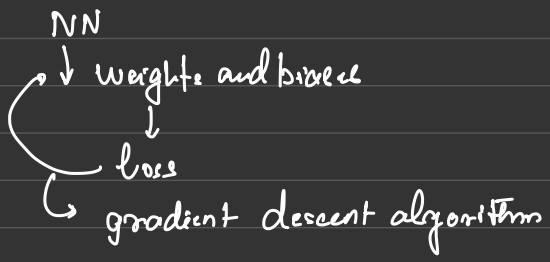
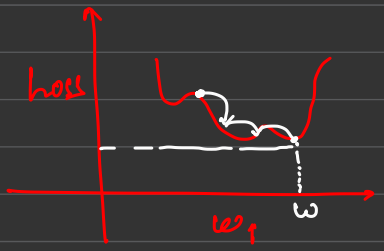
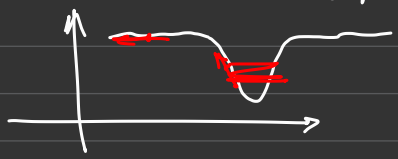
Optimizers





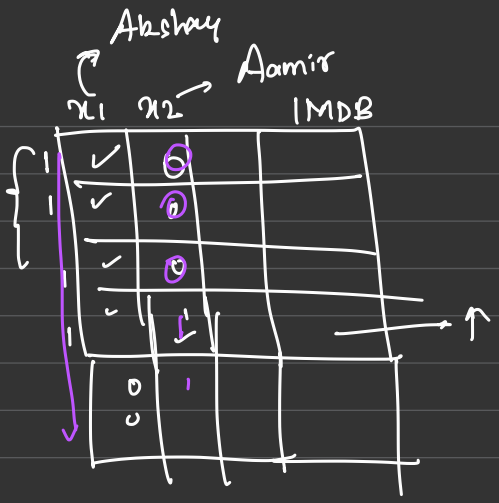
loss curves are extremely complicated functions → multiple minimas

:) be flat at certain regions
be very sharp at certain regions

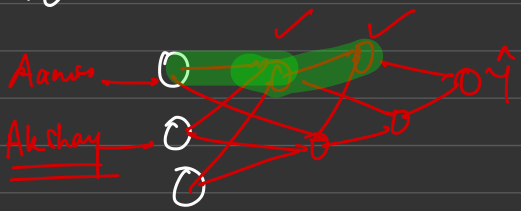


Gradient descent by design has various issues.

:) Multiple local minimas.
:) The gradients are not consistent
:) Sparse features create a problem in GD



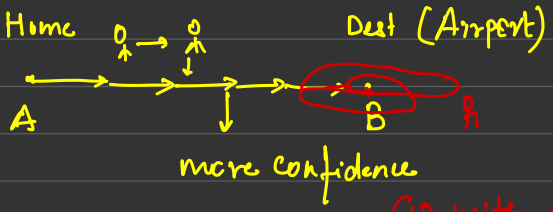
most of the rows of aamir khan's feature



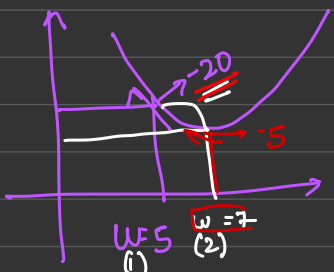
weights related to Aamir Khan
weights related to Akshay Kumar

$$w_{i+1} = w_i - \alpha \, dJ/dw_i$$

Momentum



GD



$w(2) = w(1) - \alpha \left(\frac{dh}{dw}\right)$, 0.1

$w(2) = w(1) - \alpha \times -20$

$w(2) = 5 + 2$

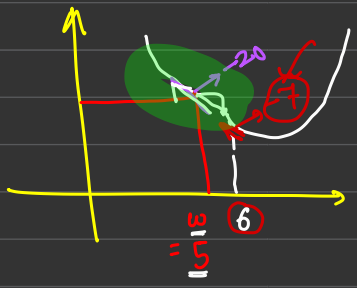
$w(2) = 7$

$w(3) = w(2) - 0.1 \times -5$

$w(3) = 7 + 0.5$

$= 7.5$

GD with momentum



$m(0) = 0$

$m(t) = \beta\, m(t-1) + (1-\beta) \times \frac{dL}{dwt}$

update → $w(2) = w(1) - \alpha(mt)$

$m(1) = 0.5 \times 0 + 0.5 \times -20$

$m(1) = -10$

$w(2) = 5 - 0.1 \times -10$

$w(2) = 6.$

$m(2) = 0.5 \times -10 - 0.5 \times -7$

$m(2) = -[5 + 3.5] = -8.5$

$w(3) = w(2) - 0.1 \times -8.5$

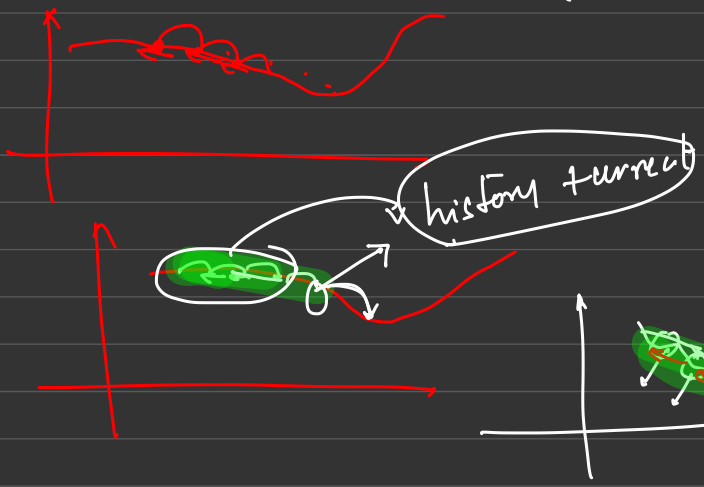
$w(3) = 6 + 0.85$

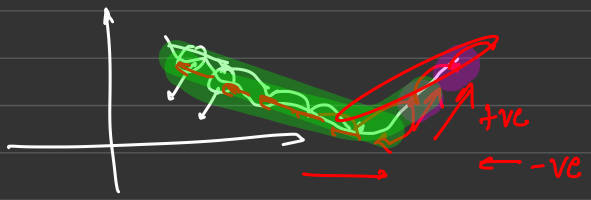
$= 6.85$

Momentum based gradient update actually takes into account the historical gradient × weightage + the current gradient × weightage.

↓

accelerates the process of the update





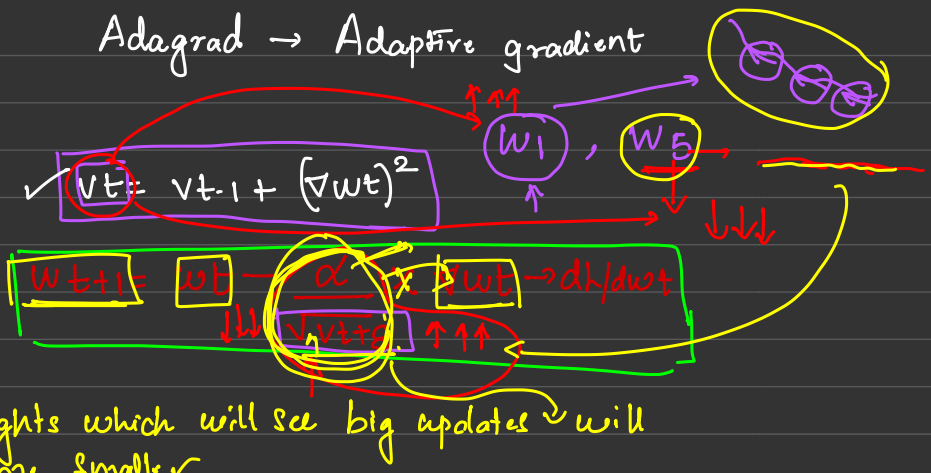
$$w_{t+1} = w_t - \alpha\, m_t$$

$$m_t = \beta \times m_{(t-1)} + (1-\beta) \times \frac{dL}{dw_t}$$

history — point / current

Even the if what min it takes longer U turns it is so fast it crosses / ignores local min

Adagrad → Adaptive gradient



Weights which will see big updates will become smaller

Adaptive gradients start punishing weight updates for weights which have already seen big gradient updates

→ dont punish/ reward weight updates for weights which have not seen big gradient updates

Rewarding (Aamir khan) ↑↑
Punish (Akshay kumar) ↓↓ → balanced learning

$$v_t = v_{t-1} + (\nabla w_t)^2$$ → too high too fast

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \times dh/dw_t$$

Rms Prop.

$$V_t = \beta \times v_{t-1} + (1-\beta) \times (\nabla w_t)^2$$

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \varepsilon}} \times dh/dw_b$$

Epsilon

Adaptive moments

Adam optimizer.

momentum

$$m_t = \beta_1 \times m_{t-1} + (1-\beta_1) \times \nabla w_t$$

$$v_t = \beta_2 \times v_{t-1} + (1-\beta_2) \times (\nabla w_t)^2$$ → RMS

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t + \varepsilon}} \times \hat{m}_t$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1} \qquad \hat{v}_t = \frac{v_t}{1-\beta_2}$$