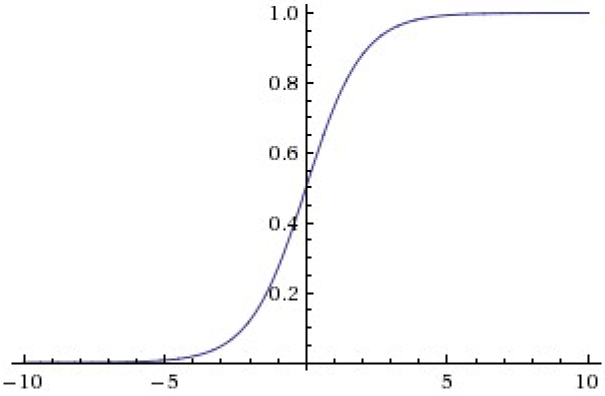
Activation Functions

Sigmoid

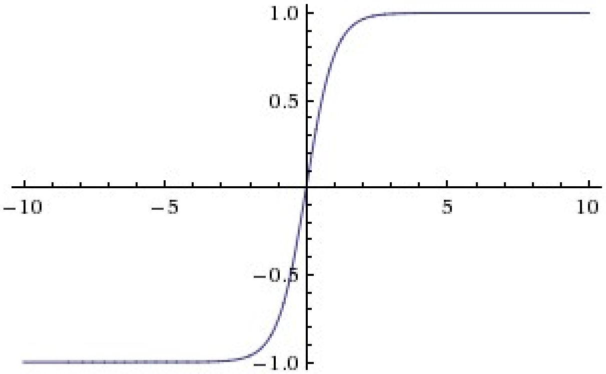


Ranges from (0, 1)

Disadvantages

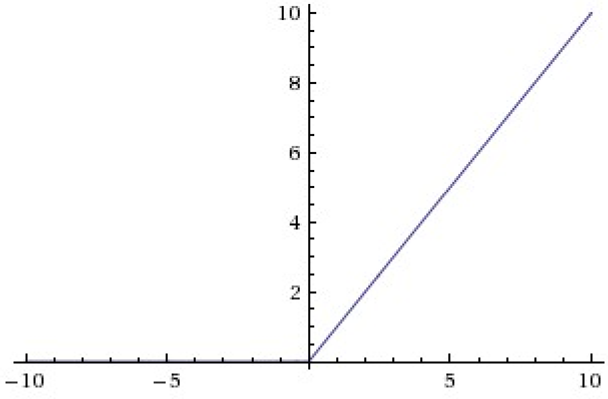
1) Saturates and Kills gradients
    a. If activation values reaches the tails of zero or one, then the gradient becomes very close to zero.
    b. If the gradient value is close to zero, the learning is very low and stops eventually.
2) Not Zero centered
    a. This causes all the weights to be either +ve or –ve
    b. Can cause over fitting/under fitting

TanH



Ranges from (-1, 1). It is zero-centered. Thus, it is preferred over sigmoid.

RELU (REctified Linear Unit)



f(x)=max(0,x)

Advantages:

1) Faster convergence of stochastic gradient descent compared to the sigmoid/tanh functions.
2) Less expensive operations compared to sigmoid/tanh.

Disadvantages:

1) In RELU, A large gradient can update weight of a neuron in a way that it can never be activated again. It is said that if the learning rate is high, 40% of the network can be dead (never activated again).

Reference

- *http://cs231n.github.io/neural-networks-1/*