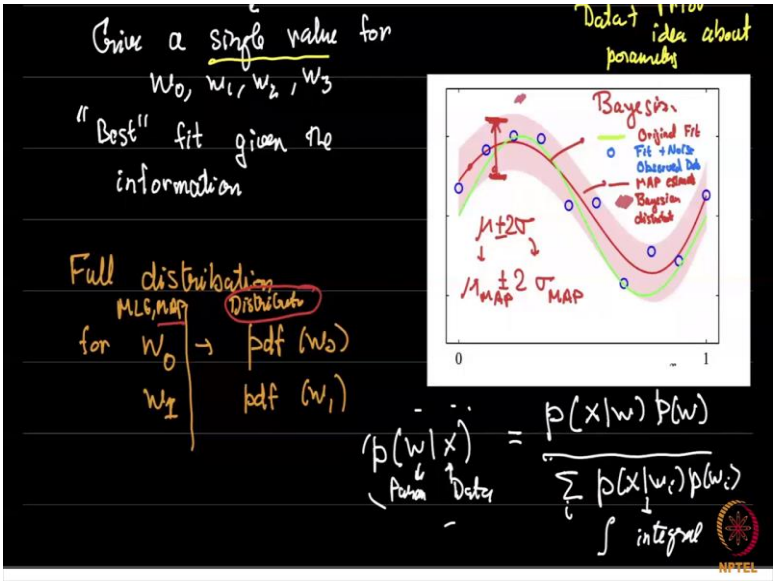
Inverse Methods in Heat Transfer
Prof. Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology - Madras

Lecture - 38
Introduction to Bayesian Methods for Inverse Problems

(Refer Slide Time: 00:19)



Welcome back, we are in week seven of inverse methods in heat transfer. In the last few videos, we saw these methods called maximum likelihood estimation as well as maximum a-posteriority estimation. Both these are what are known as Point estimates I will explain shortly what that means. In this video we want to see a couple of things, we want to see what Bayesian methods are I had actually said I would talk about that in the last video but I forgot to do that.

So, we will go from MLE to MAP to Bayesian in this video and I also want to give you an example of how you can use a computational method to calculate these. So, we are going to look at a very simple computational method within this video. we will turn it into a slightly more sophisticated computational method later on in this week. So, if you recall just to make clear what we are doing, if you recall we had done this simple example earlier of polynomial regression.

So, recall that we did this earlier in the course an example of polynomial regression where we had these points that looked like a sinusoid and we were trying to fit higher and higher order polynomials and you might remember that when you did something like a cubic, we got some

reasonable fit or let us say this is a third order fit basically something like $w_3x^3 + w_2x^2 + w_1x + w_0$.

So, with these four unknowns we were able to fit this. Now, if you think about this in a Bayesian or in a probabilistic sense, we have these four parameters and we try to find out the best parameters for this problem. So, the best parameters as we just saw in the last few videos can be done in two ways. You simply find out the least square estimate you do a MLE of this which is a maximum likelihood estimate and try to solve for the parameter.

So, this will give you some value of w_1, w_2, w_3. So, let us call this W_{MLE} or you could do a MAP estimate which incorporates a prior. This if you recall in the last video is equivalent to adding a regularization term. Since this is equivalent to adding a regularization term this is the same as the $J_{LE} + \frac{\lambda}{2}||w||^2$ and you try to minimize this and this will give you slightly different coefficients.

Now you might remember once again from our overfitting classes earlier in this course I think this was in week five or so. That when we had a higher order polynomial fit let us say instead of third order we had a ninth order polynomial fit, it went something like this it fit these points really well. So, the MLE estimate fit the points exactly but it was a bad fit, whereas the regularized fit actually went a little bit more smoothly through these points.

So, we can see that the MLE fit versus the MAP fit are different in general. But regardless of them being different in both these cases MLE and MAP or what are known as fixed estimates or Point estimates. What I mean by Point estimates is they give a single value for w_0, w_1, w_2, w_3, which is the best fit given the information. Now I will specify a little bit more what I mean by information as we go on.

But we have two versions of this information one of the information which is the MLE information, in some sense we will modify this as we go further is just the data, whereas the MAP information is data plus some prior idea about the parameters. So, we saw that is what is incorporated here that was that is what was incorporated as you saw in the last video in the exponential minus w Square by Sigma square and stuff like that.

But either case all you have is a single value, whereas let us look at what I call a full Bayesian estimate or a distribution estimate. So, in this graph if we zoom in the green line so, if we take the green line, the green line is actually the original fit the blue circles are fit plus noise, this is the observed data. Now the red line which you see here is let us say the MAP or an MLE estimate. Now MAP or MLE estimates actually tell us the best fit.

So, they actually estimate w_0, w_1, w_2, w_3 etcetera but then what is this band here. So, this pinkish band that we see here so, that band is a little bit more subtle. So, this band that we see here is what I call the full Bayesian, the Bayesian distribution. So, in this Bayesian distribution you do not say only what the best fit is, but you give a complete full distribution for w that is instead of saying w0 is a value you actually give the probability distribution function of w_0.

Similarly, instead of giving just a value which would be MLE or MAP I am going to call this the distribution give the PDF of w_1.

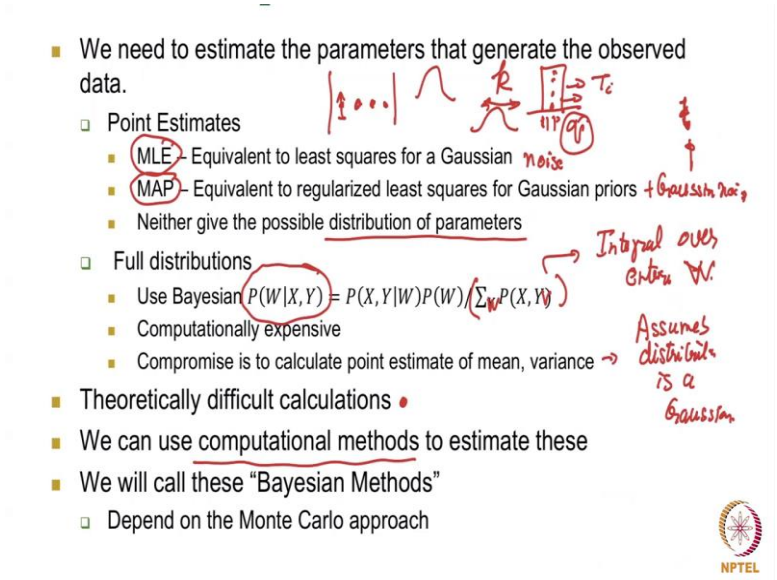
Now once you do that you actually instead of having just one value you have a range of values. So, let us say this range of values is something like the mean value which could be the MAP value, let us say $\mu \pm \sigma$. Now another way of recovering this which we will see and this is sort of the cheating that we are going to do, will be or $\pm 2\sigma$ we are going to get new MAP we will get σ of the maximum a posterior.

And just draw this range as saying well actually the parameters are not just the inverse parameters that we discovered, but they actually lay with a range. But please remember this in MLE or MAP you are giving just one value, whereas in Bayesian you are actually giving the full distribution or in the full distribution estimate you are giving the full distribution. So, now the question is why not do that just like we did MLA and MAP theoretically why not do that.

The problem of course is when we write the Bayesian $P(w|x, y)$ let me just put x, let x indicate the whole data, w indicates the parameters this is as we saw last time $P(x|w)P(w)$. Now for MAP since we were only calculating the maximum only this part was needed only the numerator was needed. But when you want the full distribution suppose you want this whole value you need the denominator also which is $\sum P(x|w_i)P(w_i)$ over all possible values of w.

Now this in the case of continuous w is actually an integral why you have to find out every single possible value of w find out the corresponding term here and then integrate over that. This as you can imagine is horrendously expensive. So, what really happens in terms of Bayesian is this is very expensive. Now we are going to find out some cheap tricks or simpler tricks in order to estimate this Bayesian because it is a useful thing. And we will talk about that in this video as well as in the subsequent videos.

(Refer Slide Time: 10:48)



So, let us summarize what we are trying to do the overall problem is like this. we want to estimate the parameters that generate the observed data. This is what we have been doing I have been using pretty much always the slab problem but you can think of the other problems we discussed you can think of a fin you made certain thermocouple measurements there are some unknown heat transfer Q and you have made some temperature measurements T_i you want to find out what was this Q or perhaps q is given and you want to estimate what is the thermal conductivity of this material.

Now you can give Point Estimates for these that is you can say well given this the mean estimate amongst all the possible estimates is the maximum likelihood. So, this is equivalent to as is mentioned here the least square. If you assume a Gaussian distribution of noise, for Gaussian noise, you can see that MLE estimate is the same as the least square system. Now maximum a Posteriori is again a regularized least square for Gaussian noise plus Gaussian priors. That is each thermocouple has some noise which is Gaussian and the parameters are themselves distributed according to a Gaussian and we have some prior, and that is what gives you regularized least square. Of course, as I just mentioned neither of them give you the full

possible distribution of the parameters. For a full distribution you have to use the Bayesian. And as I mentioned here this portion is an integral over entire w or entire y in this case.

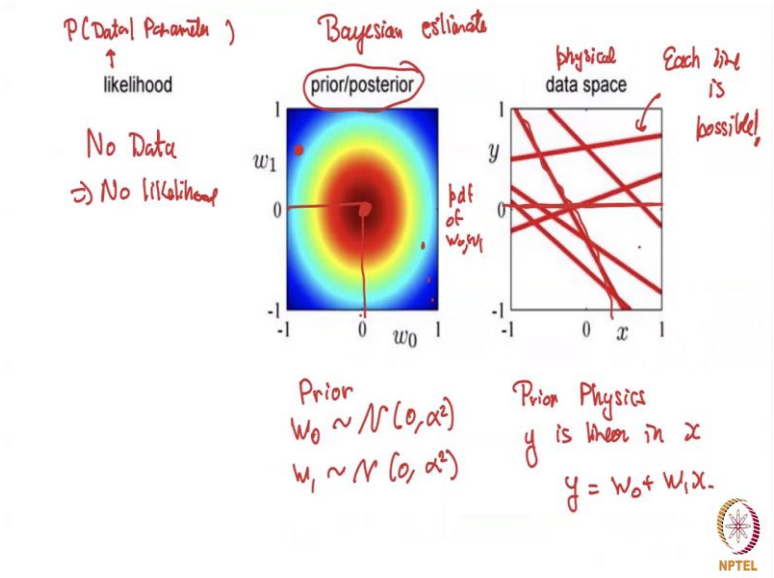
But if you in this case I put y_i should have actually put w please correct that. So, you need to well this is accurate. So, you actually have to integrate over the entire w set and once you do that this is very expensive. So, the compromise is to calculate the point estimate of the mean which would come out to as I will discuss later either the MAP or the MLE depending on how you calculate it and then you also calculate the variance. This assumes that the distribution of the parameters is a Gaussian.

So, you can see that we have made three Gaussian assumptions. we have made three Gaussian assumptions in that we assumed that the data that we collected was some original data plus Gaussian noise. Then we assumed that the parameters, prior assumption about the primate parameters is Gaussian distributed then we are assuming that the posterior distribution is also a Gaussian. Now all these things happen to be reasonable assumptions in practice but they need not always be true.

In general, these calculations are theoretically difficult to make that is because you can show that, when I do a full Bayesian even for linear regression it is very messy. So, I am skipping it the earlier plan was to include within this course but it becomes too involved and frankly we do not need to do it as I will show you computationally, we can get there a little bit faster, which is why we use like everything else we use computational methods.

I am going to call these Bayesian methods though there is a little bit of abuse of notation there and these depend in general on what is known as the Monte Carlo approach. Now before we go to the computational approach, I want to give you further intuition about how exactly this Bayesian works in the case of let us say again our linear regression problem within a slab or something of that. So, let us move on to that.

(Refer Slide Time: 14:48)



So, I am going to show a nice series of visualizations these are all from like our overfitting slides they are from Dr Christopher Bishop's excellent book on patent recognition and machine learning as I had mentioned earlier the free PDF is available online from the official book website. So, what I am going to show you is a series of images that visualize what really happens from different viewpoints, when we actually try to do a Bayesian estimate.

So, we are trying to do a full Bayesian estimate of the parameters of a problem. So, let us say we have our familiar slab or something of that sort and we have a linear regression problem. Let me show you the final picture and then we will come back to the original picture again. So, let us say this is the final picture, this is the final data. The final data is this set that you see here there is a whole bunch of axis and correlated y's.

And you want to find out what line fits best or you want to find out what is the kind of model that you wish to give for the underlying process that generated this data. So, we are going to build as we are doing this week a probabilistic model. But initially we have no data at all. All we know is some Physics of the problem that tells us. So, the only prior so, to speak is something like our slab problem is y is linear in x.

So, we could say something like $y = w_0 + w_1 x$. Now the data that we are about to generate it turns out has some specific prior values out of which we are generating it. So, I will talk about that shortly. But initially we have none of that sort. Now since we have nothing of that sort available all we have got to do. Now is to create these random lines each line is possible. So, every single line is actually a possible data set.

Now we are going to have another prior. The prior now says that well it is not quite every line is equally possible but we can say that w_0 is normally distributed with some σ and w_1 is also normally distributed with some σ. So, let us instead of calling the sigma let us call it α. So, μ and so, variances let us say α^2. So, if you look at this picture approximately α^2 is 4 or something of that sort, I think that is how we generated this data but it does not matter.

What you see is not all values of w_0 and w_1 are equally possible, it is more likely that somewhere at the centre you can get values of w_0 and w_1 whereas as you go further and further away larger and larger values of w_0 and w_1 are less likely. Accordingly, you will see very few very high slopes, but a lot of medium slopes. Similarly, you will see very few lines with very strongly negative offsets.

But you will see lots of lines which are reasonably correlated. They might be going up or they might be coming down but you will see intermediate lines, that is what this kind of prior means. Now as you see there is no likelihood remember likelihood means probability of data given the parameter, but currently we have no data at all. So, the data will come in the next slide as of now there is no data.

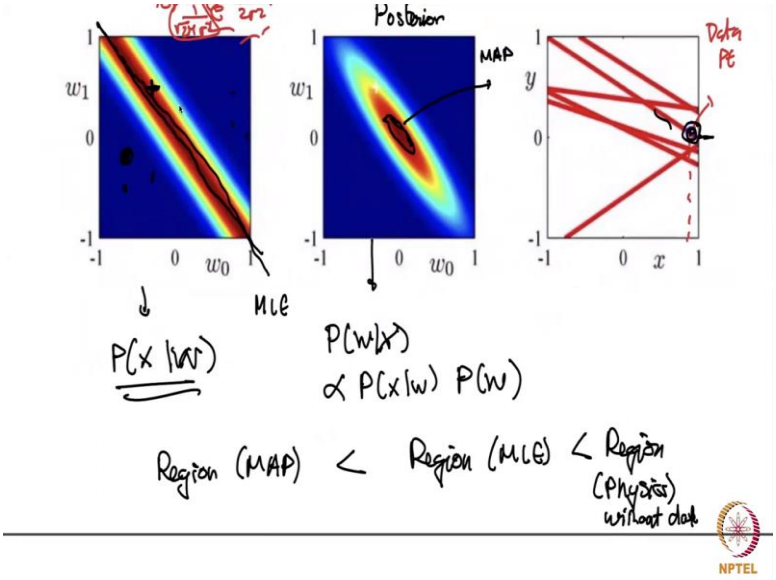
So, this means no likelihood. Now before we proceed further, notice these two pictures, these two are different pictures, this you can notice says prior as far as posterior I will explain why it says that shortly as of. Now let us call this the prior and what you see here is an image which is in the space of parameter space. So, remember we had discussed this when we came to gradient descent w_0 and w_1 how do they relate.

And this is a visualization of the probability density function of w_0 and w_1. At each point these points are very highly probable these points have low probability, that is what you are seeing in within contours, whereas this is physical data space. So, notice it is the same information each point actually corresponds to an entire line in the physical data space. So, suppose I take this point here this is $w_0 = 0$ and $w_1 = 0$. So, that will correspond to a line like this.

So, each point within the prior space or the posterior space corresponds to an entire line in the data space. So, let us move on and see what happens if we actually start giving some

information which was measurement from the slabs. Currently the only visualization here is from entirely from the physics of the problem we will. Now start visualizing what happens when we add data points one by one.

(Refer Slide Time: 20:45)



So, what you see here is we just added one single data point. So, now we have two pieces of information so, to speak. The physics of the problem which tells us that the entire data is linear and this one single data point. Now how is this data point generated? This data point was generated in the following way, we assumed w_0 is minus 0.3 and w_1 is 0.5 and we took the measurement at some x some random x.

Let us say x equal to 0.9 and we just generated,

$$y = w_0 + w_1 x + noise$$

So, it won't be exactly minus 0.3 plus 0.1 into 0.5 into 0.9 which would be approximately 0.15 I also add some noise to simulate the fact that if there is a thermocouple it is going to actually have some measurement error. Now this noise has a variance. This variance is 0.2. So, that is what we generated. So, now we are going to generate lots and lots of lines which can fit this one single data point.

Now you will say there are infinite lines that is true. So, there are infinite lines that can fit this point but there is a constraint they should fit this line. so, that the Gaussian noise should have variance 0.2 Square. So, 0.04 now what does that mean? It means that 95 percent of the lines should be within this point, which is say let us say 0.45;

$$y = 0.45 \pm 2\sigma$$

So, you should be somewhere between 0.5 and 0.85. at this point all the lines that go through this point should lie within that range, at least 95 percent of the lines that I generate should go through that. Now what represents this, this is what we call the likelihood. What is likelihood? Likelihood says that all these points now I select a w how likely what is the probability that this data set.

Remember I am going to call it capital X it basically contains the pairs of xi and yi in this case there is only one x and one y, how probable is this point given some particular choice of w. So, what we do is if you see this entire domain you keep on going let us say I select minus point one minus 0.1, I put that I put that in the Gaussian remember the likelihood was actually a Gaussian.

This was $\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{\frac{-\varepsilon^2}{2\sigma^2}}$, remember this likelihood is actually not normalized unless you do this. But we will come to that this becomes particularly powerful when we come to the MAP estimated. So, assume this I will show you a proper computation procedure for this shortly but here is the original point you see that cross there.

So, let us mark this in black in order to make it clearer. So, this was the original data point. So, this is non-noisy w0 is minus 0.3 and w1 is 0.5 you put that in here and then you generated this extra point and now you are trying to figure out given this extra point, what is the probability that the original data that generated it was somewhere here? Now for each of these choices of w0 and w1 you actually calculate a probability and that probability is what is marked here as a contour.

Obviously, this is a very expensive calculation, we are not going to do that in practice but let us say you did do it. What will happen is the probability that you actually had the original parameter is fairly high. You will see that the probability peaks right at the centre. It starts going lower later on and somewhere right at the centre of this will where you will get MLE. But this is an entire line for the maximum likelihood estimated, that is because it is not just one parameter that will satisfy this well.

There are a large number of parameters that will satisfy this well and they will all be equally probable. Now is that it is we done? No, here we have a prior. Now where did this prior come from. More than a prior this actually happens to be MAP. How does this MAP or how is this MAP calculated. So, let me show you how this was calculated. Now this gave us probability that this data points this blue data point was created by any one of these exactly.

But we have one more information, we have the prior information this one. we know that you cannot select a data point from somewhere here because that has very low probability. Most probably the data came from here and as you go further and further these become lower and lower probabilities. So, how was this generated? This was generated by a product of two probabilities,

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

which is this.

In other words what we are saying is if I had only looked at this data point all I would have been able to say is that the parameter lies somewhere here the most probable thing is the parameter lies somewhere here or that this is the probability distribution of the parameter, whereas I have some information from before. Now when you multiply these two so, let us see if we multiply this with this, you will see that the probability at the edges becomes fairly low.

Here the probability was 5 but when it is multiplied with this, the probability here decreases. Similarly, the probability at this portion becomes slow it was already low here. So, it becomes even more low. You will see that these portions become lesser and lesser possible, because all these multiply low probabilities. Similarly, these portions get a lower probability because even though this probability is low, this probability here is high.

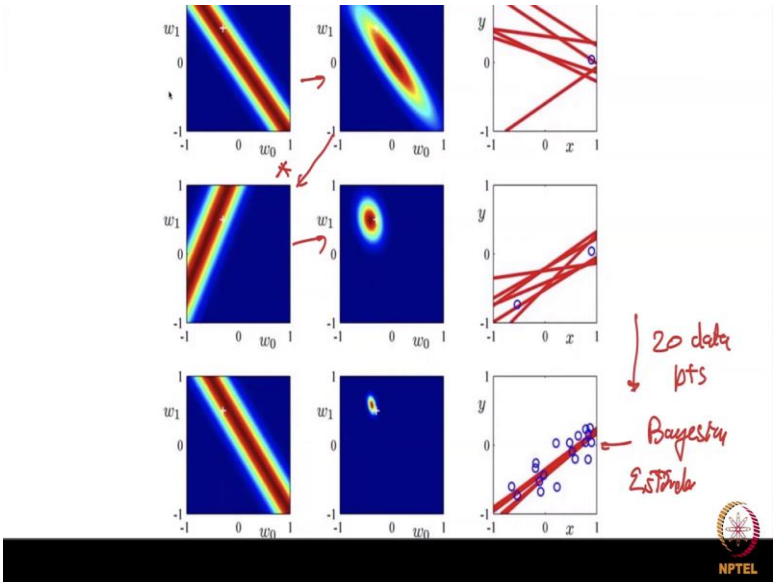
But as a combination we get this result. This after multiplication with prior now becomes the posterior. So, now our MAP region is the small what do I mean by that that if I had looped only at this data point, I would have been able to give a lot of lines which can go through it. but this data point combined with the fact that w_0 and w_1 have to be low values give us a region of probability space that is a little bit more restricted.

So, what you now notice is the region of MAP is lower than the region of the MLE, which of course is the region of just the physics of the problem, just the physics of the problem without

data. Just the physics of the problem says some line when I gave you some extra data that is just like giving a boundary condition or something of that sort you immediately said I have a large number of parameters for w_0 and w_1 but they are still smaller than a full Infinity.

Then you come to MAP and say I am in an even smaller region of parameter space. So, now we go further passing and imagine.

(Refer Slide Time: 29:55)



Now I can add one more data point. So, I seem to have repeated this. So, let us say I add one more data point here. Now what happens? Now this that I am drawing here is $P(x_2|w)$. So, likelihood of only x2 I am not including the first point as well as the second point. So, now I included the second point and. Now I draw the likelihood of that and you get another likelihood estimate.

So, this would be the MLE, maximum likelihood estimates for just x2. Now why did I do just x2 that is because we want,

$$P(x_1, x_2|w) = P(x_1|w)P(x_2|w)$$

Assuming we have independent x_1, x_2 remember we made that assumption that all measurements are actually independent of each other.

So, we will make that assumption. we will make the assumption that x_1, x_2 are actually independent. Now you might notice that this has become even smaller. Now what is this? This one is $P(w|x_1)$ multiplied by this. Now you might find this strange I mean what kind of logic is this. So, the logic is very simple,

$$P(w|x_1, x_2) \propto P(x_1, x_2|w)P(w)$$

Now this on the other hand is,

$$P(w|x_1, x_2) \propto P(x_1|w)P(x_2|w)P(w)$$

This we just showed by Independence this therefore is,

$$P(w|x_1, x_2) \propto P(x_1|w)P(w)P(x_2|w)$$

This is proportional to probability of w even explain. So,

$$P(w|x_1, x_2) \propto P(w|x_1)P(x_2|w)$$

Given w another way to say it is that this is the prior or the sorry this is the likelihood of the second data point, this is the product of this likelihood with this posterior. So, let us look at this again. you multiply these two what happens we already knew from just one data point that my data lies somewhere here I already know this. So, notice what we are doing, this is what is the basic trick you know Bayesian estimate. The trick in a Bayesian estimate is you keep on incorporating what you already know.

So, this was the posterior in step one, but this acts as a prior for step two. What does that mean? whether something is a posterior or a prior depends on what we are doing. So, for example before we started anything right at the beginning of the problem our prior was just the physics prior this before we even know this, this is just the prior. Now I multiply by this prior I get a new posterior, that is why the word is prior or posterior.

This incorporates both the physics of the problem as well as the information that I have. Now this is the likelihood, this is the data. Now this data multiplied by what was the previous prior, becomes the new posterior. posterior means after I know something. So, just imagine this as a chain rule. Suppose you have a deck of cards, the first card you pick up is an Ace of Spades, before you knew anything that card could be anything and the probability of Ace of Spades would be 1 over 52.

Now once I pick Ace of spade, I know something else. Now this Ace of Spades becomes the prior for the next thing next, I pick another card let us say that is Queen of Hearts. Now that becomes the new prior so, same thing. So, this was the posterior that is after I incorporated the

data and multiplied with this, I got a new posterior. This is the information I am going with I already know that the probability that the parameters are here is high.

Now my second data point says well, this is the probability of the second data point. I will multiply it with this. Now I will multiply it with this and once you see that you will see that the regions where are that are high become even smaller. So, now you are relegated to a much smaller space. Now you keep on adding data like this, let us say this is finally the 20th data point. So, there were 20 data points in this data set.

What you have here is just the likelihood of the 20th data point by the time you apply this to smaller and smaller regions you can see that the parameter space is really small. This is the final Bayesian estimate for our parameter space. And you can see that the lines start converging you can see very few lines or possible 95 percent of the lines will converge to a very small space.

So, let us put all of it together the prior for the previous step multiplies this gives you the posterior, prior multiplies this gives you the posterior and you repeat it 20 times and you get the final Bayesian estimate. At each step what you have is really a Bayesian estimate at each step you actually do have a Bayesian estimate. But you sort of keep on accumulating this Bayesians step by step by step and you get a final Bayesians in estimate as you can see this entire process can be computationally expensive.

We are going to show a sort of naive version of this within this course. There are very sophisticated methods for this. I am going to show you a very naive version of how to do this starting from the next video. I had something a little bit more sophisticated plan but I think that will be overkill for a course like I said this course is basically an undergraduate level course just an introduction to inverse methods and you can see that already that gets pretty complicated.

So, we will stop ourselves to a naive version of this and I will start with what is known as Markov chain Monte Carlo in the or a pure Monte Carlo, Markov chain Monte Carlo in the next video. So, I will see you in the next video, thank you.