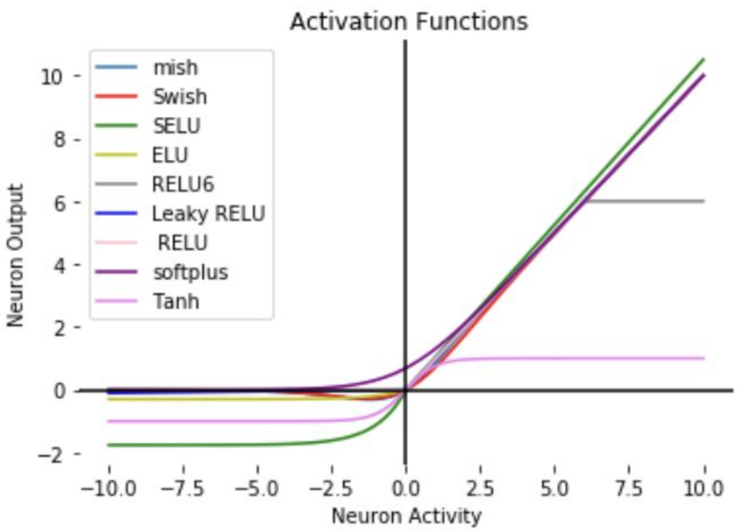
Survey of Activation Functions

By: Hiva Mohammadzadeh

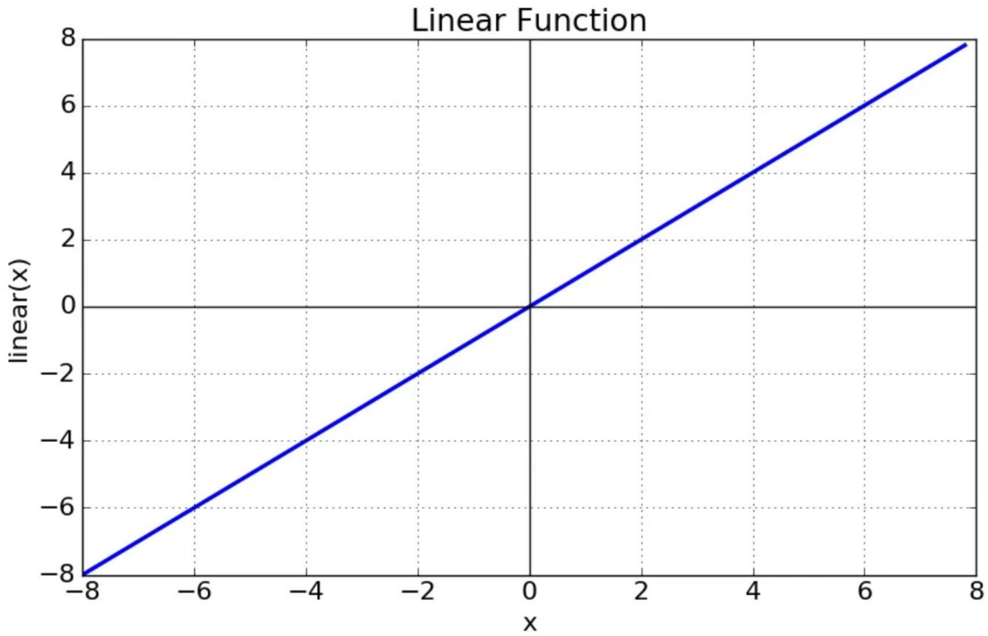
Activation Functions

- Activation functions are used to determine the output of the neural network.
- Depending on the type of the activation function, it maps the output value of the network in between 0 t 1 or -1 to 1 and etc.
- It is also used in the middle of the neural network to allow the model to learn better.



Linear Activation Func

- f(x) = x
- The range is not bounded. (-infinity, infinity)
- Not useful because it doesn't help with the complexity of the parameters of usual data

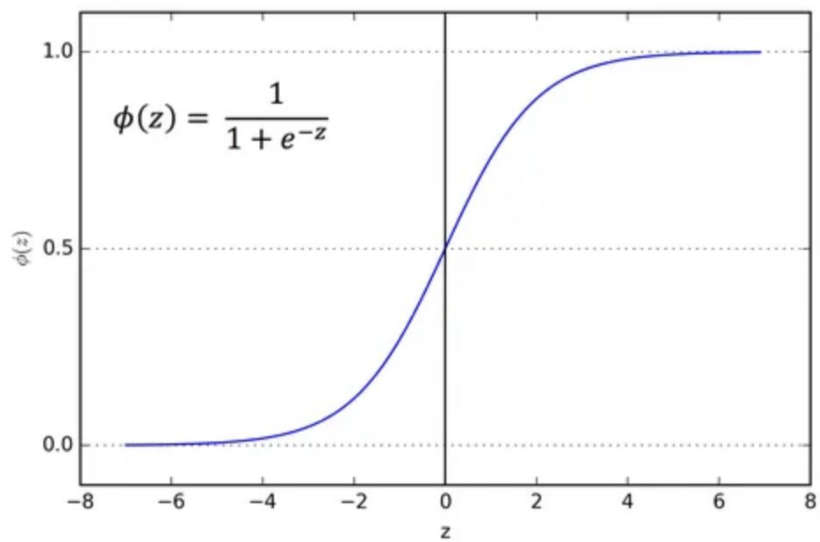


Non-Linear Activation Func

- Most used since they make it easy for the model to generalize or adapt with variety of data and to differentiate between the output.
- Some terminology:
  - Differentiable:
    - Can find the slope of the function curve.
  - Monotonic function:
    - Non increasing or non decreasing function.

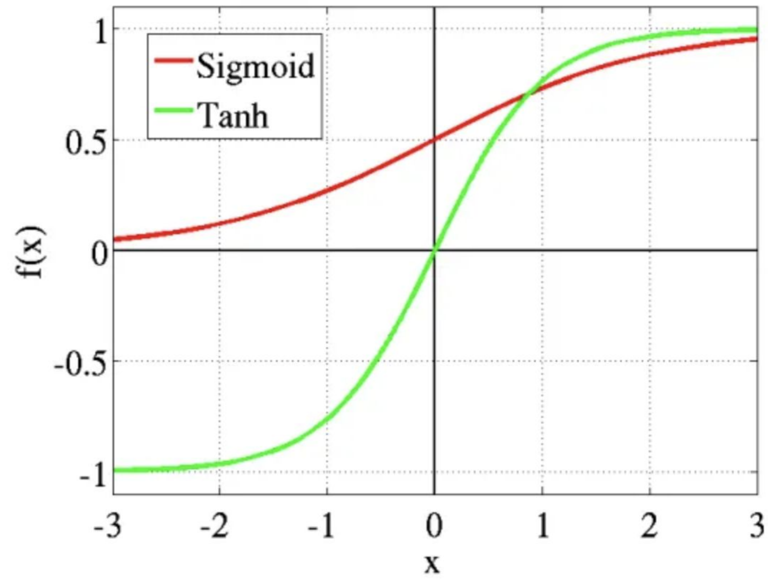
Sigmoid Activation Func

- It’s used where we have to predict the probability as an output.
- Differentiable and monotonic (but not its derivative).
- Range is (0,1)
- Used in FFNs
- Problem:
  - Can make the network get stuck in training time.



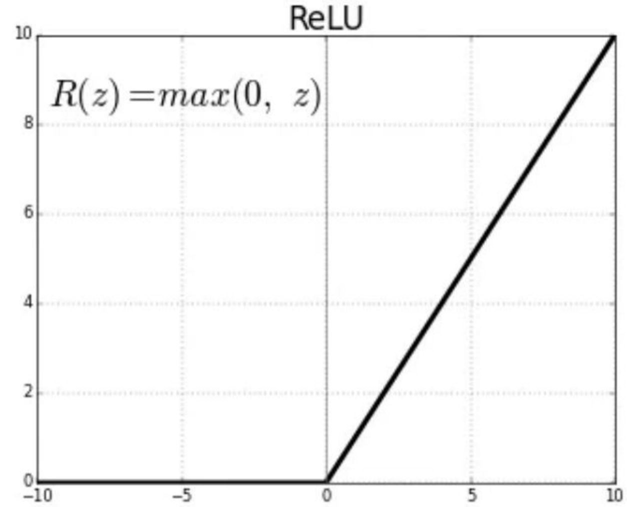
Tanh Activation Func

- Range is (-1, 1)
- Negative inputs will be mapped strongly negative and zero inputs will be mapped near zero.
- Differentiable and monotonic (but derivative not monotonic).
- Used mainly in classification between 2 classes.
- Used in FFNs



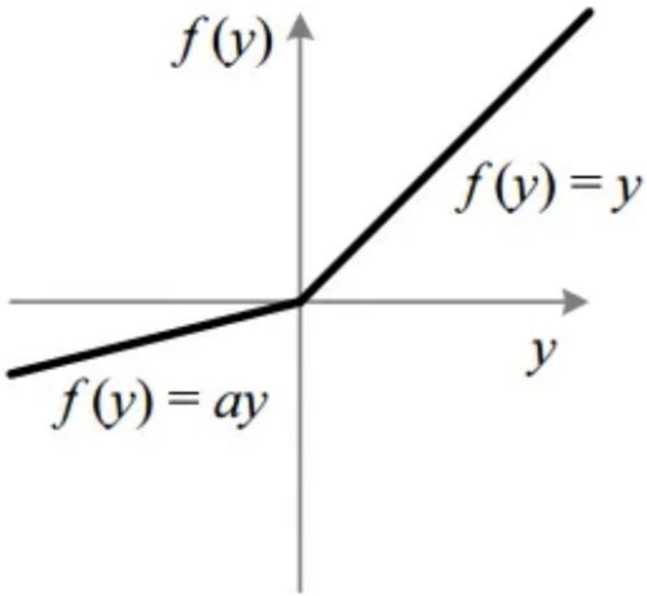
ReLU Activation Func

- Used in all ConvNets and deep learning.
- Half rectified (from bottom).
- Ranges [0,inf)
- Function and derivative are both monotonic
- Problem:
  - negative values become zero → decreases the ability of the model to fit or train from data properly.



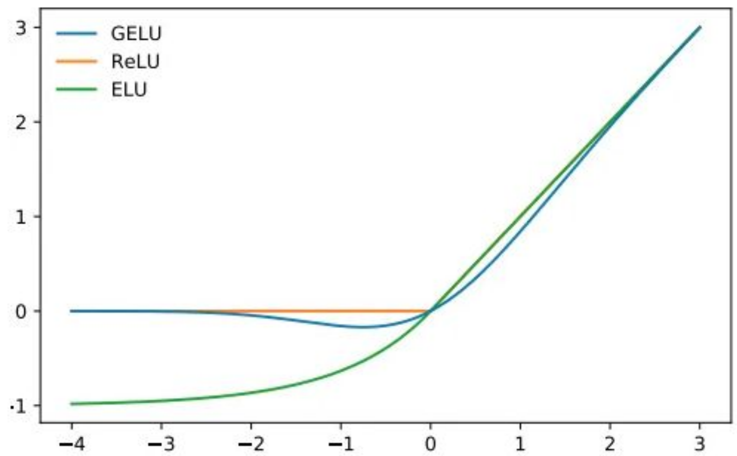
Leaky ReLU Activation Func

- Tries to fix the dying ReLU problem.
- Range (-inf, inf)
- a = 0.01 usually.
- Both function and derivative are monotonic in nature.



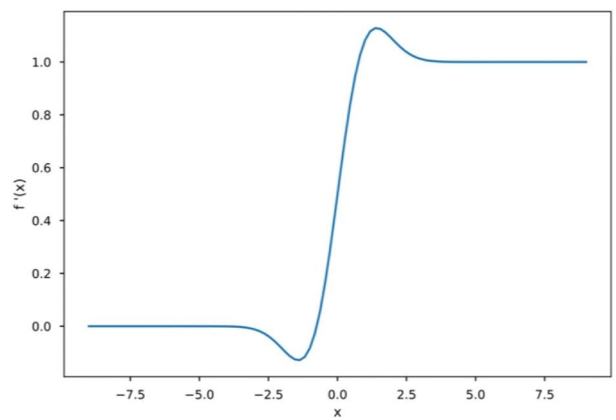
GeLU Activation Func

- GeLU is a way to smooth a ReLU. Better than ELU and ReLU.
- Non-convex, non monotonic, so it exhibits curvature at all points. Can be positive and negative.
- If used, use an optimizer with momentum.
- Faster and better convergence in Neural Networks than sigmoids.
- Has a steep slope near x = 0, which makes it more effective at learning complex patterns in the data.



GELU (μ=0, σ=1), ReLU and ELU (α=1)

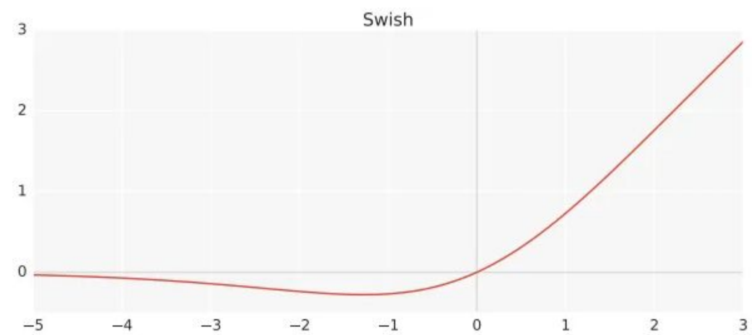
it's Derivative



Swish Activation Func

- Swish works better than ReLU on deeper models across a number of datasets.
- Smooth, non-monotonic function.
- Unbounded above and bounded below
- Has a non-zero gradient at x = 0, which allows the network to learn in this region.

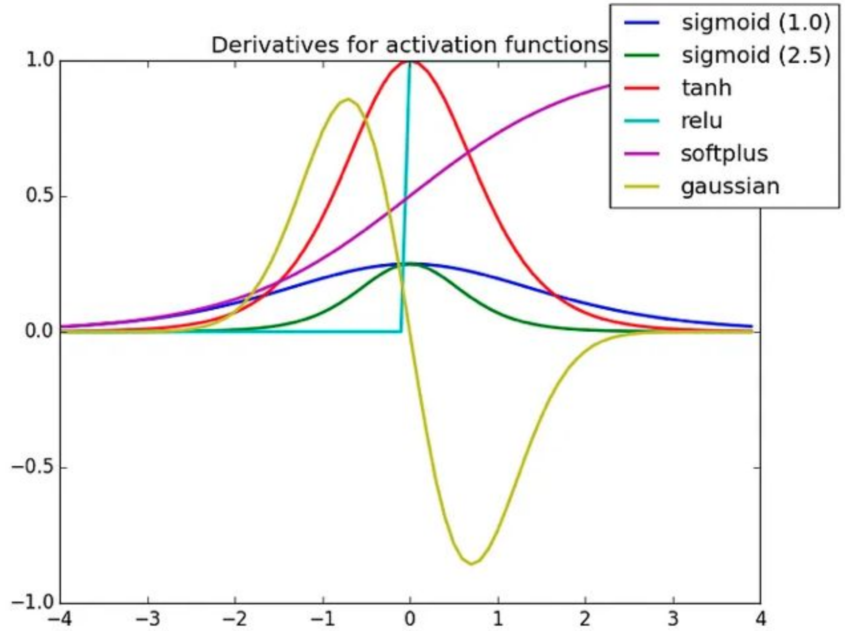
$$\text{swish}(x) = x\,\text{sigmoid}(\beta x) = \frac{x}{1 + e^{-\beta x}}.$$



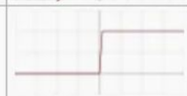
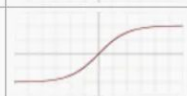
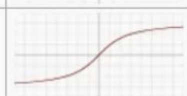
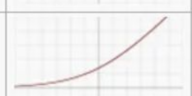
Swish Activation Function Image Source

Derivatives curves

- Use the derivatives when updating the curve, to know in which direction and how much to change or update the curve depending upon the slope.
- Smoother the derivative. Smoother the learning process in the network.



Derivatives

Name	Plot	Equation	Derivative
Identity		$f(x)=x$	$f'(x)=1$
Binary step		$f(x)=\begin{cases} 0 & \text{for } x<0 \\ 1 & \text{for } x\geq 0 \end{cases}$	$f'(x)=\begin{cases} 0 & \text{for } x\neq 0 \\ ? & \text{for } x=0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x)=\frac{1}{1+e^{-x}}$	$f'(x)=f(x)(1-f(x))$
TanH		$f(x)=\tanh(x)=\frac{2}{1+e^{-2x}}-1$	$f'(x)=1-f(x)^2$
ArcTan		$f(x)=\tan^{-1}(x)$	$f'(x)=\frac{1}{x^2+1}$
Rectified Linear Unit (ReLU)		$f(x)=\begin{cases} 0 & \text{for } x<0 \\ x & \text{for } x\geq 0 \end{cases}$	$f'(x)=\begin{cases} 0 & \text{for } x<0 \\ 1 & \text{for } x\geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x)=\begin{cases} \alpha x & \text{for } x<0 \\ x & \text{for } x\geq 0 \end{cases}$	$f'(x)=\begin{cases} \alpha & \text{for } x<0 \\ 1 & \text{for } x\geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x)=\begin{cases} \alpha(e^x-1) & \text{for } x<0 \\ x & \text{for } x\geq 0 \end{cases}$	$f'(x)=\begin{cases} f(x)+\alpha & \text{for } x<0 \\ 1 & \text{for } x\geq 0 \end{cases}$
SoftPlus		$f(x)=\log_e(1+e^x)$	$f'(x)=\frac{1}{1+e^{-x}}$