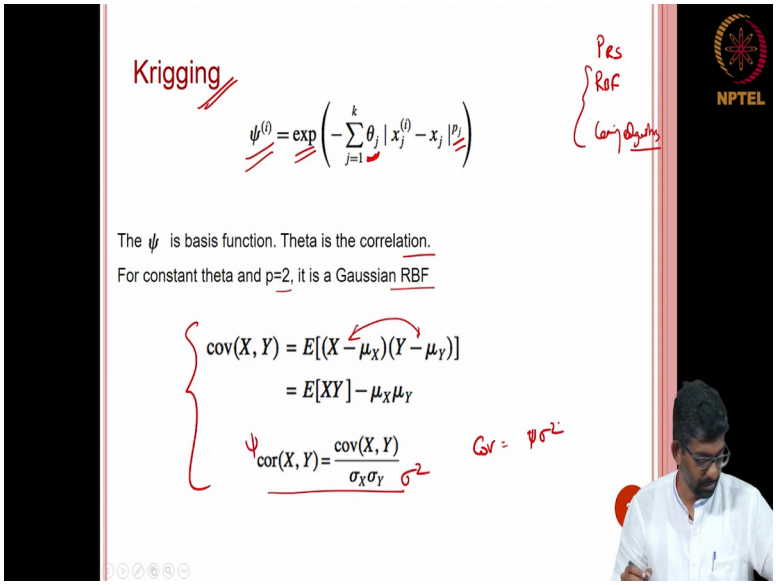
Surrogates and Approximations in Engineering Design
Prof. Palaniappan Ramu
Department of Engineering Design
Indian Institute of Technology, Madras

Lecture -14
Kriging -1

(Refer Slide Time: 00:16)



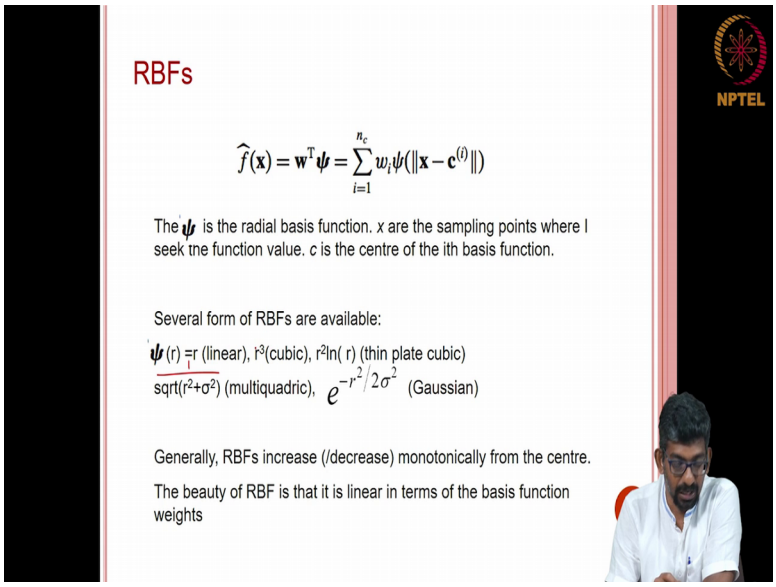
So, the next meta model that we wanted to discuss is Kriging. So, far we have talked about polynomial response surface, we have talked about radial basis function. Usually radial basis function is also very analogous to neural network. If you have done neural network ok, if you know what a neural network is the basic radial basis function equation that I showed you is nothing, but a neural network with one layer neuron that is all that is the same idea. That is how it is all these are all called learning algorithms ok. So, what they do is they yeah one question is something similar to what Kiran just now asked is, how do you decide upon your function your radial basis ok, what basis am I going to use.

So, one is through expert opinion because this is kind of your prior knowledge if you do Gaussian you need to have a prior knowledge ok. Sometimes what happens is if you do not have any information on and you take a uniform prayer because that represents a maximum uncertainty. In a similar sense if you have lot of samples then you can go for Gaussian or because there were also I showed you right like different radial basis functions, varying functions, you can capture depending on if you expect linear

relationship ok. So, efficiency versus some throttle speed or something there is a linear relationship with that, then you do not need to worry about a Gaussian function and all that you can just use a simple varying r r r square should do the job for you.

One way to do that to select your basis itself would be to solve not an optimization problem that will be very expensive. One way that people usually do that is pick a model that minimizes your error or that gives you minimum error, I should be careful it is not really minimizing the error minimizing the error means it becomes an optimization problem. Now, what you do is you will use the different models and you find out which model gives you less variation in the results ok. So, let us go back to this, lets say that you do not you have no idea on what model you need to use, then what you do is you take the data.

(Refer Slide Time: 02:30)



And you fit it with a linear model you fit it with an r cube, you fit it with this, you fit it with this and you fit it with this. So, you use a regularization parameter if you want or you do not centre it along the points ok, so you centre it at other points. So, that you can predict at the known points, you understand what I am saying right. So, you can predict at the known point then you will get the error matrix. Now, what you can do is you can get overall generalized error the global error and then you can select the model that gives you the least error that will one thing. Or you can use the press error see usually all these

ideas also came from what we call machine learning kind of idea. So, there usually what people do is they divide their sample into training data and test data.

So, training data is used to build the model and then the test data will be used for testing the model how good your testing is right. Whereas, that is where press comes into picture because what press does is if you have 20 samples it lets you build with 20 samples, but you let to do it 20 times, but that is ok, that is nothing compared to running a simulation. Running a simulation takes 4 days building a 20 model will take you 20 minutes, so that is nothing compared to the getting a simulation.

So, you do 20 models, but every time you leave 1 point out, so the first time you left the first point where the nth time you left the nth point. So, you are building it with only 19 models, your sacrificing only 120 of the information which is 5 percent. So, 20 point is the least that I am imagining that someone will work with. Even with that you are only losing 5 percent of the information instead if you divide training and test you are dividing it as 60 40. So, you are losing 40 percent of the information which you do not want ok, usually they do 20 at least 25 percent you have to use for testing.

So, instead press gives you that benefit of using all the samples for fitting and all the samples for testing as well. In that case what you can do is you can take a press which is the press which model gives you the press best press and then you can select that, but what you need to be careful about this is. This was initially proposed to the ensemble meta models and all that, but what later people realized is also where and how are you going to use these meta models. So, if you are going to build a meta model and you should further studies like optimization people showed that not a great meta model gave you a very good optima compared to oh very good meta model.

So, these are kind of d coupled though we believe because optimization only requires a smooth function. And within the smooth function the order of smoothness is there, but then as long as it can do a gradient estimation the optimizer is enough happy with that. Then what the designer need to worry is how much did you preserve your zones of interest, because the bad meta model could have preserved the zone of interest meaning the peak or the valley.

Hence the optimizer converts to a better solution than the good meta model which actually did not capture your zone of interest ok. Then this becomes a chicken and egg

problem how do I know the zone of interest you do not know the zone of interest ok. So, there people try to look into zone of interest can I know because usually zone of interest is not that difficult; meaning you know it is very simple right like I am looking for minima or I am looking for maxima. So, the more the value is the better it is, but then you could be totally unfortunate. Where you did not have a sampling point that corresponds to a low value or somewhere near that you did not have a sampling point. So, your model could entirely skip that and go elsewhere. So, we look at that so there are tradeoffs with all of this stuff.

So, that is why what is something that is important that you need to appreciate yes it is important for me to approximate the output function and all that that is what we did, but with respect to the input. Because, it is all depended on the sampling, the moment you change the sampling and then you use the same model same to any point same to any point means same number of to n points is 20, but their sampled at different locations and the previous one you will get a totally different solution.

So, they should always be kind of correlated or there should be a relationship with the x you should bring that. So, right now we do not explicitly build that relationship we are only saying that x is related to y and I am going to find a function f that will do this. While estimating you do not take that you do not take how close or how away was my new point with the existing samples, you understand what I am saying right.

Once you build your model a fat I give you a new sample you just go on plug that sample and then it will give you the value you do not find out how far was that sample with respect to your previous samples does that mean does it give you some better predictions or better confidence in your predictions. We do not explicitly do that implicitly in one sense it is built into the model because in the RBF, if you look at it says x minus the centre the centre is what usually I have the information. So, it kind of tells you that it is a Quadra it is a Euclidean distance is it is quivered distance it is either opinion.

So, the more the distance in more is a deviation that is what we understand, the closer it is to your centre I am very close to what the original value is when it goes away from it. But then when you bring the regularization you lose a little bit of that confidence also ok. So, the idea is I also want to capture this with respect to my input space that is something that is important becomes important.

So, one such thing that lets you do that is the krigging; what Kriging does it is a small modification to your RBF. The psi that you talk about it is interesting psi of i and then I give you some function some basis function here this is nothing, but your Gaussian function. The only thing is this theta did not change in your RBF and this P was fixed to 2 and it was a Euclidean distance if you remember sorry going back and forth.

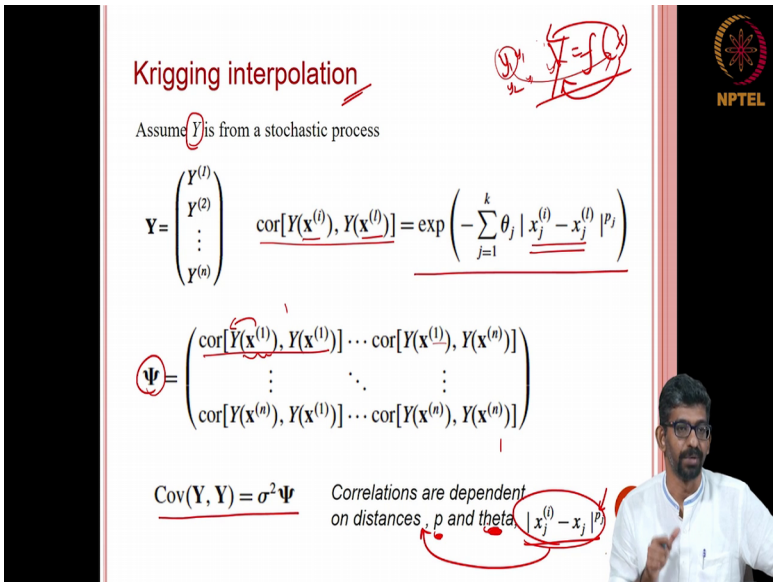
This one I did not explicitly say it is not the distance meaning it is a Euclidean distance square distance l 2 norm when you put this two things it is an l 2 norm, just quivered distance if I make this p equal to 2 and I fix this theta then nothing, but your RBF. So, what Kriging does is it introduces another parameter it says let us find this theta and this p also can vary. So, it means it is going to tell you like can it be fatter thinner, can I change the centre accordingly with respect to the point that you are giving me that is all ok.

So, there are more parameters, but it lets you it will let you do more things that is the point. The deal is it necessarily means more competition, because you can have a very thin distribution in this point and you can have a very fat distribution whereas, in RBF all the distribution where of the same bit wise they were the same. The psi is the basis function theta, what is this information is this is the correlation information that is built in will see what this correlation is. It will actually build to only with respect to your output, but it is nicely brings information into your input you will see that and for constant theta and p equal 2 it is your Gaussian RBF that is all ok.

So, it is nothing, but your Gaussian function and we are just taking these two guys. For you to just appreciate a little bit you need to know this covariance function and the correlation, basically the correlation coefficient. What is the covariance? Covariance is X Y there are two variables let us say X and Y, each X with respect to it is me product of, each Y with respect to it is me you take an expected value of this entire set of data that gives you your covariance ok.

So, covariance is between two sets of data that we are going to talk about and when you take the covariance and divided as a product of your standard deviations then you get your correlation. So, it says if X is going to vary I mean if X changes how does Y change that is one usually, the linear correlation is what we are talking about. I should have done an animation, but that is just stay with me do not look at the slide directly.

(Refer Slide Time: 11:26)



So, we are talking about the interpolation part right now, we are not talking about the regression part the regression is understood if you give this you know how to do the regression is what we understand. Your Y which is your output what you are assuming here is it comes out of a stochastic process it might not be. It comes out of a simulation where it is deterministic fine, but for the sake of Kriging what you say as is this Y is a stochastic process the moment you say stochastic, you know that it could have been some other entry also for the same input that is what we are trying to say or it could be non-linear.

So, I want to identify the correlation first what I am doing is, I am identifying the correlation between the Y entries themselves. Usually you do it with respect to X and Y you do with respect to X 1 dimension and X 2 dimension ok, but here what we are doing is we are taking X of I meaning like corresponding to X of i and X of l we take the Y and we try to identify the correlation. This we are representing using this equation do not worry about this yet, if you want to find the correlation then you can set up this matrix ok. This matrix is what your basis vector as or basic basis matrix with actually basis vector.

So, what I am trying to do is of course, this guy will be 1 the leading diagonal terms will be 1. Now I will take Y corresponding to X of 1 I will do it with respect to X of 2 X of 3 X of n are the other around X X of 2 X of 3 X of n all with respect to X 1. So now, this is

how I am building the relationship between my input space and my output space this was explicitly missing in the previous cases. I was kind of remotely bringing in this information ok, but right now what I am doing is I am bringing the correlation by building the basis vector as a function of that ok.

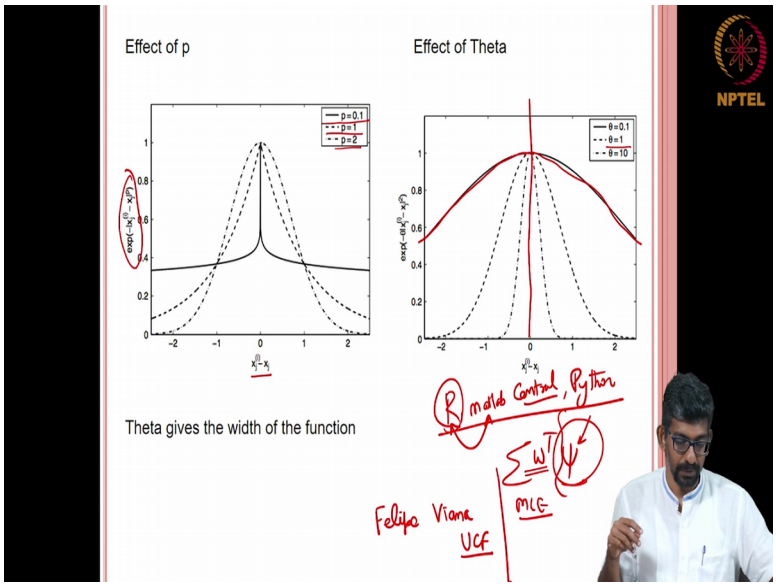
So, by doing this why I am doing because this Y this is what our overall assumption like we believe that there is a causal relationship between X and Y. That is why we write this equation if you change X Y will change this is what our under which means there is a correlation between X and Y this is what you are capturing here. But you go one level higher and what you do is you say there is a correlation between Y 1 and Y 2 Y 1 and Y 3 Y 1 and Y n. Each one of this obviously, has a correlation with X that is what your building l, because this point in space is related to X 1 and X 2.

So, I am taking that explicitly this is what, is this guy. So, Y of X 1 X 1 means with respect to X 1 itself Y which is the function of X 1 and X 2 I get that information Y which is the function of X 1 and X 3 I get that function. Because that X 1 and X 3 are going to vary from X 1 and X 2 right it could be 0 minus 1 and 1 the other point could be 1 1 1. So, with respect to X 2 it was minus 1 with respect to X 3 it was 1 these are the coordinates. So, I bring those information how far is it from X 2 how far it is from X 3 from the previous points those information's are built into this right now.

So, in one sense your kind of correlating your output, output itself with respect to each other in terms of your input ok. So, you are capturing the distances in one sense, then I can write my coefficient of coefficient of variation here in this sense sigma square times psi. But we also know that is nothing from this actually this is nothing, but sigma squared this was your psi. So, I am writing my coefficient of variation as psi times sigma squared are the other way around sigma square epsilon psi.

Now, yeah, so was put in a hurry this should have come here the correlations are dependent on the distances now the problem is changed you are finding other things now ok. So, you are going to find a correlation now, so the correlation that depended on what it is dependent on these distance information you need to choose your p to do that and your theta which is the output part. How is this going to affect why do you need to introduce, so many things.

(Refer Slide Time: 16:31)



This is the effect of p and this is the effect of theta what does p capture p essentially tells how you are going to finalize this distance that is all it says. It also tells you that X j i minus X j if it is very close to each other which means 0 the difference is 0 it is centre and I want to estimate it there, it is should give you the same value and this is my output. So, it should give me the 0.1 at that exactly at that point, when I am varying outside when I am going further away from that it is still giving me only that value ok.

So, that is corresponds to p equal to 0.1. So, these are the weights please understand weights in the sense it is the influence that you have you are not predicting from that it is the influence that it is going to have you are going to follow this influence pattern. When I am using p equal to 1 you can see how this weight is going to be p equal to 1 means this guy right this difference raise to 1.

This is how it is going to vary the more you go away what will be the value that I will get this is the value that you will get, when it is squared which is the Euclidean distance which is what RBF task it looks more like this. So, you can see how they are changing, so there multiple ways in which you can solve this even as an optimization problem or you can pick up a p yeah if you fix a p then it becomes RBF if you leave the p to be selected as an optimization problem then it becomes Kriging ok.

So, Kriging hence lets you model high order nonlinearity that is the understanding, but you learn to be careful that it does not do an over fitting that is one thing that you learn

be careful. The second part is your theta the theta kind of tells you about the fatness this tells you about with respect to the distance that is the one that it captures right it sorry, it talks about the penalization part. The second one is the theta it tells you about the functions with. When your theta equal to 0.1 you can see how heavy the or fat the distribution is, then if your theta equal to 1 it becomes thinner and if it is equal to 10 it becomes even thinner.

So, you for higher theta then I will kind of come closer to this one like a deterministic. So, what does that mean it captures your what does theta capture what is your theta capture it is nothing, but your correlation information right, what does the correlation information larger correlation means they are correlated meaning they are better correlated. So, actually if you are looking at do not look it up as a correlation coefficient because coefficient will be vary between 0 and 1 this is the correlation the general correlation information that we are talking about.

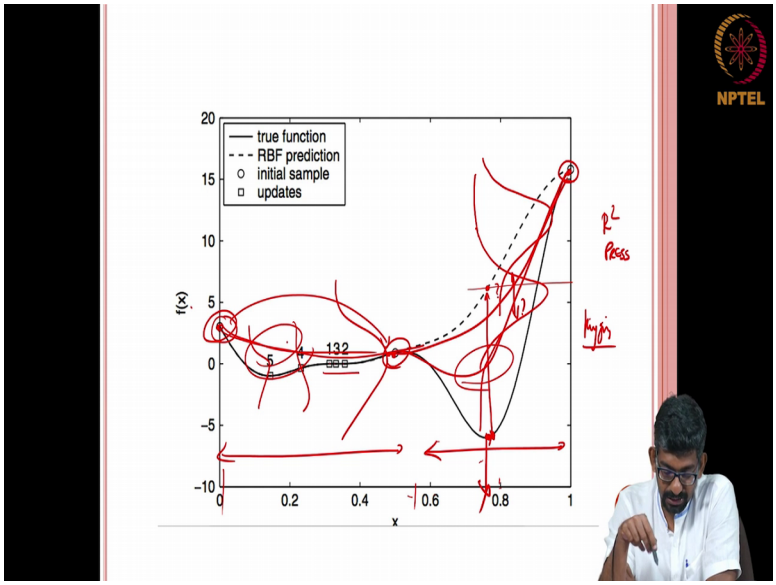
So, the larger the correlation the better it is that is what you look at ok. So, this is basically the interpolation of Kriging is what we have discussed, but you can go and estimate because the idea is the same you are just going to do summation of sorry. You just going to do a summation of this was a same idea the psi was a polynomial response surface you know what it was for RBF we use a Gaussian with the specific p in this your p also will vary and your theta will vary in Kriging that is all.

So, if you know how to deal with this guy finding this is still the same story you can use a maximum likelihood estimate and you can do that that is not a problem. Usually most of the software's like R MATLAB does not have an inbuilt Kriging right you have to download the yeah. So, in MATLAB central you have an inbuilt software I mean some people have done it and then there. So, or else we widely use something that was developed by who is currently a professor at university of central Florida we can actually share that detail.

So, you can download this tool box and you can do that or meaning o r or you can choose R ok. That also gives, python has a good toolbox, but the only thing is what I suggest is if you are already using MATLAB or you are already using R and you are going to a new software make sure you calibrate the results. The terminology sometimes are not the same in people run into issues for months together.

So, you need to have an fundamental understanding, so what you are best would be to take foresters book take one of the only example that here sees others. And take the samples with the same point see if you are getting the same error metric are the same fitting that they are getting it is very because it is easy were you to visualize also, then go ahead and try using the software that is what you should do.

(Refer Slide Time: 22:11)



So, in effect meta model in construction is over with this and also little bit of meta modeling selection we spoke about, but there are also issues with meta model building mostly with respect to the sampling point. And again you do not know the original function see for instance in this the bold line was the what the original functions. And these are the samples that have taken do not worry about the circle and the these are based on some in fill criteria that they do at later point.

Let say that divide samples and then I got this approximation is no way, because with respect to the original samples that I have the fitting is awesome you agree to that. These are the points my original sample original values are there, with respect to that my fitting is awesome. So, the general prediction variances that we have talked about what are the variance in your prediction these errors will be like close to 0 excellent error metric you will get, but if I am going to predict here I am lost completely this is the original and this is be predicted value. I won't even know this so, but you also should appreciate that these points are very closely placed ok.

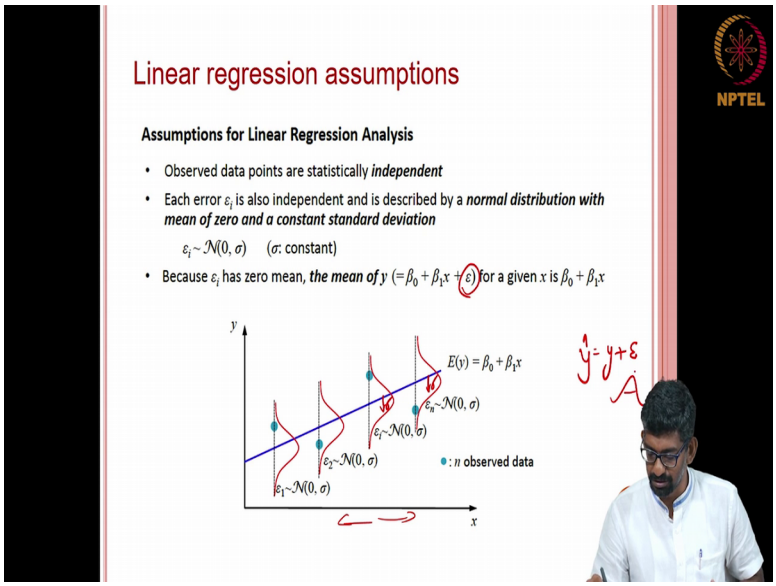
This is not a correct example because these are updates and these updates were influenced by the optima, we actually started with only these 0.3 means a simple 1 day example, simple example 1 day example enough problems, and beyond 3 beyond 2 I cannot visualize also that is another problem. So, what is now required is if I am going to predict here can you give me the confidence in your prediction not in your fitting, you understand there are two things now whatever we have talked about your R square your press error. So, I have taken 3 points and then I am fitting some curve now if I estimate my errors.

So, this was the original point this was the original point this was the original point and the curve actually passes through all the points with the regularization without a regularization whatever it is I could have got an a very good metric this is all you can do. But if you look at the design space itself the information that you are talking about actually these two are more closely placed than this guy.

So, what you might expect is if this function is non-linear I would suspect this region compared to this region this is very example tells you that that may not be the case. So, what people usually try to do is can I get a variance estimate of my prediction see this is very interesting. If I am asking what is the value at here can you give me this value that is one that is doable that we have already done you give me a f at you give me a new x I will tell you what the y is. But what I am asking is hey boss I also know that there is actually it should be the other way rounds sorry yeah it should it should be like this ok.

So, can you also tell me what this deviation is thus a confidence basically. What we are currently doing in polynomial response surface is also the same, but if you remember that curve which is similar to this distribution to standard deviations and all were the same you understand what I am saying. You recall the error that I showed you this guy this deviation is what I am talking about.

(Refer Slide Time: 26:24)



So, you are building it you are building that error into your formulation, but you are not explicitly placing it with respect to where your place your xs and it does not vary, it is a generic model you say y hat equal to y plus epsilon that is all the epsilon is a standard distribution that is it. So, people believe that you need better approximations of those errors specification approximation local approximation of those errors. So, Kriging lets you do that because it uses the correlation information, it says how does that output is related to the difference in your input space ok. And I also tell you the relationship between this guy and this guy with respect to the difference between this guy and this guy in a Euclidean sense.

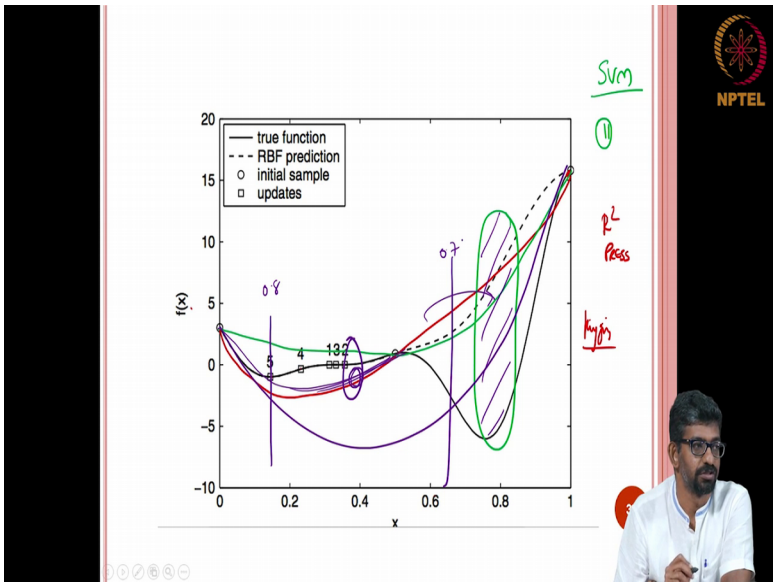
So, krigging kind of gives you that information and there is something called the prediction variance it is very interesting because you predict and then it tells you the prediction variance is large here. Then what you can do is this if you have the privilege to run one more simulation on experiment you would obviously, you predict at different locations. Here it was very large it was very large here. So, then you go and put your point here at this location and then you might get a slightly sorry you might get a slightly better curve. You can keep doing this you can keep sampling and then again you will get locked here you might lose out this surface I do not know that could also happen, you can lose out this surface that can also happen.

So, the point is what are you going to do with this meta model are you going to run and optimization solution, optimization solution you can always use a meta model run it and then go and check at that particular point whether that is what it is. But let say that you are doing a nuclear leakage you do not have a choice to go and check whether the leakage happens or not you do not need better not leak ok.

So, there are situations where you cannot do that it is not about expense you just cannot do that, but you want to find an optima. So, is the case with oil drilling this Kriging was developed by south African reception called Daniel Krige k r i g e from his name is what Kriging came into picture. They want to find these fertile areas because they want to drill with these prediction they were able to drill and they were able to they were successful that is why the method became successful.

I have some data it is really not oil, but it is some indignation based on that I have to put a multi chorded machine there and they were able to do that with. So, he kind of brought all those attributes with the relationship he said if this why it says that this is likely this oil is there and then it has to be related to the other attribute that is what those access are it is not special necessarily anyway fine. So, then what people suggested is instead of using for instance one of the models could have given this, one of the models could have given this meaning polynomial response surface could have given the dotted line the red line would have been Kriging.

(Refer Slide Time: 30:18)



I am just saying RBF could have been like this which one would you choose, because you do not know all of them passes through the points. And these are just three even in forester you go they use support vector regression support vector regressions today we are dealing with about eleven at least 11 machine learning techniques (Refer Time: 30:39) l g rig analysis this is entries. Which one would you use you go on talk to Professor Raveendran he might give you at least 5 or 6 that I will mention which one would you use sorry.

Student: (Refer Time: 30:57).

No, just now I told you right fitting error is only part of the deal, I got you the best and I will show you that did you give you the worst prediction at some point not at all the points ok. I should take back that statement because Kriging for instance also has this variance into that. So, there are multiple ways of doing this, so one is what he said best fitting something that you can do or you have a legacy you say I know my professor has been working in this area for 20 years. And we know how these functions vary and we know that polynomial response is the best or you go to some review paper where they looked at different application they said these are the functions and appropriate.

So, they said like even then they would just say the nonlinearity of the space and then they will say these are the function these are the meta models that are good ok. For instance Kriging is great, but it will choke in about 15 16 variables support vector machines will choke in about 20 20 days in space. So, there are some limitations of all this, so what people said is if you do not know anything if you know that this surrogate is going to work best for this example go ahead with that that is the best thing. If you do not know then use the information for all of them from all of them one thing that they noticed is they need not give you the best error are the best fit.

But they will save you from the worst prediction what they said is if you take this function for instance where is the maximum uncertainty that is what I want to know in the function space. The maximum uncertainties in this region compare to any other yeah compare to any other location the way that I have drawn here also there is a maximum variation. So, let us put a, let us put a fourth guy I cannot fit a surface like that, but let me put a fourth guy who is like this is very bad it cannot happen, but I am just saying

imagine that this function this point was not there. Then what happens is this space is where the maximum variation.

So, they say if you use an ensemble, ensemble means what together if you use together it will let you find the maximum variation space, but and you go and dig back you know that there is something was fishy. But then something was very small that does not mean that it means accurate prediction that they cannot guarantee, but that they say wherever there is maximum variation something is fishy. I do not guarantee because all of them might under predict that also can happen all of them could have gone like this all the three of them could have gone here.

So, very less variation here does it mean that it is a good prediction no it is not. So, I cannot assure you a good prediction, but what I can tell you is I will save you from a very bad prediction ok, but how do you do this ensemble. I have 5 bowlers 6 bowlers as when the order works 50 overs order or 20 20, I have 6 bowlers. And then I need obviously, alternate between them to bowl 20 over's whatever happens will I give 4 over for everyone who performs better on that day or in the previous over not even that day, I hope and I think likely he will perform that is his day today, so he will perform. So, I will given the second over also after the spell or I want a break I want a breakthrough I give the same person.

So, the sixth substitute person I mean in not substitute the sixth guy who is not a regular bowler bowled excellent that day I will say fine just given the additional over's the my lead bowler did not bowled well in the first two over's. So, I will stop the person from bowling that is why you need that extra stuff otherwise you are gone you do not have a 6 bowler you have a problem let us not speak cricket here, but that is the idea. So, you want to use an ensemble of bowlers there similarly here you want to use an ensemble of meta models because there also it is a prediction only you do not know what is going to happen, but based on that days performance I will build my ensemble.

So, this is going to be a weighted weight means how many over's did you give for each one of them what is the weight that you are going to give to these guys. There is going to be a prediction from each one of them you say I am going to give 0 over to this guy because these guys are doing very well fine it is a very good prediction. So, it is fine you say no, no, no this there is, but how are you going to how are going to measure the

performance obviously, in the cricket you will look at the number of runs here you will look at your errors. So, but these errors it will be good if it is a prediction error you do not look for a fitting error.

So, you can use of press for instance and then you can use the best press that is one way of looking at it or you can do a weighted of the press because the press error is itself going to vary the press error will vary a across your domain. So, at this domain RBF at this point RBF might have done well and this point Kriging might do well. So, what I do is here I say RBF you take a weight of 0.8 and here Kriging you take a weight of 0.7 and a summation of the weight should be equal to 1. So, that is one way of doing a weighted average you can either do a weighted average surrogate or a weighted surrogate ok.