Data Science and Artificial Intelligence

Machine Learning

Linear Regression

Lecture No. 05

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic

$\boxed{\hat{Y} \text{ is } \perp \text{ projection of } Y \text{ on } C(X)}$

Topic

$C(X)$ is Column space of X.

Topic

$\boxed{(Y-\hat{Y}) \text{ is } \perp \text{ to } C(X)}$

Topic

Topic

Topics to be Covered

- Topic — Gradient descent
- Topic — Question H.W
- Topic — R^2
- Topic
- Topic

By- SIDDHARTH SABHARWAL SIR


THE BEST
VIEW COMES
AFTER THE
HARDEST CLIMB
shutterstock

Gradient descent $\Rightarrow$ (GD)

- To find location of minima

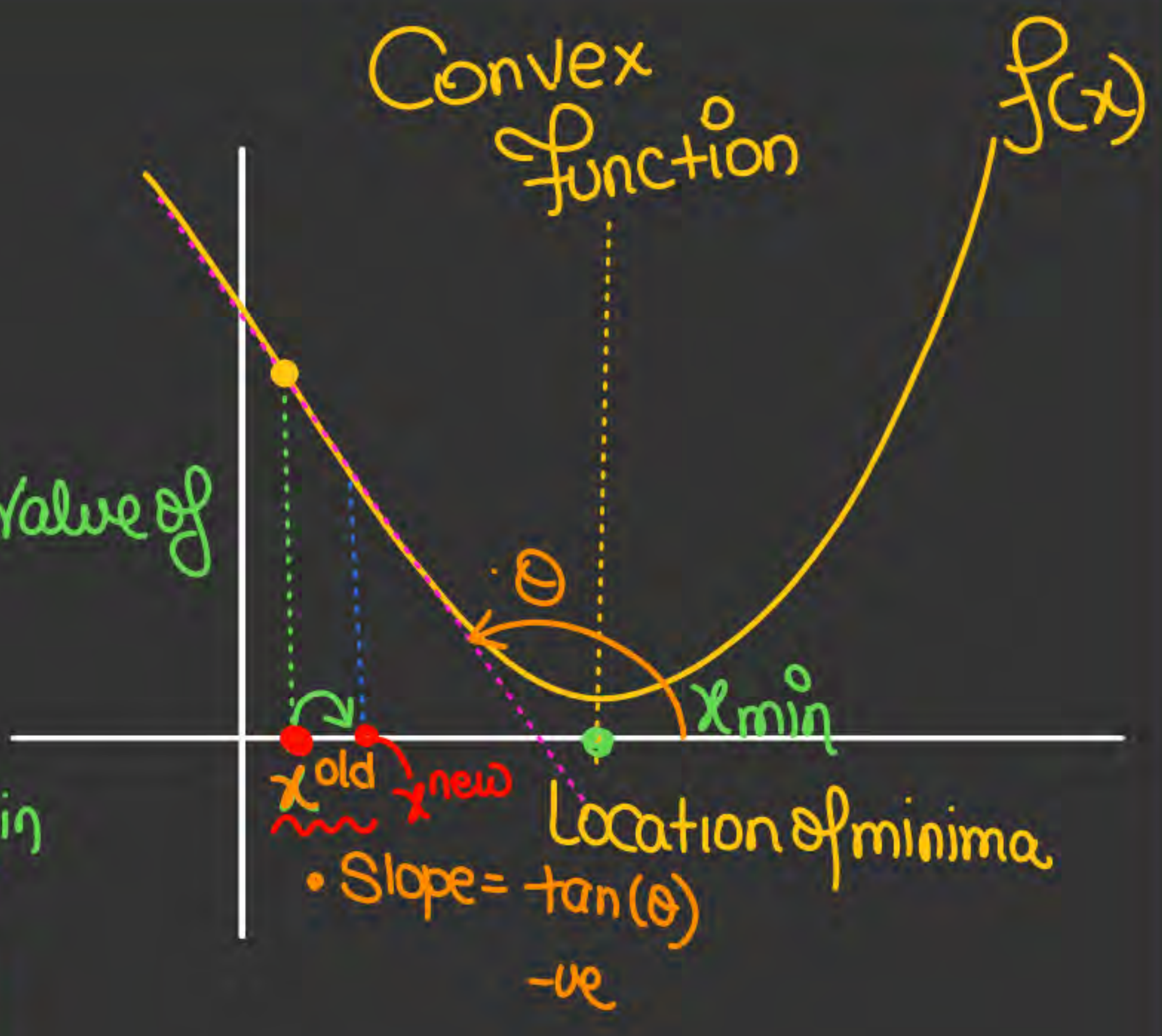
$\Rightarrow$ in GD we start with any random value of x

$\Rightarrow$ Now we have to move towards x_{min}

$$x^{new} = \left(x^{old} - \eta \underbrace{\left.\frac{\partial f(x)}{\partial x}\right|_{x^{old}}}_{\text{Slope of the } f(x) @ x^{old}} \right)$$

η is step size

+ve const value



Gradient descent $\Rightarrow$ (GD)

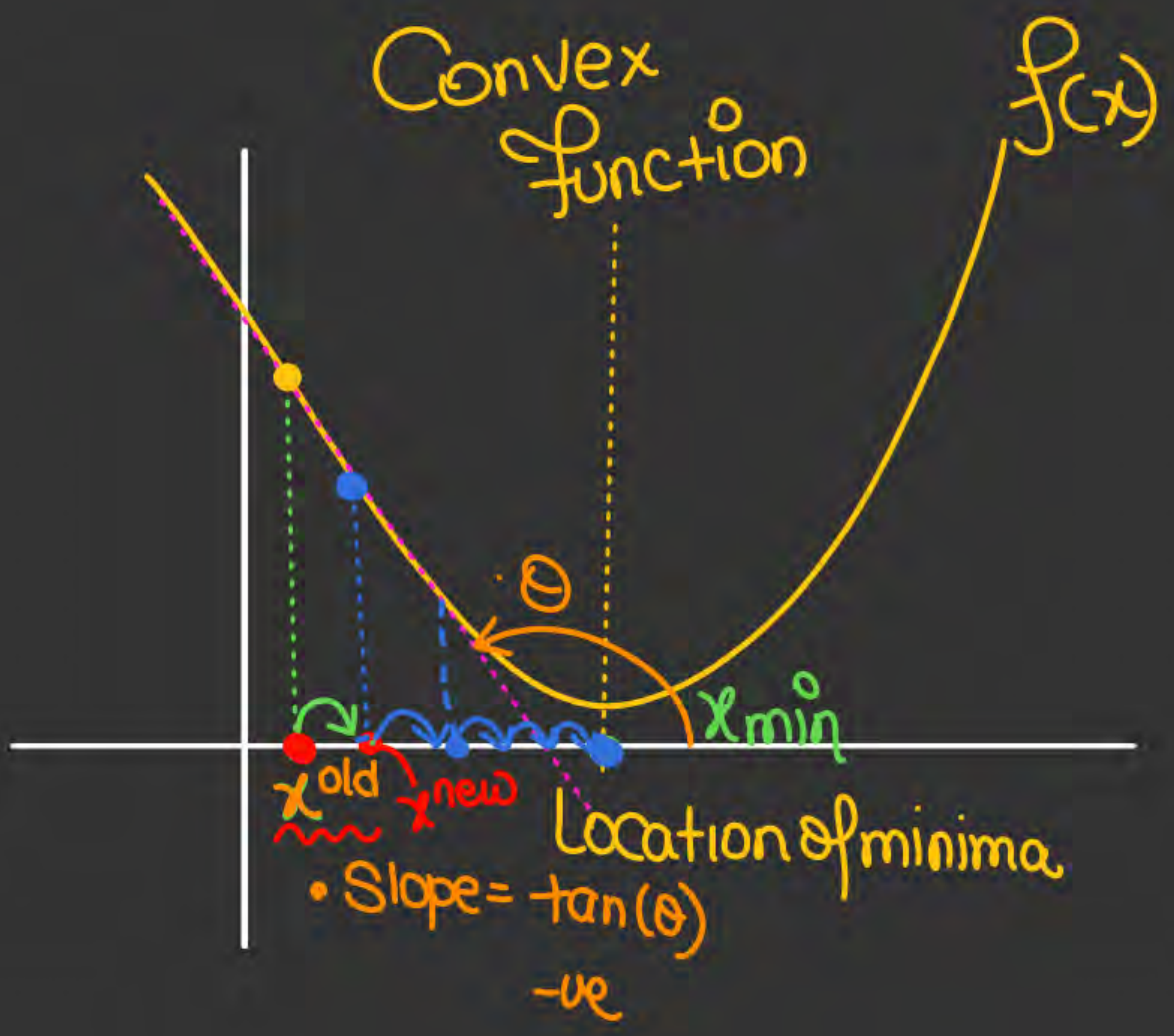
In next iteration

$$x^{new} = x^{old}$$

naya

$$x^{new} = \left(x^{old} - \eta \left. \frac{\partial f(x)}{\partial x} \right|_{x^{old}} \right)$$

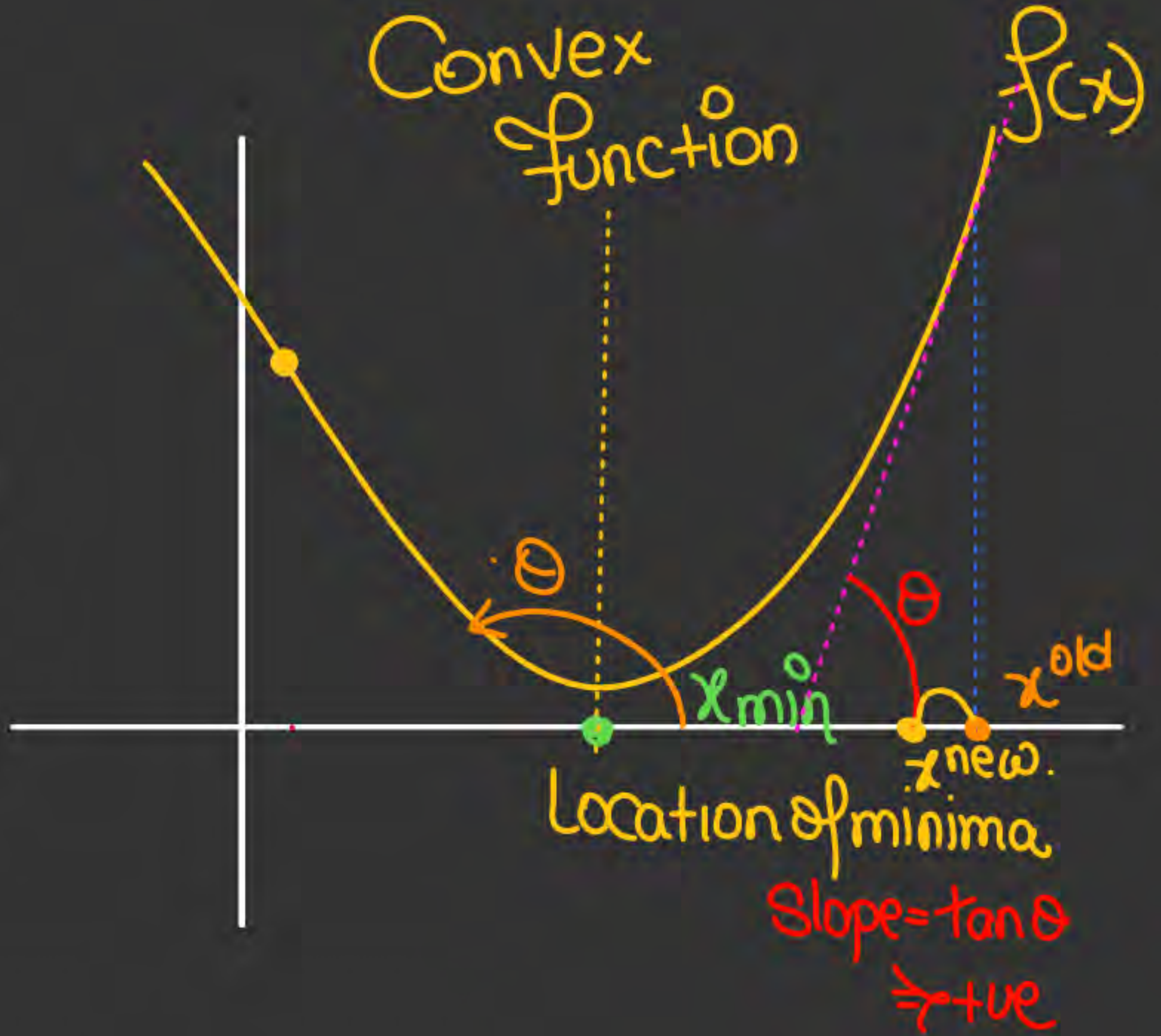
Like this we move forward we reach x_{min} where $\frac{\partial f(x)}{\partial x} = 0$ movement stops.



Gradient descent ⇒ (GD)

$$x^{new} = x^{old} - \eta \underbrace{\left.\frac{\partial f(x)}{\partial x}\right|_{x^{old}}}_{+ve}$$

$$x^{new} = x^{old} - (\quad)$$

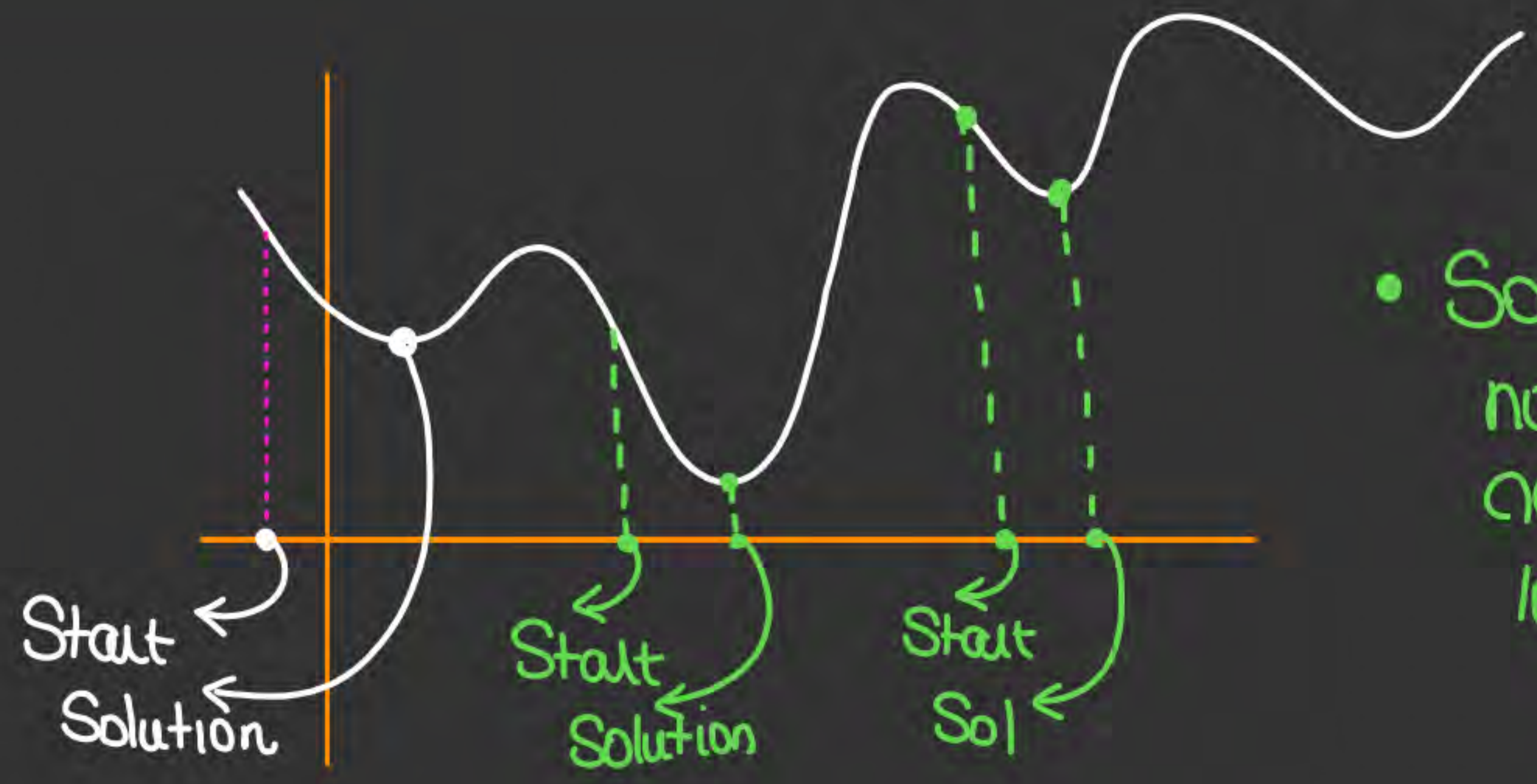


In GD, to find min of $f(x)$
we start with any random x, x^{old}

It 1: $x^{new} = x^{old} - \eta \left.\frac{\partial f(x)}{\partial x}\right|_{x^{old}}$

It 2: In It 2 the x^{new} of It 1 become x^{old} for this Iteration

Repeat same process



- So in a function which is non convex GD will give result depending on start location

Step Size Effect

"η" ⇒ step size play an important role in process.

- If η is v.v. small ⇒

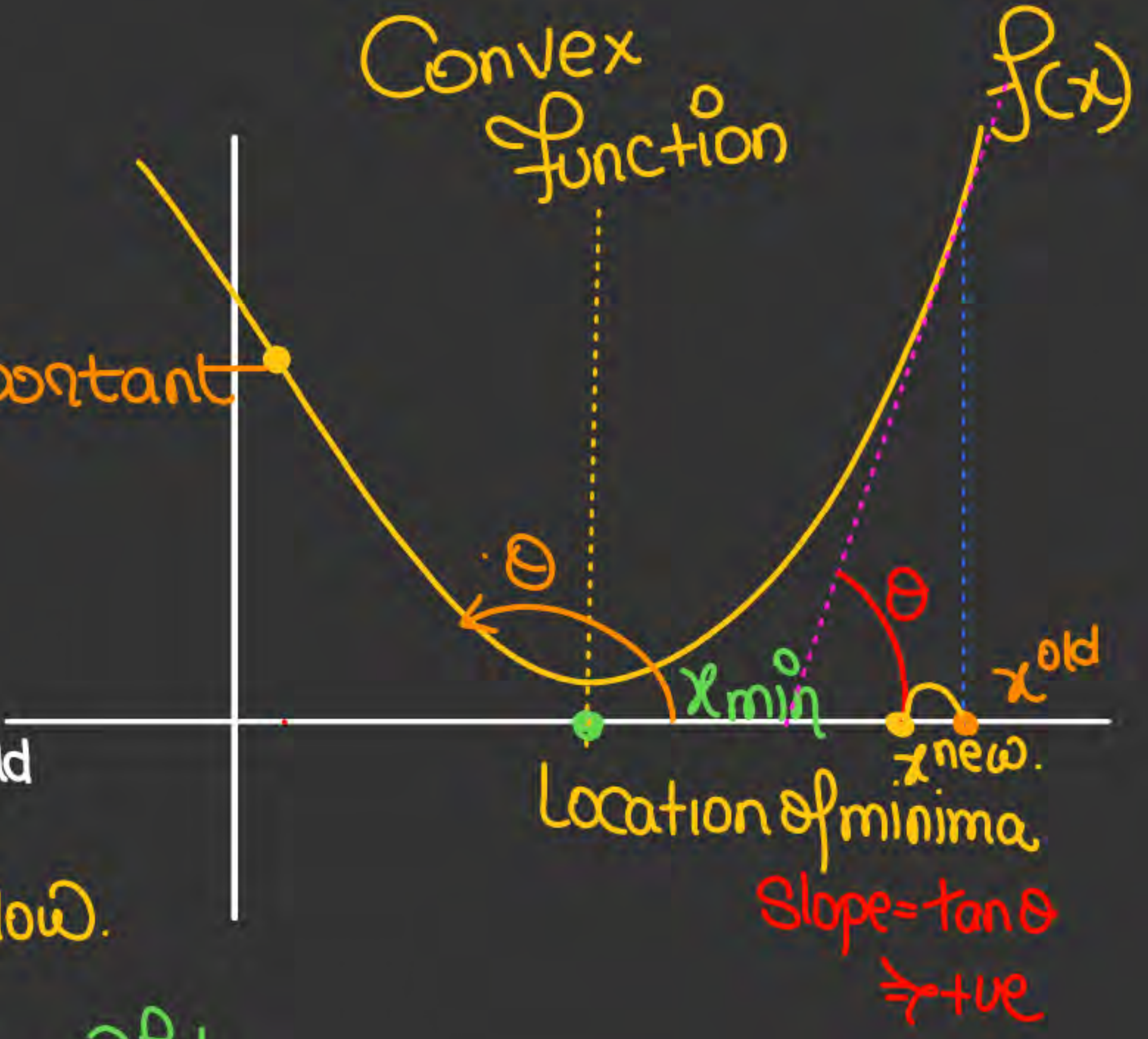
$$x^{new} = x^{old} - \eta \frac{\partial f(x)}{\partial x}\bigg|_{x^{old}}$$

Process will become very slow.

- If η is v. large.

$$x^{new} = x^{old} - \eta \frac{\partial f}{\partial x}\bigg|_{x^{old}}$$

↳ we may not get Solution.



How to represent the Loss function in the matrix format

In linear regression

1D Case $\hat{y} = \beta_1 x + \beta_0$

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4

Loss function $\Rightarrow \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

RSS

L $\Rightarrow \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2$

$$\left(\frac{\partial L}{\partial \beta}\right) = \begin{bmatrix} \partial L / \partial \beta_0 \\ \partial L / \partial \beta_1 \end{bmatrix}$$

matrix

Loss function $\Rightarrow \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

(RSS

$L \Rightarrow \sum_{i=1}^{N}(y_i - \beta_1 x_i - \beta_0)^2$

$$\left(\frac{\partial L}{\partial \beta}\right) = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \end{bmatrix}$$

matrix

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{N}(y_i - \beta_1 x_i - \beta_0)$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{N} x_i (y_i - \beta_1 x_i - \beta_0)$$

$$\left(\frac{\partial L}{\partial \beta}\right) = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \end{bmatrix} = \begin{bmatrix} -2 \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0) \\ -2 \sum_{i=1}^{N} x_i (y_i - \beta_1 x_i - \beta_0) \end{bmatrix}$$

matrix

$$= -2 \begin{bmatrix} \sum_{i=1}^{N} y_i - \left(\beta_0 \sum_{i=1}^{N} 1 + \beta_1 \sum_{i=1}^{N} x_i\right) \\ \sum_{i=1}^{N} x_i y_i - \left(\beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2\right) \end{bmatrix}$$

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \end{bmatrix} = -2 \left\{ \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} - \begin{bmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \right\}$$

So $\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \\ \vdots \\ \partial L/\partial \beta_D \end{bmatrix} = -2\left[X^T Y - (X^T X)\beta \right]$

D dimension data

Gradient descent in LR

- In LR we have to minimize Loss fxn. (L) $\to$ RSS.
- we start with any random values of $\beta = \beta^{old}$

It1 $\beta^{new} = \beta^{old} - \eta \frac{\partial L}{\partial \beta}\Big|_{\beta^{old}}$

$$\frac{\partial L}{\partial \beta}\Big|_{\beta^{old}} = -2\left[X^T Y - (X^T X)\beta^{old}\right]$$

It2 : the β^{new} of It 1 act as β^{old} of It2

$\beta^{new} = \beta^{old} - \eta \frac{\partial L}{\partial \beta}\Big|_{\beta^{old}}$

$\boxed{\text{Why Gradient descent ??}}$ → $\boxed{\text{bcoz when data is v. large then GD is used.}}$

we have Solution of $\beta = (X^T X)^{-1} X^T Y$.

#Q. If $g(x, y) = x^2 + y^2 - 4x$, find the gradient vector $\nabla g(1, 2)$

$$g(x, y) = x^2 + y^2 - 4x$$

2 variables

gradient of $g = \begin{bmatrix} \partial g/\partial x \\ \partial g/\partial y \end{bmatrix} = \begin{bmatrix} 2x - 4 \\ 2y \end{bmatrix}$

$$= \begin{bmatrix} 2(1) - 4 \\ 2(2) \end{bmatrix} = \begin{bmatrix} 2 - 4 \\ 4 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

Q.

x	y
1	4
2	6
3	5
4	7

$$\beta_{start} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$L = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\eta = .5$$

Find β after 1st Iteration

$$\beta^{new} = \left(\beta^{old} - \eta \frac{\partial L}{\partial \beta}\Big|_{\beta^{old}}\right)$$

$$\frac{\partial L}{\partial \beta} = -\left[X^T Y - X^T X \beta^{old}\right]$$

Q.

x	y
1	4
2	6
3	5
4	7

$$\beta_{start} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$L = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$\eta = .5$

Find β after 1st Iteration

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{matrix} 2 \\ 3 \end{matrix}$$

$$X^TY = \begin{bmatrix} 22 \\ 59 \end{bmatrix}$$

$$\frac{\partial L}{\partial \beta} = -\left[X^TY - X^TX\beta^{old}\right]$$

$$= -\left[\begin{bmatrix} 22 \\ 59 \end{bmatrix} - \begin{bmatrix} 38 \\ 110 \end{bmatrix}\right] = -\begin{bmatrix} -16 \\ -51 \end{bmatrix}$$

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} 16 \\ 51 \end{bmatrix}$$

$$\beta^{new} = \beta^{old} - \eta \left.\frac{\partial L}{\partial \beta}\right|_{\beta^{old}}$$

$$= \begin{bmatrix} 2 \\ 3 \end{bmatrix} - (.5)\begin{bmatrix} 16 \\ 51 \end{bmatrix} = \begin{bmatrix} -6 \\ -22.5 \end{bmatrix}$$

#Q. Let's consider regression in one dimension, so our inputs $x^{(i)}$ and outputs $y^{(i)}$ are in $\mathbb{R}$.

(a) (4 points) Linny uses regular linear regression. Given the following dataset,

H.W.

$D = \{((1), 1), ((2), 2), ((3), 4), ((3), 2)\}$

What value of θ and θ_0 optimize the mean squared error of hypotheses of the form $h(x; \theta, \theta_0) = \theta_x + \theta_0$?

#Q. Consider a one-dimensional regression problem with training data $\{x_i, y_i\}$. We seek to fit a linear model with no bias term:

(a) Assume a squared loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and solve for the optimal value of ω^*.

H.W

#Q. Consider the following 4 training examples:

X	Y
-1	0.0319
0	0.8692
1	1.9566
2	3.0343

H.W

We want to learn a function $f(x) = ax + b$ which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

#Q. The linear regression model $y = a_0 + a_1x_1 + a_2x_2 + ... + a_px_p$ is to be fitted to a set of N training data points having p attributes each. Let X 1 point be $N \times (p + 1)$ vectors of input values (augmented by 1's), Y be $N \times 1$ vector of target values, and θ be $(p + 1) \times 1$ vector of parameter values $(a_0, a_1, a_2, \ldots, a_p)$. If the sum squared error is minimized for obtaining the optimal regression model, which of the following equation holds?

H.W

(a) $X^TX = XY$

(b) $X\theta = X^TY$

(c) $X^TX\theta = Y$

(d) $X^TX\theta = X^TY$

Consider the function $J(w) = w_1^2 + w_2^2 - 6w_1 + 8w_2 - 9$. Answer questions (1-6):

1) The theoretical value of $min(J(w))$ is __________.

H.W

4) Start with the initial guess of $[w_1, w_2] = [5, 5]$. Take the value of learning rate = 0.3. The value of w_1 after 4 iterations of gradient descent will be ________.

H.W

#Q. Let's consider regression in one dimension, so our inputs $x^{(i)}$ and outputs $y^{(i)}$ are in $\mathbb{R}$.

(a) (4 points) Linny uses regular linear regression. Given the following dataset,

D = {((1), 1), ((2), 2), ((3), 4), ((3), 2)}

What value of θ and θ_0 optimize the mean squared error of hypotheses of the form $h(x; \theta, \theta_0) = \theta_x + \theta_0$?

#Q. Consider a one-dimensional regression problem with training data $\{x_i, y_i\}$. We seek to fit a linear model with no bias term:

(a) Assume a squared loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and solve for the optimal value of ω^*.

#Q. Consider the following 4 training examples:

X	Y
−1	0.0319
0	0.8692
1	1.9566
2	3.0343

We want to learn a function $f(x) = ax + b$ which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TSS = total sum of squares

n = number of observations

y_i = value in a sample

$\bar{y}$ = mean value of a sample

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The most important thing we do after making any model is evaluating the model.
- R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The value of R-square lies between 0 to 1.
- Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.
- However, we get R-square equals 0 when the model does not predict any variability in the model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.
- The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The goodness of fit of regression models can be analyzed on the basis of the R-square method. The more the value of the r-square near 1, the better the model is.

- Note: The value of R-square can also be negative when the model fitted is worse than the average fitted model. .

Considering data of P Dimensions

Adjusted R - Squares

- Adjusted R-Squared is an updated version of R-squared which takes account of the number of independent variables while calculating R-squared.
- n is the total number of observations in the data
- k is the number of independent variables (predictors) in the regression model

$$Adjusted R^2 = 1 - \frac{(1-R^2)\cdot(n-1)}{n-k-1}$$

Considering data of P Dimensions

Lets solve a question

Question 2: Given a simple linear regression model with an R-squared value of 0.64, what percentage of the variation in the dependent variable is explained by the predictor variable?

Linear Regression

Considering data of P Dimensions

Lets solve a question

Question 6: In a simple linear regression model, if the coefficient of determination (R-squared) is 0.81 and the total sum of squares (SST) is 400, what is the sum of squared errors (SSE)?

a)76

b)77

c)54

d)33

Linear Regression

Question 5: In a simple linear regression analysis, if the mean of the dependent variable (Y) is 50, and the slope coefficient (a) is 3, what is the mean of the predictor variable (X) when X and Y are centered?

a) Cannot be determined without the value of the intercept (b).

b) Cannot be determined without the value of the intercept (a)

c)Can be determined without the value of the intercept (b)

Question 9: In a simple linear regression analysis, if the sum of squared errors (SSE) is 120 and the degrees of freedom for residuals is 15, what is the mean squared error (MSE)?

Question 15: What is the purpose of the coefficient of determination (R-squared) in simple linear regression?

A. To determine the slope of the regression line
B. To measure the strength of the linear relationship
C. To calculate the p-value of the regression
D. To identify outliers in the dataset

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate?

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting

Linear Regression

Question 2: What does the coefficient of determination (R-squared) measure in multiple linear regression?

A. The correlation between predictor variables
B. The percentage of variance in the dependent variable explained by the model
C. The significance of the intercept term
D. The number of predictor variables in the model

In multiple linear regression, what is the key difference between simple linear regression and multiple linear regression?

A) Simple linear regression has one independent variable, while multiple linear regression has two or more.

B) Simple linear regression uses categorical variables, while multiple linear regression uses continuous variables.

C) Simple linear regression is used for classification, while multiple linear regression is used for prediction.

D) There is no difference between simple and multiple linear regression.

Which statistic is used to assess the strength and direction of the relationship between the dependent variable and each independent variable in multiple linear regression?

A) Mean absolute error (MAE)
B) R-squared (R^2)
C) Standard error
D) Confidence interval

What is the purpose of the residual plot in multiple linear regression analysis?

A) To visualize the relationship between independent variables.
B) To check for homoscedasticity and the presence of outliers.
C) To calculate the correlation coefficient (r).
D) To assess multicollinearity.

Linear Regression

What is the main purpose of the intercept term in a multiple linear regression model?

A) It represents the slope of the regression line.
B) It is used to control for multicollinearity.
C) It represents the expected value of the dependent variable when all independent variables are zero.
D) It is not used in multiple linear regression.

THANK - YOU