CH5230: System Identification

One step and multi-step

ahead prediction 1

Very good morning. What we have learned yesterday are the different parametric model structures. And it's not just knowing the names as I've said yesterday. What is important is to know the implications of choosing a model structure. So hopefully now we know that, when I choose an ARCS verses an OE model or a [00:40 inaudible] model. What are the implications in terms of the ability to explain the process, the parameterization that I'm choosing.And finally the ability to recover the truth.Right. And we have learned some key points yesterday. Towards the end we learned a very interesting fact that essentially although there are different model structures, seemingly different model structures but one model structure can be expressed as another model structure but unfiltered data.

Now this filtering is a very commonly used step in data analysis. In not only inidentification but also in other applications. Essentially why would you filter any data because you want to focus or you want to emphasize on certain frequency ranges.So in identification if I filter the data prior to identification. I am implicitly saying or even explicitly saying that I would like to build a model which explains the process very well over a certain frequency range. And why would I be interested over a certain frequency range because in that application the frequency ranges of interest only that band which I have chosen. Okay.

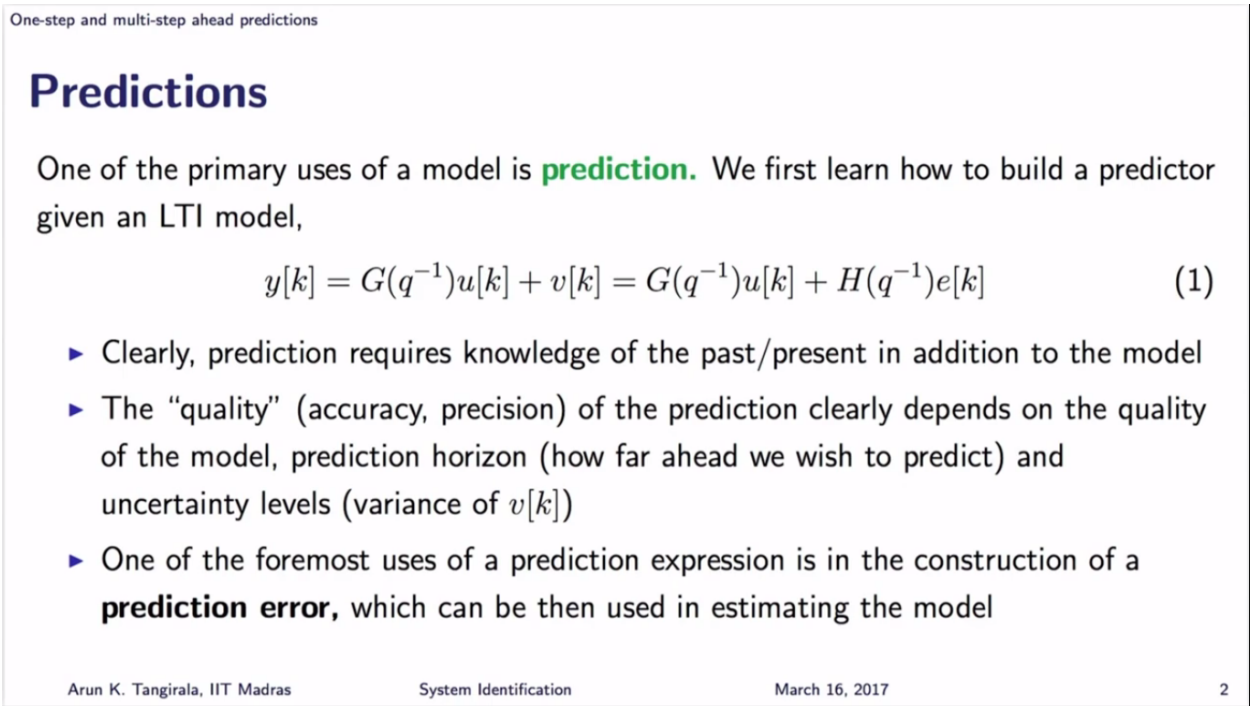
So whenever you filter the data, prior to identification, we call it as pre-filtering. And what we learned yesterday is that for example an OE model can be thought of as an ARCS model on pre-filtered data, right. Or an ARMAX model can also be thought of as an ARCS model on pre-filtered data. So this equivalence you should get used to and the main message is that you should get is changing the noise model amounts to changing the pre-filter, which is a very, very beautiful relation in many ways. It, as I said yesterday it allows us to not only view one model structure as another model structure on pre-filter data with a pre-filter being determined by the noise model. But also allows us to carry out some theoretical analysis, develop some nice algorithms like the Steiglitz-McBride algorithm and so on. Those algorithms although not necessarily as optimal as nonlinear least squares or this class of algorithms that we shell learn known as a prediction error method. They generate very good guesses. See, ultimately we're going to solve nonlinear least square problems except for FIR and ARCS model structure, the remaining model structures call for nonlinear least squares algorithms. Why is that? I'm sorry. Predictors models. We we're not talking about models.Thepredictors. So you are right. So predictors are not linear in parameters and therefore we invariably end up solving nonlinear least square problem. And all this nonlinear least square problems require a good initial guesses. We know that.

So algorithms that like the Steiglitz-McBride algorithm and so on, they actually generate good initial guesses. We'll talk about that a bit later. Ishowed you the estimation of a time series model as an example in MATLAB. When I was estimating an AR model I used AR routine but when I was estimating the

ARMAX model, I was using the ARMAX routine. Why is there a separate routine for estimating auto regressive models because whether it is ARX or AR the predictor is linear in the parameters, right.

The moment I have a moving average term as you will see today in predictions, you'll find that the predictor is non-linear. Even we saw yesterday. So let's now learn some more about these models structures in particular this how to develop predictions given a model.

(Refer Slide Time: 04:51)



And that is extremely important because of two reasons. Suppose I have estimated a model,or suppose I'm being given a model. What is ultimate use of a model. As far as we know its prediction.How do I construct an optimal prediction given a modern?That is a first reason, why we want to formally study this prediction. The second reason is given data I want to estimate the model and if I am going to estimate a model such that the prediction errors are minimized. I need to know given the model again what would be the prediction and the prediction error. So for these two reasons I need to know how to construct an optimal prediction given a model, all right.

Now if you look at the prediction theory. It's very old subject. Because predictions have been of interest to human beings for a very long time, although not documented you can guess that man must have been interested in forecasting or prediction from a lot of centuries or maybe even thousands of years. And that's

how astrology was born and so many different signs of forecasting were born. But as far as a documented literature is concerned mathematical and statistical approach to prediction is concerned. It's about I would say 100 to 250years old or maybe you can stretch it to 200. But a lot of activity happened in the early 1900's. What about literature, open literature that you see. And primarily people like Kolmogorov, Wiener,Cramer all these peopleactually have contributed a lot to prediction theory. And that's in the time series and literature was born.

So, what is the problem of prediction? Given one variable predict another. That is one way of looking at it. A more generic way of looking at it is I'll give you some information about the process in the past and you are supposed to make an estimate or an intelligent guess of what's going to happen in the future. That is what this prediction. Now this information is kind ofvery generic. What do you mean my information, typically measurements for us, for engineers it's typically measurement of a particular signal or a bunch of signals. In addition, I mean, this is the basic prediction problem. That's how it began. Then people started building models and then said we will give an model and the data, how do I make a prediction. But the original prediction problem began with given information up to a certain time point make an intelligent guess of the next instant, right. And there are lot of milestone results in prediction theory. We studied already one milestone result, which is that, what is that milestone results that we learned, when it comes to prediction? Can you recall. Sorry. Conditional expectation, right.

Conditional expectation, when I consider two random variables then given the knowledge of one random variable the best prediction of the other random variable is simply the conditional expectation. Correct. So we're going to revisit that result and then move on to asking the more relevant question which is that given this LTI description for y, which means I'm given G andH and it's understood I'm given the input. Otherwise there's no point in asking this question. And measurements up to K minus 1. I would like to predict y k,y k plus 1, y k plus 2 and so on.

(Refer Slide Time: 08:47)

Predictions

One of the primary uses of a model is **prediction.** We first learn how to build a predictor given an LTI model,

$$y[k] = G(q^{-1})u[k] + v[k] = G(q^{-1})u[k] + H(q^{-1})e[k] \quad (1)$$

- Clearly, prediction requires knowledge of the past/present in addition to the model
- The "quality" (accuracy, precision) of the prediction clearly depends on the quality of the model, prediction horizon (how far ahead we wish to predict) and uncertainty levels (variance of $v[k]$)
- One of the foremost uses of a prediction expression is in the construction of a **prediction error,** which can be then used in estimating the model

So that is a more relevant question to us.

(Refer Slide Time: 08:50)

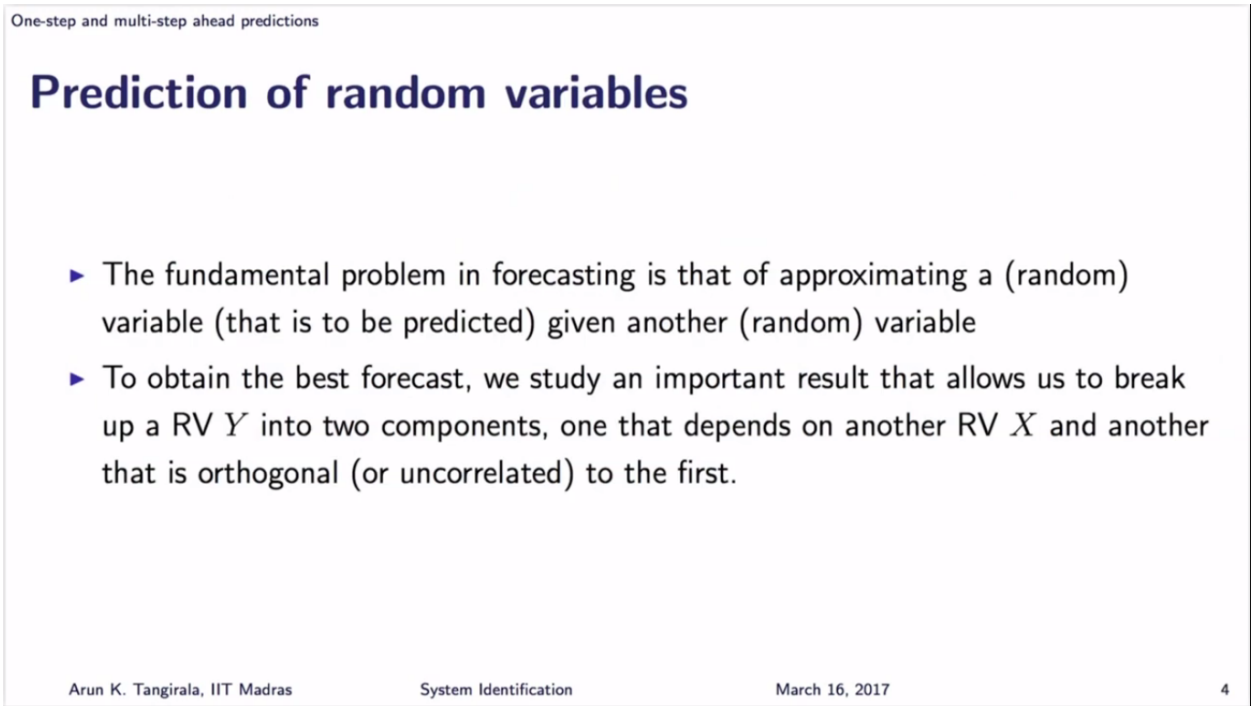
One-step ahead predictions

Predictions are often denoted by a hat; for e.g., $\hat{y}[k|k-1]$ should be understood as prediction of $y[k]$, given all the information up to $k-1$.

- In making predictions, we assume inputs are known accurately.
- Given a model (i.e., for a fixed model), the one-step ahead prediction
- The task therefore is to build a predictor for $v[k]$
  - In practice, we do not have information on $v[k-1]$, but rather have measurements $y[k-1]$ and inputs $u[k-1]$
- We first learn now **how to predict a stochastic signal** given its past/present.

So let's take. We'll come back to that.

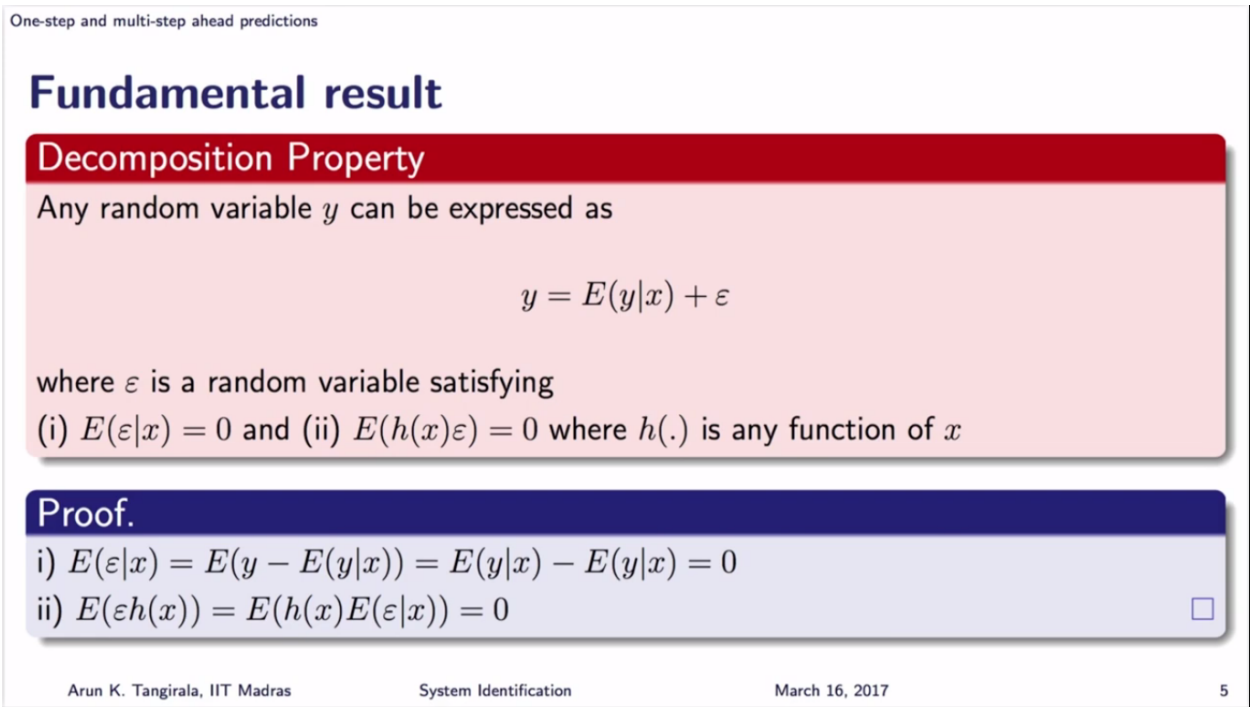
(Refer Slide Time: 08:54)



Let's actually look at the fundamental result in the theory of random variables when it comes to prediction. This is what we have just mentioned briefly. When the simplest problem in prediction is given one random variable X, what is the best prediction of the other random variable Y?Now, the first thing that we should observe because we have said both are random variables clearly whatever prediction I make, let us say I'm going to predict Y given X, whatever prediction I'm going to make of Y is going to fall short of the actual one because it's a random variable. But there are many outcomes possible. I will make some prediction but that may not be the case. Right. Which means that there is going to be some error between the prediction and the truth.

So Y is the generic outcome and Y hat is the prediction. Typically we say Y hat of X to indicate that I'm using X to make a prediction. This is the prediction error in Y. And when we seek best prediction error, we have to say what we mean by best. In this case, I'm going to find that prediction which will minimize the squared error between the outcomes and the prediction in a statistical sense average square error. We take expectation because there is not one possible value for Y. There are many possible values for Y, among the many possible values, with respect to all possible values for y, my prediction should stand very close or minimize error in a mean square error sense,right.In the mean square sense. Okay. So that is what is the primary requirement.

(Refer Slide Time: 10:46)



There are other criteria that you can employ as I said. So what this tells you is, whenever I'm making a prediction two things I realize. We are not talking of predictions of deterministic processes because that is not so much of interest and it's not so exciting. What is exciting is the prediction of random phenomena because I know I won't predict it accurately but I want to predict as accurately as possible. That's what makes it challenging, that's what gives rise to so many PhDs and so on. So that is the first thing I learned that I will never be able to predict accurately. Secondly, that I can choose the criterion of accuracy, or criterion of goodness of prediction. This is one such thing that I have written on the board which is very common. We call that as the mean square error, and if we minimize this, that I had that minimizes this is called minimum mean square error prediction, right. Find Y hat such that this is minimized.

Now before we go further,I also want to impress upon you that prediction it nothing but an estimation problem. Any prediction that you make is an estimate. But it's an estimate of what's going to happen in the future. So with respect to the time horizon that we have, so this is the time axis. Let's say, of course, we are not talking of signals yet but very quickly we'll talk about signals. If I'm standing at K and I'm given all the information up to K. So this is what I'm given. And I want to make a prediction at K plus 1,K plus 2 and so on, K plus 3.Then if I'm estimating what is going to be the signal at these instant in time then we call that as a prediction problem.

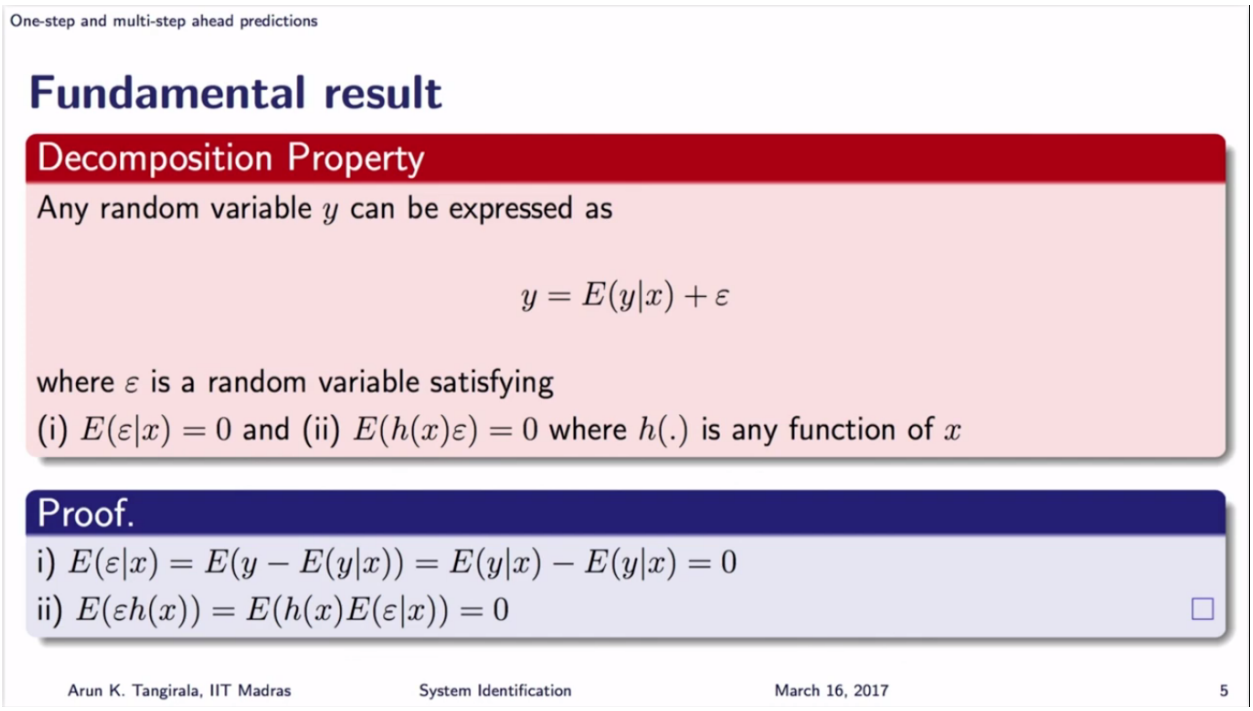
Later on we will learn there is something called filtering problem and a smoothing problem also. So the prediction problem it turns out that it is a subset of this grand problem of estimation. See the other two kinds of problems that I can have are given information up to K, estimate the truth at K. Because we assume in many cases the measurements come with error, which is a filtering problem. The other problem which is that of smoothing is given information not only up to K but also K plus 1, K plus 2 estimate what is the truth at K, which is smoothing. So, if you remember I gave you an example of going to a movie hall, right. If you had listened to the dialogue until now and making a prediction like, you know, [speaking in Hindi], even you know, the next those standard dialogues. Then you are making a prediction, right.

But if you missed out hearing on something or you heard a corrupted version, there's a lot of noise because you know fans are excited first day. I have known many who have, they religiously go on the first release. I mean in a sense it has to be there. They have to be there otherwise it's like the producer will be highly disappointed. So their lot of fans making noise and you are unable to hear what you wanted to hear. And you heard clearly before and after. And you want to estimate what happened at that time. This is smoothing. If you do notrelay on the dialogues but you relay on all the dialogues up to that instant and you want to, you have not heard clearly what is at K and you are estimating what happened. And yes. That is filtering problem free.

So this prediction, filtering, smoothing, put together is a grand problem of estimation. Any guess is an estimation problem. Correct. Prediction is one class of the estimation problem. Okay. So let us get back to random variables and then we'll come back to random signals. Right. Always we have done that.

Now there exists a fundamental result in the random variable world. It's a very beautiful result. It's called the decomposition property. On the face of it, it looks quite simple, the nature of the result. So, what does it say. It says given two random variables, I can decompose this random variable Y in two components. The first one being the conditional expectation and remaining is a residual.

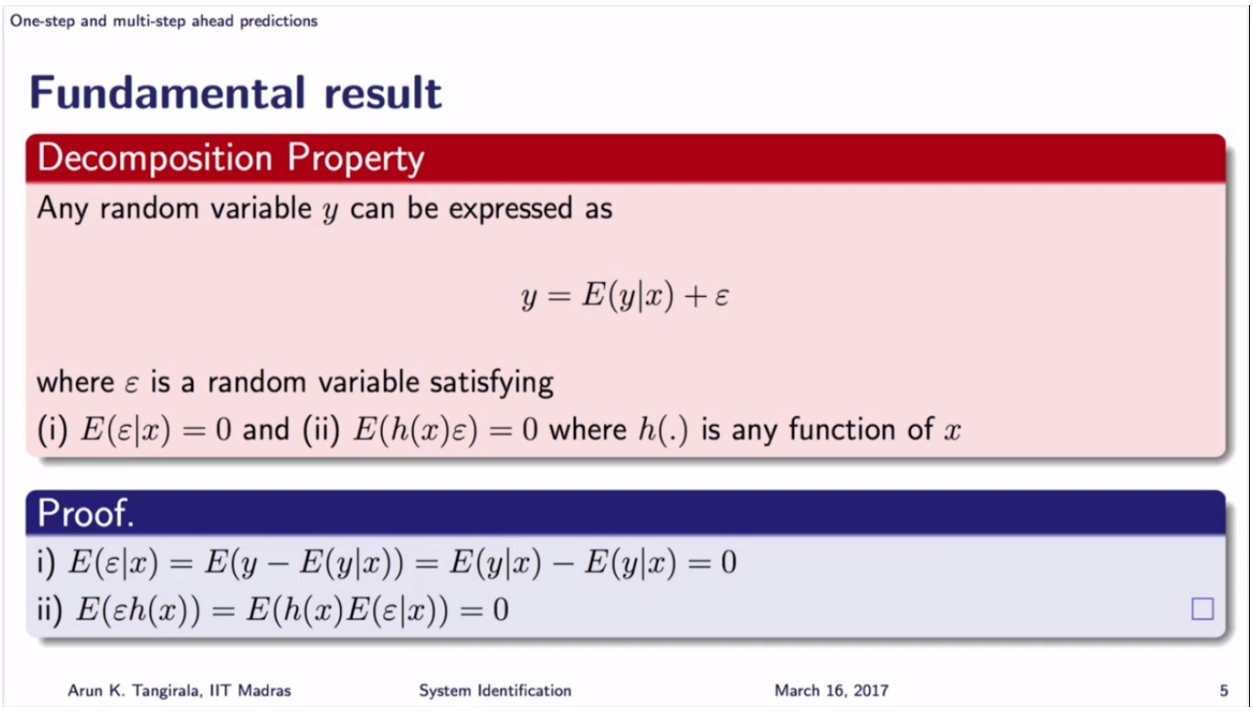
(Refer Slide Time: 15:44)



Nothing is being said about prediction here. It's a very simple decomposition result. What is so unique about this decomposition.One that the residual is such that the conditional expectation of the residual is 0.Okay. Conditional expectation of epsilon with respect to X is 0. Which means what epsilon, if you were to predict epsilon using X, in a minimum mean square error sense, you won't get anything out of it. Correct. So loosely speaking epsilon does not have too much dependency onX. I wouldn't say independent of X. Independence would mean a stronger statement, correct. At least I know it's uncorrelated.

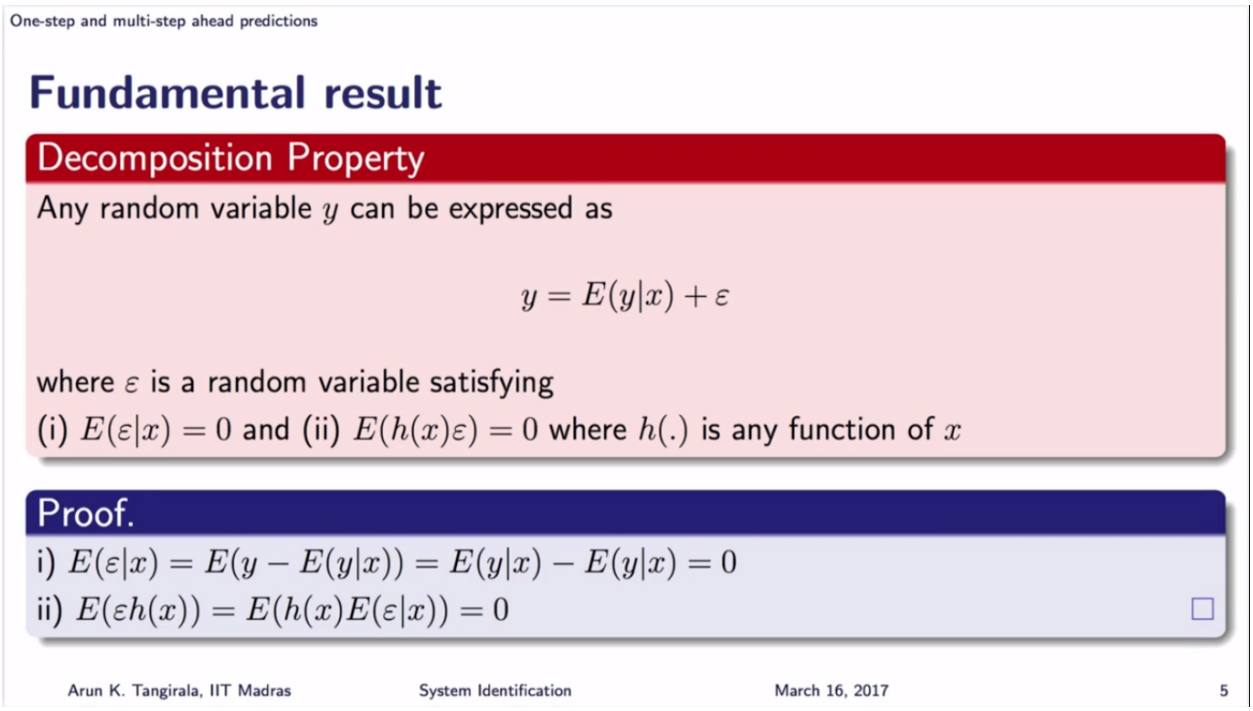
Secondly, the expectation or let me put it this way, the correlation between any function of x and epsilon is also 0, although it doesn't sound like a correlation here. But let us assume x and y to be 0 mean. So what does it say here, expectation of h of X. It should be capital Xthere, it should not be big X in fact.The ones within the bracket for conditional expectation should all be uppercase. As I will correct that notation.

(Refer Slide Time: 17:02)



So the second property of this result is that the expectation of h of X times epsilon is 0. What does it mean? How should I read it as the covariance between epsilon and any function of x is 0.h is any function. Now what does that tell me? It tells me that epsilon is not related to any function of X. Which means practically there is nothing in epsilon that I can actually predict using X.

(Refer Slide Time: 17:34)

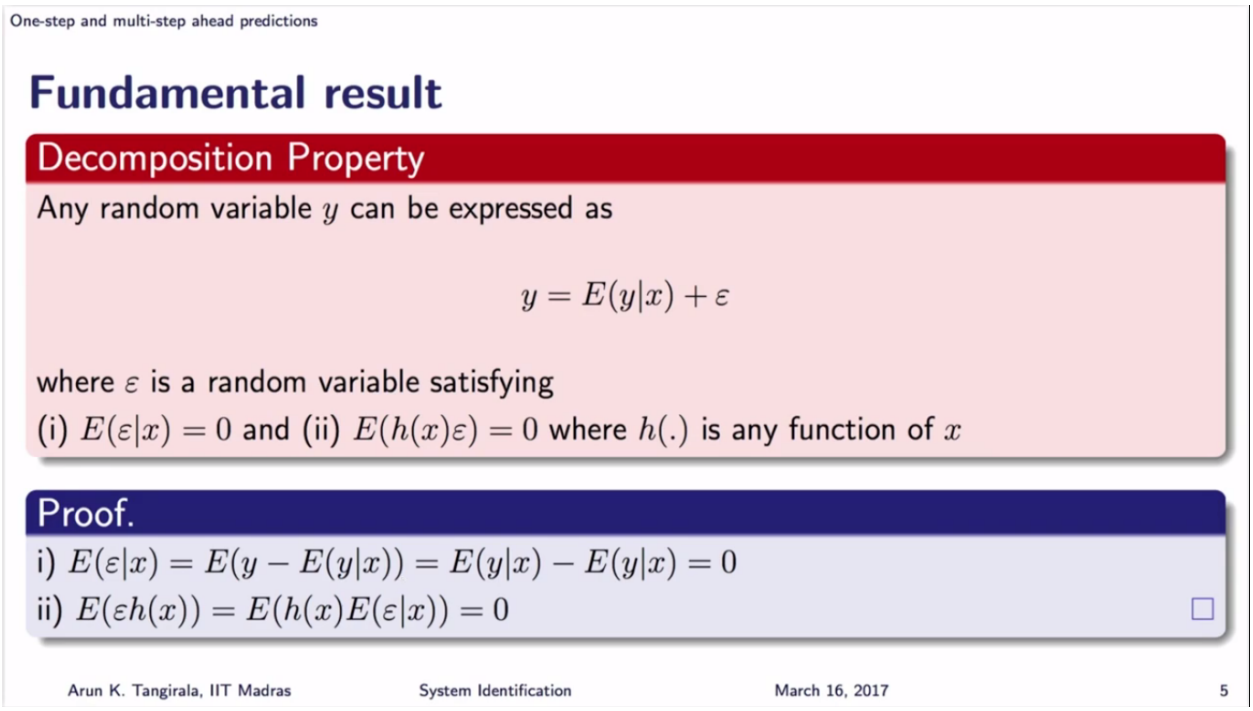


Why is this result important.Why do you think this result is, how do you connect this result to prediction? So I tell you this result is useful in prediction. Do you see the connection?

Student: Epsilon will [17:52 inaudible] how much we can predict.

Epsilon willtell me how much I can predict using X, I mean predict y using X. But the more important part is conditional expectation is the best prediction. Right.Because whatever is left out which is epsilon does not have anything to do with X. Right.So that means I have extracted whatever juice was there in y with respect to X. Can you prove this result? Yes, you can. So the proof is given here. Expectation of epsilon given X is simply expectation of Y minus expectation of conditional expectation, sorry, y minus conditional expectation and because x is deterministic. So you see here. In fact there's a given x missing here. Okay.So I'll make that connection as well.

(Refer Slide Time: 18:54)

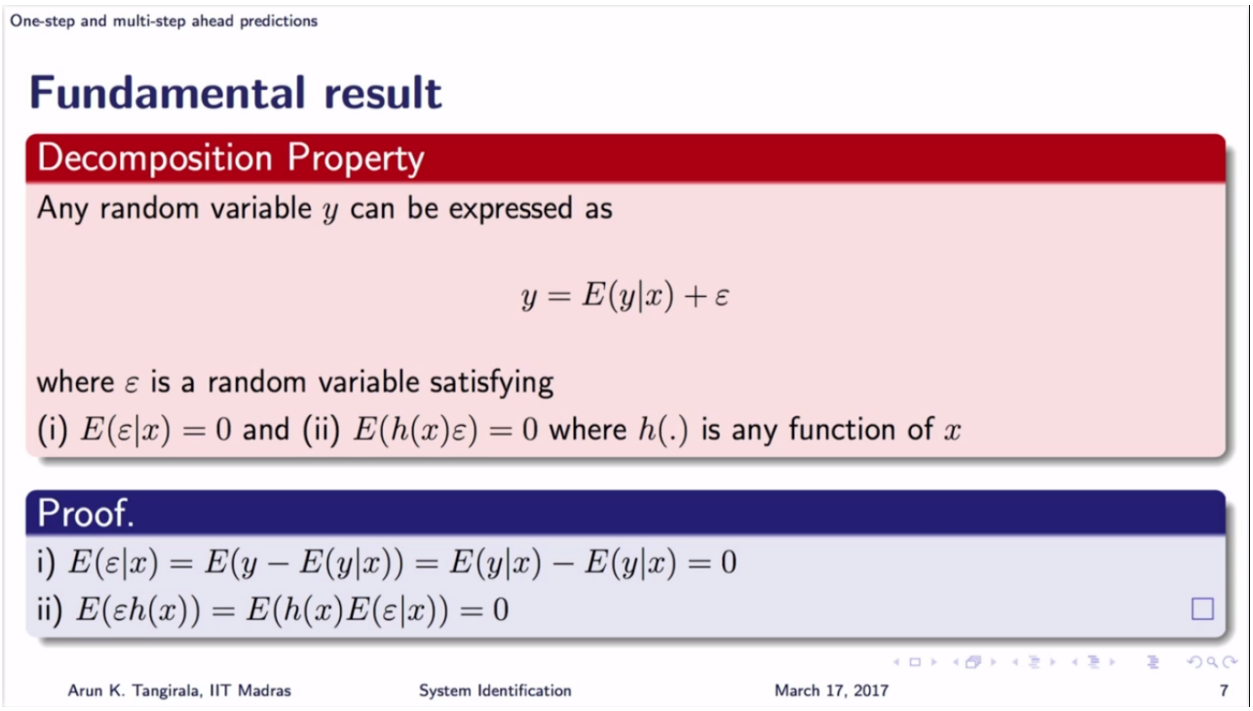


So once you write this here, the expectation of Y minus, I'm going to use uppercase, minus expectation of Y given X, this given X, this is what is its expectation of epsilon given X. Now there are two terms here. First term is conditional expectation. The second term is expectation of the conditional expectation. But since X is fixed anyway insidefixing one more time won't make any difference. So the second time is also conditional expectation and therefore you have 0.

The second one also can be proved using the iterative expectation formula. Iterative expectation says,expectation of Z, any variable Z, is what is iterative expectation result.?This is one, second one, you can use this result expectation of Z is expectation ofconditional expectation.

I understand that's a standard iterative expectation result. You can take any variable Z and say its expectation is a double expectation.In the inner expectation X is fixed and the outer one is averaging with respect to X. Now, what is Z for us here. What is Z for us here can you tell me?I just made one small correction here.

(Refer Slide Time: 20:41)

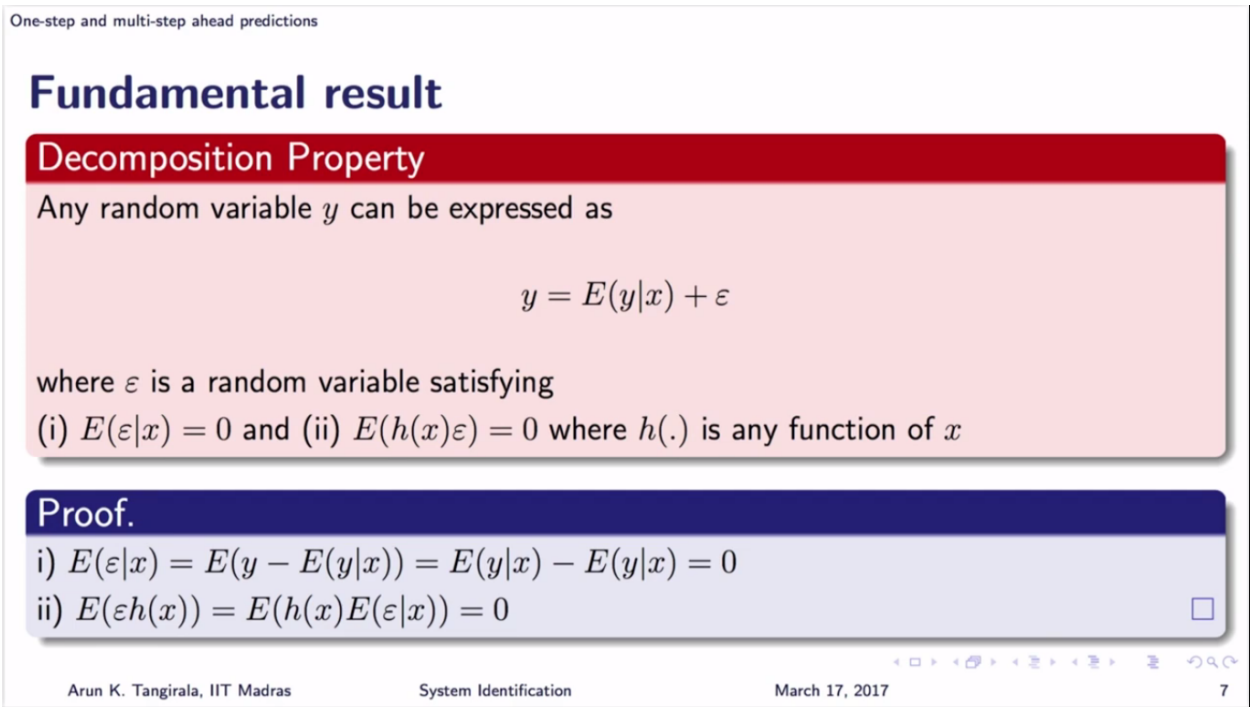


Yeah. Tell me what is Z?Why are we talking about this result? What is Z? We want to prove the second result, right. What does the second result say? Be bold. It's okay if you can. You're supposed to make mistakes here, if any. Epsilon?What is Z? Now how do you use this result to prove the second one.Correct. So, Z is epsilon times h of X. That's all. That's all. So we know all ready,so if you apply this result here, I have expectation of X times expectation of epsilon times h of X given X. Now since X is fixed in the inner one h of X can follow. It becomes a constant for that expectation. Right.

So I get expectation of X times h of X times,at conditional expectation of epsilon given it with respect to X.Now this we already know is 0. These are all standard, some other tricks that you use to prove some results elegantly. This iterative expectation result is very powerful. Any questions with respect to that. Okay. I assume you all understood. And you'll be able to prove it again.

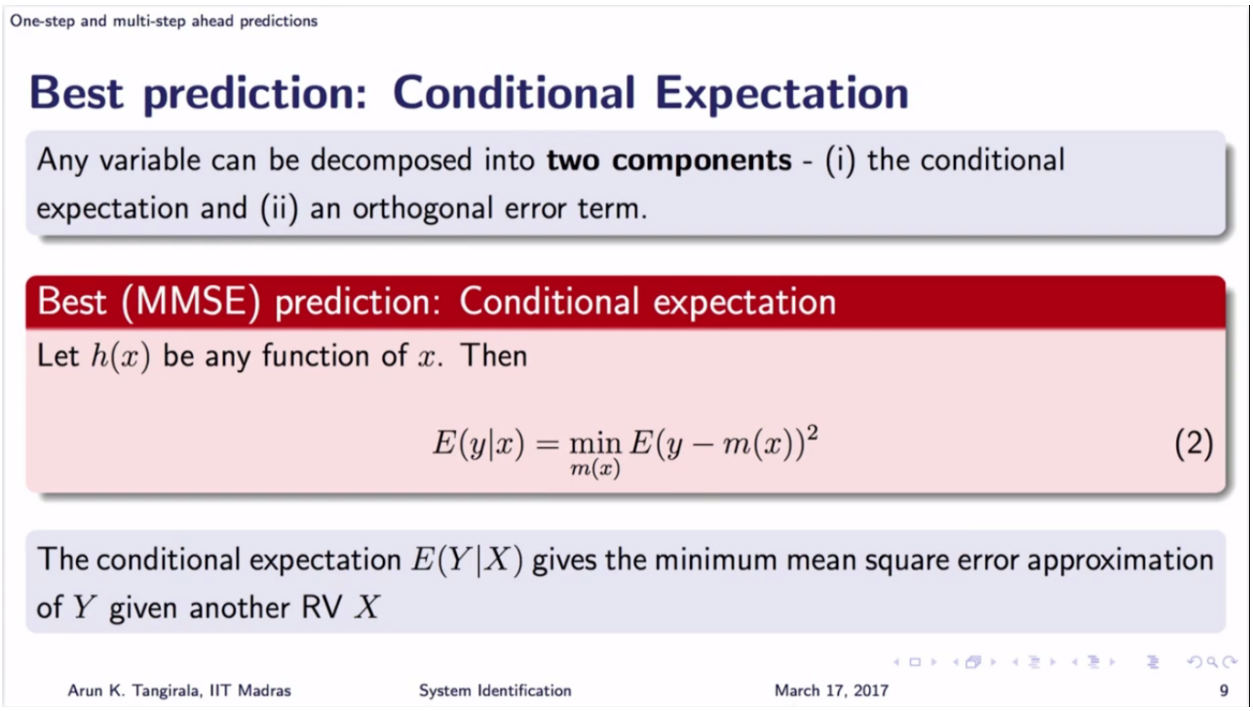
Anyways, so proofs apart, you have to understand now the message that this theorem is giving.

(Refer Slide Time: 22:33)



The message is that, given any random variable y, I can break it up into two parts. One part that is a function of x, which kind of function of x, such that it gives me the minimum mean square prediction error. It's not stated in the theorem but what is this function of x conditional expectation. We know already conditioned expectations that function of x. And the other part which does not have anything to do with x. That's great. Straight away I get my prediction result which is that the conditional expectation is the best prediction. Of course, the decomposition property doesn't tell me that this conditional expectation minimizes the means square error. That it doesn't tell me. It says that given any random variable y and another random variable x, I can break it up always into two parts. One that is a conditional expectation other that has nothing much to do with x, right.

(Refer Slide Time: 23:35)



Now thatresult can be used or you can say intuitively, you can now understand why this result holds good. As I said, you can actually prove this result without the use of the decomposition property. There are two ways of looking at it. All right. But the main point is given this decomposition property the conditional expectation stands to be the best predictor of y given x and best in the minimum mean square in a sense. This was the milestone result but we have already talked about it. This conditional expectation result is useful only if I know the conditional pdf. Otherwise how can I use it. And that is where we said, in practice this is going-- In reality this conditional expectation is mostly a non-linear function of x.

And if I know the condition pdf, I can calculate this. But if I do not know what do I do. So people started asking what if I construct linear predictors, can I work with linear predictors and then we stated a very important visit that a linear predictor is optimal only ify and x are jointly Gaussian distributed. For all other situations there is no guarantee. In general the linear prediction is going to give you suboptimal results. Okay.So the conditional expectation can we evaluated, one that either when I'm given a pdfthat is the conditional pdf or a joint pdf or something has to be told me about the nature of relation between y and x. There has to be an equation given and then I can apply this result, otherwise I cannot apply. But we will have a model. So we will use this result and I'll show you how to use it with this result.

(Refer Slide Time: 25:28)

Example 1: Prediction for MA(1) process

Consider an MA(1) process: $v[k] = e[k] + c_1 e[k-1]$

- The one-step ahead predictor is

$$\hat{v}[k+1|k] = E(v[k+1]|k) = c_1 e[k|k] \tag{3}$$

- The quantity $e[k|k]$ is not known, but has to be estimated using the observations
- How does one obtain $e[k|k]$?

$$e[k] = v[k] - c_1 e[k-1] = v[k] - c_1 v[k-1] + c_1^2 e[k-2] = v[k] + \sum_{n=1}^{\infty} (-c_1)^n v[k-n]$$

For the infinite sum on the RHS to converge, $|c_1| < 1$.

So the rest of the lecture is going to be now, how to use this condition expectation result in deriving predictions from random signals, given LTA descriptions of those random signals.