Epoch , Iteration

13 September 2023 07:07

Loss vs slope vs intercept

—xox—xox—

Epoch

Whenever entire dataset is being passed to neural net is called as an epoch.

10M

[10M] ↑ Data → NN → $\hat{y}$ → Loss ← [WB] . 1 Epoch

[10M]

[WB] → Data → NN → $\hat{y}$ → Loss ← [WB] 2nd Epoch.

Iteration.

Whenever the subset of data set is being passed to neural net is called as iteration.

[1000 R] — subset →

B1 [100] ✓ → NN → $\hat{y}$ → Loss ← [WB] — Iteration 1 (900)

B2 [100] ✓ → NN → $\hat{y}$ → Loss ← [WB] — Iteration 2

⋮

Iteration 10th.

—xox—

1000 − 100 = 900; 900 − 100 = 800 … 700 … 0 [10th]

1000 → 100, 100, …, 100 — 10 Batches

→ … →| complete

10 iteration. } 1 Epoch

1 Epoch could consist of multiple iteration

Model

1000R → 5 Epoch. 5 iteration.

25 iteration.

—xox—

(B1) ✓ 100 Random Data → NN → $\hat{y}$ → Loss ← [WB] ✓ — Loss. 100 b

[WB] (B2) 100 Data → NN → $\hat{y}$ → Loss ← [WB]

(WB) (B3)

B^{10} Loss ← 10th. [WB] ✓ 1 Epoch.

} Epoch.

B1 / 100

(GD) was only the algorithm that works on epoch concept

Stochastic gradient descent

└→ Random.

In case of gradient descent we use to take entire dataset in forward propagation

	x_1	x_2	x_3	x_4	yact
✓ R1	x	x	x	x	x
R2	x	x	x	x	x
✓ R3	x	x	x	x	x
R4	x	x	x	x	x

Random.

R3 → NN → $\hat{y}$ → Loss ← [WB]

R1 → NN → $\hat{y}$ → Loss ← [WB]

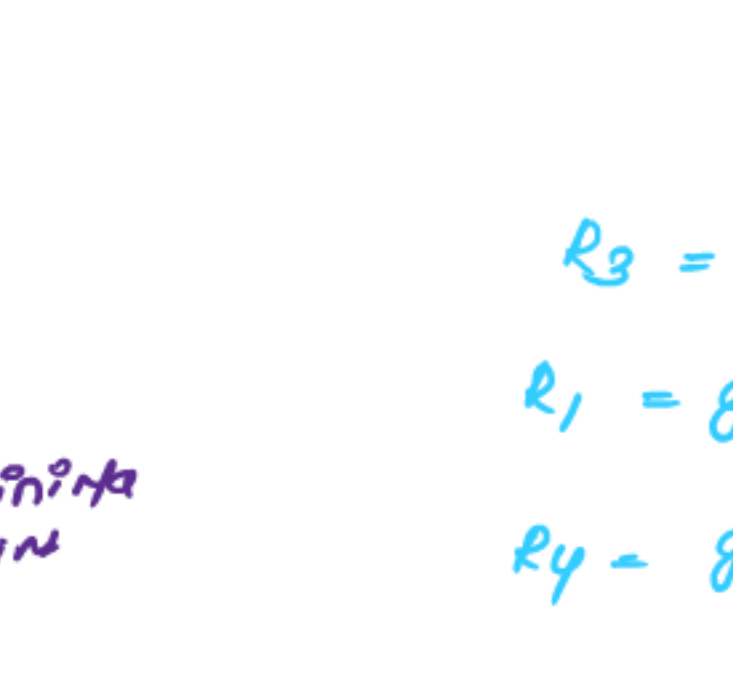
R4 → NN → $\hat{y}$ → Loss ← [WB]

R2 → NN → $\hat{y}$ → Loss ← [WB]

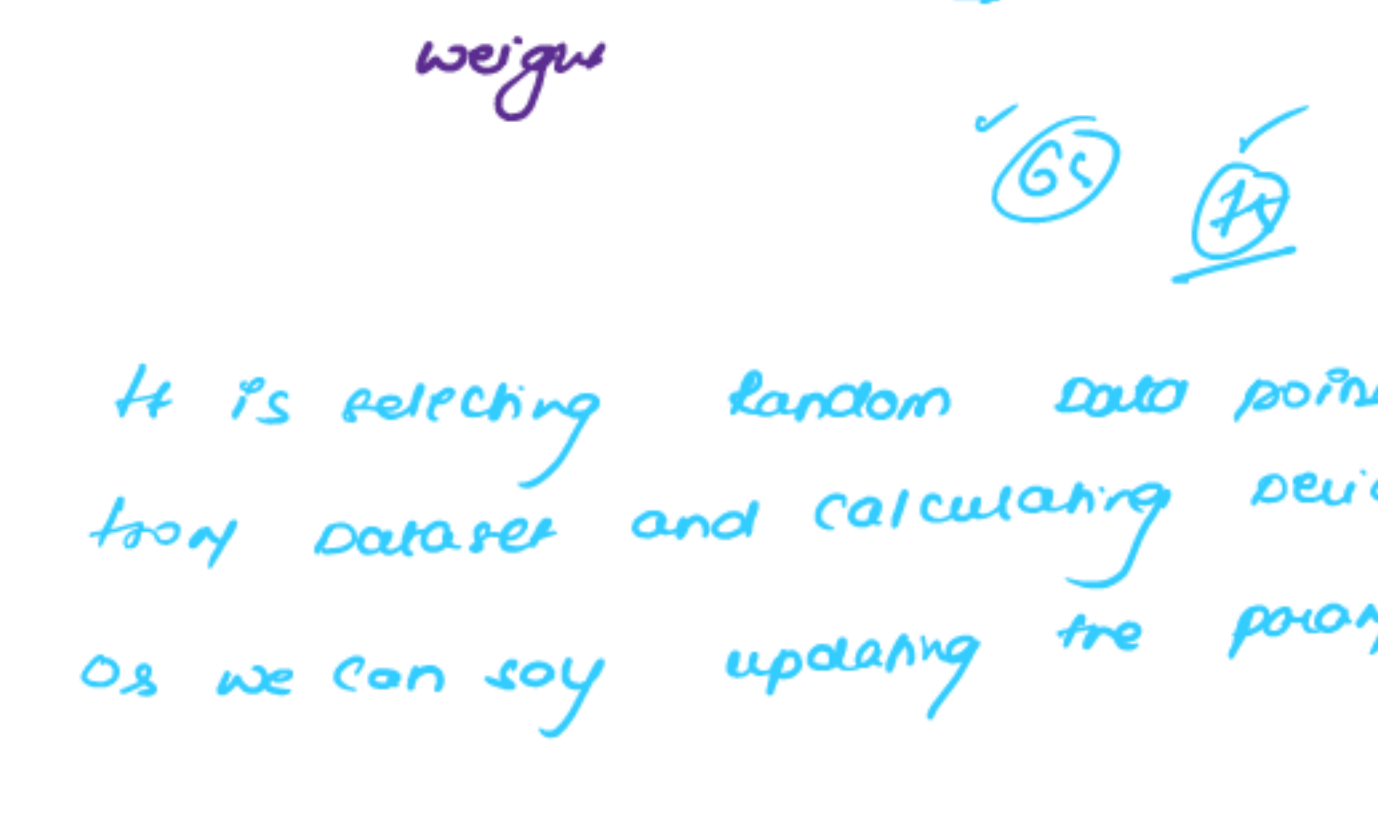
4 iteration 4 time [WB]
1 Epoch

1 M GD 1 epoch 1 WB
SGD 1 Epoch 1M WB

$w_{new} = w_{old} - \alpha \frac{\partial L}{\partial w_{old}}$ → (LR) 0 – 1, 0.01, 0.001



[SGD] is prone to multiple local minima.



(GS) (?)

R3 = good = Loss ↓
R1 = good = Loss ↓
R4 = good = Loss ↓
R2 = Bad = Loss ↑

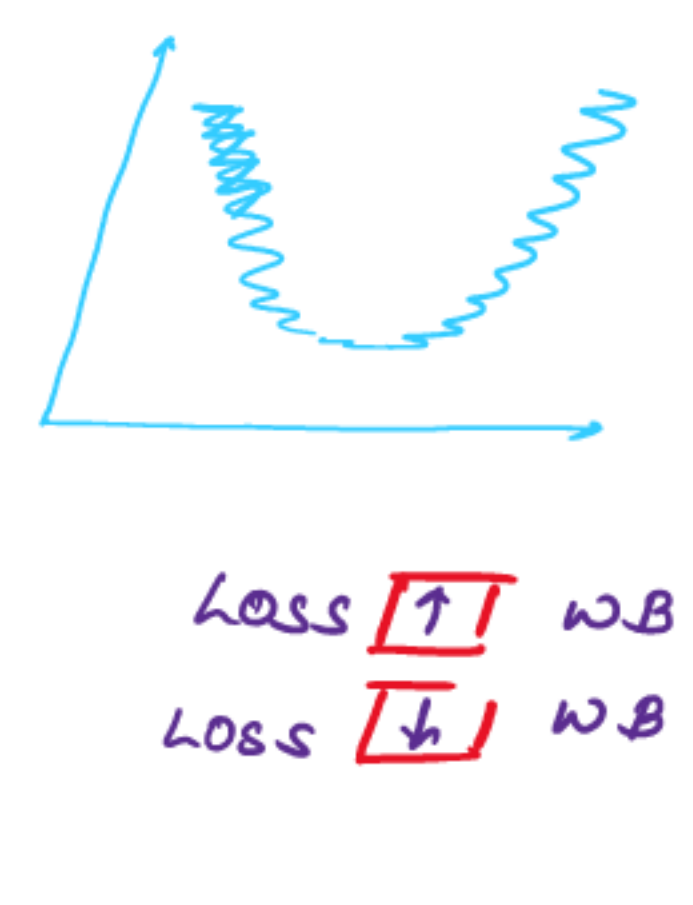
It is selecting random data point from dataset and calculating derivative or we can say updating the parameter

yact (15) $\hat{y}$ (14) = similar (WB ↓)

yact (15) $\hat{y}$ (5) = Loss (↑) (WB ↑)

* Batch size = 1
no. of Batches = Data size }

→ path is noisy
→ Computationally not expensive.

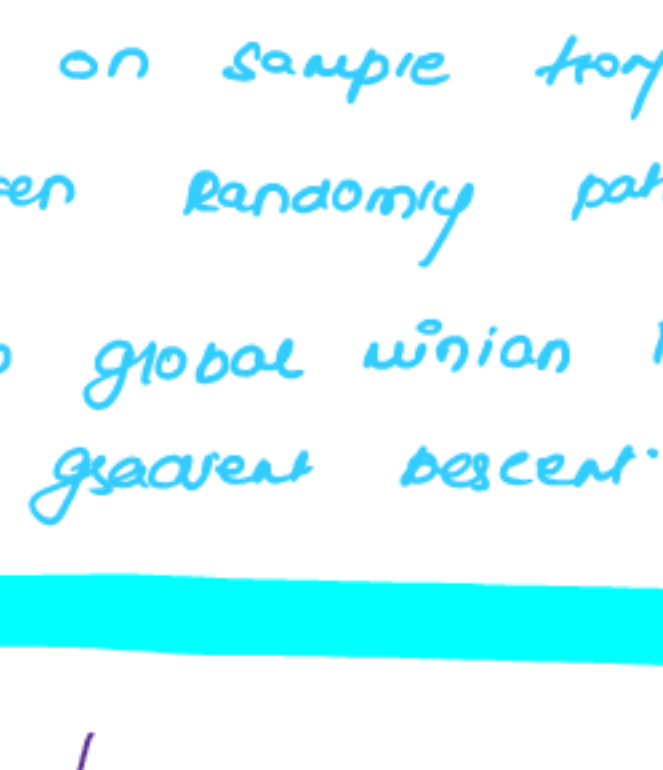


Loss ↓
Loss ↑

yact − $\hat{y}$ [WB]
15 14 → (1)

Loss [↑] WB [↑]
Loss [↓] WB [↓]

yact $\hat{y}$ [WB] (↑)
15 5 → 10



[test] (GD)

as on sample from data is choosen randomly path taken to reach global minima is usually noisier than gradient descent.