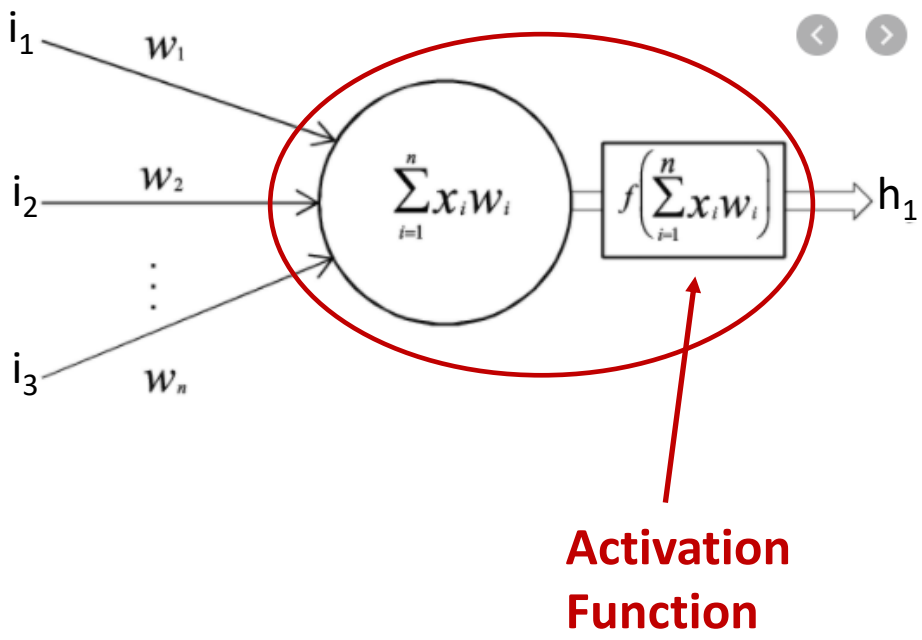
Lesson 13

Activation Functions and Their Derivatives

Activation Functions



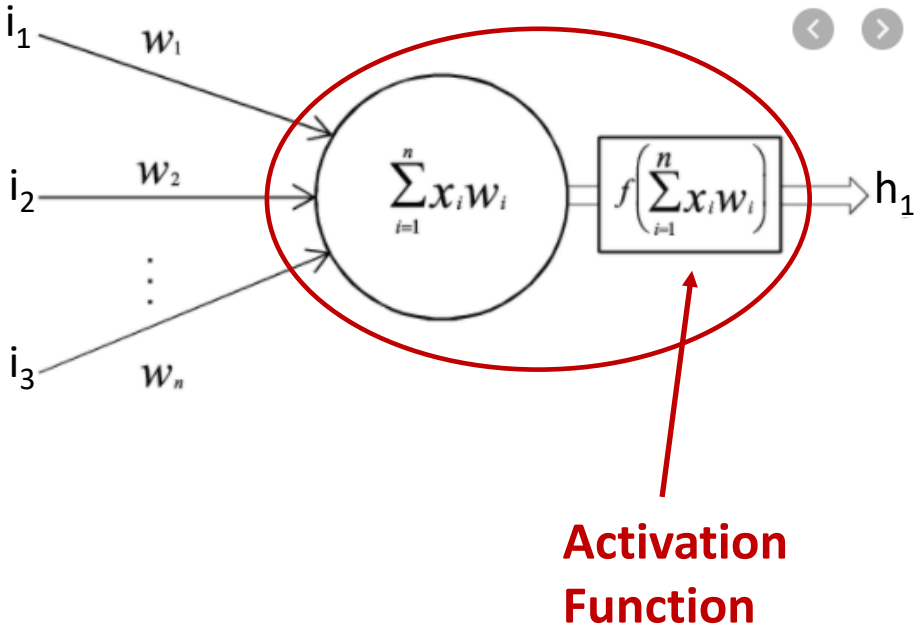
Activation Function is applied over the linear weighted summation of the incoming information to a node.

Convert linear input signals from perceptron to a linear/non-linear output signal.

It decides whether to activate a node or not.

Activation Functions

Activation functions must be monotonic, differentiable, and quickly converging.



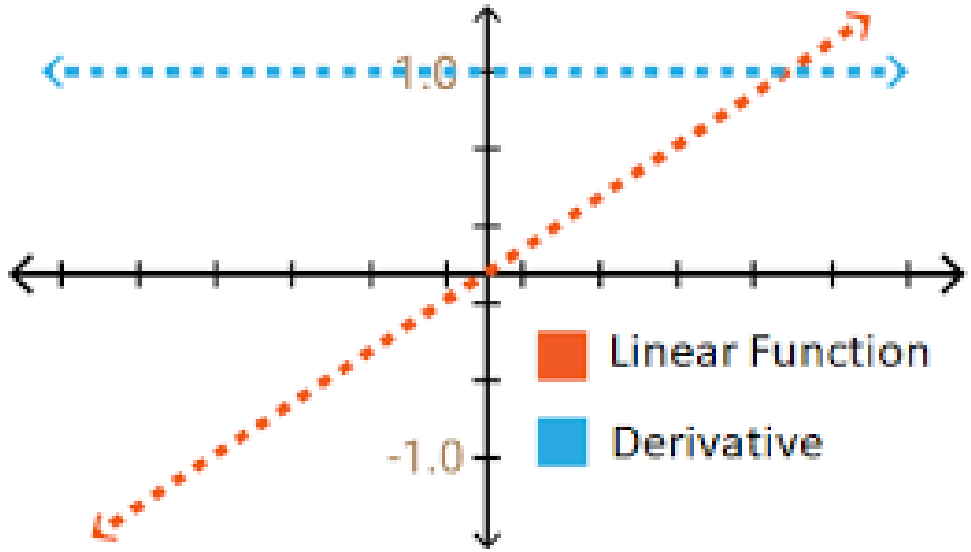
Types of Activation Functions:

- Linear
- Non-Linear

Linear

$$f(x) = ax + b$$

$$\frac{df(x)}{dx} = a$$



Observations:

- Constant gradient
- Gradient does not depend on the change in the input

Linear

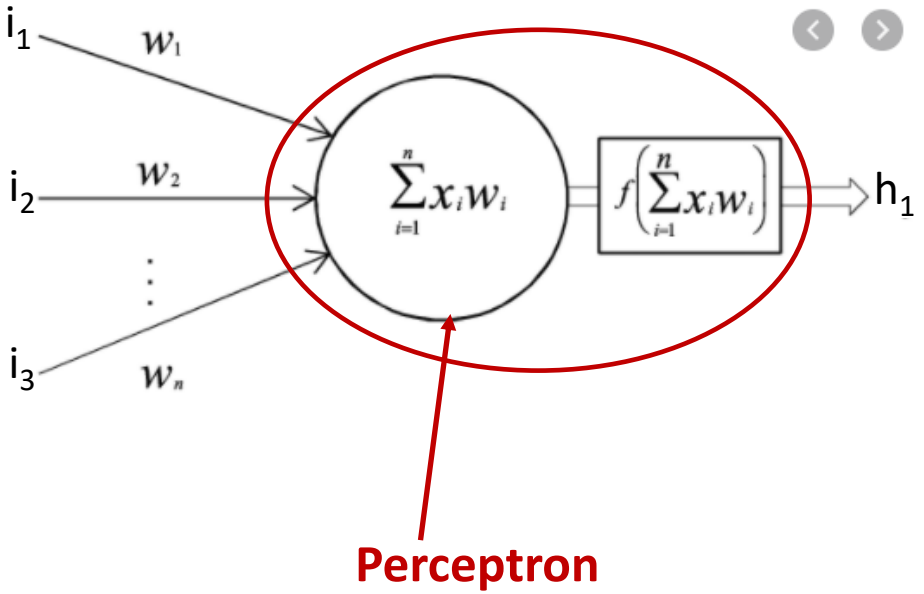
$$f(x) = ax + b$$

$$f(x) = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + b$$

Linear

$$f(x) = ax + b$$

$$f(x) = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + b$$



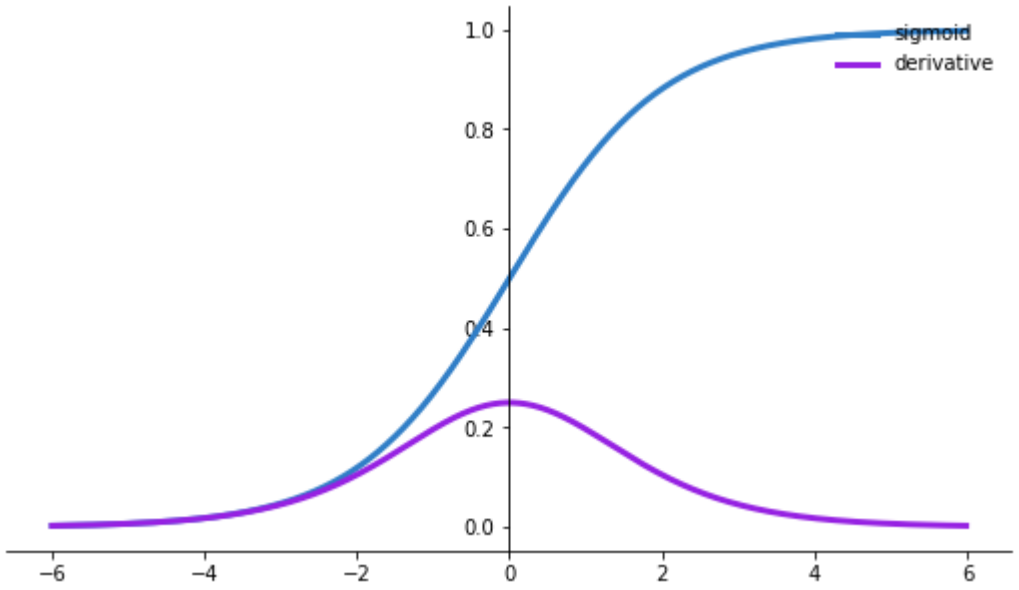
Non-Linear

- Sigmoid (Logistic)
- Hyperbolic Tangent (Tanh)
- Rectified Linear Unit (ReLU)
  - *Leaky Relu*
  - *Parametric Relu*
- Exponential Linear Unit (ELU)

Sigmoid Activation Functions (Logistics)

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

$$\frac{df(x)}{dx} = f(x)(1 - f(x))$$



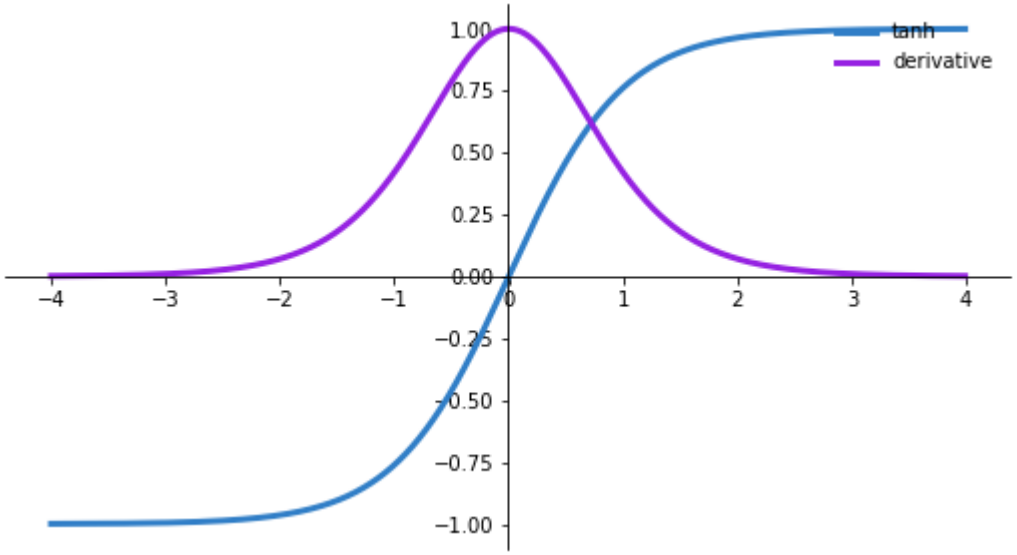
Observations:

- Output: 0 to 1
- Outputs are not zero-centered
- Can saturate and kill (vanish) gradients

Tanh Activation Function

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

$$\frac{df(x)}{dx} = 1 - f(x)^{2}$$



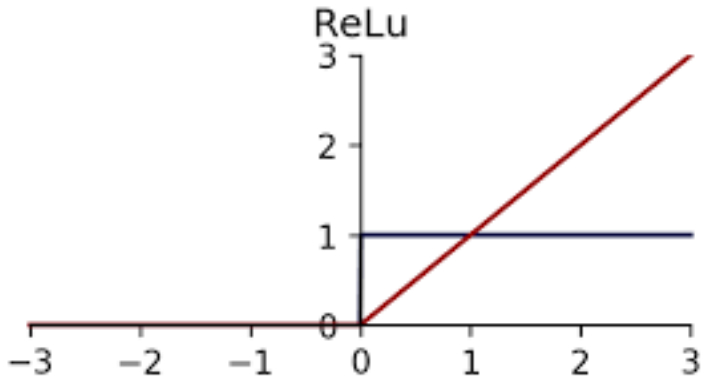
Observations:

- Output: -1 to +1
- Outputs are zero-centered
- Can Saturate and kill (vanish) gradients
- Gradient is more steeped than Sigmoid, resulting in faster convergence

Rectified Linear Unit(ReLU)

$$f(x) = \max(0, x)$$

$$\frac{df(x)}{dx} = 1$$



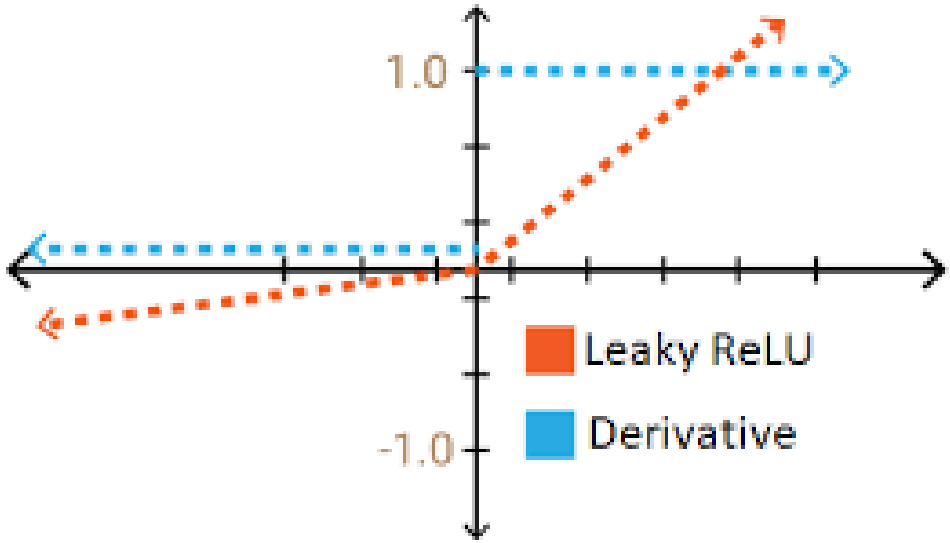
Observations:

- Greatly increase training speed compared to tanh and sigmoid
- Reduces likelihood of killing(vanishing) gradient
- It can blow up activation
- Dead nodes

Leaky-ReLU

$$f(x) = \max(0.01x, x)$$

$$\frac{df(x)}{dx} = \begin{cases} 0.01, & x < 0 \\ 1, & x \geq 0 \end{cases}$$



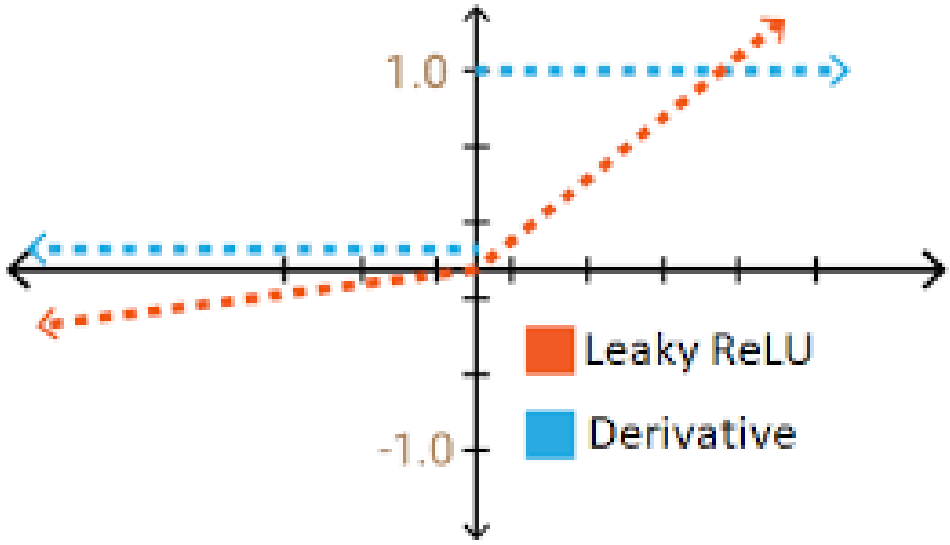
Observations:

- Fixed dying ReLU

Parameterized-ReLU

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

$$\frac{df(x)}{dx} = \begin{cases} \alpha, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

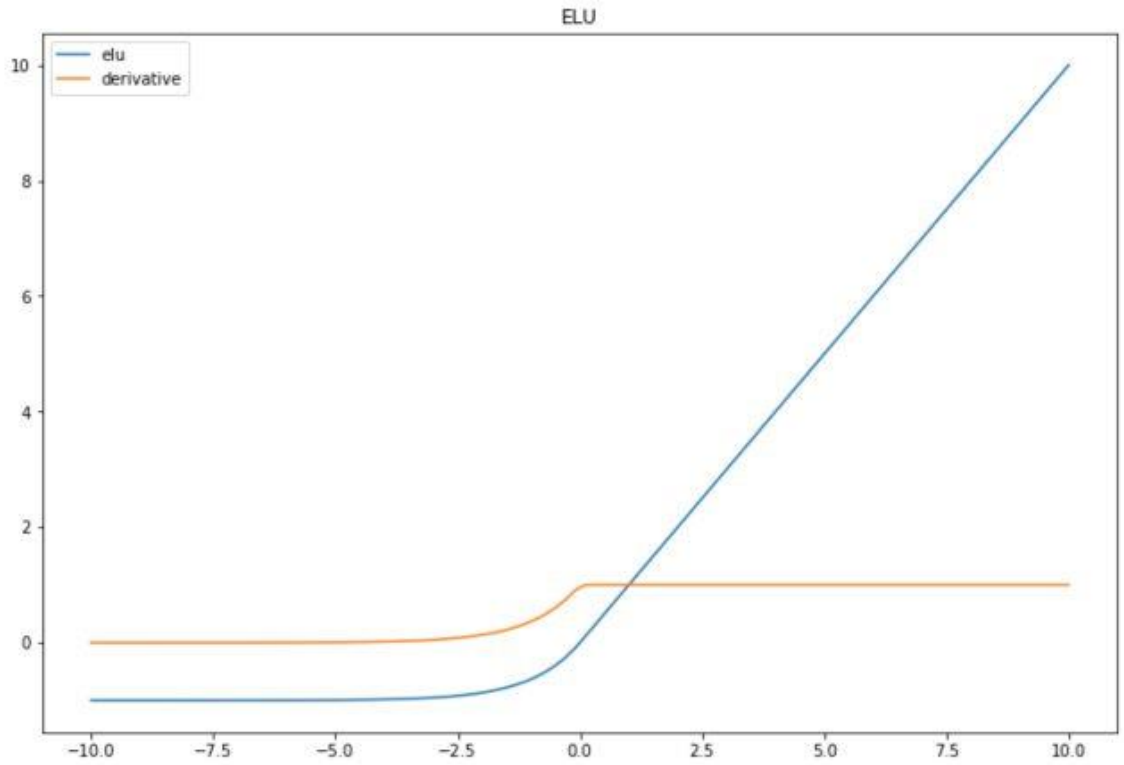


Observations:

Exponential Linear Unit (ELU)

$$\mathrm{f}(x) = \begin{cases} \alpha(e^x - 1), & x < 0 \\ 1x \;\; x \geq 0 \end{cases}$$

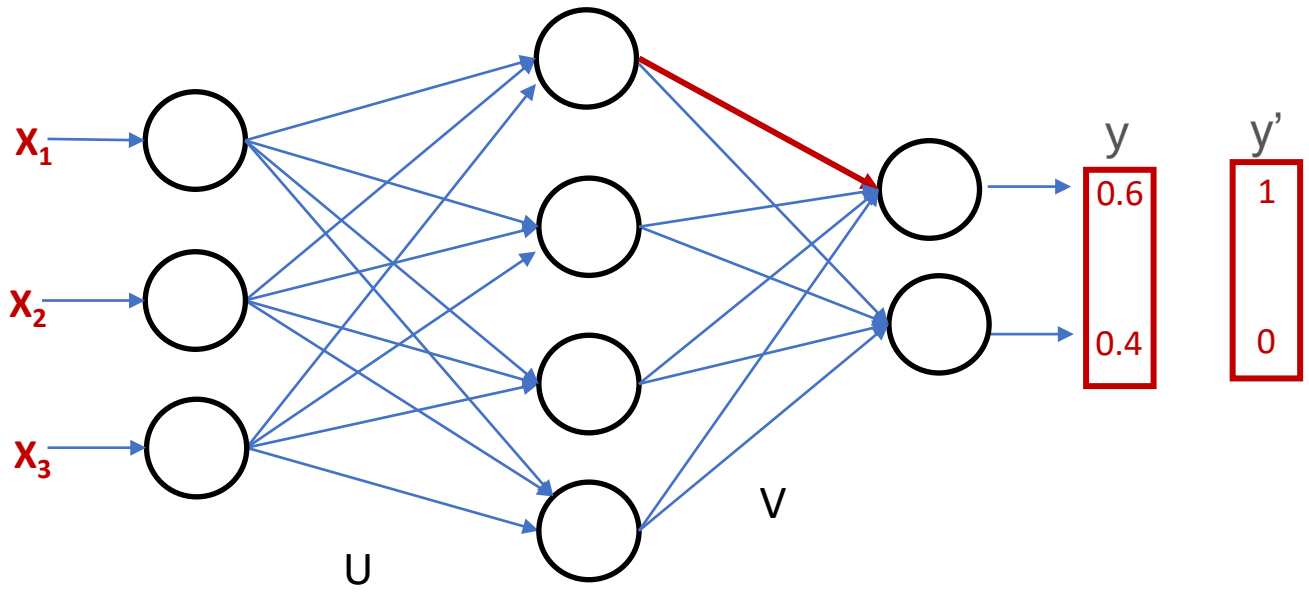
$$\frac{df(x)}{dx} = \begin{cases} f(x) + \alpha, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

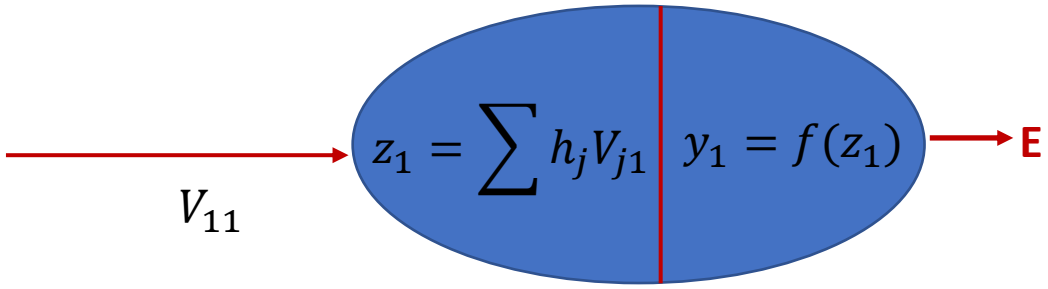


Observations:

- **It can produce –ve output**
- **It can blow up activation function**

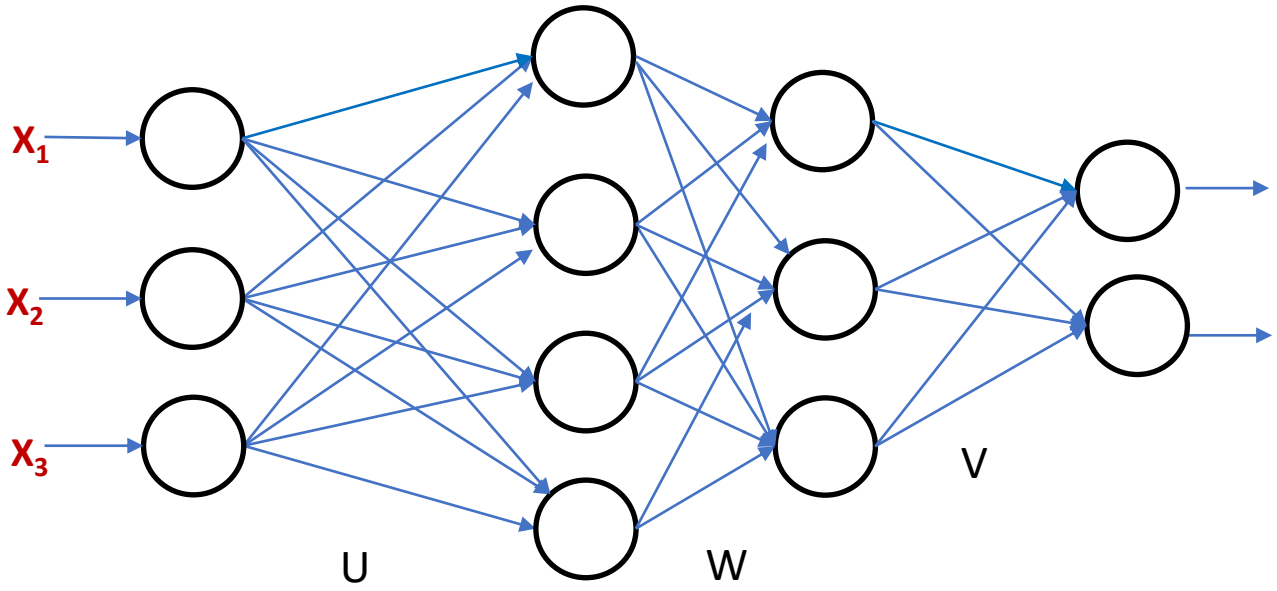
Complete Chain



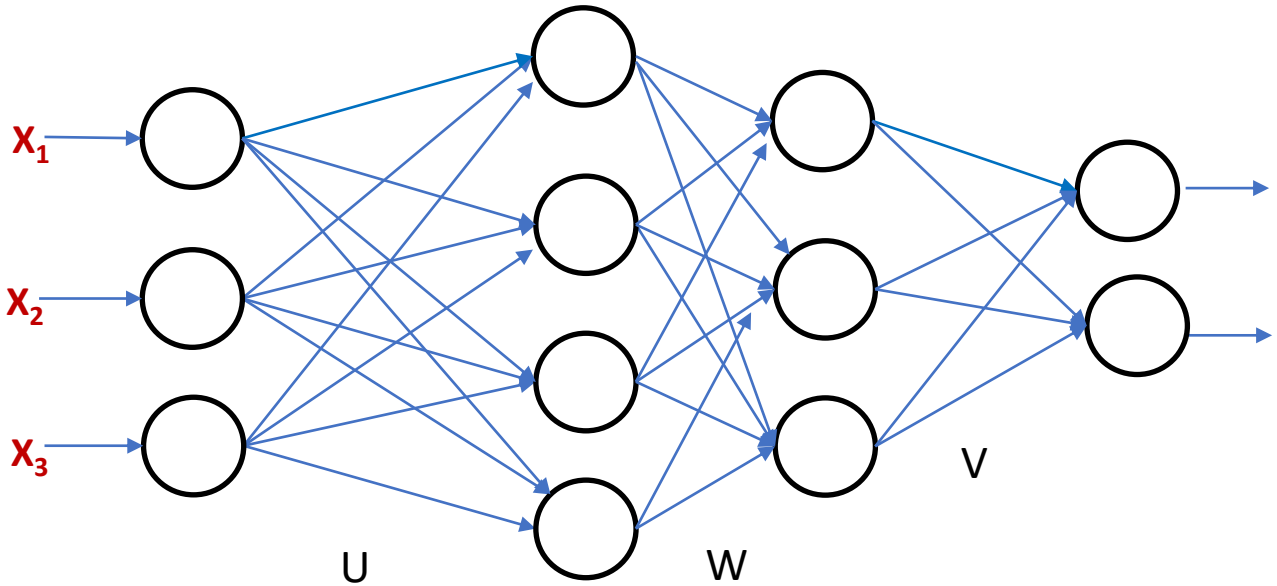


$$\frac{\delta E}{\delta V_{11}} = \frac{\delta z_1}{\delta V_{11}} \times \frac{\delta y_1}{\delta z_1} \times \frac{\delta E}{\delta y_1}$$

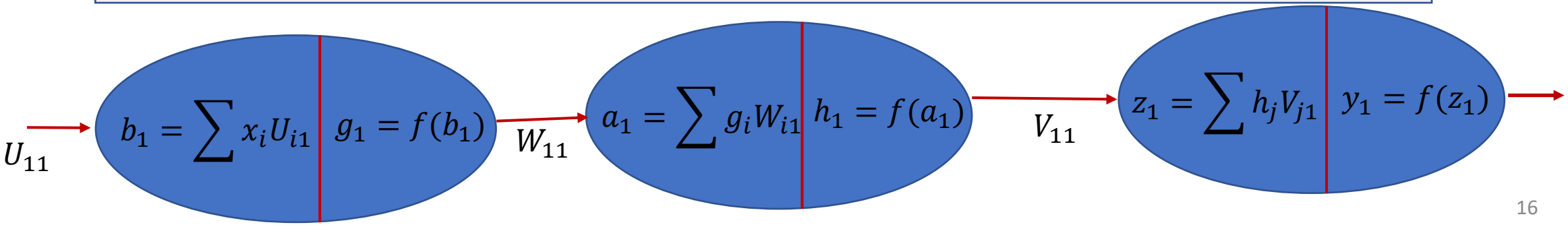
Deep Network



Deep Network - Vanishing/Exploding Gradient



$$\frac{\delta E}{\delta U_{11}} = \frac{\delta b_1}{\delta U_{11}} \times \frac{\delta g_1}{\delta b_1} \times \frac{\delta a_1}{\delta g_1} \times \frac{\delta h_1}{\delta a_1} \times \frac{\delta z_1}{\delta h_1} \times \frac{\delta y_1}{\delta z_1} \times \frac{\delta E}{\delta y_1}$$



Summary

- We learn characteristics of different Activation Functions and their gradient
- The choice of activation function depend on the nature of the problem, nature of the target output and the deepness of the network.