All activation functions are nonlinear

Purpose of activation functions🡪 to introduce non-linearities into the network

Linear activation functions produce linear decisions no matter the size

Nonlinearity allows us to approximate arbitrarily.

Universal Approximation Theorem

Batch Gradient Descent

Mini-batch Gradient Descent

Stochastic Gradient Descent

Nesterov Accelerated Gradient🡪modifies how gradient is calculated

Faster Optimizers Momentum

Adaptive Learning Rates

1) AdaGrad

2) AdaDelta

3) RMSProp

4) Adam

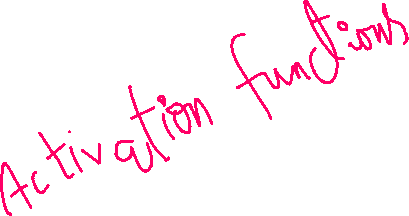
5) Nadam

Parameter Initialization

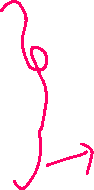
Non saturating activation functions

Sigmoid

ReLU



ELU



Leaky ReLU

Tanh

Maxout

Swish,Mish🡪for advanced models(tranformers,vision tasks)

SELU,ELU🡪deep models prone to vanishing gradients

Normalization

1. Layer Normalization
2. Batch Normalization
3. Group Normalization
4. Instance Normalization

Overfitting

1. Regularization
2. Lasso
3. Ridge
4. Dropout
5. Early Stopping