = \alpha = 0.01, 0.02, \dots, 0.03$



ReLU → Dead Neuron → Dead ReLU Problem

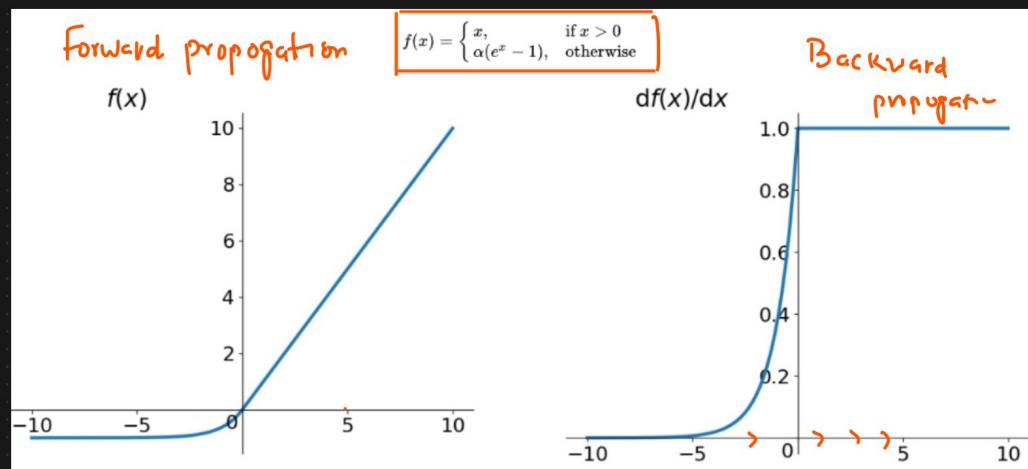
Advantages

- ① Leaky ReLU has all the advantages of ReLU
- ② It removes the Dead ReLU Problem

Disadvantage

- ① It is not zero centric

⑤ ELU (Exponential Linear Units)



Advantages

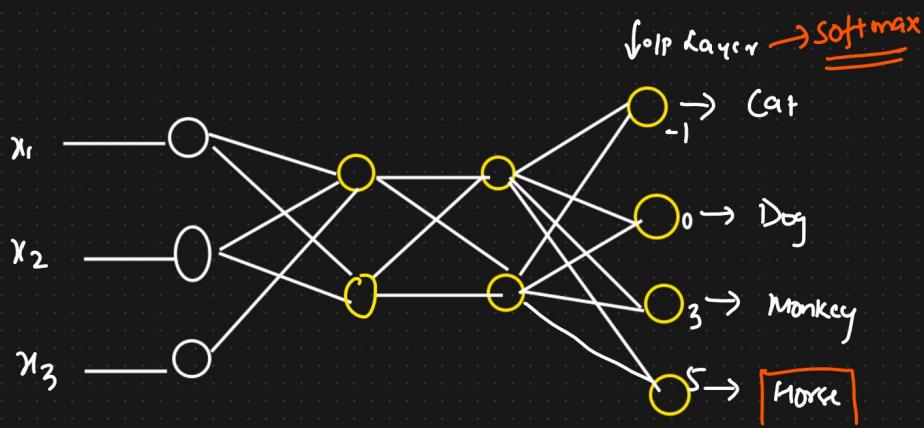
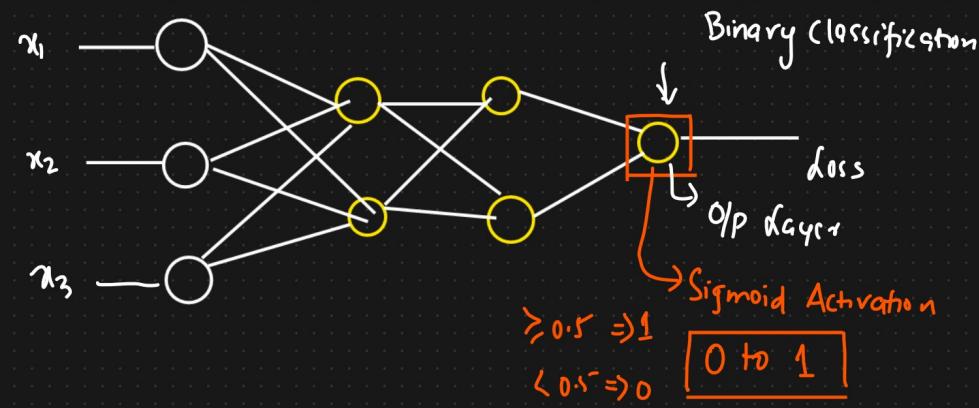
[It is used to solve
ReLU problems]

Disadvantage

- ① No Dead ReLU Issues
- ② Zero centered

i) Slightly more computationally intensive.

⑥ Softmax Activation function [Multiclass classification problem]



$$\text{Softmax} = \frac{e^{y_i}}{\sum_{k=0}^n e^{y_k}}$$

$$y_i = \theta \cdot w + b$$

$$\text{Softmax} \Rightarrow \text{Cat} = \frac{e^{-1}}{e^{-1+0+3+5}} = 0.00033 \quad \Pr(\text{Horse}) = \frac{0.1353}{0.00033 + 0.0024 + 0.0183 + 0.1353}$$

$$\text{Dog} = \frac{e^0}{e^{-1+0+3+5}} = 0.0024 \quad \approx 86\%$$

$$\text{Monkey} = \frac{e^3}{e^{-1+0+3+5}} = 0.0183$$

$$\text{Horse} = \frac{e^5}{e^{-1+0+3+5}} = 0.1353$$

⑦ Which Activation Function To Use When?

