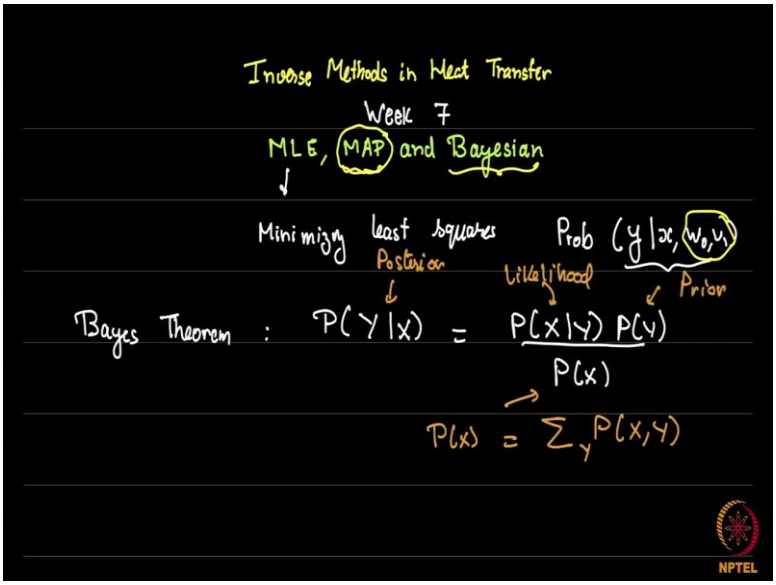
Inverse Methods in Heat Transfer
Prof. Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology - Madras

Lecture - 37
MLE, MAP Estimates

(Refer Slide Time: 00:19)



Welcome back, this is week 7 of inverse methods and heat transfer. In the last video you saw how to show that the maximum likelihood estimate using a probabilistic technique gives us the same result as minimizing the least square. So, this was the connection between the current approach that we are using which is a probabilistic approach and the functional approach that we used earlier.

Remember in our probabilistic approach all we are trying to do is to build a probabilistic model of how y the output depends on x the input and also whatever parameters we have w0, w1 etcetera. And then we try to maximize the probability of this happening what is called MLE or Maximum Likelihood Estimate and we try to find out the parameters that we can vary in order to accomplish this.

So, this is what we saw in the last video. Now in this video what we are going to do is to go two steps further, first step is what is known as maximum a post periodic and the next step is what is called a full Bayesian in estimate. So, I will do this in some detail and once again you will see an interesting correspondence with what we had done earlier in our function approach.

And the Bayesian approach I will just give you a brief outline of because it is in some sense too complicated for this class.

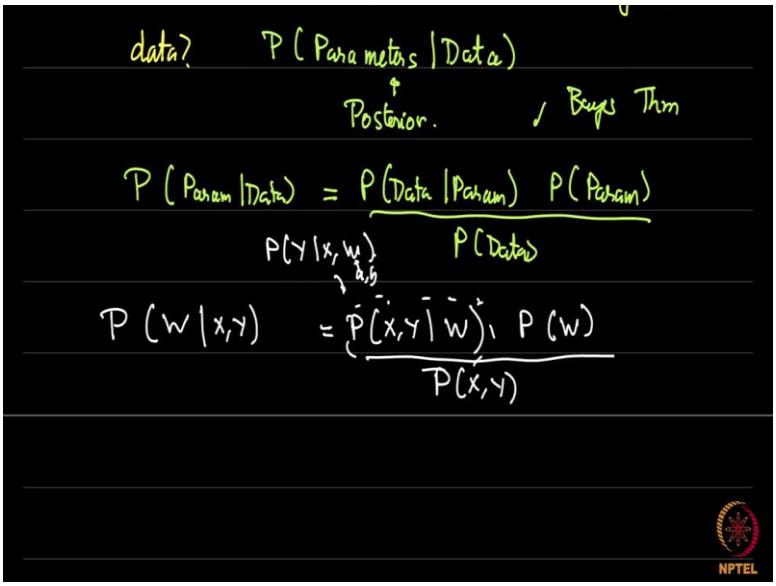
In advanced forces you can do it, it is too complicated for this class to do theoretically we will however see how to make Bayesian estimates computationally in the coming videos within this week itself. So, let us get into this topic. So, remember all of what we are going to discuss depends on what is known as Bayes theorem. we saw that earlier in the last week base theorem simply says that,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

And of course, as we saw in the examples, P of x itself is given by marginalizing over the joint distribution for all possible values of y. So, you basically understand that in case we have a case like cancer versus no cancer then we want to know do I have cancer given this test, then I sum up over all possibilities. That will I get a positive with cancer will I get a positive without cancer etcetera.

You sum up over all possible y's and that is how you get P of x. Now as it turns out and as we will see when we come to the Bayesian estimate, this quantity can often be really hard to compute. this is usually something which I called the prior, this is called the likelihood and this is called the posterior. Now of course I can switch x and y also and that is the form in which we are going to do it right now.

(Refer Slide Time: 03:42)



Now instead of using x and y which is what I did in the previous video as well, I am going to use something that is a little bit clearer. So, I am going to say data and parameters. So, typically data is called y and parameters we had called w. So, we are going to see it this way actually data itself was a combination of x and y, but let us now come to this thing of how do we estimate the parameters given the data.

Now what was MLE, I am going to rewrite MLE in this form. MLE said what is the parameter that maximizes the likelihood of observing this data. So, we saw this last name this one. So, the question really was found out probability of the data given the parameters and maximize this probability or maximize we usually did log of probability and this is what gave us the least square estimate again.

Now but this is not the right question as you can see. So, this is what was MLE right. This is not the right question because actually what we are asking is the other question. what is the best are most probable parameters given the data. So, if you see this the dependency is flipped, we say I am given the parameters, how can I generate or which is the set of parameters that generates this data and the most probable sense whereas here we are asking for a different probability distribution, probability of parameters given the data.

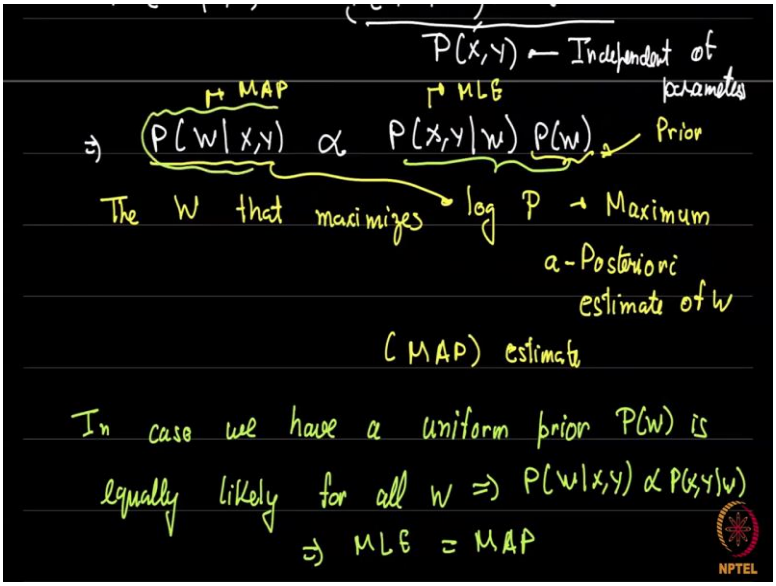
So, this like I said is called the likelihood and this is what is called the posterior. Why? So, let us write it this way probability of param given data is probability of data given parameter multiplied by probability of parameter divided by probability of data, this of course is Bayes Theory. Now let us write this in our old xy notation. So, I am going to write this as,

$$P(w|x,y) = \frac{P(x,y|w)P(w)}{P(x,y)}$$

Now you might remember earlier that in the previous video I had just written this as P of y given x, w or, a, b. I am writing it in this way because this makes the dependence a little bit clearer. There is some formal difference between these two expressions but we are going to ignore it because that will get too much into Probability Theory and why we can use it in this form but right now casually you can think about this.

I have given the data or I am given the parameters and then I can find out the probability of a given data set, here I am given the data set and I want to find out what is the best parameter that fits it. So, this multiplied by probability of w divided by probability of observing the data.

(Refer Slide Time: 08:14)



Now we can therefore write that,

$$P(w|x,y) \propto P(x,y|w)P(w)$$

where we have noticed the fact that this is independent of the parameters. Now if you find out that w that maximizes log of this probability or just the probability, this is called the maximum a-posteriority. That is in fact I should put this in small maximum a posteriority or a posteriori estimate of w other words this is called MAP, MAP standing for maximum a-posteriori estimate of w.

So, this is called MAP whereas if you maximize this you get MLE, Maximum Likelihood Estimate this is these gives you MAP. Now what is the difference between these two. the difference between these two is this quantity which is the prior. we saw in the Covid example that the prior significantly affects what probability you get here. So, in case I am going to again say this a little bit casually.

In case we have a uniform prior. So, this would mean that P of w is equally likely for all w. So, this is not quite possible, because we have w has an infinite range, but let us assume that that was the case in that case you will simply get,

$$P(w|x,y) \propto P(x,y|w)$$

And there is no other dependence on w, this will give you MLE is the same as MAP.

So, physically what we mean is this MLE and MAP are the same, if we have no information about w.

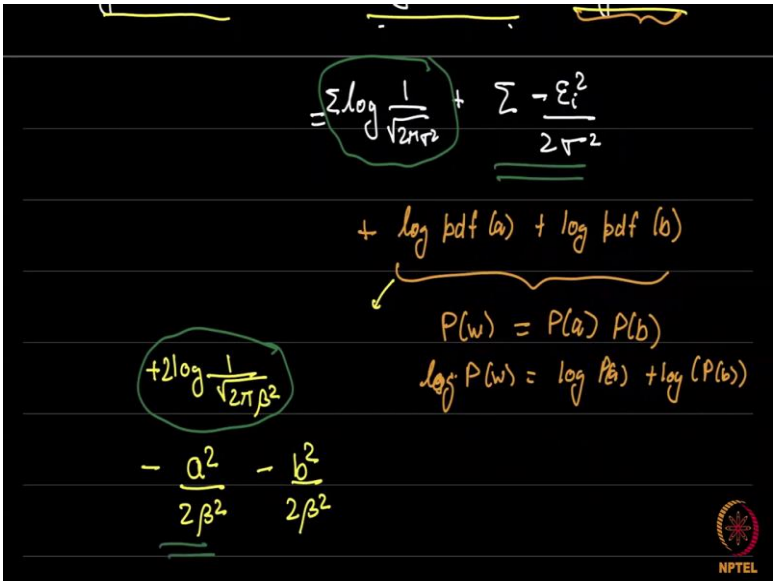
So, suppose you are estimating the thermal conductivity of a material, suppose you are estimating something else about that material or you have an MRI scan and you are trying to estimate the image and if you have no clue about what the parameters you are trying to find out are or you have no clue about their range, then you can simply do a least Square. So, this means least squares is the best estimate both map as well as MLE, but in case we have a prior that is we have some idea.

So, let me give you an idea of a prior. So, let us say we are dealing with two parameters, a and b or yeah let us just take two parameters, a and b and we know that these parameters are Gaussians or we know that they have means around zero and they have some Sigma. So, let us say I have these two parameters and we know that these are random variables. So, for example let us say we have conductivity and we know that the k is one of the parameters we are estimating.

And it is let us say we know it is made of steel and it is going to vary between let us say 50 watt per meter Kelvin to 55. If we have some such estimate then you know that you can sort of fit, I again request you to go back to the examples we did in the previous week, you know for example for the pipe etcetera. So, all these cases have some need and some standard deviation. So, the special case that I am taking here is let us say we have two parameters where the mean is zero.

And the standard deviation is instead of calling it Sigma let me call it beta. So, if we have this then there is a prior. So, then there is a prior and this prior is going to affect this a-posteriori PDF.

(Refer Slide Time: 13:48)



So, we are going to call this first the prior probability distribution function. So, the prior probability distribution function for a, for example let us say probability distribution function of a, PDF of a is,

$$pdf(a) = \frac{1}{\sqrt{2\pi\beta^2}} e^{\frac{-a^2}{2\beta^2}}$$

So, this is the $pdf(a)$, because now a is a normal variable with mean 0 and standard deviation or variance beta square.

$$pdf(a) \sim \aleph(0, \beta^2)$$

So, similarly PDF of b is also in fact exactly the same thing, because I have kept the same parameters for it. we can change this and make it different parameters too, that is not a big deal in fact we have asked one such question within the um exercise problems. So, PDF of b is,

$$pdf(b) = \frac{1}{\sqrt{2\pi\beta^2}} e^{\frac{-b^2}{2\beta^2}}$$

So, now we take,

$$\log P(w|x, y) = \log P(x, y|w) + \log P(w)$$

Why am I taking log I am doing exactly the same thing that we did in the last video, what I want to do is to maximize the posterior probability, instead of that I maximize log posterior probability, as you can see this log splits nicely also into addition. Now if you remember we had actually already calculated this. this calculator relation was Sigma of $\frac{1}{\sqrt{2\pi\sigma^2}}$. So, that should be $\sum \log \frac{1}{\sqrt{2\pi\sigma^2}}$

We can simplify this but I am skipping that step and plus you saw Sigma of log of the probabilities of e bar minus s square we call it Epsilon Square Epsilon Square and 2 Sigma i square. And for now, we will take the simple case where we assume all sigmas are the same for the new term the new term being due to this. So, if you remember there are two terms here $pdf(a)$ and $pdf(b)$.

So, I will just say,

$$\log pdf(a) + \log pdf(b)$$

why is that because the probability of w we are assuming that A and B are independent parameters all these are assumptions which need not be true. This depends on the fact that,

$$P(w) = P(a) + P(b)$$

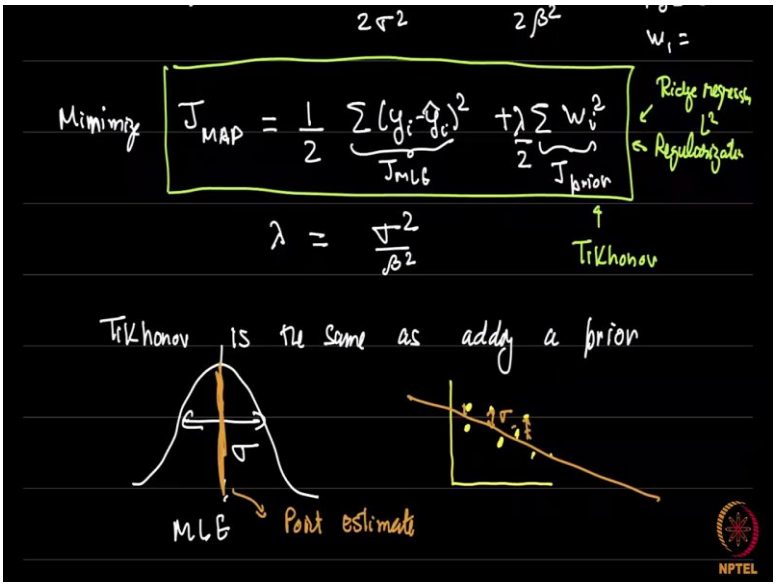
And of course, that would mean that,

$$\log P(w) = \log P(a) + \log P(b)$$

So, these terms let me write that out just these terms then come to plus, notice $e^{\frac{-b^2}{2\beta^2}}$. So, I will first put the constant term $2\log\frac{1}{\sqrt{2\pi\beta^2}}$ that handles the constant plus the next term which was $-\frac{a^2}{2\beta^2} - \frac{b^2}{2\beta^2}$. So, when I want to maximize this these terms the terms that I am circling right now these do not matter because these are constant terms.

So, this term does not matter this term does not matter the only terms that matter is these three.

(Refer Slide Time: 18:42)



So, we will maximize $-\frac{\sum(y-\hat{y})^2}{2\sigma^2} - \frac{a^2-b^2}{2\beta^2}$. So, you can. Now write this as minimize J. So, I am removing the minus I will put,

$$J = \frac{1}{2}\sum(y_i - \hat{y_i})^2 + \frac{\lambda}{2}\sum {w_i}^2$$

where Lambda is equal to $\frac{\sigma^2}{\beta^2}$.

Basically, if you multiply this equation by Sigma Square fully you get this. Now if you notice this term is exactly what we had, for what we call Ridge regression or regularization remember we call this L2 regularization it was also Tikhonov. So, we actually recovered the fact that Tikhonov regularization is the same as adding a prior. Now you can do more sophisticated things to get the kind of stuff that we got for Levenberg-Marquardt etcetera.

But I am not going to do that I mean you will basically have to show that there are covariances etcetera. So, I am not getting into that but the interesting thing here is, that we naturally recovered the idea of regularization through a maximum a posteriority. So, we can call this J MAP if you wish is J MLE plus this term which is J prior. Now what does this prior mean?

This prior basically says that depending on this beta square and the relative certainty of the data we have collected and the relative certainty of the prior that we have, what is the prior say in this case we are saying I do not want either a or b to be far away from zero. So, the more and more beta you give you more and more leeway you give for a and b to b let us say you have something in this sort it says within reasonable noise 95 percent of the data can be in a large range.

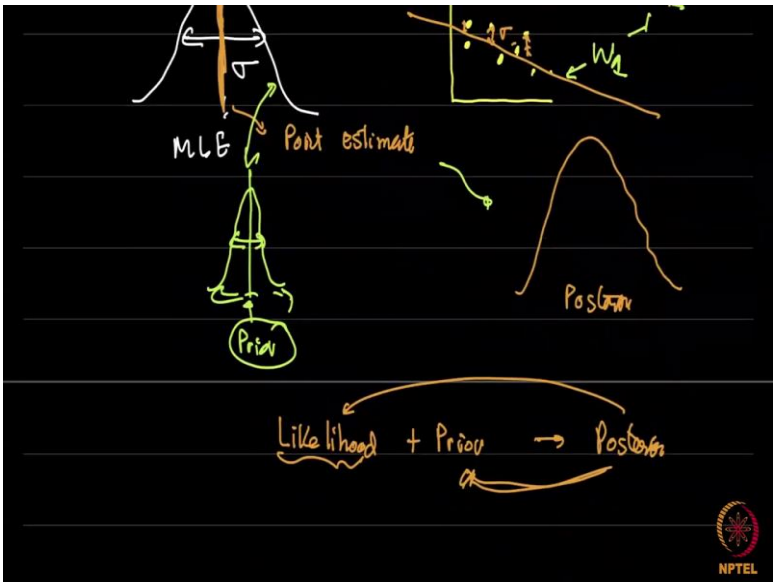
So, as you increase beta has increased beta, you can see that Lambda decreases that is the weightage of the prior decreases. but as you decrease beta the importance of the prior increases. So, another way to see it is, if this is the MLE distribution. So, the MLE distribution or the data distribution is remembered, this is with Sigma. this tells you what would be the prediction for a particular parameter let us say an if the only thing it wants to do is to fit the original data.

So, let us say the original data is like this and all you want to do is to fit some best fit line. So, once you want to fit, this best fit line you will account for all the errors here which are sigmas and you will give a distribution. You say ok my parameter lies with a particular range but the

most likely value of the parameter is the one that fits, this least square thing the best, this is called a point estimate.

I told you this earlier Point estimate means we are giving only one value for this w 1 or w 0 or both as we estimated.

(Refer Slide Time: 23:37)



Now apart from this we have another piece of information which is the prior. The prior says you might get some value of w1 here which is the slope. But I have a prior requirement that the slope should not lie out of a particular range or the more and more you want to push it out of this range. Let us say you are estimating the thermal conductivity of a material and it is steel and your best fit actually happens to say that thermal conductivity has to be very low it has to be one.

Now this is a bad idea because you know the material is Steal and it is way it has ah you know it has very low manufacturing probability that thermal conductivity is going to be extremely low. So, we want to incorporate this prior information. So, what we do is this prior information affects the likelihood and put together it gives you a posterior which is somewhat of a balance between the two.

So, you can say that likelihood plus prior leads to posterior. If the likelihood is strong then the posterior distribution will strongly bias towards the likelihood, if the prior is strong the posterior would strongly bias towards the prior. what does that mean in our Sigma and beta

terms. So, let us say the likelihood is strong means Sigma is very low. Sigma is very low means, I have very low errors in what is happening here, my thermocouples are really accurate.

Then if Sigma is low, Lambda becomes low. if Lambda becomes low, this term dominates and you would like to minimize this strongly and therefore your prior your posterior will strongly bias towards the likelihood. Similarly, if beta is low, then this term becomes very high this term becomes very high, your map will strongly bias towards the prior. So, what we have right now is in this video we have shown theoretically that regularization is the same as incorporating some prior information about a material or about the parameters the via regularization.

So, the regularization that we did earlier could be rederived. We also saw that the posterior is affected both by the prior as well as the least square estimate that sits there. Now in the next few videos this might all be theoretical and I have a couple of small theoretical questions within the exercises, but in the coming videos we will actually see how to make this computationally operational.

So, this cannot be calculated theoretically for most cases we will come and see how to do this computationally using a chain of techniques or using a set of techniques called Markov chain Markov chain Monte Carlo techniques or generally Monte Carlo technology. So, I will see you in the next few minutes, thank you.