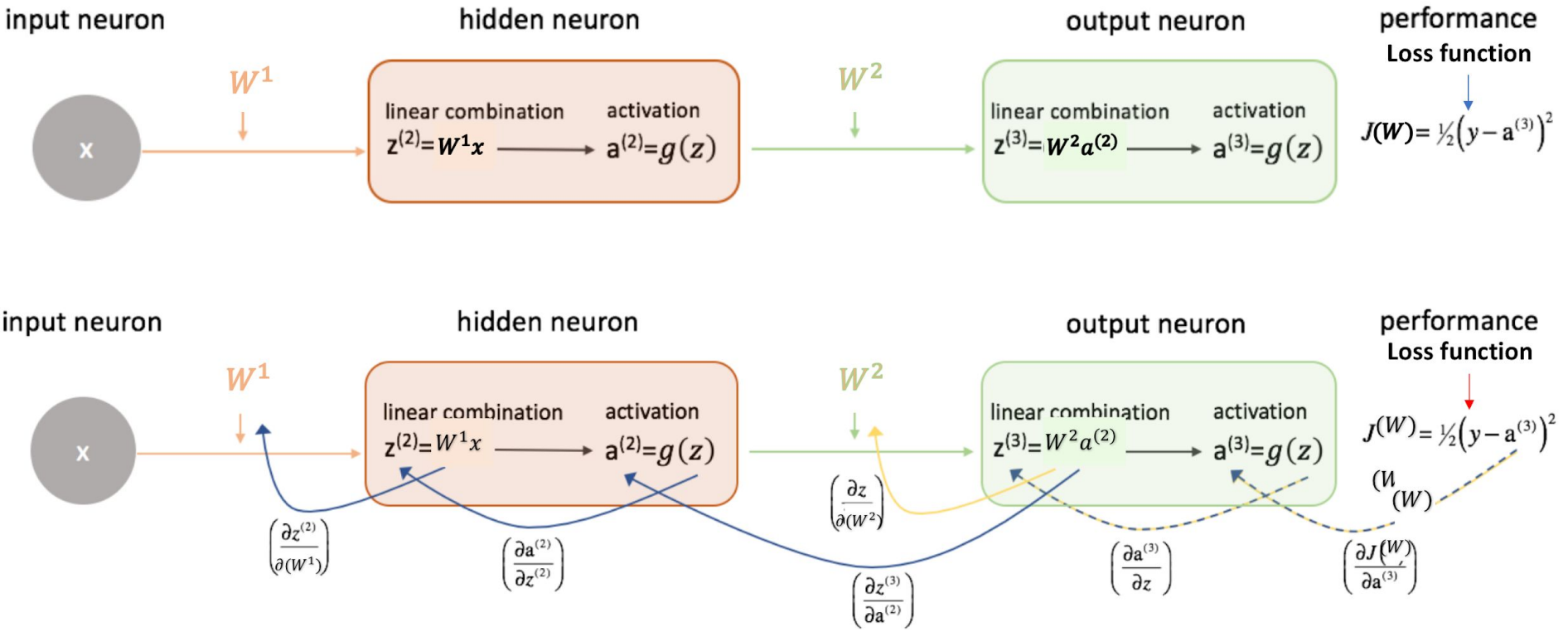
MLP

Binary classification

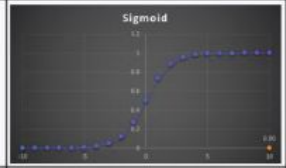
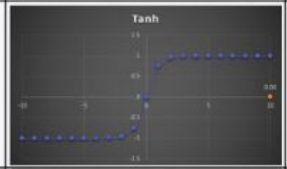
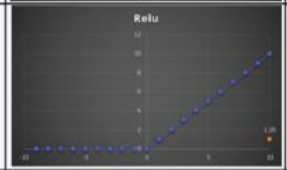
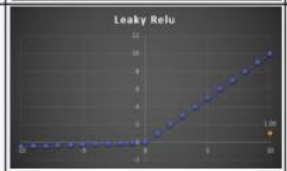
Semen Sorokin

Propagation



Activation functions

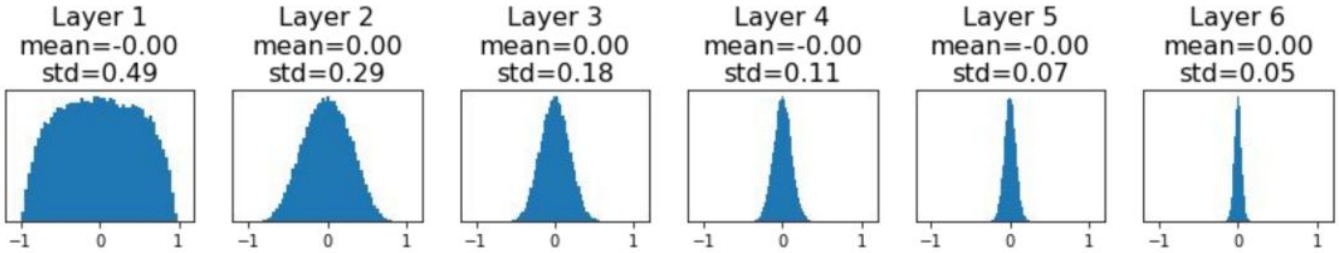
Result without any activation function [-inf, +inf]. Gradient doesn't work.

Name	Plot	Equation	Derivative
Sigmoid		$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
Tanh		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	$f'(x) = 1 - f(x)^2$
Rectified Linear Unit (relu)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Leaky Rectified Linear Unit (Leaky relu)		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$

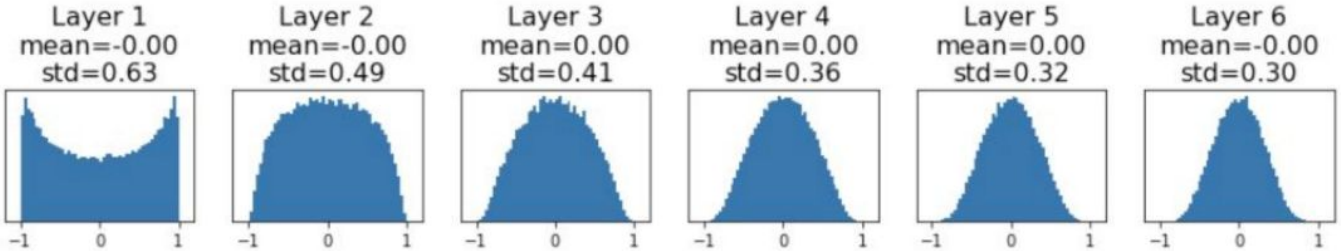
More info

Initializing weights

weight = np.random.rand(in_features, out_features) * 0.01



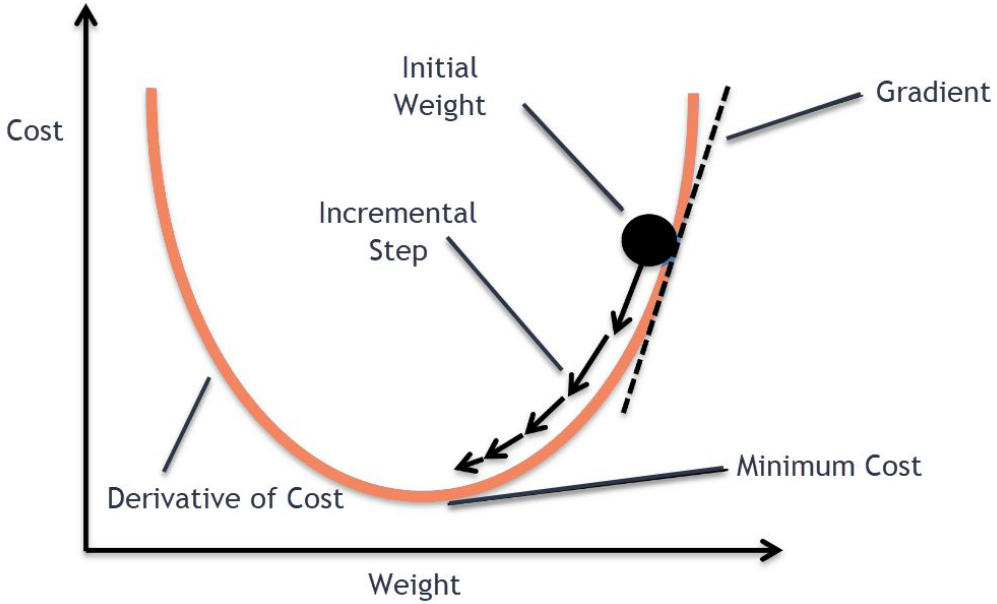
weight = np.random.rand(in_features, out_features) / sqrt(in_features)



More info

(Stochastic, Batch) Gradient Descent

The word ‘stochastic‘ means a system or a process that is linked with a random probability.

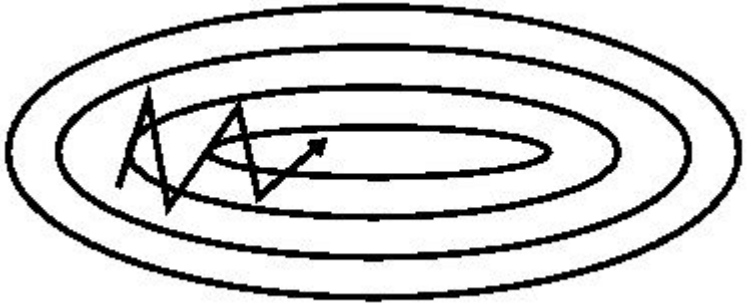


More info

GD+...

1. Momentum
2. Nesterov accelerated gradient
3. Adagrad
4. Adadelta
5. RMSprop
6. Adam
7. AdaMax
8. Nadam
9. AMSGrad

Without momentum



With momentum

More info