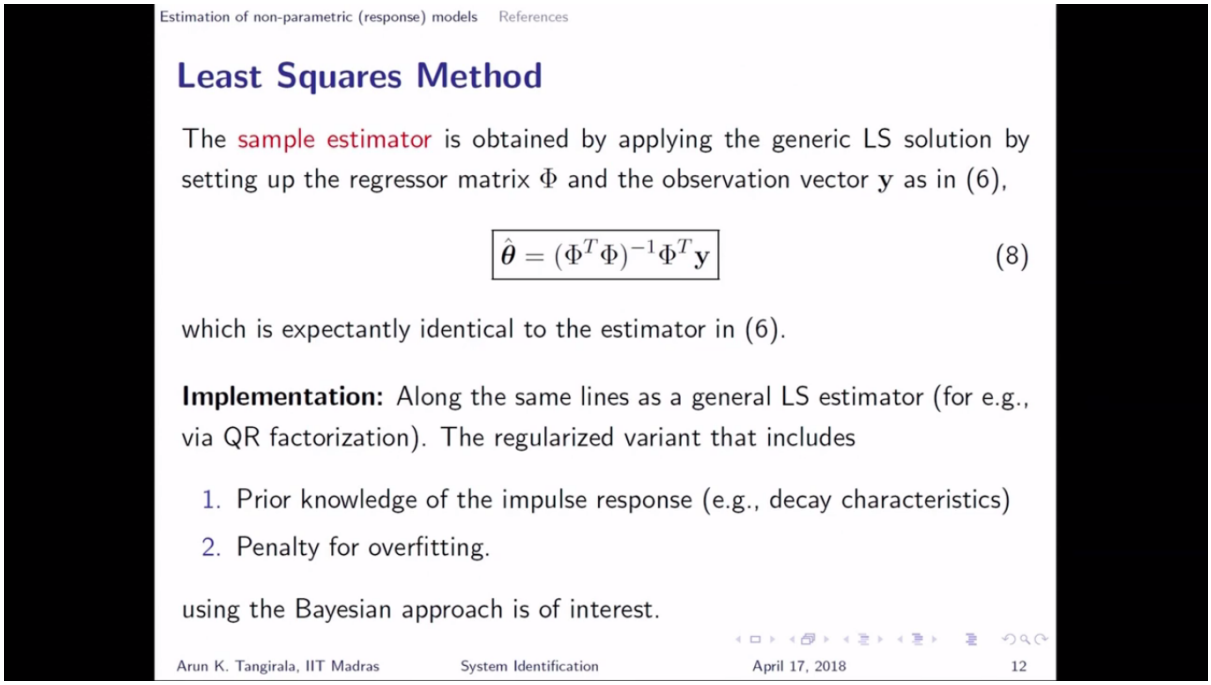
CH5230: System Identification

Estimation of non-parametric model

Part 4

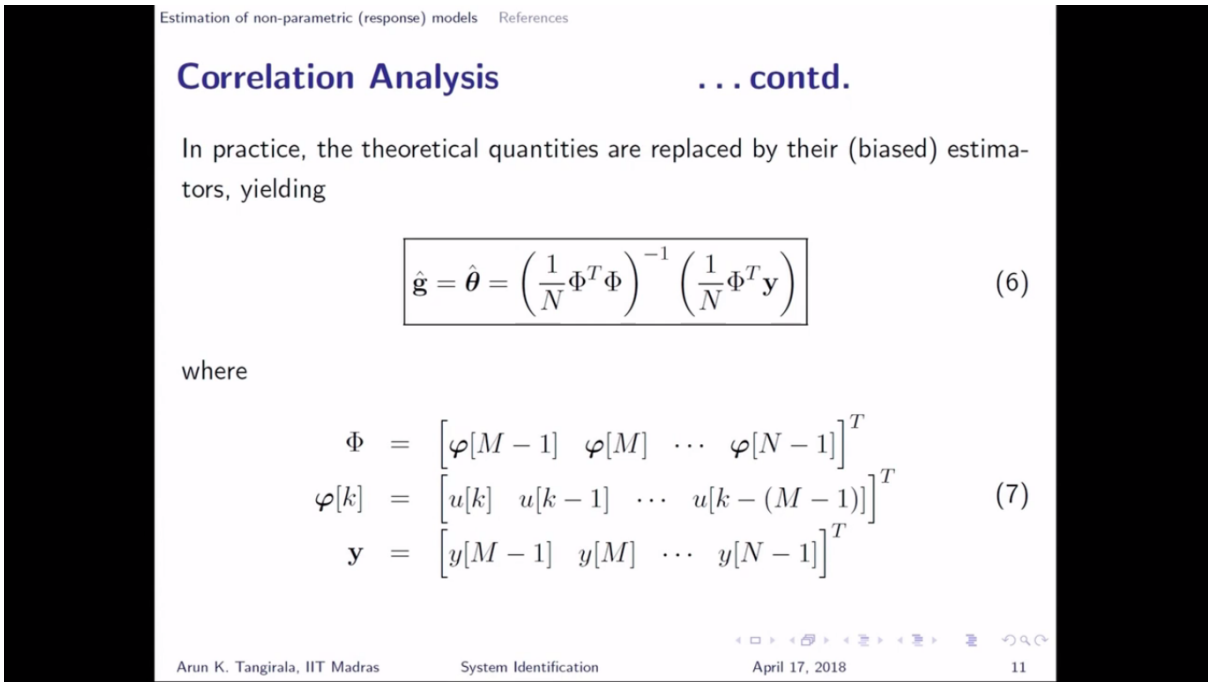
So we'll know more on the least squares method which we have talked about again in a generic way and in a specific way yesterday as well, in the previous lecture. So as usual you set up your Phi matrix exactly the way I have shown here.

(Refer Slide Time 00:25)



You set up your Phi matrix which is a regressor matrix and set up the y vector as you see on the screen.

(Refer Slide Time: 00:35)



And simply plug into the generic solution. So you have theta hat equals Phi transpose Phi inverse times Phi transpose y. Now obviously the estimators coincide because as I pointed out earlier the conditions for that we have applied in the covariance method and the property of the least squares estimator coincide which is that the residuals-- leftovers are orthogonal to the regressors. Numerically the way the solution is calculated is not using this formula. As I mentioned even in the least squares,

lectures on least squares. The implementation is done using a QR factorization which is numerically robust and efficient also.

Now let's talk about the regularized variant. I've already talked about the properties of this least squares estimator, we have talked about consistency, efficiency and so on. So let's move on to the case where some kind of regularization is important. So what is regularization?Let's spend some time on that. All right.So let's spend some time on what is meant by regularization. How many of you are familiar with the concept of regularization and parameter estimation? You're familiar? Anyone else? You also a familiar a bit. Okay. The others are new to this concept. Okay, fine.

So to understand legalization it's very straightforward. You start with your usual lease squares estimation. Right.So you generally you minimize this. This is the approach in least squares. And although I talk of regularization,the context of least squares. The idea of regularization applies to other estimators also. Okay. But it's easy to talk of the regularizationthe concept of, in the context of regularization.

If you look at this objective function its entire focus is on minimizing approximation errors only. Pretty much like in classical control. I am sure most of you have done a basic course on control. What is the objective in classical control? To maintain the output very close to the set point or minimize the distance of output from its set point. There we talk of control error. Here we talk of approximation error or prediction error. But as you know there is a strong duality between estimation and control. So what does a classical control do? The classical control places its entire focus on keeping the output very close to the set point. It doesn't bother much about the so-called control effort.

What do we mean by control effort? What is a prize that to pay for moving the manipulated variable. In control the manipulatedvariable movements are realized by what are known as actuators which are physical devices. They have to move up and down. And there is bill. There is a cost associated with that movement. There is a wear and tear associated. So many things that are associated with making their move. They are not coming for free. So over the period of time people started asking,"Should I only focus on this or should I also worry about the control effort." And that's when the optimal control was born which gave importance to both.

The deviations of the output from set point as well as the control, the extent of control moves that you're making. And in optimum control, the extent of input moves that you make is quantified typically by the energy associated with it. And energies are usually quantified by squared two norms. So now let's come back to estimation. In optimal control you also have an objective function which consists of two terms. And then the optimization problem is solved to get the so-called optimal input which has a good trade-off between control effort and the penalty for deviating from the set point. And you can give rates to both.

Likewise in estimation if the goal is only on minimizing prediction errors then the optimizer doesn't know whether you are achieving this at the cost of a very complicated model, that is an over parameterized model or a simple model. What is a model that you are using the optimizer doesn't know. It simply looks at the model that you are given and the objective function that you have specified and tries to drive the parameters so that this objective function is minimized. Pretty much akin to classical control.

Then as a user you have to do go through trial and error. You start with simple model, see whether, see what is the prediction that you obtain or approximation that you obtain and then you slightly sophisticate the model with one more parameter and see if there is some improvement in the

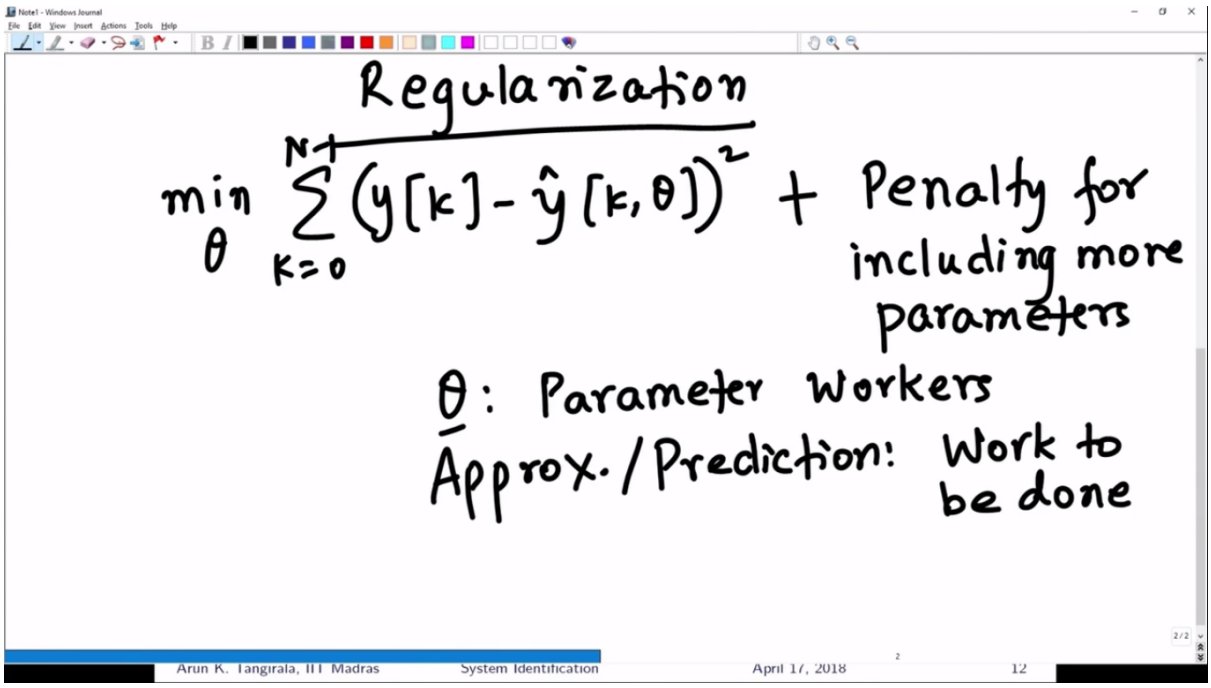
prediction that is in the fit and so on. So you do a stepwise manner which is very well known in stepwise regression also. The other way, the smarter way of doing it is telling the optimizer that,"Hey, look. I have this issue also to worry about. I want not only good approximations but also parsimonious models." What do we mean by parsimonious models?Models with minimum number of parameters.

Right now, the least squares objective function as we have written on the screen doesn't tell optimizer this story. It's only telling me that half, one side of the story.So you wanted to have one more term in your objective function which serves as a penalty. So here you want to have a penalty for including more parameters. And at this point a nice analogy helps. So think of thisthetas as some parameter power. That is, give some personification for them.

So they say parameter m or parameter humans or whatever it is. So you say parameter workers for you. Thetas are parameters workers.We don't worry about the gender. And why are we hiring these parameters in the model, what is the purpose? To get some work done. What is a work that needs to be done? Prediction or approximationthat is the work that needs to be done. Right? So approximation or prediction is the work that has to be done. Work to be done.

So you have recruited these parameter workers to get the work done for you. But do you think, they'll do it freely for you? Imagine the parameter [08:05 inaudible] for you. Do you think anybody will do things like this really for you? Nothing comes for free we know that. So what is it? There is a cost associated with hiring each of this worker, if you don't factor that in either it is assumed that you are a billionaire, trillionaire and you don't care or that you are unaware of the fact that there is a cost associated with.

(Refer Slide Time: 08:36)



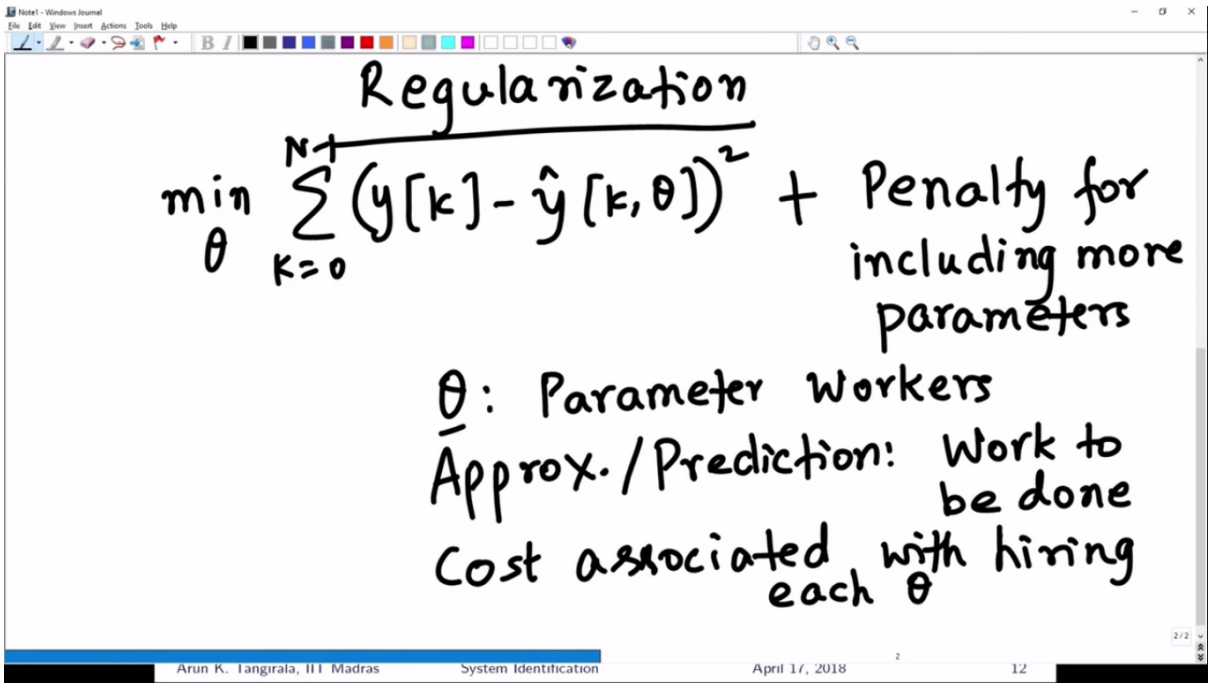
The reality is that there is a cost associated and the budget for this work is limited. Right? The information content in the data is limited. The information is not going to grow. It is not.

Why are we worried about the information? Because it is the food for your parameter estimation and these thetas are actually going to be estimated based on that information. More the parameters that you employ, more actually is the cost associated with hiring them.

As a result what happens you end up paying lower salaries?Right? You end up paying them lower money for the work that needs to be done. So what happens as a result, these workers go unsatisfied. They are not satisfied with the wages that you're paying. So what do you want to do now? One option is, if you want to satisfy a parameter worker, what you need to do. If you want to really satisfy, you hire only one person and pay him the full amount that you have. But at what expense are you doing that? You hire only one person. Let's say you have to shift homes from one house to the other home and you have to get your work done.

If you hired more people you can get more work done quickly but then you have a fixed budget. If you hire one person you can pay that one person very well but then you won't necessarily get your work done in the same constraint that you have. So there is an inherent trade-off that is the same trade-off that you have here. So there is a cost associated with each hiring each theta. Right?If you don't factor that cost then there is an issue. You may end up getting a lot of work done but you realize after the work is done, "Oh, boy. I don't have enough money to pay all of them in a satisfied manner." What does it mean in the estimation?In estimation it means that you may have, you may have employed a very sophisticated model, a model with a lot of parameters. They will do a great job of fitting the data for you but at the cost of large errors in your parameter estimates.

(Refer Slide Time: 10:55)



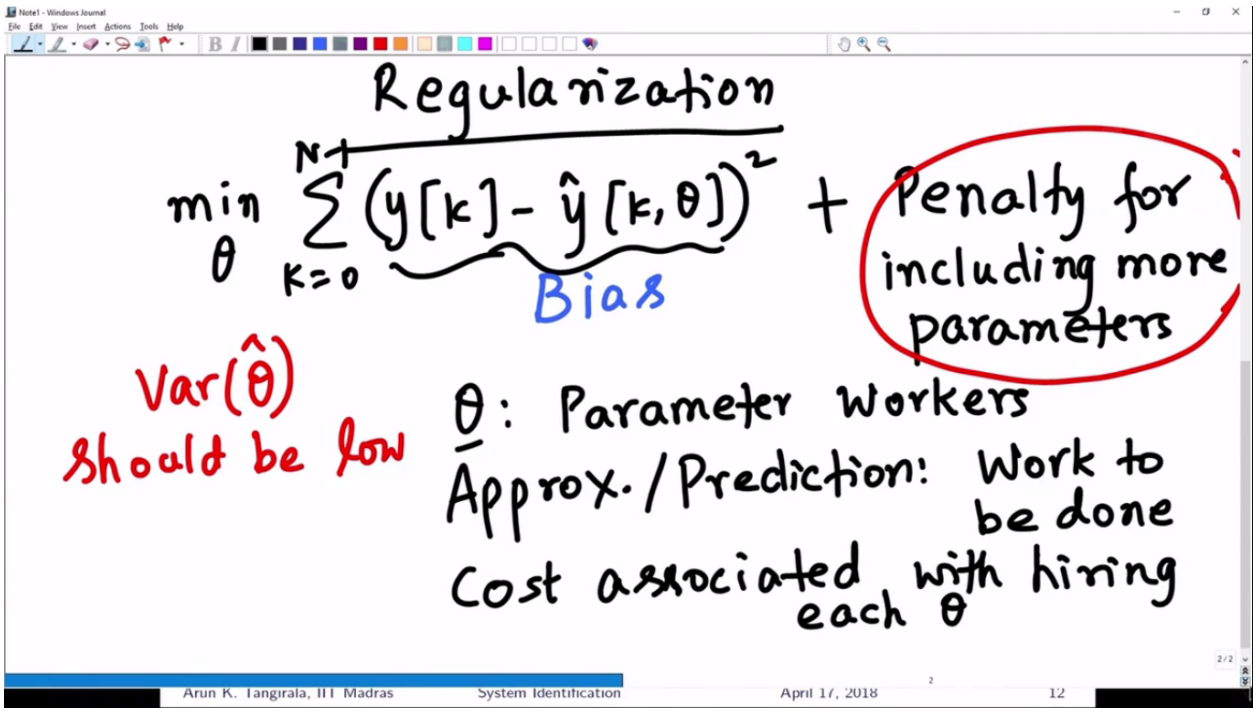
That means they go hungry. So you don't want that. That is the thing that, that is a trade-off that is involved.We say, there is a trade-off between bias and variance. The bias that we are talking about is this of.Sorry. The bias that you're talking about is this late. All right.So I am worried about this bias. I want to reduce this bias. But I also want the variance to be. Variance in what?Parameter estimate. So I want variance in theta hat also should be low. And unfortunately you may not be, I mean, it's not possible to reduce both. Just now as we discussed, if I want to completely eliminate the bias, that

means if I want a perfect fit. What do I need to do? I have to include number of parameters in the model. The dimensionality of theta will go up. Correct?

Then the variance will also shoot up. We know that variance of theta hat is directly proportional to the number of parameters that we have. So the essence, that the optimizer has to be informed both of the bias and the variance. And this penalty here should reflect that. It should reflect the variance or it should somehow reflect the dimensionality of your model. Somehow it should tell the optimizer that it has to search for the trade-off between the first term and the second. And there should be. They are conflicting with each other. And this is the idea. This is essentially called the regularization.

Now the hunt is all about finding a mathematical way of choosing this penalty function. Okay? So that is the basic idea.

(Refer Slide Time: 12:59)

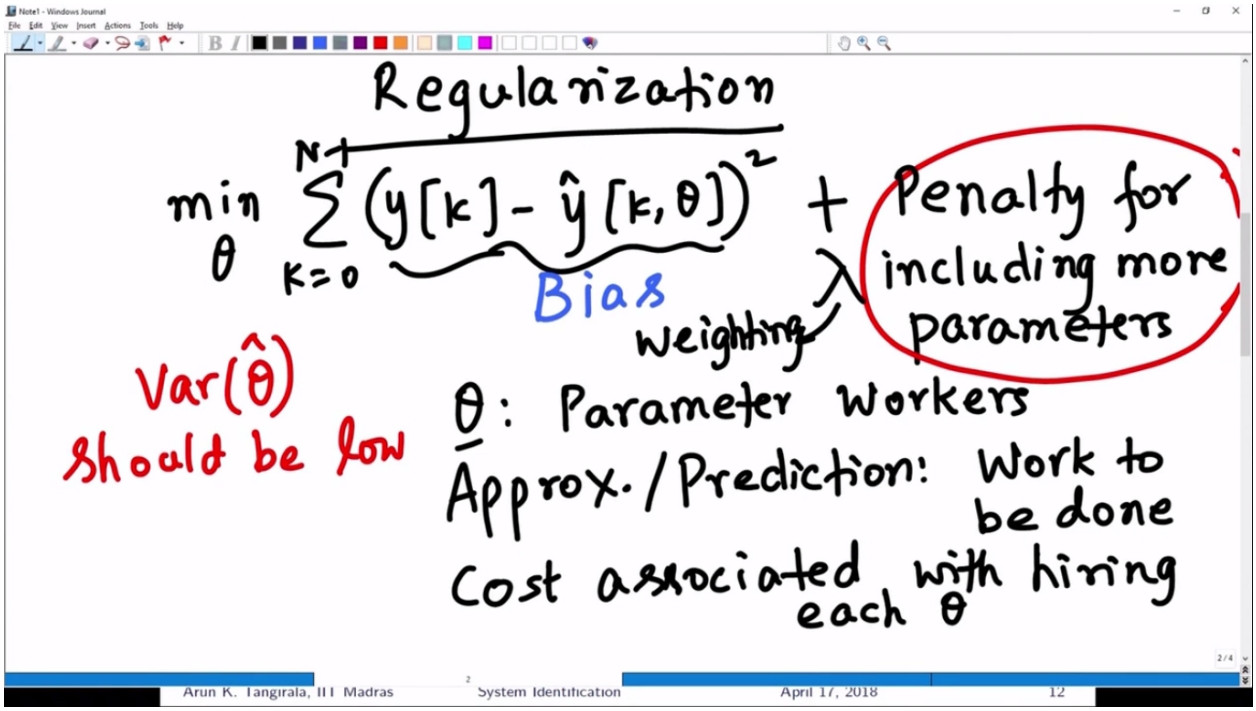


So when you talk of penalty functions, then you have one, the standard penalty function that was introduced long ago by Tikhonovis. So let's call the penalty function as some f of theta vector. And f of theta, you can choose for example, to be the squared two norm of theta because that denotes the energy associated with this theta and as you increase the theta typically you should expect this penalty function also to increase. Correct? And by the way I have also written previously. So in the previous case, I have also, I've just included qualitative penalty here.

But you have to give weightings to both these teams. Normally you have a lambda associated with this which is the relative weighting for this penalty. If lambda is too high then you're telling the optimizer I'm heavily constrained on the budget that I have. That means the optimizer should give, real,much more importance to the model complexity than to the minimizing the prediction error. If lambda is very low then that means you are giving more importance to getting your work done.

And you saying, it's okay. There is some constraint but it is not so heavy, not so serious.

(Refer Slide Time: 14:39)



So this is, when you choose f of theta as this is called the Tikhonov regularization. Take an old regularization. And it was introduced by Tikhonov long ago. However.So let me point out the merit and demerits of this, the merit of choosing this penalty function is, it keeps the objective function convex.

Remember that is also another requirement that we have from a practical viewpoint. We want to solve convex optimization problems. The least squares objective function as it stands is convex. There is no issue with it. The linear least squares to begin with. When we add a penalty function we should also mean make sure and make every effort to preserve the convexity of the objective function. And choosing a penalty function like this may maintains a convexity. So that the objective function has a unique minimum and all the nice properties of convex optimization problems apply to this situation as well.

So that is why this penalty function became very popular and achieved the desired results. But it fell short of doing one thing which is that it does not necessarily drive those details thetasto zeros that I have unnecessarily included possibly. See that that is another requirement. I would like the optimizer to automatically like IT field we have people who are working and people who are on bench and people are laid off. The least squares that this with the Tikhonov regularization perhaps puts them on bench but doesn't lay them off. Doesn't tell them, give them the slip and say your services are no longer required. It doesn't tell that. So we want the optimizerin other words to serve as a selector.

Ideally man wants to be as lazy as possible. We want the algorithm to do things for you. Instead of doing a trial and error we are now doing it the more automated way by including a penalty function with the hope that the optimizer with select only those parameters workers that are necessary for your problem. So it does not serve or you can guess it does not select. Necessarily of course, the appropriate or the correct thetas, correct parameters that have to be included in the model or to be retained in the model.

So I may have put in maybe 200 parameters.But I want to optimizer come back and tell me,"Hey, you included 200 parameters. But I found only 10 parameters are necessary. I have driven the remaining

190 to 0." This method does not do that because, what it is trying to minimizes, it is trying to minimize the energy. And in doing so, it may assign small values to theta not necessarily zeros. Because that can also achieve minimum energy for theta you are saying minimize the energy of thetas and say, "Okay. They are not completely put to sleep. I'm actually going to keep them mildly awake so that you don't have to feed them too much food. Maybe some juice is enough." But you still have to supply juice to them, some kind of food.

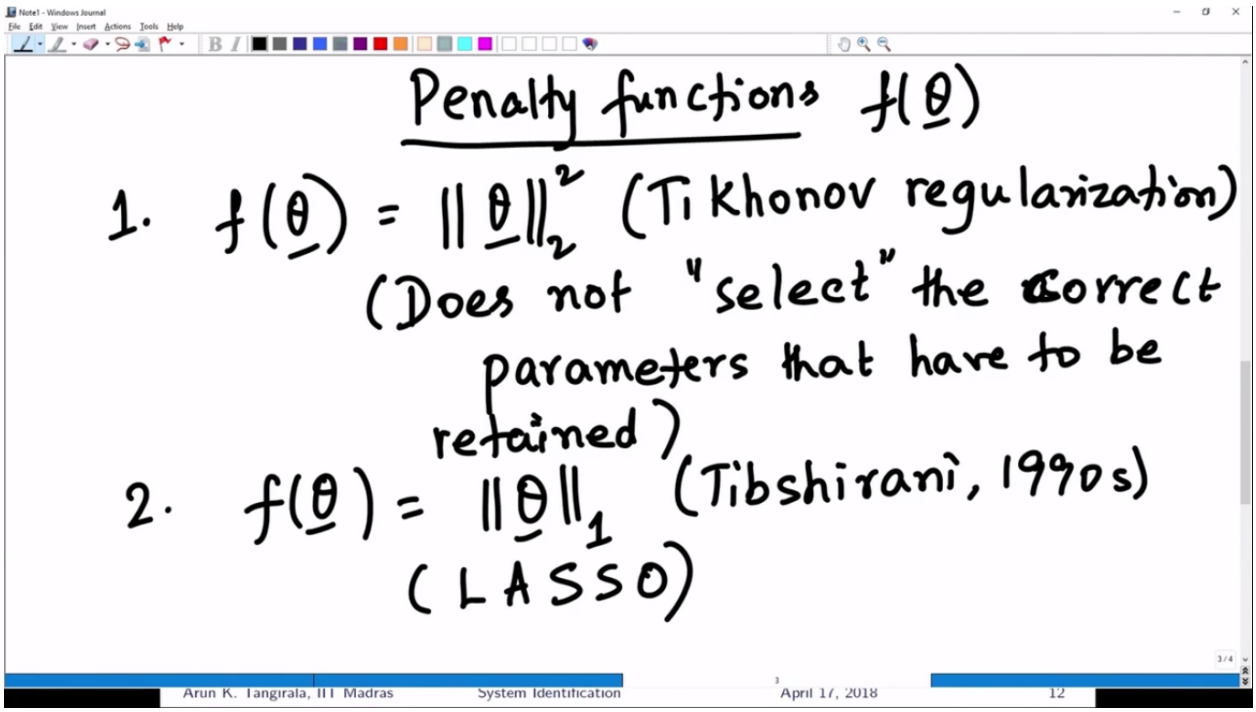
Ideally you want to optimize that to come back and say,"Hey, they're all sleeping. You don't have to feed any of them. Only those are awake have to be fed." So this was the problem that was studied and later on a new penalty function was proposed which is the one nom of theta. And this was by, this was a work by Tibshirani in mid 90s.Okay? Around 1990s, 96 like that. Tibshirani, he works at Stanford. And what he showed is that with this finality function you can now use this to select variables. Sorry parameters. Okay. Although, he calls it as a variable selection.

You can think of it as a variable selection or parameter selection. Doesn't matter because, if parameters are turned off then the associated variables in the model are also turned off. So he called this variable selection method but you can think of this as a parameter selection method as well. And he named this as LASSO. LASSOstands for least absolute shrinkage selection operator. Least absolute shrinkage you can see. Right? Because in the one norm of theta is essentially some of absolute values of theta, you are minimizing that, therefore it's least absolute, shrinkage. You're shrinkageit.

Not only I shrinking in the process of doing that you are also, now you have created a selection operator which is what you are looking for in many problems. Yes. So he proved that to complete. So he proved that with this penalty function if the true data generating process, so imagine that let us say, in the context of a FIR models, let us say it’s a fifth order FIR model that is the data generating process but unknowingly I have, I'm trying to model this as a fiftieth order FIR model.

So if you compared with the true data generating process 45 parameters has zero value. Right? The Tikhonov regularization does not necessarily drive those 45 additional thetas to zeros.

(Refer Slide Time: 20:37)

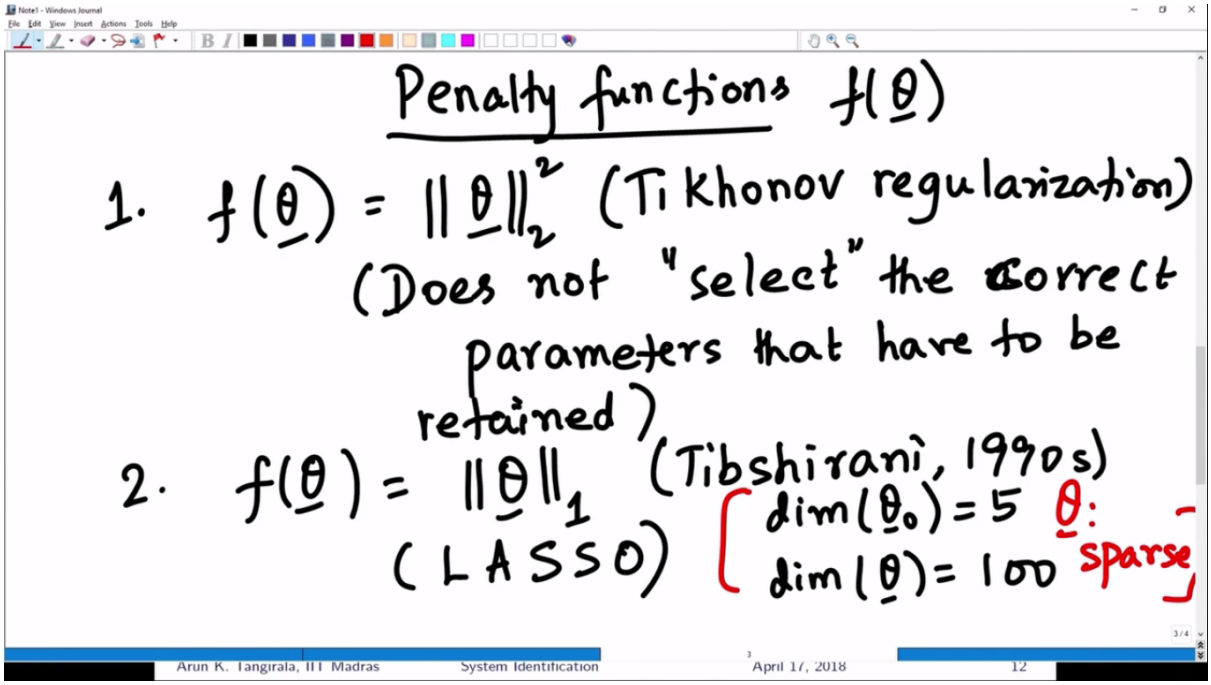


Theoretically, okay. Whereas with this penalty function the optimizer will search for thetasin such a way that ultimately those 45 extra thetasthat you have included will go to zero. Under some conditions of course but that is what was proved. So the proof exists. So that's the main difference. That means it has selected exactly those five ones for you.

Provided the true one is commensurate with the model that you have chosen. In reality what will happen is, sorry the true one may not exactly match, we know with the model selected because the model is after all an approximation. Then you will get what is known as a Sparse Approximation.So we say here that with respect to the true theta for the example that we just talked about. Let's say that the dimensionality of theta not is only 5. That means there are only five unknowns to be estimated. Whereas the dimensionality of your theta that you have chosen your model let say is 100.

So if you know compare. And if you ask, how many compare with the truth theta not, compare theta with true theta not, only five of them are non-zeros. The remaining 95 are zero value. So we say that in this case theta is Sparse. Right? So we say that theta is Sparsein this example. Sparsemeaning most of them are zeros. And that is where this entire Literature onSparse optimization took off. But of course there is another school of,there is another approach that led to Sparse optimization that is from a signal processing viewpoint will not touch based on that.

(Refer Slide Time: 22:34)



We are only knocking the doors of Sparseoptimization through the LASSO channel. And for those of you who have done courses on machine learning and so on you must have heard of LASSO. It's not new to you. But in the context of system identification what is the role of LASSO,regularization. That means if you have a large number of parameters in your model and you have unknowingly included thenLASSO will come to the rescue. What does Tikhonov regularization do for you? Tikhonov regularization does a different kind of regularization where it doesn't drive those unnecessary thetas to zeros but two very small values depending on the value of lambda or the rating that you have chosen.

Now regardless of the regularization that you choose, the fact is that now you are seeking a trade-off between bias and variance. So you compare the regularized one with the un-regularized one. Then what is it that you gain by regularization? What is it that you lose by regularization? What do you gain by regularization is lower varianceof the parameter estimates. Because it will try to include only that many parameters that are necessary as much as possible while keeping the remaining thetas at bias.

In that process it reduces the variance but at the increased bias. That means you may not actually get the same level of approximation error that you get with the un-regularized one. Naturally, because it's like a constrained optimization problem.Regularization is more or less like a constrained optimization problem. Right? For example if you go back to this objective function that we had earlier.Although, I have written this in this form for those of you who have taken the basic course on optimization you can straightaway lambda is a Lagrange multiplier. Right?Which usually comes about, when you're solving a constrained optimization problem.

So essentially the regularization is like solving a constrained optimization problem and the solutions, the optimum that you achieve with constrained optimization problems are definitely not as good as the one that you get with unconstrained optimization problem. So that is the story here. So let me just quickly go back to our discussion. This is what we have discussed, the only difference between the equation 9 and what we have chosen. So here we have fallen back to Tikhonov regularization. Where the two norm is being represented as theta transpose theta but there is also a D. There is a weightingmatrix D. This waiting matrix D ensures that, that allows you to give different importance to different thetas.

So if you know something Apriori, if you know that the first few impulse response coefficients are certainly going to be important to you than the last ones. Then you can incorporate that information or additionally you can incorporate other prior knowledge that you may have, which I will talk about briefly in the next lecture. This prior knowledge that you may have is, maybe uncertain and you have to turn or you have to give this a Bayesian flavour. In fact, this regularization approaches can be given a Bayesian flavour where you can think of this additional term that you have theta transpose D theta as in the Bayesian framework giving some prior knowledge.

Remember Bayesian estimators. Take some prior knowledge. Then take your data pass the prior knowledge through the data so that the posterior is less certain. That is the fundamental difference between the philosophy of Bayesian estimators and the remaining three estimates that you have learned. In Bayesian estimators you assume that prior knowledge, some prior knowledge of theta is available. It is not that I don't have because in these squares MLE and method of moments, you pretend that you don't know anything about theta and you want to get everything from the data.

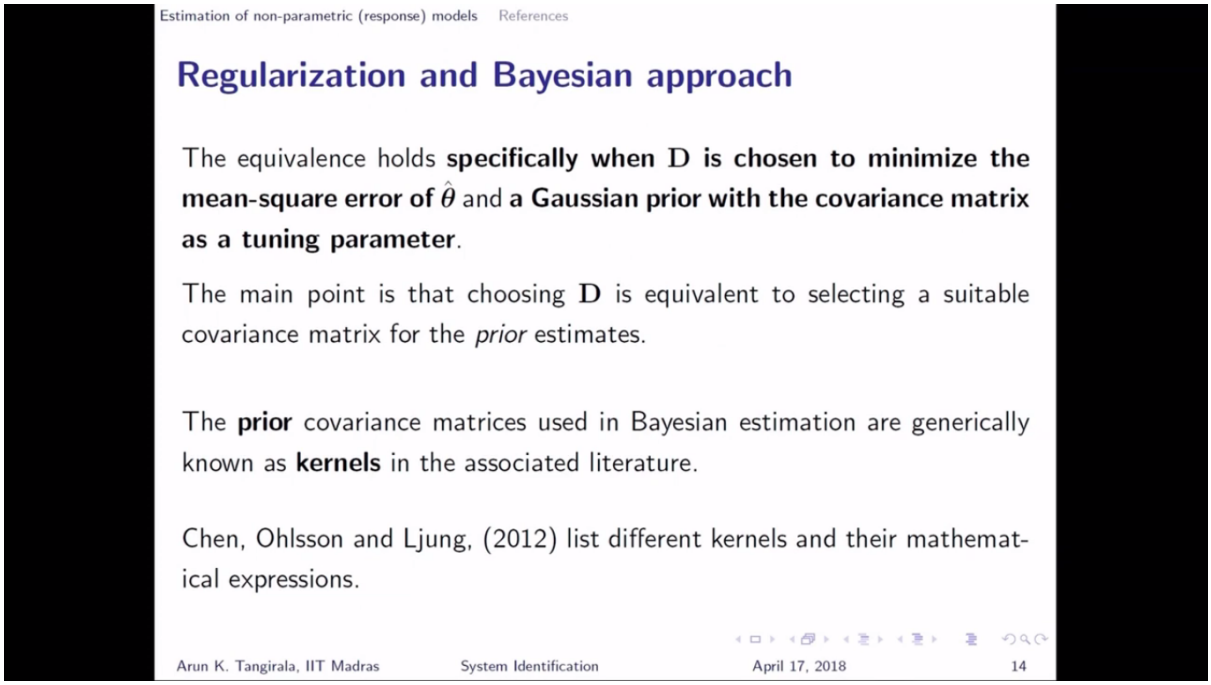
Whereas in Bayesian, you say that I may have some prior knowledge which is also very practical. But that prior knowledge is not accurate. It has some uncertainty in it. And Bayesian estimators allow you to incorporate that prior knowledge. And also take data, fuse the prior knowledge and data together to get you a better estimate. Better estimate meaning the final estimate is also uncertain. We know that it is whether you use MLE least squares or Bayesian, the final estimate that you obtain has errors in it.

But the nice thing that happens in Bayesian is right from step one it acknowledges that your knowledge is uncertain and will remain uncertain with fixed number of observations. That is what is the philosophy Bayesian estimators. Now what you can show is that the regularization, Tikhonov regularization particularly that you see in equation 9 which as an analytical solution equation 10, is equivalent to doing a base in estimation with so-called Gaussian prior for your thetas.

And the fundamental point that you should keep in mind is. Two points.One, Bayesian philosophy rests on the assumption that thetas are random. Whereas until now with the least squares and so on we have maintained that even in MLE that thetas are deterministic. So philosophically there's a world of a difference and therefore some people don't like this analogy. This equivalence, although mathematically they're equal.

So the equivalence of this technology globalization with Bayesian estimation and Gaussian prior is only mathematical not philosophical equivalent.Philosophically they assumed completely different things. So that is something to keep in mind. What the recent works have done is, they have taken this mathematical equivalence keeping aside the philosophers and say you can debate forever,don't let's not worry about it. Mathematically they're equal. Let us exploit that mathematical equivalence and design that weighting matrix D.

(Refer Slide Time: 29:10)

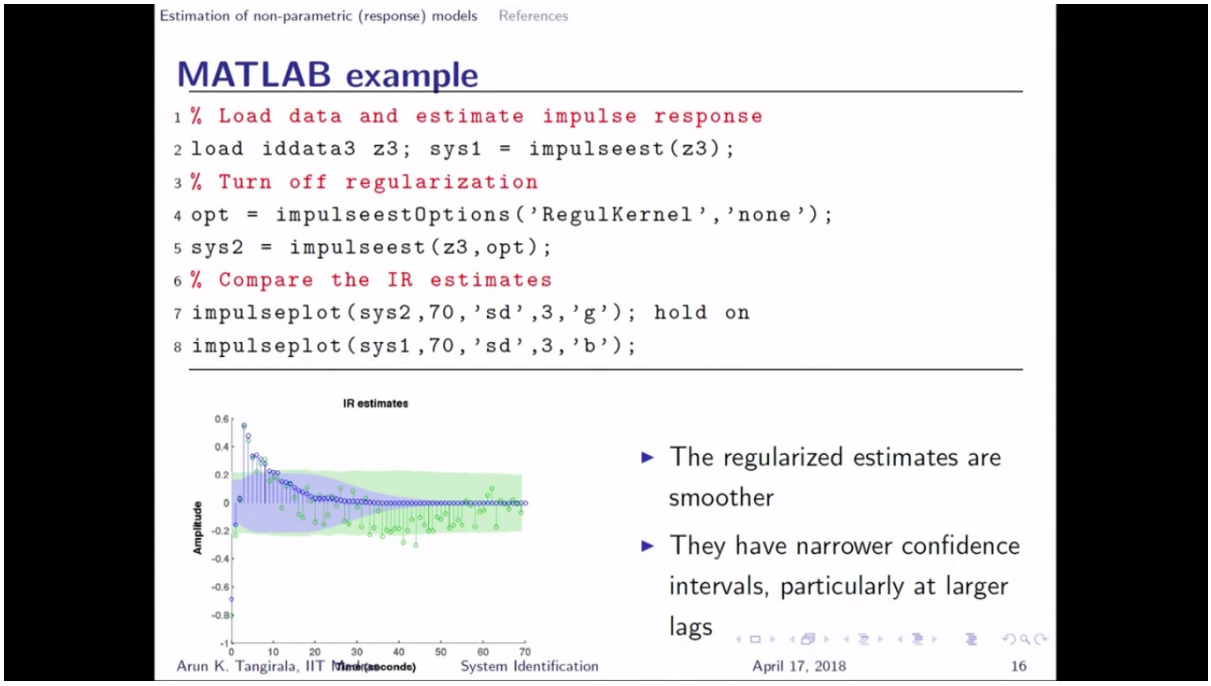


And they've used this Bayesian approach to design the D. And what you'll get eventually is a nice word you say an estimation methodology based on what are known as prior covariance matrices which I shall talk about in the next lecture. But I just want to end this lecture with this example here and technicalities of which we shall talk about in the next lecture.

So here what I have done is I have just loaded the data from iddata3, it'sa dataset that's available in the [29:28 inaudible] tool box and I'm showing you impulse responses estimates with and without regularization. So with regularized-- There are two bands that you see, a green band and a blue band. Right? And also a green colored set of estimates and blue colored set of estimates. The blue colored ones are with regularization and the green colored ones are without regularization.

You can see that you obtain as smoother estimate with regularization and the variance of the estimates are much lower than the variable. The bands are essentially showing you the errors, error bands. The error bands are much wider with regularization. But as I told you at the expenses of some slightly higher bias that is what your impulse, the regularization helps you. I've give you the MATLAB commands. Impulse set options is through which you control the inclusion of regularization or omission of it.

(Refer Slide Time: 30:37)



So what you want to do is, you want to, if you look at line four, it says regularization kernel is set to none. That kernel is what I'll discuss in the next lecture but setting that to nun means you are switching off the regularization. That is what is. Yes.

STUDENT: We will have a higher bias with regularization but will that bias go to zero as end [30:50 inaudible]

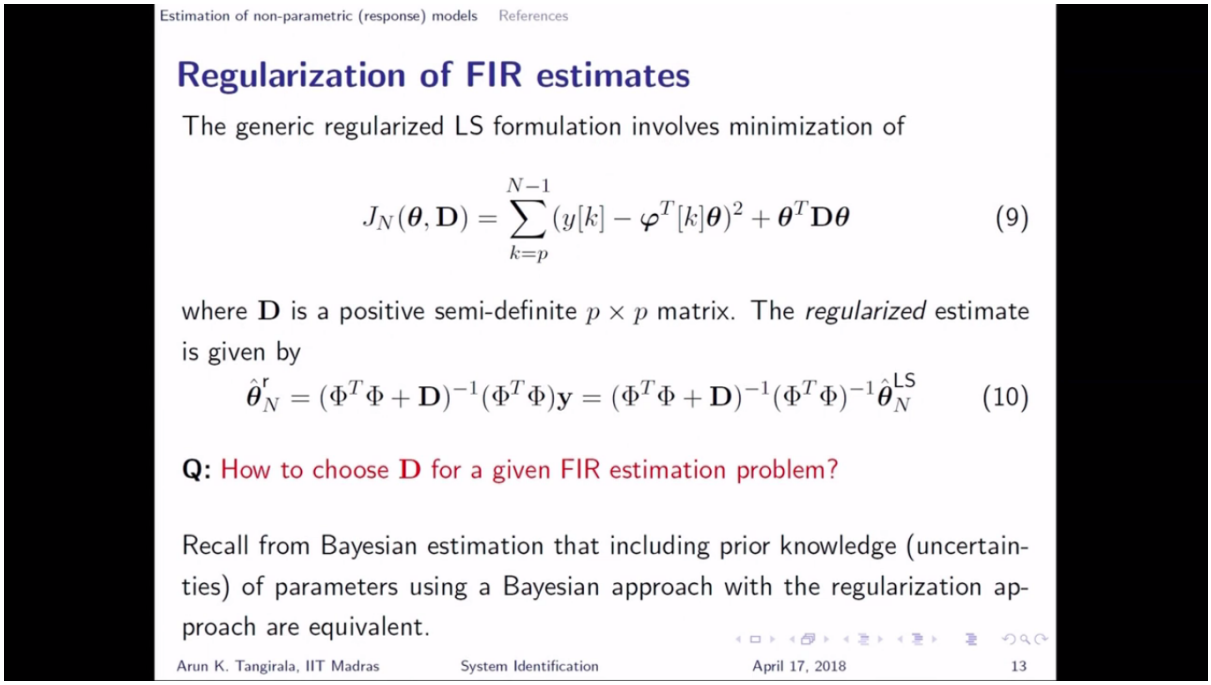
Asymptotically the bias can vanish. Yes. But for finite samples it has a larger bias.

STUDENT: By adding that extra term we are losing the analytical expression which was guaranteeing asymptotic [31:04 inaudible]

So the asymptotic, the analytical expression is this. I don't have the expression for variance of theta hat. I will answer your question probably in the next lecture. The expression in 10 only gives you analytical expression for theta hat. But you can also derive an analytical expression for variance of theta which will probably answer your question. Right? Whether as N goes to infinity the bias will also goes to zero. Right? For finite samples it has a larger bias. Certainly. Bayesian estimators are like that.

For example, Bayesian estimators, they have large biases for small observations but the bias goes to zero as N goes to infinity.All right. So which means that these regularization methods will produce larger bias for finite observations but that bias will go to zero as N goes to infinity. Yes. Which one? What is the issue? Here, it should not be Phi-- So what is the issue with the matrix this thing.

(Refer Slide Time: 32:18)



It should be Phi transpose y. I will correct them. Yeah. I will do it. Thanks. I will correct that. I will just check that. It should be Phi transpose y. There should be no Phi here. We will correct that. But the next part of it is correct. Okay. Thanks, I will correct that. Okay. So we end the lecture here.