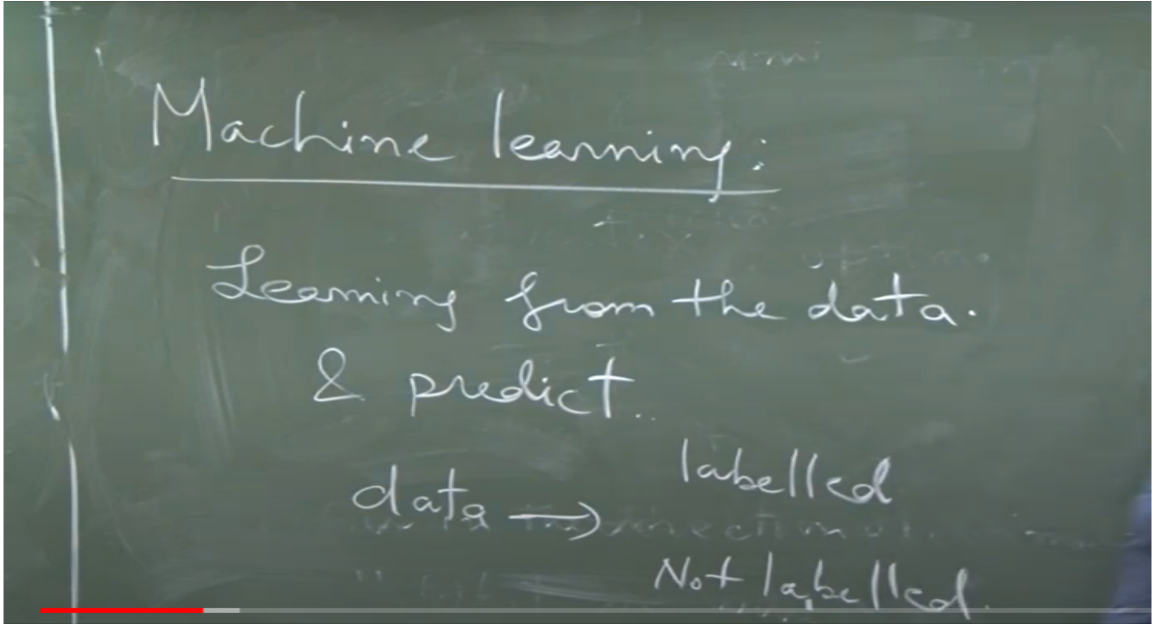
Linear Programming and its Applications to Computer Science
Prof. Rajat Mittal
Department of Computer Science and Engineering
Indian Institute Of Technology, Kanpur

Lecture – 48
Linear Regression through LP

Welcome to another lecture on linear programming and today we are going to take some nice application of linear programming and this is I guess the favorite topic of half the young generation Machine learning everyone is learning Machine learning and it will not be my course will not be complete if I do not talk about how linear programming can help Machine learning. So, you will see some basics of Machine learning what Machine learning is about and then very natural problems in Machine learning which can be tackled through linear programming. Let us start what is Machine learning? Again it is a part of artificial intelligence, but for me Machine learning in some sense is learning from the data. You are given data about environment around you about some particular problem and then you want to predict the output you might be given rainfall of previous 10 months and you want to predict what will be the rainfall next month you might be given crop productivity for 10 areas and you might be asked what happens to a new area and how much crop can it produce all those questions can be tackled with the help of data. You are given data you learn from it and predict. Again I am not an expert in Machine learning.

So, whatever I say about Machine learning take it to the pinch of salt, but these are all kind of intuition of what we are doing. This is Machine learning and here data can be of two types labelled or not labelled. These are broadly 2 categories it could be semi labelled and other things also, but broadly this is the case. Let me explain it with an example.

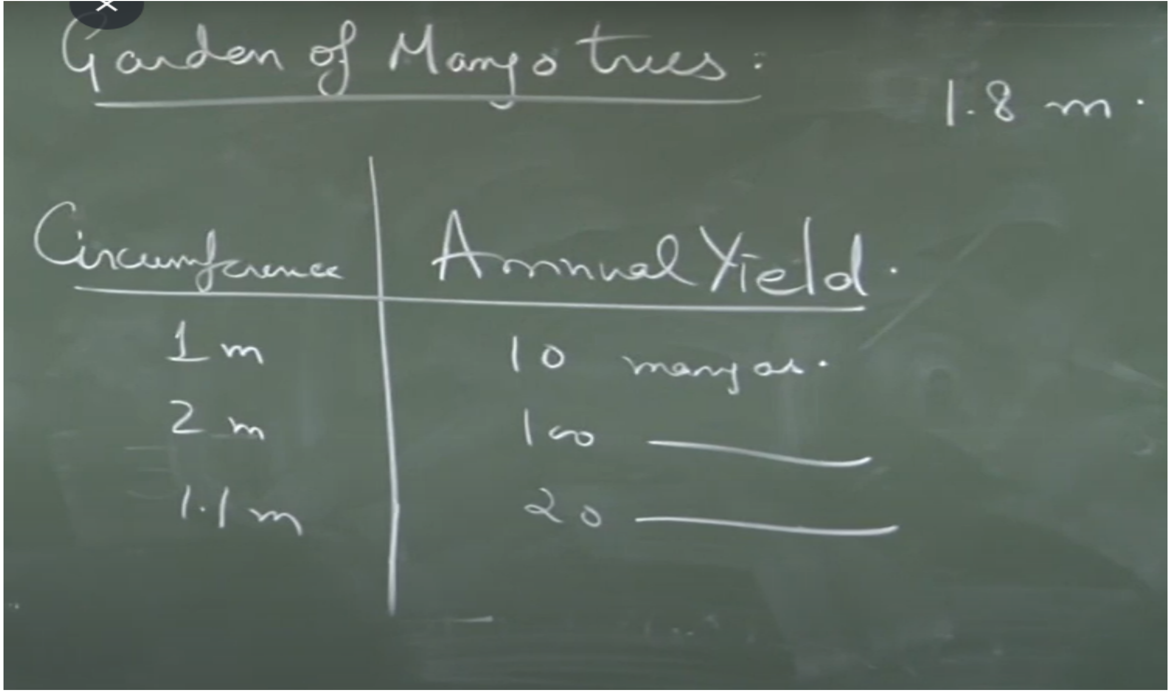


Machine learning:

Learning from the data.
& predict.

data → labelled
Not labelled.

Suppose you have a garden of mango trees and you believe that if the tree is thicker it produces more mangoes. So, you have built up a table which keeps this data in record many trees can say oh if the circumference was 1 meter you get 10 mangoes that is probably less, but let us start with that 10 mangoes per year. If the circumference was 2 meters then you get 100 mangoes, if the circumference was there was another tree which had the circumference was 1.1 meter you got 20 mangoes so on and so forth right. This is a problem where some data is given and then you have now in the next year or so you have another tree whose circumference is let us say 1.8 meter and you want to predict what is going to be your output. This is a standard Machine learning problem. In this case this is a label data because for the all the previous iterations we are given what our output is, what the label is on 1 meter, what the label is on 2 meters and so on and then you want to predict the label of the new thing.



Garden of Mango trees:

1.8 m.

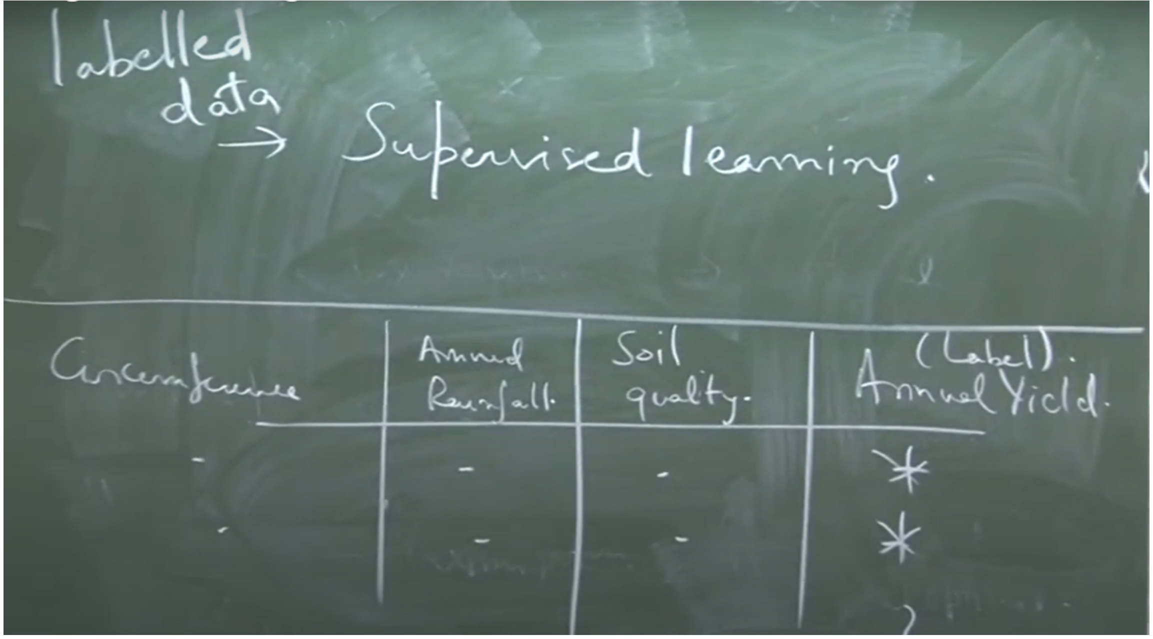
Circumference	Annual Yield.
1 m	10 mangoes.
2 m	100 ———
1.1 m	20 ———

This is example of label data and we call it supervised learning. On the other hand in many cases you might have two kind of data you might have let us say you have a collection of trees whose outputs are given and you want to understand or the circumference is given and you want to classify them as saying these are productive trees, these are unproductive trees.

So, you want to classify them on the basis of something. So, that would be a unlabeled data and that would be unsupervised learning, but for today for our applications we will be interested in supervised learning. So, it would look like we have some data, we have output on that data and using those data and output we want to give an output on some new data. This is the idea in the supervised learning. Notice here simply I have taken just one dimension of data, but probably the annual yield depends on multiple factors.

So, a general problem will have many columns of data. So, you can say circumference, annual rainfall, I do not know what else can your productivity depend on, soil quality, some water something and then finally, you will have the label on it. In our case the label is the annual yield, this is going to be the label. So, now you will have data here,

here you will have a label, you will have data here, you will have a label and then question would be given some new data, what should be the annual yield? And again I am not going to go into detail of physics or biology of this thing, whether the annual yield depends on all these things, we will assume that annual yield depends on just these three things and then what we can predict.

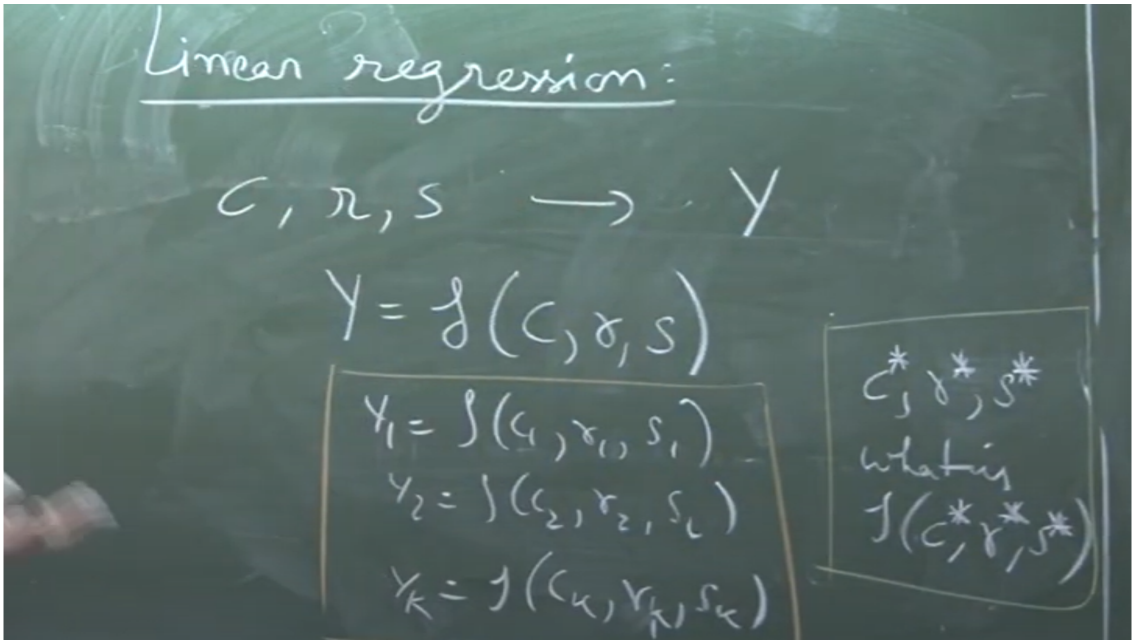


labelled data → Supervised learning.

Circumference	Annual Rainfall.	Soil quality.	(Label). Annual Yield.
-	-	-	*
-	-	-	*
			?

The first problem I am going to talk about is linear regression. Let us look at the problem again and see if are we going to, are we trying to solve a very difficult problem. We had this data, we had numbers associated with, so let us say first tree, second tree, third tree, you have these numbers, you have the annual yield, right. I can say that I have three parameters on which annual yield depends, that is circumference, soil and soil quality and then I have the yield. And I can say that yield is a function of these three parameters.

So, then I am given Y1 is equal to f C1 R1 S1, let us say you are given k data points, this is the data which you have recorded and the problem is now given some new data point, what is f C star R star S star. So, what is my yield for this particular data? If my tree had this circumference, it observed this much rainfall, it grew up in this kind of soil quantity, how many mangoes should I expect? I know the past trend, this is the past trend, what is going to be the future.



Linear regression:

$c, r, s \longrightarrow y$

$y = f(c, r, s)$

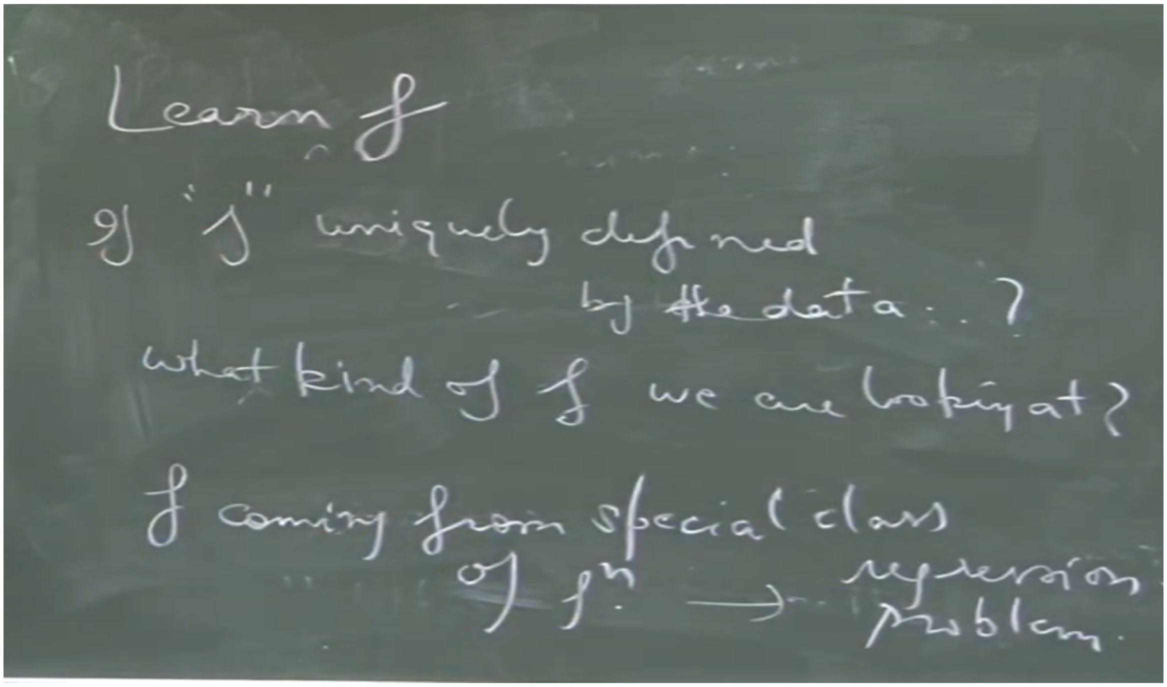
$y_1 = f(c_1, r_1, s_1)$

$y_2 = f(c_2, r_2, s_2)$

$y_k = f(c_k, r_k, s_k)$

c^*, r^*, s^*
what is
$f(c^*, r^*, s^*)$

Now, ideally you would like to learn this function f, right. And if we can learn f, then our problem is solved, once we know what f is, we can put C star R star S star and get the value of f of f of C star R star star. But notice defined by the data, if you look at it by Mathmatical perspective this much data is not enough to define function f there can be many many functions that can be fitted which has different values on these quantity. So what actually is the answer? If we allow f from all sort of fuctions then these questions does not even make sense.



Learn f

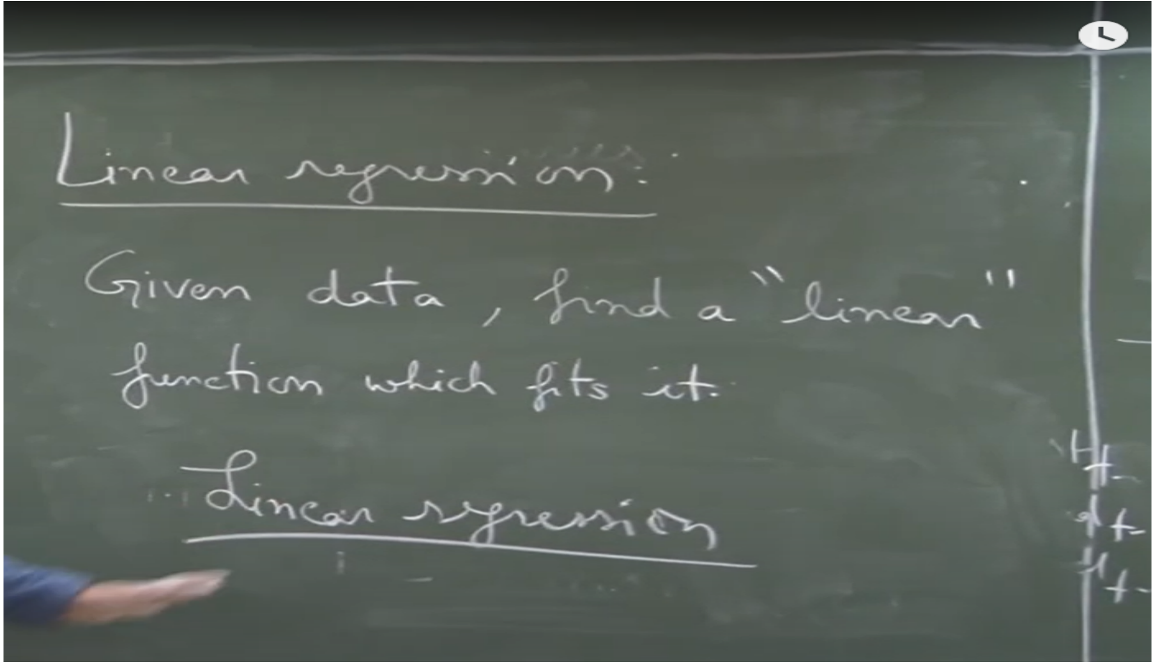
If "f" uniquely defined by the data . . ?

what kind of f we are looking at?

f coming from special class of f^n $\longrightarrow$ regression problem.

What kind of f are we looking for? So, to make this question meaningful, we look at f coming for special class of functions. It could be trigonometric functions, it could be linear functions, it could be quadratic functions, but you restrict your class of functions, you say that you know this is this natural process where the mangoes depend on circumference, rainfall and soil quantity only in a linear way or in a quadratic way, instead of saying that this could arbitrarily depend on these three quantities. Because if this arbitrarily depends on all these, there is no way to successfully predict this thing. So, if we want to have a chance of solving this problem, we have to restrict our attention to special class of functions and this is called the regression problem. We want to find the best function in some class, which fits our data.

If that is the case, if this is what we are looking for, given that we have been talking about linear program and everything, the one case which could be possible for us to solve is linear regression. You might already guess from the name what we are going to do. Given data again of this kind or this table, whatever suits, whatever seems familiar to you, given the data that means, our parameters and the label on that, find a linear function which fits it. This is the problem of linear. That is a very good question.



Linear regression:

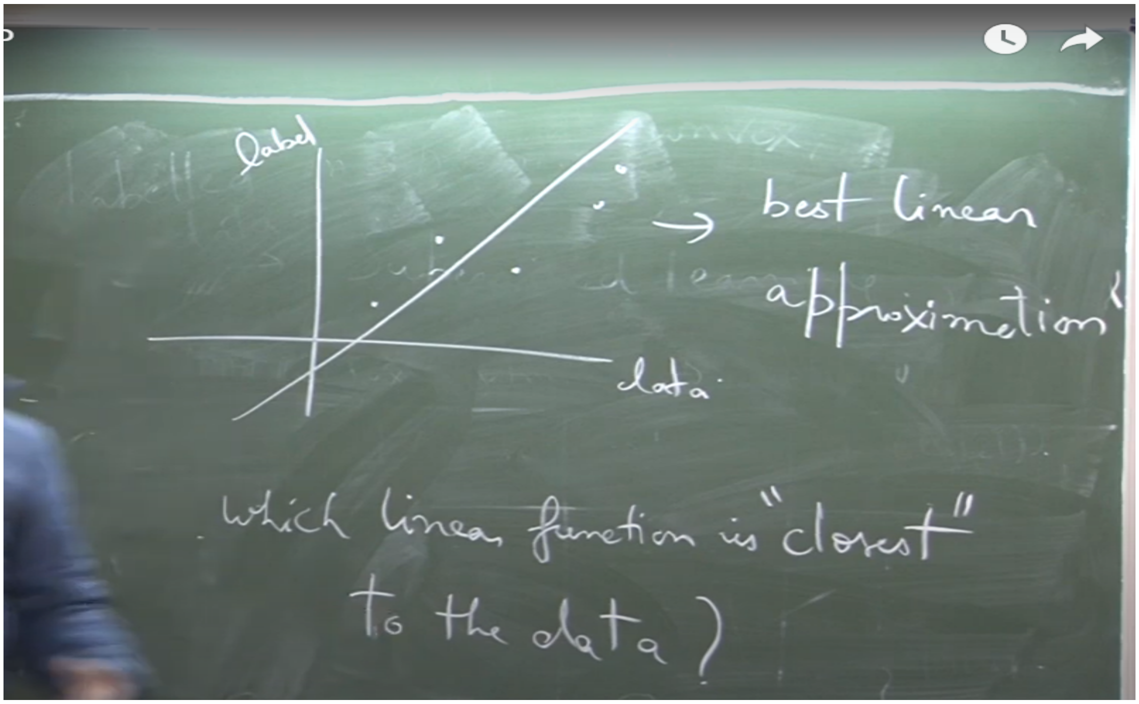
Given data, find a "linear" function which fits it.

Linear regression

So, Tufan asks how do we know that the function has to be linear or quadratic or trigonometric. That is not in our hand that would be decided by the expert people in these areas. What we are going to say is that if you feel that linear functions are the kind of functions you are interested in, then we know how to solve it. There are different ways to solve it, but using linear program you will be solving linear regression problem. Different kind of regression problems.

Machine learning people are interested in many different kind of regressions. Some people that might be experts in gardening or something will say this should depend linearly. Given these parameters should depend linearly. Then it seems interesting.

Exactly. So, as far as I understand people with domain knowledge will be able to decide who are of that domain, whether this function should be linear or of certain kind. So, they will say this is a function of certain kind. If that certain kind is linear, then we are in play. Then we can say something.

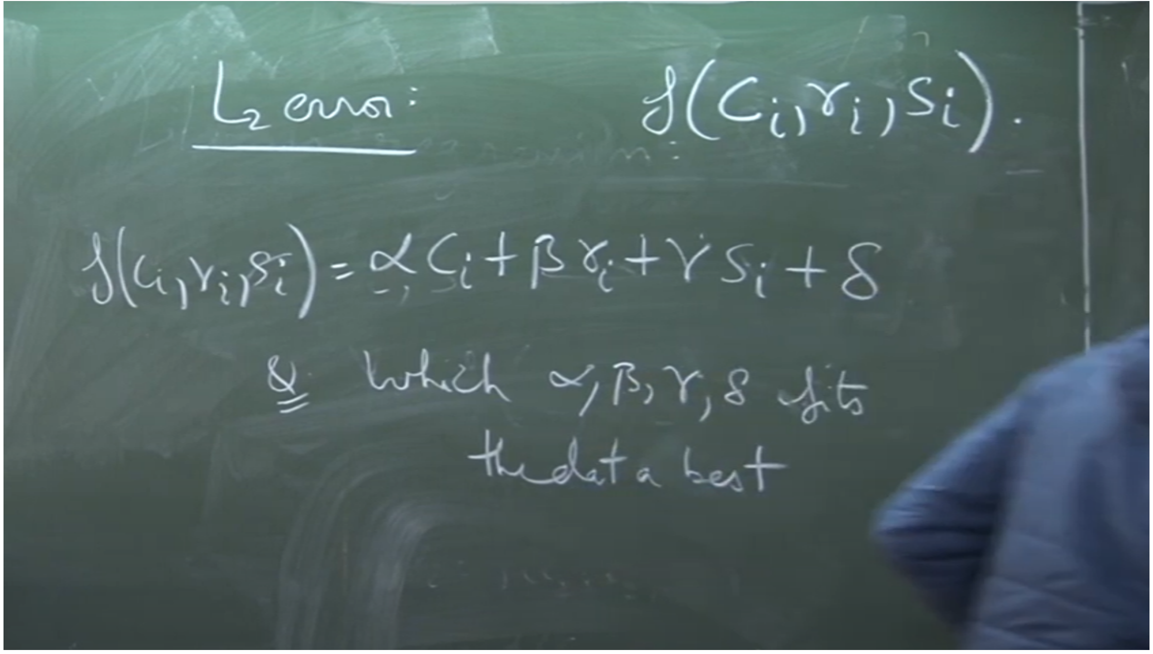

label
data
→ best linear approximation
which linear function is "closest" to the data ?

Very nice point. Very nice point. So, Tufan is asking what does it mean that the linear function fits it? Does it mean that it gives the exact answer? Does it mean approximates? And from previous experiments or slightly from even from experience, it will turn out that your data will not fit in a linear function. So, what does it mean? Let us look at the same problem graphically. We can say that we have different points. So, this is my label, this is my data.

I am simply assuming my data to be one dimensional, but it could be in multiple dimensions. And now I am asking for a linear function which goes through all these points, but that is definitely not possible. I cannot necessarily have a linear line passing through all these points. As you pointed out what we need is the best linear approximation, exactly. Another way to put it is which linear function is closest to the data.

The problem seems well defined, but there is still one thing which I have not defined. Can you guess what? I have some data, I have some points in my space let us say R to the power n, I want to find the line which is closest to the data. Have I defined everything? Am I missing something? If you think about it closely mathematically, what does it mean closest? Which kind of distance am I talking about? Is it Euclidean distance? And even here there is a line and there is a point, right. So, what does it mean? Where is it? Just because I drew the picture in the Euclidean plane probably that is not the right measure. So, which distance should I be looking at? And actually the answer is not very concrete.

Again it depends on what application, what domain you are talking about. You can talk about closeness in multiple areas, in multiple mathematical definitions, some suit some applications, some suit some applications. The most natural I would say would be the Euclidean distance or you would say the L2 error, right. So, assume that I have remember I am looking for a function on 3 variables. What is a linear function on these 3 variables? It looks like some alpha ci plus beta ri plus gamma si plus some delta, right.

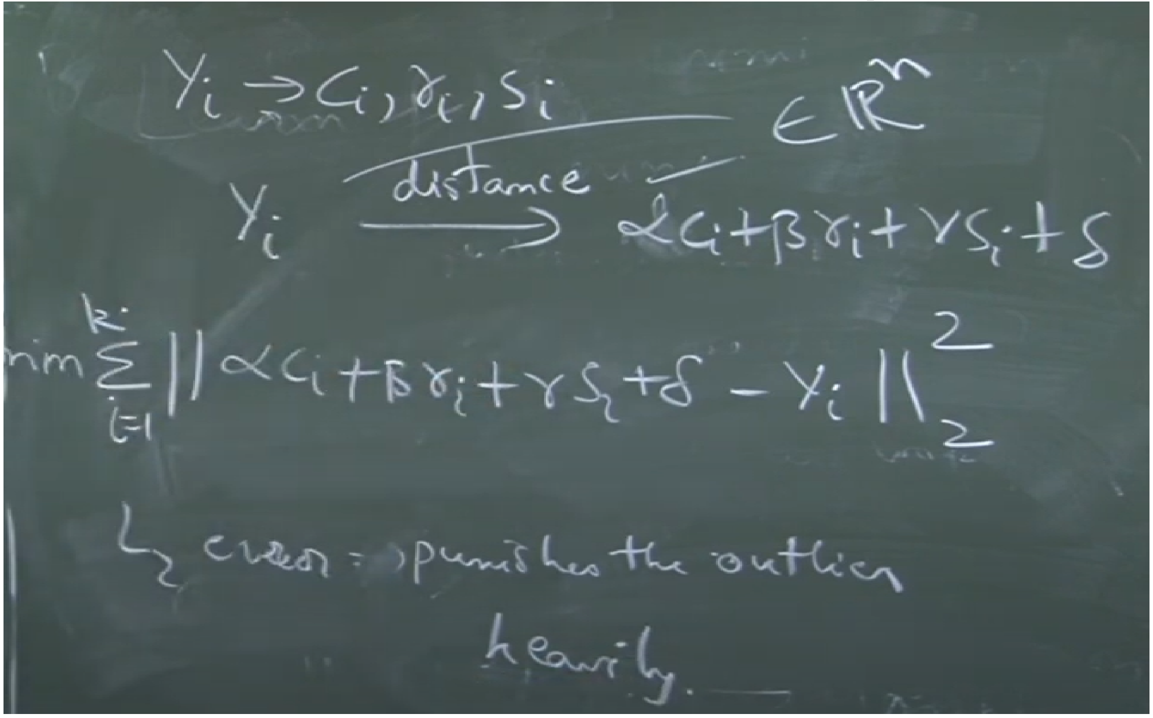


L_2 error: $f(c_i, r_i, s_i)$.

$f(c_i, r_i, s_i) = \alpha c_i + \beta r_i + \gamma s_i + \delta$

& which $\alpha, \beta, \gamma, \delta$ fits the data best

Some people might object and say these are affine functions, but we want our class to be as big as possible. So, we will say we can take care of these. So, this is the definition of linear functions for us. So, this is my linear function and I am interested in finding out which alpha beta gamma delta fits my data best. And we were at the juncture where you wanted to define the error with respect to this.



$Y_i \to c_i, r_i, s_i \in \mathbb{R}^n$

$Y_i \xrightarrow{\text{distance}} \alpha c_i + \beta r_i + \gamma s_i + \delta$

$\min \sum_{i=1}^{k} \| \alpha c_i + \beta r_i + \gamma s_i + \delta - Y_i \|_2^2$

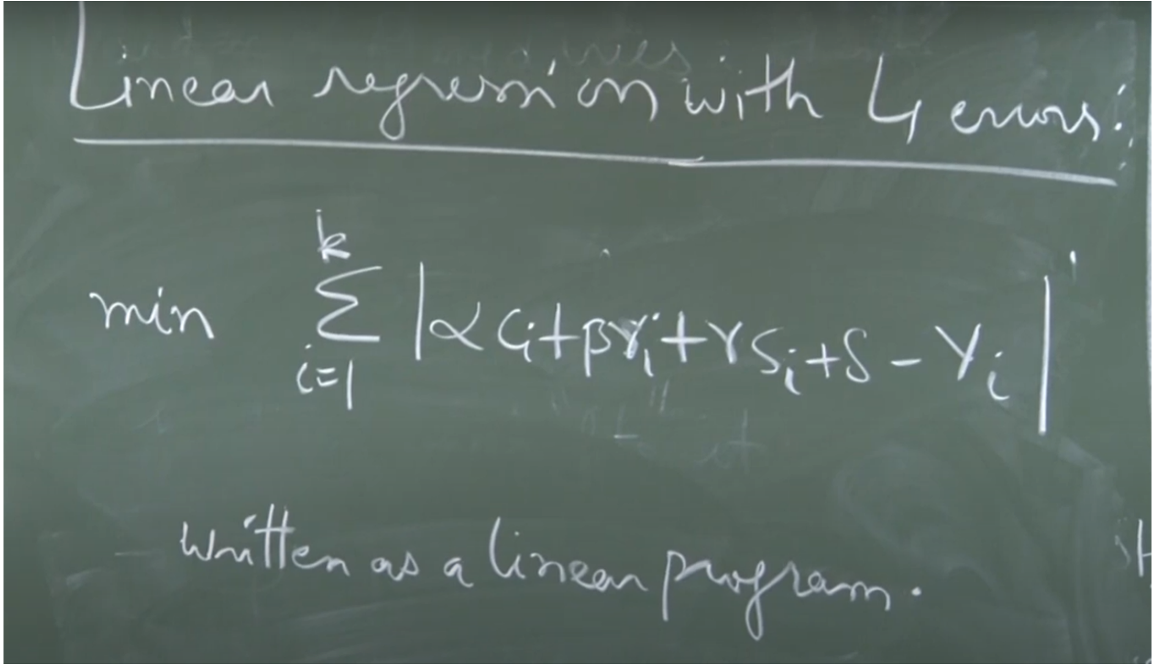
L_2 error $\Rightarrow$ punishes the outlier heavily.

And there is no price in guessing if you your data was yi on ci ri si, you want to kind of have a distance between yi and plus delta. In the case of L2 error, this will be this is a point in r to the power n. These both these points are points in some Euclidean n dimensional space that is the natural notion of distance would be the vector distance between these two. Remember what is the vector distance? This is a vector, this is a vector, you subtract the first coordinate of this with first coordinate of this square it. First coordinate of this, second coordinate of this with second coordinate of this square written at them up.

The Euclidean distance between the distance between these two points as the normal distance that you can take it. And now you might want to minimize the error over all the k points. This is linear regression where your error is L2 error. Unfortunately, this is this notion of error is not very helpful. What is the reason? The idea is if you think about it, L2 error punishes the outliers heavily.

Since, I am squaring it if I have a so suppose this is the point and let us say I have an outlier which is a point at outside. This will make this line really bad. Just this point it will not care about what these points are this will make there are really large. On the other hand something like this will fit this data better. Now if you have chances of outliers you might want to reject them then you want to say that I really do not want to give that much weightage to just one single point.

In that case we do not talk about L2 error. What we talk about is linear regression with L1 error. With norm 1. So, what does it mean? It basically simply means minimize the absolute value of this quantity and this reduces the focus on outliers. So, what we are going to show that this actually can be written as a linear program.



Linear regression with L_1 errors:

$$\min \sum_{i=1}^{k} |\alpha c_i + \beta r_i + \gamma s_i + \delta - y_i|$$

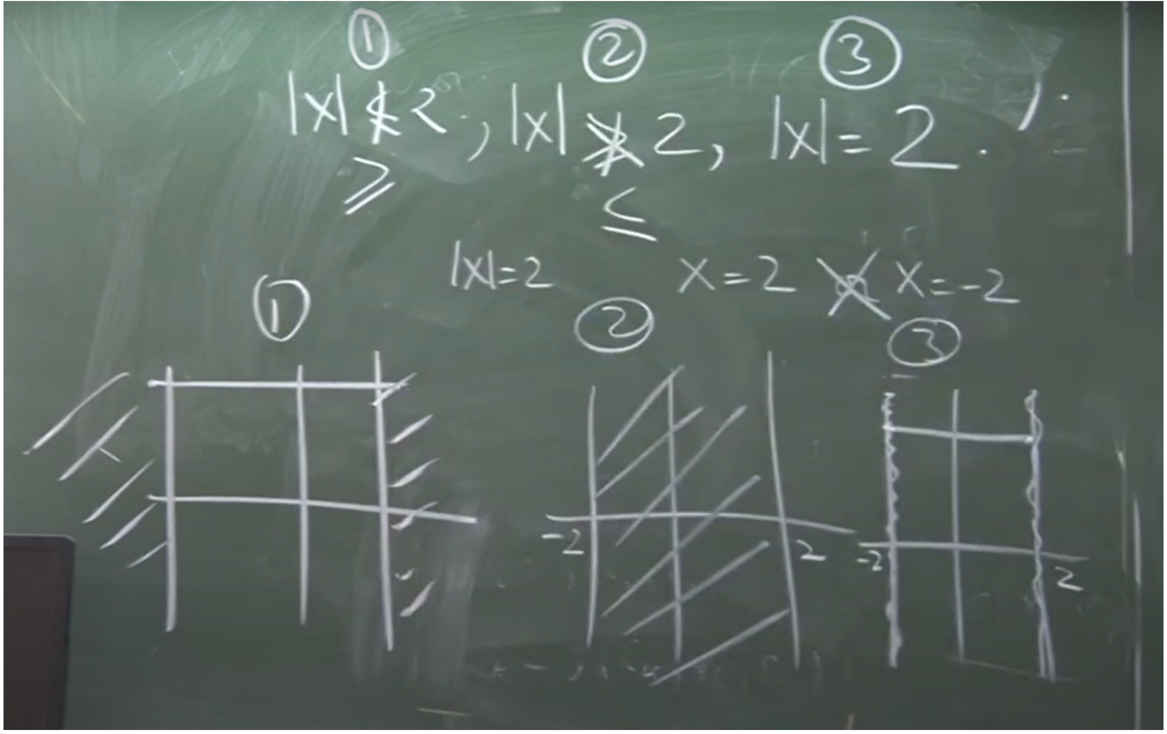
Written as a linear program.

So, we can solve linear regression with L1 error using a linear program. The problem is clear we are given k data points we want to find a linear function. In our case the linear function means these coefficients alpha, beta, gamma, delta such that it fits our function best that means the L1 error is minimized. The good thing is this looks like a linear programming problem, right. But this is not really a linear program what are we missing? First thing of obviously is the absolute value.

We are not allowed to have absolute values in this linear program. Somehow we should get rid of absolute values and then generally we are familiar with linear programming problems where there are constraints. Here there are no constraints, but that is ok we can optimize over r to the power n. So, this is not really a big problem. And you will see that while we reduce the absolute value sign we will get some constraints, but how to get rid of absolute value sign? How to convert this into a linear program? This is a very interesting trick and to motivate that interesting trick.

Let me first ask you mod x less than 2 look at these 3 constraints mod x less than equal to 2 mod x more than equal to mod x mod equal to 2 which of them can be represented as linear constraints. So, we have I can have 1 I can have 2 or I can have 3 which of them can be written in terms of linear constraints. But remember you are writing or, or is not allowed, right. As when we write constraints of linear programming the constraints are end not or. So, can I write these as a bunch of linear constraints? Very good that is very good answer Tufan and he has put the thing at the right perspective.

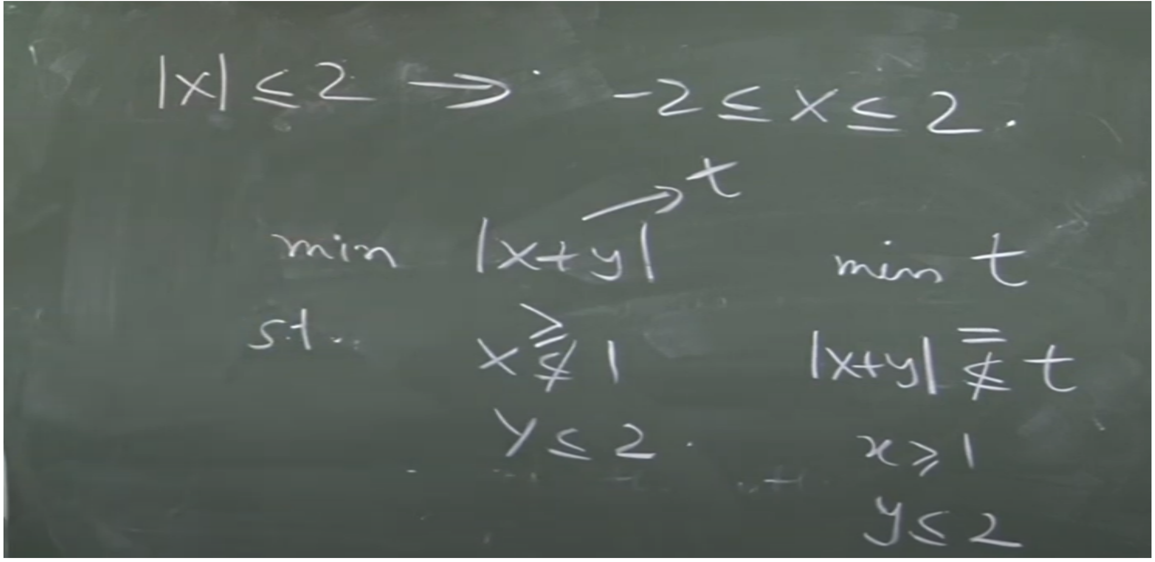
Let us look at these feasible regions. If you look at the feasible regions of these only one of them is going to be nice and nice means convex. So, for 1 my feasible region is this, for 2 my feasible region is this. Remember this is x equal to minus 2 this is x equal to 2 and this is the feasible region for the third set of constraint. And notice what did I ask you which of them you can write it as a bunch of linear constraints. But if they can be written as a bunch of linear constraint we have studied that are feasible region is going to be convex.



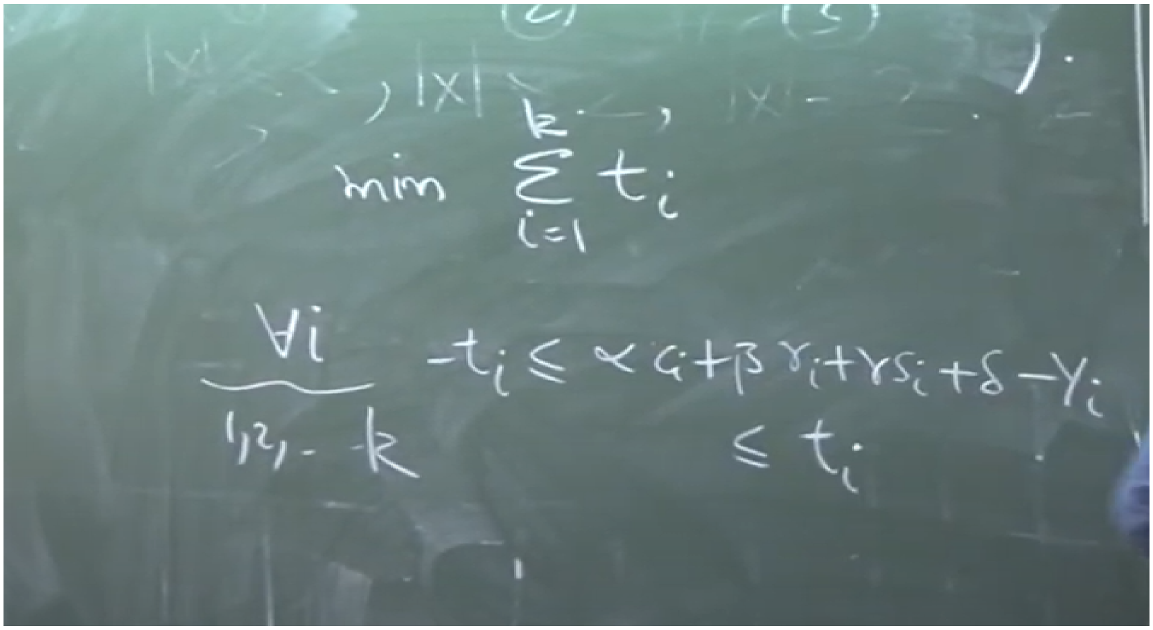
Which of this is convex? Clearly this is not convex because this line is not here, this is not convex because this line is not here, this is the only place which is convex. So, that means I can write this as a bunch of linear constraints, right. Let us take the next step remember our task is to convert this into a linear program. No it is not a ball it is the area outside the ball it is the complement of that.

Oh sorry sorry I actually interchanged oh sorry. My pictures were opposite thanks to Vanva pointing out 1 greater than equal to 2 is this less than equal to 2 is this. And I can write less than equal to 2 as a bunch of linear constraints that is the norm 1 ball or right. So, that is that is why it is convex right. Now, great still it is not clear to us how to remove this absolute value.

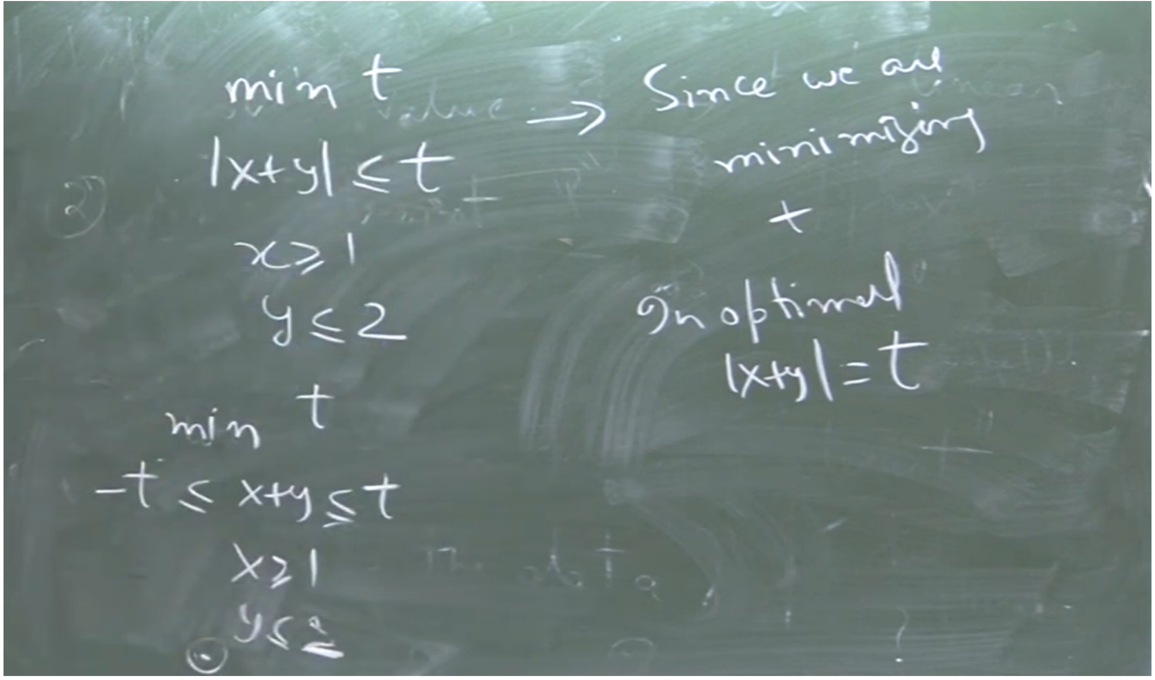
So, let me ask you one more question. Is this a linear program? Let me stick with my notes and let us say this is the problem. Notice I can if I have this way if I have this I can write it as a bunch of linear constraints. So, let me just say first let me put this absolute value in constraints. What is the way to do it? Let us say this is t. So, I can say minimize t such that mod x plus y is less than equal to t.


$|x| \leq 2 \Rightarrow -2 \leq x \leq 2.$
t
$\min |x+y|$
$\text{s.t. } x \overset{\geq}{\nleq} 1$
$y \leq 2.$
$\min t$
$|x+y| \overset{=}{\nleq} t$
$x \geq 1$
$y \leq 2$

Ideally I should be saying equal to t and then y less than equal to 2. So, if this was equal you have no problem in believing me that this is the case right because I have just write wrote this is a new variable and written equality. Now this minimization comes to our rescue. My claim is I can change this equality into inequality because I am minimizing this. I am saying that this linear program and this linear program are still the same changing equality into inequality does not change the value because since I am minimizing t in my optimal solution automatically this is going to force equality.

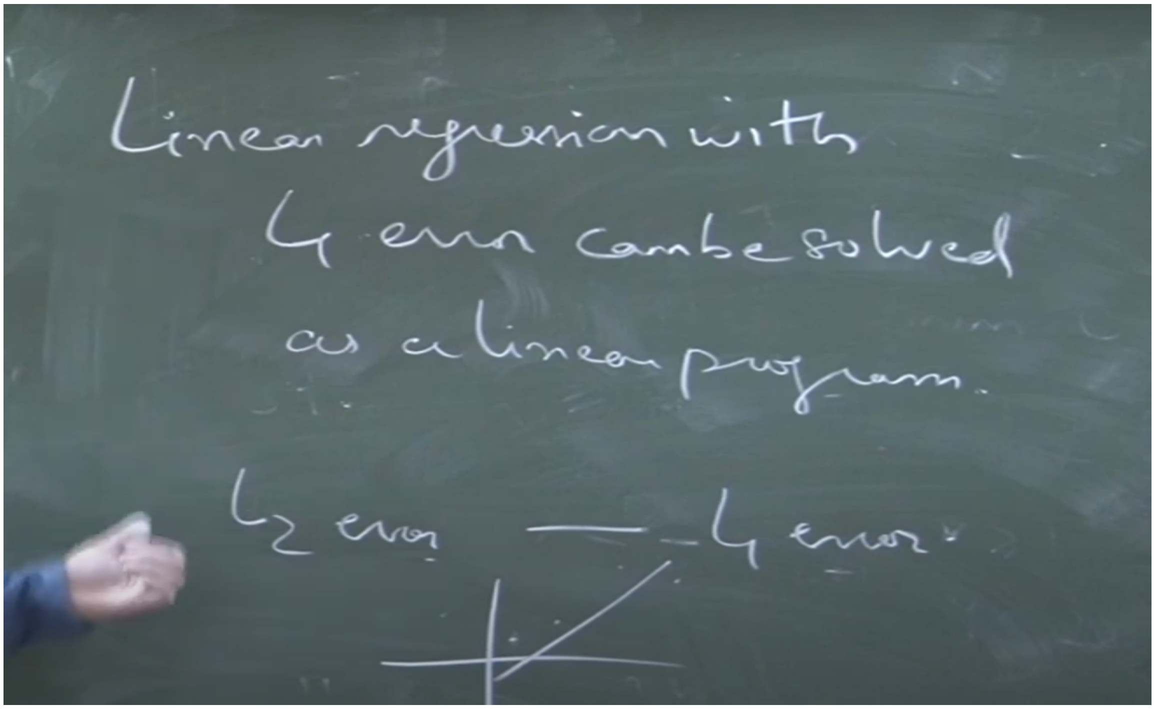

|x|
, |x|
, |x|
$\min \sum_{i=1}^{k} t_i$
$\forall i$
1,2,.. k
$-t_i \le \alpha c_i + \beta r_i + \gamma s_i + \delta - y_i \le t_i$

If this is not equality t does not appear anywhere else I can reduce t and get a better optimal value. In optimal mod of x plus y will actually become equal to t. It was not equal to t I can reduce t and reduce and get the better value. So, that means these two are equal, but this constraint we know how to handle.



This is minimization t. So, a very similar problem like this we can convert it into a linear program and this is same idea by which even this can be written as a linear program. How will we do it? What we are going to do is we are going to introduce new variables for each of the terms here. So, each of the terms will become some ti and then I want to minimize this such that for all i remember that this quantity should be less than equal to t i. So, I am writing it as.

So, there will be k constraints I varies from 1 to up to k. So, we will have k constraints and we are going to minimize and automatically since we are minimizing this quantity at the optimal I might this quantity will either be minus ti or ti. So, my minimization will ensure that I do not lose anything when I change equality to inequality. What it shows is that this quantity can be written as a linear program. So, to summarize linear regression with L1 error can be solved as a linear program. We have a linear program for solving linear regression with l 1 error.



Linear regression with L_1 error can be solved as a linear program.

L_2 error — L_1 error.

The only comment I want to make is between L2 error and L1 error. I justified saying that L2 error penalizes the outliers a lot. Remember we said oh there are point there is a nice line, but if there is a outlier L2 error become really large and if you do not want to give that much weightage to this outlier we will prefer L1 error. I am not saying in all situations L1 error is better than L2 error depending on the applications sometimes we would want to use L2 error for linear regression sometimes you would want to use L1 error. If we want to solve linear regression with L1 error then we can do it using a linear program.