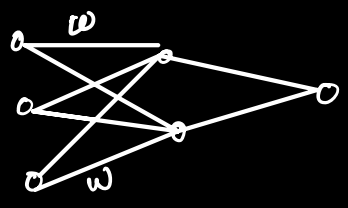
Today's Agenda

1) Activation Functions
2) Optimizers



$f = \boxed{wx+b} + (I)$

$f(x) = x$

Not Activating / Linear

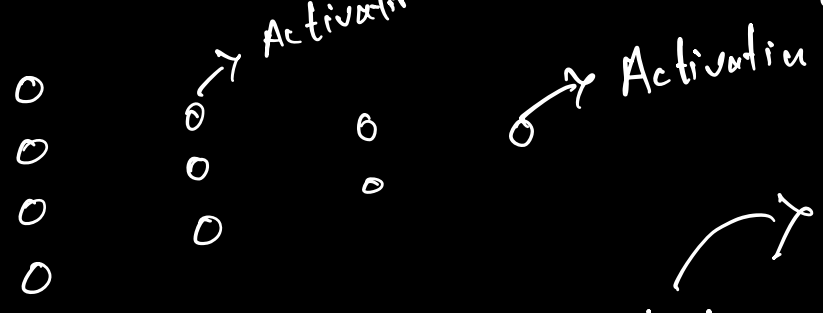
Linear 5%
Non Linear 95%

Linear Regression

Activating → Mathematical Transformation

$\boxed{wx+b}$

{linear} Act fun



Hyperparameter → Before the output in all layers

Parameter → Rules are fixed

Binary Classification

Rules
- Last Activation → Sigmoid
- Loss Function → BCE

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}}$$

$$w_{new} = w_{old} - 0.00000065$$

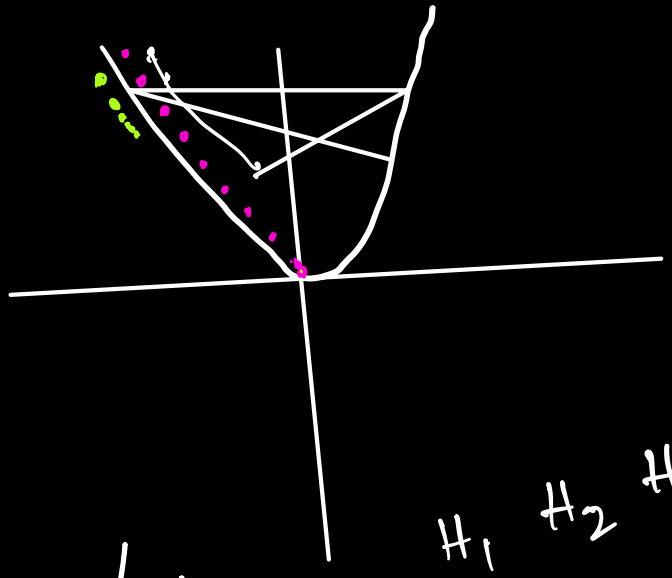
$$\boxed{w_{new} \approx w_{old}}$$

No Weight Update

No Training

Vanishing Gradient Problem

Loss will become stagnant far from minima.



For a particular hidden layer

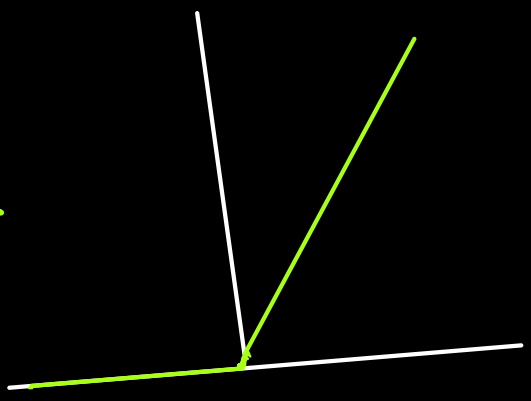
1. All gradients are +ve.
2. All gradients are −ve.

H_1 H_2 H_3

$H_1 \rightarrow$ +ve / −ve

0 → 0 0 0

Relu Activation Func



$$f(x) = \max(0, x)$$

Negative Information is dumped

+ve information is passed.

Information loss

$$z = wx + b$$

$$= z * \text{Act. Fun}^c$$

$$= z * 0$$

$$= 0$$

Dead Neuron

Optimizers

1) Increase the performance of network
   - → W/B update
   - → Act Func update

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}}$$ → slope

Types of Optimizers

1) Gradient Descent → 1) Batch Gradient Descent
2) Stochastic Gradient
3) Mini Batch Gradient Descent

Batch Gradient Descent

1000 → Data Points
Rows in Data

Epochs / Iterations

↓ complete dataset seen by the network

↓

Epoch 1 { 1000 datapoints / 1 iteration } weights will update

Epoch 2 { 1000 datapoints / 1 iteration } weights will update

Epoch the big cycle
Iteration ↳ small cycle

Pass all data points at once and then weight updation

8GB RAM → 1 billion data points

* OOM *
Out of Memory Issue

Advantages

Disadvantages

1) Convergence will happen

1) Resource Intensive
Huge RAM

CPU → RAM
GPU → VRAM

1 Epoch = 1 Iteration

Example :- 1000 datapoints → 100 Epochs

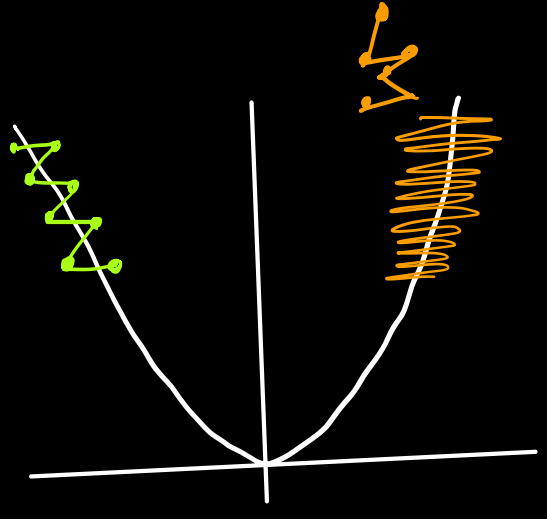
Weight Updation → 100 times it will update

How many total weights updation → 1000 × 100
→ 1000,00

Stochastic Gradient Descent

1000 datapoints

Epoch 1 { 1 iteration / 1 data point
⋮
1000 iteration / 1000 datapoint



Dis Advantages

1) Time taking
2) Resources Wastage
3) Jitters / Noise

Advantage

1) No Resource Issue

Mini Batch Gradient Descent

1000

BS → 90

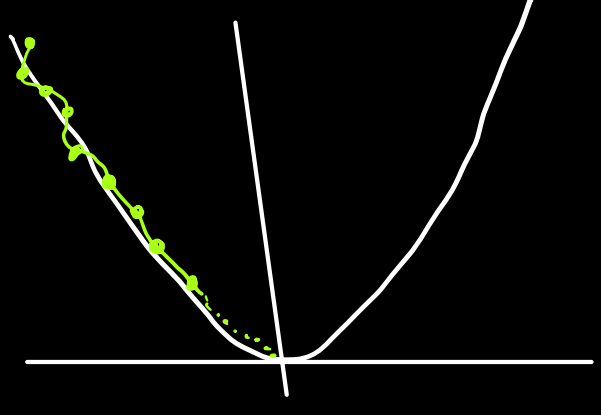
Iterations →

990 = 11 iteration

↳ 10 datapoints

12th iteration → 10

Epoch 1 { iteration 1 / Batch Size 2



Batch Size

1000 datapoints

BS → 100

Iterations → 10

$$= \frac{\text{Total data points}}{\text{Batch Size}}$$

if integer then iterations otherwise int +1

Advantages

1) less Noise
2) Convergence will happen
3) Resources Efficient
4) Time less (SGD)

Disadvantages

1) Noise exists

Batch Size
↳ hyperparameter