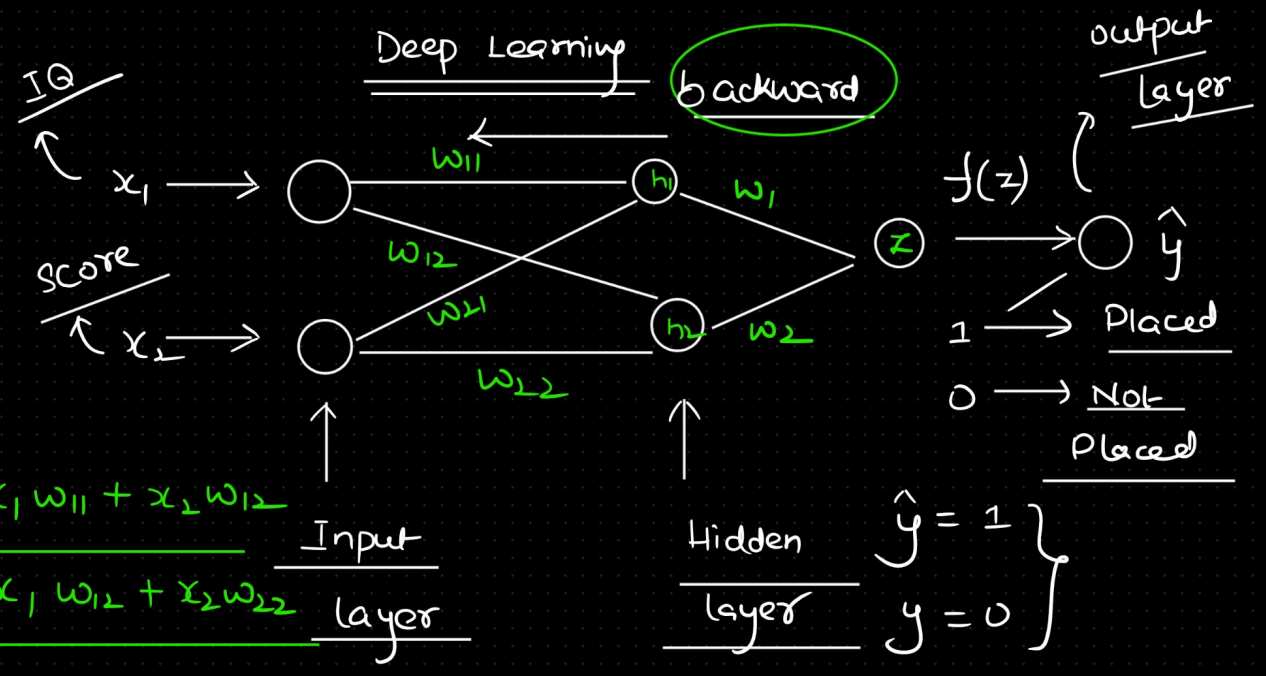
Optimizers

Optimizers are algorithms that adjust a model's parameters during training to minimize a loss function and improve performance.



$h_1 = x_1 w_{11} + x_2 w_{12}$

$h_2 = x_1 w_{12} + x_2 w_{22}$

$z = w_1 h_1 + w_2 h_2$ → feedforward

↓

Activation function

Gradient Descent Algorithm

how fast the update will happen

Learning Rate

Derivative

optimizer

Minimize loss function & improve the model performance

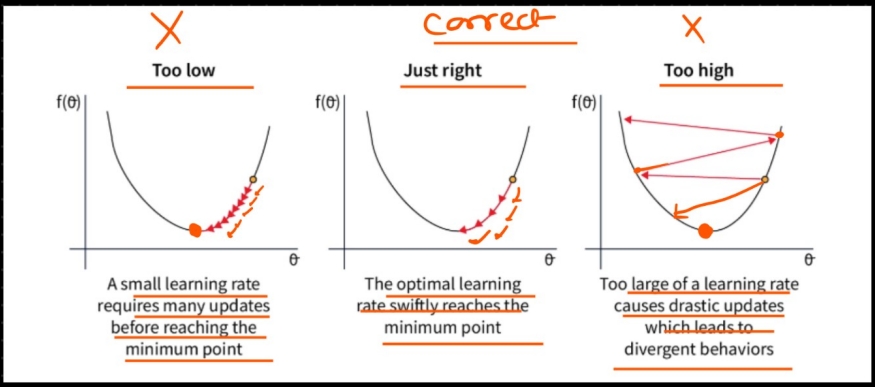
$$w_{new} = w_{old} - \eta \frac{\delta L}{\delta w}$$

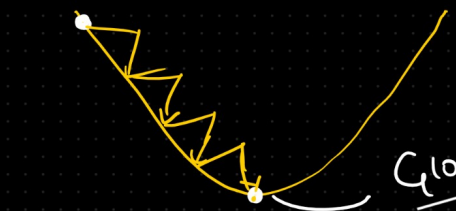
$$b_{new} = b_{old} - \eta \frac{\delta L}{\delta b}$$

→ most optimal weights

Learning Rate → hyperparameter

(hyperparameter tuning ≈ 0.001 (before training)





Global minima

Error ≈ 0

Gradient Descent

entire data

very huge

Epochs

Training

Vs

Iterations

① Computationally very expensive

1 Epochs → 10 iterations

* Batch size ⇒ 100

10000 data records

100

10 iterations

② Training time is also huge

SGD (Stochastic Gradient Descent)

↓

1 training sample at a time

1000 Records

1000 iterations → 1 epoch

Reduce ← the Computational task

Mini batch Gradient Descent

batch ⇒ 100

Num of records ⇒ 1000

iterations ⇒ $\frac{1000}{100} = 10$

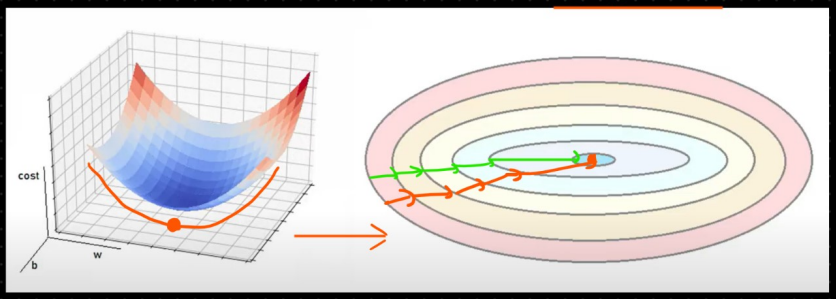
20 epochs

↳ for every single epoch, 10 iterations

Version of Gradient Descent

- Gradient Descent → entire data
- Stochastic Gradient Descent → single training sample at every iteration
- ** Mini batch Gradient Descent

Contours



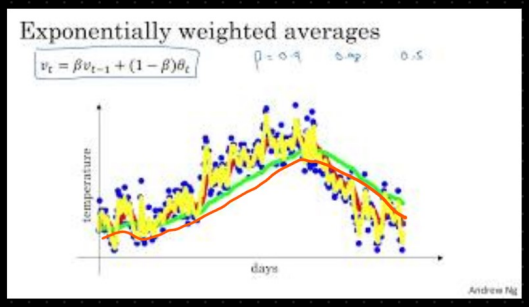
Exponential average

↓

Momentum optimizer

↓

Smoothing



$0 < \beta < 1$ $\quad \beta = 0.9$

$$v_t = \beta v_{t-1} + (1-\beta)\theta_t$$

$v_0 = 0$

$$v_1 = \beta v_0 + (1-\beta)\theta_1$$

$$v_1 = (1-\beta)\theta_1$$

$$v_2 = \beta[(1-\beta_1)\theta_1] + (1-\beta)\theta_2$$

$$v_4 = (1-\beta)(\theta_4 + \beta\theta_3 + \beta^2\theta_2 + \beta^3\theta_1)$$

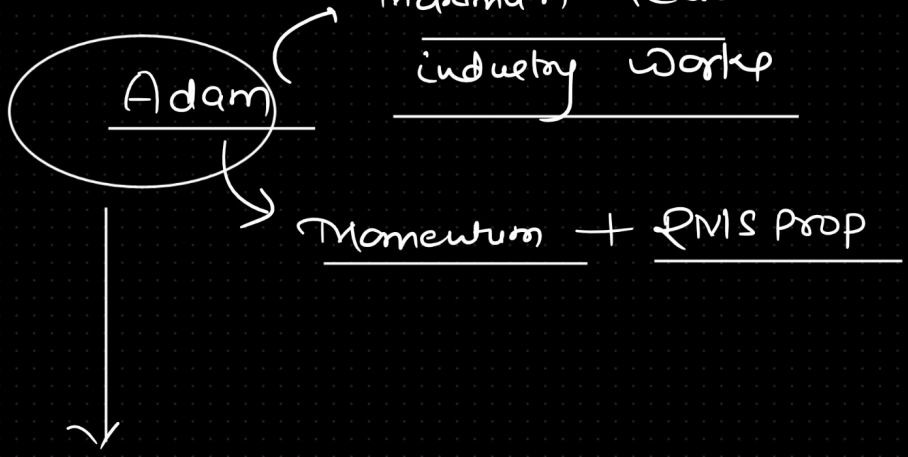
$$W_{new} = W_{old} - \eta V_{dW}$$

What is the impact of β??

Adagrad (Adaptive Gradient Boosting)

$\eta = 0.001$ (fixed before training)

dynamic $\alpha \leftarrow \eta$ (Adaptive)



better generalizability

Note:- In every optimizer, the only difference is in the updated of the parameter.