Optimizer

1. G.D. $\Rightarrow$ $W_{new} = W_{old} - \eta \frac{\partial L}{\partial W}$ $\Rightarrow$ Entire Data
2. SGD $\Rightarrow$ Single Point at a time
3. Mini-Batch GD $\Rightarrow$ Batch
4. Momentoum (SGD with momentum)
5. NAG (Nesterov Accelrated graident)
6. Adgarde (Adaptive gradient)
7. RMS Prop (root mean Squre ProPogation)
8. ADAM

- Reduce the Loss
- Train a tariable Parameter in Best Possible way.
- Fast convesny to global minma

1. epoch $\Rightarrow$ Complete cycle with entire Data
   FP + BP
2. Iteration — SGD, Mini Batch GD
3. Batch size

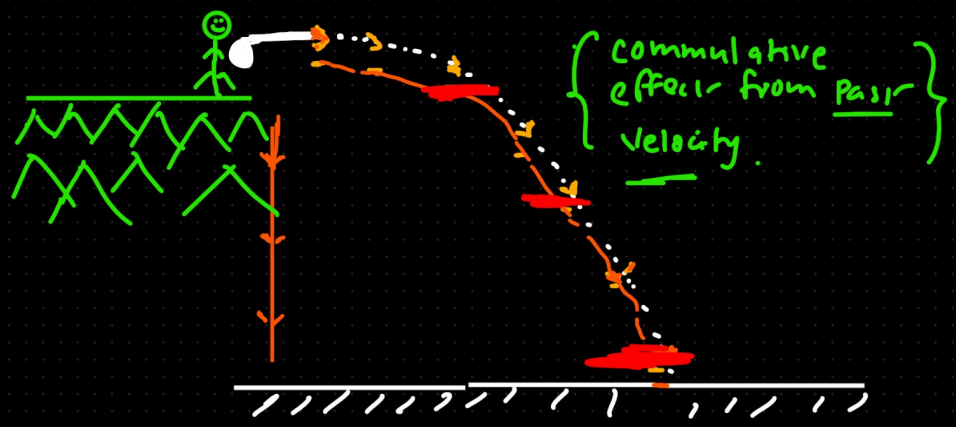
AD

Minute change in Gradient Decent itself

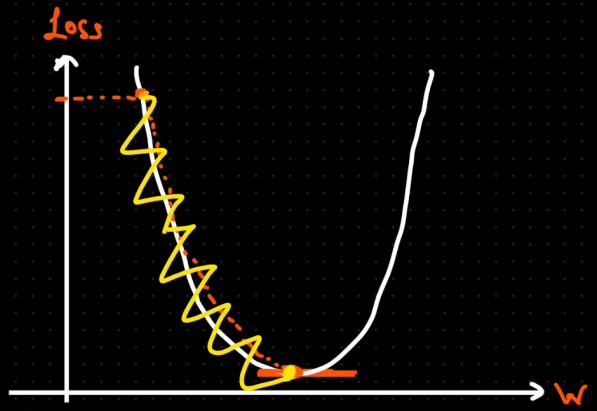
1 Momentum (SGD with momentum)

↳ Phy (Accerlation) change into the Postion

$M \times V_f =$



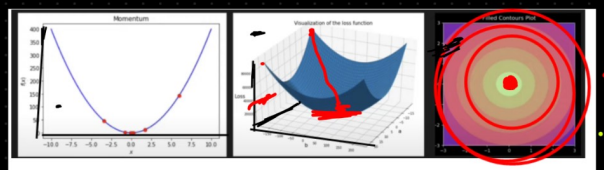
GD, SGD, Mini GD



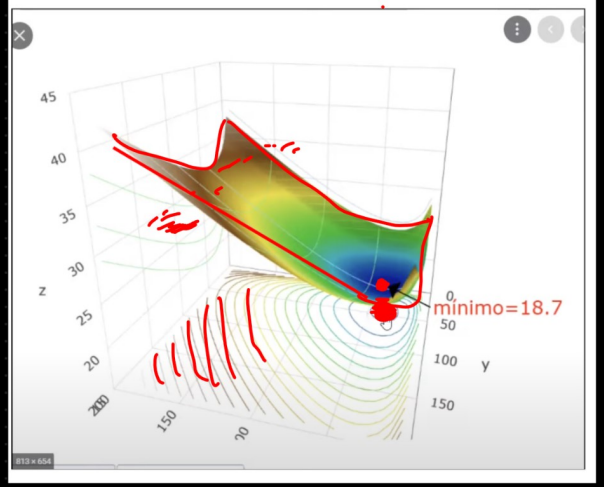
GD ⇒ More time to converge

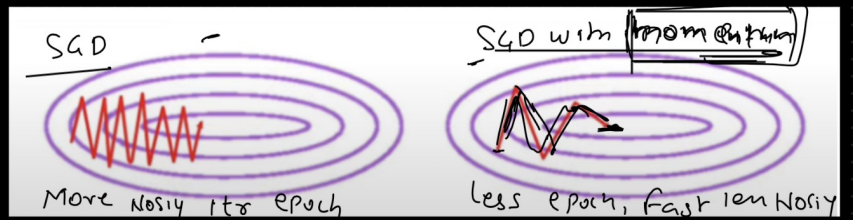
More memory

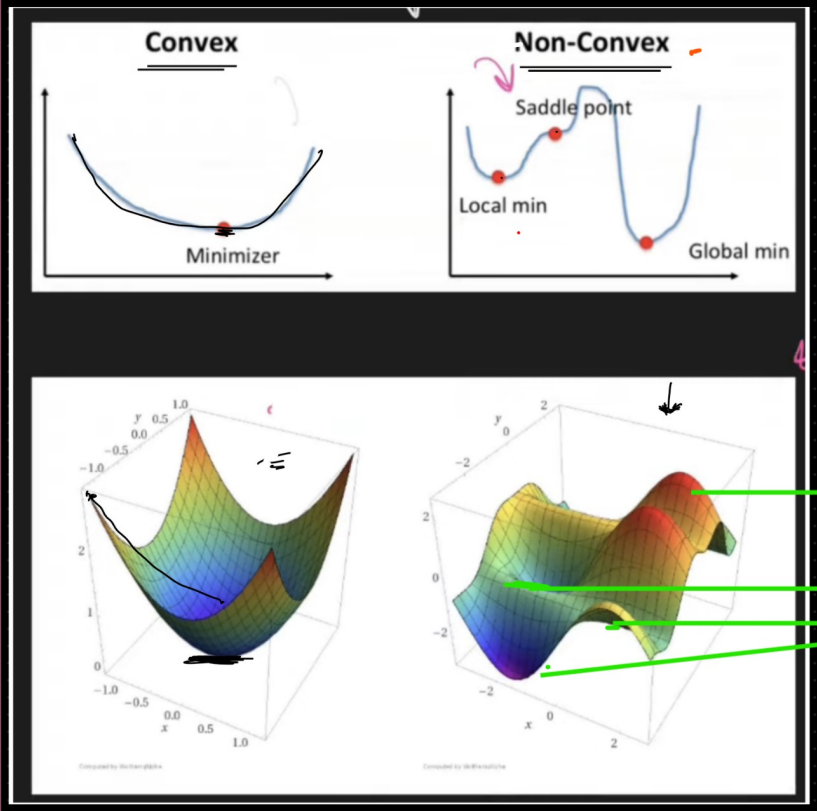
SGD = epoch, iteration, Noise



← Contour







Flat Surface ⇒ Slope = 0 (won't be W update)

Curvature

Saddal Point

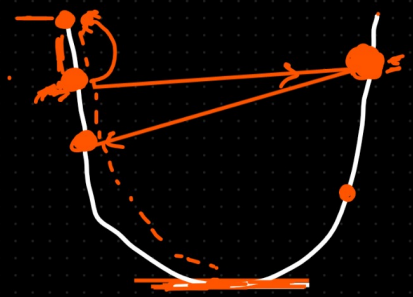
Local minima

Noisy

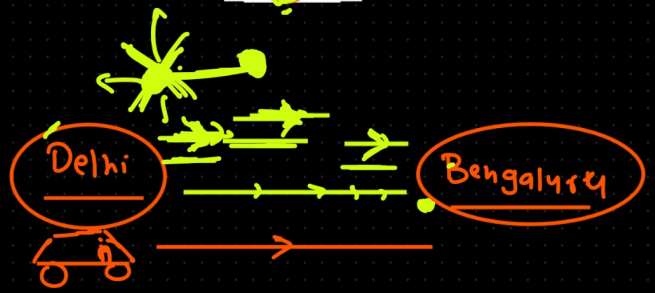
Global minima

G.D.

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



Right Dir, fast



SGD

Momentum

$$W_{t+1} = W_t - v_t$$

Mom = M × V

$$V_t = \beta V_{t-1} + \eta \frac{\partial L}{\partial w}$$

β = decay factor

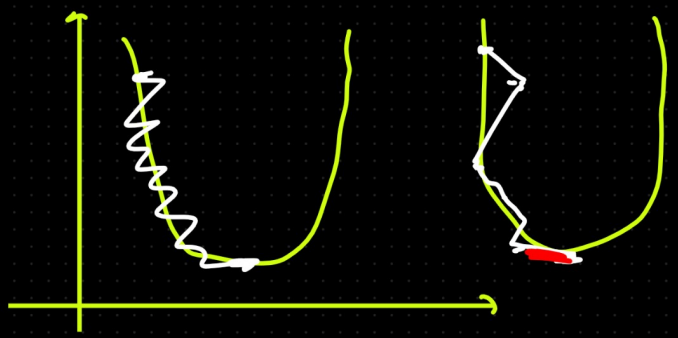
$0 < \beta < 1 \Rightarrow$ 0.9

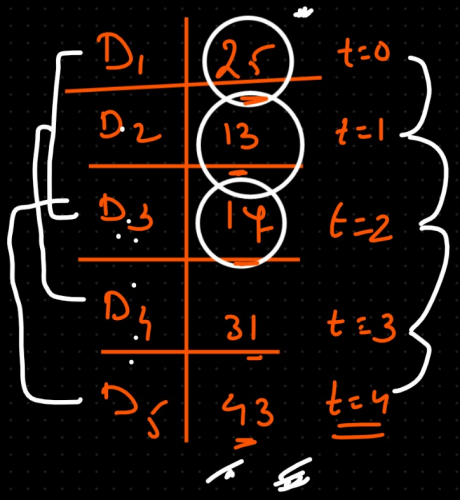
$$W_{t+1} = W_t - \left[\beta V_{t-1} + \eta \frac{\partial L}{\partial w}\right]$$

$\Rightarrow$ EWMA

Adding the momentum

Considering a history of gradient





SMA $\Rightarrow$ Window = 3, 5, 7, 8, 9, 10, ... n

MA $\Rightarrow$ trend, Pattern, Smoothing

EWMA

= EWMA =

DL optimizer

$$V_t = \beta \times V_{t-1} + (1-\beta)\theta_t$$

- V_t: at t time EWMA
- V_{t-1}: at t-1 time EWMA
- θ_t: Value at t time

$V_0 = \theta_0 = 25$

$V_0 = \theta_0 = 0$

$0 < \beta < 1$

$\beta = 0.9$

$V_1 = 0.9 \times 0 + (0.1) \times (13)$

$V_1 = 0 + 1.3$

$V_1 = 1.3$

$V_2 = 0.9 \times 1.3 + (0.1) \times (17)$

$= 1.17 + 1.7$

$$V_t = \beta V_{t-1} + (1-\beta)\theta_t$$

$$V_0 = 0$$

$$V_1 = (1-\beta)\theta_t$$

$$V_2 = \beta V_1 + (1-\beta)\theta_t$$

$$= \beta(1-\beta)\theta_2 + (1-\beta)\theta_2$$

$$V_3 = \beta V_2 + (1-\beta)\theta_3$$

$$= \beta\left[\beta(1-\beta)\theta_2 + (1-\beta)\theta_2\right] + (1-\beta)\theta_3$$

$$= \beta^2(1-\beta)\theta_2 + \beta(1-\beta)\theta_2 + (1-\beta)\theta_3$$

$$= (1-\beta)\left[\beta^2\theta_1 + \beta\theta_2 + \theta_3\right]$$

θ_1 value $q+1$, θ_2 value $q+2$, θ_3 3

$0 < \beta < 1$

$\beta = 0.9$

0.5×0.5

0.81

?

$0.81\theta_1 + 0.9\theta_2 + 1\times\theta_3$

NAG ⇒ Nestrov Accelarated Gradient (Updated Version of Momentum)

G.D. ⇒ $W_{new} = W_{old} - \eta \frac{\partial L}{\partial W}$

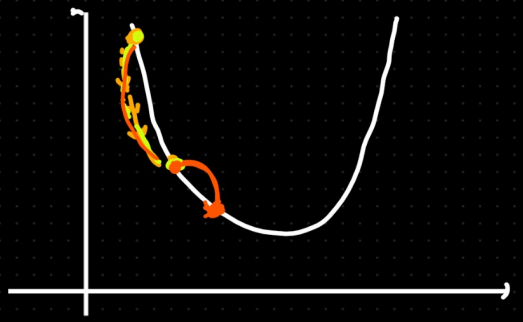
Momentum in GD

$$W_{t+1} = W_t - V_t$$

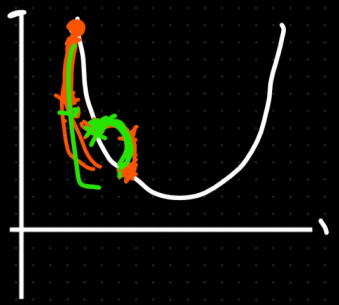
$$V_t = \beta V_{t-1} + \eta \frac{\partial L}{\partial W}$$

$$W_{t+1} = W_t - \left[\beta V_{t-1} + \eta \frac{\partial L}{\partial W}\right]$$

history of Gradient + gradient at point



NAG ⇒ look ahead factor



Gradient at point

$$\beta V_{t-1} = W_t - W_{la}$$

$$W_{la} = W_t - \beta V_{t-1}$$

$$V_t = \beta V_{t-1} + \eta \frac{\partial L}{\partial W_{la}}$$

$$V_t = (W_t - W_{la}) + \eta \frac{\partial L}{\partial W_{la}}$$

$$W_{t+1} = W_t - V_t$$

Adagrad (Adaptive Gradient)

GD ⇒ $W_{new} = W_{old} - \eta \frac{\partial L}{\partial w}$ → Learning rate.

Dynamically

- Momentum
- NAG
  - Updated momentum
  - Lookahead factor

Adding past history of Gradient ⇒ Acceleration

$$W_t = W_{t-1} - \eta \frac{\partial L}{\partial w}$$

$$W_t = W_{t-1} - \frac{\eta}{\sqrt{V_t + \epsilon}} \times \frac{\partial L}{\partial w}$$

η → lr, $\frac{\partial L}{\partial w}$ → Diff.

$$V_t = V_{t-1} + \left(\frac{\partial L}{\partial w}\right)^2$$

Past gradient sum of square

ϵ = Small value

$\sqrt{V_t + \epsilon} \neq 0$

$\epsilon \neq 0$

= RMS-Prop ⇒

Adadelta

Dynamically ↓

small

$$W_{t+1} = W_t - \frac{\eta}{\sqrt{V_t + \varepsilon}} \times \frac{\partial L}{\partial W}$$

Adagrad

= Adagrad

$$V_t = V_{t-1} + \left(\frac{\partial L}{\partial W}\right)^2$$

- RMSprop ↓

$$V_t = \beta V_{t-1} + (1-\beta)\left(\nabla W_t\right)^2$$

ADAM ⇒

foundation

- VGD / SGD / MBGD
- Momentum → NAG ← phy
- Adagrad → RMSProp ← Dynamic learning rate

ADAM ⇒ Momentum + Dynamic learning rate

$$= W_{t+1} = W_t - \frac{\eta}{\sqrt{V_t + \varepsilon}} * m_t$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w}$$

$$V_t = \beta_2 V_{t-1} + (1-\beta_2)\frac{\partial L}{\partial w}$$

Friday → 8 to 11 pm

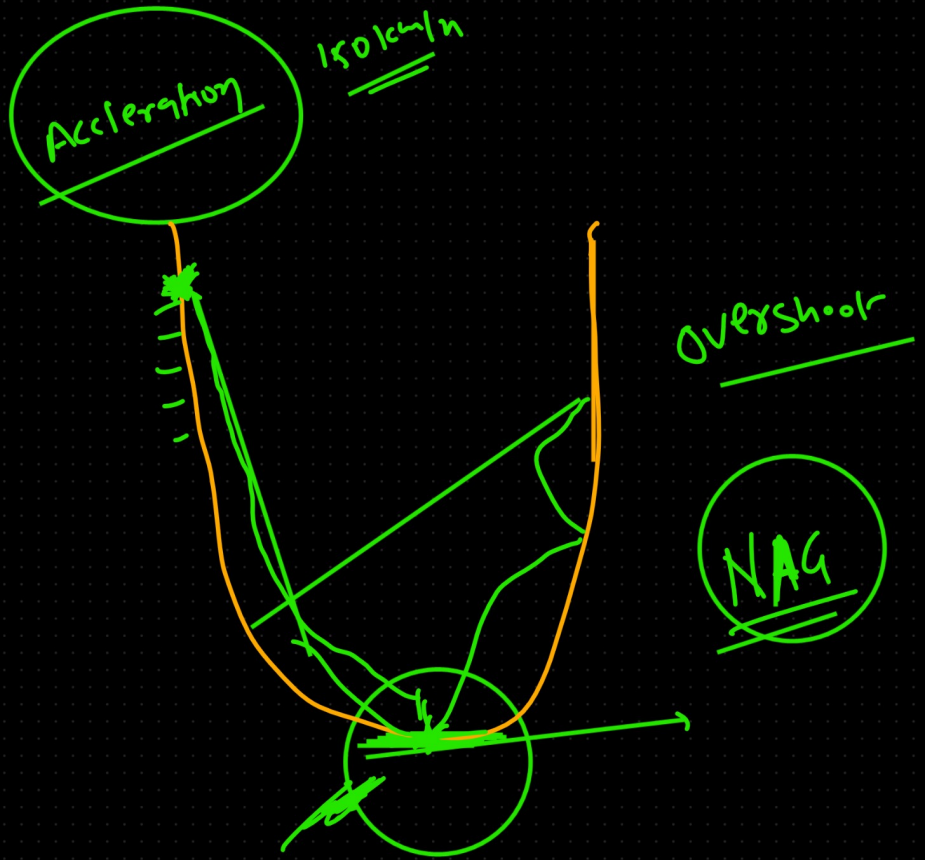
- hyperparameter tuning in ANN

Sat/Sun →

- vanishing / exploding grad -
- BN / Reg / Dropout -

overfitting

- weight initialization -


Acceleration
Isolation
Overshoot
NAG