



Get unlimited access

Open in app



Published in Analytics Vidhya



KS Haarish Dharan

Follow

Mar 23, 2021 · 5 min read · Listen



Save

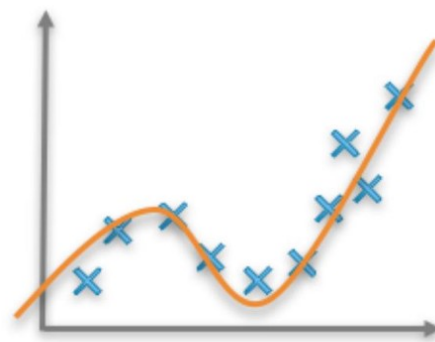


ACTIVATION FUNCTIONS IN NEURAL NETWORK(MATHEMATICAL INTUITION WITH GRAPHS)

Activation functions are used in neural networks in order to map the input and output in a more perfect manner and most of the time(about 99% time) we use non-linear activation function rather than linear activation function because a non-linear graph can do much better job at mapping than the linear graph.



Linear function



Non-linear function

Best fit linear and non-linear models

[Source](#)

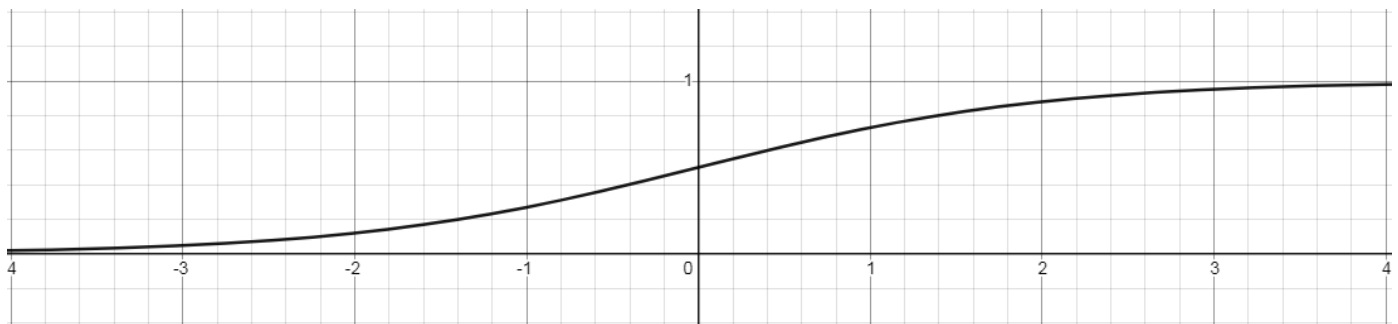


[Get unlimited access](#)[Open in app](#)

make easier to interpret the results. Sigmoid functions are also an important part of a logistic regression and classification model. Logistic regression is a modification of for two-class classification, and converts one or more real-valued inputs into a probability, such as the probability that a customer will purchase a product.

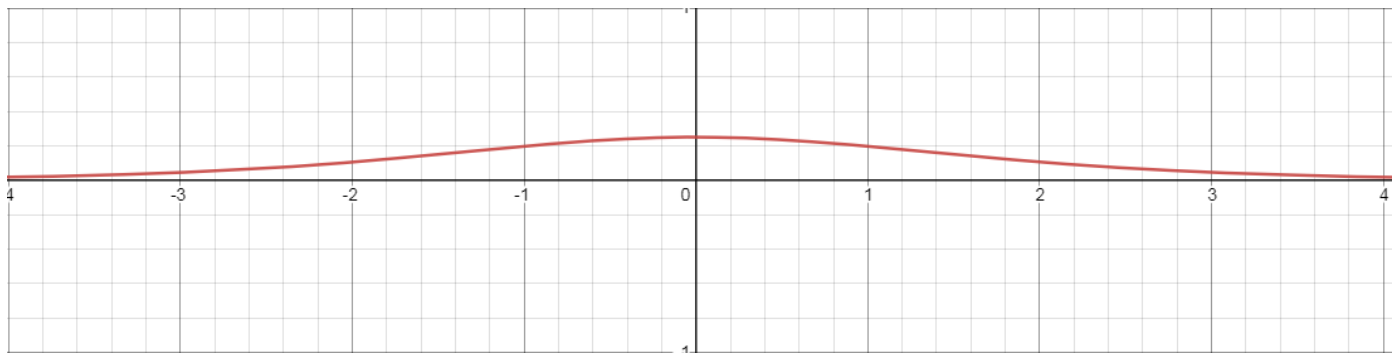
Mathematical formula of sigmoid function:

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



from Desmos online graphing calculator

if you observe this graph you could tell this is an S shaped curve and it is continuously differentiable at all points but if we observe the graph of derivative of this function



from Desmos online graphing calculator

The gradient values are significant for range -3 and 3 but the graph gets much flatter in other regions. This implies that for values greater than 3 or less than -3, will have very small gradients. As the gradient value approaches zero, the network is not really learning.

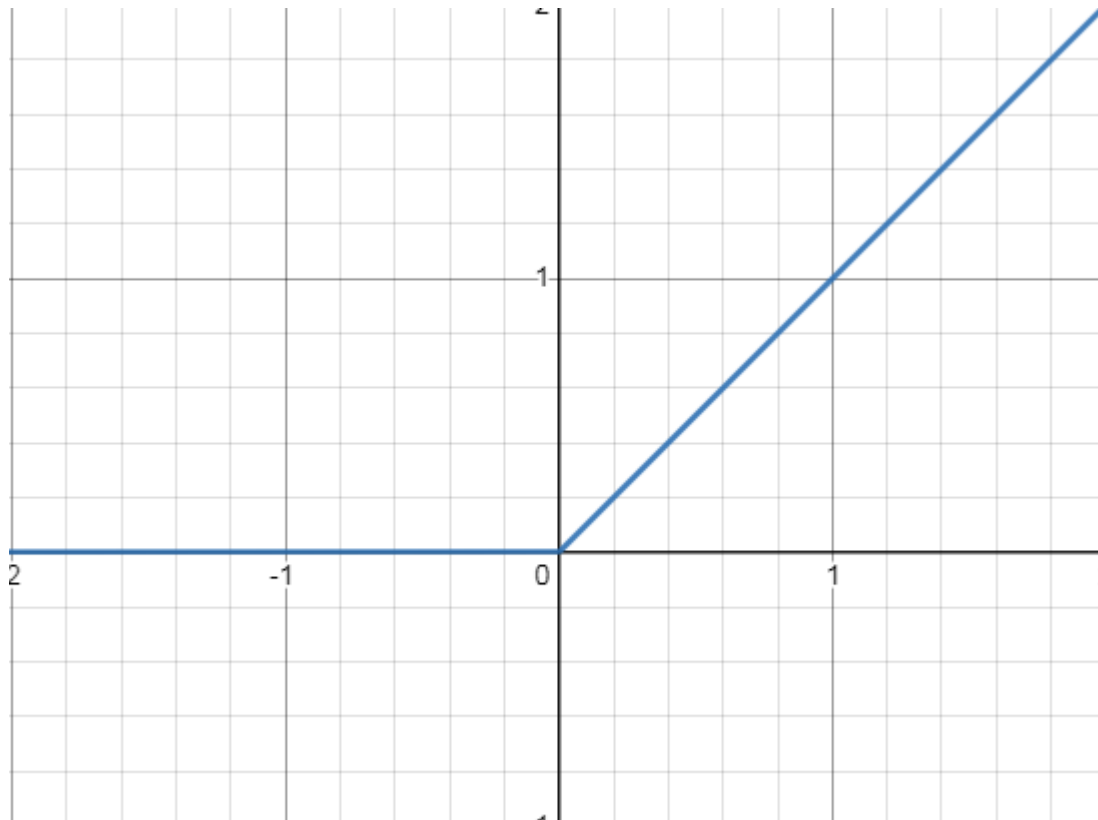


[Get unlimited access](#)[Open in app](#)

RELU:

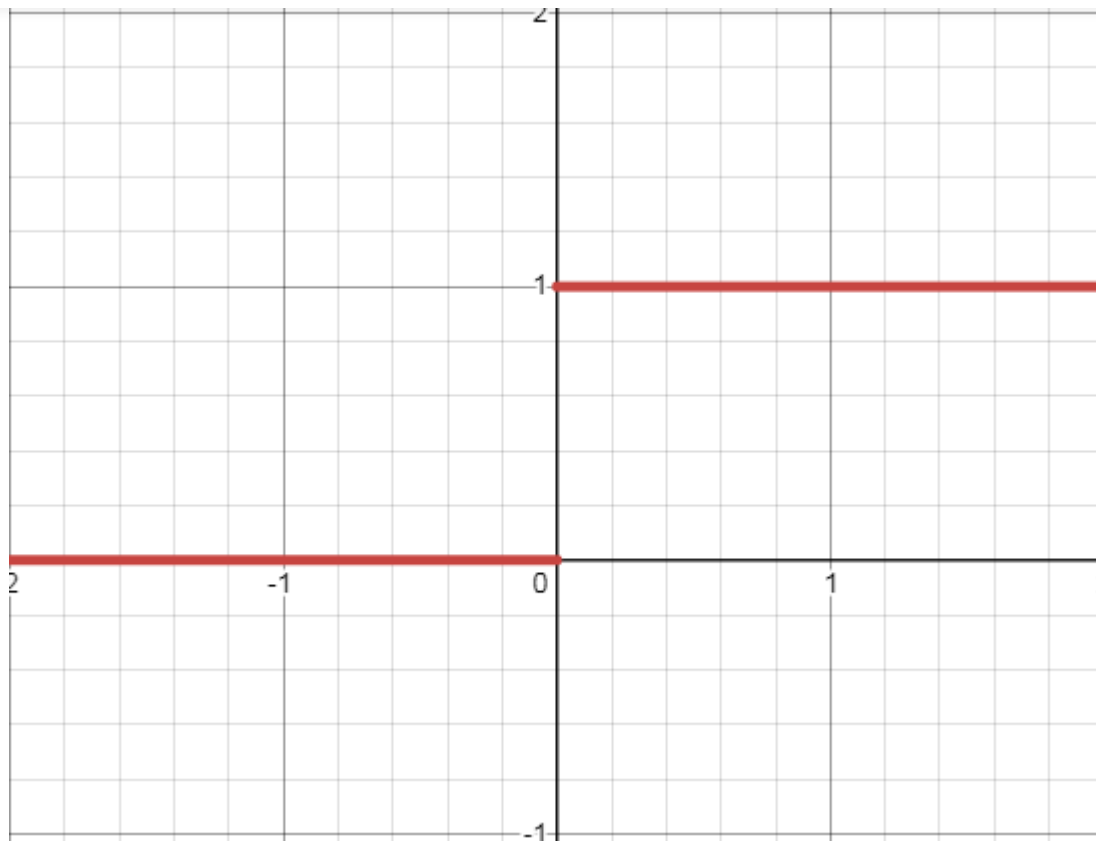
ReLU stands for Rectified Linear Unit. The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time.

$f(x) = \max(0, x)$ (i.e for any value of x above 0 it returns the same value whereas any value of x below 0 it returns 0)



from Desmos graphing calculator



[Get unlimited access](#)[Open in app](#)

from Desmos graphing calculator

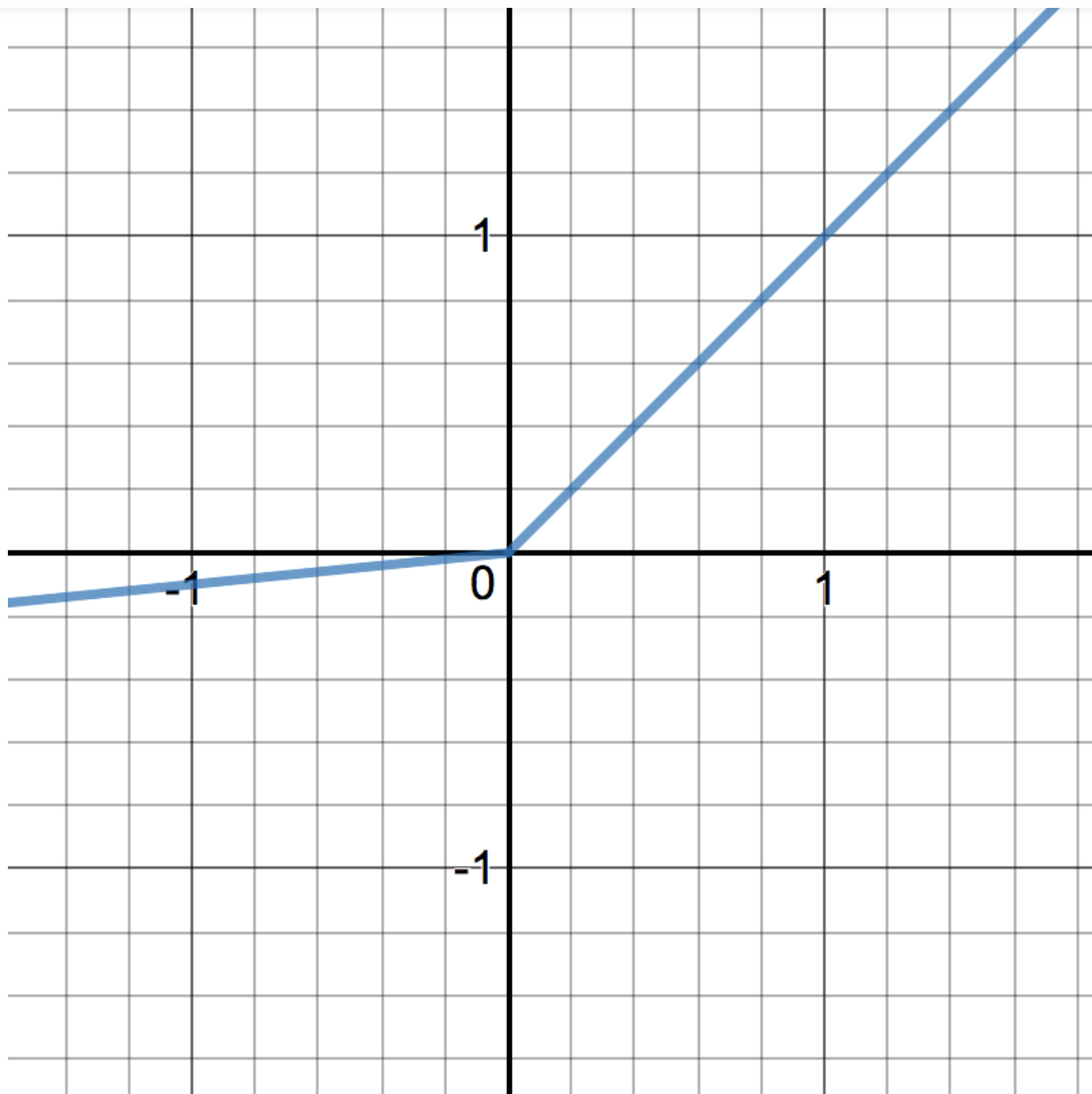
The problem with the ReLU activation function is If you look at the negative side of the gradient graph, you will notice that the gradient value is zero. Due to this reason, during the backpropagation process, the weights and biases for some neurons are not updated. This can create dead neurons which never gets activated.

LEAKY RELU:

Leaky ReLU introduces a little modification in the ReLU activation function instead of assigning 0 for the negative values of x , we define it as an extremely small linear component of x .

$$f(x) = 0.01x \text{ when } x < 0 \text{ and } f(x) = x \text{ when } x \geq 0$$



[Get unlimited access](#)[Open in app](#)

Source

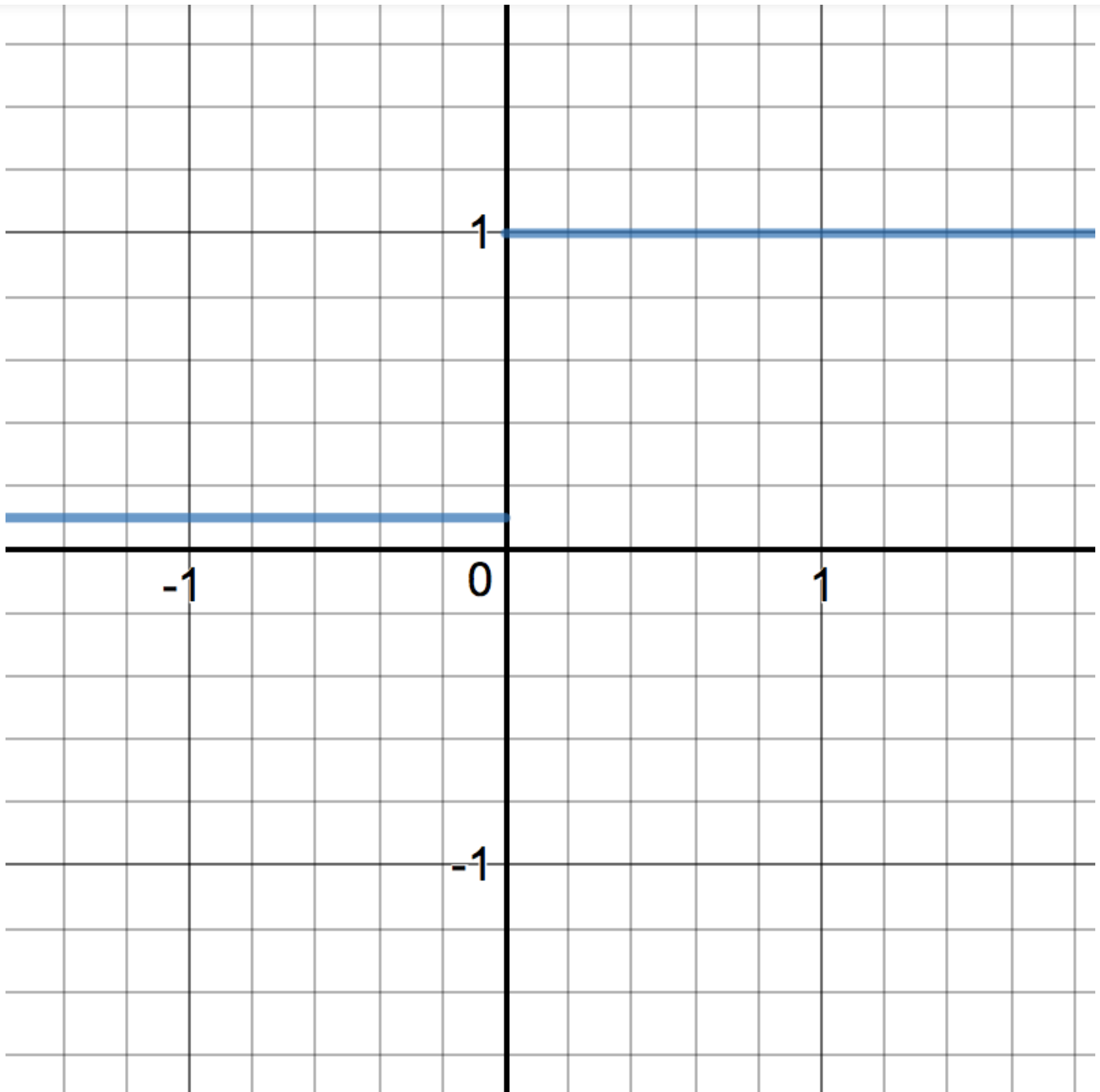
$$f'(x) = 0.01 \text{ when } x < 0 \text{ and } f'(x) = 1 \text{ when } x \geq 0$$





Get unlimited access

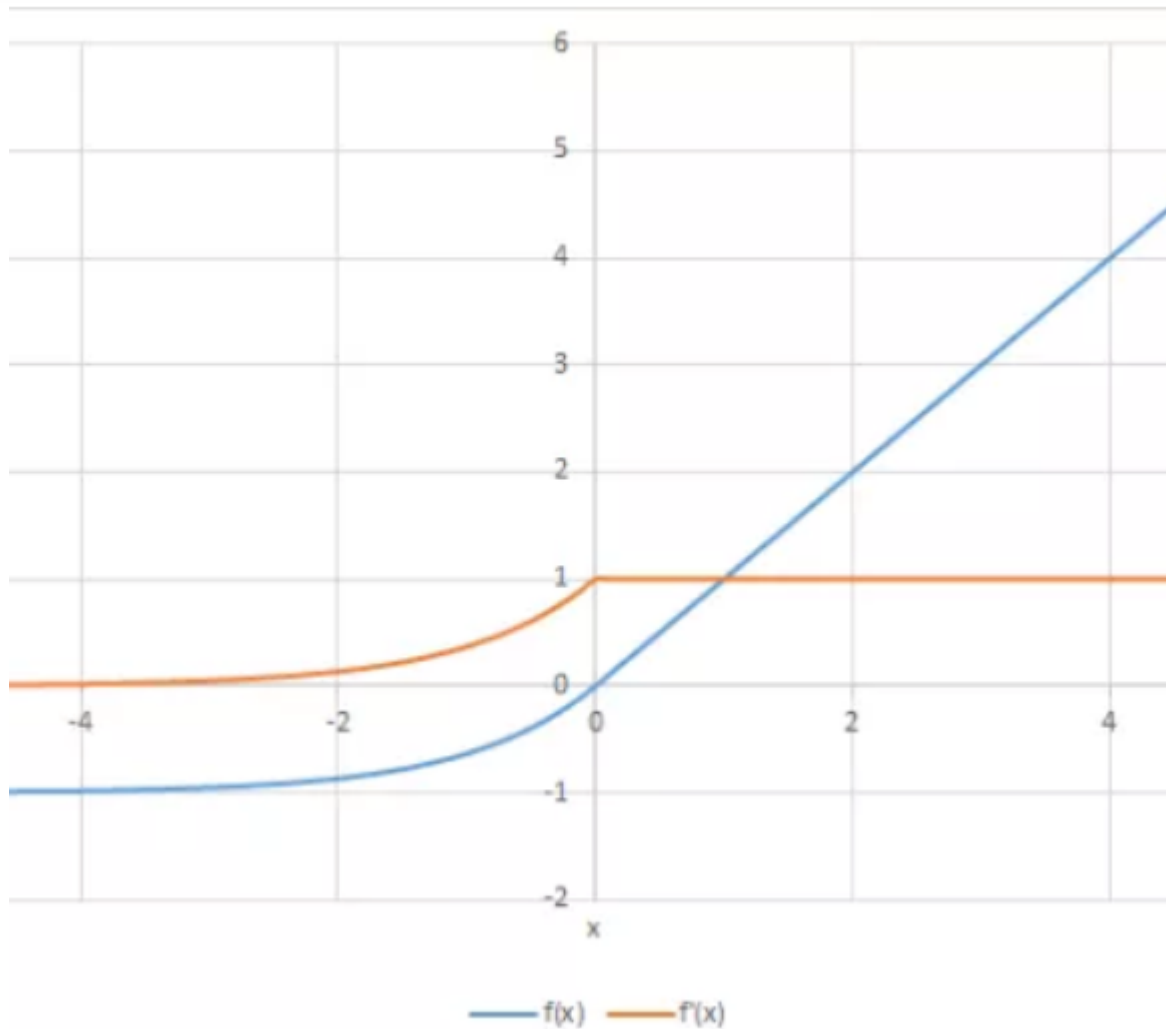
Open in app

Source

By making this small modification, the gradient for negative values of x comes out to be a non zero value. Hence we would no longer encounter dead neurons in that region.

Elu(exponential linear unit):



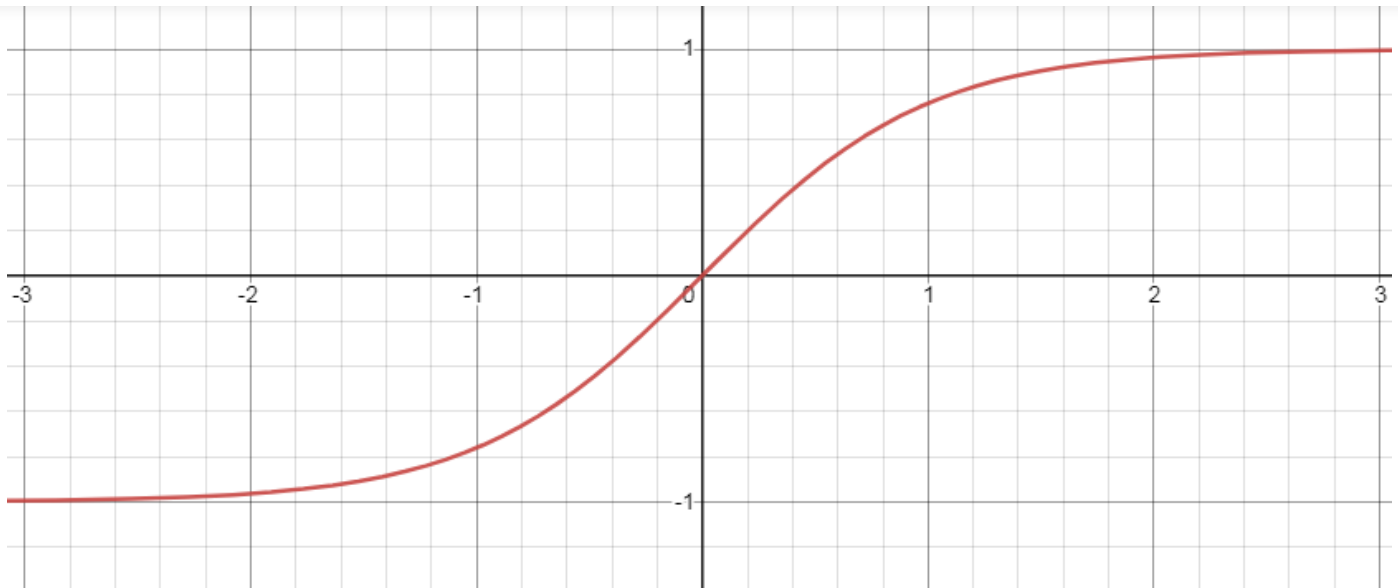
[Get unlimited access](#)[Open in app](#)

Source

Tanh:

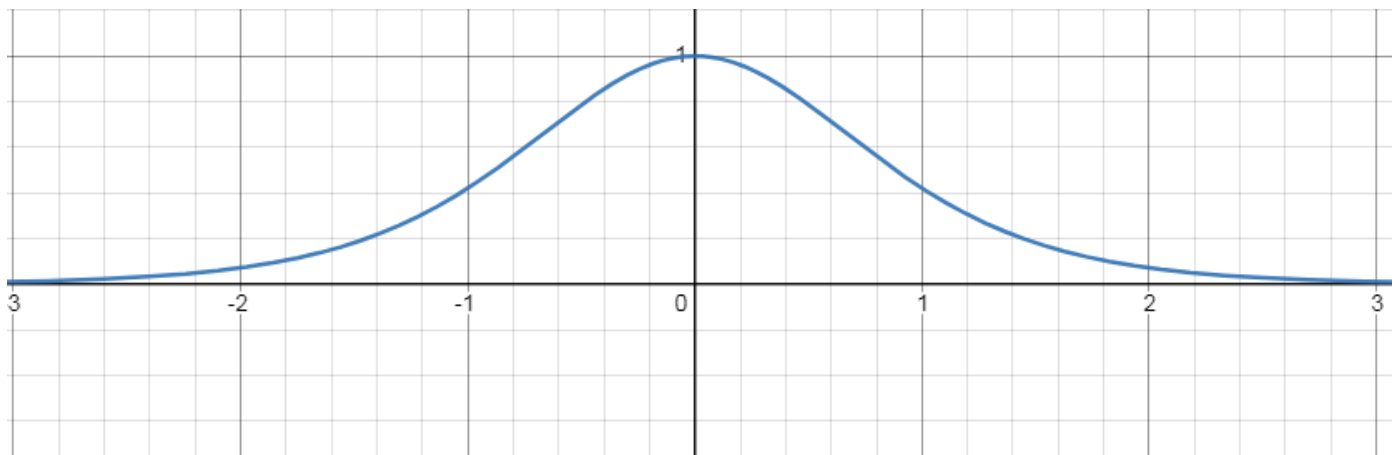
The tanh function is very similar to the sigmoid function. The only difference is that it is symmetric around the origin. The range of values in this case is from -1 to 1. Thus output will not be of only positive.



[Get unlimited access](#)[Open in app](#)

from Desmos graphing calculator

Apart from range all other properties of \tanh function are the same as that of the sigmoid function. Similar to sigmoid, the \tanh function is continuous and differentiable at all points.



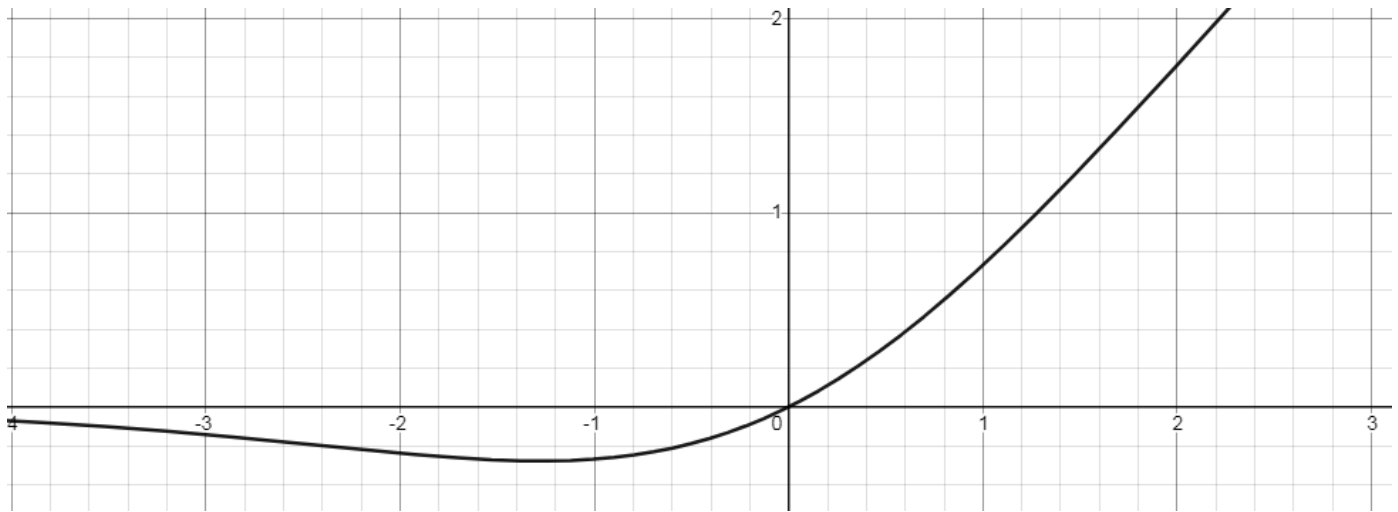
from Desmos graphing calculator

The gradient of the \tanh function is steeper as compared to the sigmoid function. Usually \tanh is preferred over the sigmoid function since it is zero centered and the gradients are not restricted to move in a certain direction.



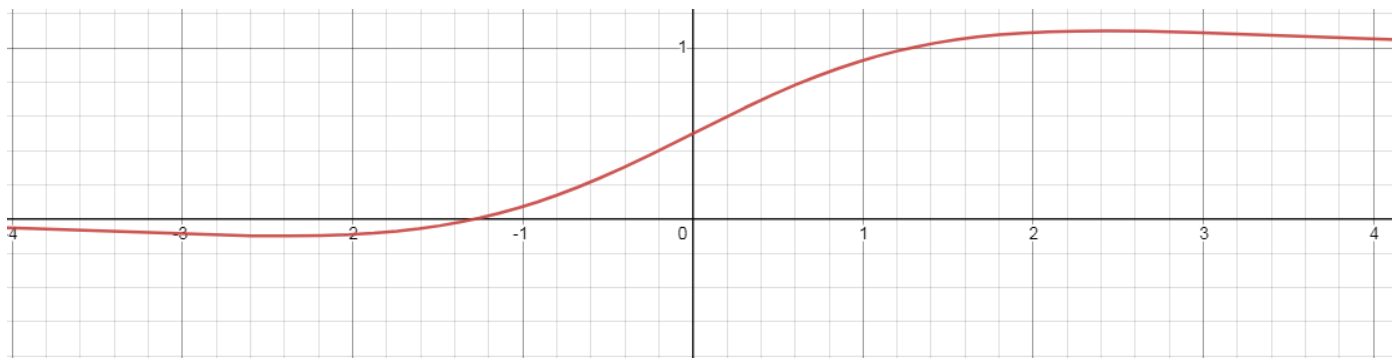

[Get unlimited access](#)
[Open in app](#)

$$f(x) = x * \text{sigmoid}(x)$$



from Desmos graphing calculator

$$f'(x) = f(x) + \text{sigmoid}(x) * (1 - f(x))$$



from Desmos graphing calculator

the curve of the function is smooth and the function is differentiable at all points. This is helpful during the model optimization process and is considered to be one of the reasons that swish outperforms ReLU.

SOFTMAX:

Softmax is the activation function that is used in the final output layer and mainly used in multi-classification problems, as it will report back the “confidence score” for each class.



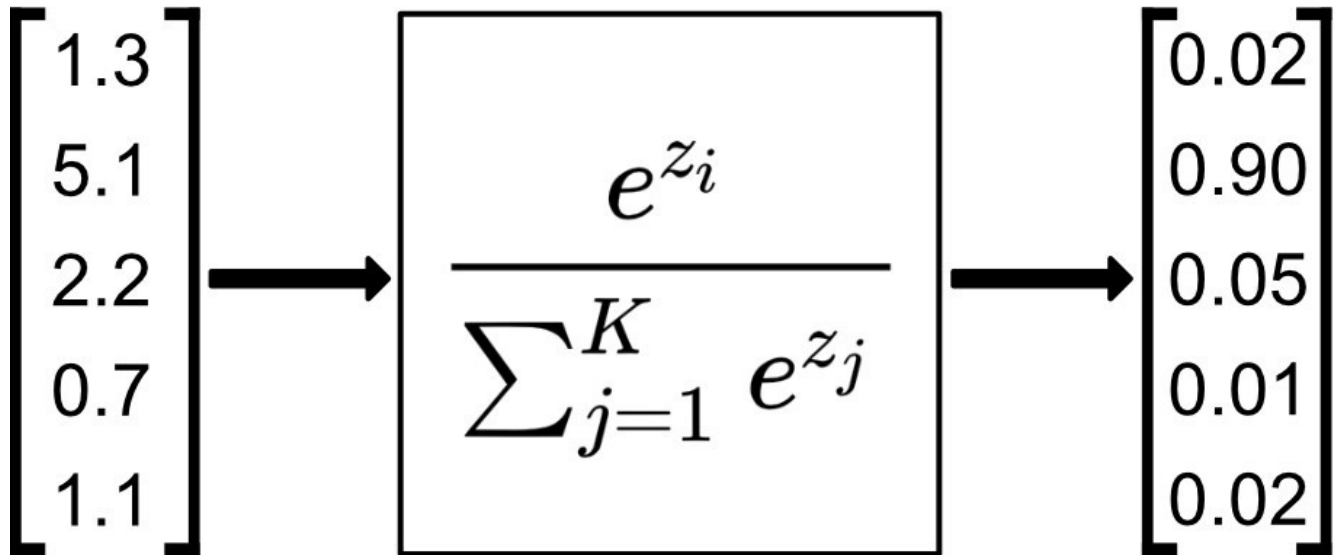


Get unlimited access

Open in app

Output
layerSoftmax
activation function

Probabilities



Any suggestions and corrections are always welcome.

Thank you for reading this lengthy one.

Sign up for Analytics Vidhya News Bytes

By Analytics Vidhya

Latest news from Analytics Vidhya on our Hackathons and some of our best articles! [Take a look.](#)

Get this newsletter

Emails will be sent to alaminbhuyan321@gmail.com.

[Not you?](#)





Get unlimited access

Open in app

