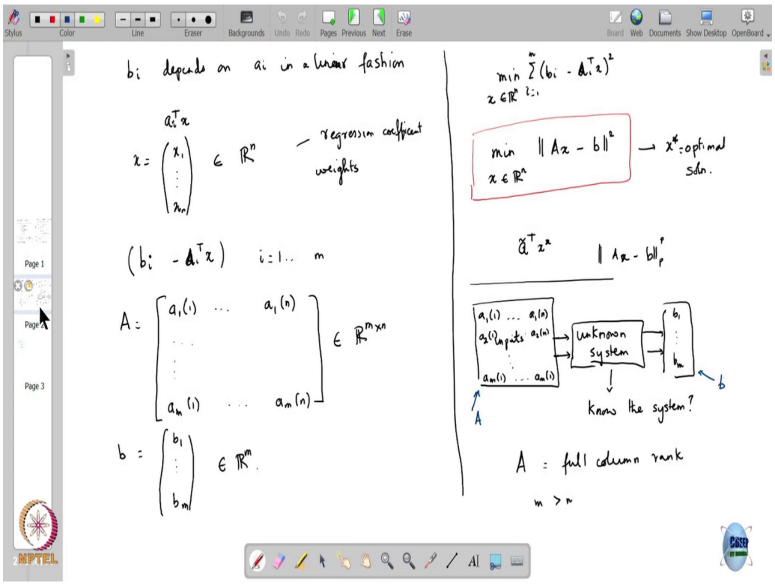
Optimization from Fundamentals
Prof. Ankur Kulkarni
Department of Systems and Control Engineering
Indian Institute of Technology, Bombay

Lecture – 5B
Least square regression (continued)

(Refer Slide Time: 00:16)



b_i depends on a_i in a linear fashion

$a_i^T x$

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ — regression coefficent weights

$(b_i - a_i^T x) \quad i = 1 \dots m$

$$A = \begin{bmatrix} a_1(1) & \dots & a_1(n) \\ \vdots & & \\ a_m(1) & \dots & a_m(n) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$$

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} (b_i - a_i^T x)^2$$

$$\boxed{\min_{x \in \mathbb{R}^n} \|Ax - b\|^2} \rightarrow x^* = \text{optimal soln.}$$

$\tilde{a}^T x^*$ $\quad \|Ax - b\|_p^p$

$a_1(1) \dots a_1(n)$, $a_2(1)$ inputs $a_2(n)$, $\vdots$, $a_m(1) \dots a_m(1)$ ← A → unknown system → $b_1, \vdots, b_m$ ← b

Know the system?

A : full column rank

$m > n$

So, broadly speaking the problem is something like this. So, you have an unknown system to which you have provided inputs right. Inputs denoted in this sort of fashion n inputs here next time you provide again another set of n inputs and you have got some outputs. And what we would like to do is know what the system is like.

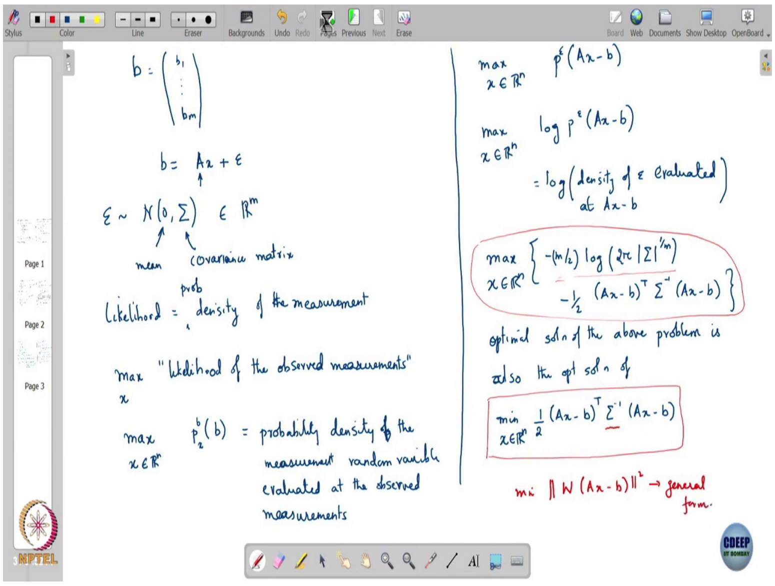
And this is the sort of problem that results from this kind of formulae. Now, it is important here if when you go about solving this problem it is important to that a the matrix a is full column rank ok. And in particular in its in may one way that can be ensured is that you have in if you are talking of a system and taking and you have provided m different inputs means you are taking measurements m times of the system.

That means the number of measurements is greater than n which is the length of the input or the number of parameters that you are trying to put. In that case this can be solved very easily and has a non trivial solution ok. Now, how is this related to what we have been talking about? The reason I brought this up is if you look at this optimization problem which is what the entire exercise of regression brought it brought as down to is this optimization problem. This what kind of optimization problem is this?

This problem has a twice differentiable objective right and it is being optimized over the entire space of R n; now the full space of R n is obviously, an open set. So, this is an optimization of a twice differentiable function over an open set. Now, because of the nature of the function here it is just a simple quadratic function this also admits a closed form solution, in using the pseudo inverse of a.

But, that is because of the choice of norm that we are chosen, but in general if you take some other type of loss function you would not necessarily have a closed form solution. So, an example of this is one example of a optimization over an open set as an unconstrained optimization problem. Let me do another example of a similar flavor, in this case we actually here the problem is slightly different here the problem is that you actually have a model for the noise ok.

(Refer Slide Time: 03:38)



$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

$$b = \underset{\uparrow}{Ax} + \varepsilon$$

$$\varepsilon \sim N(\underset{\substack{\uparrow \\ \text{mean}}}{0}, \underset{\substack{\nwarrow \\ \text{covariance matrix}}}{\Sigma}) \in \mathbb{R}^m$$

Likelihood = prob density of the measurement

max "likelihood of the observed measurements"
x

$$\max_{x \in \mathbb{R}^n} \; p_x^b(b)$$ = probability density of the measurement random variable evaluated at the observed measurements

$$\max_{x \in \mathbb{R}^n} \; p^\varepsilon(Ax - b)$$

$$\max_{x \in \mathbb{R}^n} \; \log p^\varepsilon(Ax - b)$$

= log(density of ε evaluated at Ax - b)

$$\max_{x \in \mathbb{R}^n} \left\{ -(m/2) \log\left(2\pi |\Sigma|^{1/m}\right) - \tfrac{1}{2} (Ax - b)^T \Sigma^{-1} (Ax - b) \right\}$$

optimal soln of the above problem is also the opt soln of

$$\boxed{\min_{x \in \mathbb{R}^n} \frac{1}{2} (Ax - b)^T \Sigma^{-1} (Ax - b)}$$

$$\min \|W(Ax - b)\|^2 \rightarrow \text{general form.}$$

So, suppose you have we have you have taken m measurements and you have gotten this b the vector of measurements like this b equal to b 1 to b m these are m measurements that you have taken of a system. And we know that b should be equal to A x plus some epsilon where epsilon is your measurement ok.

So, what we do not know is x. So, you provided this inputs A this is your that is your matrix A here you got these output outputs or measurements b here and we know that b and a are related in this sort of way b is equal to x, but there is some measurement noise involved which is epsilon. And suppose we have for I am going to assume that we know the also that the noise takes has Gaussian density.

So, noise is distributed as a Gaussian random variable with the mean as the 0 vector and covariance matrix sigma ok. So, this is the mean of the Gaussian random variable the covariance matrix. So, when you get a string of these m measurements all corrupted by noise like this, but we know the density of the noise we have we know that it is has a Gaussian density. One philosophy for guessing what the x that result that resulted in these measurements is to say is to look at the likelihood.

Likelihood is simply the density likelihood ok; likelihood in this case is simply the density of the measurement the probability density of the measurement ok. So, if you are if you have, these are discrete random variable it would be the probability of seeing that particular measurement.

And then say ok, well this is the probability of seeing this particular measurement with when; you know when I gave these inputs then you ask and my there is a hidden parameter or an unknown parameter here which is x. You ask what parameter would have maximize the chance of seeing this particular measurement ok?

So, you look for the parameter x that maximizes the likelihood of seeing this particular parameter ok. So, you maximize effectively the you can say is the likelihood maximize find the x that maximizes the likelihood of the observed measurements. So, you find likelihood of this.

Now, what would that be? That would be basically in this case we also your maximizing over x in R n the probability density evaluated at the measurement that you have to see. This is the probability density of the measurement random variable evaluated at the absorbed measurements.

Now, this is the probability density of this. So, this well the probability density well it is derived directly from the probability density of the noise itself right and the system model that you have built. So, consequently it implicitly depends on x ok. So, I will write a small x here just to denote that just to indicate that there is a dependence on the x side.

So, what we are doing is; we are trying to say what kind of choice of x would have given me the highest likelihood of seeing this particular outcome ok. There are criticisms about this particular philosophy of finding the x that is a separate matter that is not for a discussion. I want to get to the optimization problem that it this implies. So, let us just let us take this forward.

So, this P x of b is actually nothing, but is can be written in the other in a different sort of way well, I is maximize over x in R n which is the probability of an epsilon getting realized probability density of epsilon evaluated. So, this is now its call denote this. So, this was probability density of b this is the probability density of epsilon. So, this is probability density of epsilon evaluated at A x minus b right.

So, what is the probability that epsilon turned out to give you can say take value A x minus b. You are finding the x that gives me this is the largest set value ok. Now, it is quite convenient in this case to now instead of looking at the probability you actually it is convenient to take the log of the probability because log is a monotone function that does not affect the optimization.

So, you can just simply instead of doing a maximum likelihood you do maximum log likelihood. So, do log of probability of the density of density of the random variable epsilon evaluated at A x minus b ok. So, this log of density of epsilon evaluated at A x minus b.

And this so because this is a Gaussian random available this epsilon was remembered in R m and this is Gaussian random variable the log of this I can tell you its you get a it has as an expression. Its minus m by 2 times log of 2 pi the determinant of sigma raised 1 by m minus half A x minus b transpose sigma inverse A x minus b.

We remember sigma is the covariance matrix of random variable. So, what you are doing effectively is what you are doing effectively is doing a maximization of this function over x in R n. Now, if you look at this function you will notice that the first term here the first term here has nothing to do with x at all. But simply a constant it does not affect the optimization problem it only shifts the function lateral laterally or in one direction. It has no it has no bearing on the optimization so this.

So, as far as finding the optimal solution the optimal solution of the above problem is also the is also this will optimal solution of this problem, where which is maximizing minus half A x transpose sigma in A x minus b transpose sigma inverse A x minus b maximizing this second term.

Now, that is maximization of a negative quantity something that is negative. So, there is a minus sign outside maximization of minus of this. So, that is what I can do is that is actually equivalent to doing minimization of the same thing multiplied by minus 1. Minimization of this sigma inverse A x minus b.

So, the x star that maximizes this quantity the first optimization problem here is this optimization problem is the same as the x star that solves this particular. These are you can get the it is a same x star that solves. The optimal value would not will differ because we have dropped additive constants that were there in the objective function there have been rocks.

So, they have to be added they need to be adjusted for, but the x star is the same the optimal solution will be same alright. So, this is also another optimization problem over an open set ok. This is very similar to the problem of minimizing the norm of A x minus b except that now the norm has been scaled with this sort of matrix in between sigma inverse right. So, it is or scaled or rather the correct word is skewed by this matrix A x minus b by this matrix sigma inverse.

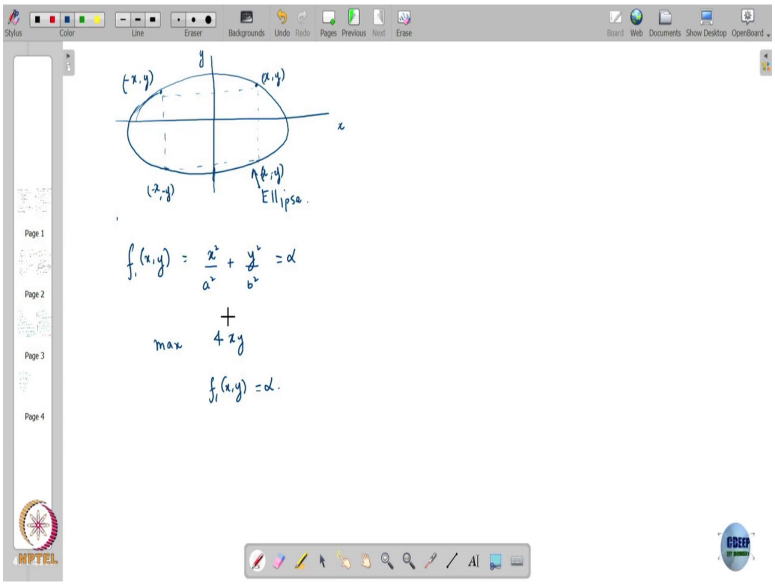
So, in general all these problems become tend to have the sort of form they have the form of where you are minimizing the norm of some W times A x minus b the whole square where, W is your weighted matrix is a matrix of width ok, this is the general form of this problem ok.

This was another example of a problem that is can very commonly used in practice in many different disciplines. For example, power system state estimation this is the standard model just. So, I can I wanted to demonstrate that as application of optimization over open sets ok.

Now, let us come to let us take this forward and let us see if there is there are problems other than problems that directly look like optimization problems over open sets, but we will somehow get reduced to those ok. And this surprisingly is actually a very general class.

So, I will we can derive a very general theorem about this, but that is not as revealing and as illuminating as doing an example. So, what I will do is to an example and then I will state the general result. So, hopefully we can at least complete the example in today's class.

(Refer Slide Time: 16:09)



$$f_1(x,y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} = \alpha$$

$$\max \quad 4xy$$

$$f_1(x,y) = \alpha.$$

So, the problem is the following you have suppose you are in 2 dimensions you have here x axis you have here y axis ok, and suppose we have this thing here this is a ellipse ok. It is given by this equation let us let me denoted by f 1 of x comma y this is given by x square divided by a square plus y square divided by b square equal to alpha suppose ok.

So, any point on this ellipse any x comma y that that lies on the surface of this ellipse must satisfy this equation and converse any point satisfies the equation you can located on this ellipse. The problem for us is to is simply to find the rectangle with maximum area that can be inscribed in this ellipse ok.

So, how do you find such a rectangle? What I need to do is, I act how do I define such a rectangle? Well, the rectangle can be defined in a very easily, what we will do is we will make for simplicity we will make sure that the rectangle alliance with the coordinate axis that we have chosen ok.

So, that is possible, so what I mean by that is suppose you take a point x comma y here. I will look at the corresponding point minus x sorry x comma minus y look at this point minus x comma y minus y and this point which is minus x comma y. Take these four points and let us look for the rectangle of this sort of form that has the maximum area ok.

So, what is the area of this rectangle? With this corner with these end points or these corner points it will be 4 x y 2 x times 2 y right. So, what we want to do is therefore, find x and y such that you maximize the function 4 x y, but then x and y should lie on the ellipse ok. So, which means that x and y must solve this ok.

Now, what does this mean? Now, what is the implication of this? So, if you look at this particular problem you are maximizing a function that is differentiable, but then you are not maximizing it over an open set, you have to maximize this only over those points that are on the ellipse. That are on the locals of that of this function f 1 of x comma y equal to alpha right.

So, we are so now, if you look at the set of points that form this ellipse this is actually a close set, this is not this is not an open set and that then therefore, become this become therefore, a problem that is not in the previous category ok; however, what I will show you is that it is possible to still reduce it to some something that we have seen before. So that is.