Improving Deep Neural Networks

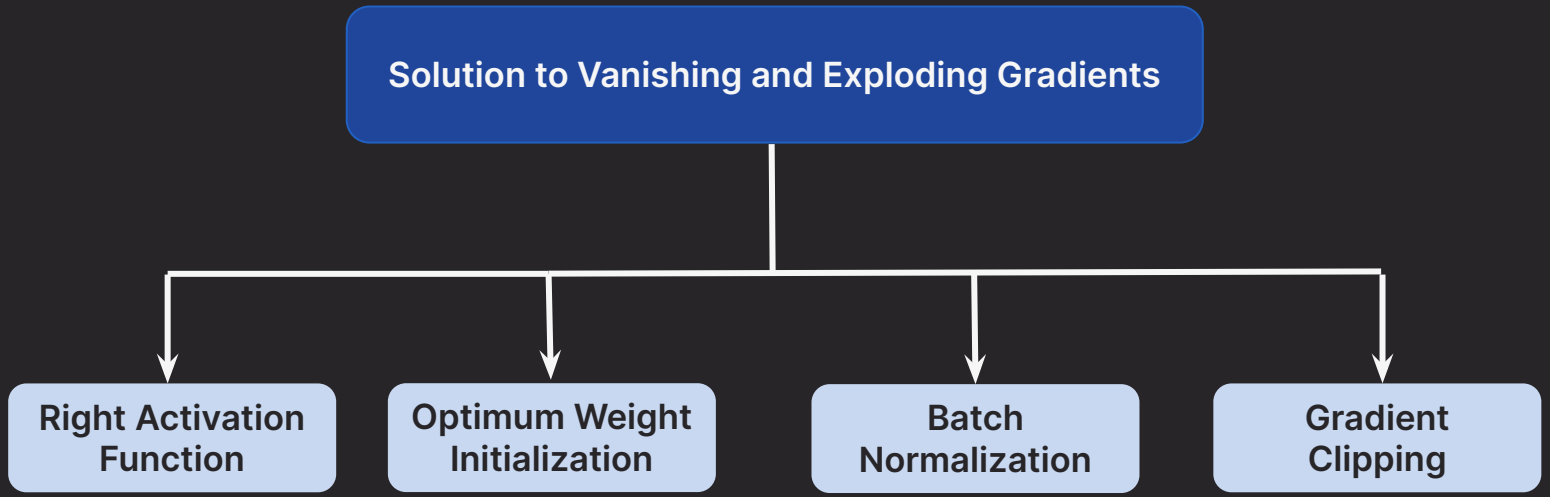
Video 2: The Solutions to Vanishing and Exploding Gradients - Part 1

Analytics Vidhya

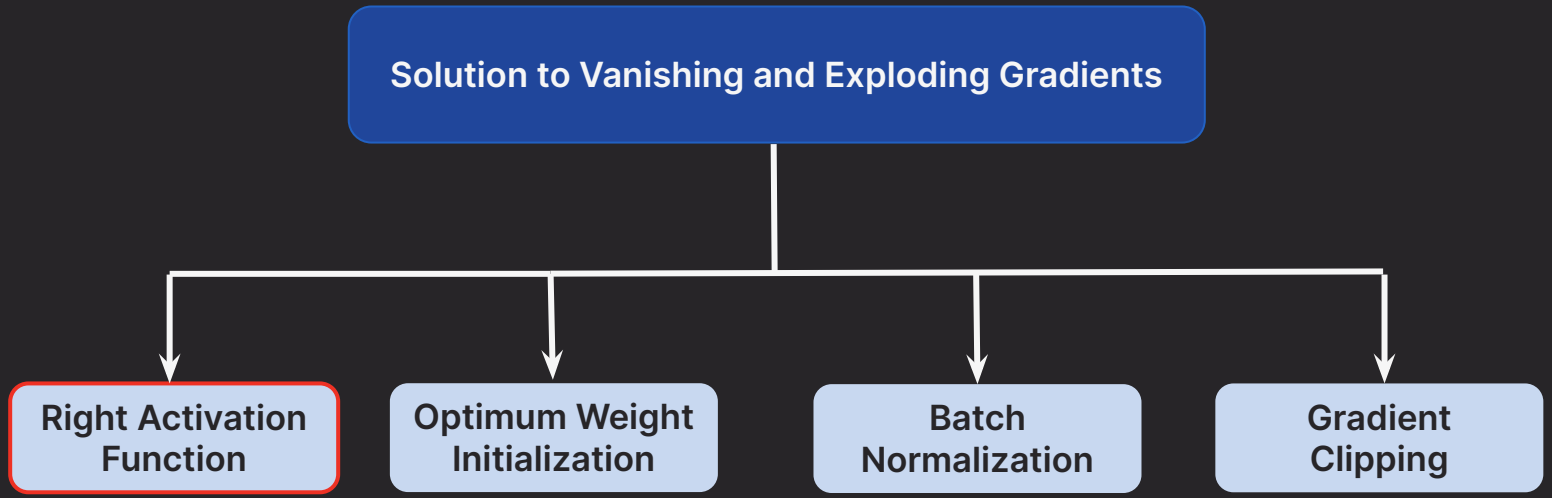
Vanishing and Exploding Gradients

Analytics Vidhya

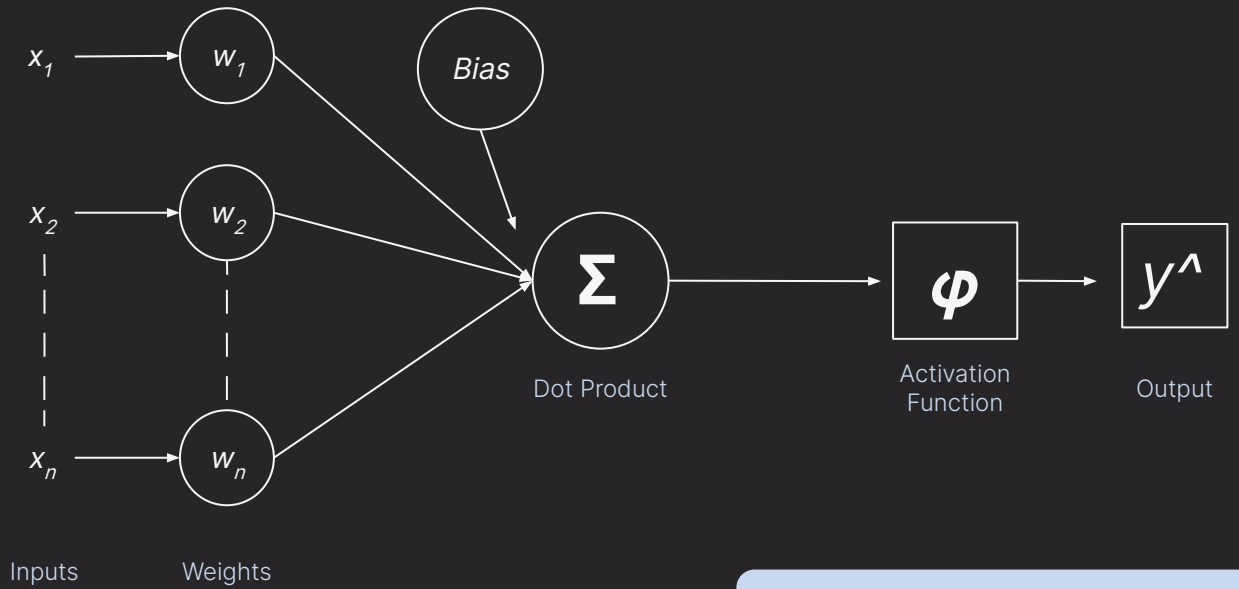
Vanishing and Exploding Gradients



Vanishing and Exploding Gradients

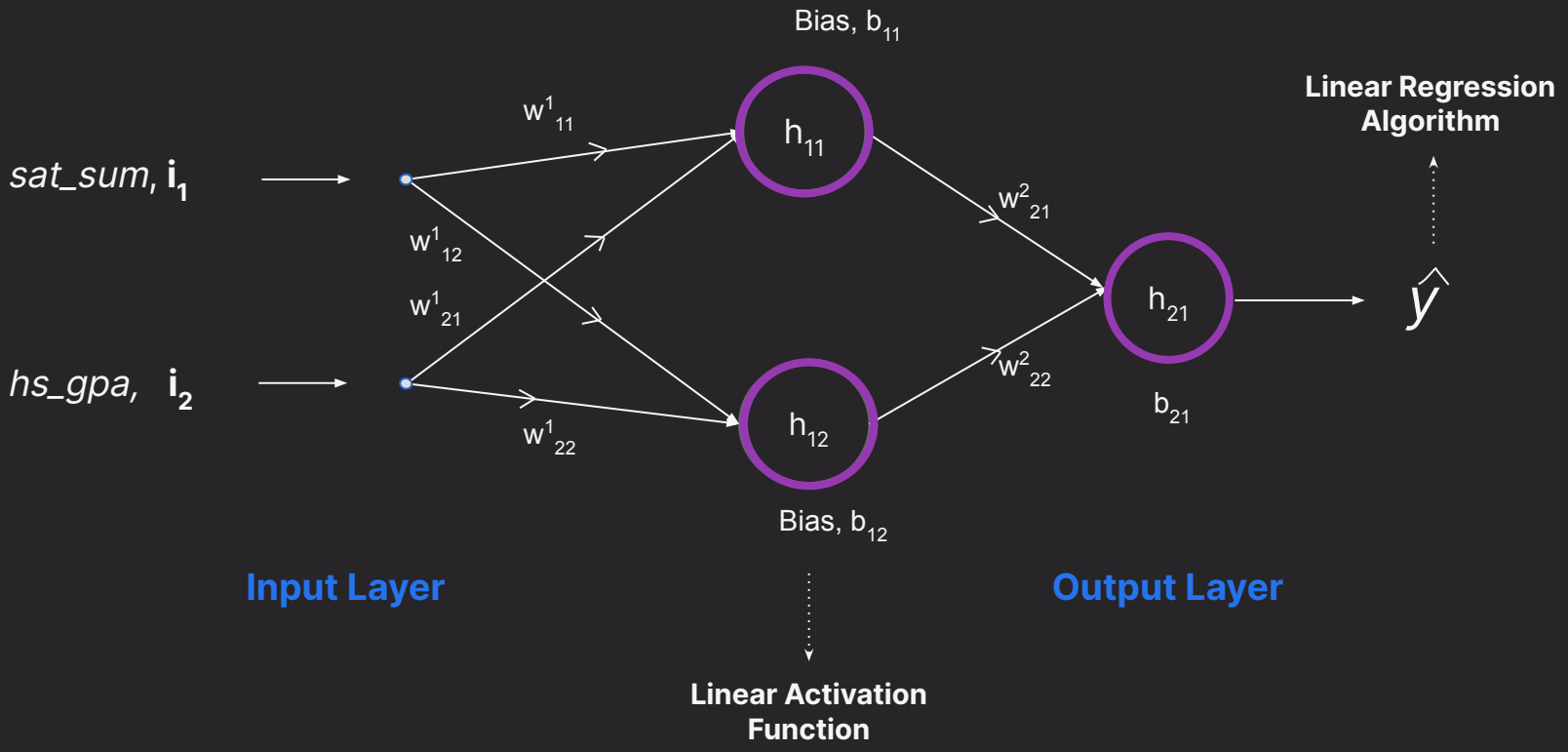


Recap



Linear Regression Model

Recap

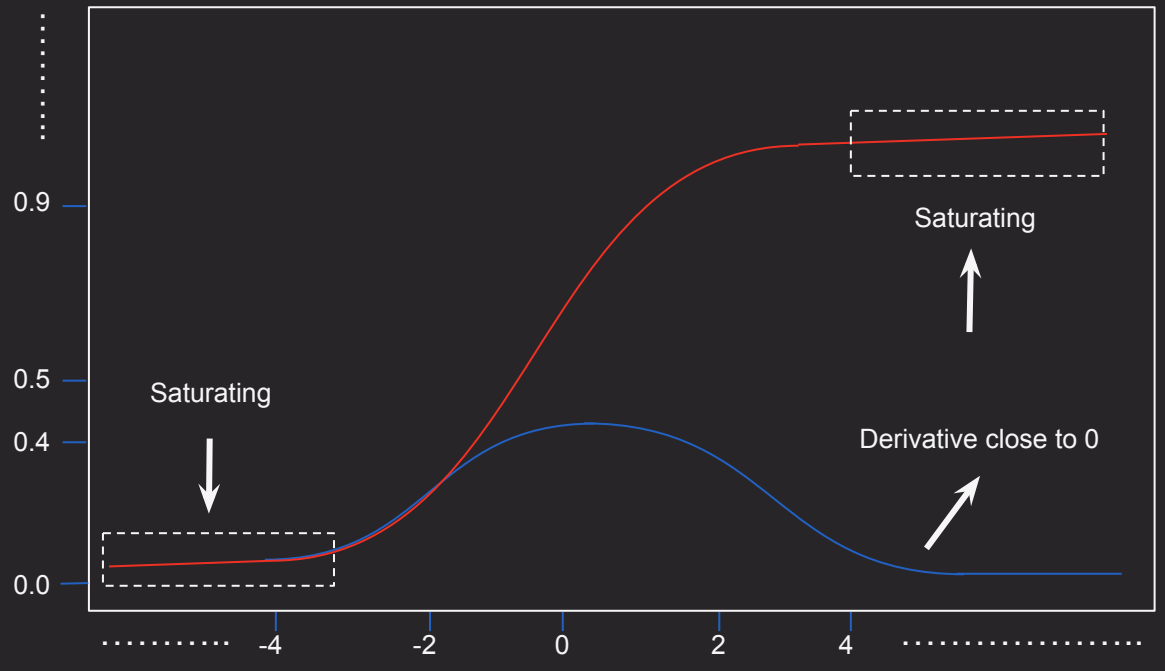

Bias, b_{11}
w^1_{11}
h_{11}
Linear Regression Algorithm
sat_sum, i_1
w^1_{12}
w^2_{21}
w^1_{21}
h_{21}
$\hat{y}$
hs_gpa, i_2
w^2_{22}
w^1_{22}
h_{12}
b_{21}
Bias, b_{12}
Input Layer
Output Layer
Linear Activation Function

Sigmoid for Non-Linearity

Researchers preferred using a sigmoid activation function

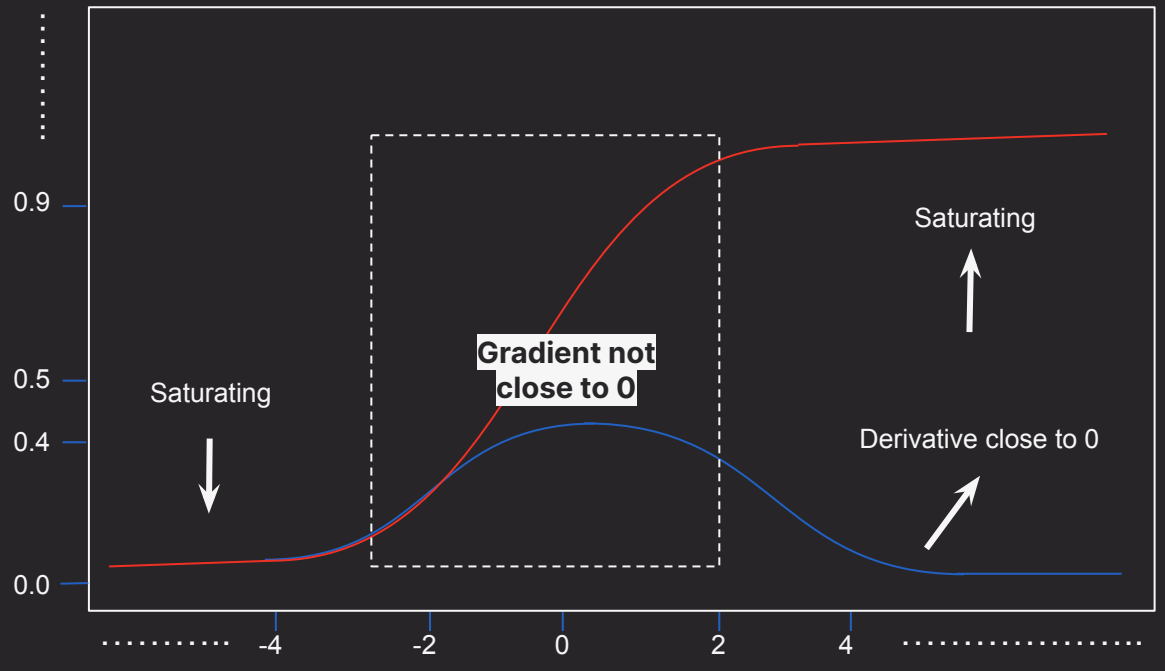
Sigmoid function narrows input ranges to 0-1 output

Sigmoid for Non-Linearity



Sigmoid Function and its derivatives

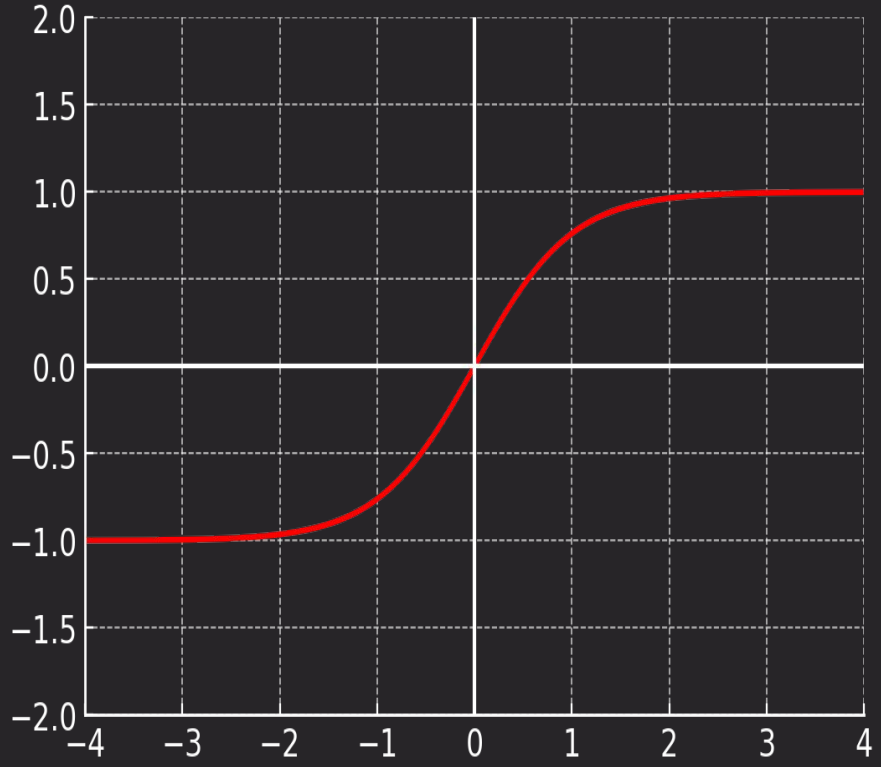
Sigmoid for Non-Linearity



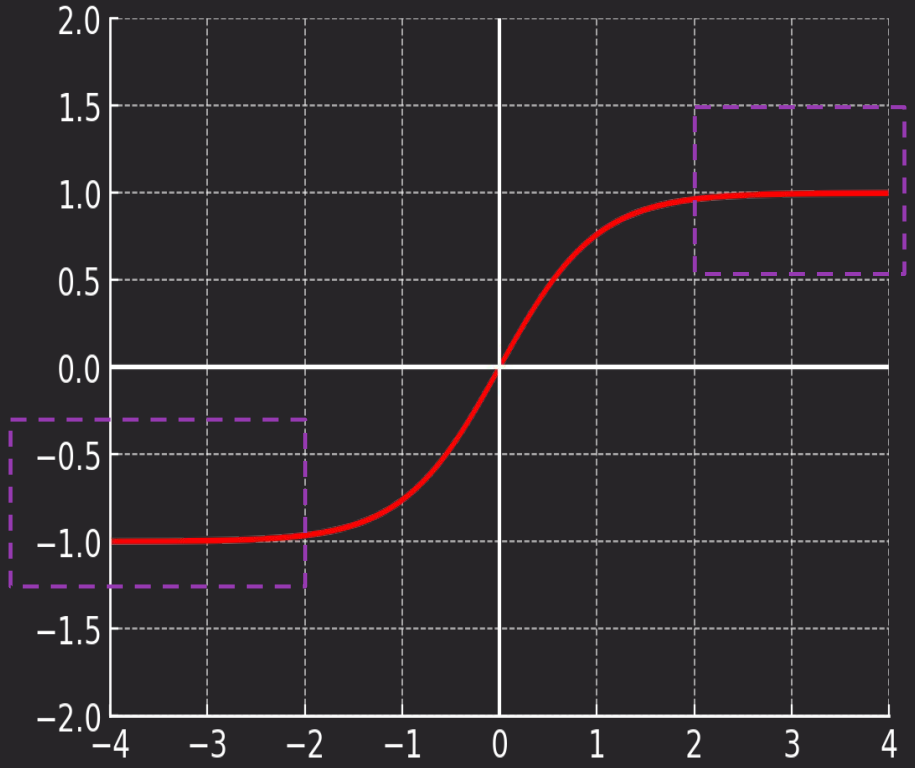
Sigmoid Function and its derivatives

Other activation function: TanH

TanH Function

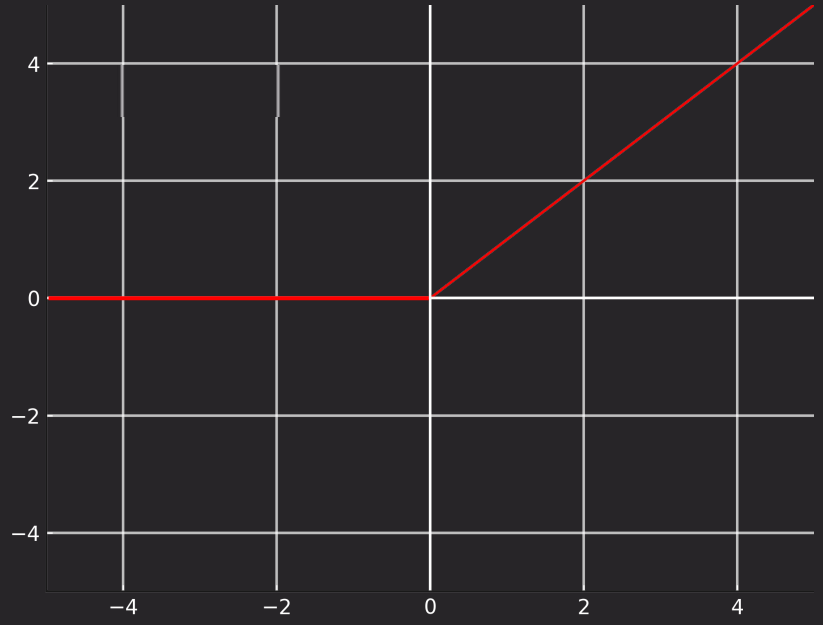


TanH Function



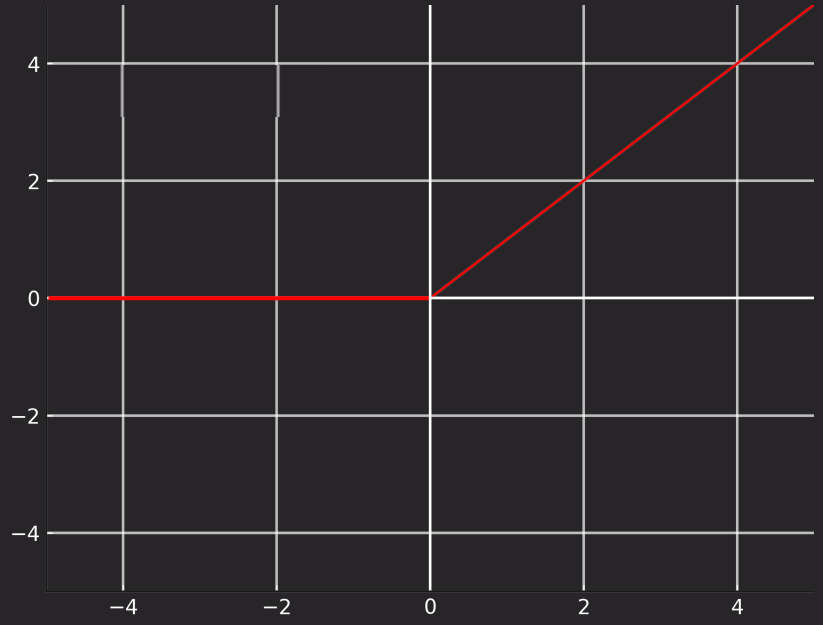
Other activation function: ReLU

ReLU Function



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

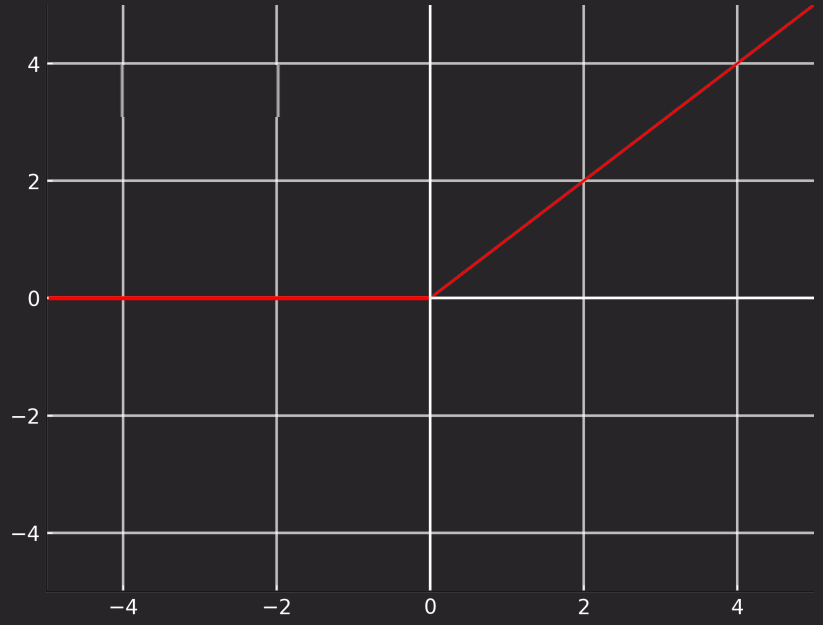
ReLU Function



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

ReLU is faster and more effective.

ReLU Function



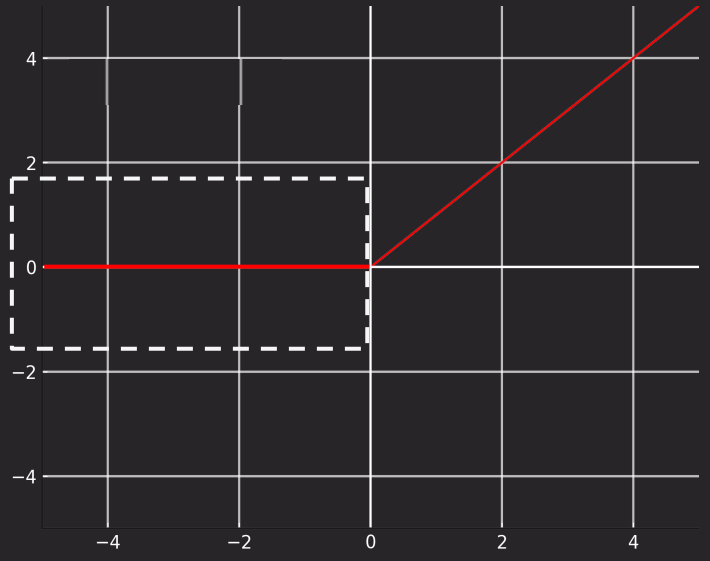
$$f(x) = \max(0, x)$$
$$f(x) = x, x \geq 0$$
$$f(x) = 0, x < 0$$

ReLU is faster and more effective.

ReLU is computationally efficient.

Analytics Vidhya

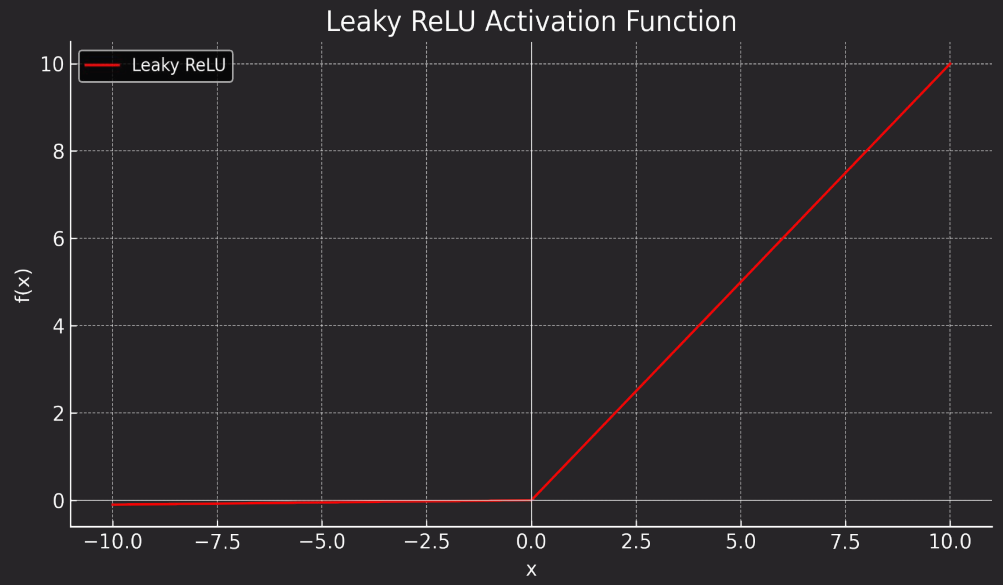
ReLU Function



During backpropagation process, the weights and biases for some neurons may not be updated.

Leaky ReLU and ELU Activation Functions

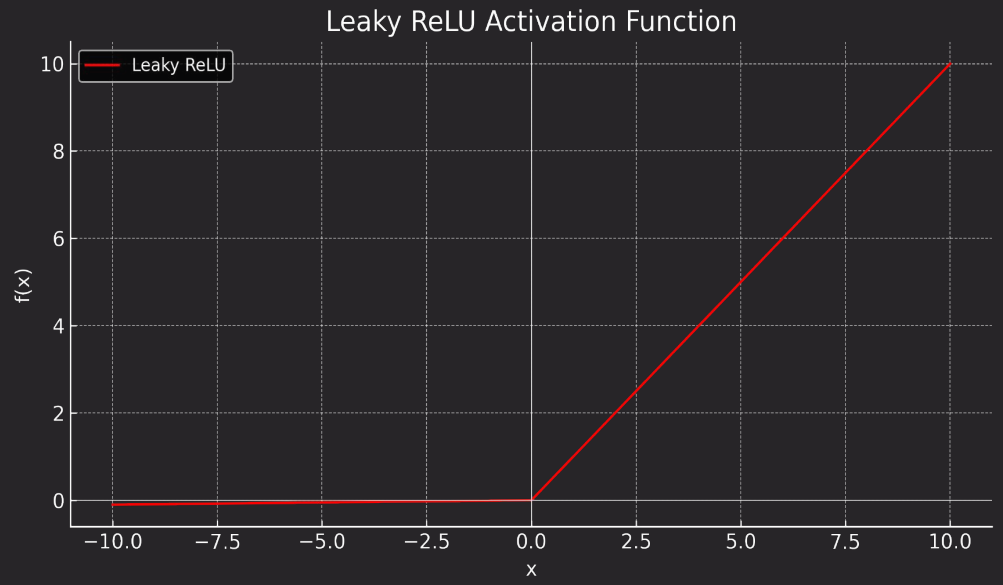
Leaky ReLU Function



$$f(x)=\max(\text{alpha}*x\ ,\ x)$$

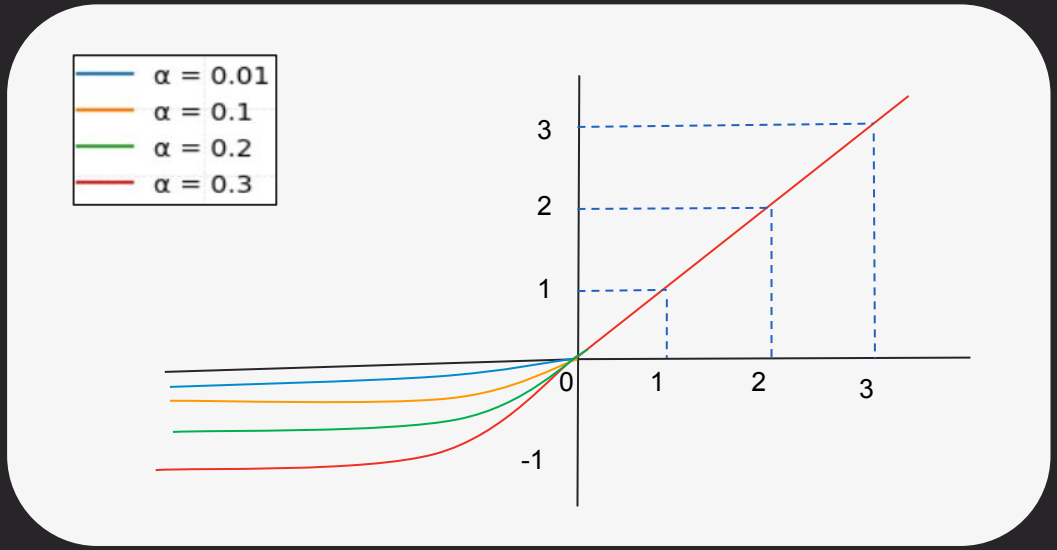
Analytics Vidhya

Leaky ReLU Function



$$\text{Leaky ReLU} = \begin{cases} x, & if\ x > 0 \\ a \times x, & if\ x \leq 0 \end{cases}$$

Exponential Linear Unit



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

Analytics Vidhya

The Current Approach

- ReLU is used to reduce the problem of vanishing gradient
- The sigmoid activation function is used in the last layer of the binary classification problems.
- TanH and sigmoid are used in the hidden layers of some complex neural architectures.

Analytics Vidhya

Up-Next: Optimum Weight Initialization Techniques