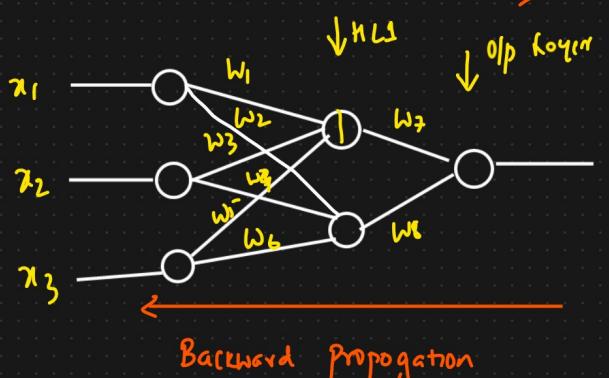


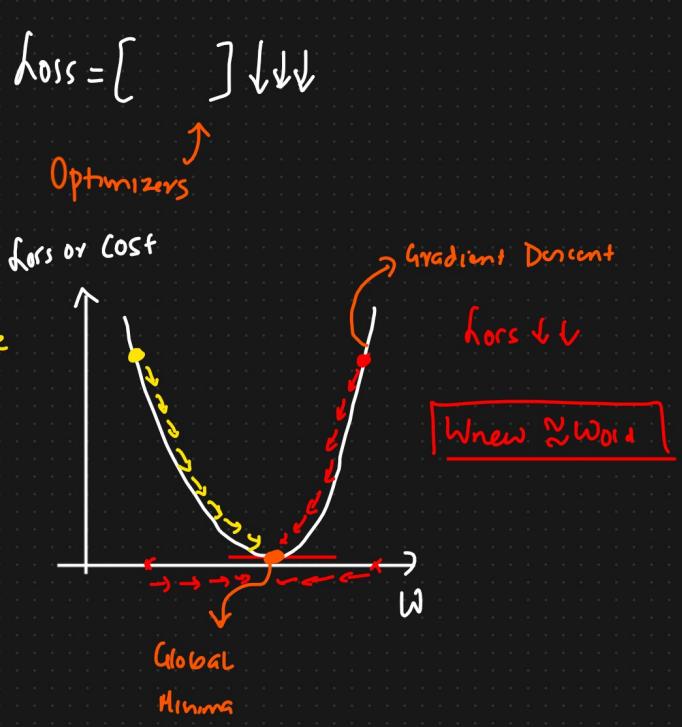
Optimizers

- ① Gradient Descent
- ② Stochastic Gradient Descent (SGD)
- ③ Mini batch SGD
- ④ SGD With Momentum
- ⑤ Adagrad and RMSProp
- ⑥ Adam Optimizers

Gradient Descent Optimizer



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



MSE

$$\text{Loss fn} = (y - \hat{y})^2 \quad \text{Cost fn} = \frac{1}{h} \sum_{i=1}^h (y_i - \hat{y}_i)^2 \quad \text{Dataset} = 1000 \text{ Data points}$$

Epochs, Iteration

1 Epoch = 1 Iteration

1 Epoch { 1000 datapoint → \hat{y}_i = Cost function w/w
Weights will get updated }

1 Epoch = 1 Iteration

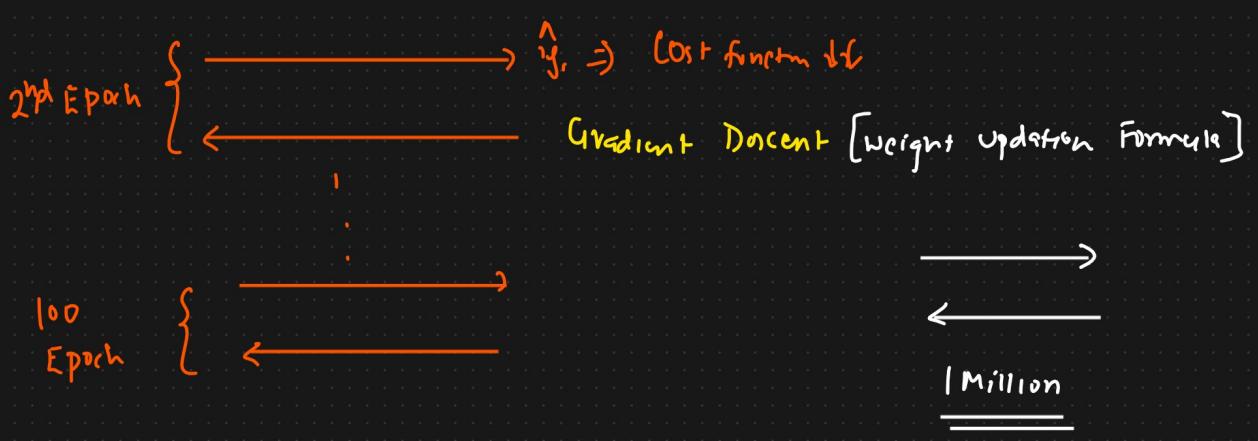
Gradient Descent optimizer [Weight update]

10 Iteration

$100/10 = 100$

100

← → ← → ← → ← → ← →



Advantages

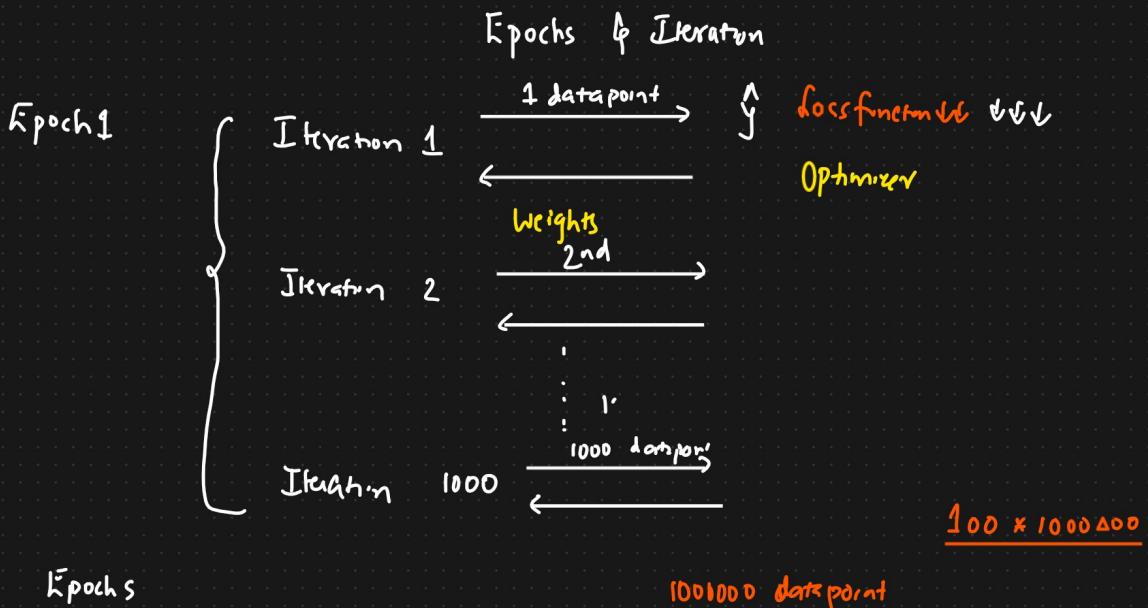
① Convergence will happen.

① Huge Resource RAM, GPU
 \Downarrow

Resource Intensive -

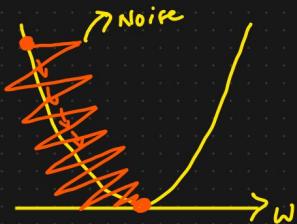
② Stochastic Gradient Descent (SGD)

1000 datapoint



Advantage

① Solve Resource Issue



Disadvantage

① Time Complexity \uparrow

- ④ Convergence will also take more time.
- ④ Noise gets introduced

$\rightarrow \hat{y}$ Cost

③ Mini Batch SGD

Epoch, Iteration, Batch-size

$$\text{No. of iterations} = \frac{100000}{1000} = 100 \text{ iterations}$$

Data points = 100000

batch.size = 1000

MSGD

$$\text{Cost fn} = \sum_{i=1}^{1000} (y_i - \hat{y}_i)^2 \downarrow$$

Epoch 1

Iteration 1

Optimizer \Rightarrow Mini Batch SGD

change the weights

1000

\downarrow

[8gb] 1

Iteration 2

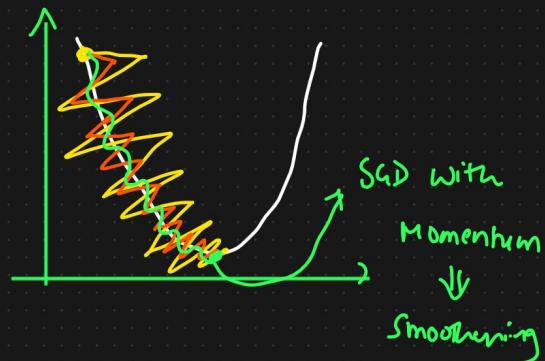
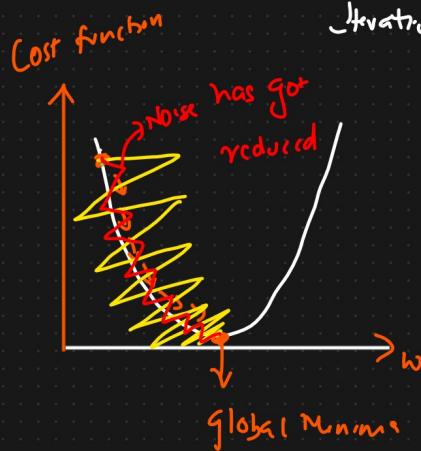
[16gb] 5000

Iteration 3

Iteration 100

Iteration 100

Batch Size



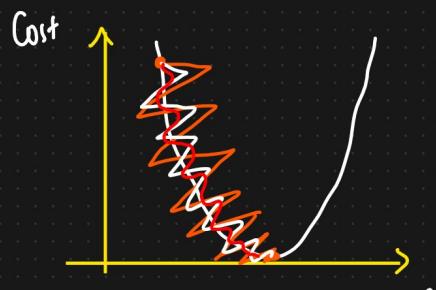
Advantages

- ① Convergence speed will increase
- ② Noise will be low when compared to SGD
- ③ Efficient Resource Usage (RAM)

Disadvantage

- ① Noise still exists

④ SGD With Momentum



Weight Update formula

$$w_{\text{new}} = w_{\text{old}} - \eta \frac{\partial h}{\partial w_{\text{old}}} \quad \text{Learning Rate}$$

$$b_{\text{new}} = b_{\text{old}} - \eta \frac{\partial h}{\partial b_{\text{old}}}$$

$$w_t = w_{t-1} - \eta \left[\frac{\partial h}{\partial w_{t-1}} \right]$$

Exponential Weight Average {Smoothing} } \Rightarrow ARIMA, SARIMAX

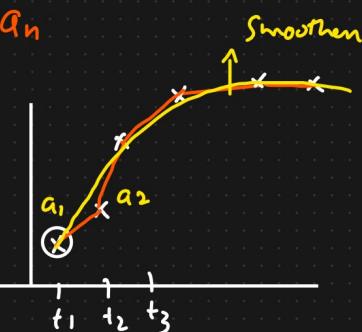
Time = $t_1, t_2, t_3, t_4, \dots, t_n$ Time Series

Values = $a_1, a_2, a_3, a_4, \dots, a_n$

$$V_{t_1} = a_1$$

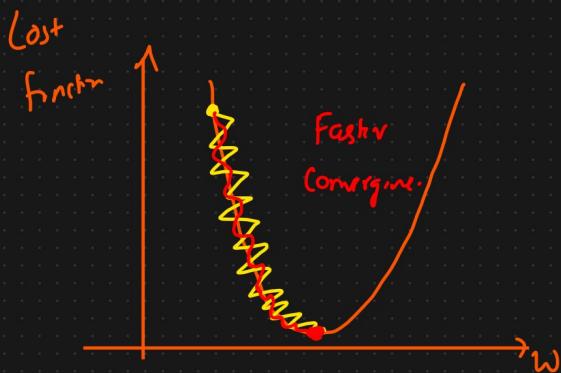
$$V_{t_2} = \boxed{\beta} * V_{t_1} + (1-\beta) * a_2$$

$$\begin{aligned} \beta = 0.95 \\ V_{t_2} = 0.95 * a_1 + (0.05) a_2 \end{aligned}$$



$$V_{t_3} = \beta * V_{t_2} + (1-\beta) * a_3$$

$$= 0.95 \left[0.95 * a_1 + (0.05) a_2 \right] + (0.05) * a_3$$

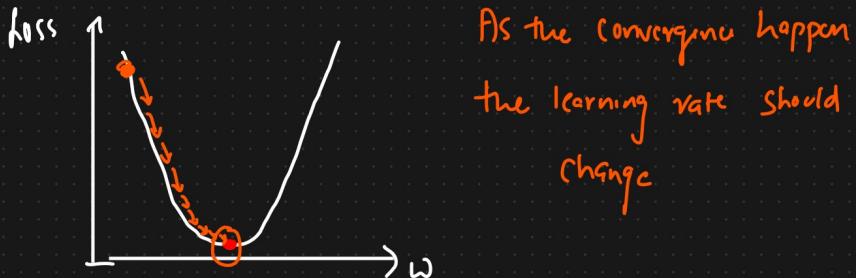


Advantage

- ① Reduces the noise
- ② Quick Convergence

⑤ Adagrad = Adaptive Gradient Descent $\eta = \underline{\text{fixed}} \Rightarrow \underline{\text{Dynamic learning}}$

$$w_t = w_{t-1} - \eta \left[\frac{\partial h}{\partial w_{t-1}} \right] \quad \begin{array}{l} \text{Learning} \\ \text{Rate} = 0.001 \end{array}$$



$$w_t = w_{t-1} - \eta' \left[\frac{\partial h}{\partial w_{t-1}} \right] \quad \begin{array}{l} 0.00001 \approx 0 \\ d_t \uparrow \downarrow \end{array}$$

$$\eta' = \frac{\eta}{\sqrt{d_t + \epsilon}} \quad \begin{array}{l} \eta' \downarrow \text{if } d_t \uparrow \\ \epsilon \text{ is a small value} \end{array}$$

$$d_t = \sum_{i=1}^t \left(\frac{\partial h}{\partial w_i} \right)^2$$

$t=1 \quad t=2 \quad t=3 \quad w_t \approx w_{t-1}$

$$\eta = 0.01 \quad \eta = 0.005 \quad \eta = 0.003$$

Disadvantage

- ① $\eta' \rightarrow$ Possibility to become a very small value ≈ 0

⑥ Adadelta And RMSprop

Exponential Weight Average

$$\beta = 0.95 \quad S_{dw_t} = 0$$

$$S_{dw_t} = \beta * S_{dw_{t-1}} + (1-\beta) \left(\frac{\partial h}{\partial w_{t-1}} \right)^2$$

$$\eta' = \frac{\eta}{\sqrt{S_{dw_t} + \epsilon}}$$

Dynamic LR + Smoothing [EWA]

$$w_t = w_{t-1} - \eta' \frac{\partial h}{\partial w_{t-1}}$$

⑦ Adam Optimizer

SGD with Momentum + RMSprop [Dynamic LR + Smoothing]

$$w_t = w_{t-1} - \eta' V_{dw}$$

$$b_t = b_{t-1} - \eta' V_{db}$$

Weight Updation

Bias Updation

$$\eta' = \frac{\eta}{\sqrt{S_{dw_t} + \epsilon}}$$

EWA

$S_{dw_t} = 0$

$$S_{dw_t} = \beta * S_{dw_{t-1}} + (1-\beta) \left(\frac{\partial h}{\partial w_{t-1}} \right)^2$$

[EWA]

$$V_{dw_t} = \beta * V_{dw_{t-1}} + (1-\beta) \frac{\partial h}{\partial w_{t-1}}$$

$$V_{db_t} = \beta * V_{db_{t-1}} + (1-\beta) \frac{\partial h}{\partial b_{t-1}}$$

\Rightarrow Momentum
↳ Smoothing