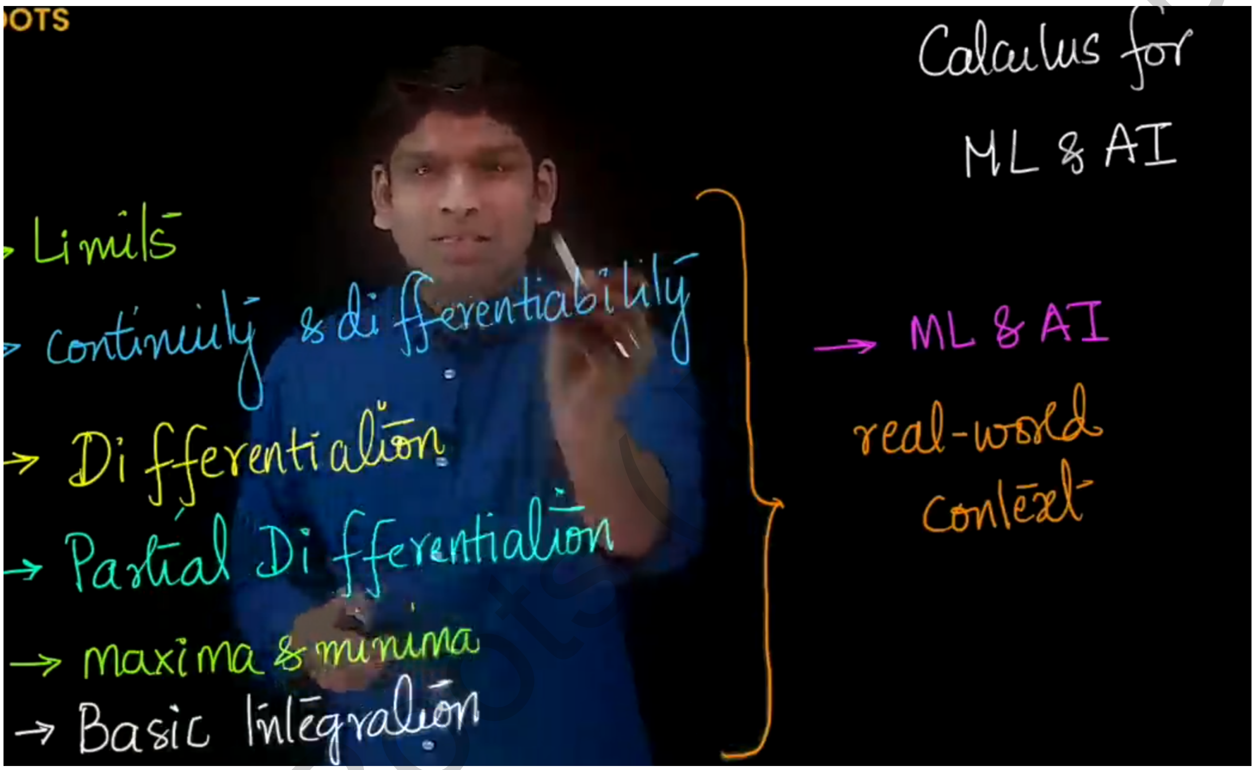
# 3.1 Introduction

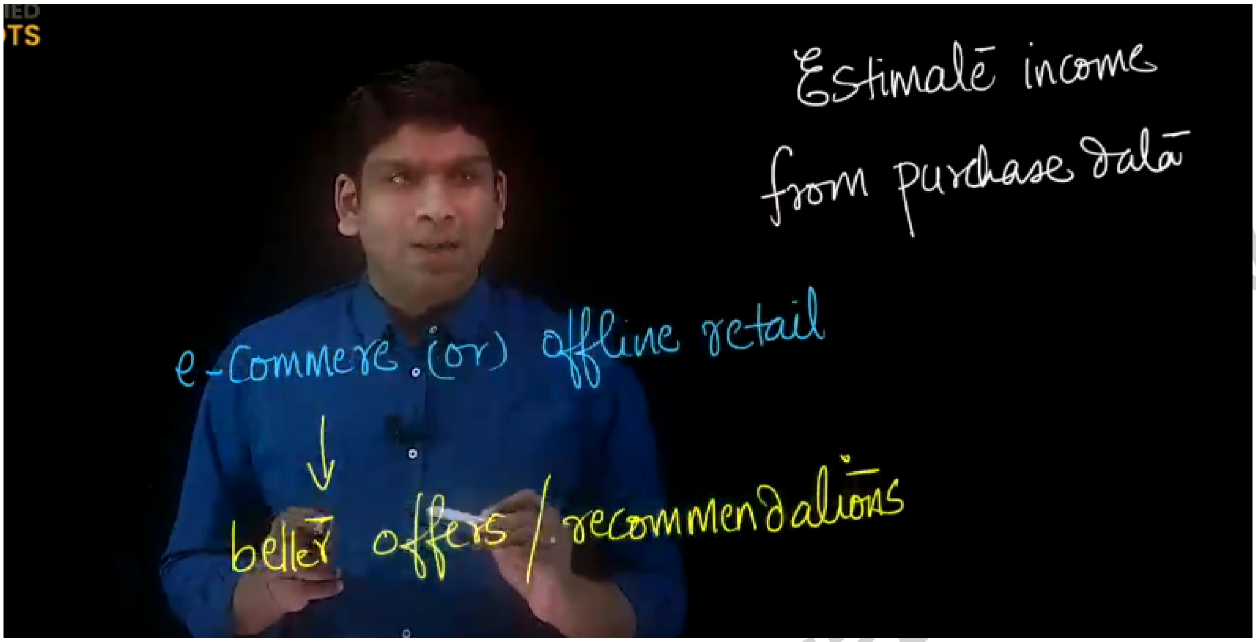
At timestamp   in video




- In this chapter we study calculus for machine learning and AI applications in real world context.
- We cover all the above concepts

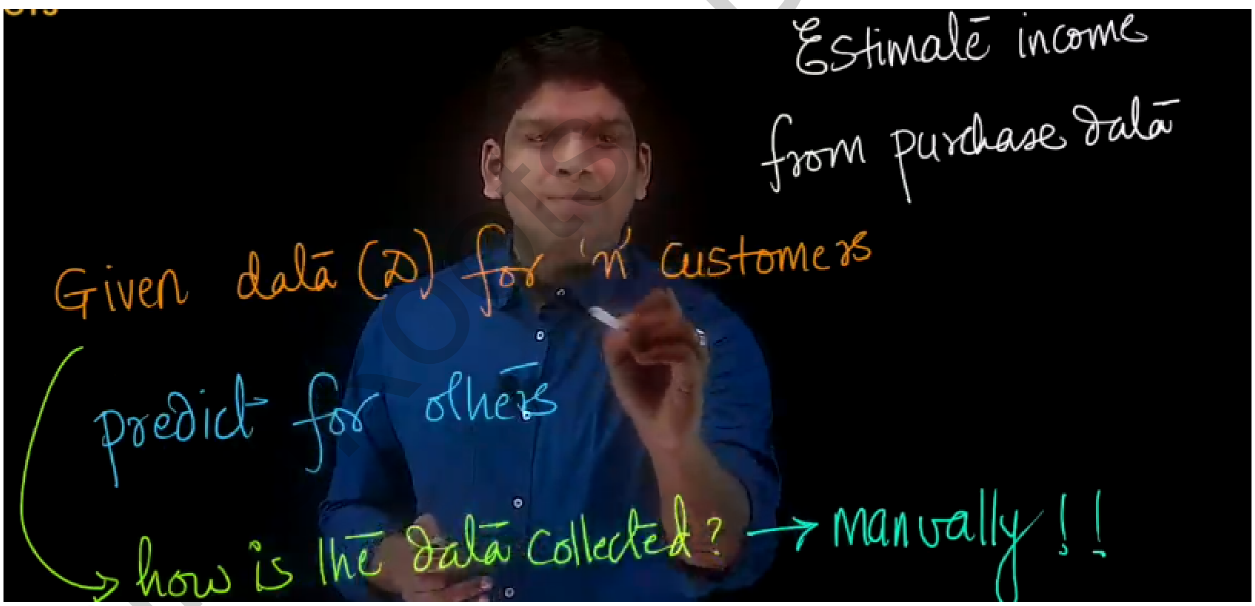
# 3.2 Real world problem: Estimate income using purchase data

At timestamp  1.24 in video



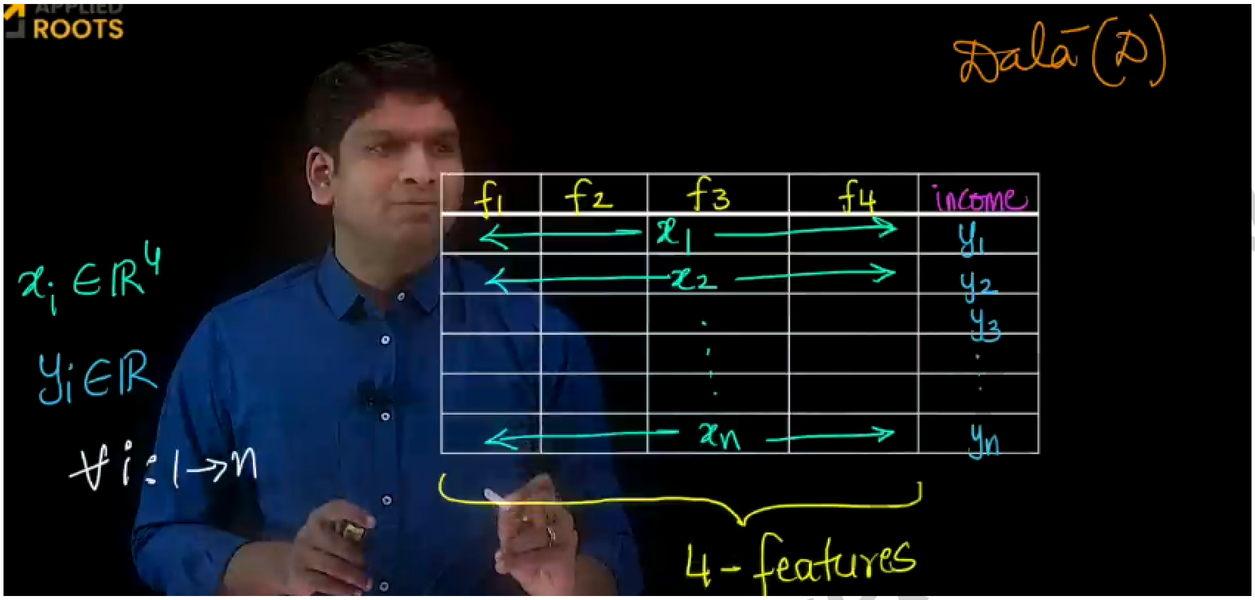

- In this chapter we estimate income from purchase data of an ecommerce company.

At timestamp 2.24 in video



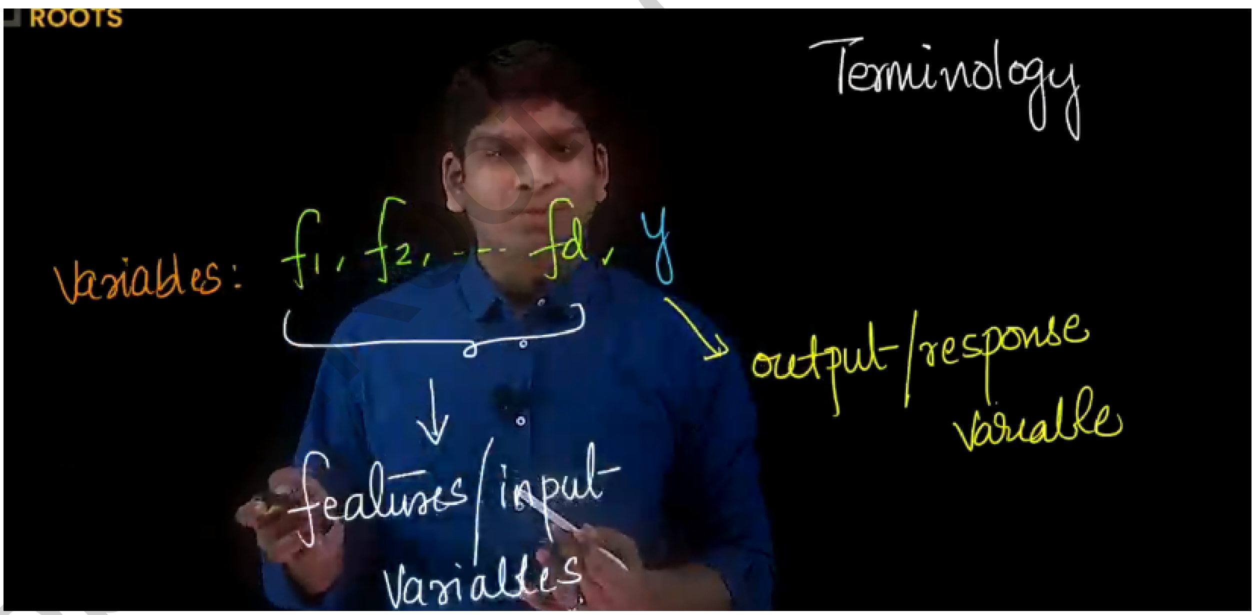

- We collect customer data manually .Given data of n customers ,we train our model and try to predict the income of others.

At timestamp 4.30



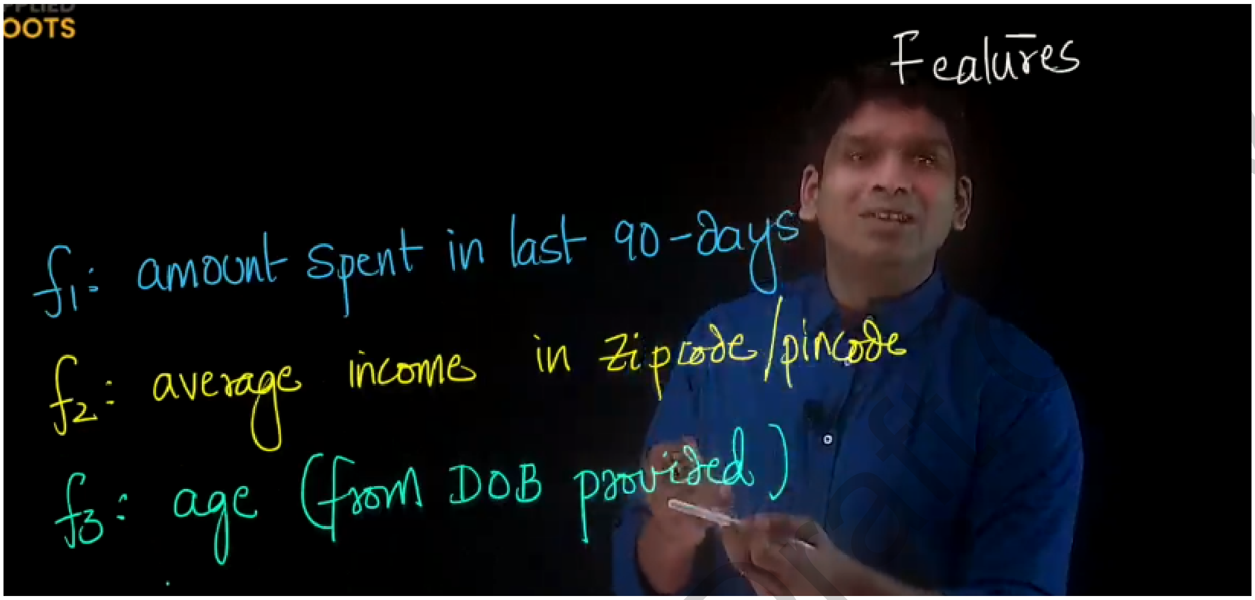

- Let's assume 4 we have four features and we have to predict income based on these features .Each row corresponds to data of one customer.

At timestamp 6.24



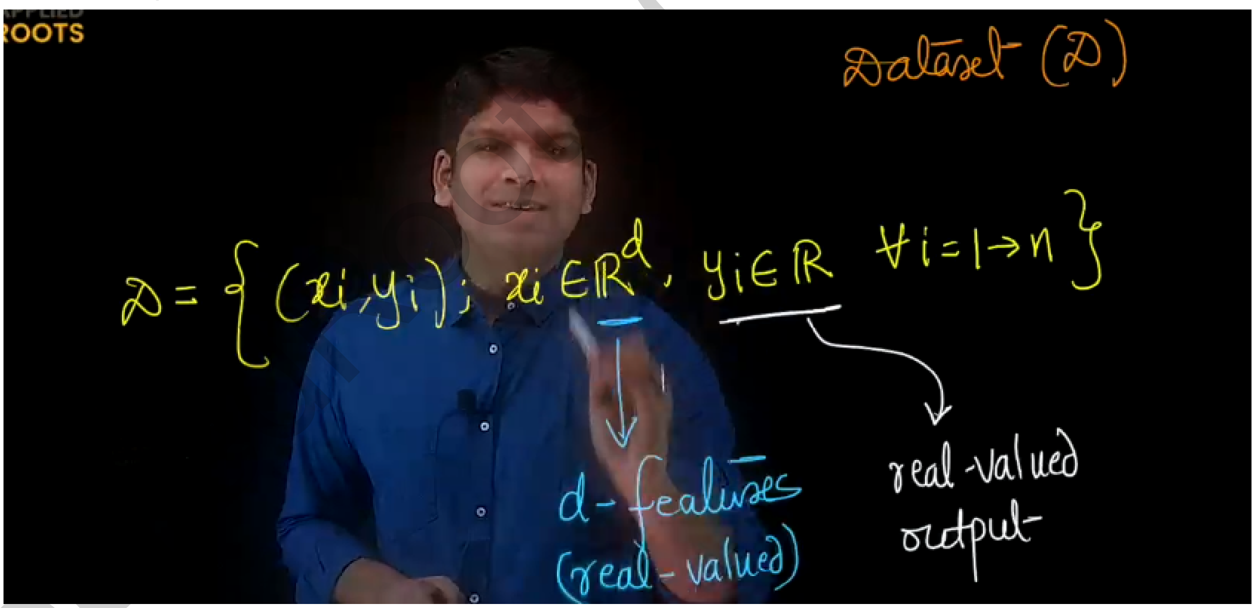

- We have four features we call them as input variables and based on these features we predict output or response variable.
- Once we build a mathematical model we give the model input features and it outputs us the response.

At timestamp 9.53




- The features can be any data that we have collected as shown above

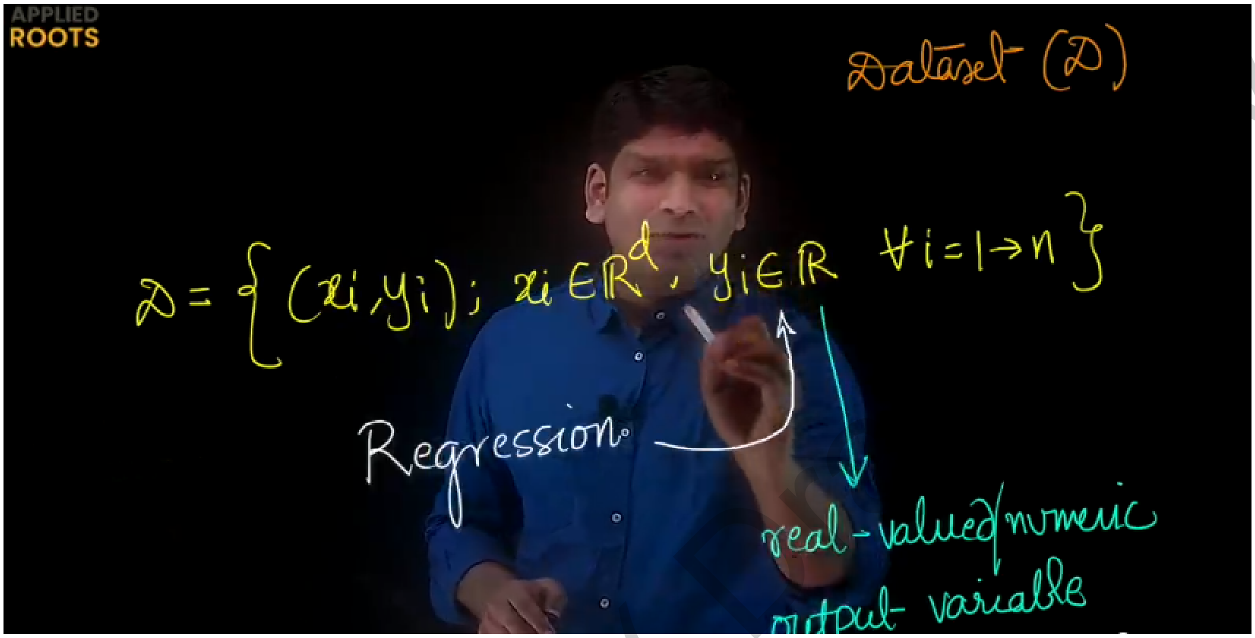
At timestamp 14.22




- In this chapter we consider all features to be numeric going further we discuss categorical features as well.

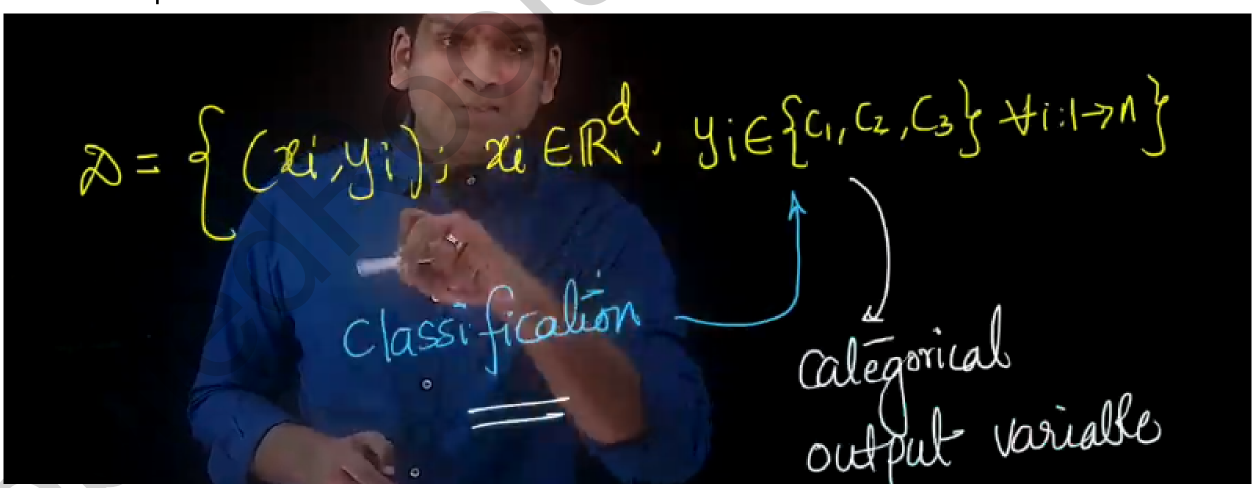
# 3.3 Visualize the Data

At timestamp 14.53



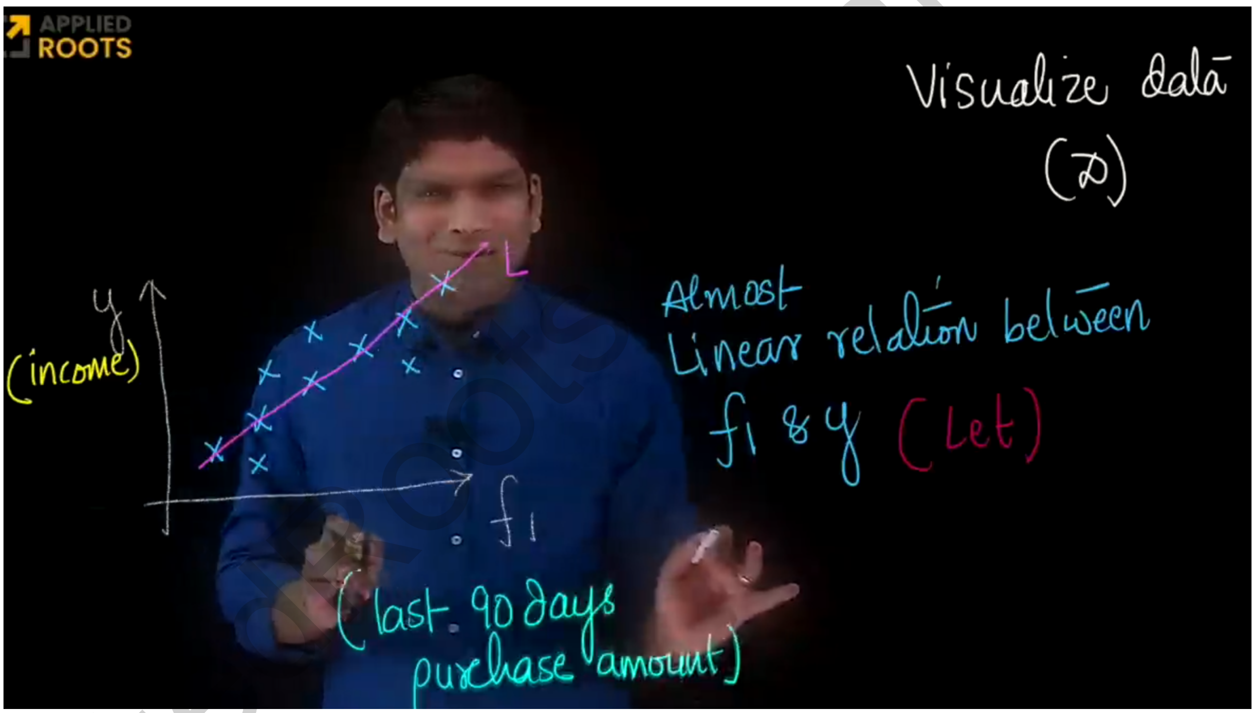

- If the output variable is real valued then the problem at hand is a regression problem .

At timestamp 15.40



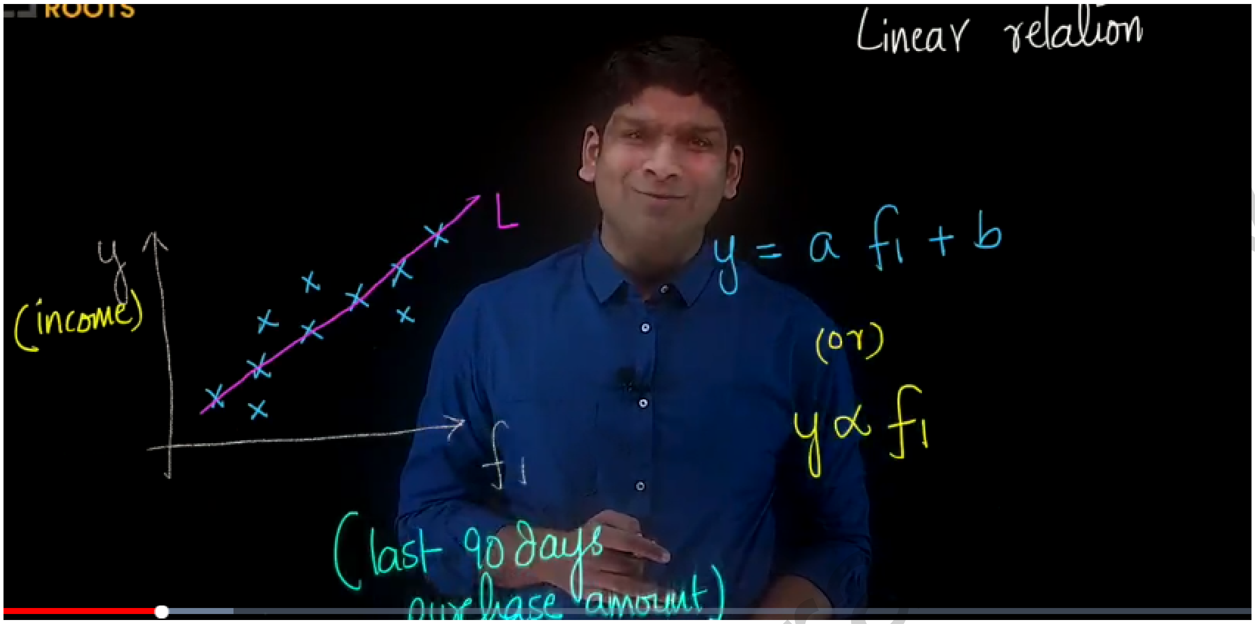

- If the response variable is categorical ,this means the problem at hand is classification problem.(like fish sorting problem)
- If y belongs to more than one class it is called a multiclass classification problem.

At timestamp 2.54  in video

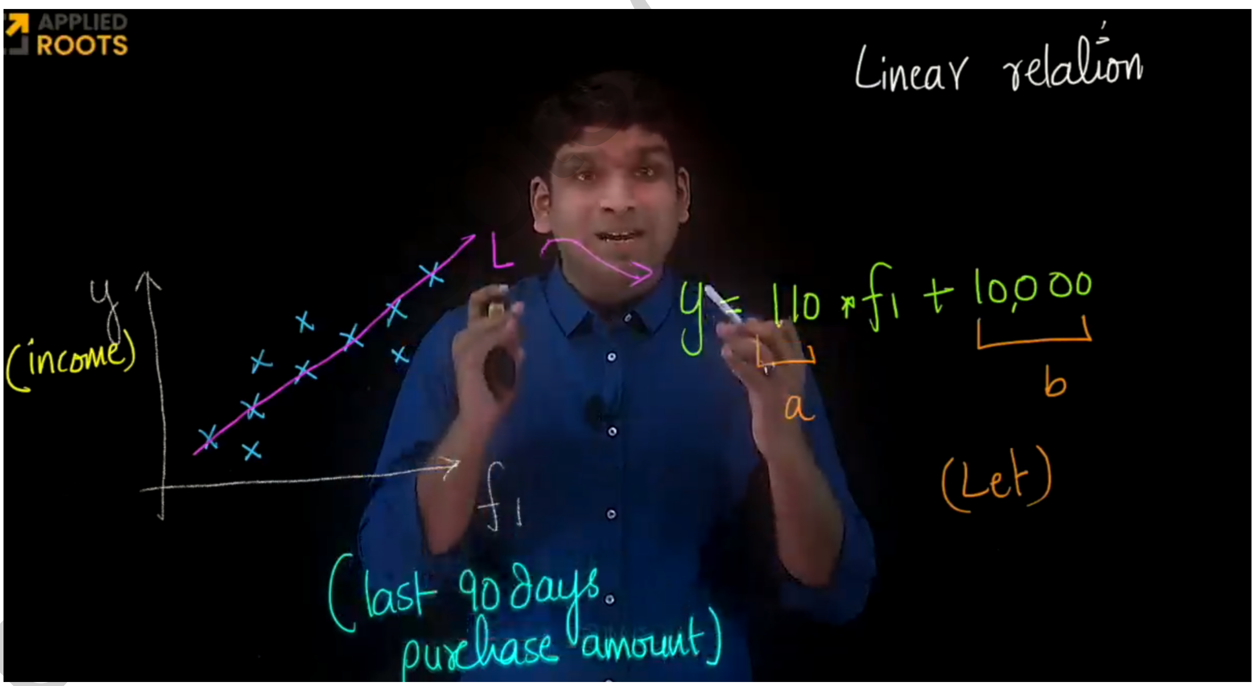



- As we are going to solve the regression problem ,let's try to visualize our data in 2 D,So we have only 1 feature as shown above along the x axis.
- We can clearly see that the feature f and response variable y are having almost a linear relationship.

At timestamp 4.02 in video

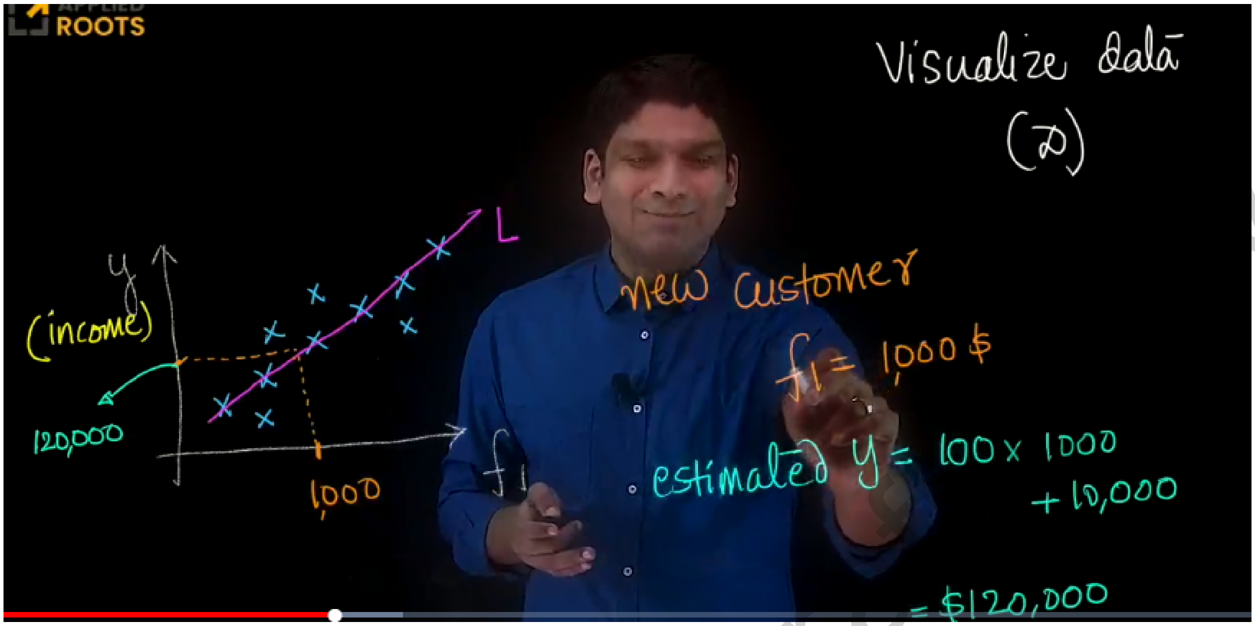



- So we can represent the relation as y=a.f+b(as a line).



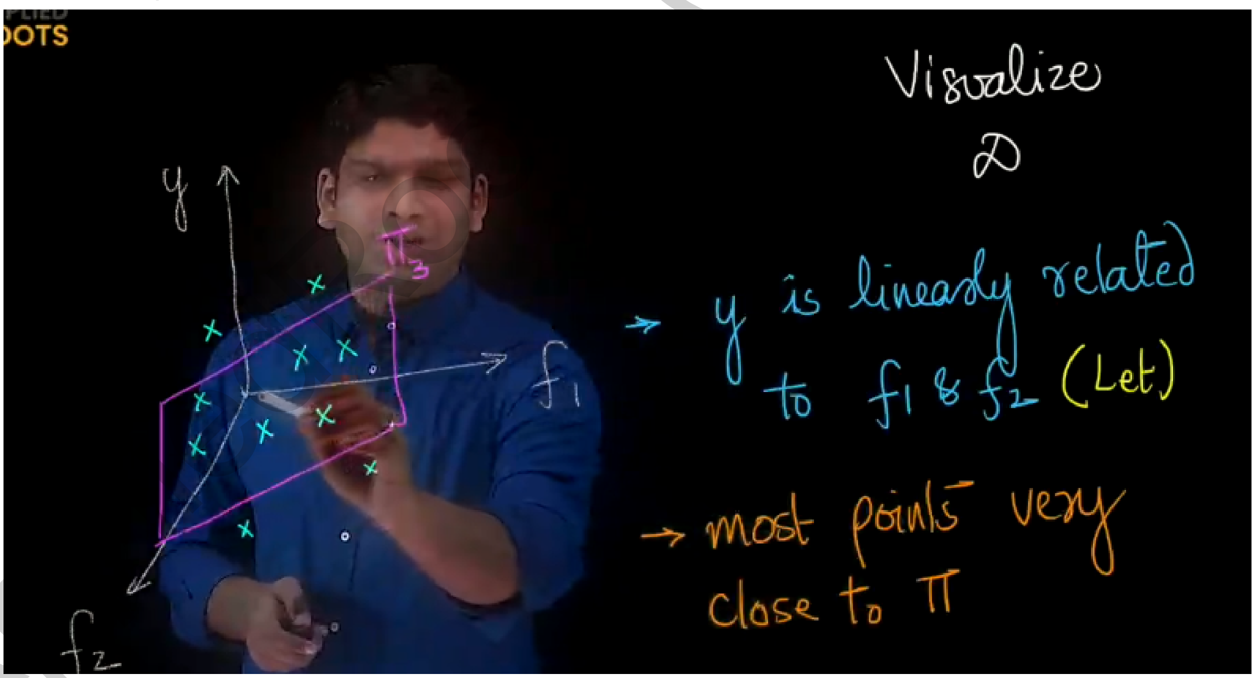

- Let’s say we trained and fit the model and we got the above line as our model.
- y=110.f1+10000 is our mathematical model

At timestamp 7.21 in video

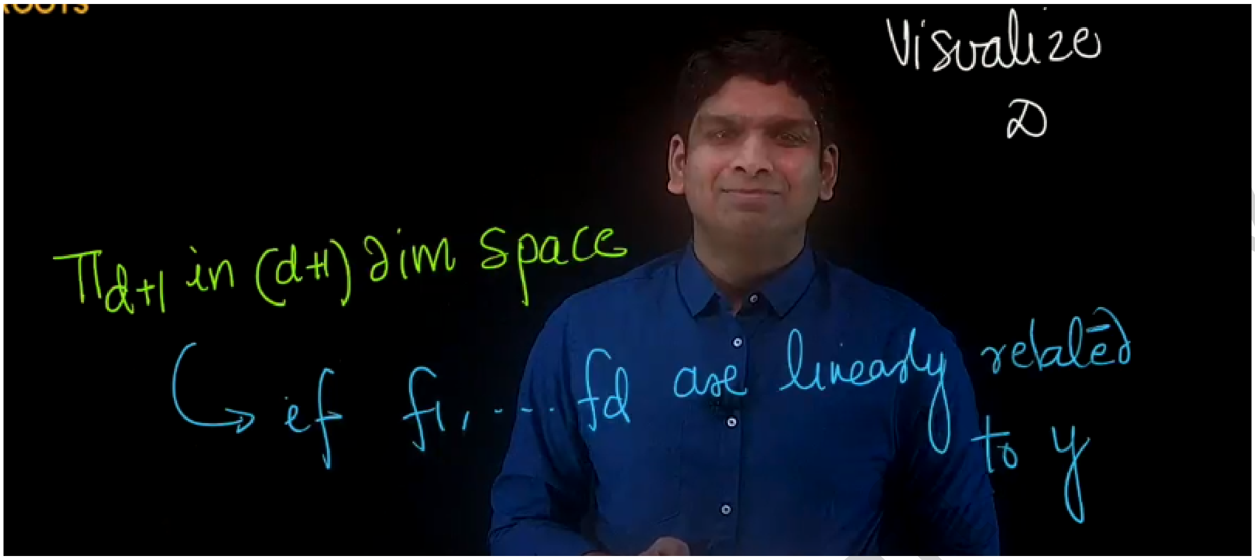


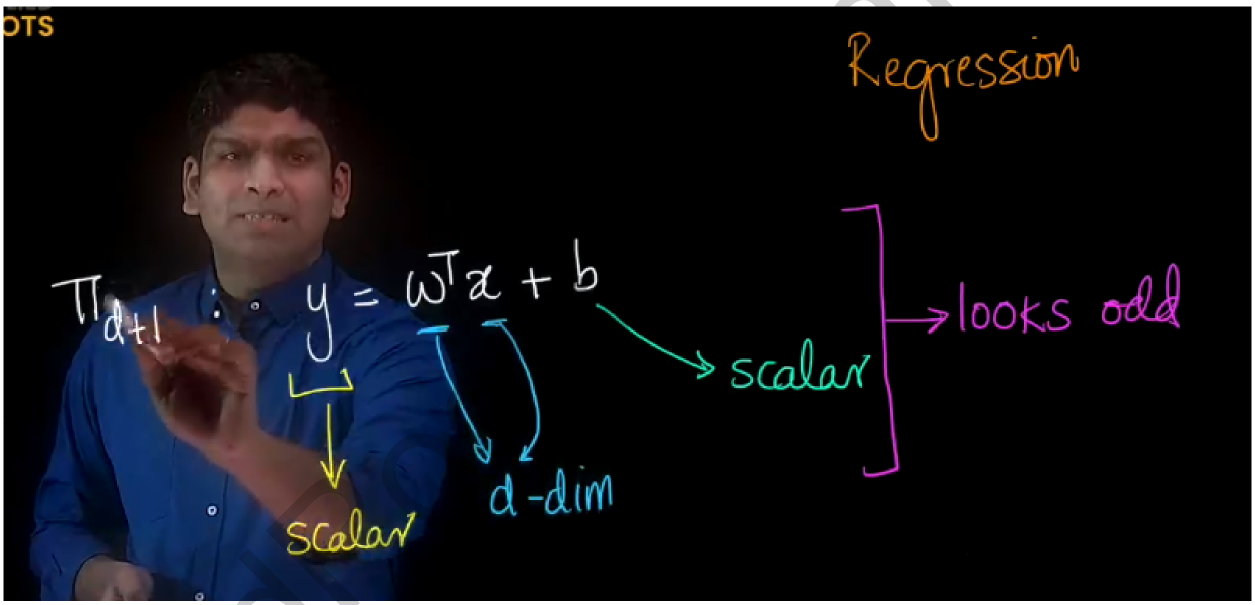

- If we consider the above line as our model then when we have data about a new customer we can just substitute it and we get the income of the new customer.

At timestamp 7.31 in video



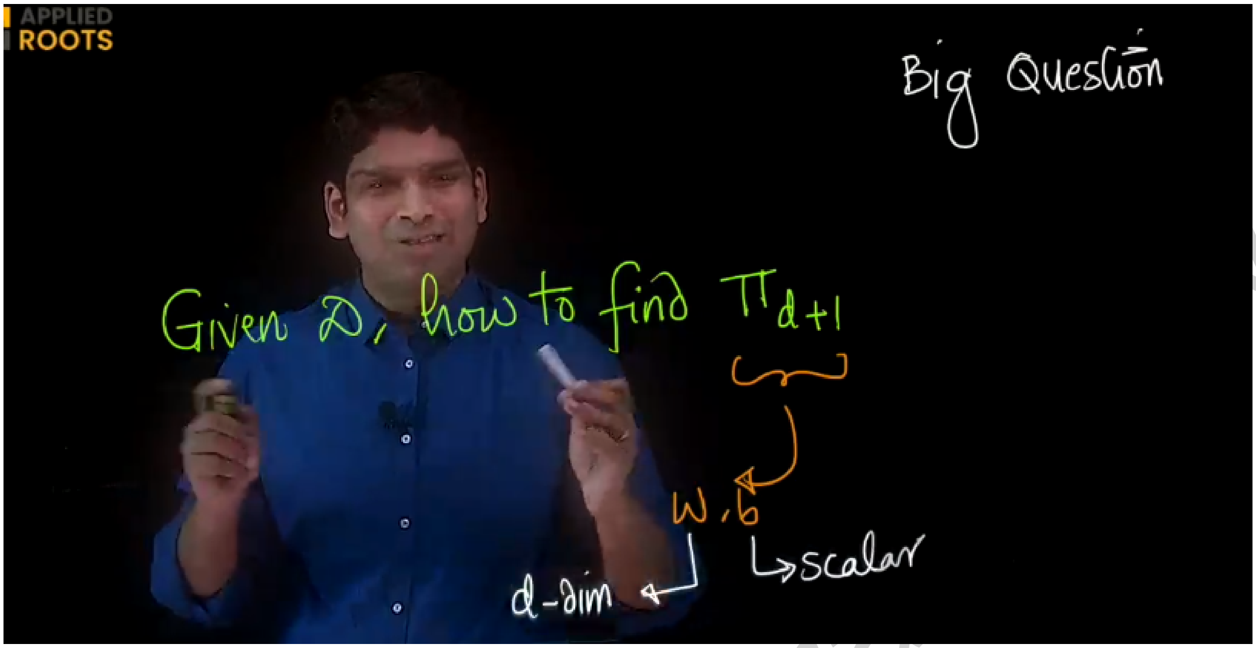

- Let's visualize data in 3D ,we have two features f1 ,f1 and our response variable is y.Assume that y is linearly related to f1 and f2 then we can fit a plane to our data.We try to find a plane such that most points lie close to the plane or on the plane itself.
- If we have 1 feature we use a line as our model,if we have two features we use a plane.




- Similarly if we have d features which are linearly related to y we can have a hyperplane in d+1 dimensional space.
- We cannot use lines and planes if features are not linearly related y

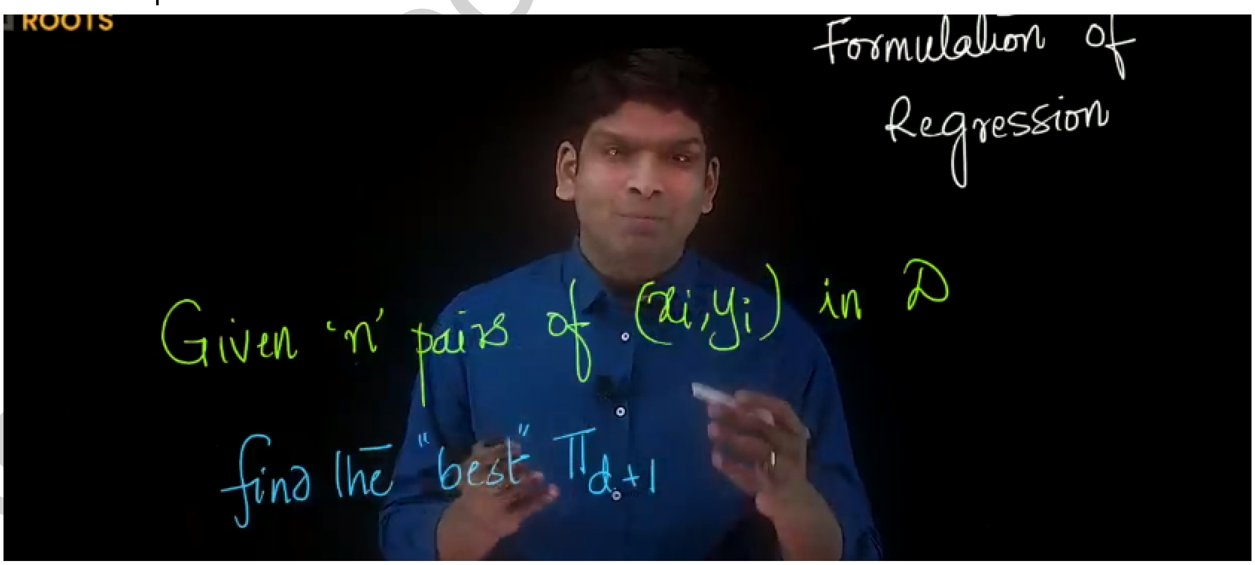
At timestamp 20.19 in video



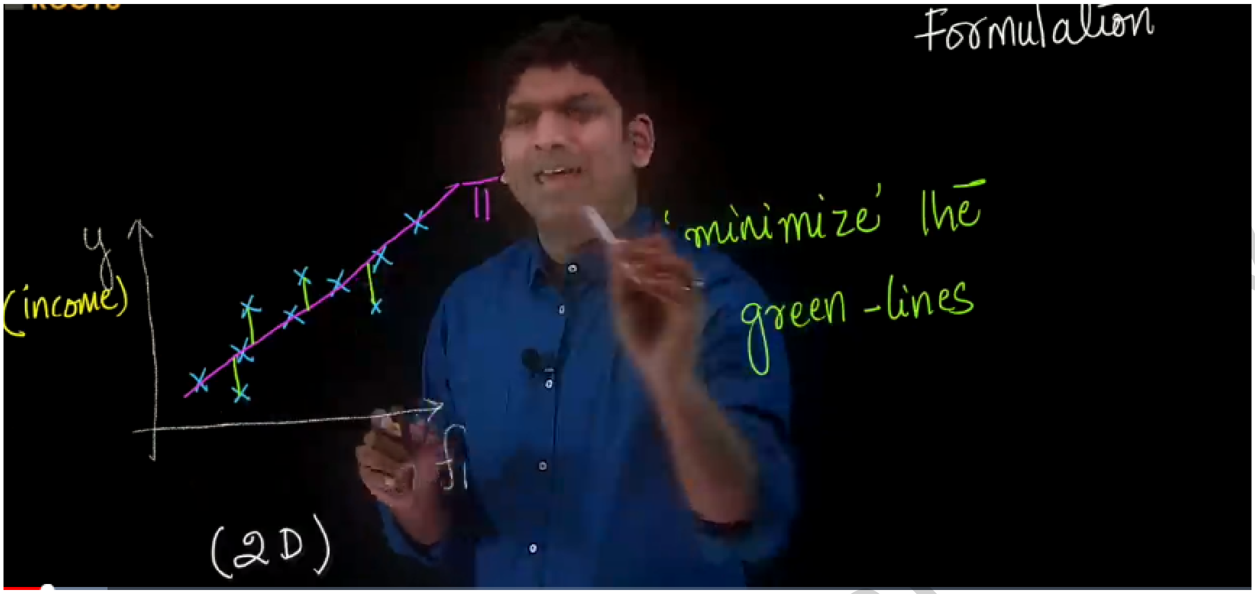

- The challenge is how do we find the plane .Lets find out.

# 3.4 Formulation of regression

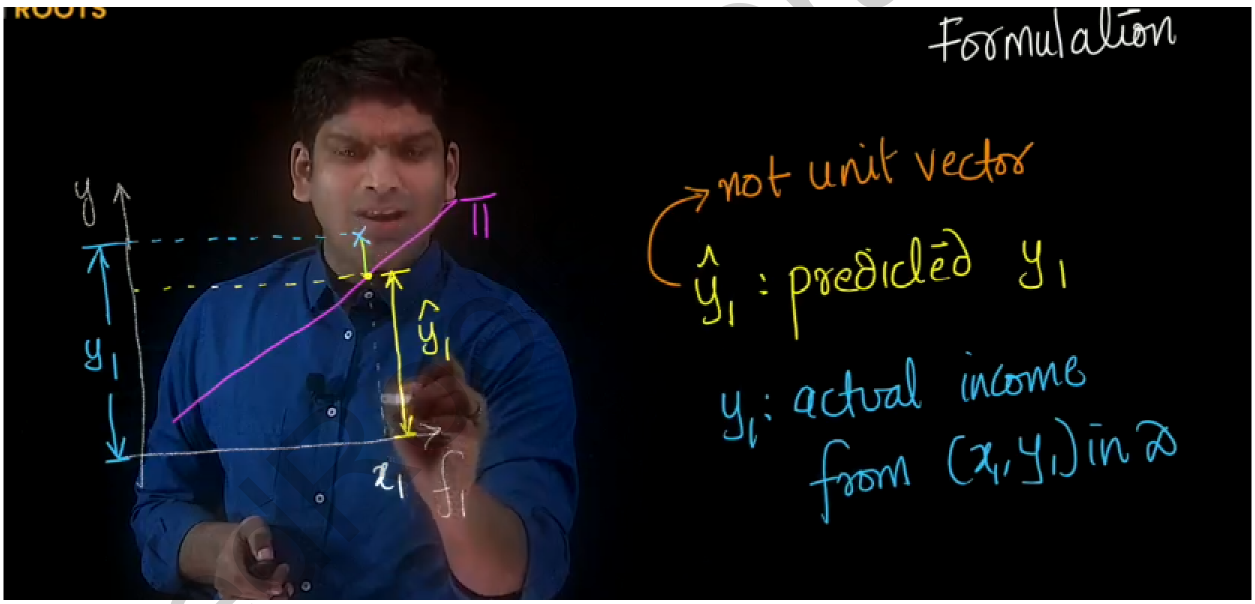
At timestamp 0.11 in video



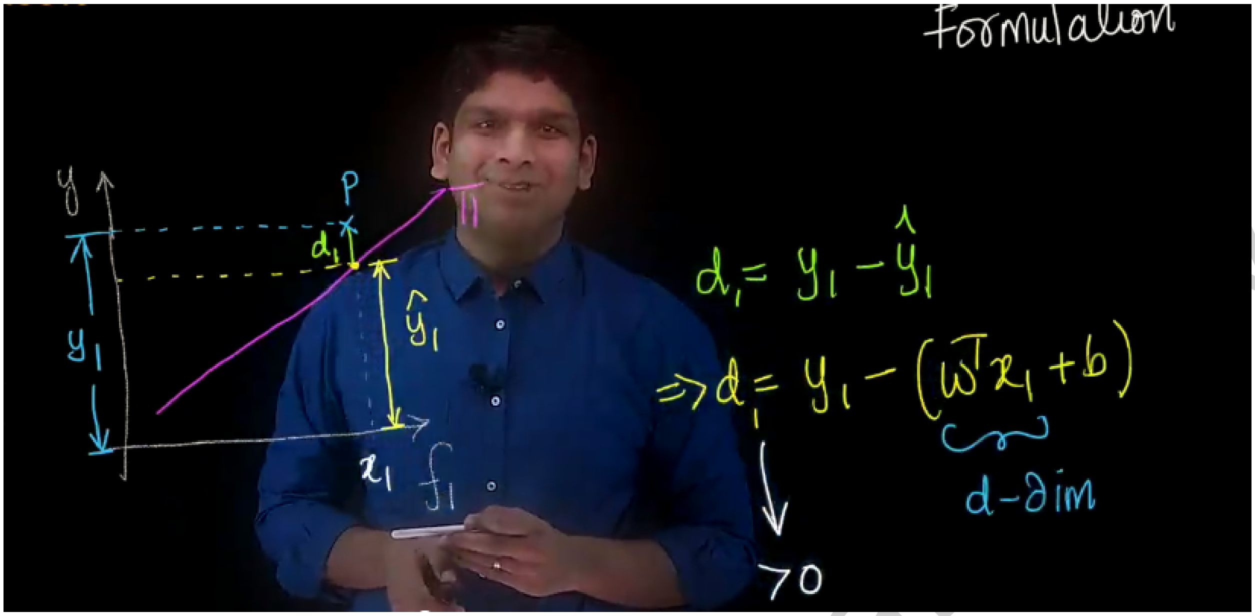

- Given (x,y) n pairs in dataset d,we have to find the best plane.We will try to formulate a regression problem from a mathematical point of view.




- In order to find the best plane we want to minimize the distance from the point to the plane .

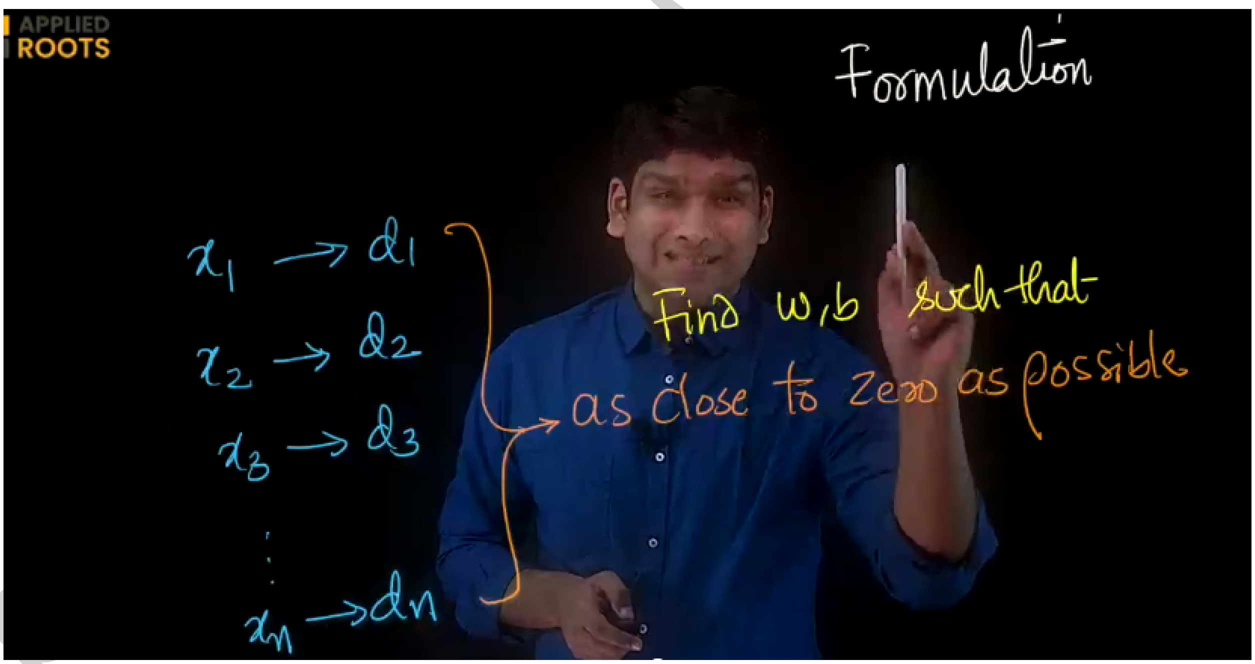



- The predicted value for y is the value which line gives us ,the actual y value is given in the dataset.Our predicted and actual y values may differ based on the line we have as our model



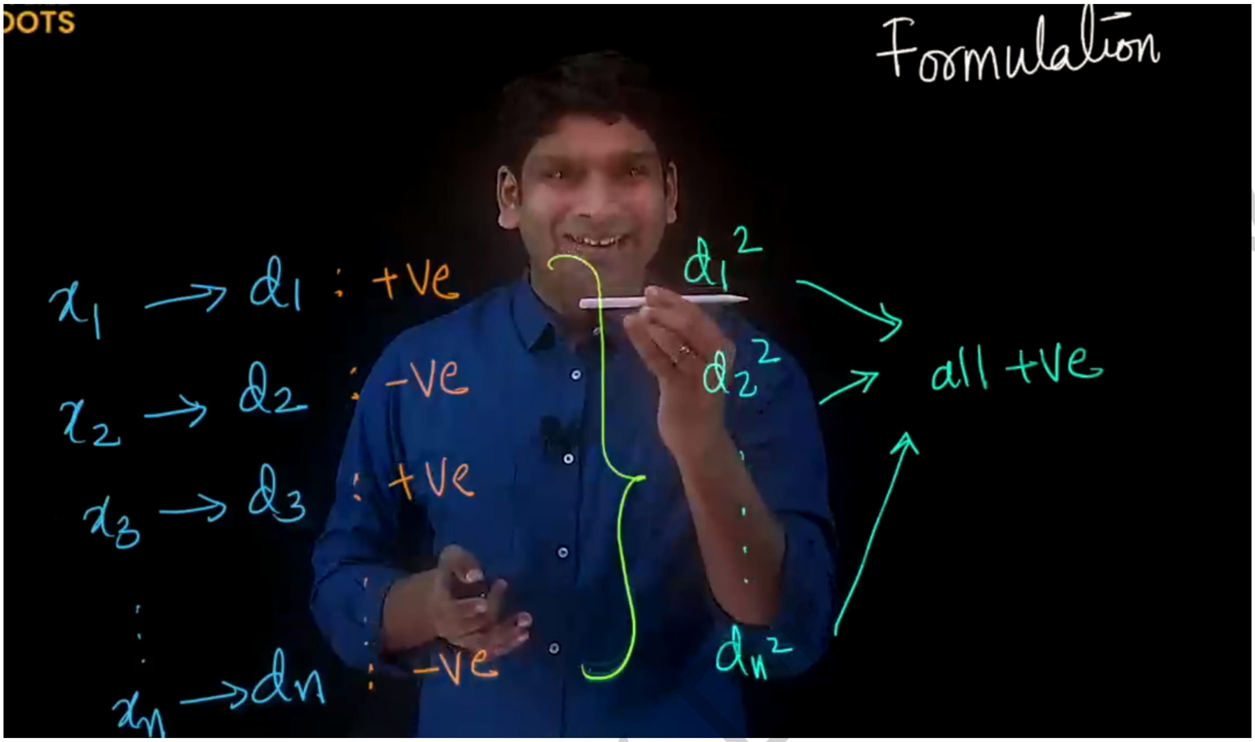

- When we say minimize the distance from point to the line we are actually trying to minimize the value d as shown .(we are minimizing the difference between actual predicted y)

At timestamp 12.36 in video



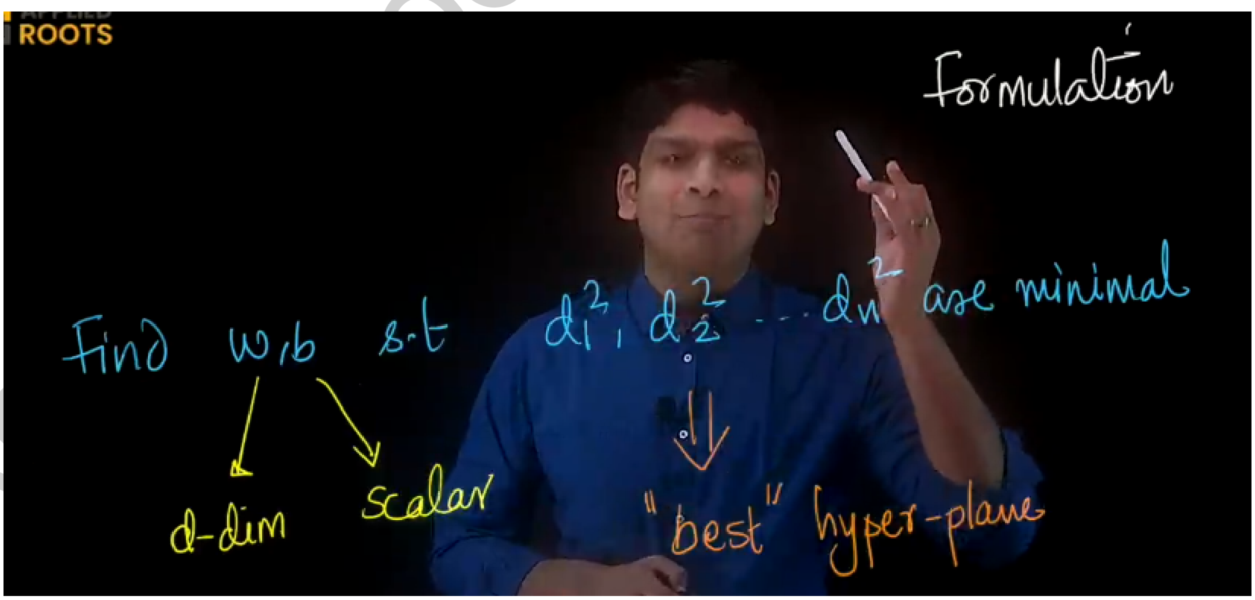

- Our objective here is to find w,b such that the distance (d) from point(x ) to the plane should be close to zero.
- We want the sum of distances as close to zero as possible.If distances are closer to zero the points are closer to hyperplane

At timestamp 13.57



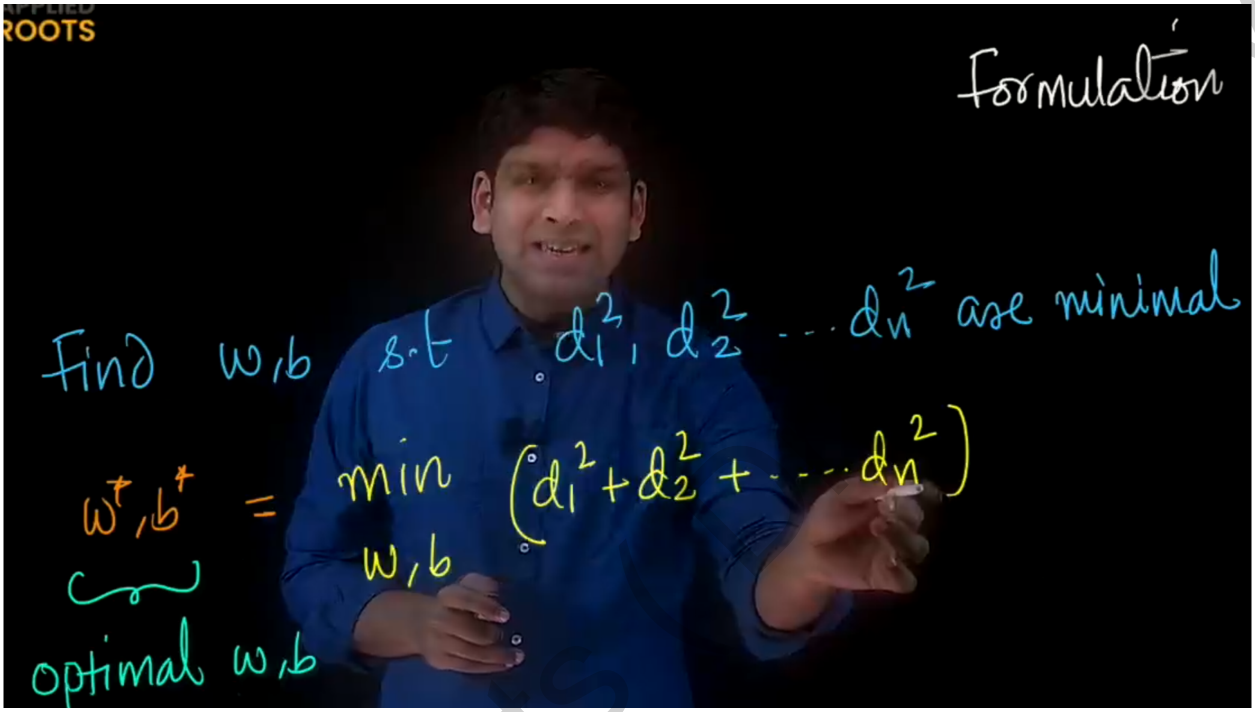

- For points which are either side of the plane the distances will be both positive and negative ,if we sum up the distances it will not be appropriate so we consider the sum of squares of distances.
- We want these sum of squares of distances as close to zero as possible

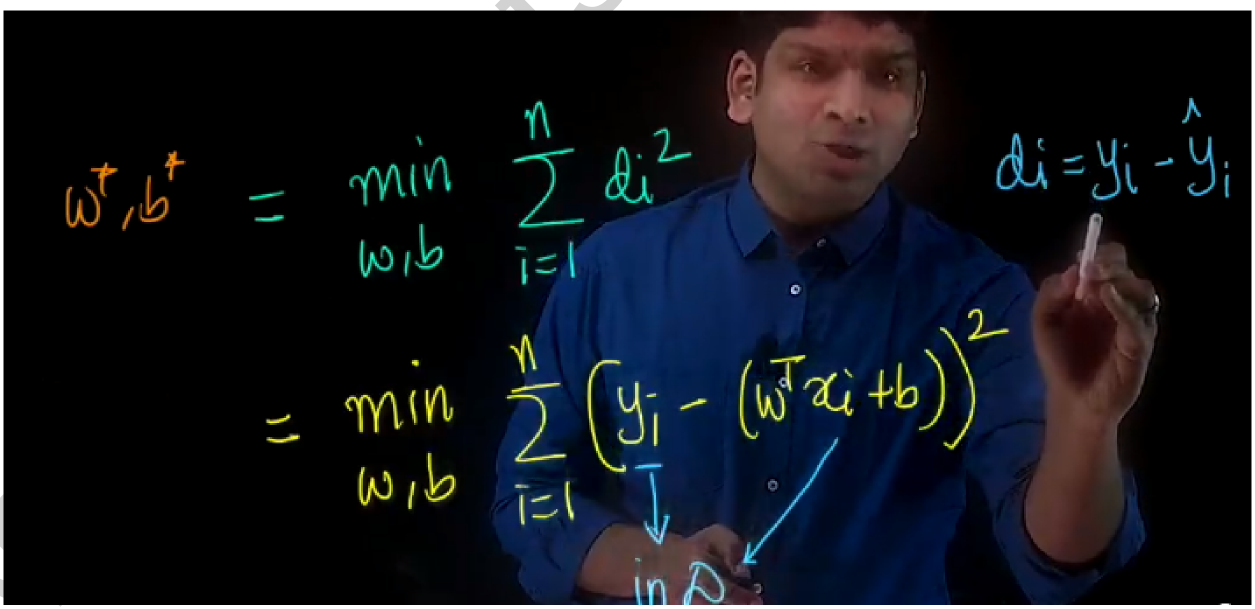
At timestamp 15.42



- Our whole problem boils down to finding w,b such that we can minimise the squared distances as much as possible.
- By minimising each if di^2  we arrive at best hyperplane.

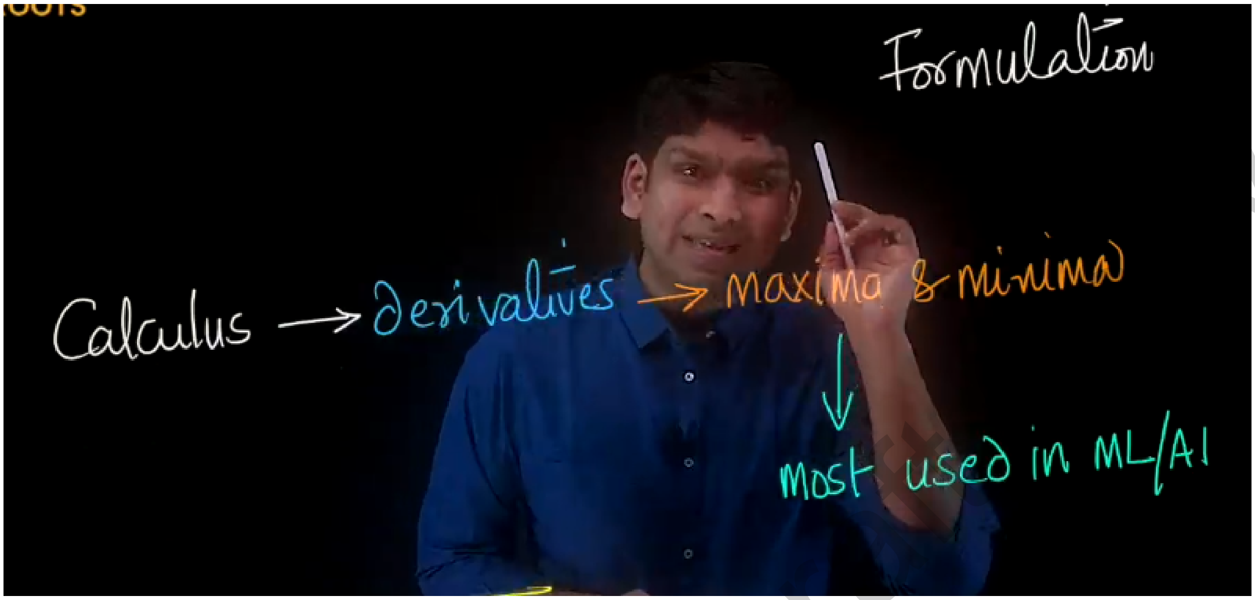
At timestamp 18.30 in video




- We have to find optimal w,b such that the sum of squared distances will be minimised.
- Here we are minimise loss so that we can obtain w,b

At timestamp 22.35 in video



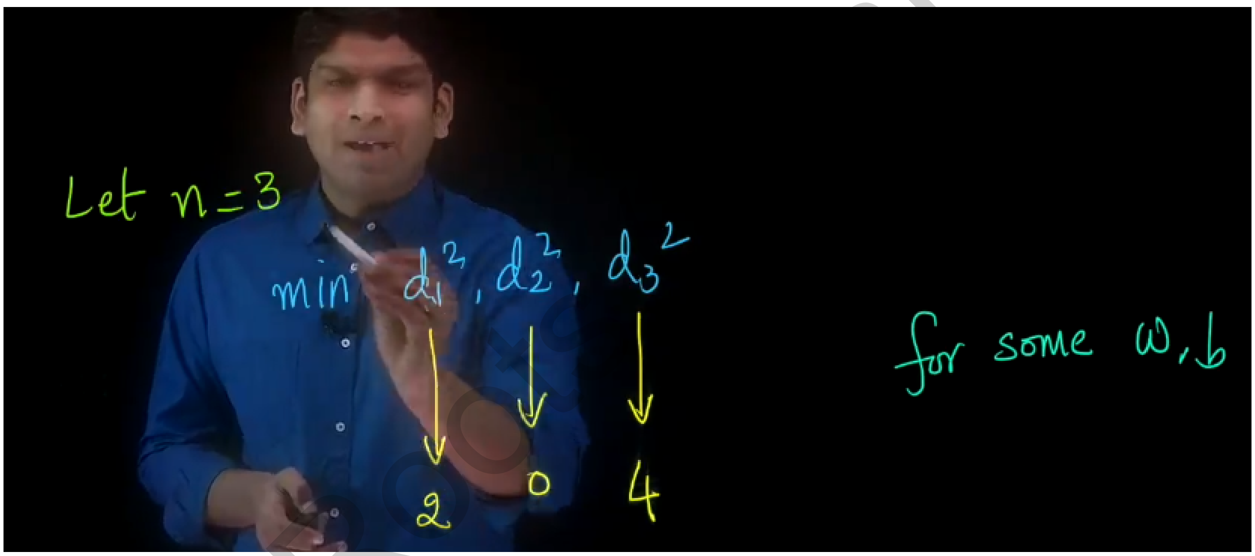

- To solve this regression problem we use maxima minima from calculus.



At timestamp in video

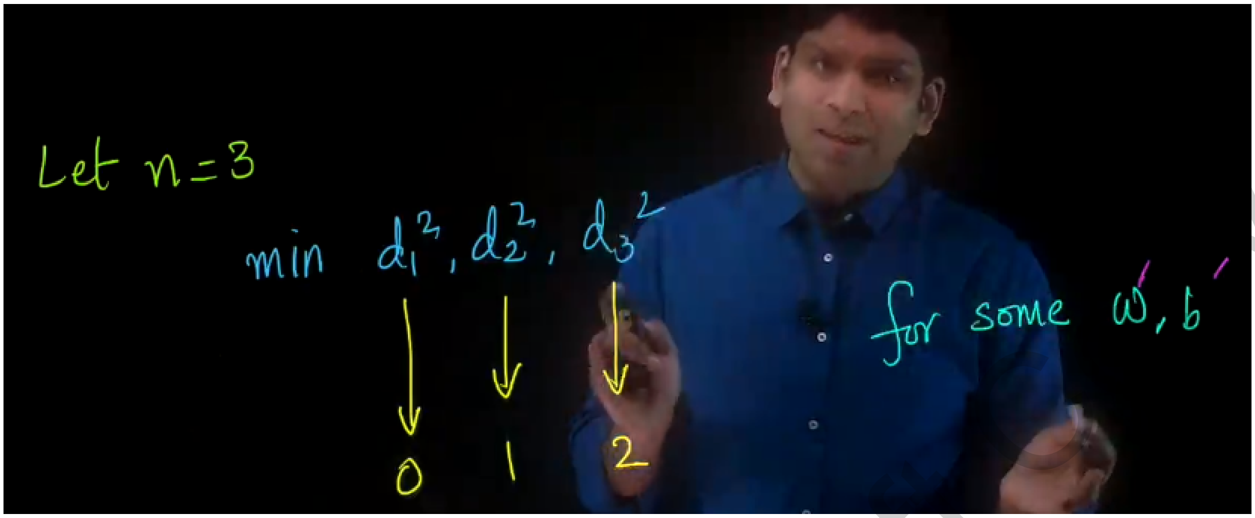
# 3.5 Minimisation of multiple values

At timestamp 2.08  in video



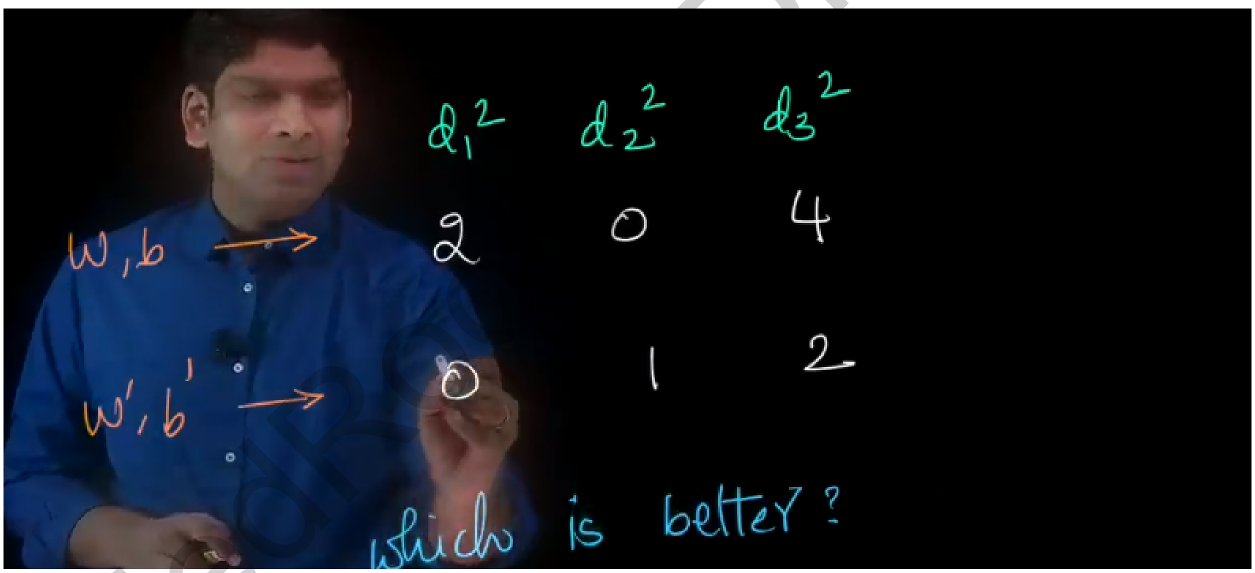

- Imagine we have 3 points and corresponding to each point we have squared distances as shown above.Lets assume a plane with some values of w,b as our model.(remember we can determine a plane if we know and b)

At timestamp 2.55 in video

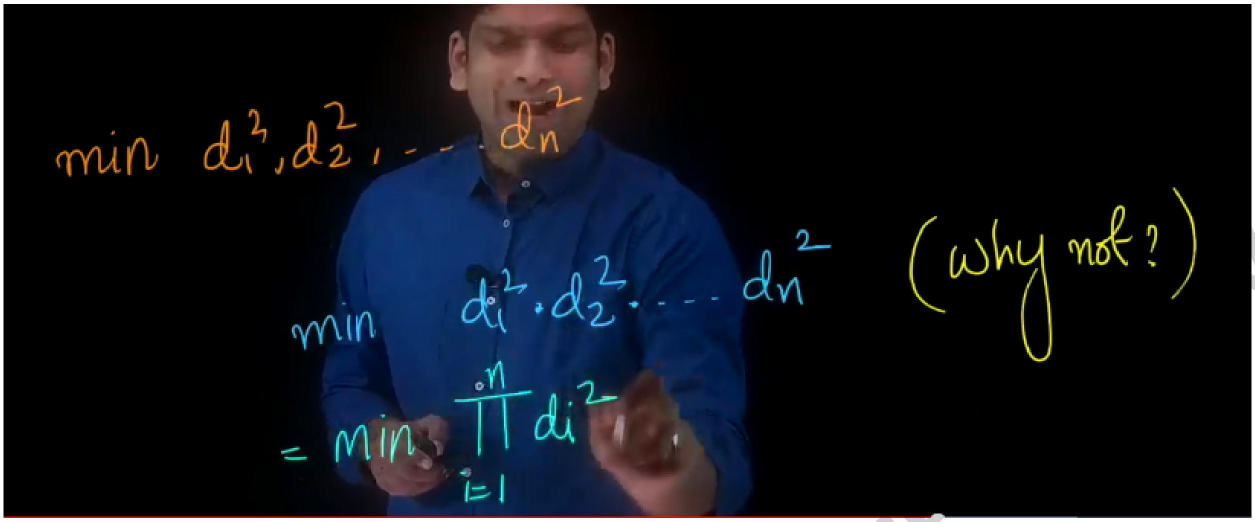


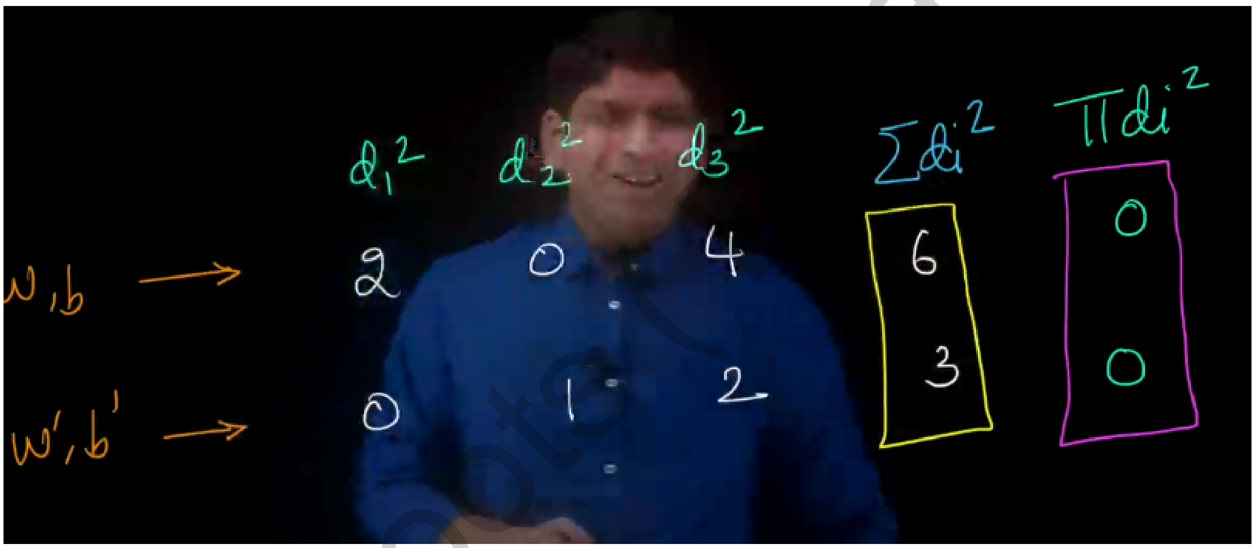

- For some other plane with w',b' the squared distances are as shown above
- As the plane changes the distances change.

At timestamp 3.40 in video



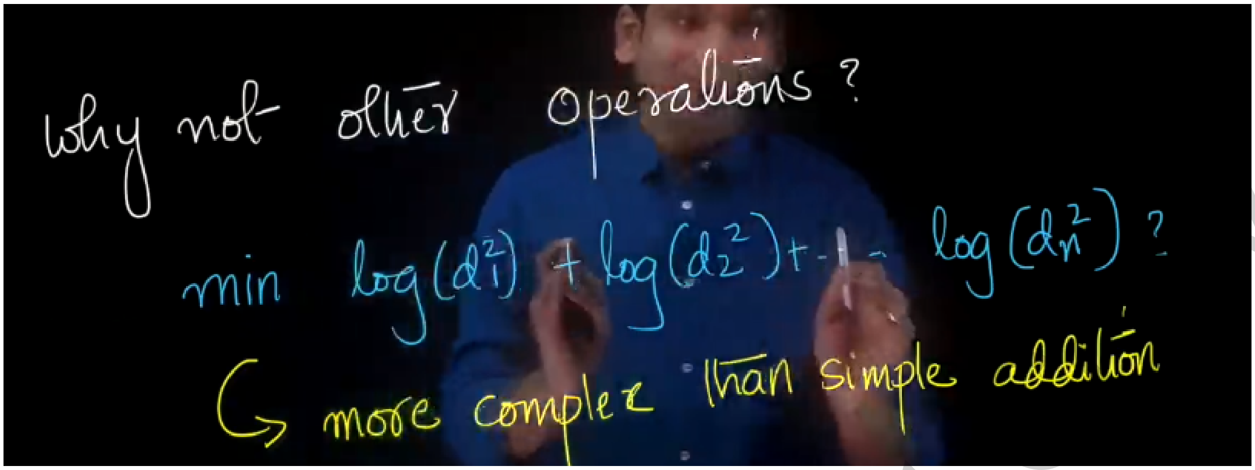

- We choose the plane which minimises total sum of squared distances.




- Minimising each squared distance is equivalent to minimising the sum of squared distances  and we cannot use the product of the squared distances even if one distance becomes zero(actual and predicted y are same ) total product becomes 0.

At timestamp 19.47  in video

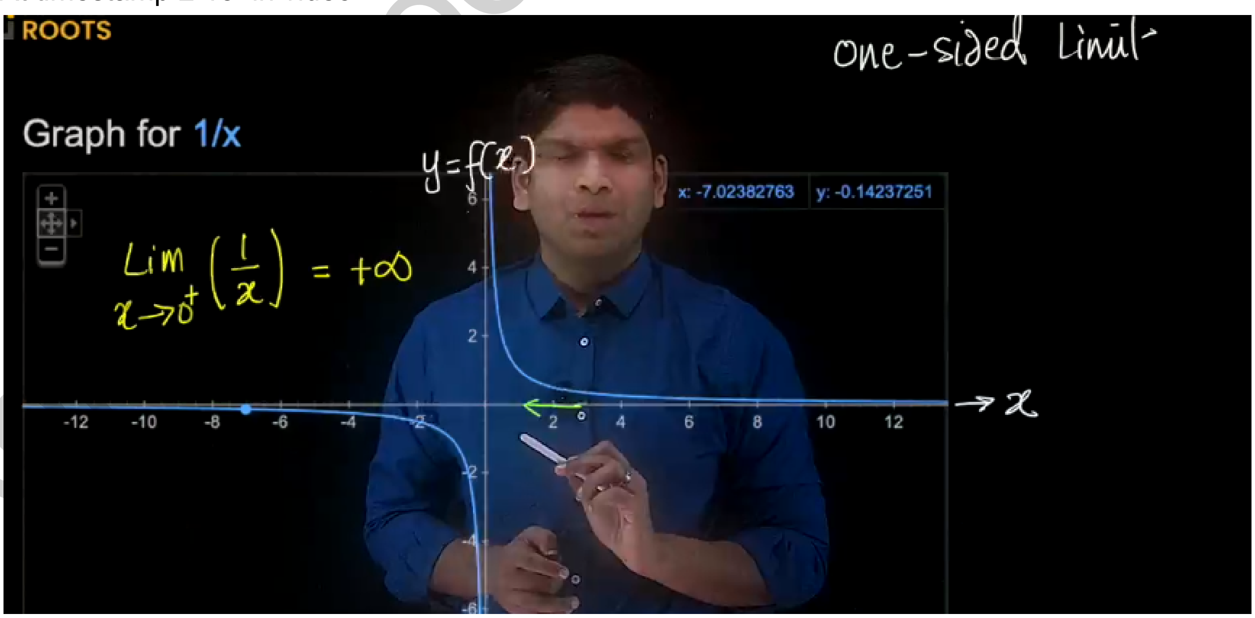



- We can use other operations than addition such as log. We don't want to increase the complexity so we use addition.

# 3.6 Limits, Range and Domain of a function

In this chapter we learn calculus ,limits and range for most popularly used functions in AI/ML

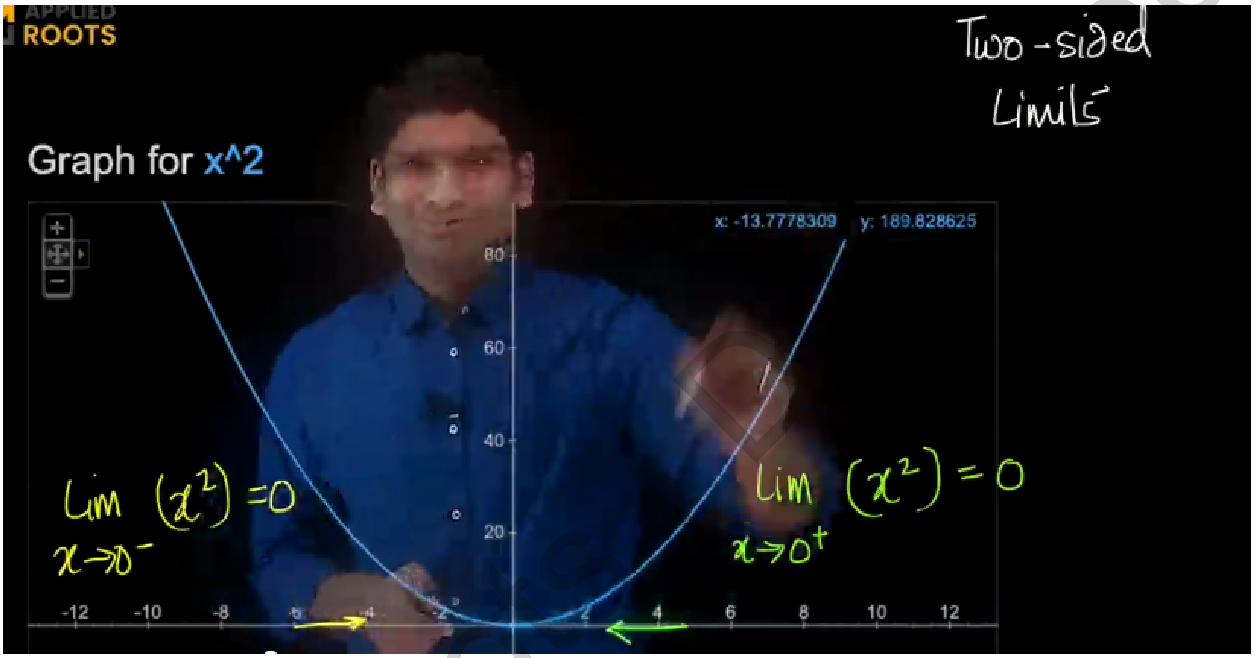
At timestamp 2.15  in video



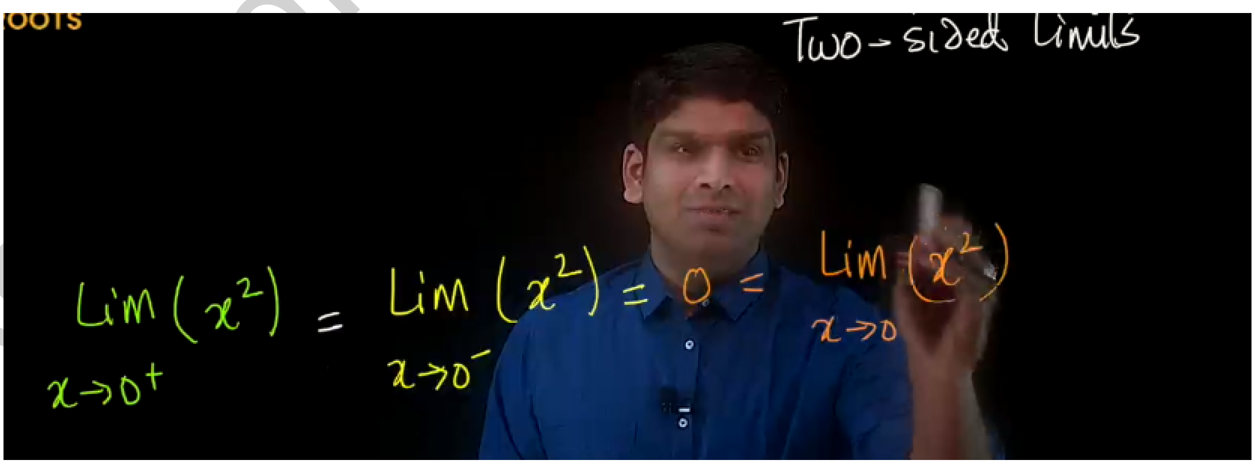

- Consider function f(x) = 1/x and lets understand the concept of one sided limits

- As you are coming closer and closer towards 0 from the positive side of the x axis,the curve is moving towards infinity.The same fact can be represented mathematically using limits as shown above and it is often referred to as the right/positive sided limit .
- As you are moving closer from the negative side of the x axis towards 0,the curve is tending towards minus infinity.The same fact can be represented mathematically using limits as shown above and it is often referred to as the left/negative sided limit .

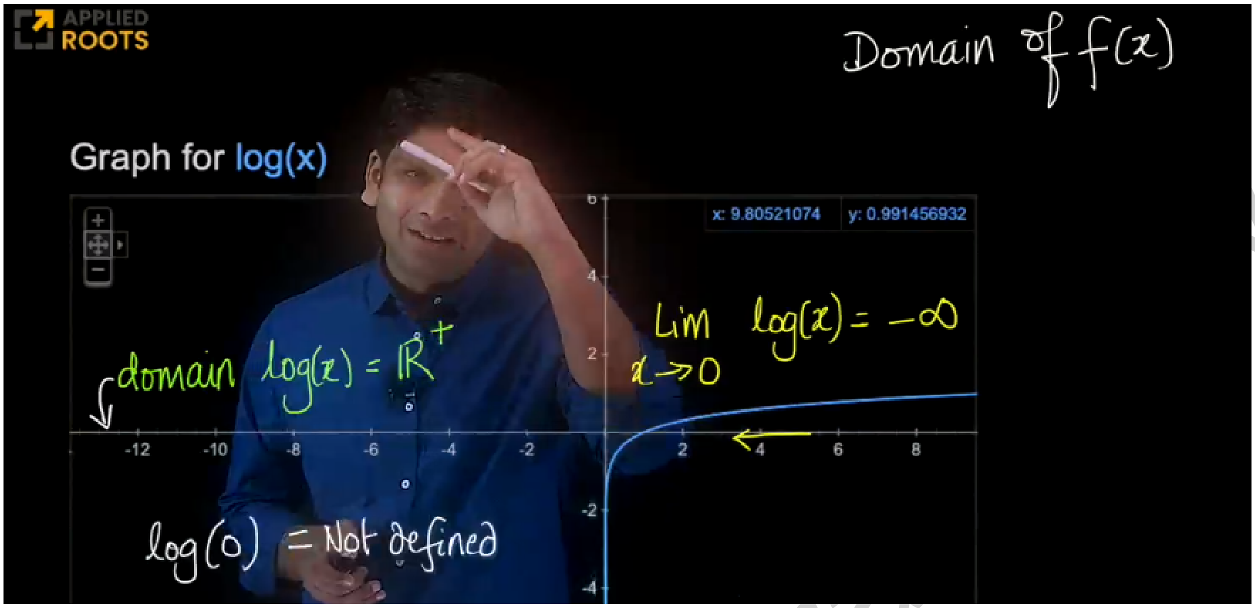
At timestamp 5.49 in video




- The above curve is a parabola,There is a similar concept called two sided limit
- As you are moving closer and closer towards 0 from the positive side of the x axis,the curve is moving 0 so the right hand limit is 0 similarly the left hand limit is also 0.
- We represent the above behaviour mathematically as shown below.

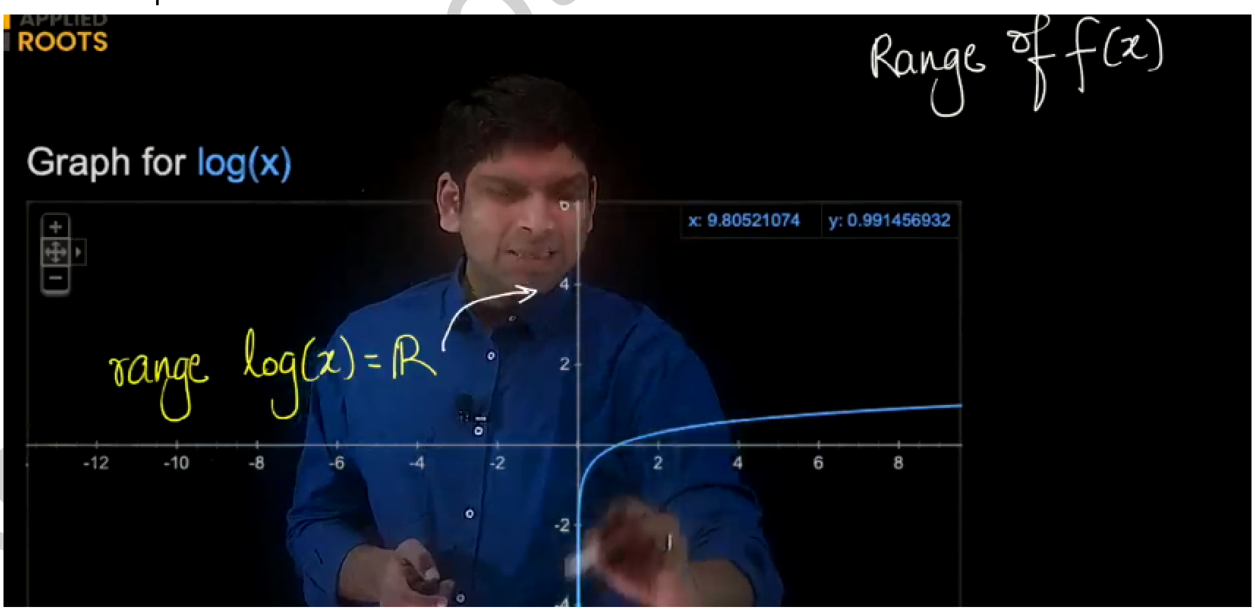


At timestamp 9.05  in video



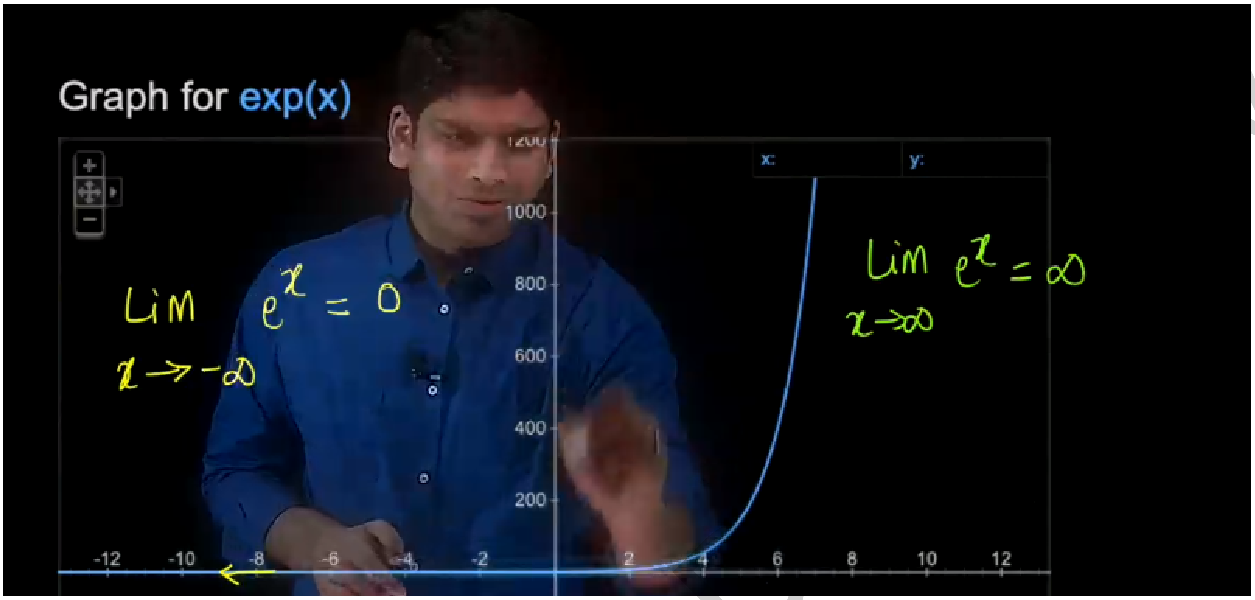

- The above is log(x),we use this function a lot in machine learning.we know that log of 0 and -ve numbers is not defined, log is defined only for positive numbers .The set of possible values which a function can take is called as domain and the domain for log(x) is positive real numbers.
- From the plot we can see that the right handed limit when we move closer to 0 from the positive x axis the function is tending towards - infinity.

At timestamp 11.41  in video

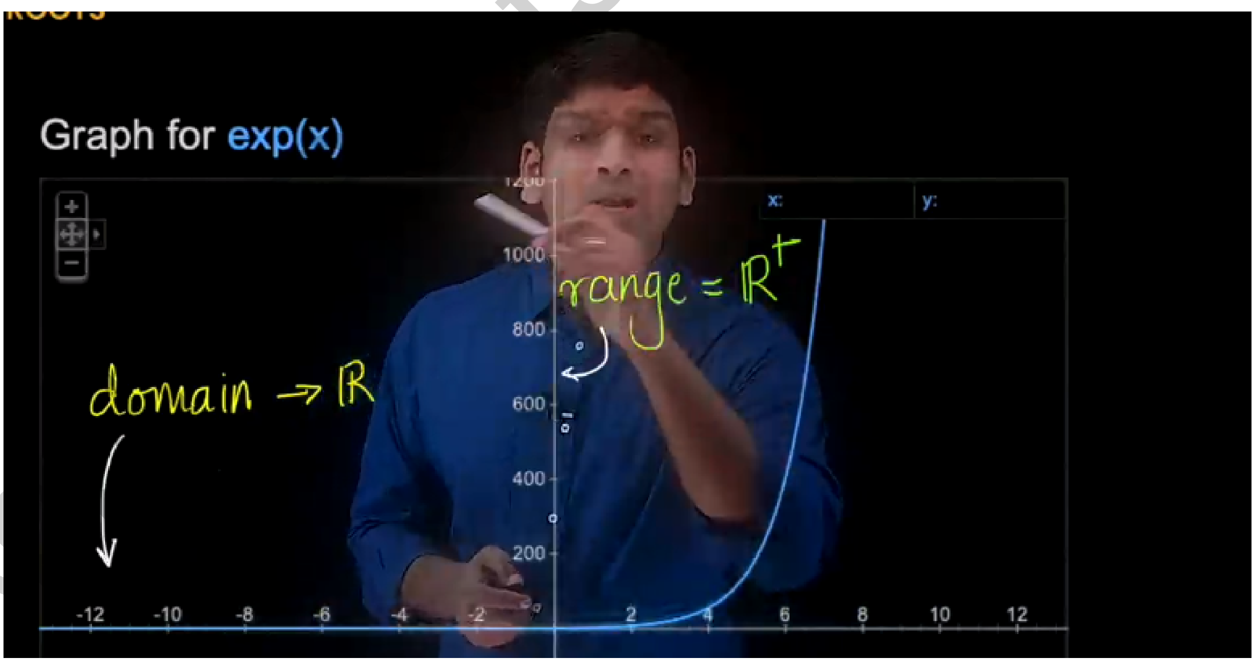



- The values taken by the function on y axis is called range ,log(x ) can be positive ,0 and negative .So the range of function y=log(x) is all real numbers.

At timestamp  12.40  in video

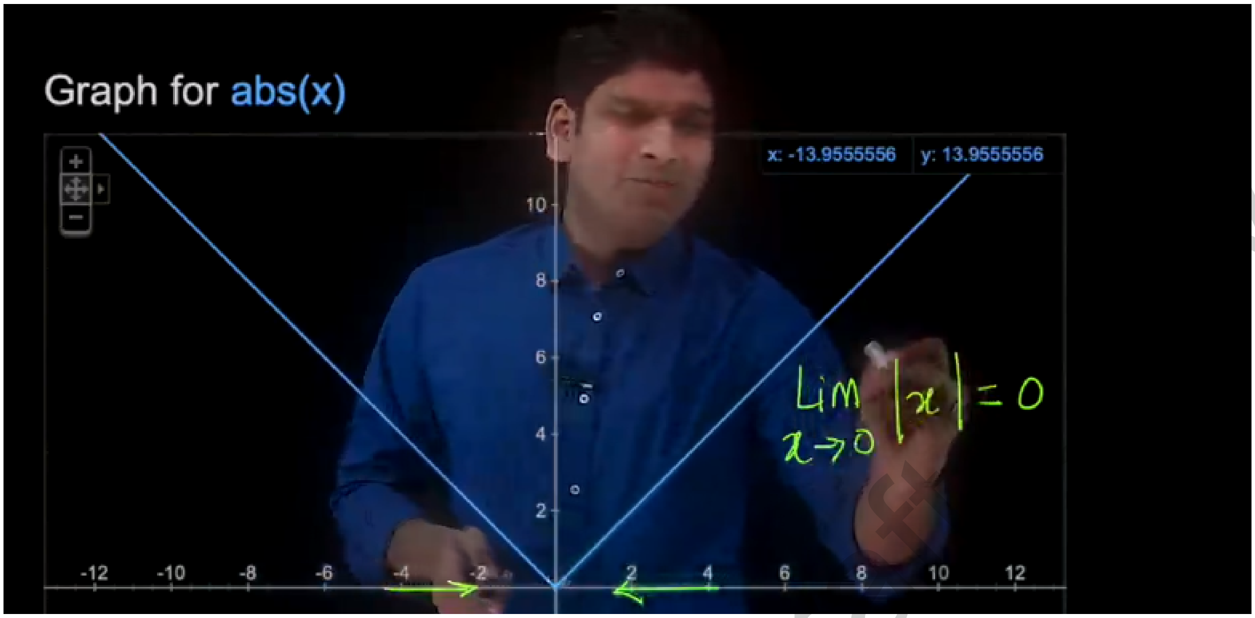



- Let's consider the function e^x,as we move closer and closer to  -∞ the function tends towards 0.It can be mathematically represented using limits as shown above .
- Similarly as we move towards +∞ the function also moves towards +infinity.

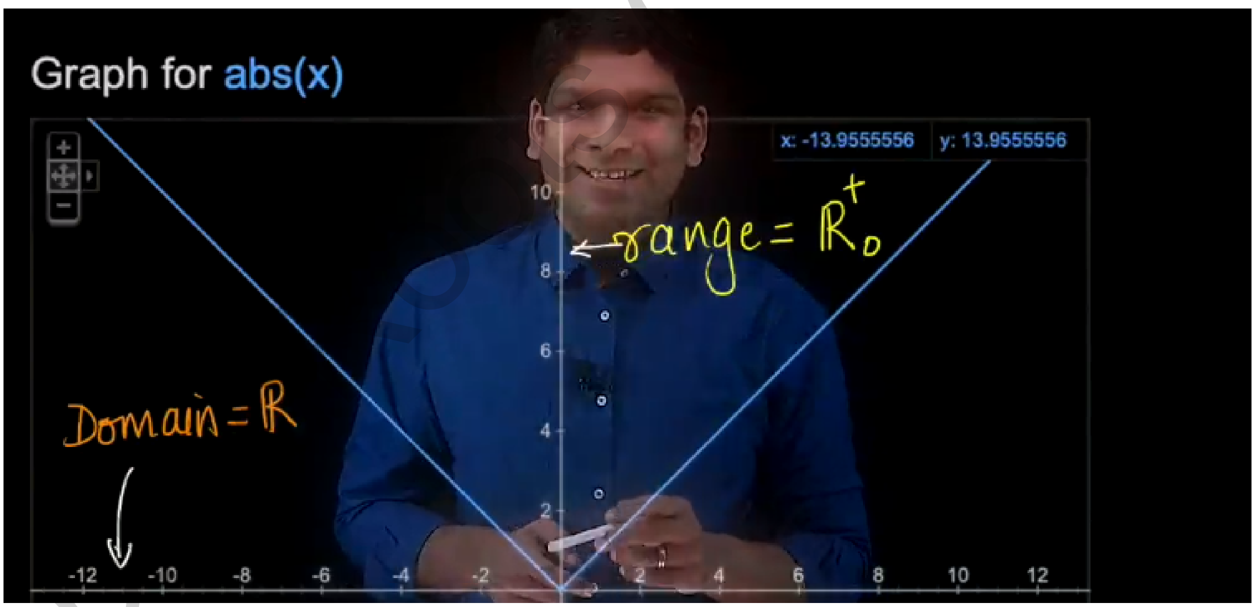



- The domain of the function is set of all real numbers  and the range is all positive real numbers.(R+ means all positive real numbers excluding 0)

At timestamp 16.01  in video

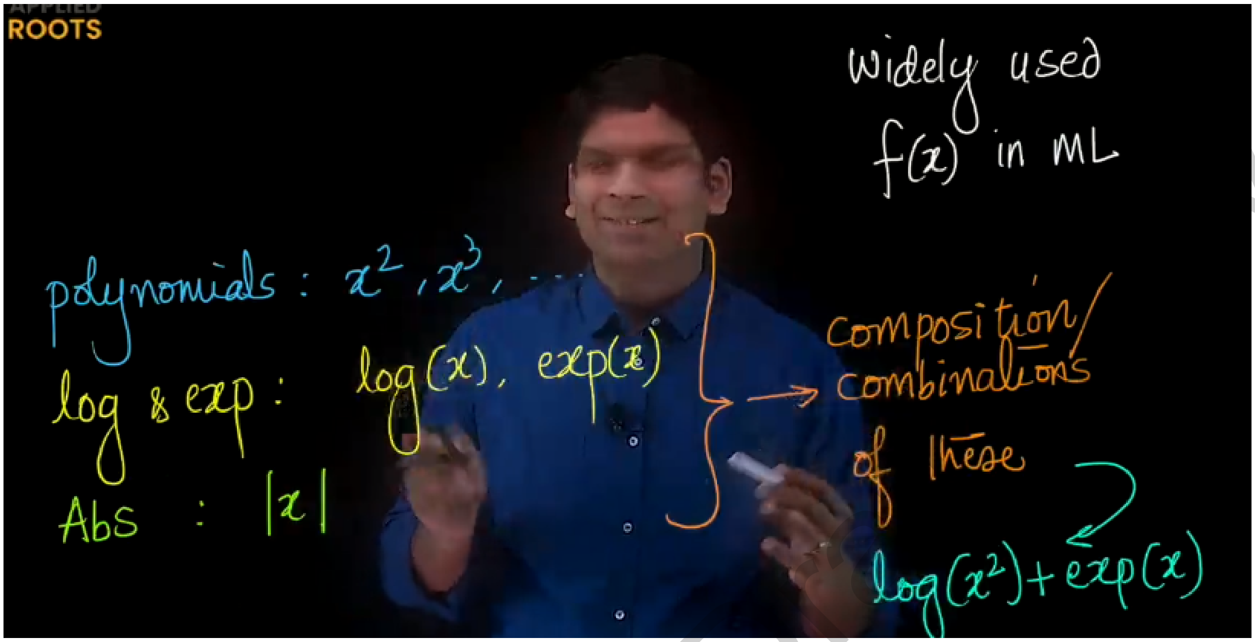



- The above function is absolute value function,the both sided limits of the function are 0 as shown above.




- The domain of the function is all positive real numbers and the range of the function is all positive real numbers including 0 because at x=0 the function gives 0 .

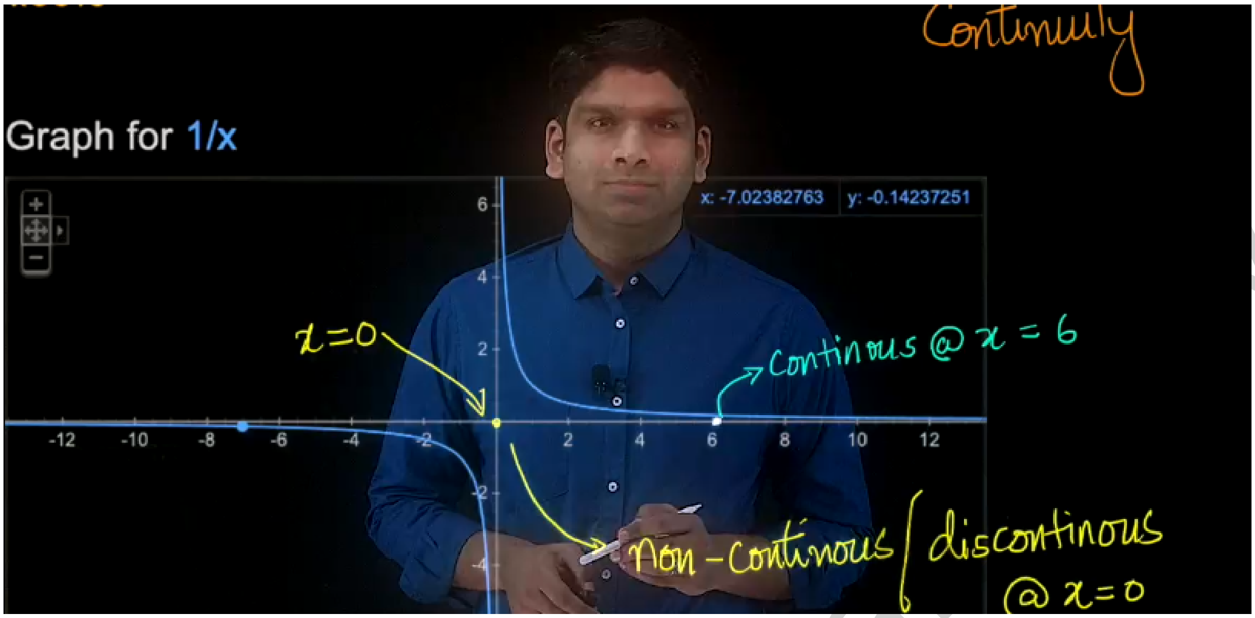
At timestamp 17.59 in video




- We use the above shown functions extensively in ML.

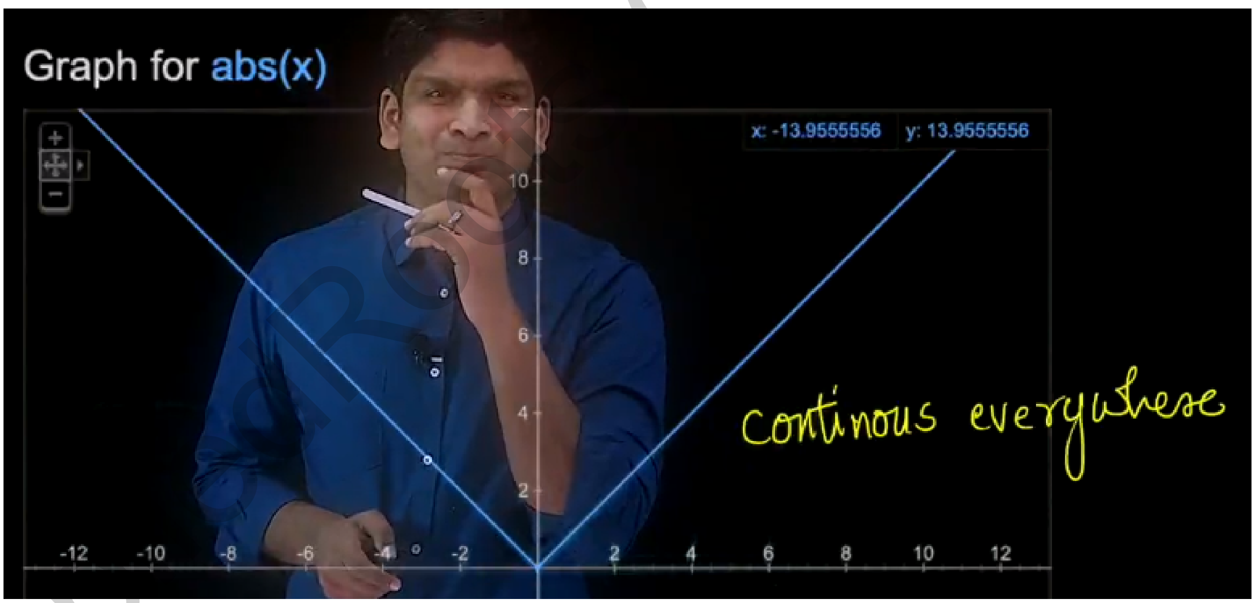
# 3.7 Continuity of functions

At timestamp 0.19 in video



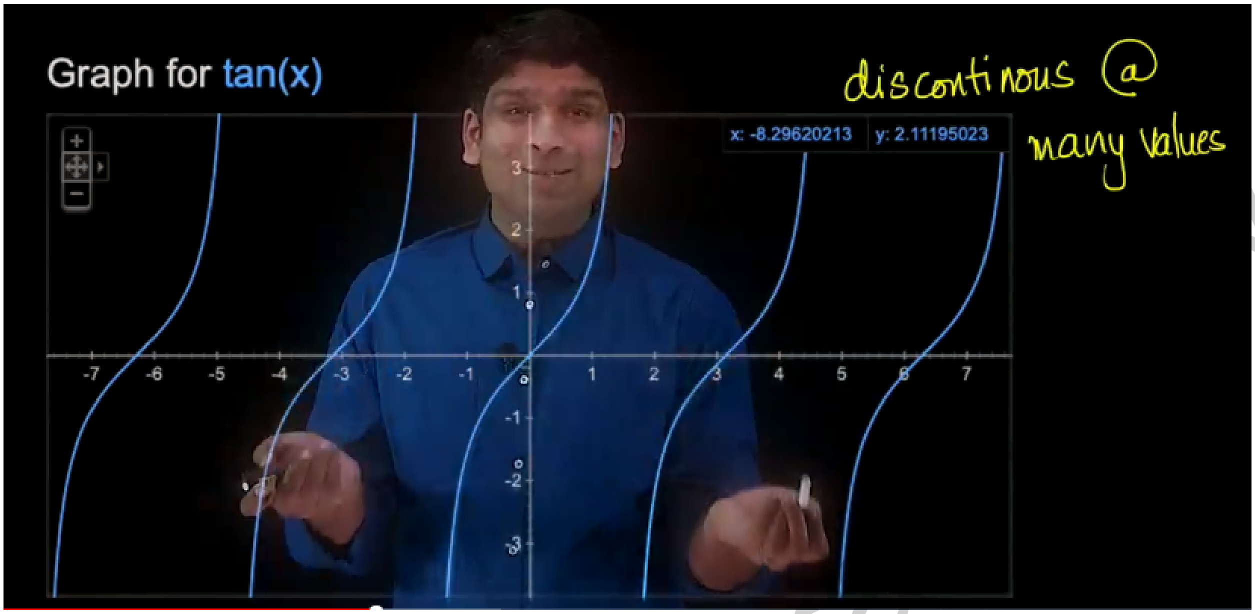

- Let's understand the concept of continuity of functions.
- Consider the function 1/x and from the plot you can see that at x=6 the function is continuous ,but at point x=0 the curve is discontinuous.

At timestamp 1.20 in video



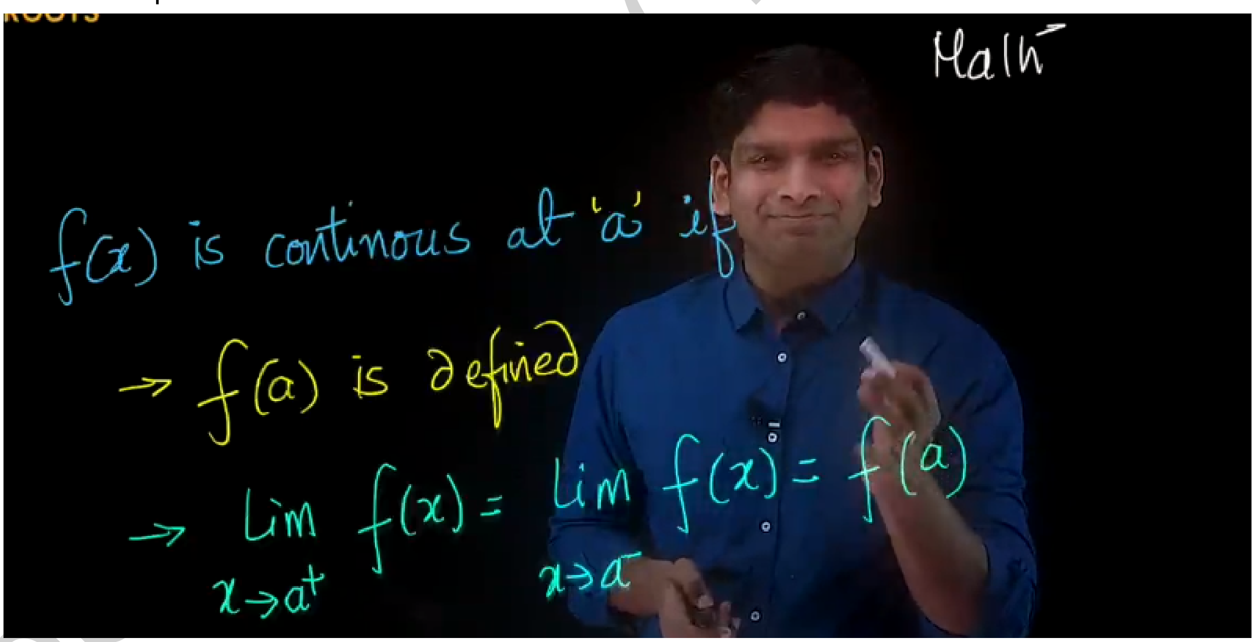

- If we consider the above function it is  continuous everywhere even at x=0 th function is continuous,there is no gap or break or discontinuity.

At timestamp 2.08 in video




- For the function tan(x) you can clearly see from the above plot that there is a lot of discontinuity.

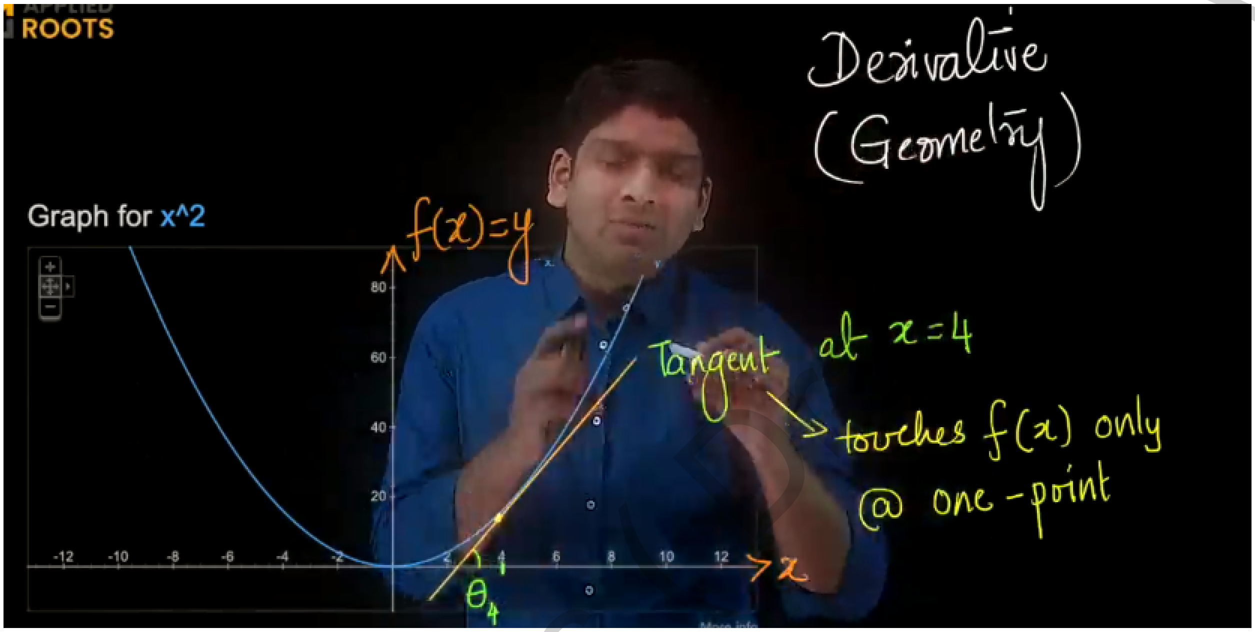
At timestamp 3.32




- Now that we have seen enough examples ,we define what continuity is mathematically as shown above.
- A function f(x) is said to be continuous at a point x = a, in its domain if the following three conditions are satisfied:
    1. f(a) exists (i.e. the value of f(a) is finite)
    2. Lim x→a f(x) exists (i.e. the right-hand limit = left-hand limit, and both are finite)
    3. Lim x→a f(x) = f(a)

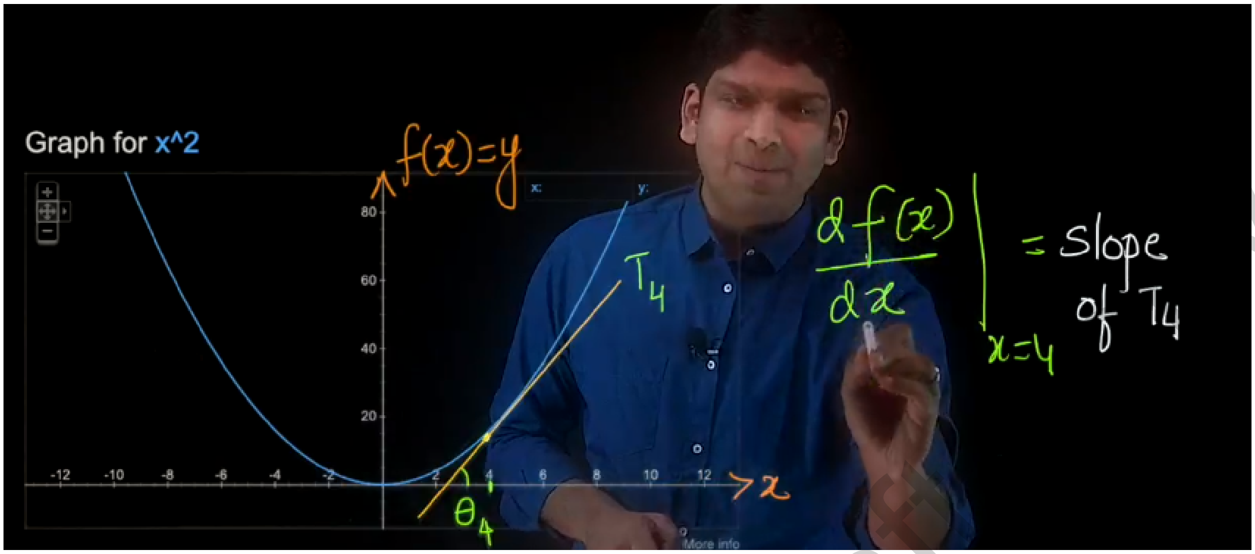
# 3.8 Derivatives: geometric intuition

At timestamp 0.24  in video



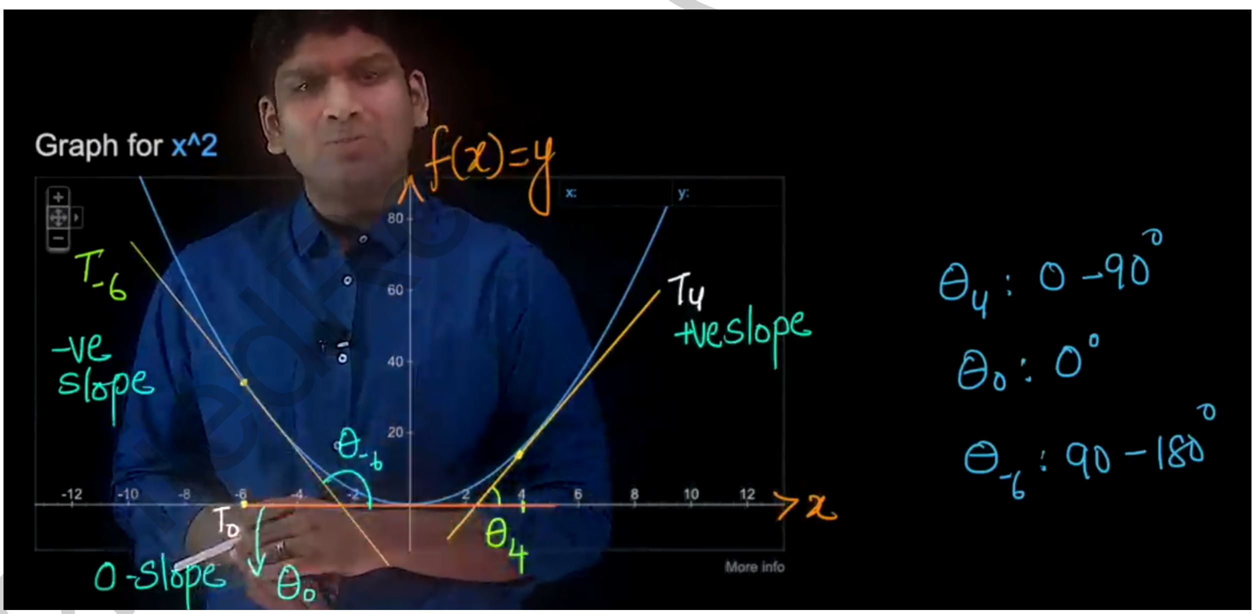

- Let's understand the concept of derivatives,consider the function f(x)=x^2 which is a parabola as shown above.
- A tangent is a straight line that touches the curve exactly at one point.
- If we consider point x=4 corresponding y will be 16 ,if we draw a tangent to the curve at this point x=4 and we call it as T4.This tangent makes some angle with x axis we call it $\theta$4.

At timestamp 3.35 in video




- The derivative of f(x) with respect to x at x=4 as the slope of tangent T4 and it is denoted mathematically as shown above.
- The slope of T4 is nothing but tan$\theta$4.
- The slope of tangents at different points on the curve could be different.

At timestamp 6.19 in video

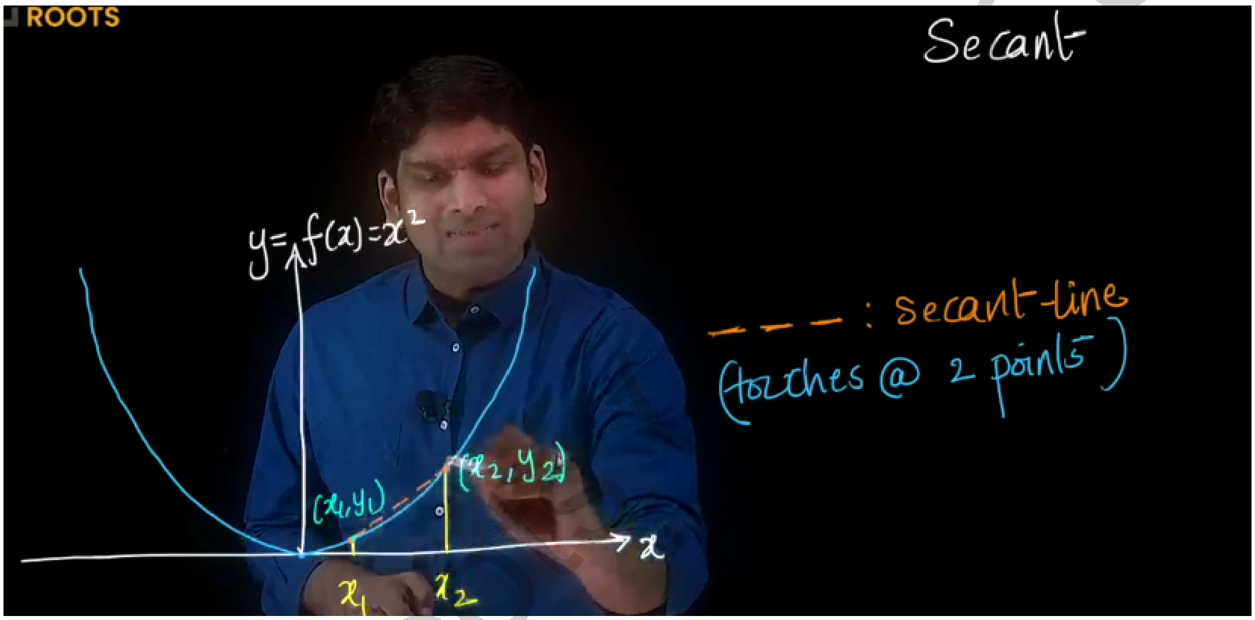



- Consider three tangents and angles made by the tangents with x axis as shown above.We can clearly see that T4 is having positive slope ,T6 is having negative slope and T0 has zero slope.
- If slope at a point is positive it means that the underlying curve is increasing on the other hand if the slope is negative at a point the curve will be decreasing.

- If the slope of a tangent at a  point is zero at a point it's slightly complicated and we will discuss it.

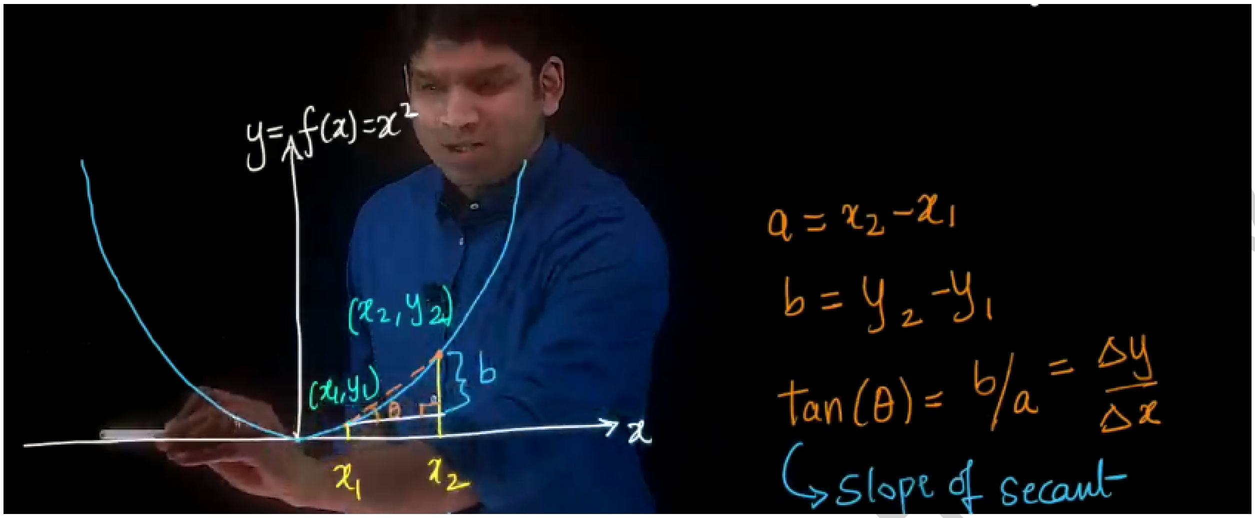
# 3.9 Derivatives: rate of change + Math

At timestamp   0.55 in video

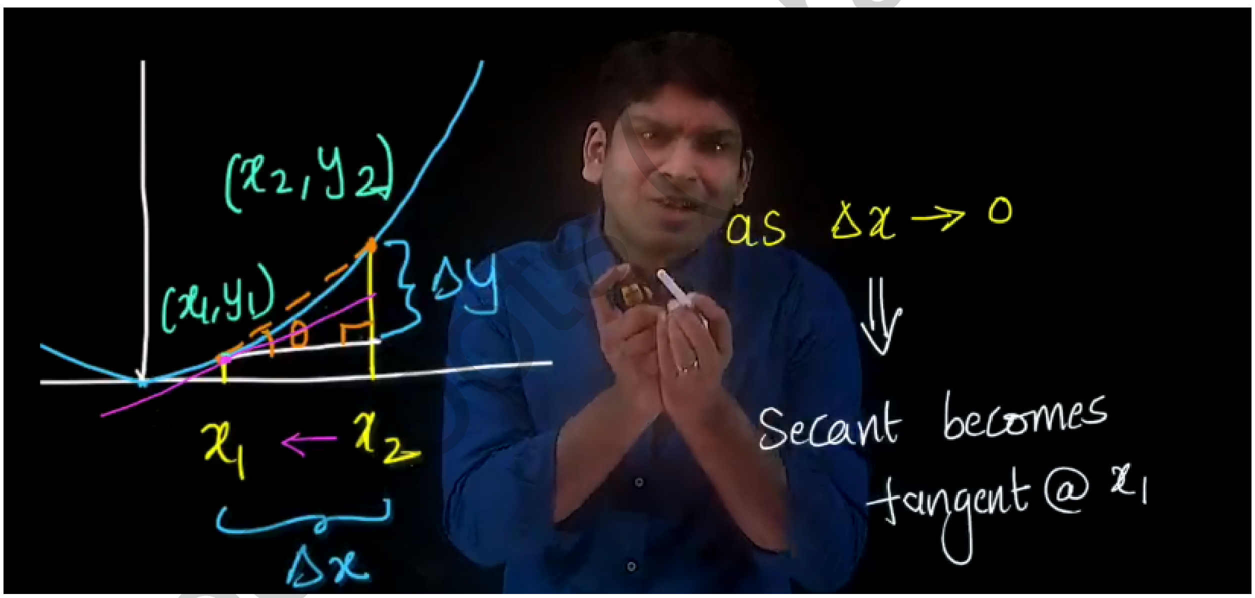



- Let’s understand derivatives from a rate of change perspective .Consider two points on the curve x1,x2 and their corresponding coordinates as shown above .
- The line between two points is called as secant line

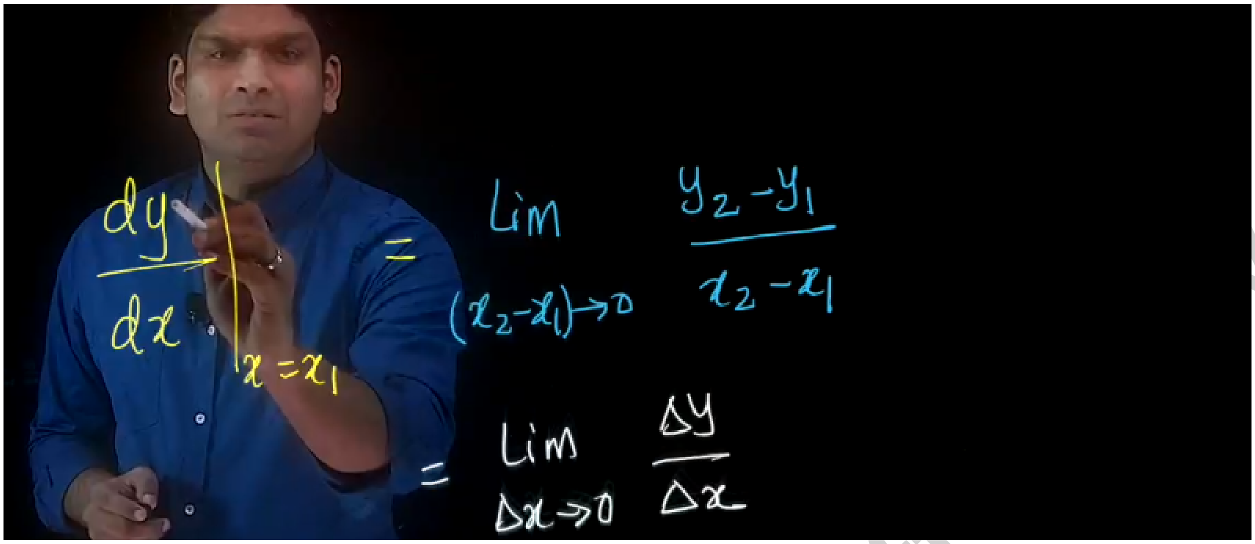
At timestamp  1.32   in video




- Slope of the secant line can be calculated as shown above.

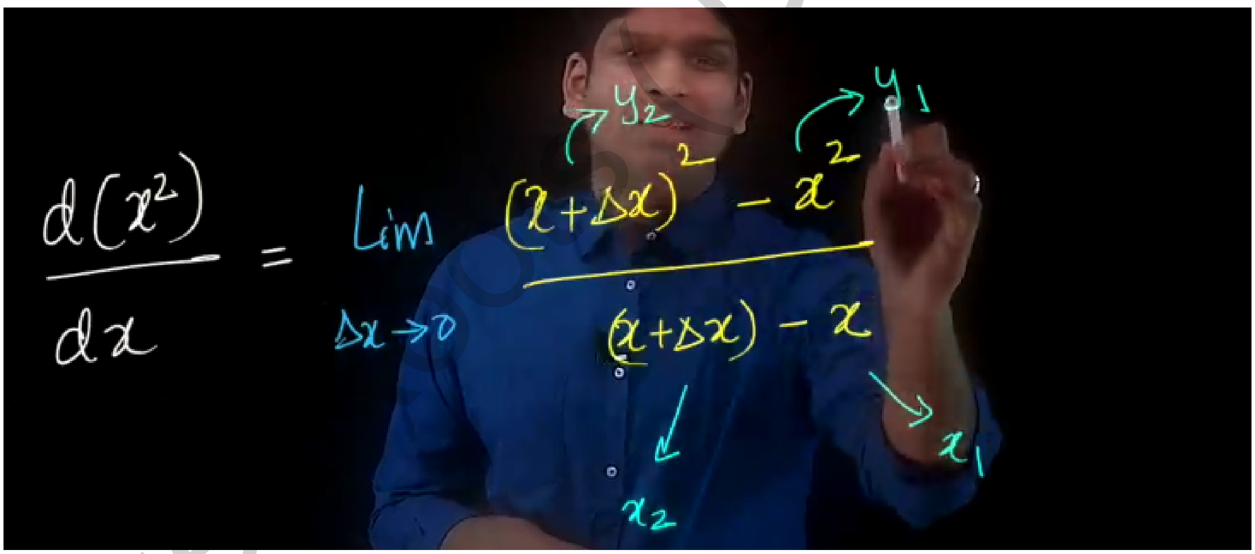


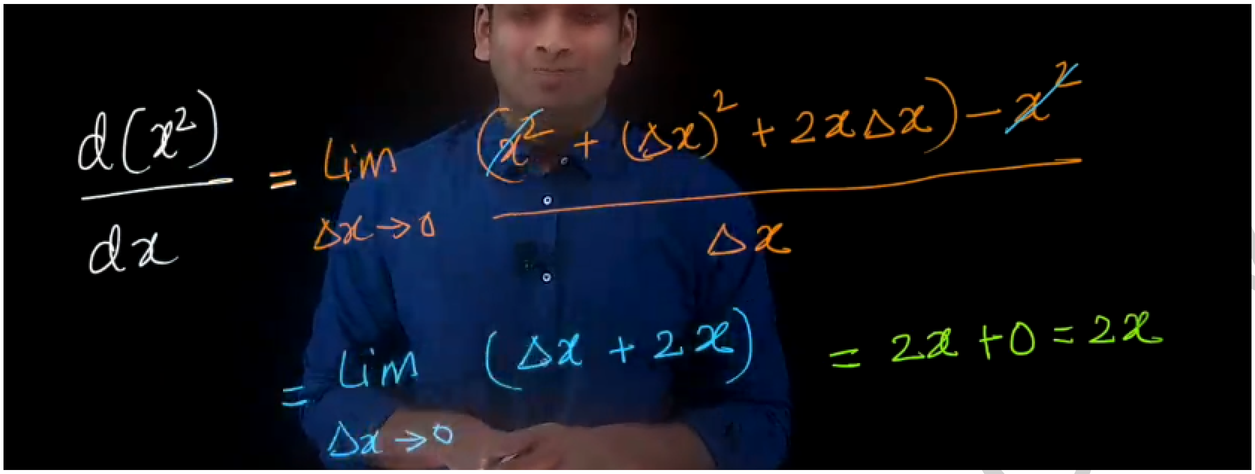

- When(x2 - x1) $\Delta x \to 0$ secant becomes tangent x1.

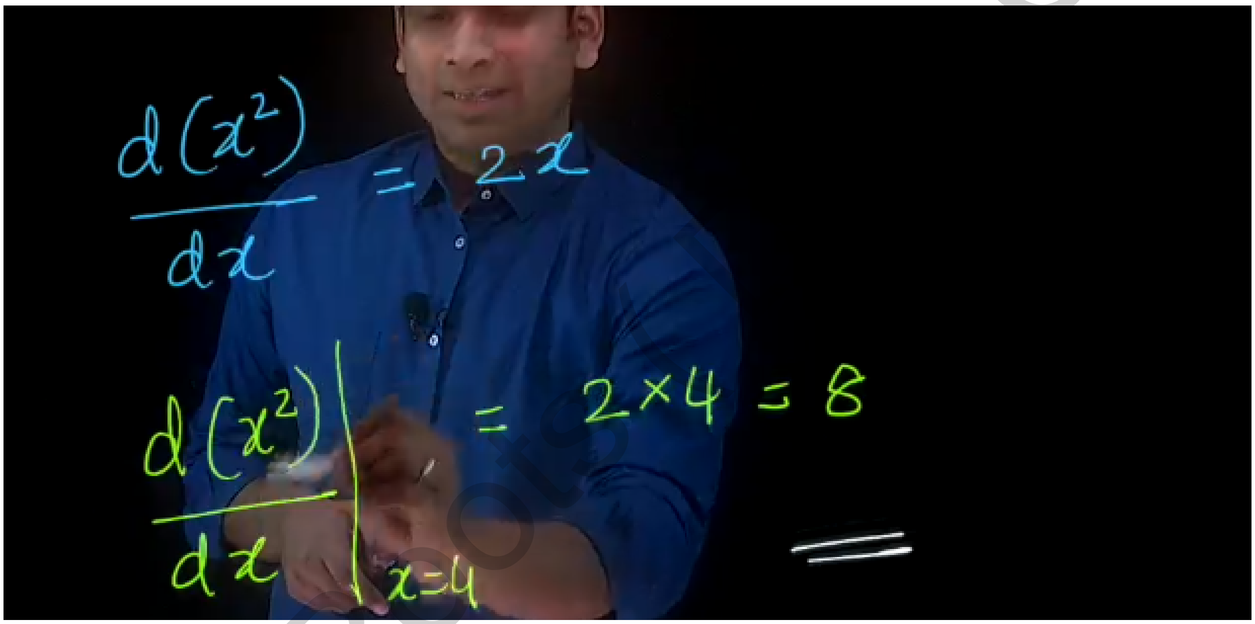



- The derivative of function y=f(x) at x=x1 can be written using limits as shown above
- So the derivative is the rate of change of y around the point x=x1.Instead of computing the derivative at a particular point we can compute it at all points.

At time stamp 9.56



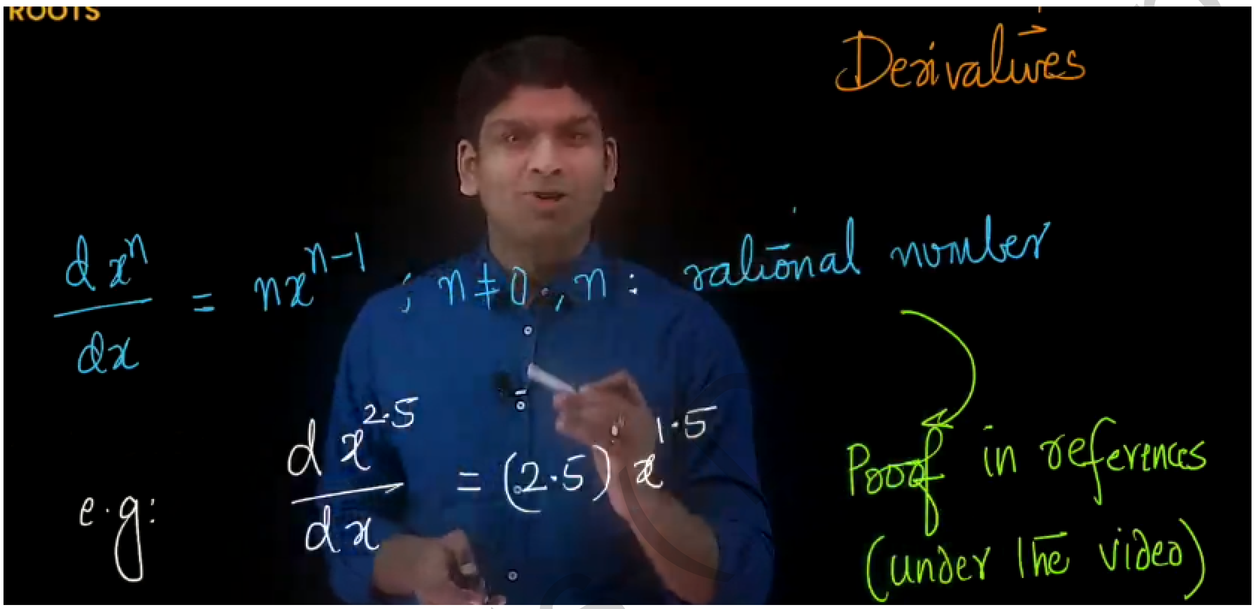


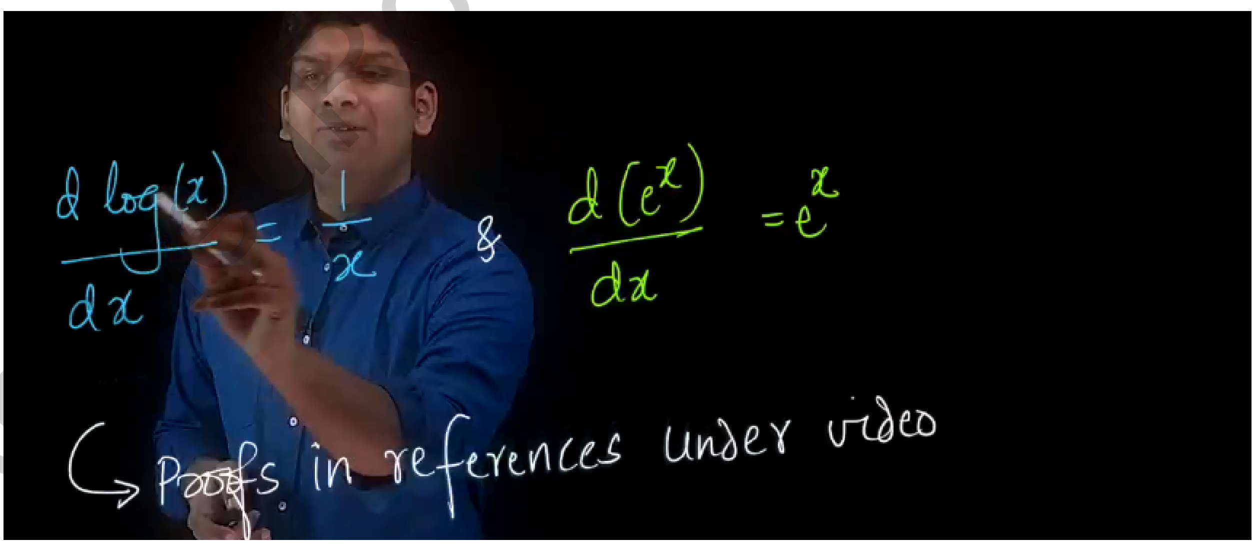

- The derivative of the function can be calculated as shown above .
- For computing the derivative at a specific point we can substitute the corresponding x value.

# 3.10 Derivative of Common functions

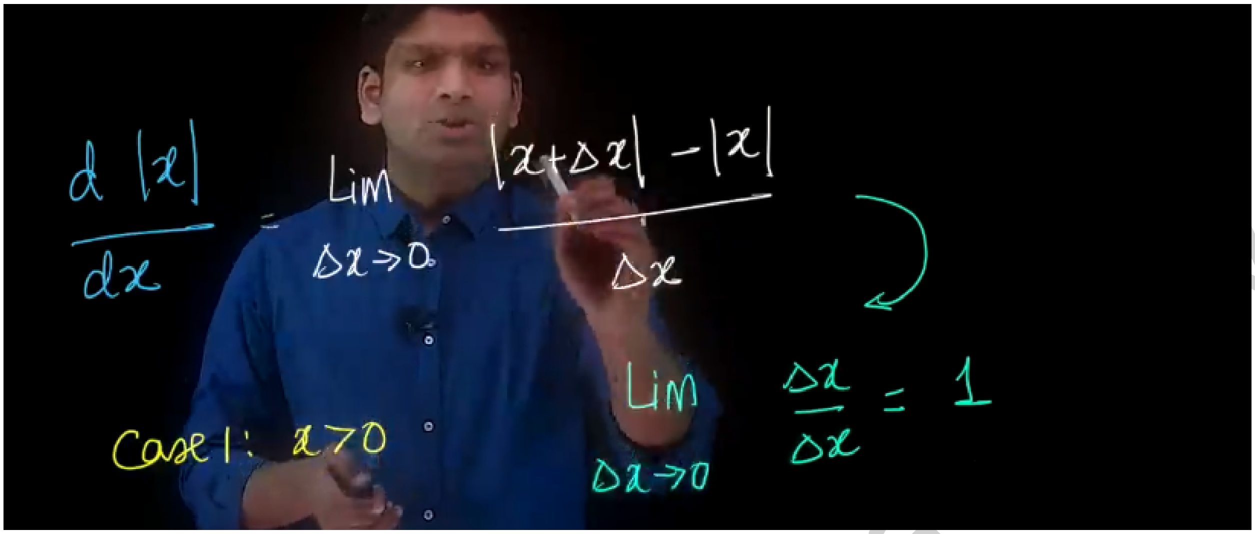
At timestamp  0.10 in video



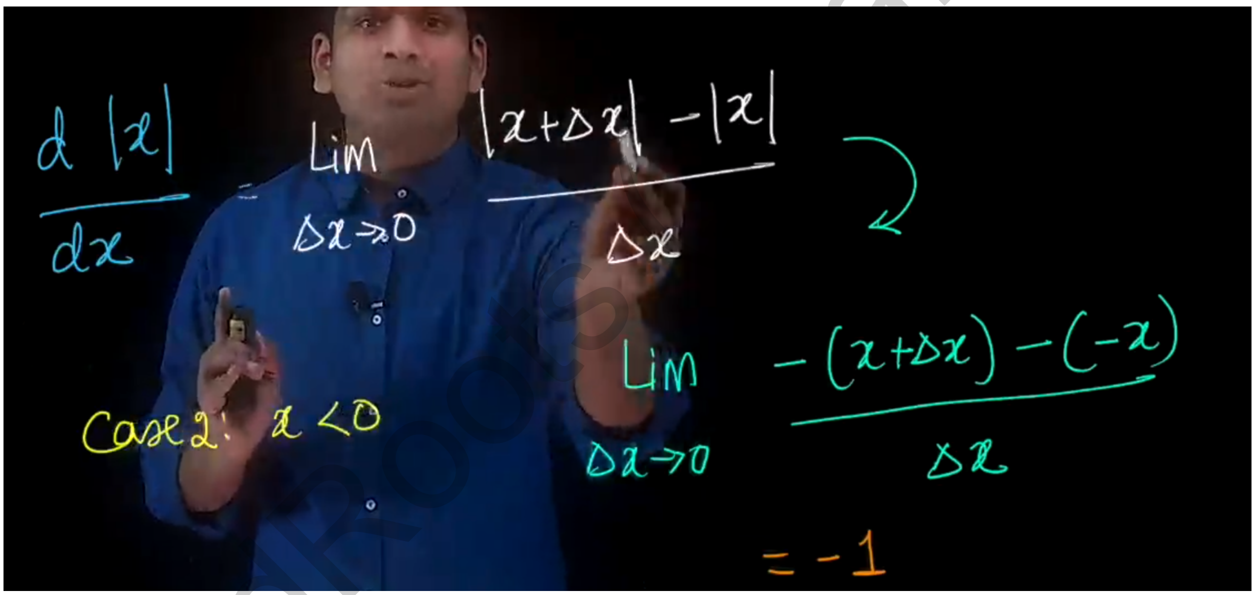

- Let's see the derivatives of most commonly used functions in Machine learning.The function shown above is a polynomial function and the derivative can be obtained as shown.



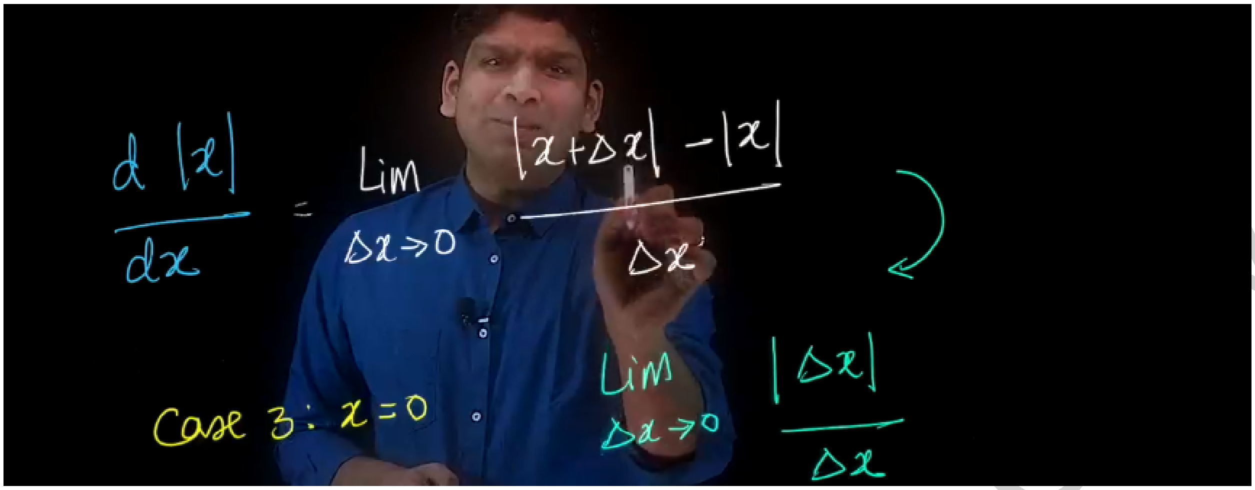

- Derivatives of log x and e^x are as shown above,using limits you can solve these easily.

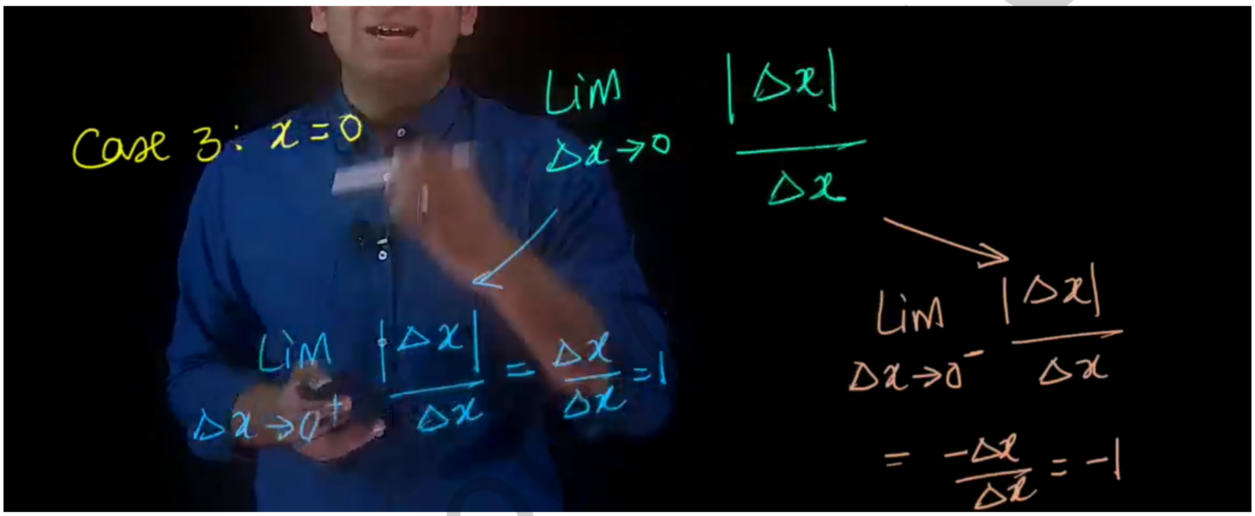



- The derivative of abs(x) can be calculated as shown using limits.
- Incase of absolute value function if x>0 the derivative is +1




- When we have  negative x value x<0 then the derivative is -1.

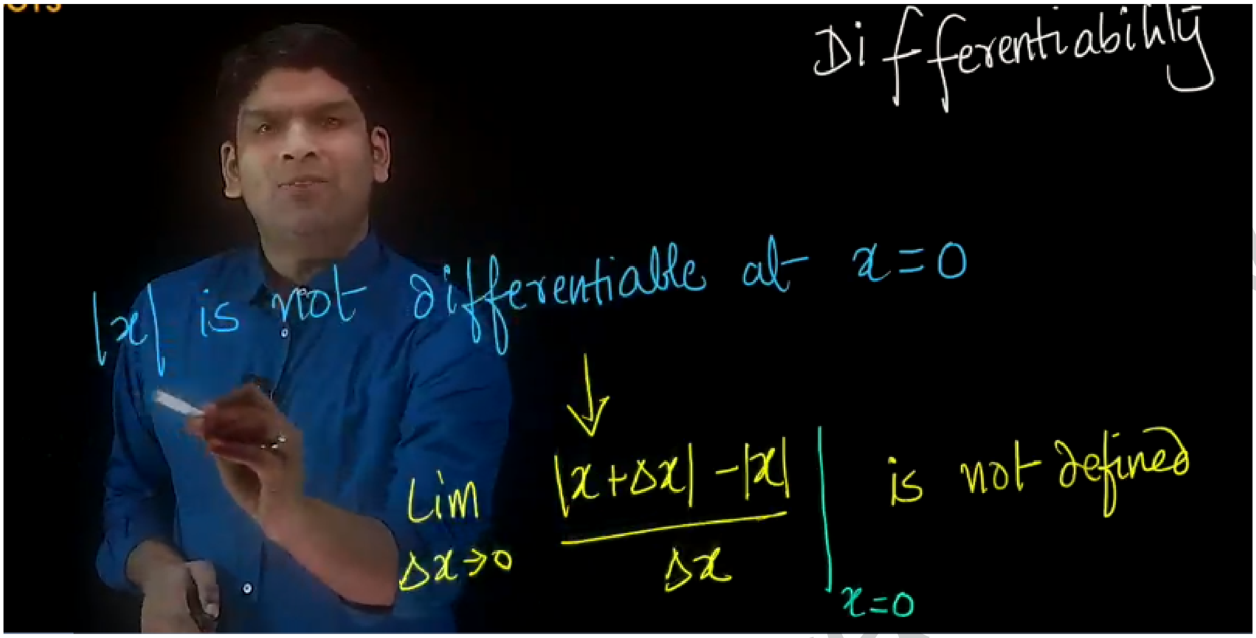



- When  x value x=0 then the derivative is not defined because the left handed limit is not equal to right handed limit.

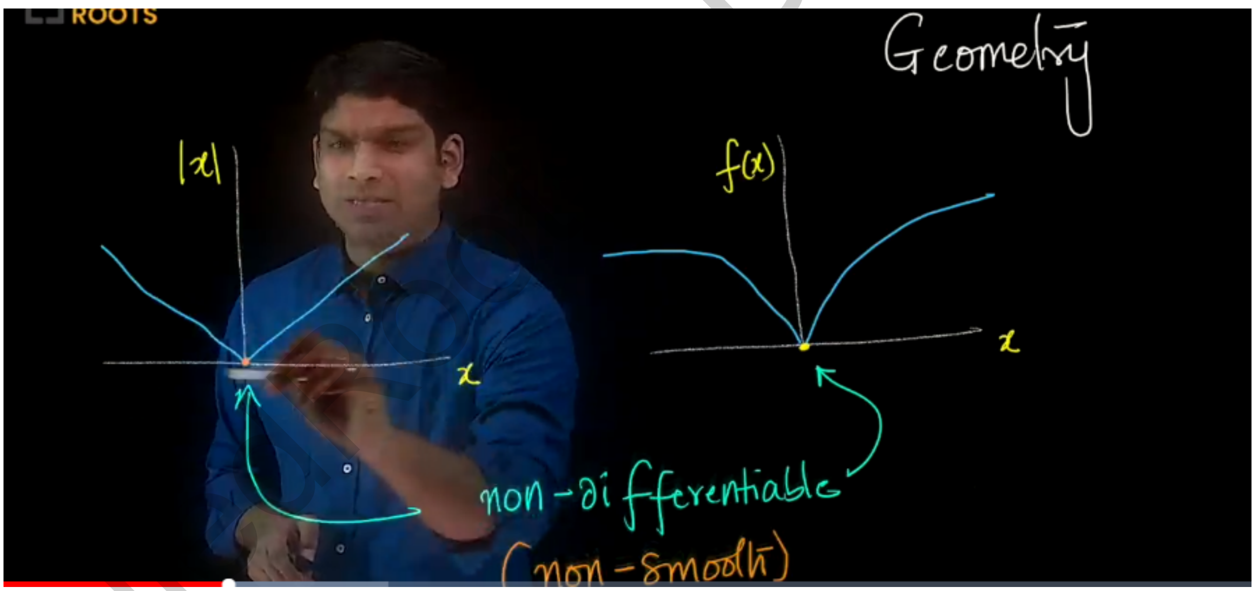
# 3.11 Differentiability of functions

At time stamp 0.10



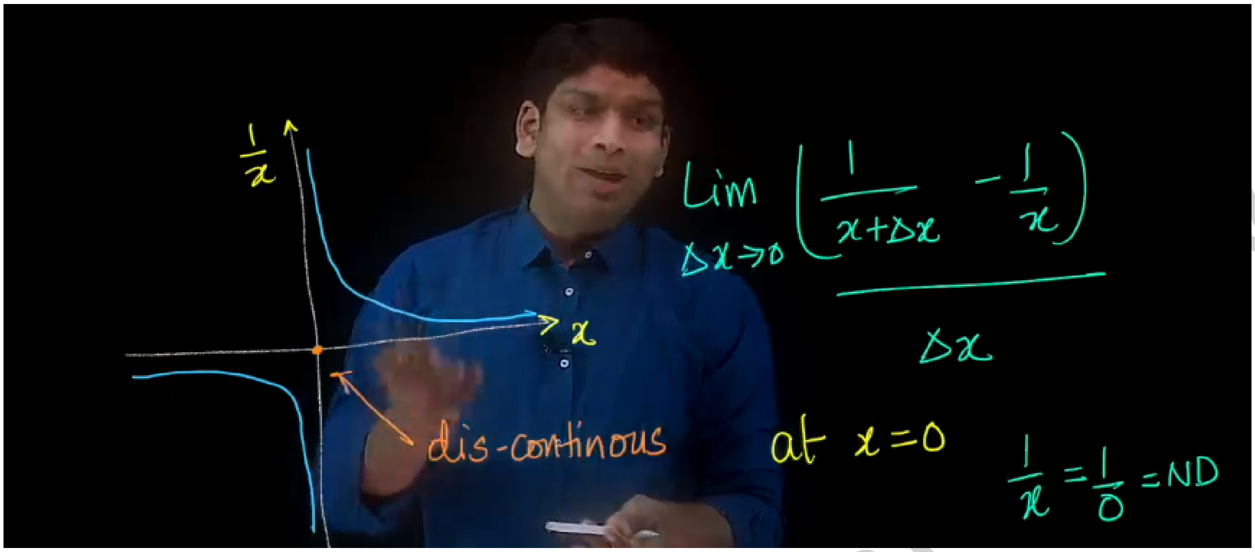

- Absolute value function is not differentiable at x=0 but it might be differentiable at other points

At timestamp 1.59



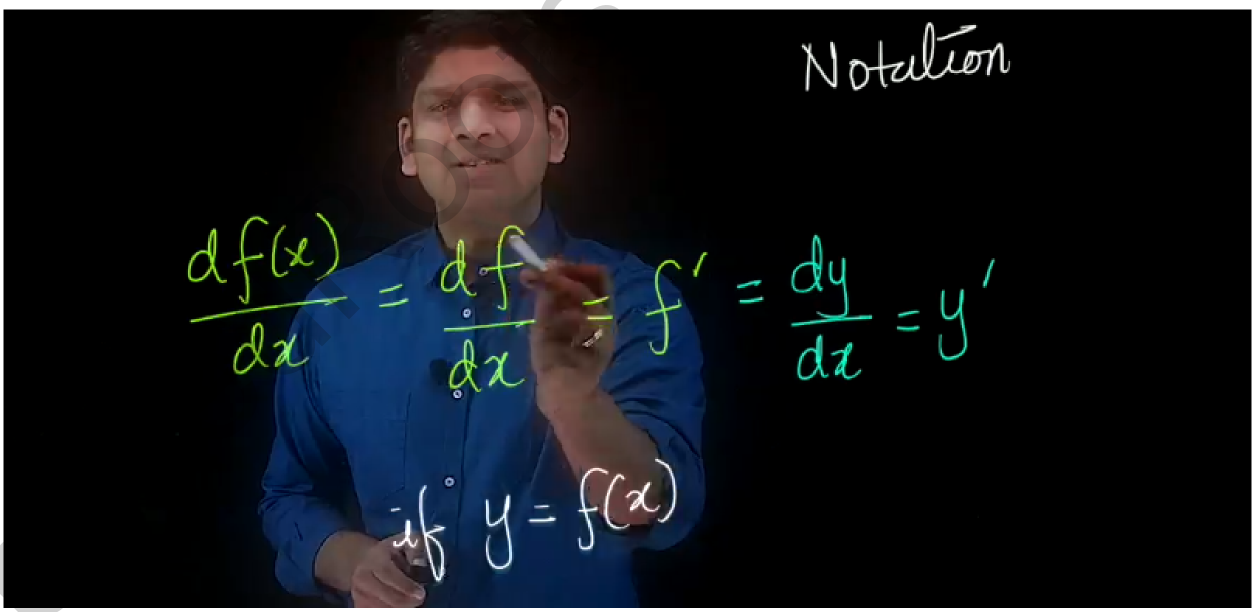

- You can see for above functions are differentiable at other parts of the curve but at x=0 they are not.Wherever the graph or curve is non-smooth we tend to face the problem of non -differentiability.

At timestamp 3.02

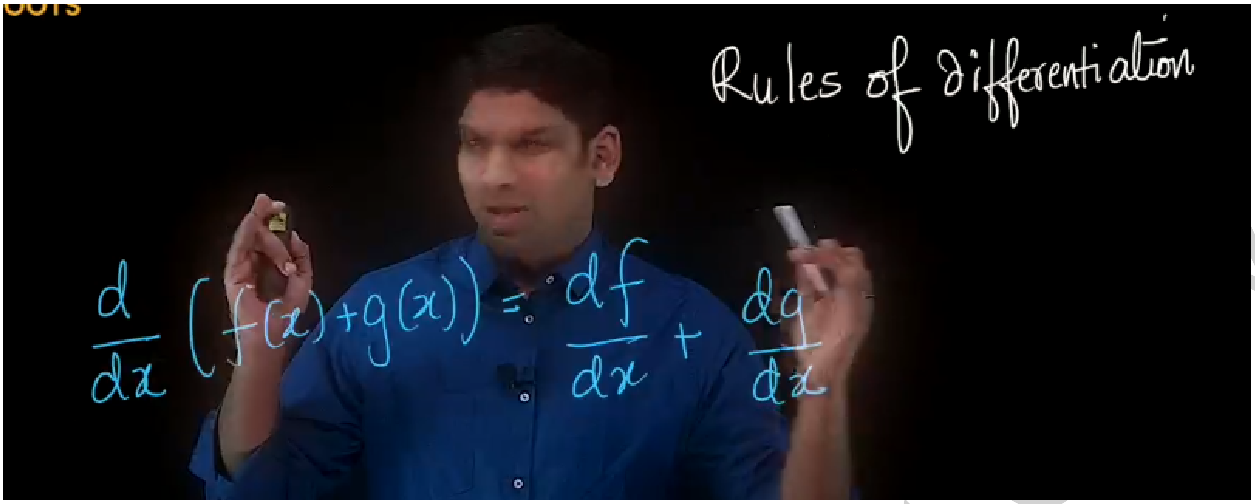


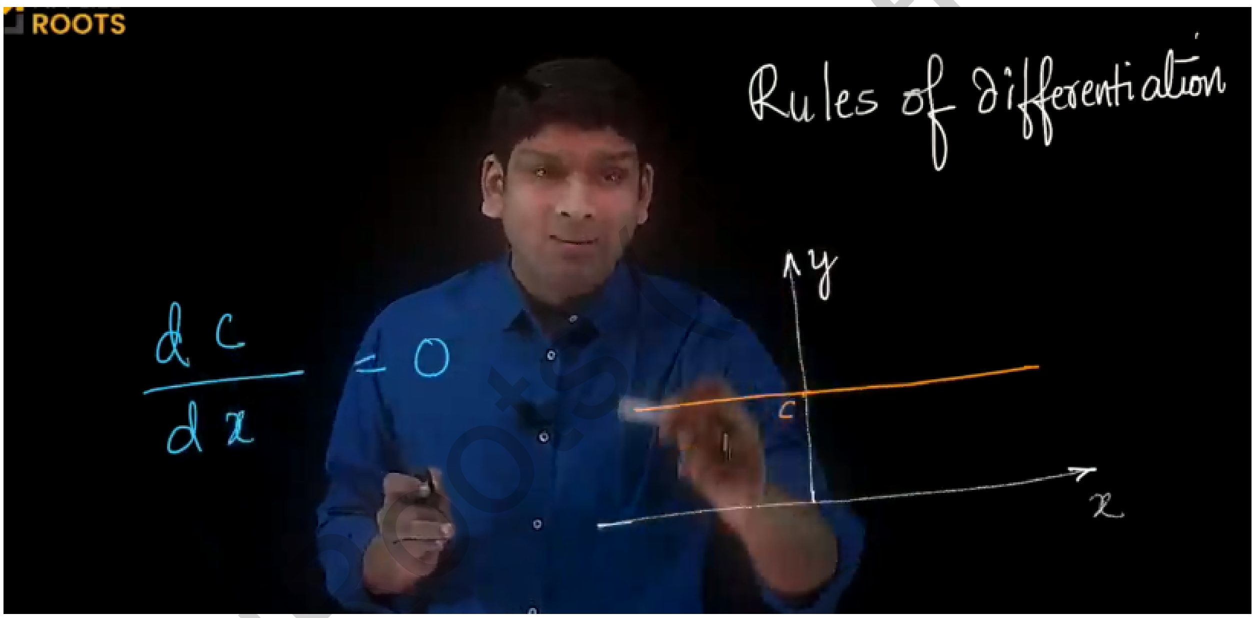

- We face the problem of non-differentiability when the function is discontinuous,because the positive and negative limits will not be the same .

# 3.12 Rules of differentiation

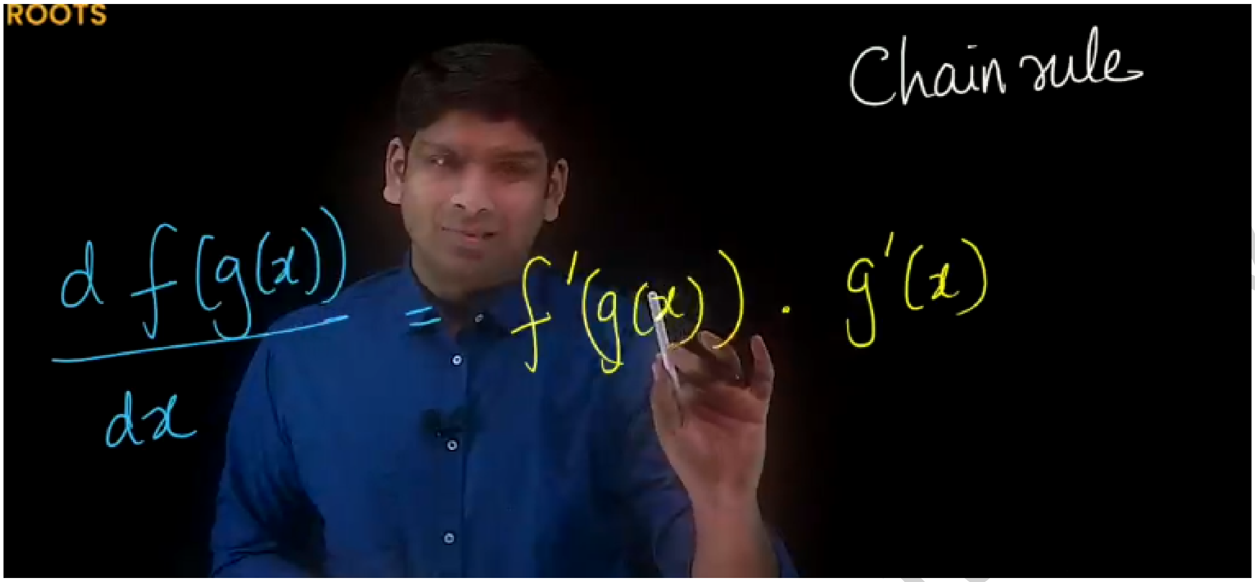


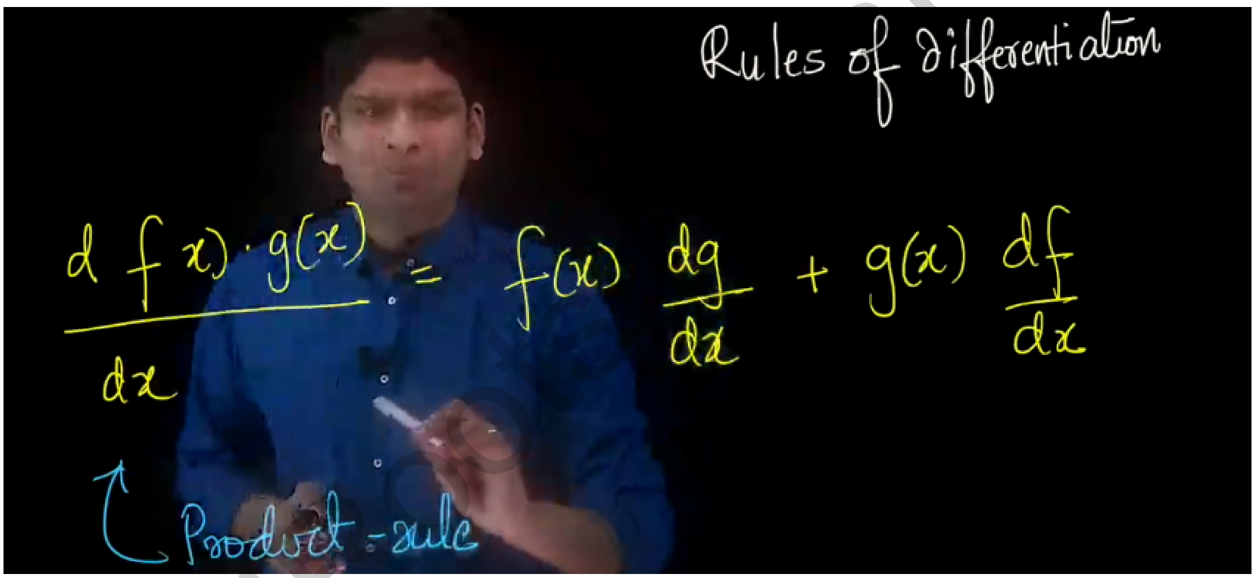

- The derivative of y=f(x) with respect to x can be represented in any of the above ways.
- We have rules of differentiation which help us when we try to differentiate large equations.These rules are listed below.



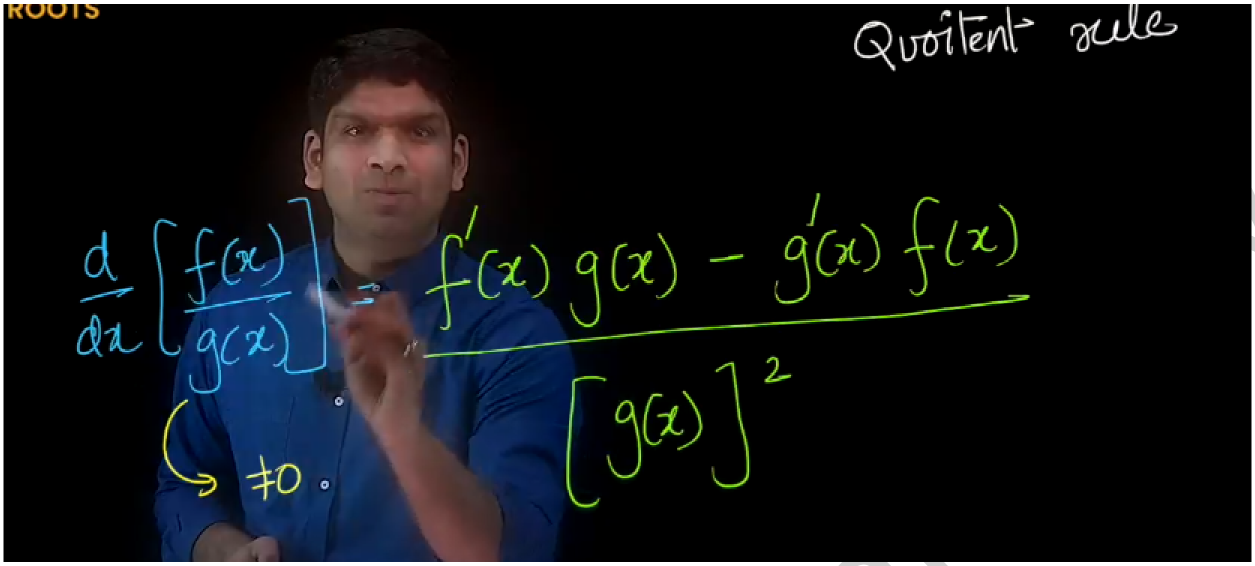

- The derivative of constant is with respect to x is 0.




- When we have to find the derivative of a product of two functions we use product rule.

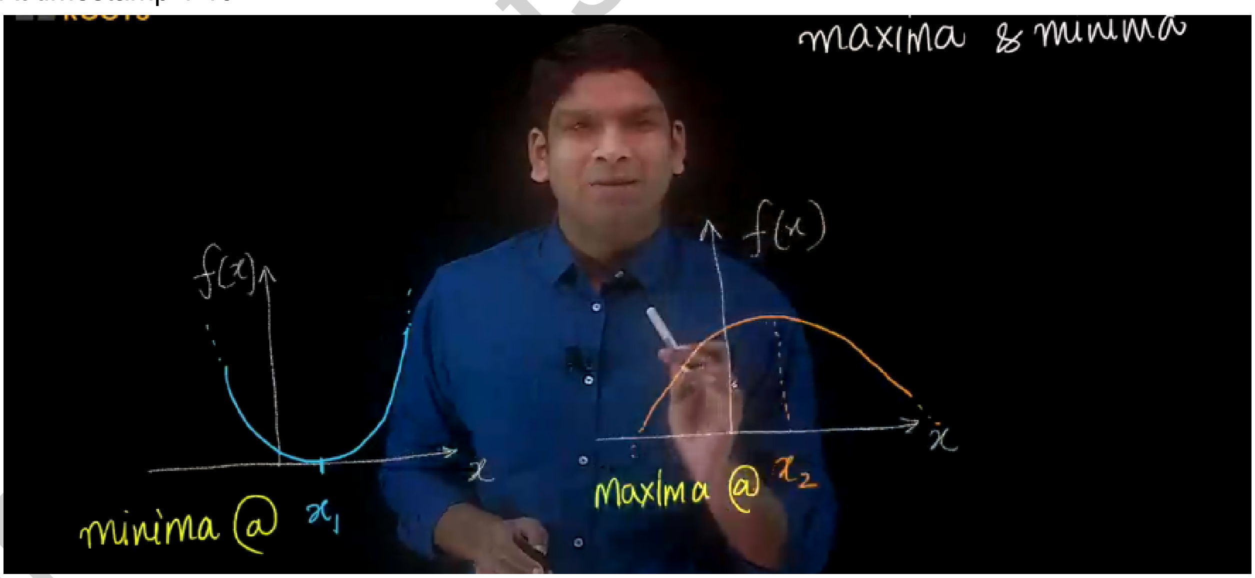
At timestamp 13.07




- Quotient rule is used when we have to find the derivative of two functions where g(x) is dividing f(x).Here g(x ) cannot be 0.

# 3.13 Maxima and Minima

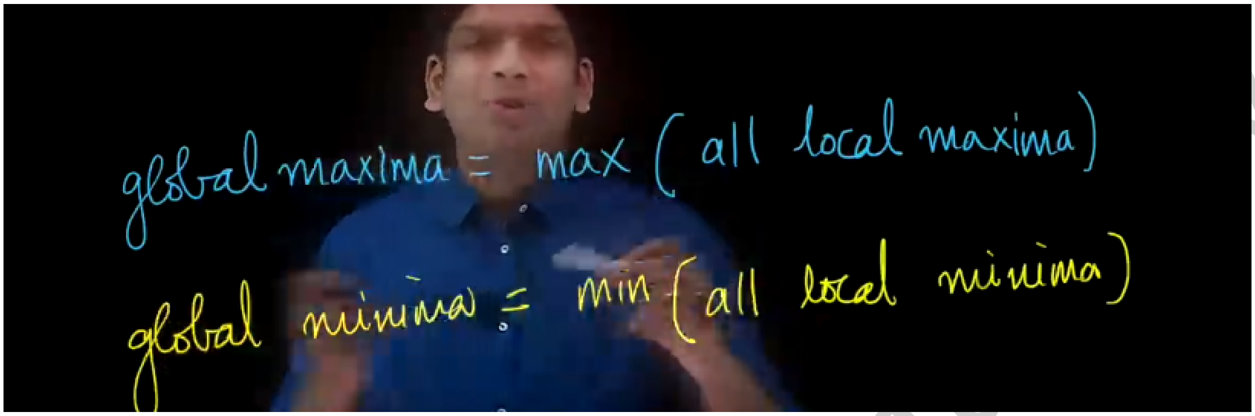
At timestamp 1.19

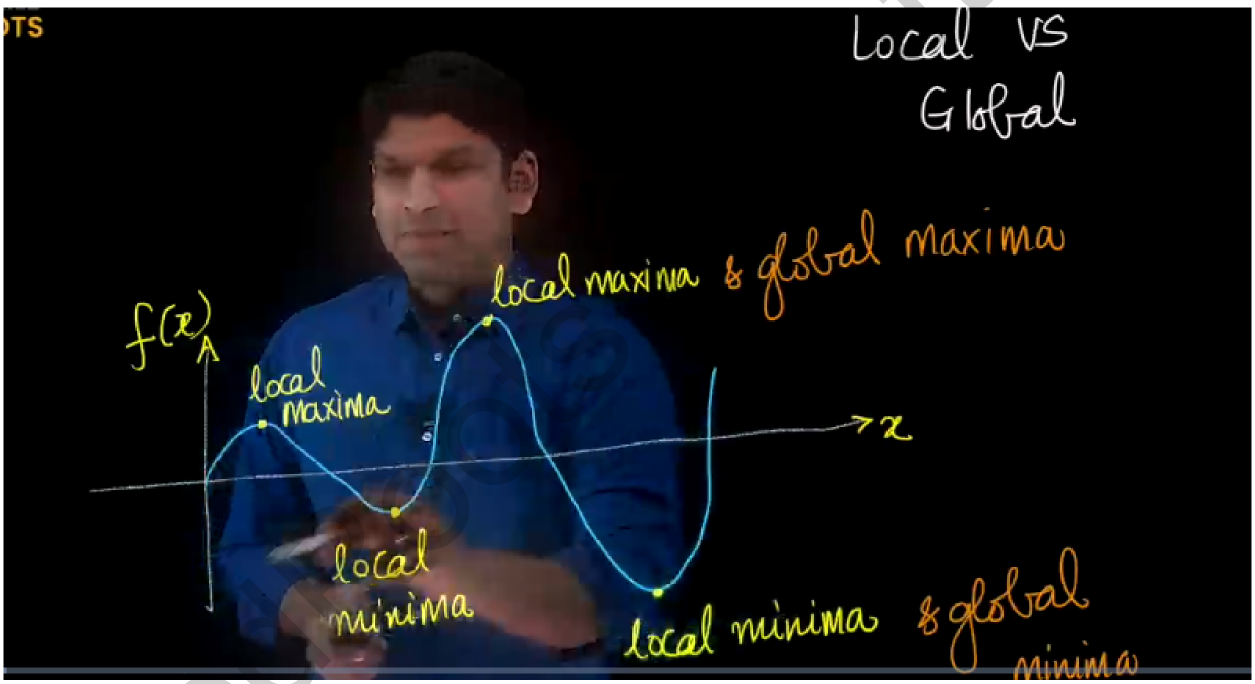



- All the calculus we have studied till now is primarily to study maxima and minima.In Ml most of the time we try to minimize or maximize certain functions or mathematical expressions.
- In the functions shown above the first function doesn't have maxima but has minima at x=x1 and the other doesn't have minima but has maxima at x=x2.

- Function can have maxima and minima some functions can have multiple maxima and minima some functions may not have any minima or maxima.

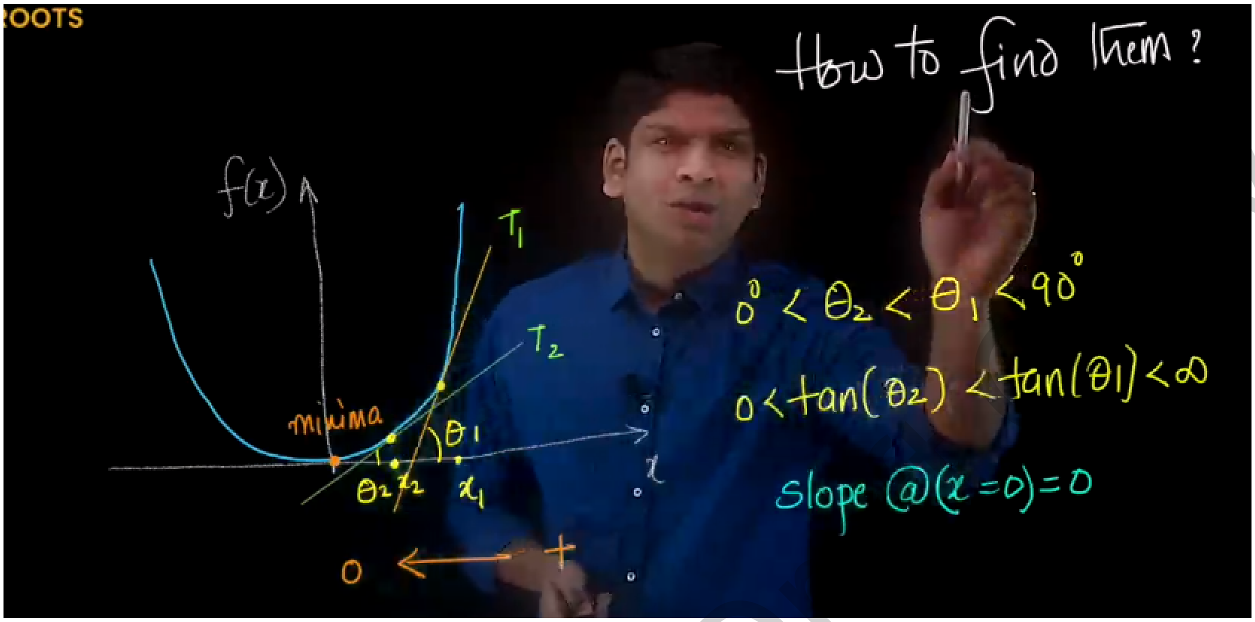
At timestamp 2.19



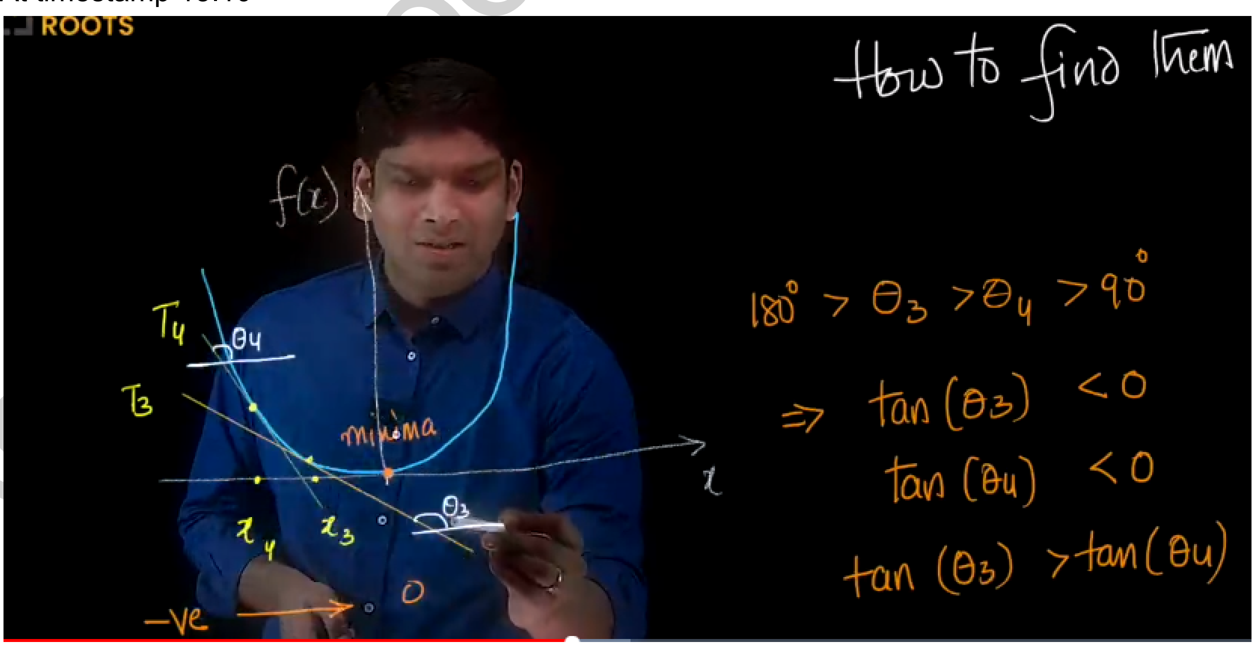

- Whenever we have local maxima or minima we use a certain term called optima and it can be either maxima or minima.
- We can have local optima which can be either local maxima or minima.Similarly we have global optima which can be global maxima or global minima.

At timestamp 6.09



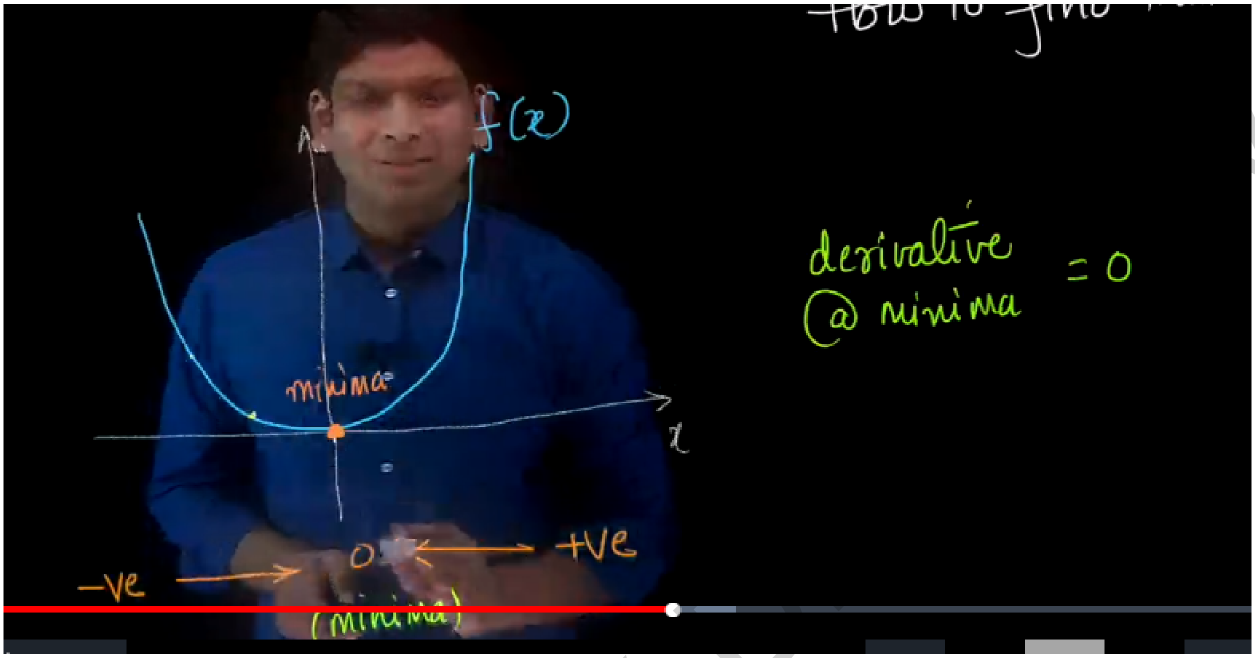

- Consider a curve as shown above y=x^2 and two tangents T1,T2 at points x1,x2 on curve and making angels **ϴ**1 and **ϴ**2 with x axis respectively.Let's see how to find minima mathematically.
- We can clearly see that the slope of tangent T2 is less than slope of T1.Slope is nothing but the derivative of the function at x=x1,x2
- The derivative of x1 and x2 are positive but derivative of x1 >derivative of x2.
- At x=0 the tangent is nothing but x-axis,so the slope of tangent at x=0 is 0.

At timestamp 13.10



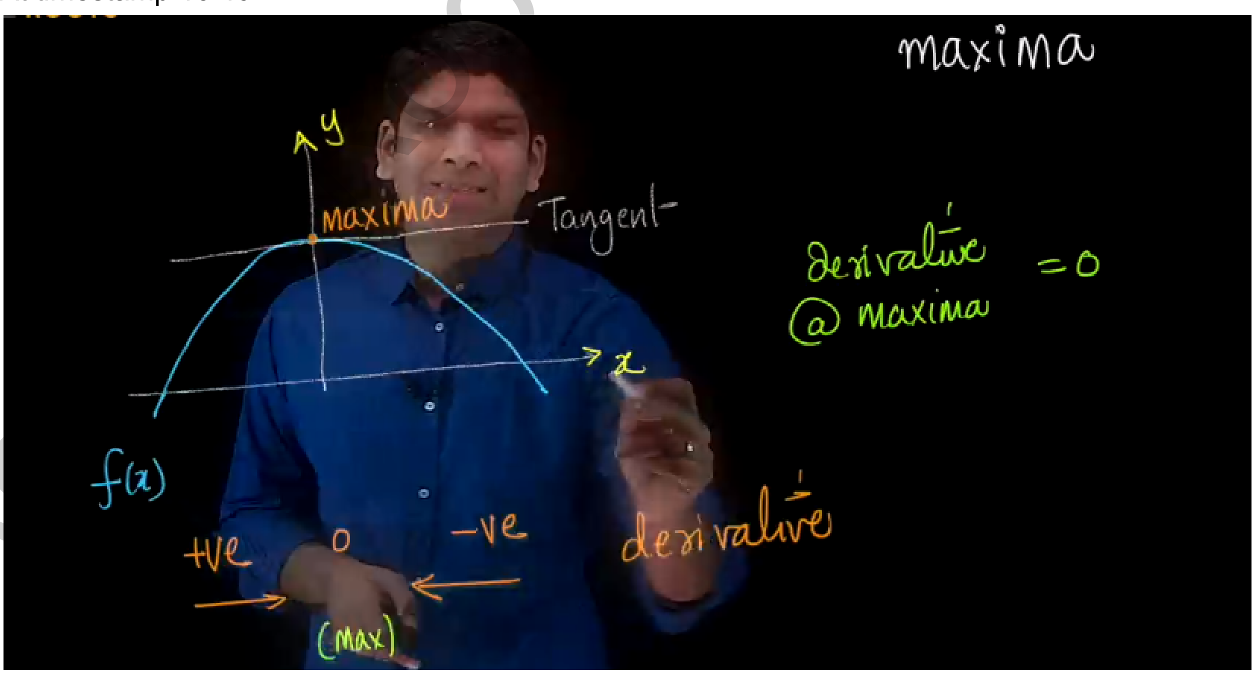
- You can observe a similar trend on the left side of the curve as well as shown above.

At timestamp 13.58



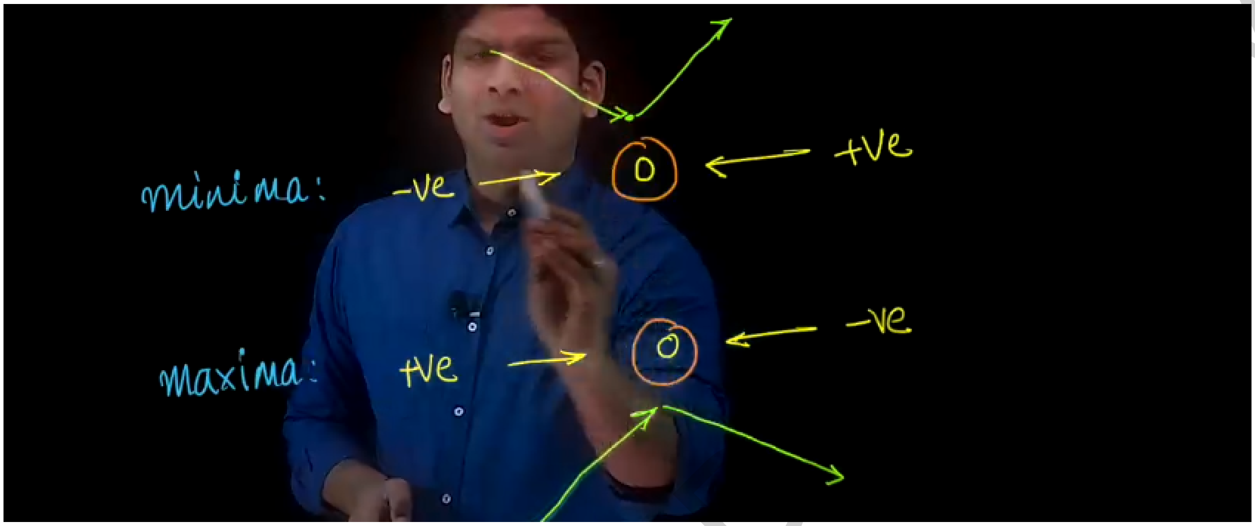

- We can observe that on one side of minima the derivatives are all positive and on other side derivatives are all negative at minima the derivative is zero
- We find the minima at a point  where the derivative is zero and function should be increasing on right side of minima and decreasing on left side of the minima

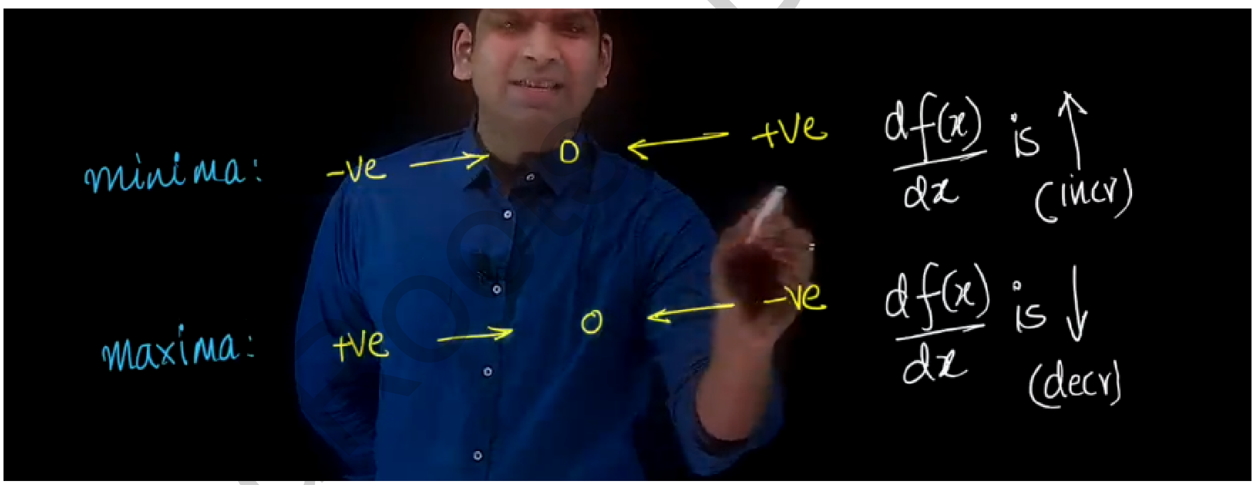
At timestamp 16.40

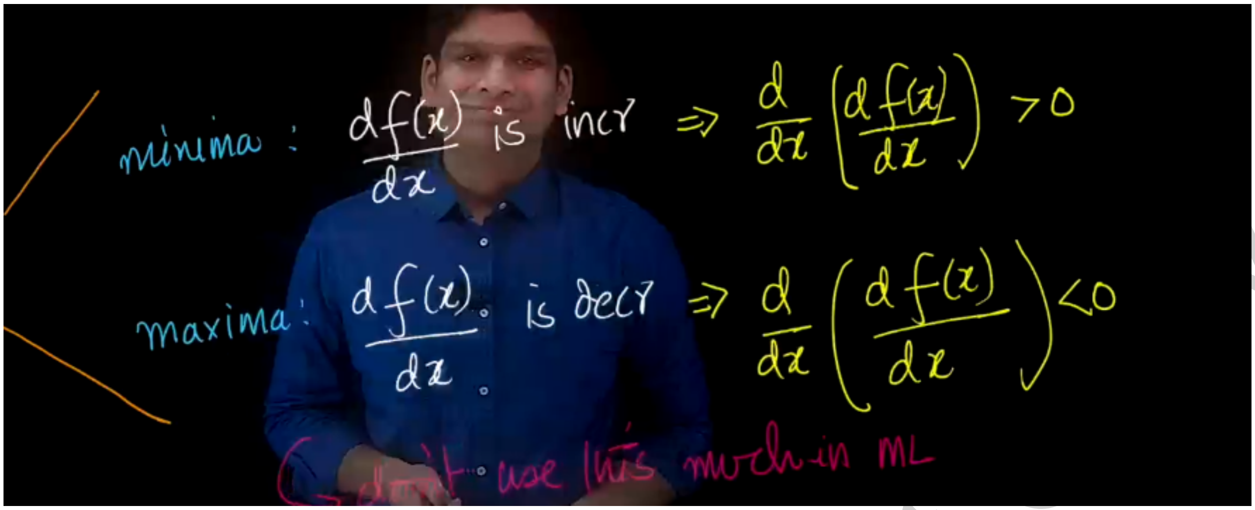


- Just like the derivative at minima is zero the derivative at maxima is also 0.You can see that the angle made by tangent with x-axis is 0 so the derivative or slope is 0.
- We find the maxima at a point  where the derivative is zero and the function should be decreasing on the right side of the maxima and increasing on the left side of the maxima

At timestamp 18.54



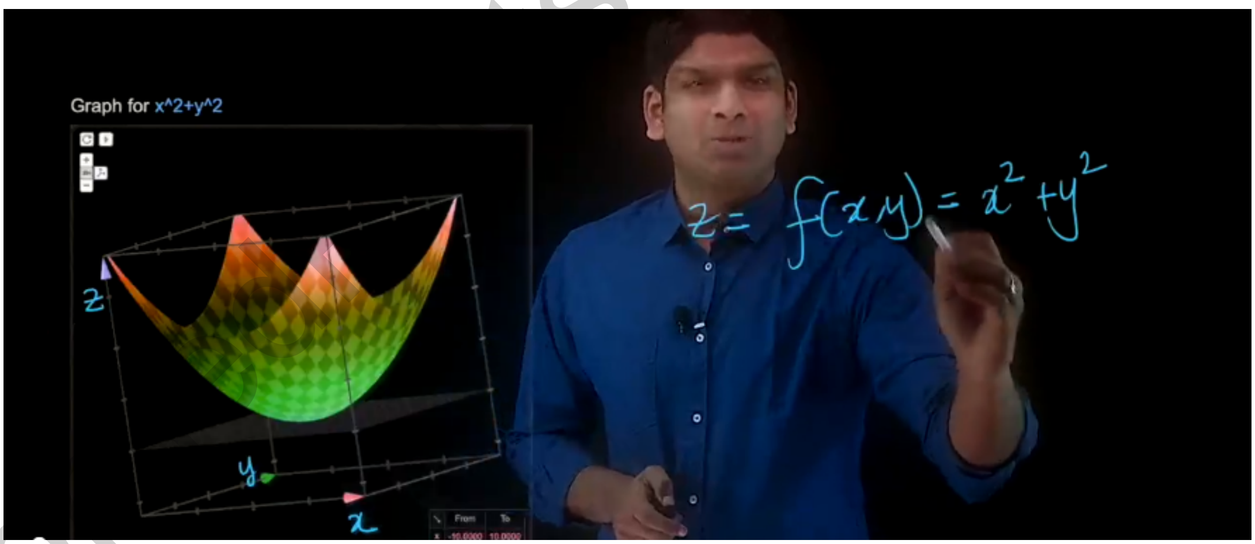






- When the derivative is zero we can analyse whether it is maxima or minima as shown above.

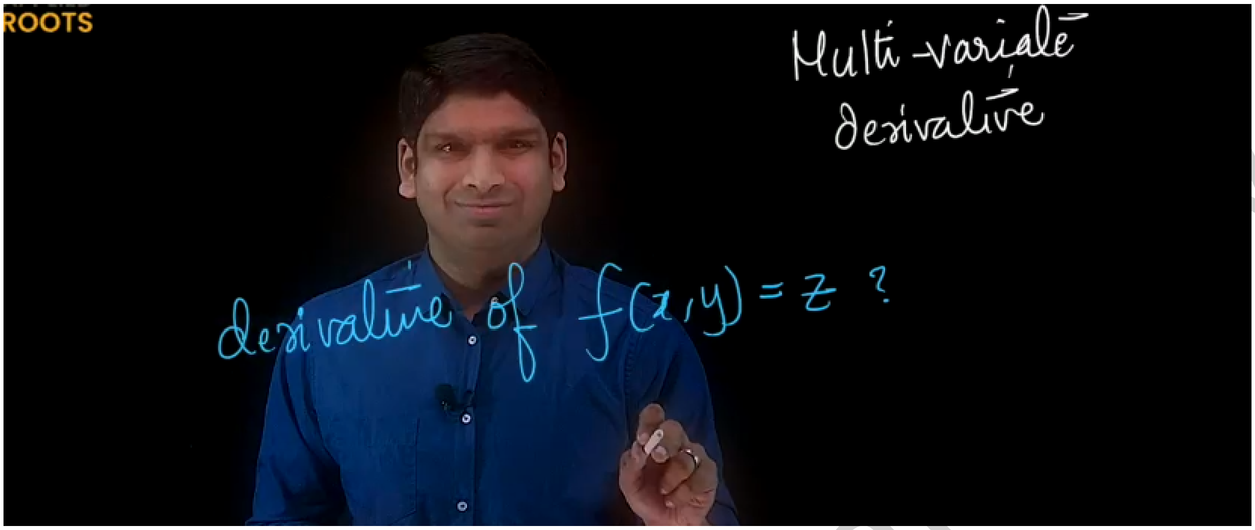
# 3.15 Partial derivatives & Del

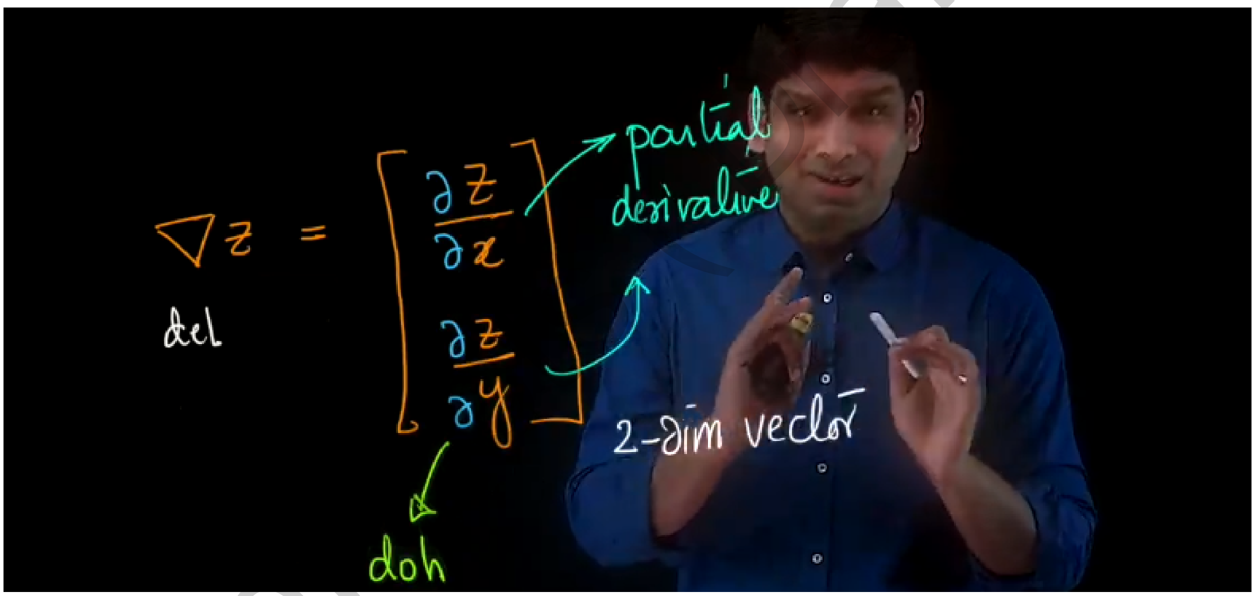
At timestamp 2.4

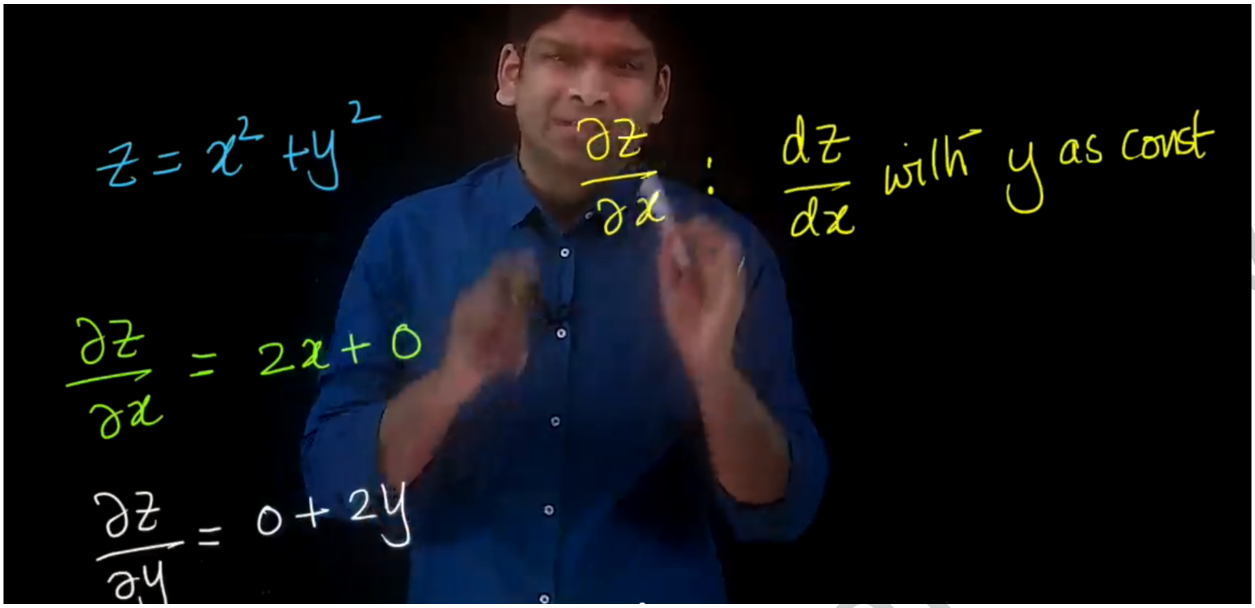


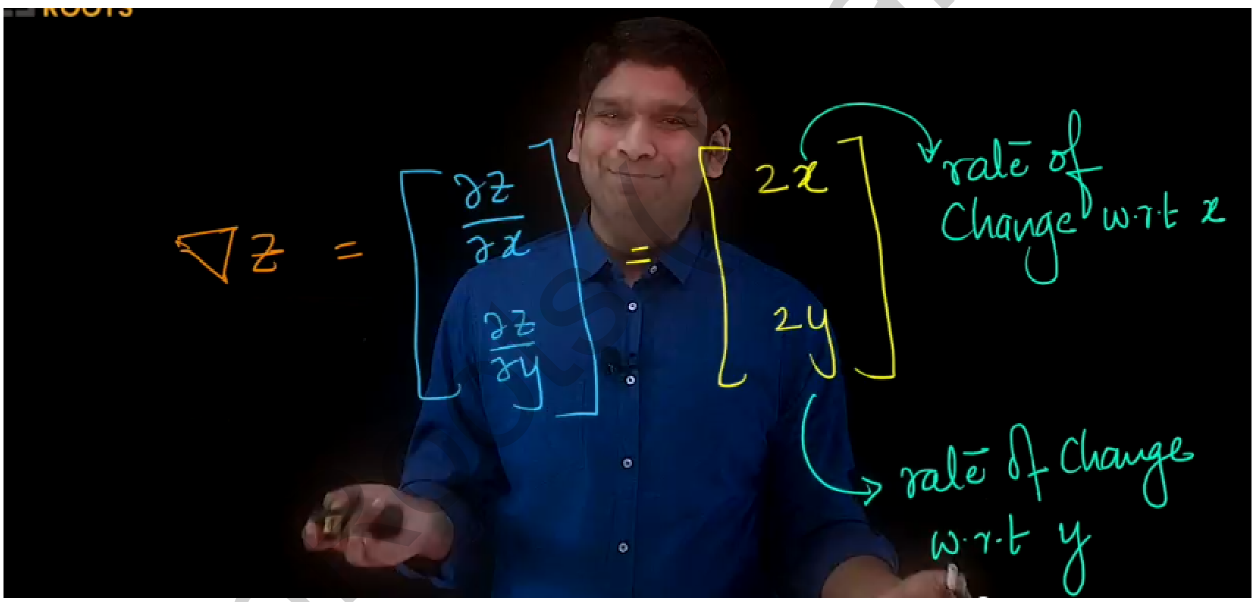

- When we have multivariate functions as shown above we use partial derivatives .
- The above function is quadratic,we have three variables x,y,z so we can plot the function in 3D.

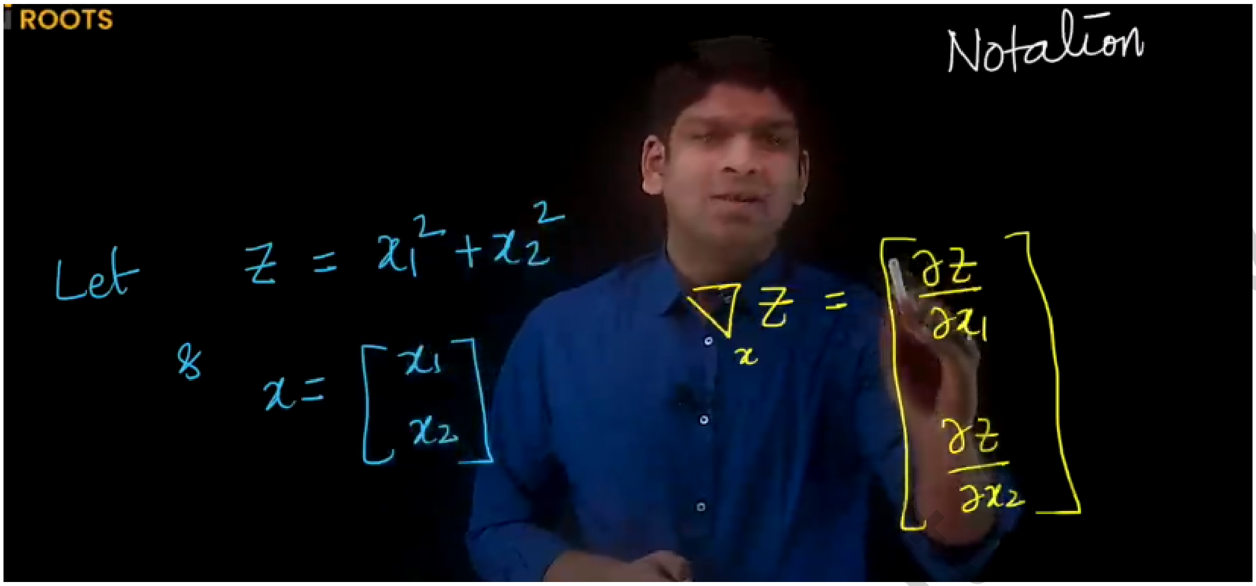
At timestamp 3.51




$z = x^2 + y^2$
$\frac{\partial z}{\partial x}$ : $\frac{dz}{dx}$ with $y$ as const
$\frac{\partial z}{\partial x} = 2x + 0$
$\frac{\partial z}{\partial y} = 0 + 2y$

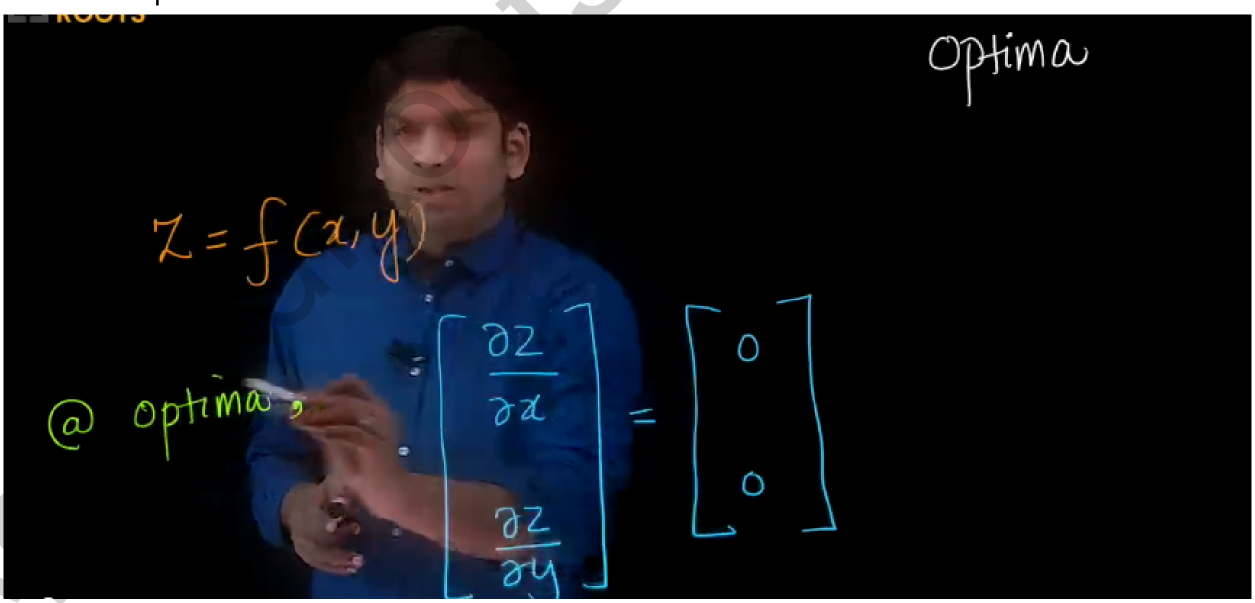

$\nabla z = \begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$
rate of Change w.r.t $x$
rate of Change w.r.t $y$




- We have seen how to find derivatives when we have one variable,here we have more than one variable so we use a concept called partial derivative as shown above.
- We can extend the same concept even for n dimensions.

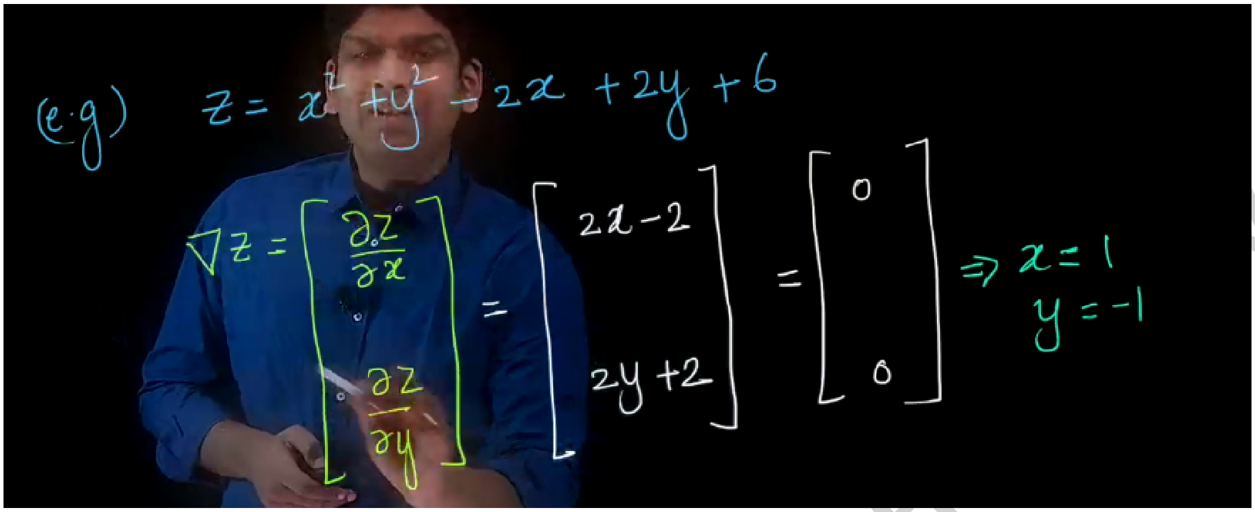
# 3.16 Optima using Partial derivatives

At timestamp 0.53

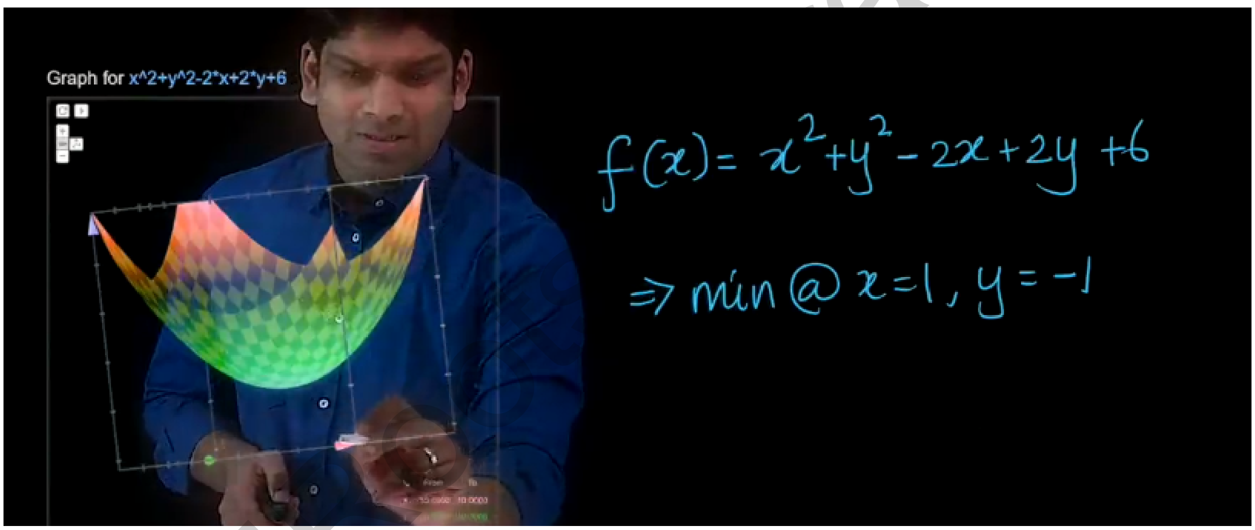



- At optima each of these partial derivatives will be zero(since we vector of partial derivatives each component must be equal to 0)

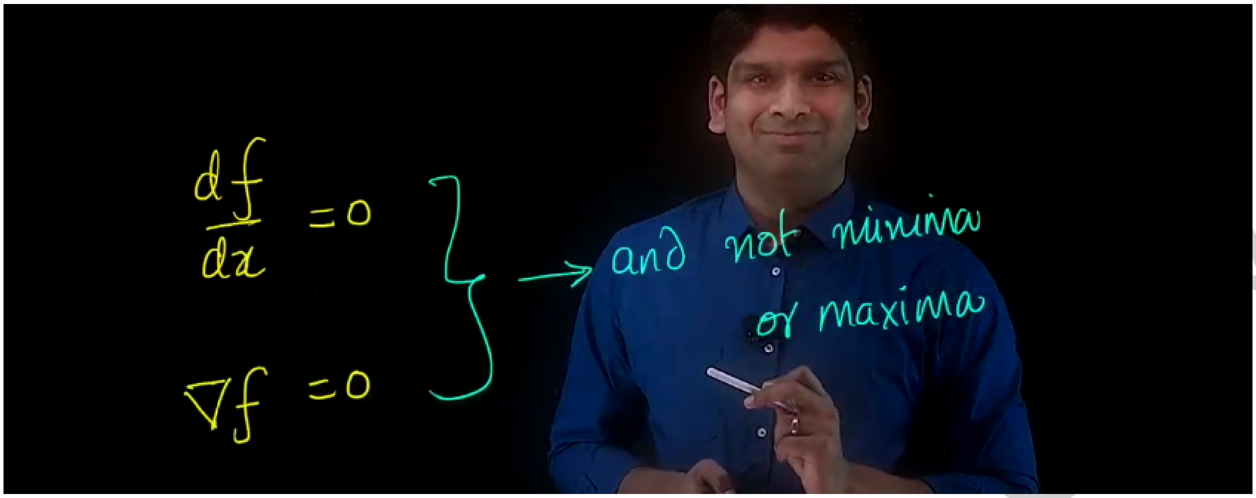
At timestamp 2.01




- Above is an example of calculating partial derivatives and by equating them to 0 we got optima at x=1 and y=-1.The optima can be either minima or maxima.



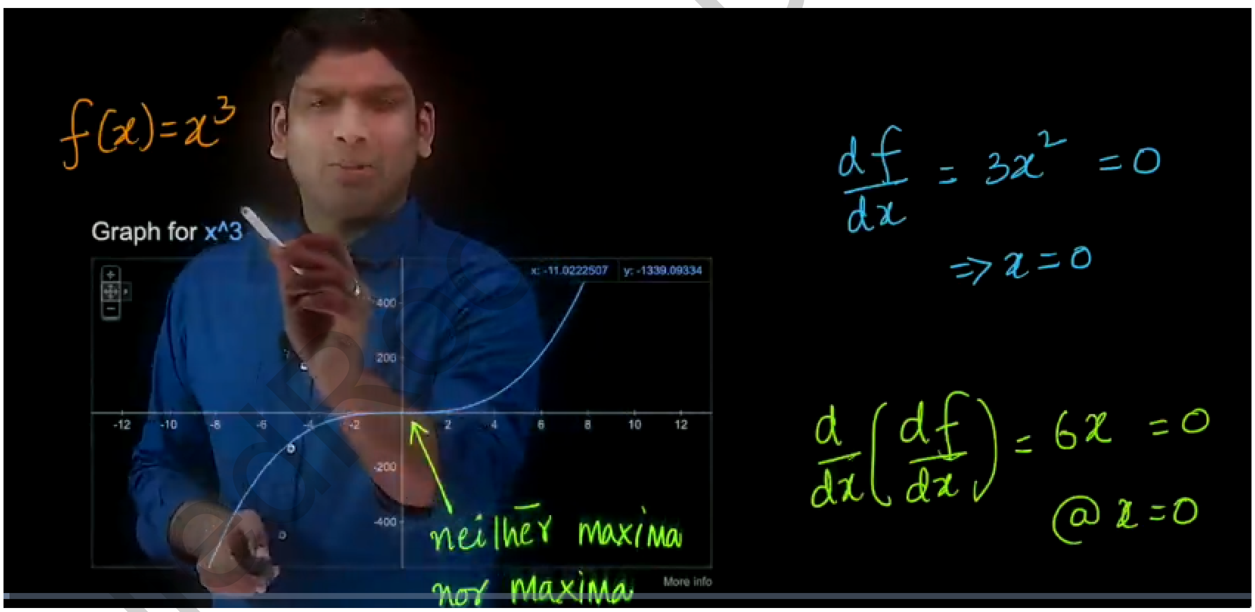

- If we plot the function we can clearly see that the optima that we arrived at using partial derivatives at x,y=(1,-1) is minima.




- There is a special case where the derivative of the function(can be regular derivative or vector of partial derivatives) is 0 but it doesn't have either minima or maxima.

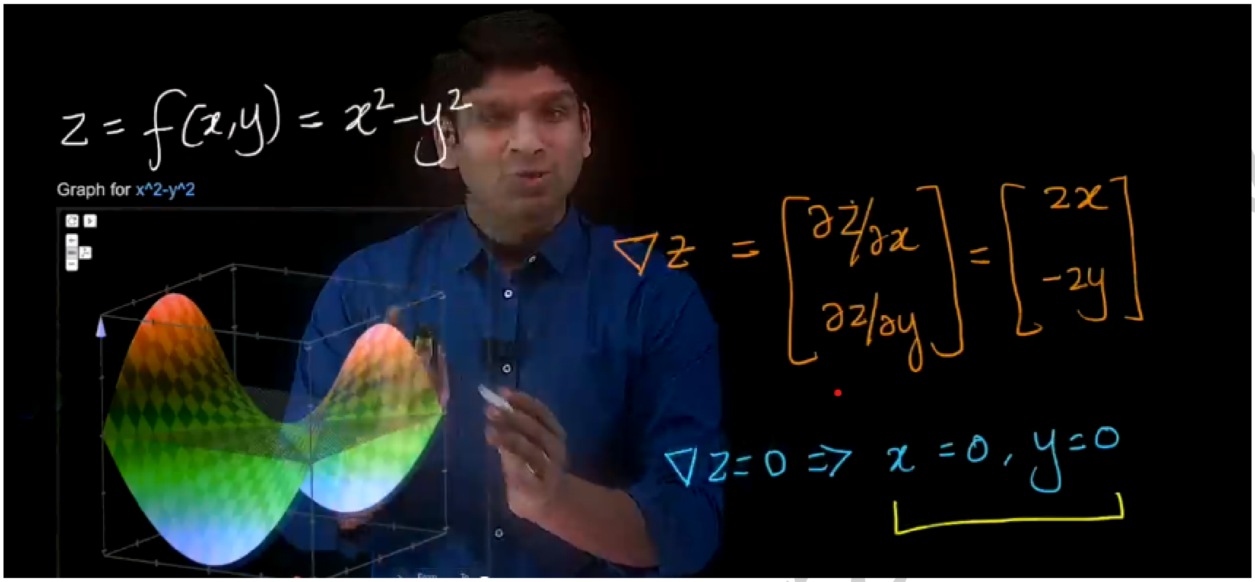
# 3.17 Saddle point

At timestamp 1.15

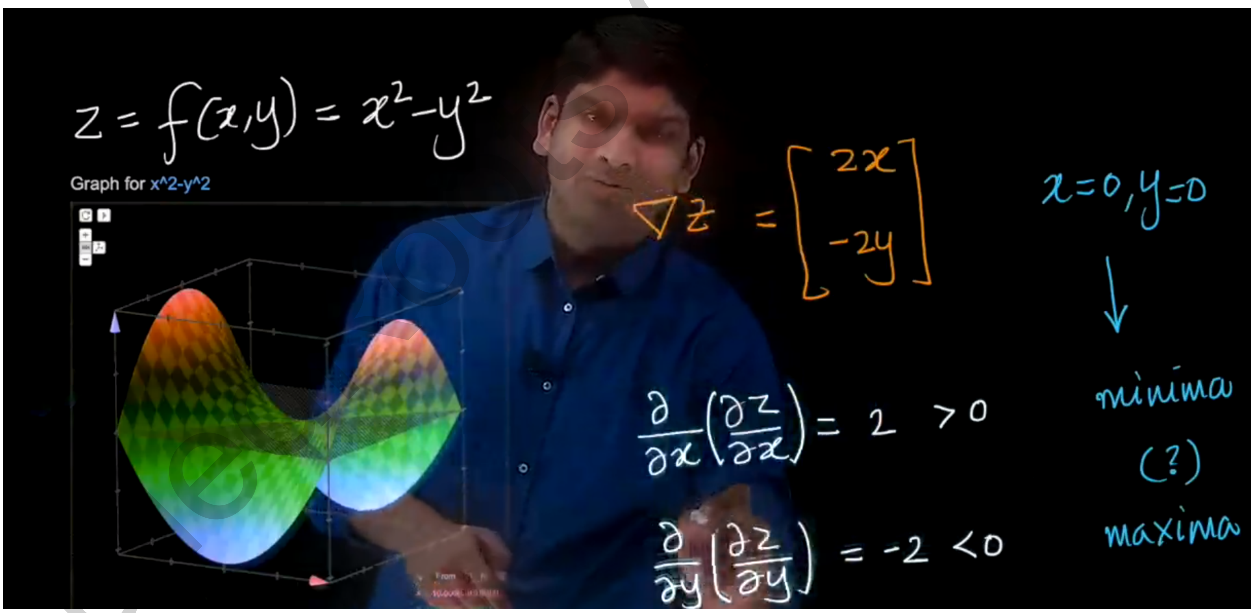


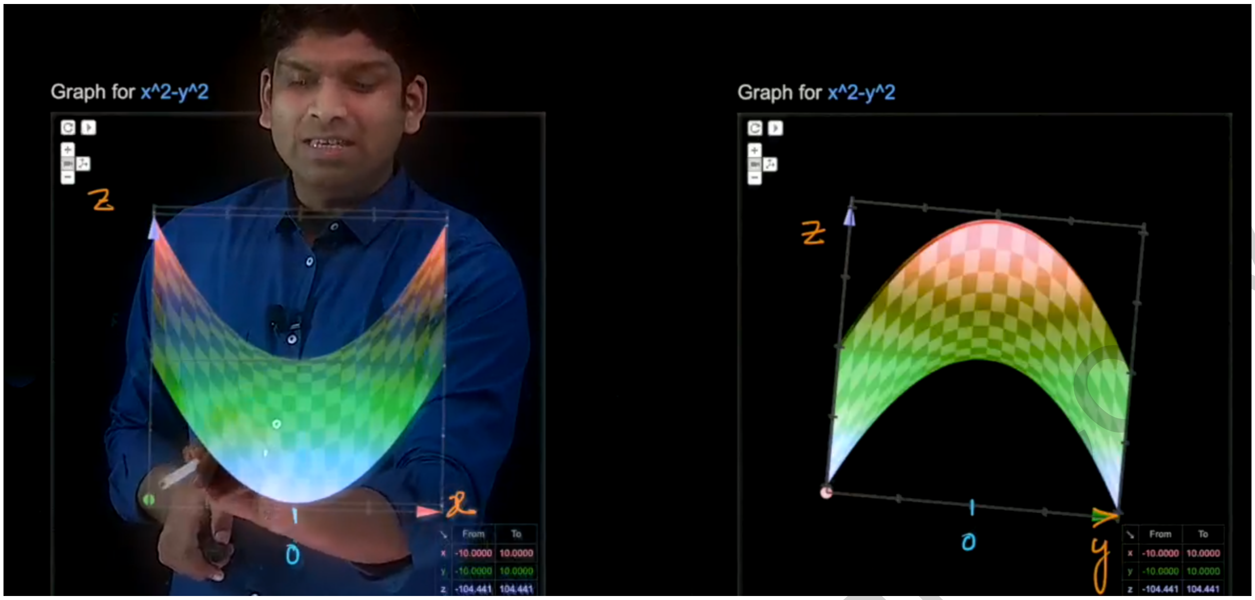

- For the above function you can clearly see that the derivative of the function at x=0 is zero but it's either minima or maxima .

At timestamp 3.45




- We we have more than one variable in our function as shown above it's slightly tricky ,you can find the partial derivatives and can say that at x,y=(0,0) we have optima.But there is no optima



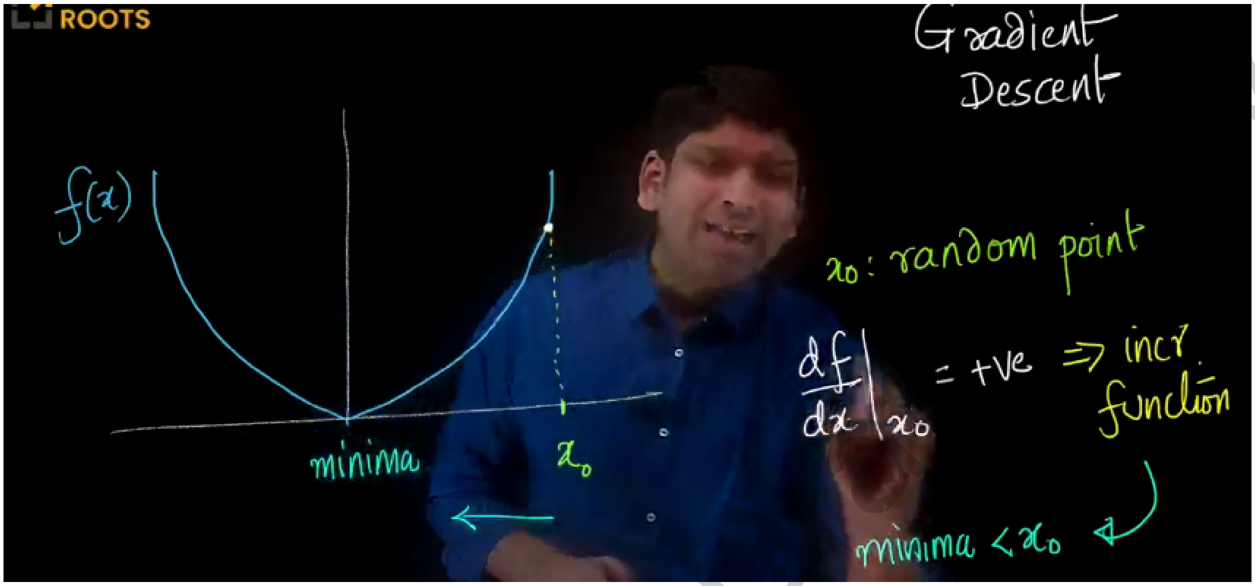




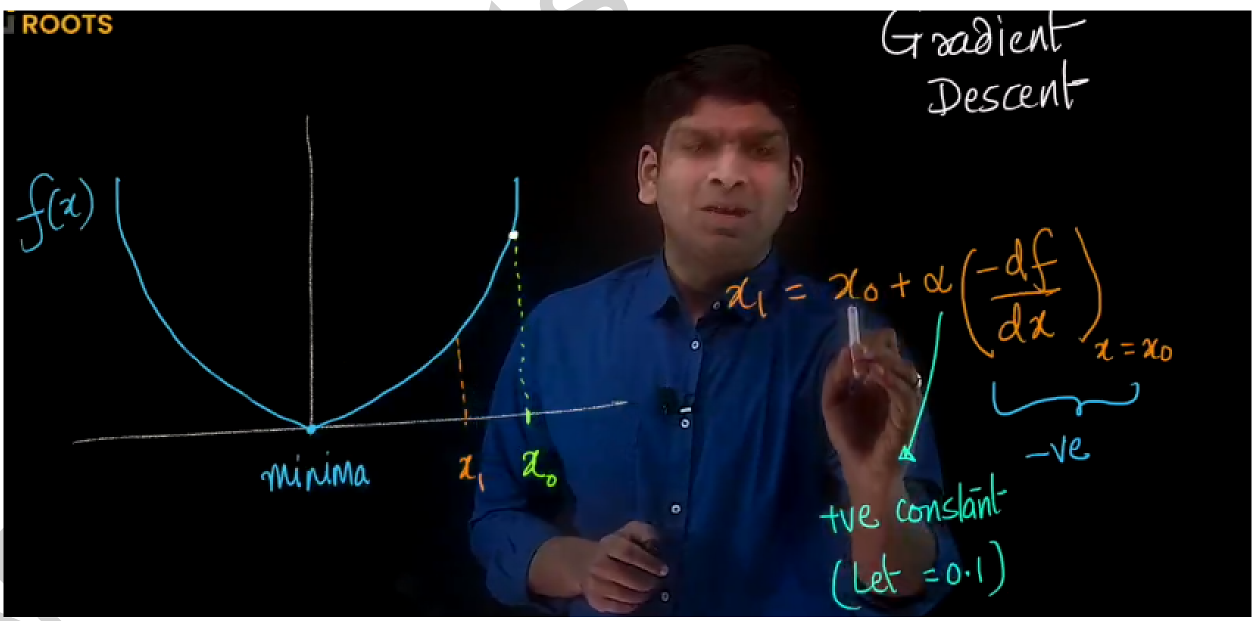
- When we calculate second order partial derivatives you can observe that x is increasing(positive) and y is decreasing(negative).From x perspective the point is minima and from y's perspective it is maxima.
- Such points are called saddle points and we cannot say that it is either maxima or minima.

# 3.18 Gradient Descent

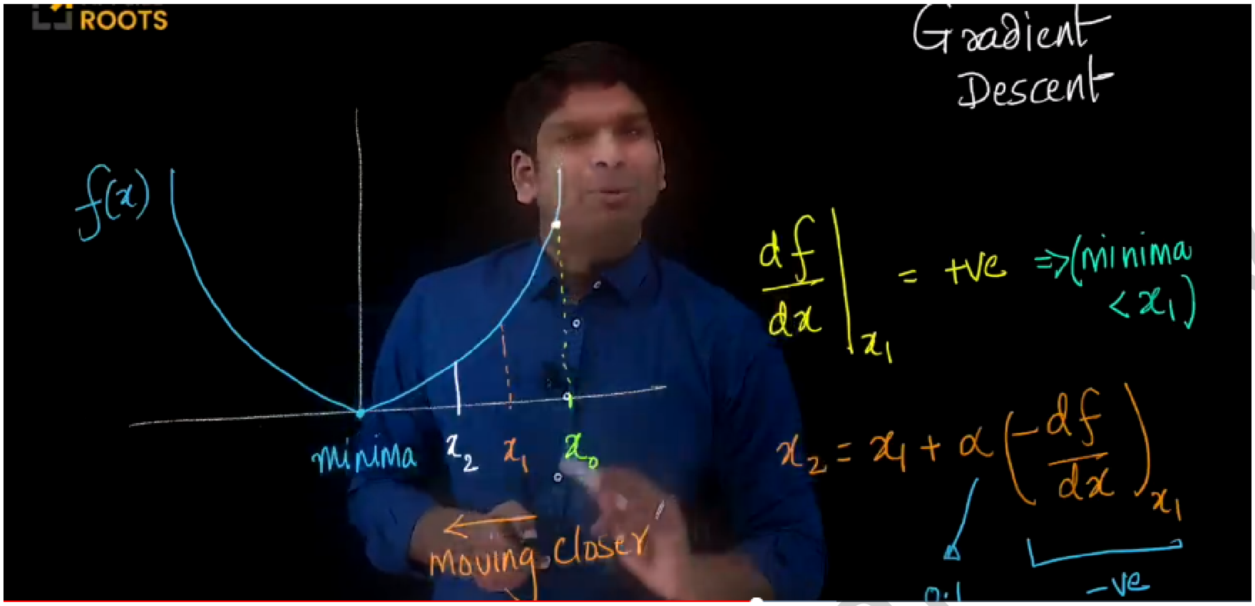
At timestamp 3.48

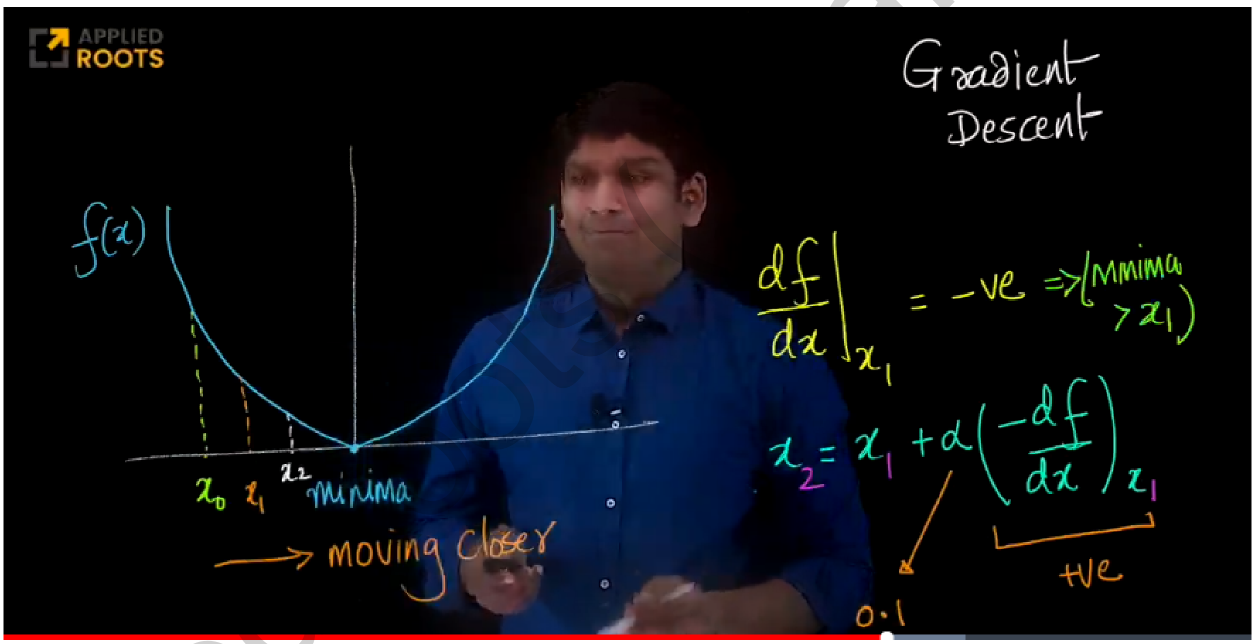



- Let's consider the above curve and we want to find the minima ,we consider a random point x0 and find the derivative at that point .Since the derivative is positive we know that the function is increasing and minima<x0.



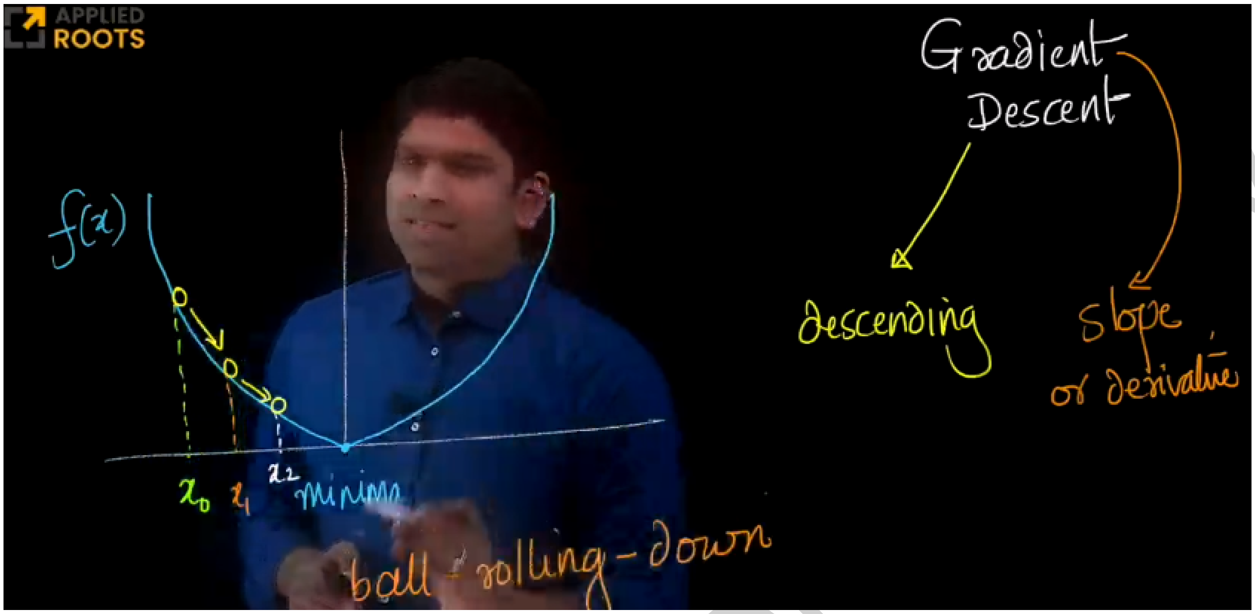

- We know that we have to move towards minima and we do this by using the above equation.




- We keep on updating x using the equation until we reach the minima as shown.

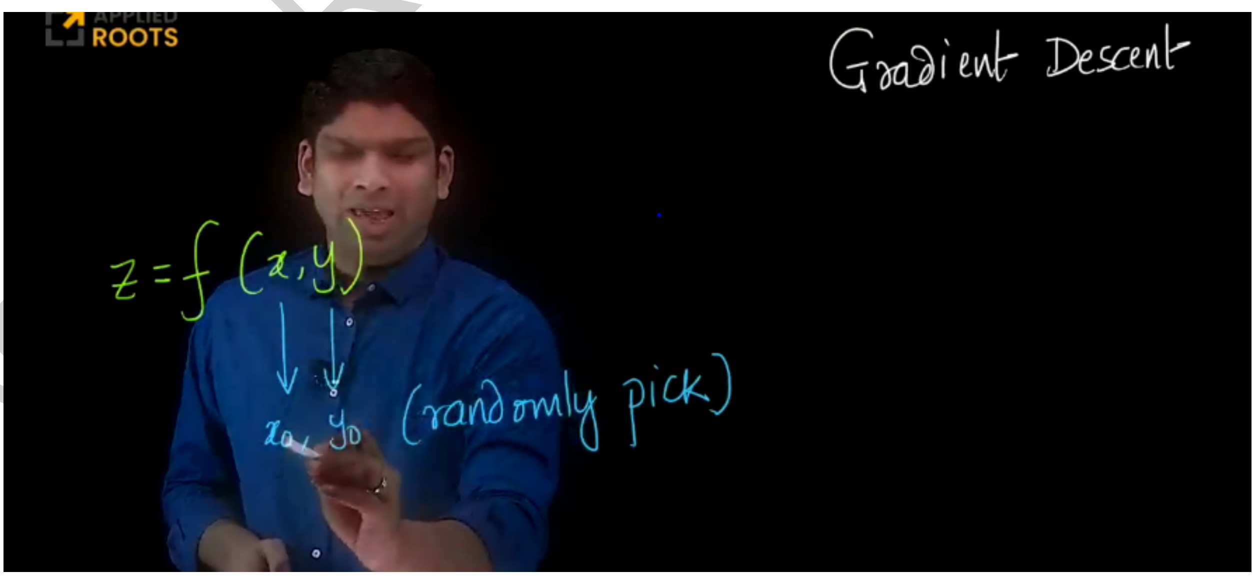
At timestamp 13.45

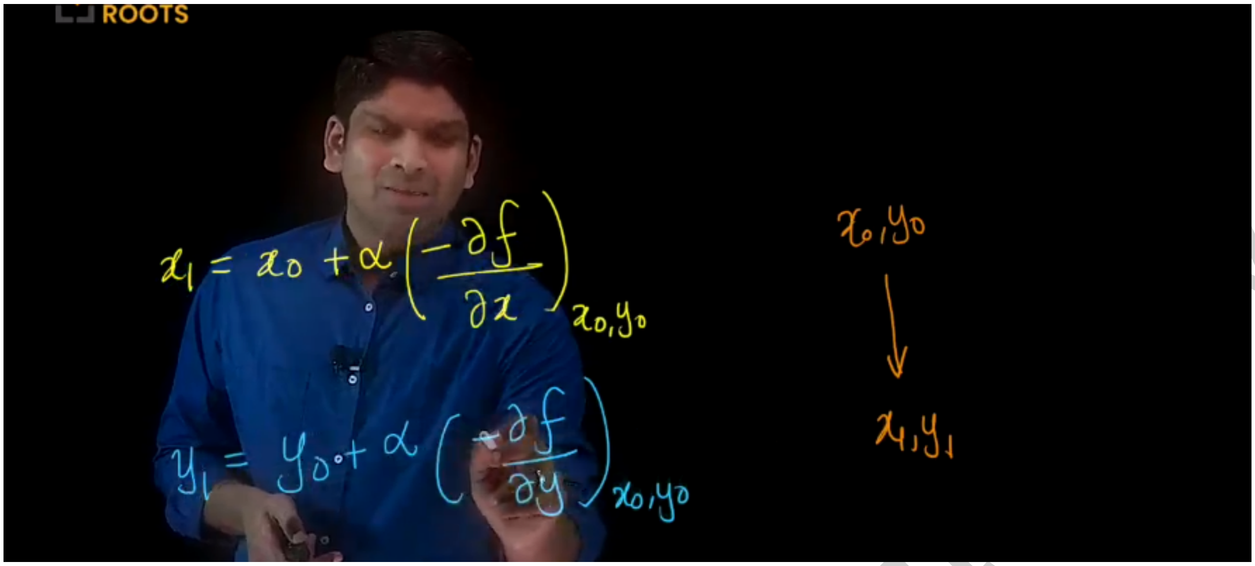



- The algorithm is called gradient descent because we are using gradient or slope or derivative and we are slowly descending towards minima by starting at a random point on the curve.

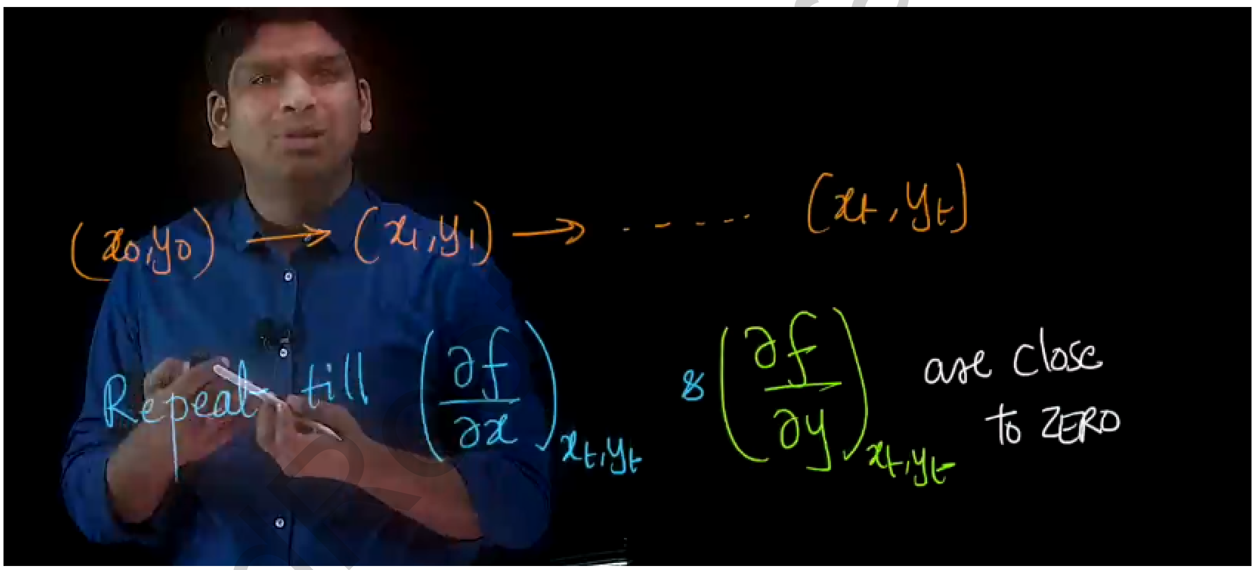
# 3.19 Gradient Descent with multiple variables

At timestamp 0.23





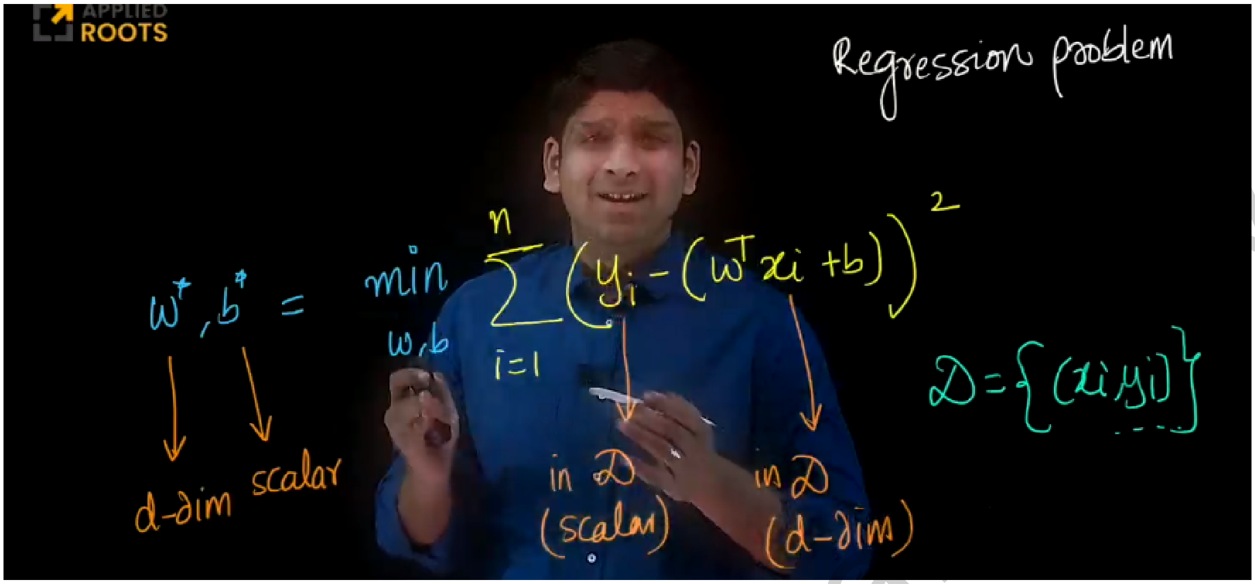

- We are trying to understand gradient descent with multiple variables as shown above



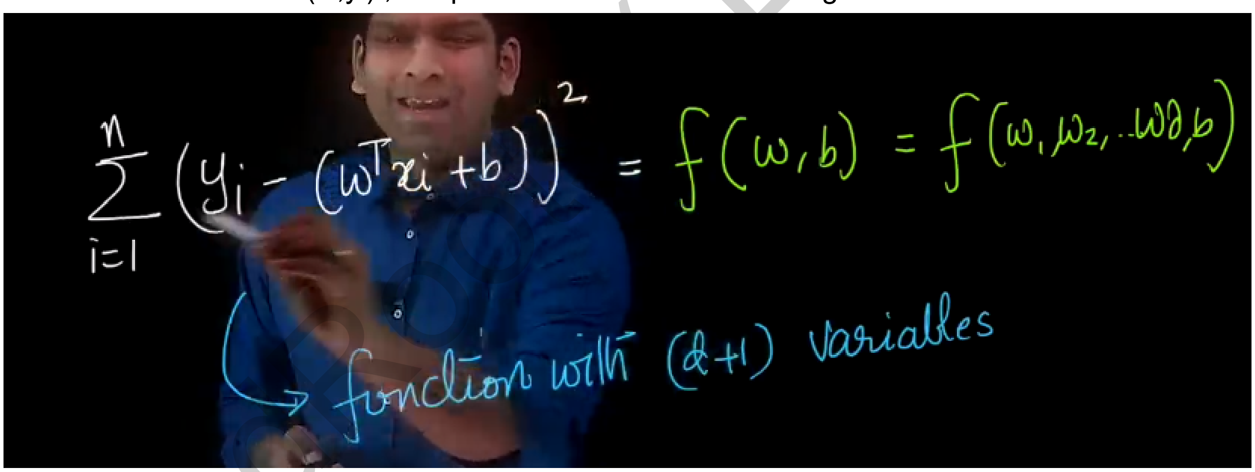

- We start with x0,y0(we randomly choose them) and we keep on updating the values using the above equation until we get partial derivatives as close to zero as possible,then we consider that we have reached the minima.

# 3.20 Regression using Gradient Descent

At timestamp 0.23

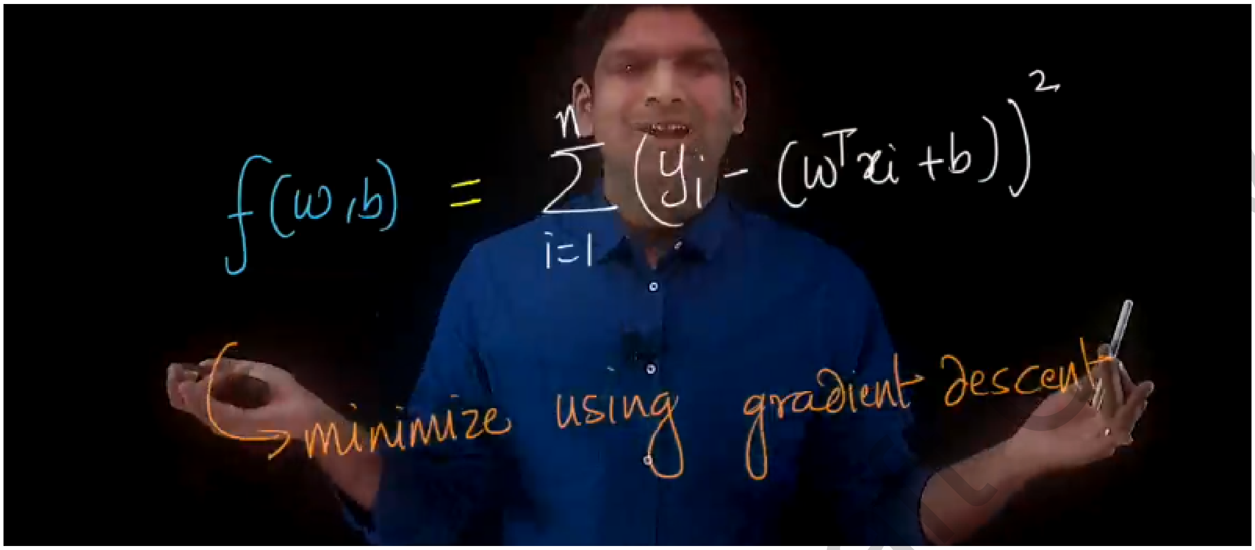



- Let's solve our regression problem using gradient descent .We arrived at the above optimization problem where we have to find optimal w,b such that our sum of squared distances should be as minimum as possible.
- Given dataset D(xi,yi) ,The problem at hand is minimizing the above function



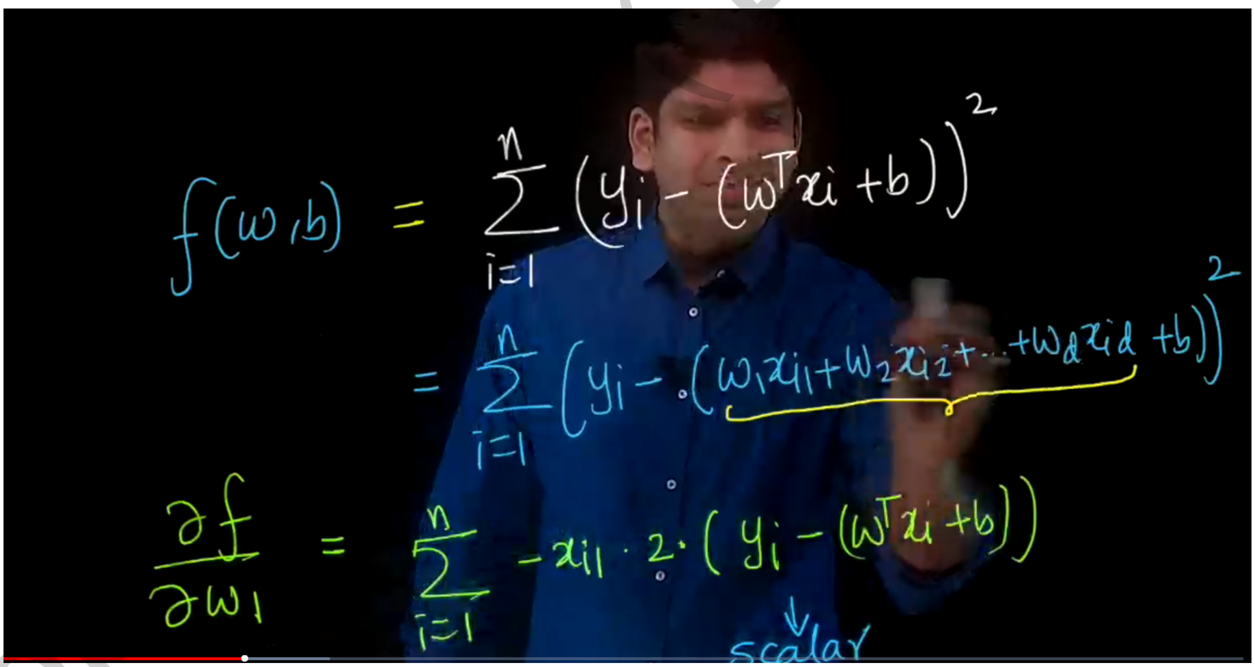

- We can write this as function of vector W and scaler b

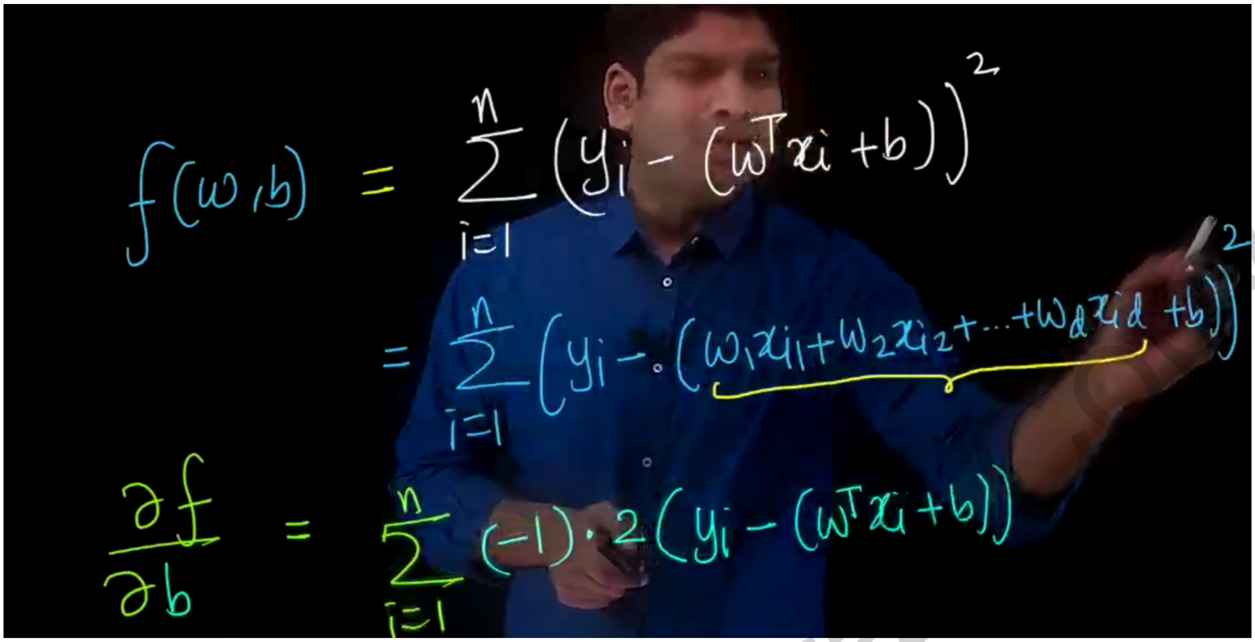
At timestamp 3.40



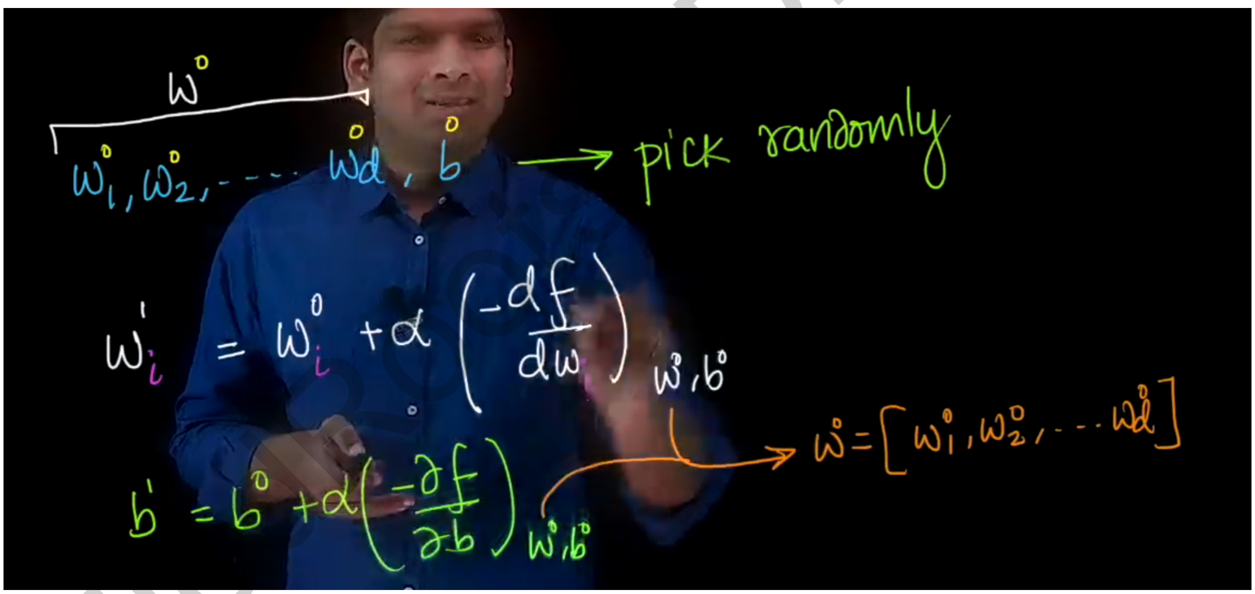

- We can use gradient descent here since we have two variables w,b and we have to find the minima.

At timestamp 4.16

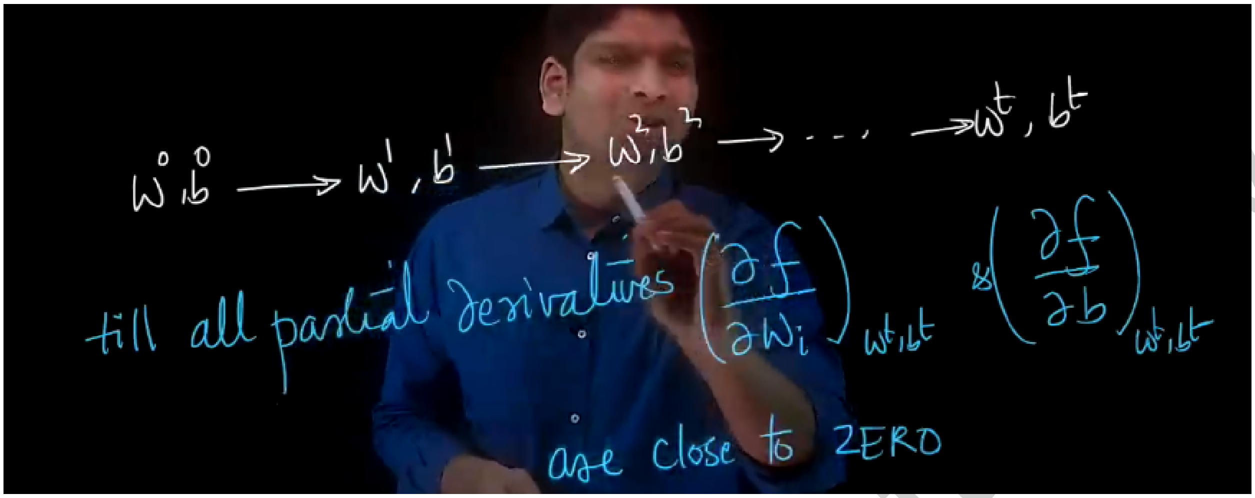





- We calculate the partial derivatives as shown above.



At timestamp 12.42




- Initially we randomly pick w vector and b,then we use gradient descent and keep on update the values and move towards the minima as shown above.
- We use gradient descent even when we have more than two variables.So given a regression problem we can find the best line/plane using gradient descent while minimising the loss.