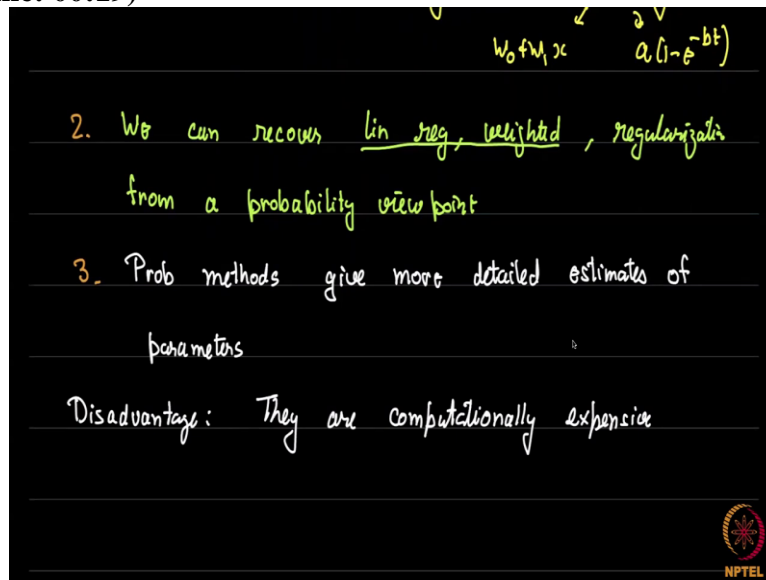


**Inverse Methods in Heat Transfer**  
**Prof. Balaji Srinivasan**  
**Department of Mechanical Engineering**  
**Indian Institute of Technology - Madras**

**Lecture - 36**  
**Maximum Likelihood Estimate**

(Refer Slide Time: 00:19)



Welcome back, we are in week seven of inverse methods in heat transfer. This week will continue from last week and if you recollect last week, we looked at probability theory, within that we had looked at Basic Probability Theory as well as Bayes theorem and the idea of probability distributions. Now what we will do this week is to combine these two ideas and come up with inverse methods that use Probability Theory.

What you will see that is that these inverse methods using probability ideas is actually a superset of our previous methods. one advantage that we will see is that they are not so, they are the same methods regardless of linearity of model. So, this is one major advantage of probabilistic methods that is they are the same methods regardless of whether your underlying model is linear or non-linear.

For example, if you remember when we had linear regression versus when we had non-linear regression, we had to use entirely different methods and entirely different iterative techniques, whereas whether the model is linear like  $w_0 + w_1 x$  or non-linear like  $a(1 - e^{-bt})$  in both these cases we will use exactly the same probability ideas. One other Advantage is we can see that we can recover many of our earlier methods.

For example, linear regression, weighted linear regression and our regularization from a probability viewpoint. In fact, one of those which is the first two I will be doing right within this video. So, from a probability viewpoint, we will be able to recover all these previous methods. A third advantage that you will see right now is, that probability methods give more detailed estimates of parameters.

What I mean by that is instead of saying that thermal conductivity is 50 watts for meter Kelvin, it will say it lies between 45 to 55 which is a characteristic a hallmark of probabilistic techniques. So, these are the three primary reasons within this course, that we will be looking at probabilistic techniques, the first being that they work regardless of the linearity of the model. We are able to recover all the earlier methods right from the probabilistic technique and they also give more detailed estimates of the parameters that we are interested in in solving in inverse problems.

The only problem is, so the disadvantage of probabilistic methods is that they are computationally expensive. So, because of this computational expense in many practical realms, it is not really possible to do a full-scale probabilistic model though we will see some approximations of this.

Now in general when we go to inverse methods or you go to the inverse methods literature you will see that a lot of the inverse methods literature is actually written in probabilistic language. So, that is one of the reasons for us to study these methods. some of the initial notation might be slightly confusing hopefully if you have gone through in detail last week's lectures you will be able to access this a little bit more easily.

**(Refer Slide Time: 05:20)**

### Inverse problem in a slab

1. Consider one-dimensional steady-state heat conduction in the slab. Estimate heat flux ( $q$  ( $W/m^2$ )) and boundary temperature ( $T_1$  ( $^{\circ}C$ )) using least squares regression (LSR). The experimental temperatures at various location are shown in Table 1. The length ( $L$ ) and the thermal conductivity ( $k$ ) of the slab are 70 mm and 14.4 W/mK, respectively.

Location of thermocouples (K-type)	x, m	Experimental temperature, $^{\circ}C$
1	0.01	15.46
2	0.02	14.59
3	0.03	12.66
4	0.04	12.55
5	0.05	11.57
6	0.06	11.42

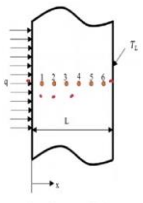



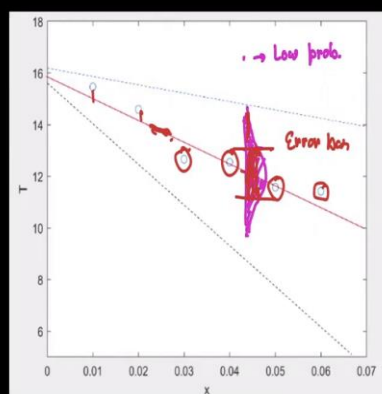
Fig. 1 Geometry of slab.



So, what we will do now is go back to our slab problem and formulate it as an inverse problem but in the probabilistic language. So, let us go ahead and do that. So, if you recall we were looking at this simple slab problem, again I am repeating the same problem again and again just. So, that you have some orienting case we had made six thermocouple measurements and we were asked to find out what is the heat flux which is coming in.

Given that you have not really given boundary conditions but we have just given some internal temperatures.

**(Refer Slide Time: 05:59)**



Assumptions

- The errors are additive  
i.e.  $y_i = \hat{y}_i + \epsilon_i$   
↑                    ↑                    ↑  
Measured        Model of reality        Error  $\epsilon_i$
- Error has zero expectation  
 $E(\epsilon_i) = 0$   
Multiple experiments  $\rightarrow \bar{\epsilon}_i = 0$

Now we handle this as you remember by doing a simple least square fit here are some couple of wrong guesses so, to speak. They are wrong not really but it is just that they look off even to our eye and we somehow came up with minimizing the least squares. Now what we are

going to do over the next few minutes is try to rederive you know why minimum of least Square.

So, our fundamental question right now is going to be why least square? is there some other way in which we can justify the fact that we are trying to minimize the square of the difference between our prediction which was  $\hat{y}$  and the ground truth which was  $y$  or the experimental method? So, this is what we did and we want to justify this and we will see that we can actually justify this particular problem via probability Theory.

So, what we are going to do is to make a series of assumptions about what is true about our experiment. So, what we notice of course is that there are all these thermocouples which are sitting here and they are all making measurements. Now the first assumption that we are going to make about not just this problem but any problem that we go ahead and solve is that the errors in this problem are additive.

I will explain what that means what we are assuming is that at the base of all this is some reality. So, we see this is what is measured but underlying it is some reality, which is our model of reality, but on top of it we are adding an error. So, there is an error on top of it. This is error or noise due to some factors that we either do not understand or we cannot control. So, we either cannot control these errors or we are not able to understand where these errors came from.

So, the Assumption here is at each point so, at each measurement location. So, we had six locations here, at all of these six locations, we have some error which is being added which is why our prediction need not look exactly like whatever measurements did. So, we do not expect something that goes through here, because actually we are trying to find we are trying to find a model for this noise and we are going to try to denoise.

So, as to obtain some actual reality so, that is the model that we are trying to get to and this is because this cannot be done uniquely is why we have an ill-posed or an inverse problem. Now we are going to make further assumptions about how this noise behaves. So, the second assumption here is that the noise the error has 0 expectation another way to say it is error for each one of these epsilons remember this notation is  $\epsilon$  what does that mean.

So, again physically you need to understand it, it does not mean that the sum of the errors here is simply 0, what it means is if I take multiple realizations if I do multiple experiments. So, let us say we have the slab. So, we have the slab and I did one experiment and this was the measurement that I got. Now I did another experiment. So, I might get slightly different numbers like 15.44, 14.62 etcetera. So, this is experiment one this is the only experiment by the way which we have access to.

But right now, we are thinking about what reality is because reality is not that whatever I measure is the truth but whatever I measure is the truth plus some error. This error will vary I mean just like when you take a slab and you measure its length, you would see that sometimes you'll measure it a little bit less and sometime in measure it a little bit more. So, similarly what we are assuming is that  $\varepsilon_i$  in our language from the last week is a random variable error is a random variable.

And the error has 0 expectation  $E(\varepsilon_i) = 0$ , that is sometimes you will measure more than the actual temperature and sometimes you will measure less than the actual temperature. And overall, over infinite experiments, you will get an average error to be 0 which is what is written here. that is the  $E(\varepsilon_i) = 0$ . The third assumption we are going to make is errors are distributed normally.

So, now that we have decided that the error is a random variable, we have to now make a guess about what its distribution is like. So, for our case I mean this need not always be the case for a general experiment, but for our case just like we did last time, we are going to assume that the error has a probability distribution and the probability distribution is a Gaussian. Now when you say Gaussian you have to give two things.

What is the expectation or  $\mu$  of that Gaussian and what is  $\sigma$  of the Gaussian. So, we are going to say epsilon I, that is error at the ith thermocouple is distributed normally. So, that we will say drawn from remember we had used this notation the last week. So, normal distribution which has mean or expectation 0 and its variance is Sigma Square. So, when I write 0 Sigma this means variance equals Sigma Square.

So, right now at least in the beginning we are assuming that each of these thermocouples is measuring error at a different variance. Now what would that mean what that means is this. So, if I assume that, this is the truth, this line here is the truth the red line here. What I am assuming is that each time I measure the temperature I am going to actually measure a Gaussian or I am going to select from a Gaussian around this.

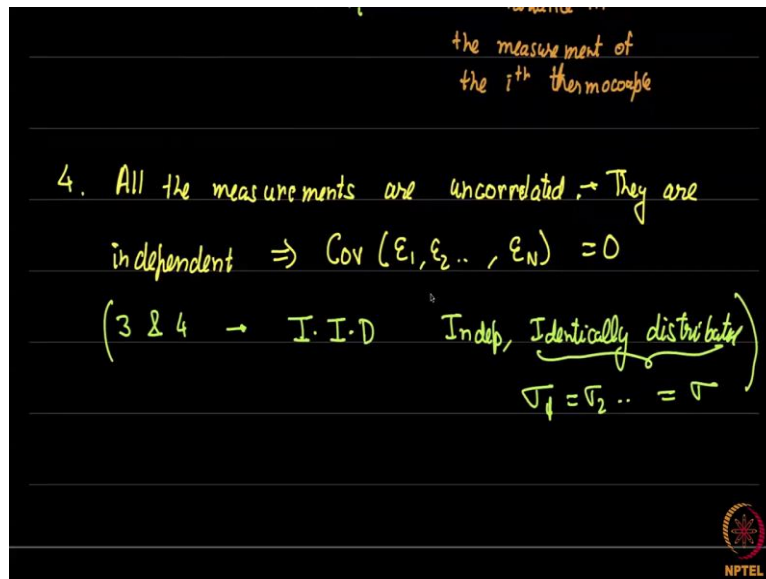
So, the temperature I actually measure could be anywhere in this range. Now as we know that the Gaussian has a low probability further and further away the chance that your error will be this high this is low probability. Chance that you lie typically you will lie between  $+2\sigma$  -  $-2\sigma$ . So, this you can draw as an error bar which you might have seen. So, 95 percent of the times when I make these various experiments that is I keep on repeating this taking the slab taking a different error etcetera.

95 percent of the times I will lie somewhere here a few of the times you will go outside of this. So, that is the Assumption we are making. So, that is a reasonable assumption. Now notice that something interesting happens here. what happens here is that we actually account for in this model the fact that there is a variation. So, we are actually accounting for the fact that we have measurements that go away remember in the function approximation we never talked about why is it going away in the function approximation he said my values are just farther away.

You know I guess what I will do is, I will do the best I can and fit something that goes in between that is not what we are doing here. We are actually giving a process we are giving in detail not in much detail but at least in little bit more detail a process which accounts for the fact that there is going to be variation from experiment to experiment and person to person but you still want the parameters of this line.

So, that is the key here in the probabilistic approach modelling the error itself as some probability distribution and in our case a Gaussian. So, let us move further from here.

**(Refer Slide Time: 15:39)**



So, if we write this  $\epsilon$  in detail. Now we know what the normal distribution is. So, we can now write the PDF I am going to call it  $f$ ,  $f$  corresponds to the PDF of the error,  $f(\epsilon_i)$  at this point remember what it means, physically all it means is I am finding out what the shape is that is if I randomly pick something and say um the error is so, much what is the probability that the error lies in this range or the error lies in this range, you can actually give a number by integrating this function.

So, you must again revert back to the previous week in case you do not understand what this

means. So, the Gaussian distribution here is  $\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\epsilon_i^2}{2\sigma_i^2}}$ . Now why is this Epsilon  $i$  square

because ah Nu is 0. So, it is simply a Gaussian it is not a normal what is called a univariate Gaussian it is not any 0, 1 which we also say as the standard normal distribution because Sigma is not necessarily 1 in this case.

Now what does Sigma represent this of course represents the variance in the measurement of the  $i^{\text{th}}$ , in our case thermocouple, it could be anything else of course the formulation is General. So, the  $i^{\text{th}}$  thermocouple that we have in our slab each of them might be from different places and when you statistically measure what happens let us say you take 100 measurements each from here you can actually do it inside the slab or outside the slab.

Or wherever it is you repeat those measurements you will see each one of them might have a slightly different variance if you recall earlier from this course, we did that a data set also of

key sizes if you remember. And I had given that as an example of a weighted linear regression where each measurement had a different uncertainty. So, now we are explicitly accounting for accuracy of the instrument.

So, in this case basically this is a measurement of the accuracy of the instruments typically people use the term  $\beta_i$  equal to  $1/\sigma_i^2$  and  $\beta_i$  is called the Precision but I will stick with  $\sigma_i$  for now. I also use  $\lambda_i$  a few times when I did weighted linear regression. So, now we actually have errors you I can call this three a because this is a direct consequence of three and the next keen thing is that we are going to assume that all the measurements are uncorrelated.

More importantly I am going to make a stronger assumption I am going to say, in fact that they are independent, they are not just linearly and correlated they are independent. Of course, if they are independent this means covariance of  $\varepsilon_1, \varepsilon_2$  all these variables how many hours 0 but remember covariance is 0 does not mean independent but independent does means they are covariate.

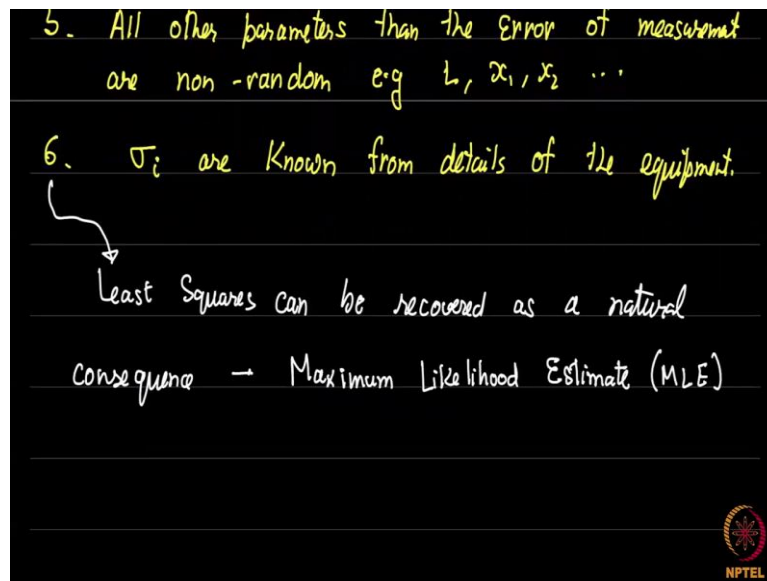
More strongly since we have made these assumptions three and four put together say that these measurements are what is called I I D. IID would be independent and identically distributed, strictly speaking we can use this in two ways we can say identically distributed but with different sigmas but generally identically distributed would mean that all  $\sigma$ ,

$$\sigma_1 = \sigma_2 = \dots = \sigma$$

So, we can use this for now for this derivation, initially I will just assume different sigmas and we will see the special case where all the sigmas are the same. but the reason I use this term was for you to recollect this thing IID which means independent and identically distributed. So, the final assumption we are going to make for the final two assumptions in some sense slightly related assumptions is.

**(Refer Slide Time: 20:54)**





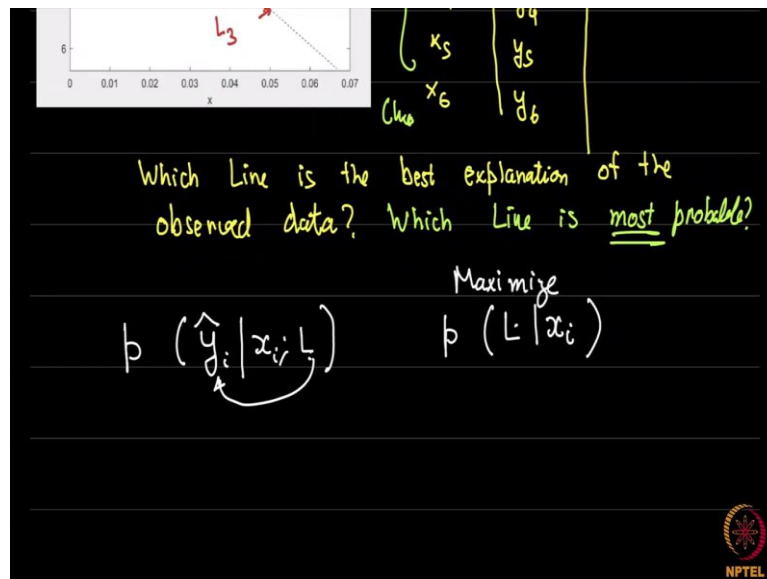
All other parameters than the error, well the error is not a parameter but everything else the error of measurement other than that are non-random. So, for example we are going to assume that the length of the slab is fixed. Length of the slab the locations  $x_1, x_2$  all these are not random parameters we actually know the locations. Obviously when you make an actual measurement these  $x_1, x_2, x_3, x_4, x_5, x_6$ , these are not actually fixed there are errors that we can account for those as well.

But all of it is in some sense put together and kept in this one single parameter  $\varepsilon_i$ . So, let us just make that assumption and one final thing which is also related in some sense like I said is  $\sigma_i$  are known this is not something I need to write explicitly but let us assume that the variances are known somehow from details of the equipment.

So, when we say measurement error that is not something that we are trying to figure out that is something that is given to us. It said that the thermocouple accuracy is so, much and so on and so forth. So, if we do all this. So, what, so, if I know all this so what. So, from here we can show that we can actually recover least squares I should not put an arrow here I should actually put it here.

From all these least squares can be recovered as a natural consequence. What does that mean? we are going to see something called the maximum likelihood estimate, which I will explain shortly. And this is another way of handling inverse problems or another way of looking at this Bayes problem. So, this is called M L E. So, this Maximum Likelihood Estimates. So, let us see what we mean by maximum likelihood estimate.

(Refer Slide Time: 23:49)



Let us look at these three lines. Line one we will call it L1 line L2 and line L3. Now original data was like this we had some  $x - x_1, x_2$  let me write this fully  $x_3, x_4, x_5, x_6$ . We had some  $y$  data  $y_1, y_2, y_3, \dots, y_6$  and we want to figure out which one of these three lines is the best explanation which line, So, this is the way we are going to post the question is the best explanation of the observed data like we discussed right at the first week this is somewhat like a detective problem.

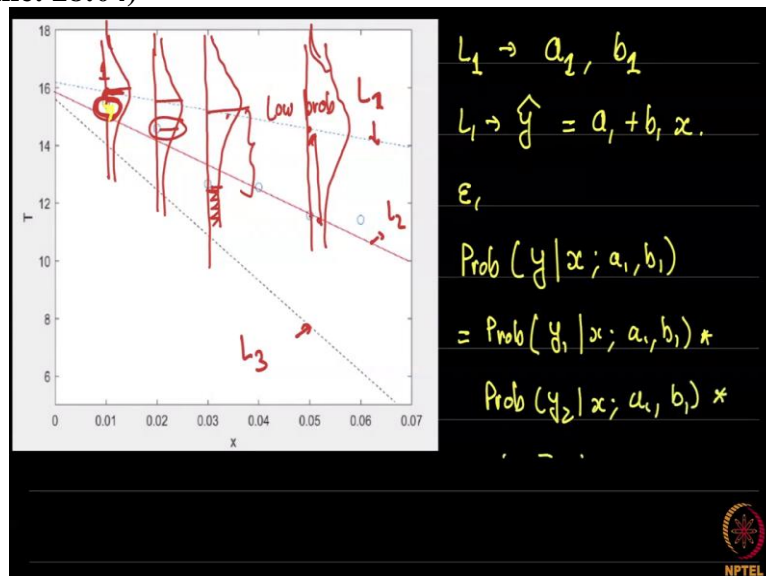
So, let us say somebody goes, sees a murder or does not see a murder, sees a dead body and they look at all the clues which are around them. So, this we treat as if they are Clues at  $x_1$ , I saw  $y_1$ , at  $x_2$ , I saw  $y_2$ . Now I have different possibilities someone L1 could have killed, L2 could have killed, L3 could have killed. Now if L1 killed how well does all my observation fit the fact that L1 did it. If L2 generated this data how well does all my data get explained by L2 and if L3 did it how well does it explain it.

So, this is the way we are going to look at it look at it and here is where probability Theory really helps. when we uh looked at least squares we had like a very rough tool we just said the gap between L1 and this data is too much. So, maybe that is not. So, great but that is still too crude whereas the what we are going to ask is which line is most probable. So, please notice this word most probable, we will modify the meaning of this through some of the videos that we do.

We will see three level different levels of the same idea. The first level is this very simple idea, I have already observed this data this was the data, that I showed you right at the beginning I have already observed concrete numbers here. which model best explains the data? So, how do we write it in probabilistic language we want to say I observe I am already given  $x$ , I want to maximize  $y$ -hat. The  $y$ -hat that fixes this  $x$  for some given guesses.

So, given some line so, here is the observed data now each  $L$  will correspond to some  $y$  hat. In fact, if you want let me write it more clearly, I want to show this I want to maximize there are different lines for different for all the given  $x$ 's all these lines each of them has a different probability I want to maximize that probability. So, I am going to write it in some technical term. So, let us see how each one of these data points is generated by for a given line. So, the generation process will be like this.

**(Refer Slide Time: 28:04)**



First, we are given some  $x_i, y_i$  in our case  $i$  goes from 1 to  $n$  where  $n$  is 6. Now I am going to choose some hypothesis line. This is very much like how we did gradient descent also. So, this hypothesis line is chosen by choosing  $w_0$  and  $w_1$ . So, for example I choose one  $w_0$  and  $w_1$ , I get this line I choose one  $w_0$  and  $w_1$ , I get this line and some other one I get this third line. So, we choose that.

Now third step this is the key step randomly assigned some error to it. This is how each experiment takes place remember we had initially modelled our entire process of generating this data as  $y_i = \hat{y}_i + \epsilon_i$ . So, there is some truth which is our line and then this was just

randomly assigned this is how each experiment takes place. So, you do this and you pick some error.

And how do you pick this error, you pick this error by drawing from this distribution imagine this is like I said this is just like a ball or this is like a bag and you are just drawing some ball from it and it will give you a value of Epsilon. How we do this in practice we will discuss later this week but this is the probabilistic process. So, we are modelling reality as if it is doing this. So, keep. So, evaluate the probability that we will observe the given data for the choice of parameters into in step two.

So, this is the way it happens in practice. So, we come here let me repeat this graph suppose I want to evaluate the probability that L3 explains this data. So, this is the way I will do it, I will come here, I will come to 1, 0.1 and then I will say line one right or sorry yeah line one. So, I will come to line one let us say it has the parameters let us call it  $a_1$  and  $b_1$  instead of  $w_0$  and  $w_1$  this is easy. So, line one then generates  $\hat{y} = a_1 + b_1x$  all right.

Now once  $\hat{y} = a_1 + b_1x$ . Now I know what the probability of this point is how do I evaluate it we will. Now draw the probability distribution of  $\varepsilon_1$ .  $\varepsilon_1$  is a Gaussian which has mean 0, that is centred around right around line one and some given Sigma. Now all we need to do now is to evaluate the possibility, that you will get this point here for that given  $\varepsilon$ .

Similarly, I will come to 0.2 and I will draw again the  $\varepsilon_2$  distribution here and I will find out the possibility that this point will come similarly here you will. Now see if I put it centred around line 1, this becomes very low probability. you see the Tails as we saw in Gaussians and if we come to this point for example again  $\varepsilon$  is centred here, this is again very low probability.

Now since these events are independent, we can say that probability of observing the data given these locations for these parameters  $a_1$  and  $b_1$ , so this language I will explain once more the probability of observing this is a product of the probability that the first point is observed because these are independent, the first point is observed with the same parameters multiplied by the probability that the second point is observed given these parameters so on and so forth.

**(Refer Slide Time: 33:50)**

Since the errors are independent,

$$p(y|x; a, b) = \prod_{i=1}^6 p(y_i|x_i; a, b)$$


pdf ↑ Whole dataset

$$y_i = \hat{y}_i + \varepsilon_i$$

↑ Fixed ← Random

$$P(y_i|x_i) = p(\varepsilon_i|x_i)$$

One data pt

$$= \frac{1}{\sqrt{2\pi}\sigma_i^2} \exp\left(-\frac{\varepsilon_i^2}{2\sigma_i^2}\right)$$


So, let me write this in a mathematical language since the errors are independent. We can write that probability that we will see the specific measurements we made or actually the range of measurements if we want to say it a little bit more precisely but this is actually just the PDF. So, for observing the PDF for the given input data with some choice of parameters. So, let us say we made some choice again I will use a and b maybe just for now.

So, that we do not get confused with the subscripts is the product, So, product we denote by  $\prod$ , if you remember just like  $\sum$  is for some the product over all six measurements or n measurements in general of, I should have called this  $P(y_i|x_i; a, b)$ . Now this  $P(y_i, x_i)$  if you remember  $y_i = \hat{y}_i + \varepsilon_i$ . So, since only this portion is the random portion and this portion is fixed the  $P(y_i|x_i)$ , I am not proving this but you can imagine.

This fairly straightforward thought process is the same as giving the error. So, once I give you a point that is the basic model and I give you an error obviously you can predict what was measured so the same thing here. So, all it says is the only random portion in our measurement is the fact that there was a random error. So, this tells you that this value here is what we wrote

earlier  $\frac{1}{\sqrt{2\pi}\sigma_i^2} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$ .

So, before we get lost in the symbols all we are saying is this if I want the probability or the probability distribution function at this point given that the ground reality was this, then this point is going to have a very low PDF because it is far away power from the mean. So, from

the mean here which is just this line. now if on the other hand I took this model the L2 model this point would become more probable.

That is the error required to reach this point is fairly low, whereas the error required to reach this point is fairly high and we know that the error is normally distributed. So, it cannot be that high or at least high errors are not that likely, if you keep on demanding from a data set that every point is coming because there was a lot of noise. So, we can always explain saying hey your model is bad, let us imagine you give this as the model of reality.

Or let us take this because this is less messy L3 is the model of reality and I want to explain these six points which are sitting here and I say oh why is your data. So, far away from reality you say because there was a lot of noise. At the same time, you also say that my instrument's accuracy is very high its variation is slow. So, then you will say maybe your model itself is not, this is a much better model with low noise.

This might be a decent model with high lines, but this is a much better model. let us come to this again.

(Refer Slide Time: 38:26)

Handwritten notes on a blackboard:

$$\Rightarrow \text{Max} - \sum_{i=1}^6 \frac{\epsilon_i^2}{2\sigma_i^2} \quad \Rightarrow \text{Min} \sum_{i=1}^N \frac{\epsilon_i^2}{2\sigma_i^2} \rightarrow \epsilon_i = y_i - \hat{y}_i$$

$$\text{Minimize } \frac{1}{2} \sum_{i=1}^6 \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \text{ to maximize prob}(y|x)$$

Likelihood  
How likely is y given x?

NPTEL

So, this is for one data point and this is for the whole data set. why is that because if each data point has this error, then the whole data set as I have written before is product of individual data points which turns out to be in this case,

$$\prod \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-\varepsilon_i^2}{2\sigma_i^2}\right)$$

But this still does not look like anything like B squares.

Now here is where a trick is so, we take log of both sides. So, I say,

$$\log P(y|x) = \sum_{i=1}^6 \ln \frac{1}{\sqrt{2\pi\sigma_i^2}} + \sum_{i=1}^6 \frac{-\varepsilon_i^2}{2\sigma_i^2}$$

This is just by taking log there should be a logarithm here. So, if I want to write that log here is ln. So, that is what I am saying 1 by square root 2 pi Sigma i square you can make it minus but it will shortly become irrelevant.

Now since a log is a monotonic function if I am finding out which parameters maximize probability, this is the same as asking which parameters maximize log of probability because if log is maximized, probabilities also maximized. So, this means we need to maximize this plus this. Now this is fixed, why of course 2 pi is fixed we have no control over it and sigma is from the instrument we cannot do anything about it.

So, what you can do is you maximize,

$$\text{Max} \sum_{i=1}^6 \frac{\varepsilon_i^2}{2\sigma_i^2}$$

Since there is a negative here, we can say this is the same as minimizing,

$$\text{Min} \sum_{i=1}^N \frac{\varepsilon_i^2}{2\sigma_i^2}$$

Now I am going to make it 1 to n just for n points but what is  $\varepsilon_i$ ?  $\varepsilon_i$  remember was,

$$y_i = \hat{y}_i + \varepsilon_i$$

which means  $\varepsilon_i$  is simply the error,

$$\varepsilon_i = y_i - \hat{y}_i$$

So, what we are saying is minimize  $\frac{1}{2} \sum_{i=1}^6 \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$  let me say to maximize the  $P(y|x)$ . Now this maximizing this y given x is called likelihood. What it means is, how likely is y given x,

how likely were you to observe this data set or these temperatures at these particular locations that is what it physically means.

(Refer Slide Time: 42:45)

The image shows handwritten mathematical derivations on a blackboard background. At the top, the log-likelihood function is written as  $\log p(y|x) = \sum_{i=1}^6 \ln \frac{1}{\sqrt{2\pi} \sigma_i} \exp\left(-\frac{\epsilon_i^2}{2\sigma_i^2}\right)$ . The terms are circled, and the expression is simplified to  $\sum_{i=1}^6 \ln \frac{1}{\sqrt{2\pi} \sigma_i^2} + \sum_{i=1}^6 -\frac{\epsilon_i^2}{2\sigma_i^2}$ . Below this, it asks "Which parameters maximize probability" and shows the equivalent maximization of  $\log(p)$  as  $\Rightarrow \text{Max} -\sum_{i=1}^6 \frac{\epsilon_i^2}{2\sigma_i^2}$  and  $\Rightarrow \text{Min} \sum_{i=1}^6 \frac{\epsilon_i^2}{2\sigma_i^2}$ . A boxed equation states: "Minimize  $\frac{1}{2} \sum_{i=1}^6 \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$  to maximize prob(y|x)".

Now notice this for all sigmas being the same, this immediately gives us minimize  $\frac{1}{2} \sum (y_i - \hat{y}_i)^2$ , of course, the sigma square is equal element this is of course the least square loss function. That is performing least Square is equivalent to maximum likelihood estimate. So, you might get a little bit lost in the math, even though if you review it the math is actually just two steps. So, what it means is this see amongst all the infinite lines that could have fit this data amongst all the infinite lines that could have fit this data only one of them has the maximum probability given this data set.

What that means is each of these lines has a certain probability that it explains this data. because when I keep the line here it means this data has a probability, this data has a probability, this data has a probability. And I need to multiply all these six probabilities in order to get the probability that this line could have generated this data. Now given all this given all the infinite lines there is one line which gives you maximum probability of explaining this data of Maximum likelihood of this data.

Now that line also happens to be the same line that minimizes these squares. So, therefore minimizing the least square is equivalent to maximizing the likelihood. Now what happened to this, this if you recall is exactly what we did for with this is weighted least squares. That is if each point each of the points here had a different variance, that is let us say there is a point with very high variance like this one.



And that is only because the thermocouple was very noisy in that case the probability of this and the probability of this might actually be very close to each other even though or even this if this is something which is highly accurate, you actually have to blow up the difference. And if this is less accurate you have to actually reduce the difference. And as we discussed in weighted least squares, we need not weight each point the same because each point has a different variance.

So, from Maximum likelihood we are able to obtain both normal least squares as well as weighted least squares in one shot. Now this of course was a little bit of theory and it turns out this is not the only way or this is not the end of how to use probability theory. In the next video we will say that this is fine but what if I already have an idea of what this line is supposed to look like. What if I have in other words a prior.

Remember the prior example that we did in the covid case in the last week that strongly affects what our final estimate is. So, we will try to incorporate ideas for example you are estimating a thermal conductivity of a material you would like to incorporate the idea that you already know that the material thermal conductivity should look like the thermal conductivity of let us say steel in that case that should be incorporated into the actual information.

And we will see how probability can do that and what the corresponding equivalent of what we already did is. So, let us take a look at that in the next video, thank you.