Pattern Recognition

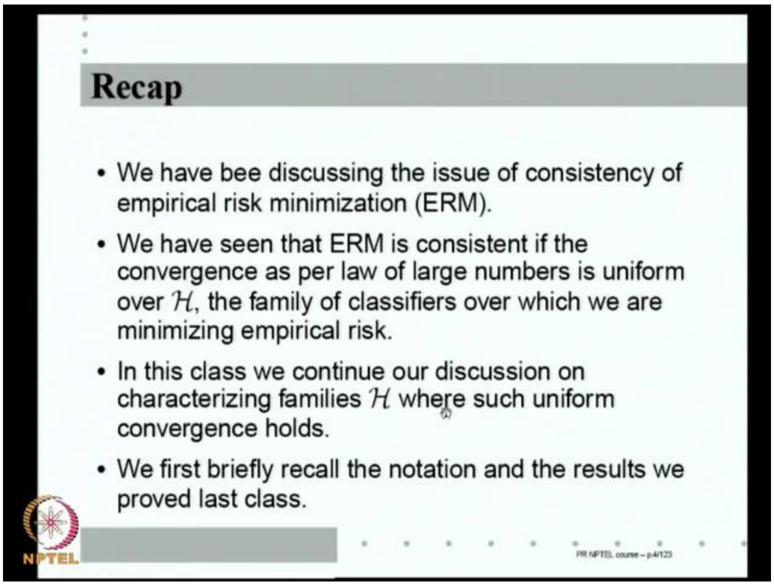
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 23

Consistency of Empirical Risk Minimization; VC-Dimension

Hello and welcome to the next lecture in this Pattern Recognition course, we have been looking at certain issues of statistical learning theory, some simple introduction to basic issues of statistical learning theory. Specifically, we have been looking at the issue of consistency of empirical risk minimization, this most algorithms which are essentially based on minimizing empirical risk. If they find global minimize of empirical risk will that be a good enough classifier to learn right, will the global minimize of empirical risk be same as the global minimizer of true risk. That is the consistency of empirical risk minimization that we have been considering.

(Refer Slide Time: 00:59)

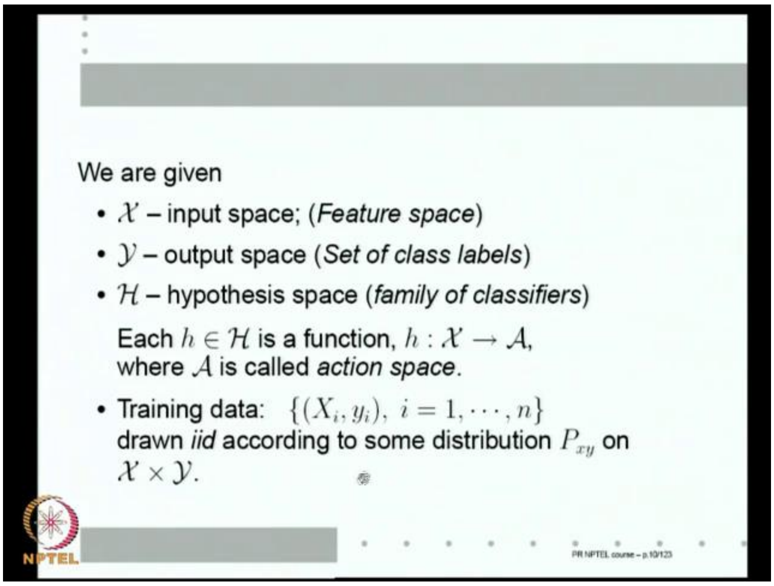


So, specifically we have seen that empirical risk minimization is consistent, if the law of large numbers convergences uniform over H. We call that H is the family of classifiers over which you are minimizing the empirical risk, given any one function h little h, we know the empirical risk r hat n h is nothing but, a sample mean estimate obtained on the sample of the training examples of the true risk.

Hence, for any given H law of large numbers intuits, law of large number guarantees that r hat n h converges to r h is as n tends to infinity. So, the issue is that with convergence uniform over the family h, we seen in the last class that here empirical risk minimization is consistent if this convergence is uniform. So, essentially then what we want is to characterize the families, H characterize the family of functions what should the family of classifiers h should satisfy. So, that such a uniform convergence holds right that is the discussion that we are going to continue in this class.

Recall that we are doing all this only for 2 class classification problems, the others what we are doing here maths become much more complicated, if we consider real valued functions. So, simply taking H to be a family of binary valued functions defined over our feature space x, and for that class we are asked the question what should be the family of classifier satisfy, so that the needed uniform convergence holds.

(Refer Slide Time: 02:56)

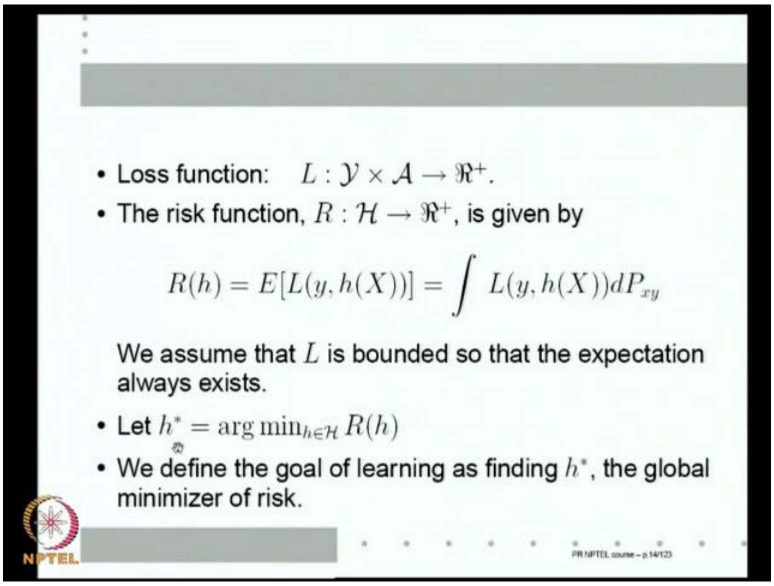


So, since this is mathematically involved let us once again review our notation and if you think that we proved last class. So, let us start how are we specifying our learning problem, we are given script X which is our input space, for the pattern recognition problem is the feature space all the feature vectors belong to X. So, normally it is a D dimensional Euclidean space for us, Y is the output space that is the set of class labels for us and in this particular case it is binary. We can take 0 1 plus 1 minus 1, it really does not matter we are taking 0 1.

Then the space of functions of classifiers over which we search is being given the symbol script H, we call it the hypothesis space. In general the hypothesis space consists of functions that map X to another set called A the action space this is, so that you know by the learning directly binary valued classifier functions or discrete functions. All of them can be viewed in the same framework that is the reason why we had this H. But, for the purpose of this class where we are looking at when the family H is such that the convergence under law of large numbers is uniform.

We are only considering binary valued functions, so actually y 0 1 and our A is equal to Y. Then the training data we are given is to pull X i, y i there are n training examples. They are drawn iid according to some distribution on X cross Y, as we said already in this framework there is no target concept, so examples come like this. So, that any noise in examples for example, the same X can have different Y with different probabilities. For example, when the class conditional densities overlap, all such things are taken care of by taking this distribution to be some unknown distribution on X cross Y.

(Refer Slide Time: 04:50)



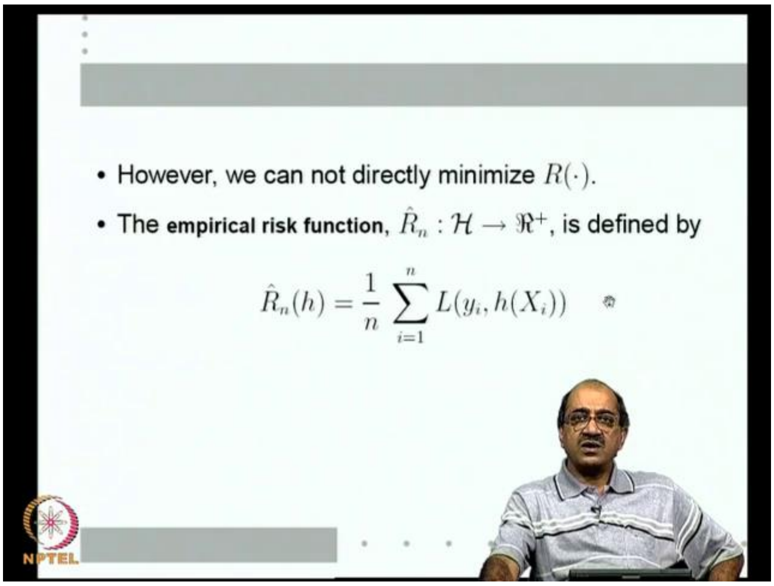
We are also given what is called a loss function, a loss function maps Y cross A to R plus that is (()) real line, the idea is L of y comma h of X as tells me the loss I suffer with the function h on a random sample X y. h of X is what the function will say on X and y is a in a statistical sense the true thing to say, so L of y comma h of X is my loss. I still kept the more general definition of loss function, as I said for the purpose of this class discussion this A is Y.

Then we define the risk function, which assigns a number to every classifier given a classifier h, the risk of h is the expectation of the loss. That is you take average of L y h X over all X and y, where the average is with respect to the same distribution with this examples are drawn. So, that is our entire issue of having representative samples, our samples are drawn with respect to some distribution P X y and I know how the loss is actually calculated.

And ultimately I am evaluating each classifier by taking expectation with respect to the same distribution with which examples are drawn. So, risk of h risk is expectation of loss, where expectation with respect to d P X y, so given this our objective is the global minimizer of risk. So, we have given it the symbol h star, h star is the value of little h that will globally minimize R h. So, ultimately our goal of learning is finding the function h star which is the global minimizer of risk.

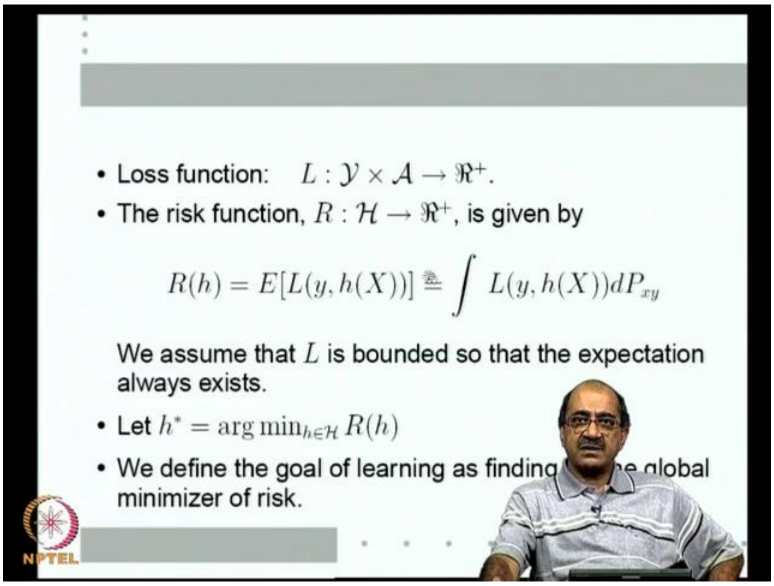
Of course, as I said the h star may not be unit there might be more than one function that achieves global minimum of risk, but that does not make any difference to us because we are going to distinguish different functions in h only in terms of their risk values. So, if two different functions have the same risk value as far as you are concerned they are the same function. So, the fact that the problem minimizer may not be unique is of no consequence to us, but the problem is that we cannot directly minimize R. Why cannot directly minimize R, because given A h. We cannot even calculate R of h, because that depends on the unknown probability distribution, so since risk for a given h cannot even be calculated.

(Refer Slide Time: 07:29)



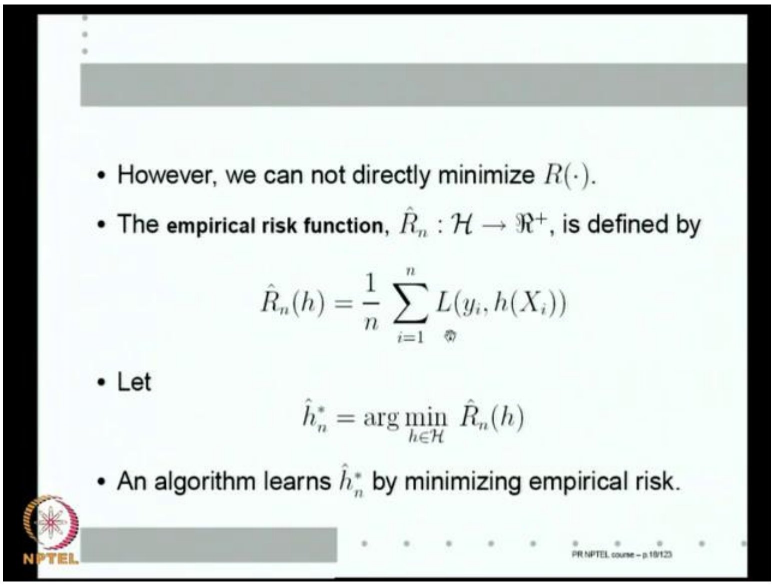
We cannot directly minimize R, so what we decided is we have a empirical risk function, which also assigns a number to every classifier like this. This is the average of L y i h X i average taken over all the training samples.

(Refer Slide Time: 07:48)



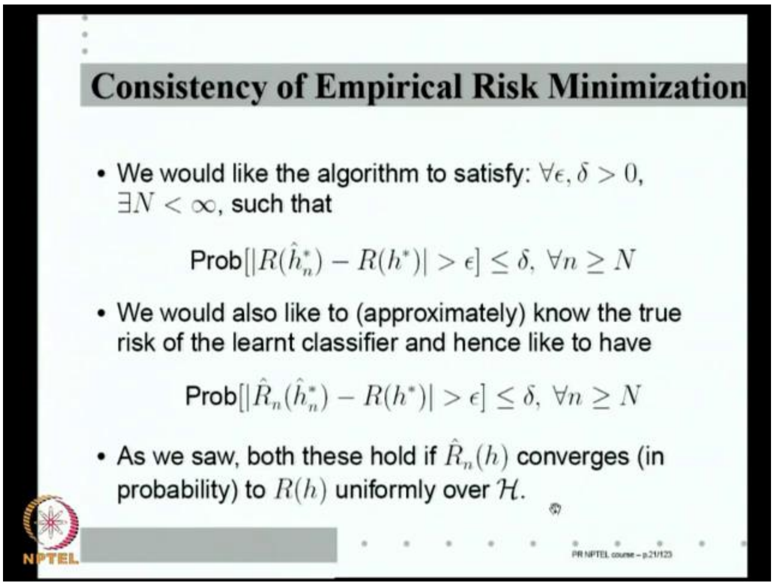
So, essentially risk is the expectation of L y h X i, so L y h X i can be thought of as a random variable, it is a function of the random variable X comma y random vector X comma y. So, if I want its expectation, if I do not know the distribution I can always approximate the sample mean. So, I get samples X i y i calculate the value of this function and take the average and that is what the empirical risk is.

(Refer Slide Time: 08:13)



So, empirical risk is nothing, but the sample mean estimator of the expectation of loss obtained through n iid samples. Since, X i y i are the given samples and we know L given any h I can calculate R hat n and hence in principal I can find A h where this is minimized. So, let us call that h hat star n, as I already explained to you while the notation may look cumbersome, basically the hat denotes that it is an estimate of h star the actual global minimizer and n denotes that the estimate is obtained through a sample of n iid examples. So, an algorithm all learning algorithms are essentially minimize empirical risk, using some optimization technique and the algorithm learns h hat star n by minimizing the empirical risk.

(Refer Slide Time: 09:19)

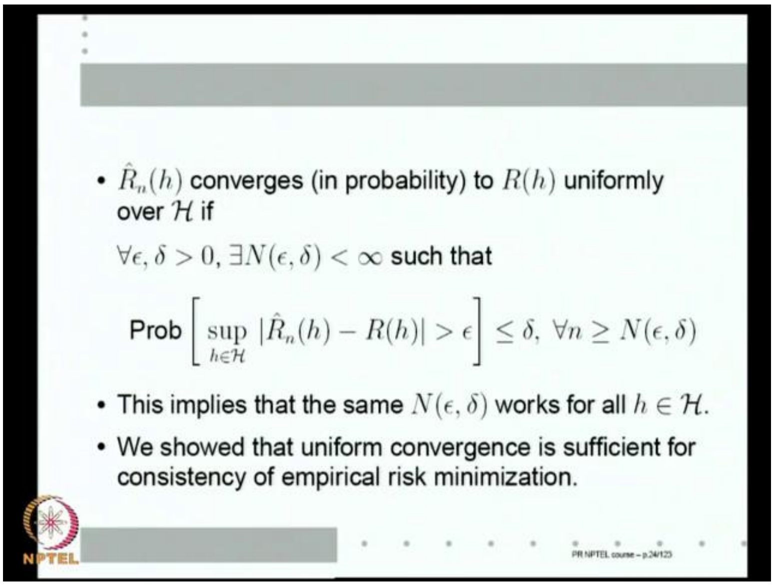


So, this is the setup that we have and what is that we want for consistency, we like the algorithm to satisfy the following, given any epsilon delta there should be some n that is the found in number of examples. Such that, probability R of h hat star n minus R h star the absolute difference between them being greater than epsilon is less than delta if n is greater than capital N. R h hat star n is the true risk of what I learnt h hat star n is the minimizer of the empirical risk, that is what I have learnt r of h hat star n is the true risk of what I have learnt R h star is the true risk of the optimal classifier.

Really, as I said h star may be not unique, but R h star is unique in the sense is that global minimizer. So, R h star stands for the global minimum of risk, so I want the risk of what I learnt should be closed to the global minimum of risk. This will tell me that h hat star n is good enough, thus we h hat star n being close to h star is simply that their risk values are close, this is what I want.

As you seen we will we also would like to have r hat n of h hat star n close to r of h star, not only R of h hat star n, but even R hat n of h hat star n is being close to R h star. Because, this way after I finish my learning I know h hat star n and I can calculate r hat n of h hat star n. So, if this is also there, then who will know the true risk or at least approximately know the true risk of what we have learnt. This is the issue of consistency of empirical risk minimization and as we saw last class both these will hold if R hat n h converges in probability to R of h uniformly over h.

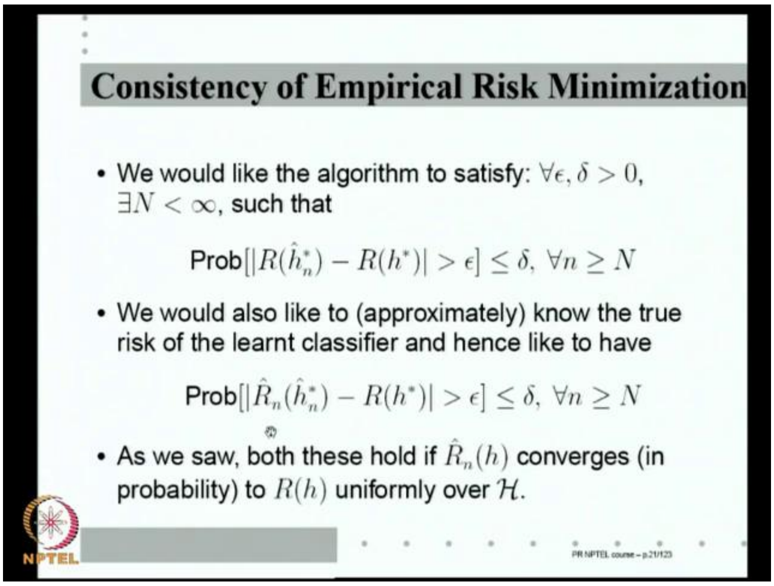
(Refer Slide Time: 11:03)



What do you mean, R hat n h is the sample mean estimator of R h, R h is the expectation of loss, R hat n h is the sample mean of loss over the iid samples. So, anyway we know R hat n h converges to R h where each individual h as n tends to infinity, but what we want is that this convergence should be uniform. That is given an epsilon delta, there exist one capital N which can depend on epsilon n delta, such that for all h say this supremum of the difference in R hat n h minus R h over all h greater than epsilon, is less than delta if the number of examples is greater than capital N.

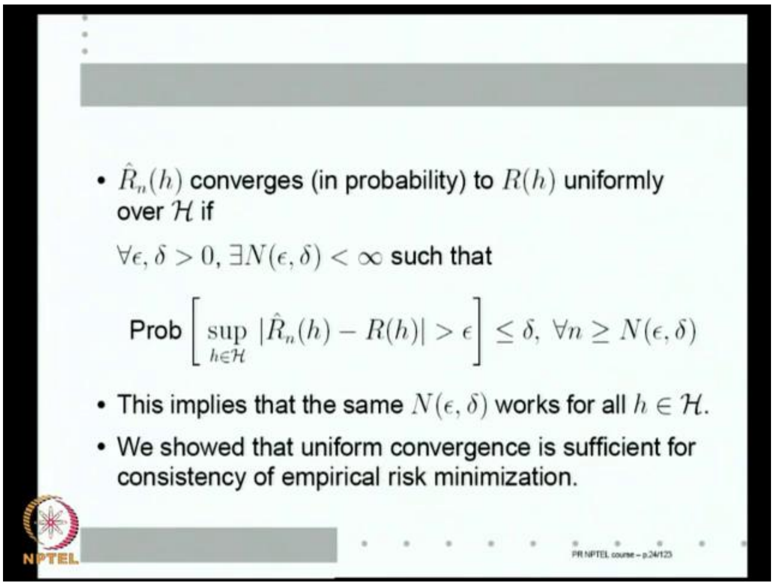
What it means is given an epsilon delta I can calculate one number capital N epsilon delta which works for all h. Works for all h meaning, no matter which h for which I am i am estimating the risk, if I estimate with this number of n samples I will get to n epsilon risk. So, that is what is meant by uniform convergence, we discussed this last class. Now, we showed last class that uniform convergence of R hat n h to R h is sufficient for consistency of empirical risk minimization, we showed that if R hat n h converges uniformly to R h. Uniformly, over the class of classifiers h and which empirical risk is minimized, then both these things that we wanted earlier.

(Refer Slide Time: 12:34)



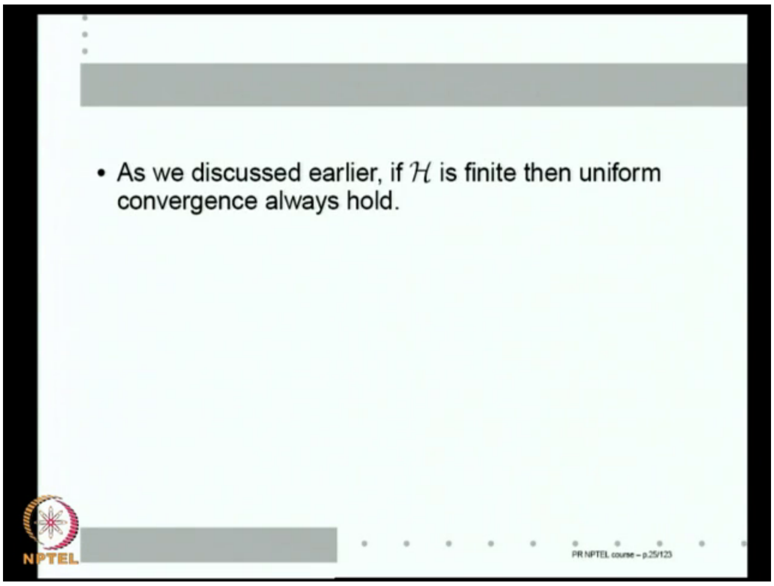
Both these are satisfied.

(Refer Slide Time: 12:38)



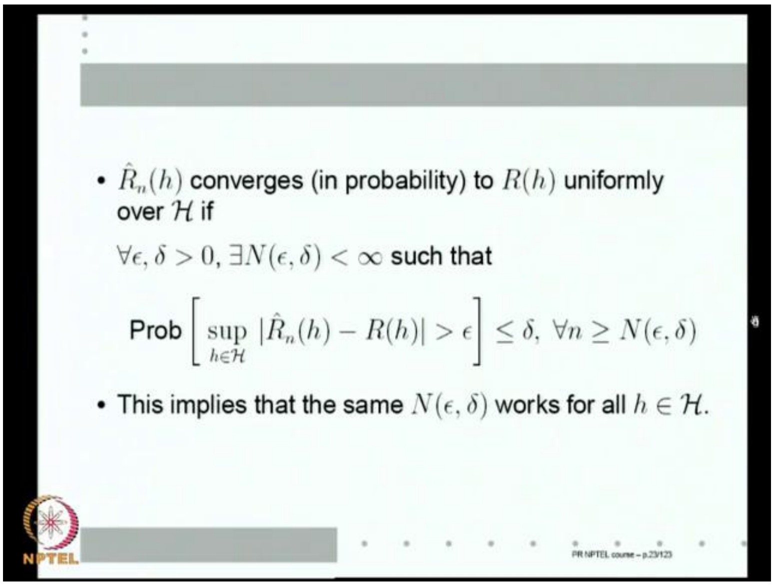
So, in that sense the uniform convergence is sufficient for consistency of empirical risk minimization, we also mention that uniform risk, the uniform convergence also necessary for considering empirical risk minimization. We did not prove that, but we mentioned we told that it is also necessary, hence empirical risk minimization is consistent. If and only if the class of classifiers h that we that we choose to minimize consistency of empirical risk over is such that the R hat n h converges to R h uniformly over this class. So, now our question is how to decide given A h whether or not this uniform convergence holds?.

(Refer Slide Time: 13:18)



Now, for that we first looked at finite, so basically if the class of classifier h is finite then uniform convergence always holds.

(Refer Slide Time: 13:32)



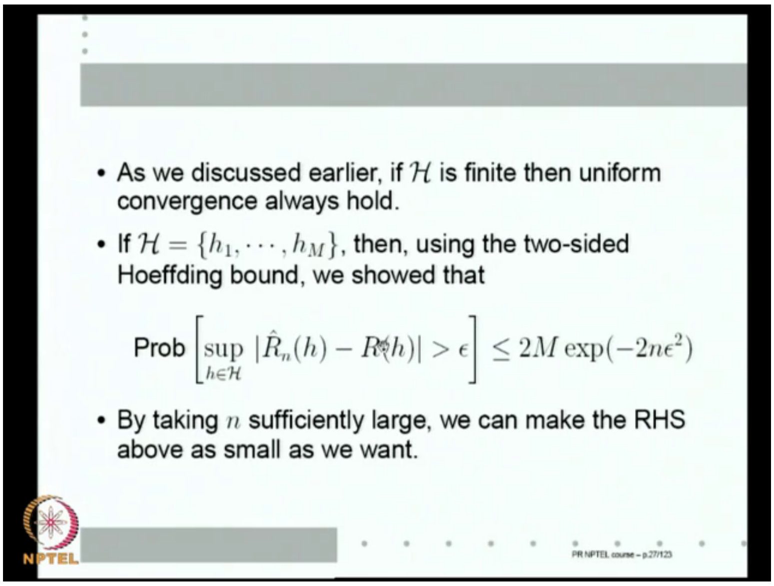
So, the idea is that for uniform convergence the supremum over this of R hat n h minus R is greater than epsilon should be less than delta, for the for if I take one capital N that capital N should work for all h. For every h because the law of large numbers, they would be a capital N may be dependent on h, but if they are finitely many h's I can take the maximum of all those n's, so this will always work.

That is the whole idea of y, if h is finite then uniform convergence always holds. In specifically if h is h 1 h 2 h m then using the, so called two sided hoeffding bound, we showed last class that probability supremum over h belonging to script h. Absolute value of R hat n h minus R h greater than epsilon, this probability is bounded above 2 m exponential minus 2 n epsilon square. Matter of fact hoeffding bound tells us that for each h R hat n h which is the sample mean estimator and R h is the expectation.

So, the sample mean minus expectation being greater than epsilon that can be bounded above by exponential minus 2 and epsilon square, that is what the hoeffding bound is. If there are m functions then as you see using a union bound, it simply adds a factor of m on the right hand side. Now, this shows that this probability can will go to 0 as n tends to infinity, because as n tends to infinity this factor goes to 0, which means that if I take n sufficiently large no matter how large m may be if small n is sufficiently large.

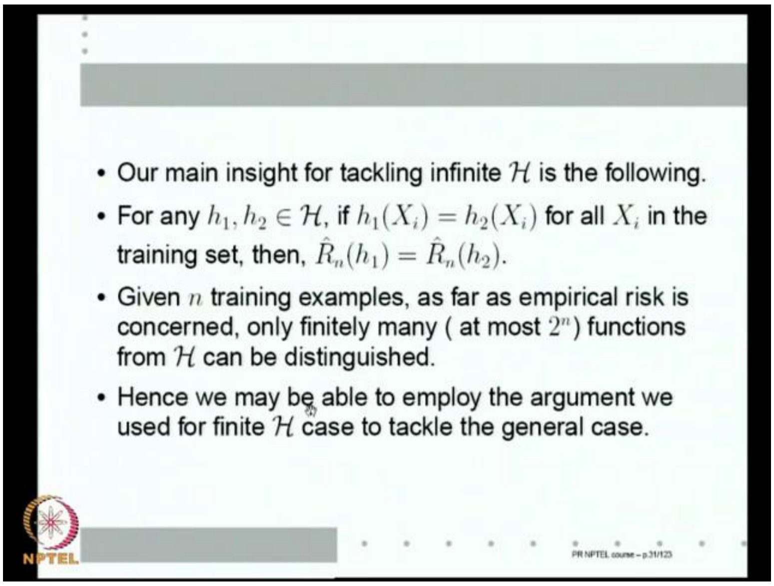
Then this exponentially decaying term will ultimately make the right hand side less than any delta we want. And hence, if we take n sufficiently large we can make right righthand side as small as you want. And hence the needed uniform convergence holds, we can make probability supremum h belonging to h R hat n h minus R is greater than epsilon to be less than delta. For any given delta for a particular n which is dependent epsilon delta because essentially I want this to be less than delta.

(Refer Slide Time: 13:59)



So, I have to calculate n, so that n can only depend on delta epsilon, m is a constant anyway. So, if h is finite then uniform convergence always holds.

(Refer Slide Time: 15:53)

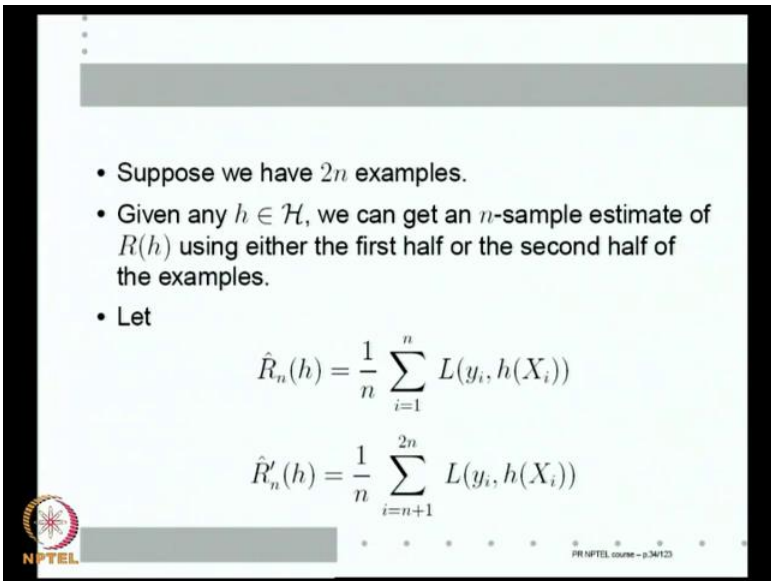


Now, how do we tackle the infinite h case, so as we very briefly discussed towards the end of last class our main insight for tackling the infinite h case is the following. If I take any 2 functions h 1 h 2, in my bag of classifiers, they are functions of such that h X i is equal to h 2 X i for all X i in the training set. They take the same values on the training set then R hat n h 1 is equal to R hat n h 2. So, essentially what it means is that just using empirical risk over the n sample, I cannot distinguish among all possible functions in X.

There will be many functions, which may take the same values on the specific training examples, in which case R hat n h 1 will be equal to R hat n h 2. So, based on empirical risk I cannot distinguish R from this, how many can I distinguish I have got n samples and h s are all binary valued functions. So, any given h if I look at all possible values it can take on the n samples, it is one n-bit binary number, every h if I look at all possible values it can take on the entire set of training data, so if it is the total number of n-bit binary numbers. So, they can at most be 2 power n different functions, that can be distinguished because given n samples. There are only 2 power n different possible binary to pulls, that the n example that the function n examples can take, means given n training examples as well as empirical risk is concerned we can distinguish at most 2 power n functions from h.

So, which means we can use the same type of arguments as in the finite h case to tackle the general case, because we are ultimately looking at only empirical risk and empirical risk can distinguish only among finitely many functions. So, even though the supremum over all the infinitely many functions in h, because R hat n h can take only finite can distinguish only among finitely many functions we should be able to employ the argument that we use for finite h case.

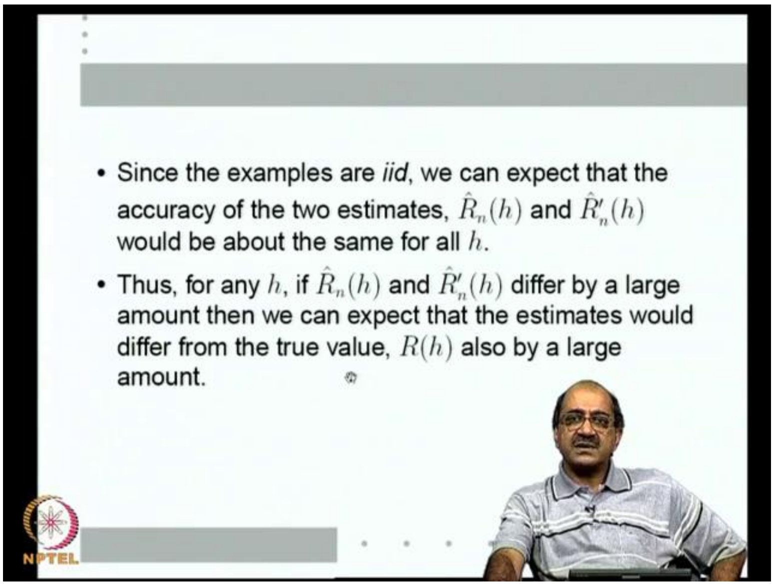
(Refer Slide Time: 18:17)



So, this is what we are going to do, so let us particularize this argument, so we will use a very clever commercial argument which was first used by Vapnik and Chervonenkis which is called the symmetrisation argument. So, the basic idea is the following, suppose we have two n examples, instead of n examples given any h belong to h suppose we want an n sample estimate of R h we want a R hat n h. So, we want to estimate the expectation of loss using a sample of n iid examples, but we have 2 n examples.

Hence, I can get 2 estimates I can get one estimate using the first half and one estimate using the second half of the examples. So, let us give some names to the 2 estimates, let us call the estimate obtained through the first half, that is i is equal to 1 to n, as R hat n h and for i is equal to n plus 1 to 2 n as R hat n prime h. So, we have 2 estimates both are n sample estimates, one is R hat n one is R hat prime n, this is obtained over the first n this is obtained over the second n. Now, the basic idea of the symmetrization argument is the following.

(Refer Slide Time: 19:36)

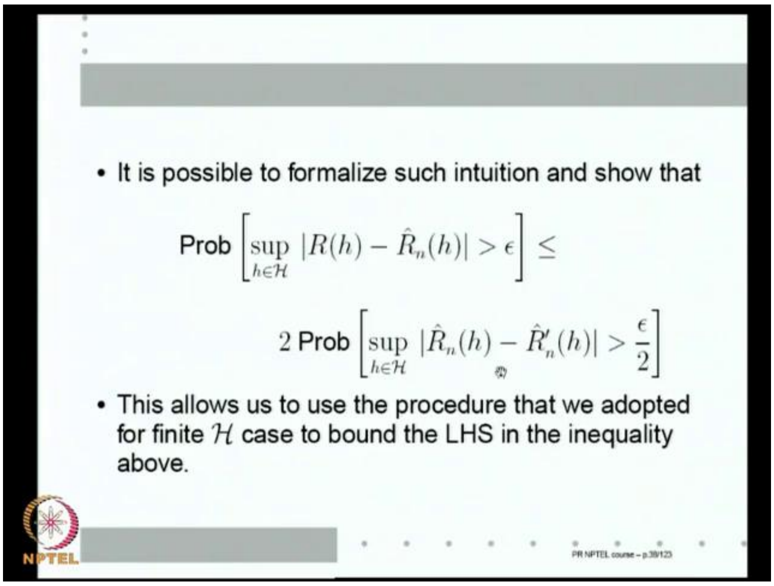


Since the examples are iid we expect the accuracy of the 2 estimate to be about the same, because both are obtained through an iid examples. So, it is like saying I tossed the coin 10 times and finds the estimate of heads using 10 sample, let us toss it again 10 times and find another estimate. We do not really expect the accuracy the estimates to be vastly different, because both of them are obtained based on 10 toss. So, when the examples are iid we can expect the accuracy of the 2 estimates R hat n h and R hat prime n h would be about the same.

Now, what does this mean, so the probability of they differing from the two values about the same, so either both of them are bad or both of them are good so to say. Which means if both of them are good both of them are closed to R h and hence they are closed to each other, conversely if they are very far from each other then they we also expect that the estimates are far from the true values.

Because, they are not really good estimates, the estimate is good then R hat n and R hat n prime n should be about the same, if both of them are close to R h they should also be close to each other. Conversely if there far away from each other, then any one of them is likely to be far away from R h. This is say of course, very hand waving intuitive argument, but I hope the basic idea is clear, because examples are iid whether a estimate an n sample estimate using one sample of n or another sample of n probabilistically i should get about the same accuracy.

(Refer Slide Time: 21:31)



Now, this argument can be made very precise and rigorous combinatorially and thus we can prove this. We can show that if I want probability supreme over all h, R h minus R hat n h greater than epsilon, that can be bounded above by twice the probability of R hat n h minus R hat prime n h greater than epsilon by 2. Essentially, the idea is that if R hat n and R hat prime n both epsilon close to R h, then you know they cannot differ by more than 2 epsilon.

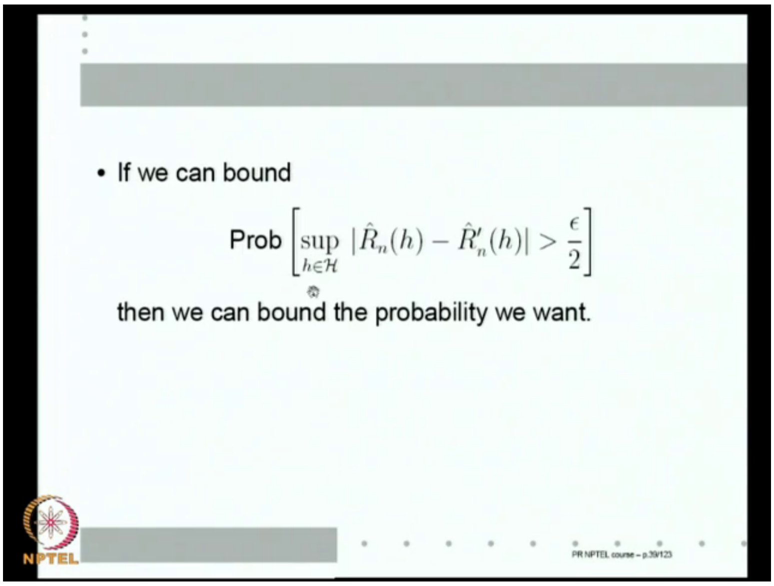
Similarly, if they are if they are further away than epsilon by 2 then it is possible for R h minus R hat n h 2 be greater than epsilon. While, we are not proving it because the proof is a little involved, I hope the basic idea of this inequality is fine, this is called the symmetrization argument. So, this is the probability that we want to bound ultimately supremum R h belong to h R h minus R hat prime n h, we want to show that as n grows large this can be made less than any delta that we want.

Now, we can bound this probability by this probability, what is the point the point is here even though R hat n h distinguishes between only finitely many h's, R hat R h can distinguish between all h. So, this supremum directly I cannot argue this has to be taken only over finitely many h, but here I have only 2 n samples totally R hat n h and R hat prime n h both are functions of only 2 n samples. So, given 2 n samples they are only finitely many h s that can be distinguished, for the rest of them R hat n h is equal to R hat prime n h.

So, this difference will anyway be 0, so to find the supremum I need to find it over only finitely many h's. So, using this symmetrization argument this supremum which I do not know how to bound is been converted into probability over another supremum which supremum can be taken over only finitely many samples. Because, R hat n h and R hat prime n h both of them are calculated from 2 n samples. So, if I take any h 1 h 2 which take the same values on the 2 n samples then both R hat n h and R hat prime n h will be same, so to say.

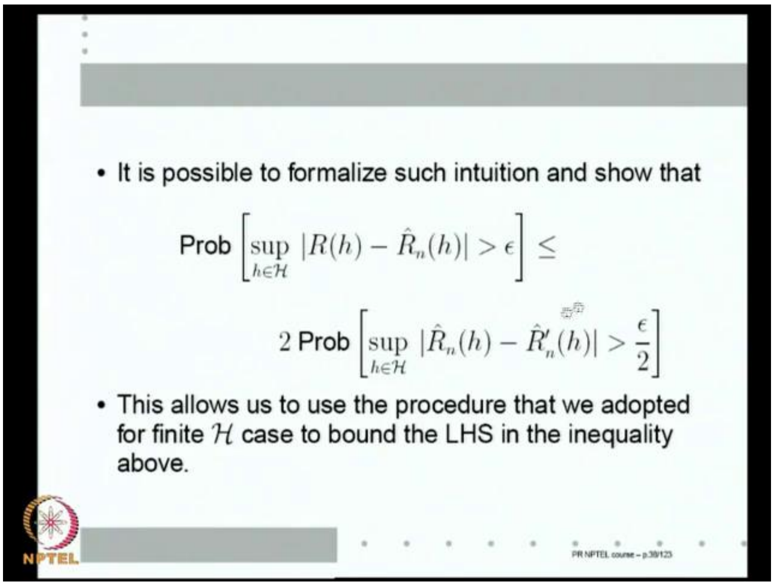
So, essentially one can argue that they cannot be more than finitely many h's over which the supremum needs to be taken. So, that is the basic idea this allows us to use the procedure to be adopted for finite h case to bound the L H S in the inequality, because this can be bounded using only finitely h cases. Once I bound this I have bound this, this is what I am interested and I have bounded that above by this. Now, this can be bounded using only the supremum over finitely many h. So, let us go on and see how I can do that.

(Refer Slide Time: 24:49)



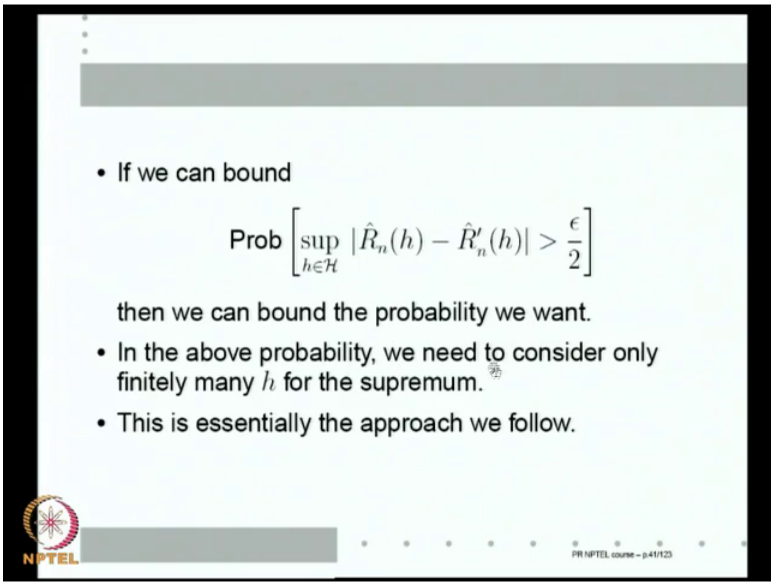
So, if I can bound supremum h belong to h R hat n h minus R hat prime n h greater than epsilon by 2.

(Refer Slide Time: 24:55)



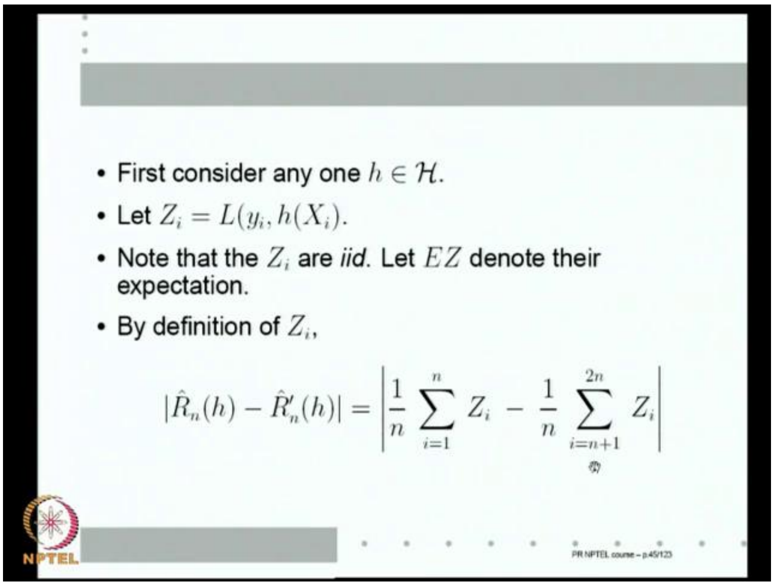
Then I can bound the probability, I want this is the probability I want to bound, now because of this if I can bound this probability I can bound this probability.

(Refer Slide Time: 25:03)



So, this is what we want to bound and as I explain in the above we need to consider only finitely many h for supremum because this depends only on the empirical risk. So, that is the that is the approach we want to follow, let us ask how to you know make it rigorous. How do I mean taking supremum over finitely many which finitely many how do I know exactly how many functions I have to take the supremum over all that.

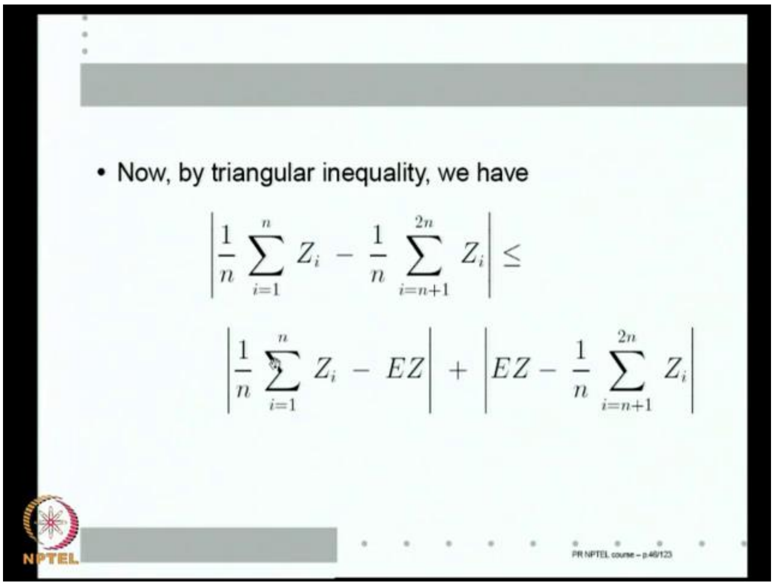
(Refer Slide Time: 25:37)



So, to go there first let us consider any one h belonging to h, let us give a name Z i 2 L y i h X i. Z i is some random variable, essentially the function of X i y i. So, I am just denoting L y h X i by the symbol Z i. So, where X i y i is a example, note that Z i are iid, because examples are iid, let expectation Z be the expectation that will be the risk of h.

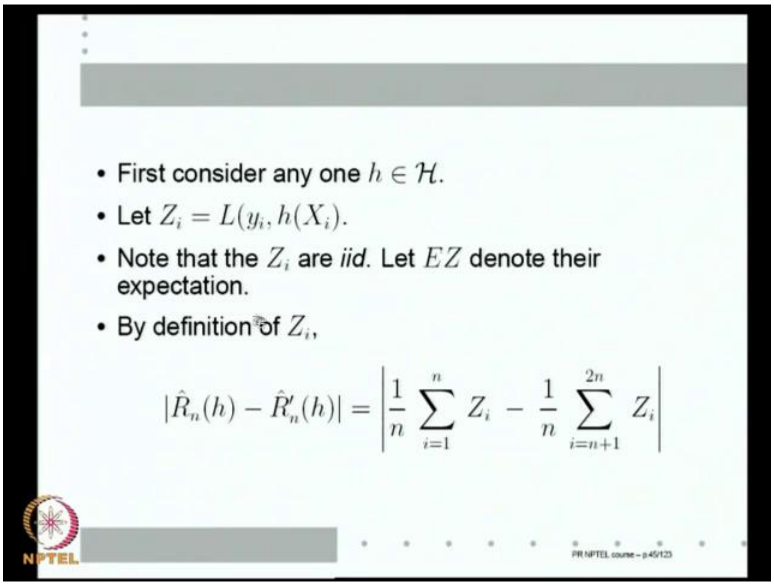
So, using this new Z i R hat n h minus R hat n prime h is nothing, but 1 by n, i is equal to 1 to n, Z i minus 1 by n, i is equal to n plus 1 to 2 n Z i. So, essentially we have some random variable Z, which is L y comma h X we have iid realisations of Z, Z 1 Z 2 Z 2 n I am finding the mean of Z. First using the first half of the samples, next using the second half of the samples. I want to know how to bound the probability, that this difference is less than some epsilon.

(Refer Slide Time: 26:53)



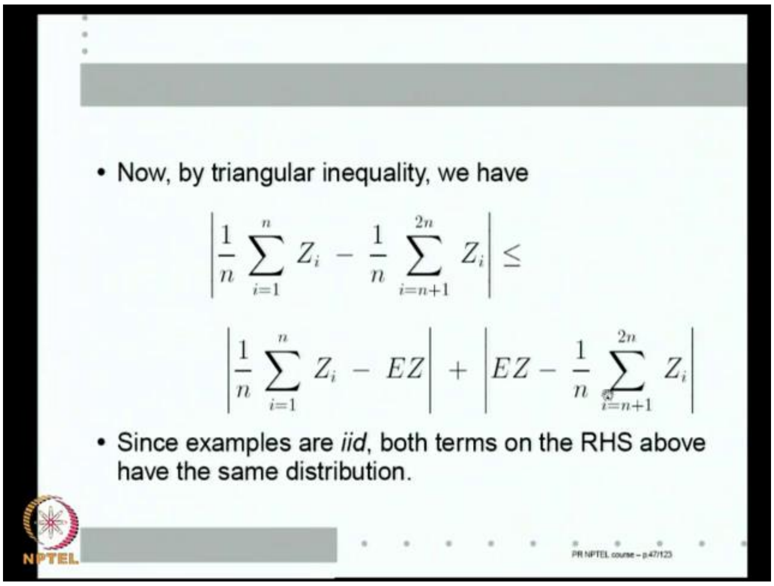
Now, how do I do this, I can use simple triangle inequality.

(Refer Slide Time: 26:59)



Remember that we are considering a single h here, so we are not yet getting into the supremum.

(Refer Slide Time: 27:08)



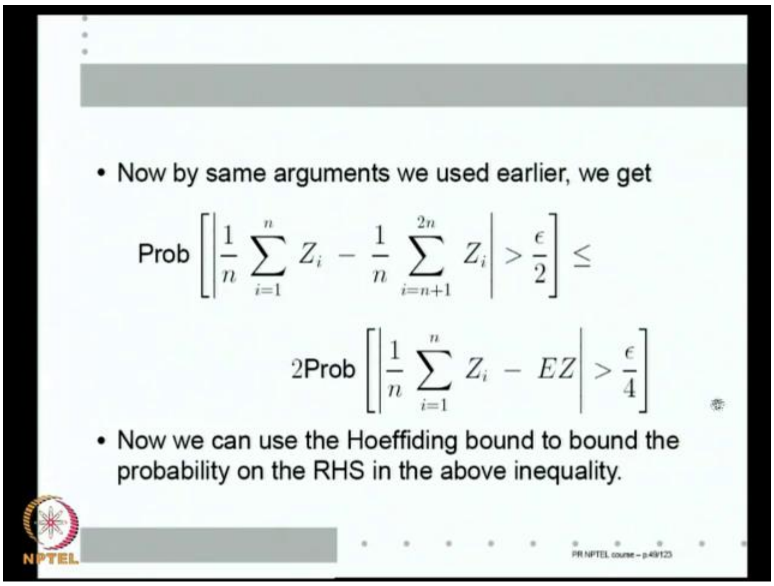
So, there is one random variable Z i here or rather iid random variable Z i all of them are iid here, now so this is what I want to bound. So, I can absolute value of a minus b can be written as a minus c plus c minus b, which is then bounded above by absolute value of a minus c plus absolute value c minus d. So, I have just added and subtracted expect value of Z, why this is a nice idea, because Z i or iid realization of Z and this is the sample mean and this the true expectation.

I know how to bound this using law of large numbers for example, hoeffding inequality, this also same and both of them are essentially have the same distribution, because this is a sample mean estimate of the unknown mean using n samples. This sample mean estimate of the unknown mean using n samples, so the probability distribution of this and this is the n fold probability distribution of Z i.

Now, I know how to bound each of these terms, we already seen that using hoeffding inequality, because this is less than equal to this plus this. We use this argument already more than once earlier, so for this the probability that both them are greater than epsilon by 4 is bounded above by this greater than epsilon by 4 and this greater than epsilon by 4. Actually, we can make it better than that because these 2 are essentially the same random variables, but let us just choose the argument that we used earlier.

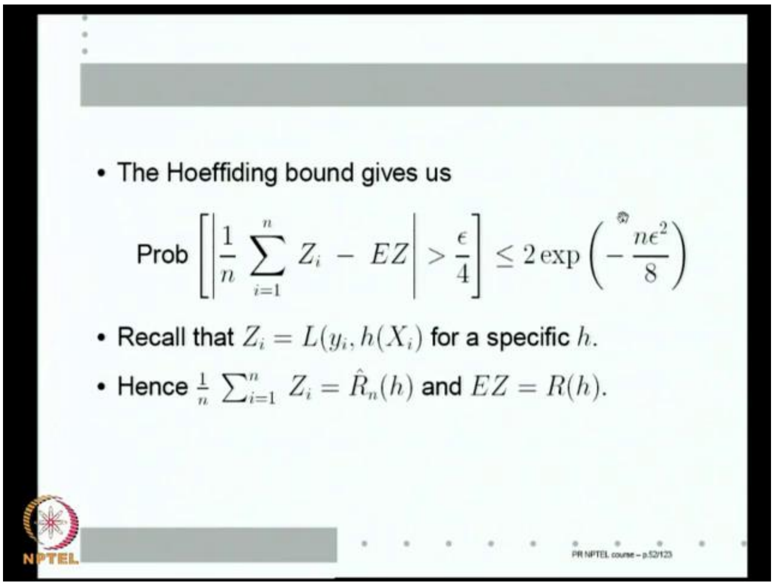
So, twice the probability that greater than epsilon by 4 is a bound by the probability that this is greater than epsilon by 2. The actual integrity of this argument we seen in the last class, so I as I told you we will use this argument again and again. As I said actually this two is not needed really, it is not needed because we can use what are known as chernoff bounds, because both of them are exactly identical things this is mean minus n sample estimate of sample mean.

(Refer Slide Time: 29:23)



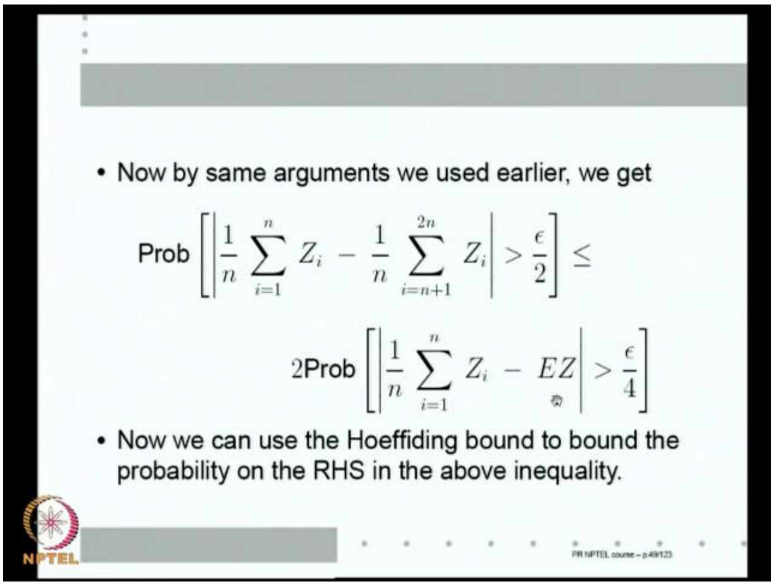
So, but because I have already shown you one argument of how to do this, we will stick to that argument I will just put it two because these factors do not make any difference towards. So, we know that this greater than epsilon by 2 can be bound above by twice probability, this greater than epsilon by 4. And this we know how to bound Z i refers to only one h and using law of large numbers and the hoeffding bound we can bound this.

(Refer Slide Time: 29:45)



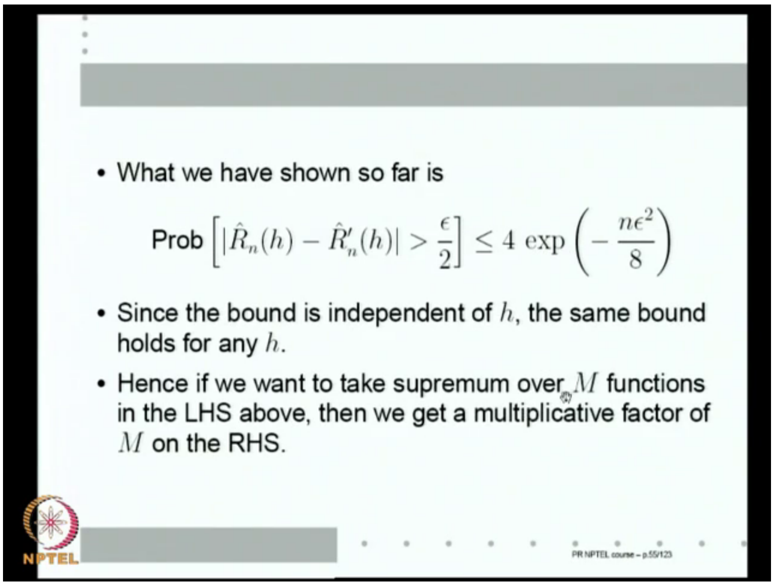
What is that bound hoeffding bound tells us that the sample mean minus expectation greater than something is less than 2 exponential minus n times that something square. So, I get epsilon squared by 16 and the 2 in the numerator cancels once, so I get 2 exponential minus n epsilon square by 8, this is same hoeffding bound we used earlier. So, note that Z i is L of y h X i for a specific h, hence one by n i is equal to 1 to n Z i is the R hat n for that h expectation Z is R h. So, what do I shown is R hat n h minus R h greater than epsilon by 4 is less than or equal to 2 exponential minus n epsilon square by 8, this we know from hoeffding bound.

(Refer Slide Time: 30:35)



So, going back what is this for originally this is what we want this is R hat n h, this is R hat n prime h greater than epsilon by 2 I want. That is twice the probability of R hat n h minus R is greater than epsilon by 4, this we have bound using hoeffding bound like this.

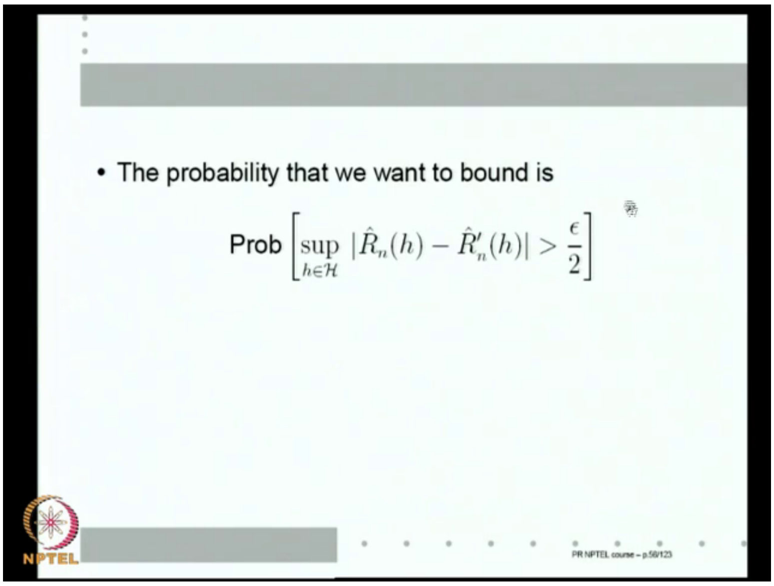
(Refer Slide Time: 31:04)



And hence, the original one R hat n h minus R hat n prime h greater than epsilon by 2 is twice this probability it becomes 4 exponential minus n epsilon square by h. This is for a specific h, but any given h this is true, but the bound is independent of h it does not depend on h. The right hand side does not depend on h because it does not depend on h the same bound holds with any h no matter what h I want I can use the same bound.

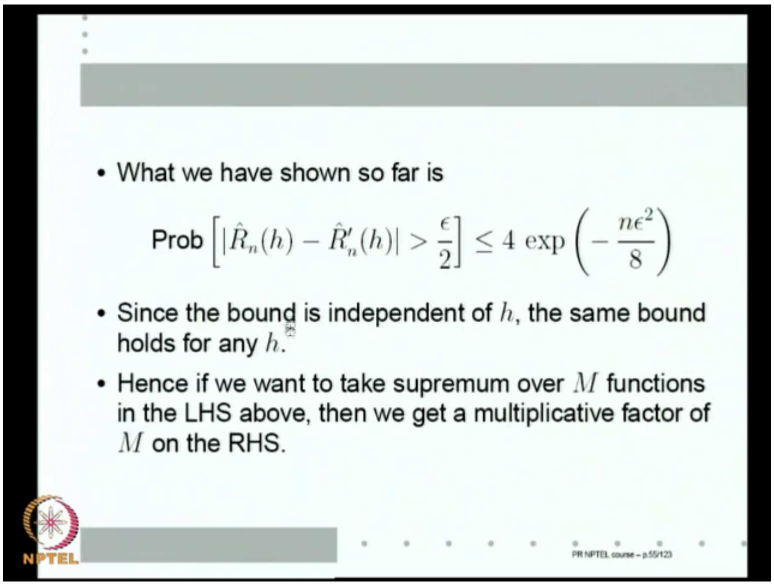
So, if I change the h in the left hand side I do not have to change the bound here, which means like in our finite case if we take supremum in the left hand side probability over m functions. Then we get a multiplicative factor m here, that is what we did earlier. So, essentially if I am taking supremum over h for over some m different functions h here then this simply becomes 4 m exponential minus an epsilon square by 8.

(Refer Slide Time: 32:00)



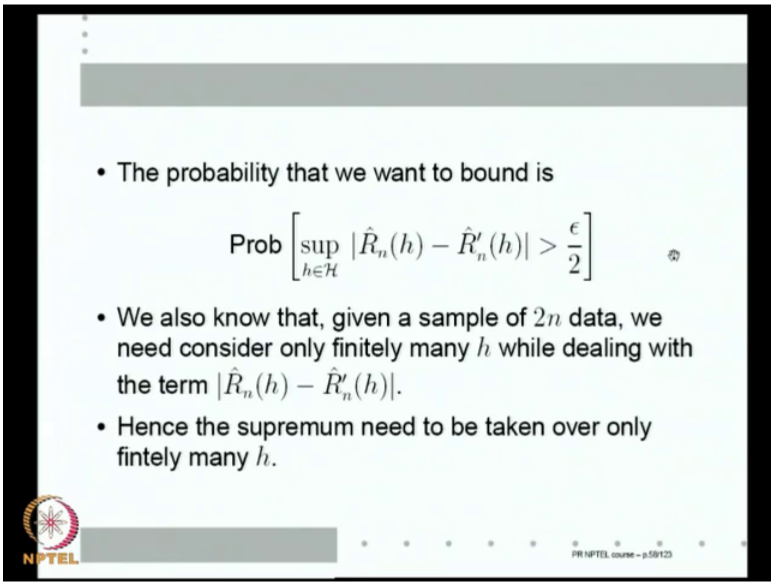
So, let us go over this argument again the probability that we actually want to bound is this.

(Refer Slide Time: 32:07)



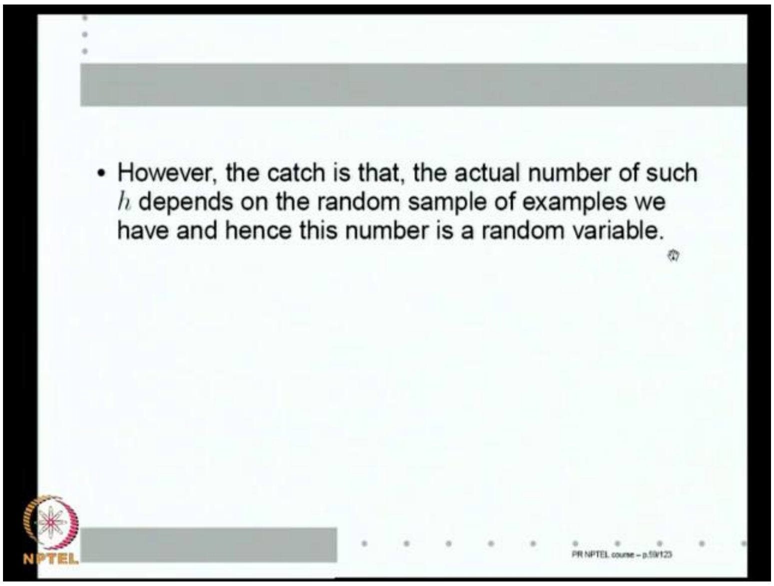
What we bounded is for a particular h, but with a bound that is independent of h, but we actually want to bound is this.

(Refer Slide Time: 32:12)



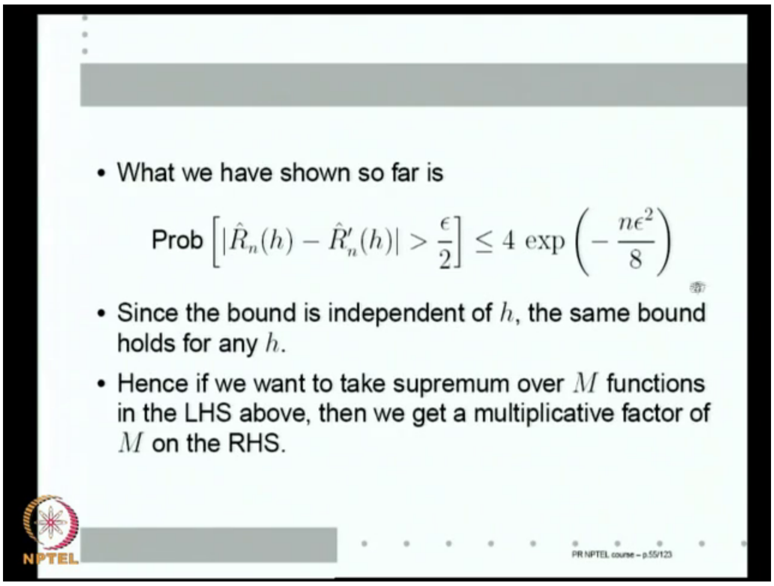
We also know that because these are R hat n's given sample of 2 n data will it consider only finitely many h, while dealing with them. While dealing with this term I need to consider only finitely many h. So, I can take the supremum of finitely many h and when I take supremum of finitely many h I think that I can put that finite number on the right hand side and supremum need to be taken over only finitely many h.

(Refer Slide Time: 32:40)



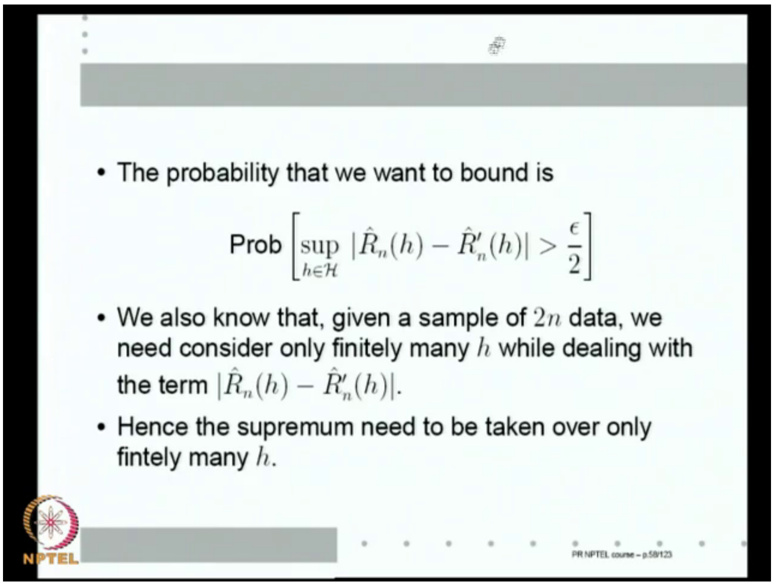
The real catch everything is, so far looks very nice.

(Refer Slide Time: 32:45)



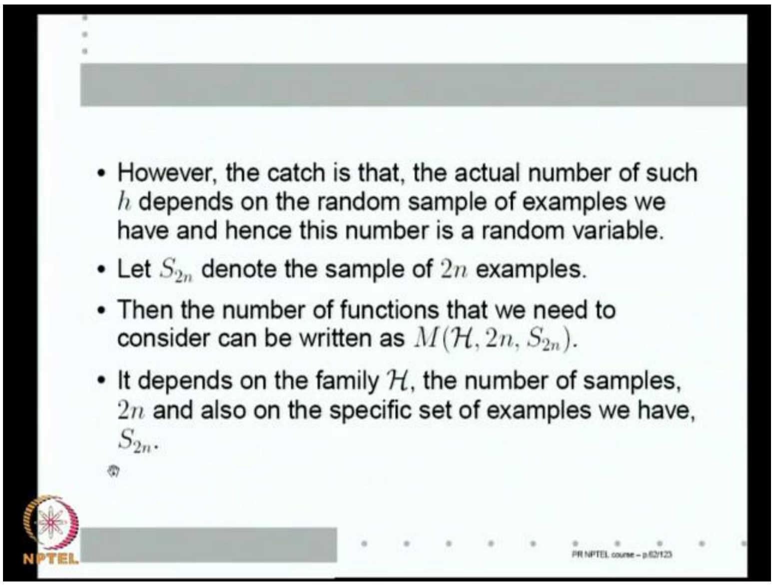
So, basically it looks like I have for one h I have this, and I want the supremum.

(Refer Slide Time: 32:53)



And supremum need to be taken only finitely many things. So, I can just use the that bound and put that number there.

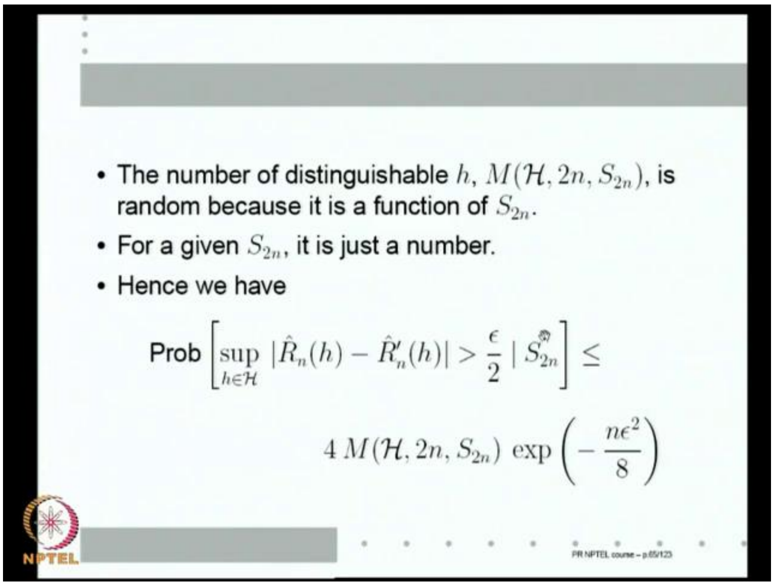
(Refer Slide Time: 33:05)



Of course, obviously, things cannot be that simple, the catch is the following this actual number of such h depends on the specific random sample of examples you have in that sense that number is a random variable. Thus it depends on what are the specific X i that I have gone in this 2 n sample. So, if S 2 n denote this a sample of 2 n examples, then this number of function that we need to consider has to be written as a function of h 2 n and S 2 n.

This function obviously, depends on what is the maximum number of functions I need to consider for this supremum, very much depends on what functions are there in h. So, it is a function of h certainly it of course, depends on number of samples, but in addition to that it also depends on the specific examples we have. So, it can directly use this number on the right hand side for this probability bound because it is a random variable, what do i do about that.

(Refer Slide Time: 34:17)



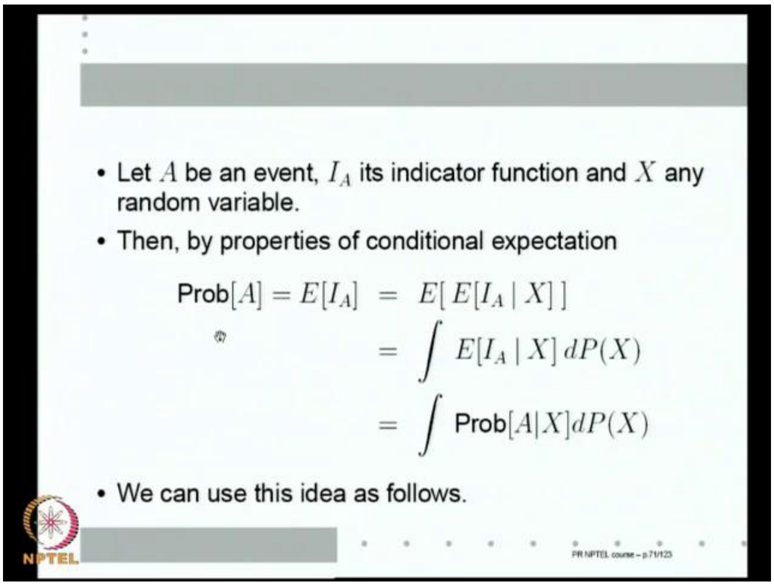
What I do about that is the number of distinguishable functions in h distinguishable based on a sample of 2 n samples, which now we given this symbol. M of H 2 n S 2 n is random, because it is a function of the specific 2 n samples. What it means is if I can consider a particular 2 n sample, then it is a given number. That is what it means for any given particular sample of 2 n it is a specific number, but the a sample of 2 n examples is a random variable random vector.

And hence this number is also a random vector meaning if I can condition my probabilities on a specific sample of 2 n then I can use this number. So, we can certainly write now supremum h belonging to H R hat n h minus R hat n prime h is greater than epsilon by 2 conditioned on the random variable S 2 n will be less than equal to 4 times this number now.

M H 2 n S 2 n exponential minus n epsilon square by 8, earlier we shown that for any one h R hat n h minus R hat n prime h greater than epsilon by 2 is bounded above by 4 times exponential minus an epsilon square by 8. Now, given a particular a sample of 2 n examples the random vector S 2 n we know there are this many distinguishable functions for the supremum. So, now the supremum probabilities bounded above by this number of course, this is nice as far as it grows, but this is not the probability we want to bound.

This probability we do not want to depend on the sample, so we want the unconditional probability here not conditioned on S 2 n, but we now know how to bound this probability conditioned on S 2 n. So, can I somehow use the bound on the probability of the supremum being greater than epsilon by 2 conditioned on S 2 n. To find a bound on the probability that the supremum is unconditional probability, this supremum is greater than supremum is the difference between these two is greater than epsilon by 2. So, this is the next question, we can do it using some standard probability tricks, the trick is like this.

(Refer Slide Time: 36:38)



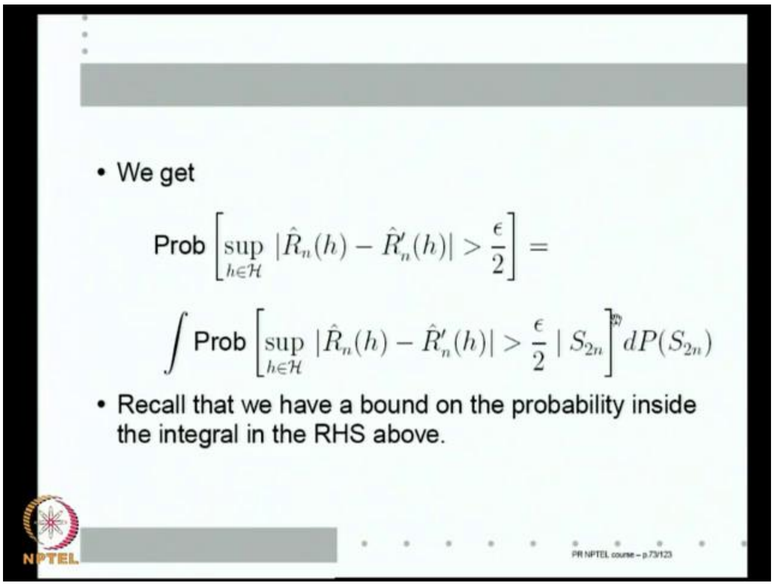
Given any event A, let A be any event and let I subscript A be the indicator function, I A is 1. If A is occurred at 0 otherwise, so this is how events can be converted into random variables, and let X be any other general random variable. How it is related to A is of no consequence to us right now. Then we already seen some properties of conditional expectation, so probability of A is expected value of I of A. The indicator random variable takes only two values 1 and 0.

So, expectation of a bounded random variable the probability takes value 1 and indicator random variable value is 1 whenever A occurs. So, expected value of I of A is probability of A that all of us know and we know expectation is expectation of conditional expectation, so expectation of I A can be written as expectation of expectation of I A given X. The outer expectation is with respect to distribution of X, because the inner conditional expectation returns a random variable as the function of X.

So, the outer expectation is with respect to distribution of X, so I can write this like this. So, I am taking the outer expectation, there is an expectation integral with respect to the distribution of X of the function inside the expectation expected value of I A given X. Now, I A is still a bounded random variable, so expectation of binary random variable is still probability, so this is nothing, but probability A given X d P X. In the sense this is a generalization of the standard some rule we have in probability, we can always write if B 1 B 2 B 3 is a partition of omega.

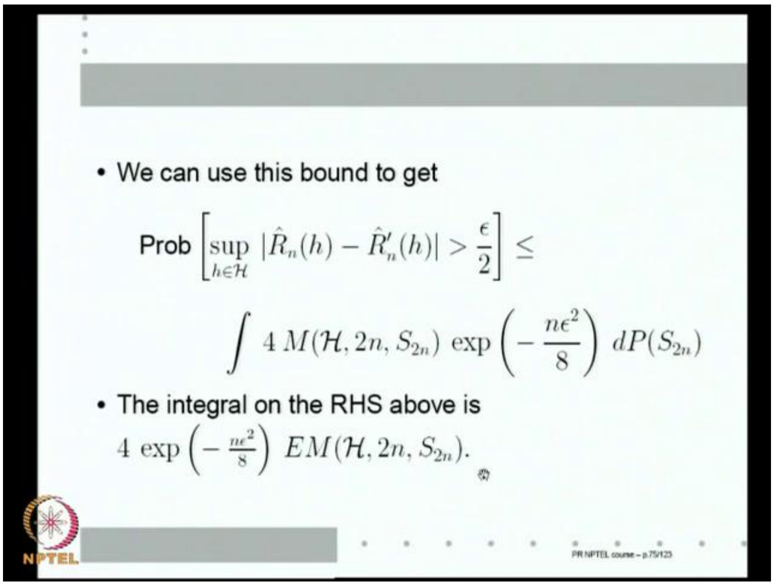
Then probability of A can be written as probability of A given B 1 into P B 1 plus probability A given B 2 into P B 2 plus probability A given B 3 into P B 3, this I can do for any finite partition. Essentially, conditioning random variable says that this rule, this total probability lies is called can be extended to an uncountable summation also by using conditioning on random variables. So, coming back if I have any event and I know the probability of the event conditioned on some random variable. Then by taking expectation of this with respect to that random variable distribution I can get the unconditional probability of a that we can use now.

(Refer Slide Time: 39:08)



We want probability of this event, we know the probability of this event conditioned on the random variable S 2 n. So, I can write the unconditional probability of this event, at the probability of the same thing conditioned on S 2 n into d P S 2 n integral, so it is an expectation integral or less to n. Now, the this term this probability supremum R hat n minus R hat prime n greater than epsilon by 2 condition on S 2 n, we have a bound for this, we have a bound for the probability that is inside this integral. So, we can substitute that bound if I substitute that bound what will I get, so this is what we have.

(Refer Slide Time: 39:54)

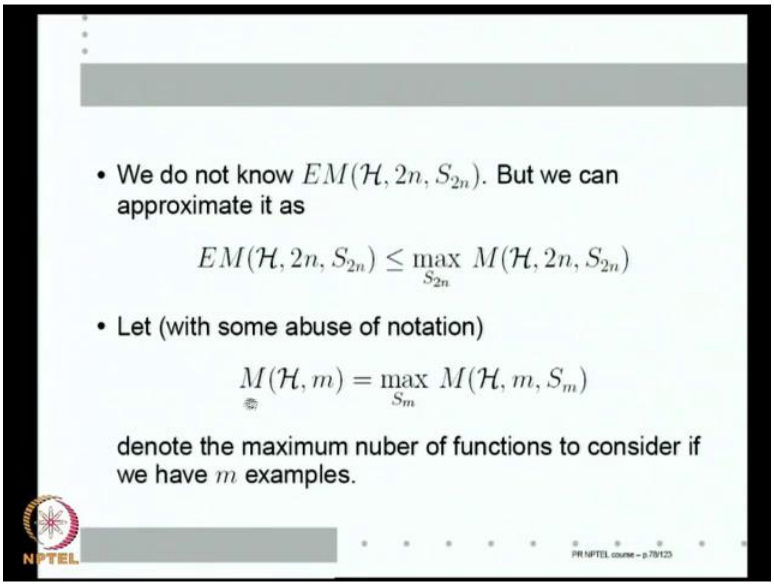


So, this probability is equal to this, but this probability is less than or equal to something else, so by putting that this probability is less than or equal to this is the bound we had this is the bound we had earlier, so we substituted that bound. Now, what is this integral 4 times exponential minus n epsilon square by 8 will come out of the integral, because that does not depend on S 2 n. So, what I am left with is M of H comma 2 n comma S 2 n into d P S 2 n integral, now M H 2 n S 2 n is the random variable that is a function of S 2 n, it is like some g of S 2 n.

So, if I take, so g of S 2 n into density of S 2 n integral that will give me the expected value of that. So, the integral and the R h s is nothing, but 4 times exponential minus n epsilon square by 8 the expected value of M H 2 n S 2 n is the expectation with respect to the various S 2 n's, S 2 n is the random various 2 n samples I can get. So, now I got a proper bound this can be bounded above by 4 times exponential minus n epsilon square 8 multiplied by expected value of M H 2 n S 2 n.

Where, M H 2 n S 2 n is the number of distinguishable functions based on distribution functions, H based on a sample of length 2 n where the sample happens to be specifically S 2 n. Am I done? This is the probability if I bound this probability I can bound the other probability I am interested in this probability now I bounded.

(Refer Slide Time: 41:26)



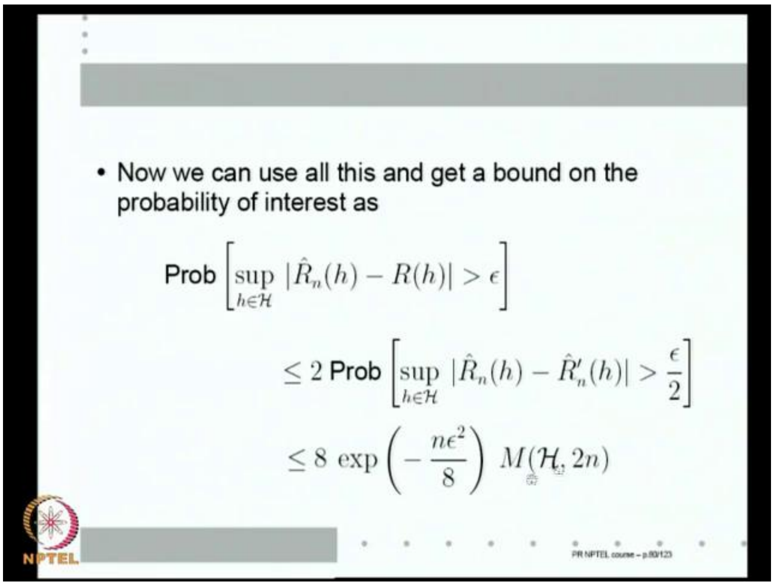
Am I done? Unfortunately I am not done because I do not know expected value of M H 2 n S 2 n of course, I do not know M H 2 n S 2 n anyway, for a given S 2 n also I do not know I have I have not told you how to calculate. The expectation anyway cannot do because I cannot take the expectation with respect to S 2 n, because that involves the probability distribution with X y X and y come. But, what we can do is we can approximate the expectation, we can bound the expectation itself by the maximum of M H comma 2 n S 2 n is the maximum is taken over all possible samples S 2 n.

The expectation of any random variable has to be less than or equal to the maximum value that random variable takes. And hence if expectation can always be bounded above by the maximum over S 2 n of M H 2 n S 2 n, because I really do not know at this point of time how to calculate this. And hence I do not know how to calculate this, but will tell you later on that calculating this is not that difficult (())not; that means, there are ways to conceptualize this calculation.

So, if for a for any given 2 n sample I know how to calculate this, then I can calculate or if I can somehow find the maximum possible. Then I am done, I do not need the distributions, I do not need the expected value. So, let us say it is of course, an abuse of notation, but we will say a max of M H m S m for any integer M, we write it as M H comma m of course, m is originally given as function with 3 arguments.

Now, I am calling 3 2 argument function, but this kind of abuse of notation is often done. So, we will write M of H comma m little m as max over all possible m length samples of the number of distinguishable m length samples for any particular samples. So, this denotes the maximum number of functions to consider if I have m examples right. So, with one specific m sample there might be some number of maximum functions to consider, with another specific m sample there might be some other number of functions that we need to consider to take the supremum. We are saying over all possible multiples of examples we could draw, what is the maximum number of distinguishable functions? That is what this is we can certainly bound this above by that, so that is what we called M of H m.

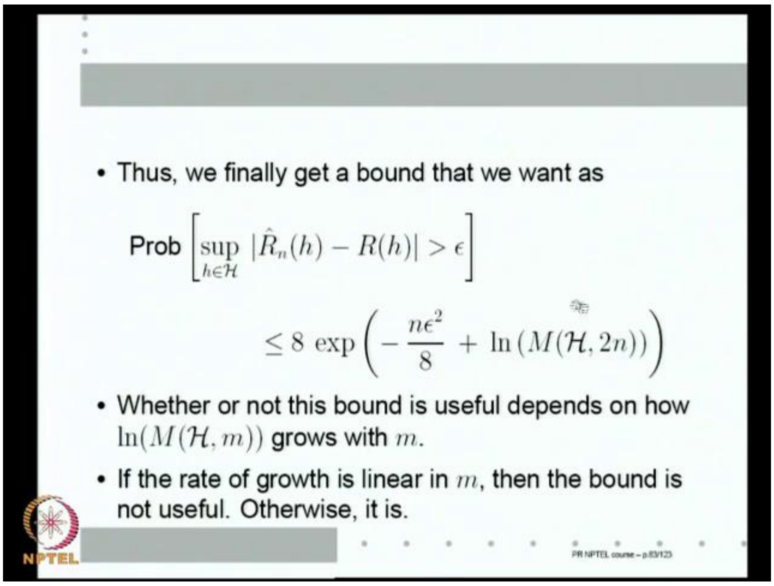
(Refer Slide Time: 43:59)



So, now we can use all this to get a bound on the probability of interest, see this is what we started with we want to get supremum h belonging to H R hat n h minus R h greater than epsilon, because we want to know whether R hat n converges to R uniformly over h. Using symmetrization argument, we can bound this by twice the probability that R hat n h minus R hat n prime h greater than epsilon by 2, where these are sample mean estimate this empirical risk or the sample mean estimate is obtained on a 2 n sample using the first n and the second n.

Now, with what we have done, so far this in turn can be bounded above by 8 exponential minus n epsilon square by 8 M times h comma 2 n, where M H comma 2 n because this is based on these are calculated based on 2 n samples. I have to put 2 n there, M H comma 2 n is the maximum number of distinguishable functions in h over all possible samples of length 2 n. So, this is our final bound now, let us write this bound, so this M H 2 n I want to push it under this expectation. So, under this exponential. So, I can always write it as e to the power L n M H 2 n, so then it comes into this exponential function.

(Refer Slide Time: 45:28)

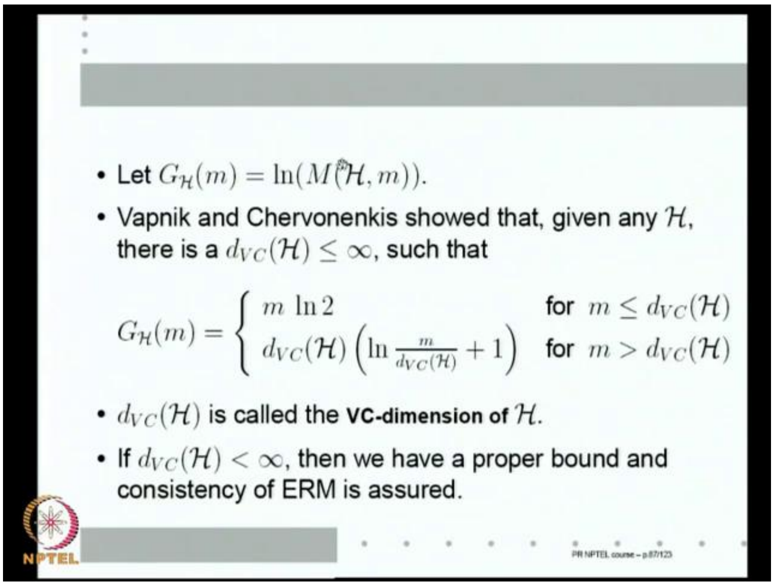


So, by doing that I can write this bound as 8 exponential minus n epsilon square by 8 plus L n M H comma 2 n is this a good enough bound, will it let me show that this probability can be made as small as I want by taking n as large as I want is this good enough. Well, it is good enough or not depends on how this number grows with this, this is a log of the maximum possible number of distinguishable functions based on 2 n samples.

So, we want to know this number; obviously, depends on 2 n and hence on m, so for a given m we want to know how L n M H comma m grows with m. If the logarithm of capital M grows linearly with m then this bound is not useful at all, if this logarithmic term grows linearly with 2 n or linearly with n, then this we have inside the exponent we have two terms. One term linearly decreases with n, another term linearly increases with n, so we cannot say that this can be made as small as we want by taking n sufficiently large.

On the other hand if this grows less than linearly with n, let us say it grows only a logarithmically with n. Then we are done because this falls off linearly as n this grows only logarithmically as n and hence the R h s can be made as small as we want. So, whether or not this bound is useful, useful in the sense whether or not we with this bound we can show that this probability can be made less than delta. Depends on how L n M H comma m in general or L n M H comma 2 n will grow with the number of samples, so everything now hinges on how this function grows with m..

(Refer Slide Time: 47:25)



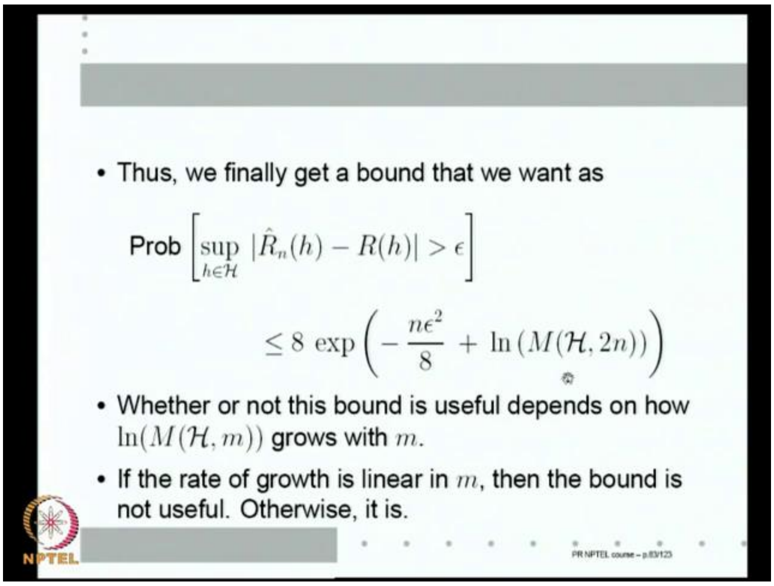
So, here is another great result which unfortunately will not be able to prove, so let us give a name for this thing, let us call G H of m to be this logarithm of the number of distinguishable functions from h over all possible m samples. Let us call that G H of m, it only depends H and m, so I put h as subscript and m as the main argument. In a in a very seminar paper in 70s Vapnik and Chervonenkis showed that for any family h, there is a number which we called d subscript V C of H which may be infinite.

That is why I said less than or equal to infinity such that as long as m is less than or equal to d V C of H G H m grows linearly in m. m l n 2 basically what it means is the capital M itself grows as 2 power m, that is why l n of that will become m l n 2. We already know that given m samples the maximum possible distinguishable things on any n sample is 2 power n. So, what this result shows is of course, in the beginning the maximum distinguishable number of functions grows as 2 power m as the intuition says, so till some number which we called d V C H.

Till m reaches the d V C H G H m grows linearly or capital M grows as S 2 power m, but after reaching the size d V C H the G H m this is l n of capital M will grow only logarithmically enough. Actual, growth is d V C H into l n m by d V C H plus 1, but the growth is logarithm which means capital M will grow only as linearly in m. So, this d V

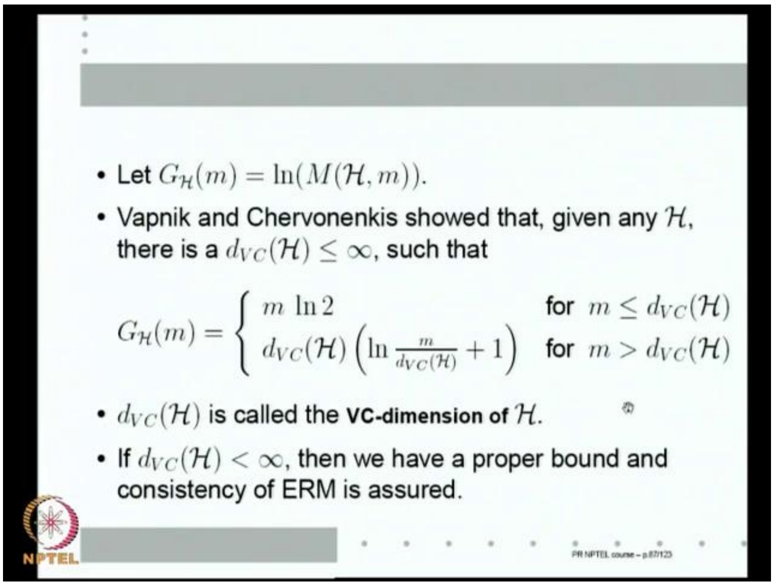
C of H is called the V C-dimension of H, because a number that is that we associate with h is called the V C-dimension of H. So, what this says is that if the if the V C-dimension is less than infinity, then we have a proper bound because there is some number some number which we call d V C of H after that this l n will not grow linearly.

(Refer Slide Time: 49:58)



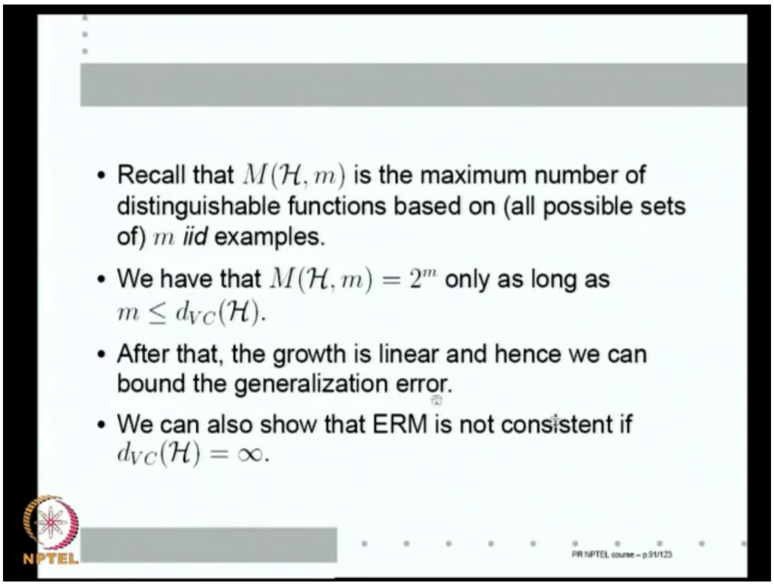
So, after some n which is determined by the d V C of H, this number will not grow linearly, and hence this whereas, this is falling of linearly this will grow only logarithmically and hence this can be made less than delta.

(Refer Slide Time: 50:15)



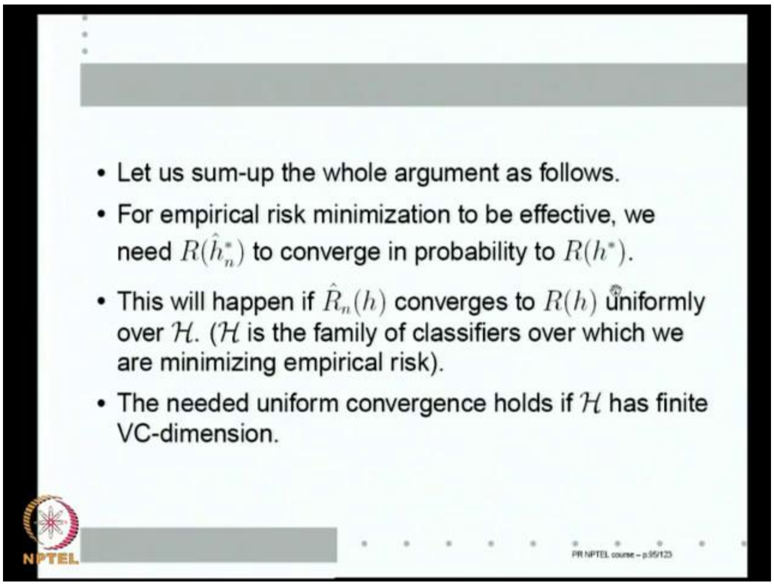
So, if the V C-dimension of H is less than infinity then we have a proper bound and consistency of E R M is assured. We still have not comeback to define this, but we will we will see a d v the V C-dimension of H in next class much more detail, but all it says is give me any H, there is some number which we call the V C-dimension of H. And the number of distinguishable functions grows exponentially only till as long as the sample size is less than V C-dimension of H after that it does not grow exponentially.

(Refer Slide Time: 50:49)



So, M H comma m is the maximum number of distinguishable functions based on all possible sets of m iid examples. So, what we have is that this maximum possible number we know it is the real maximum is 2 power m, it as 2 power m till m reaches d V C of H. After m reaches d V C of H the growth is linear and this means once the because the growth is linear we can bound these generalization error. Of course so, if d V C h is less than infinity we can bound d V C h is equal to infinity we cannot bound by our method. But we do not know whether there is any other bound, but one can show that if d V C H is infinity V C-dimension of H is infinity then we cannot bound and E R M is not consistent.

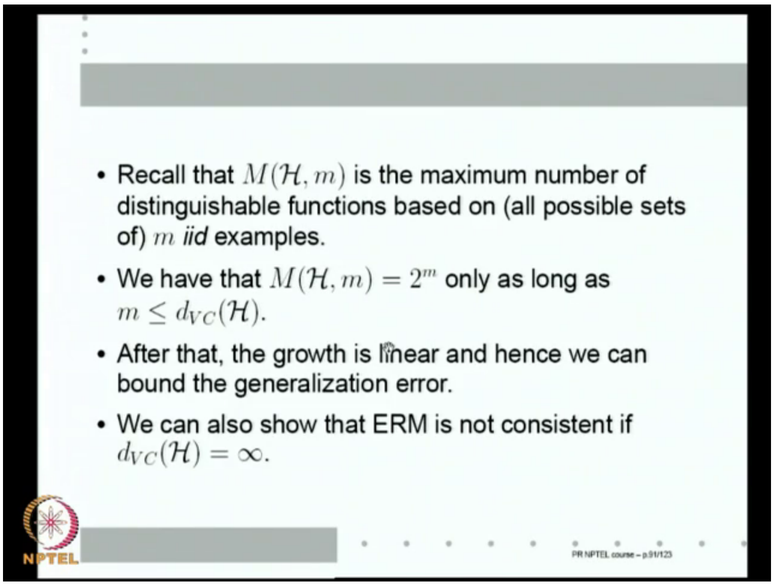
(Refer Slide Time: 51:38)



So, let us sum up the whole argument that we got, so far, for empirical risk minimization to be effective, what we want. So, we are we are finding the global minimizer of empirical risk, and we were asking how close is it to global minimizer of true risk. Now, closeness is only in terms of the true risk of a function right, so I am asking is the true risk of what a l n, what a l n is h R star n, minimizer of empirical risk. Is the true risk R of h hat star n close to the global minimum possible risk which is R h star.

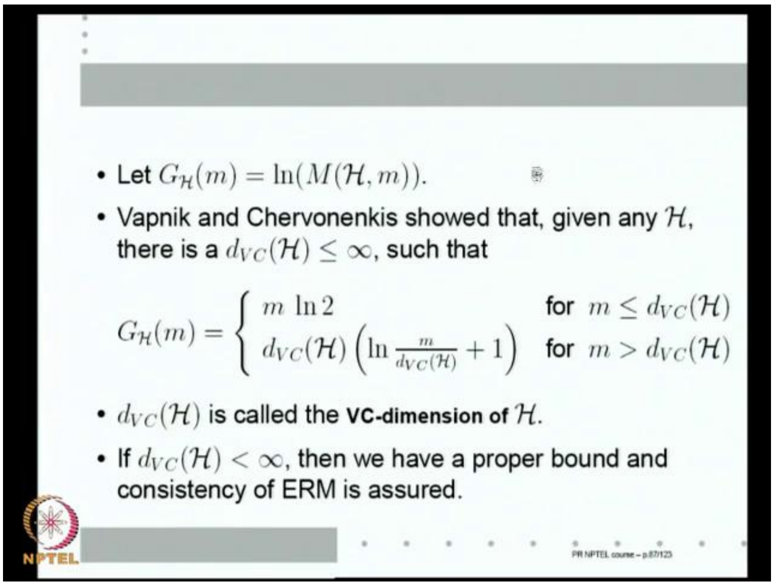
So, for empirical risk minimization to be effective this is what we want, we seen that this will happen, if the R hat n h which is the sample mean estimator of the true risk based on n iid samples converges to true risk. That is the expected value of loss uniformly over h we go anywhere that the sample mean the the mean of the expectation of loss function that we want can be obtained in the limit as the as the sample mean using law of large number. But, the question you are asking is is this convergence uniform over H? Where H is the family of classifiers over which we are minimizing empirical risk. What we are now formed is that the needed convergence holds if H has finite V C-dimension, where what is V C-dimension?

(Refer Slide Time: 53:14)



So, far V C-dimension for us is if M H comma m is the maximum number of distinguishable functions based on all possible sets of m iid examples. Till m is less than equal to d V C of H this maximum stays as 2 power m and after that it is less than 2 power m.

(Refer Slide Time: 53:36)

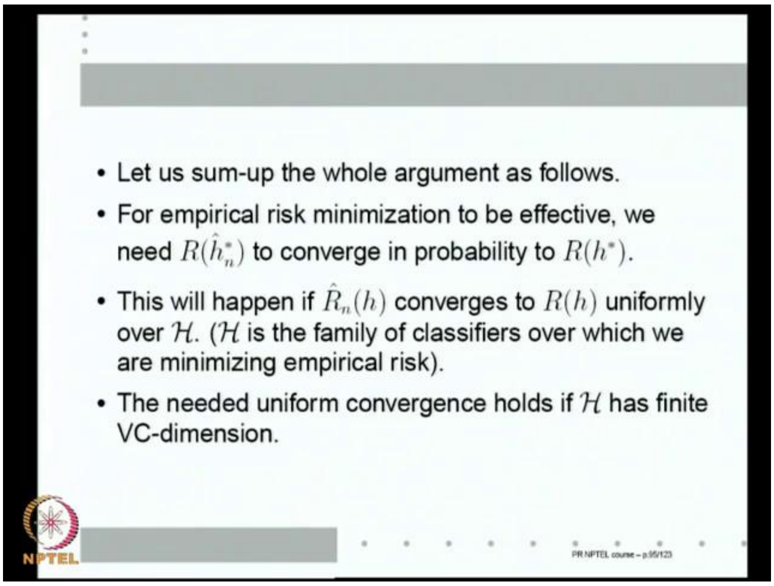


Why should there be such a number because this price have showed that that is the only possible way this function, the maximum number of distinguishable functions from a family of classifiers based on m number of samples. That number if you take logarithm of that number that function has only two possible growth rates it grows as m l n 2 that is capital M grows as 2 power m till some integer.

After that it grows logarithmically in m what they showed is given any H, there is such an integer that we can associate with H. Rather that till that time this grows exponentially after that it does not grows exponentially. Of course there can be H for which there is infinite in which case it forever grows as 2 power m, that is the reason why you cannot learn if the V C-dimension is same.

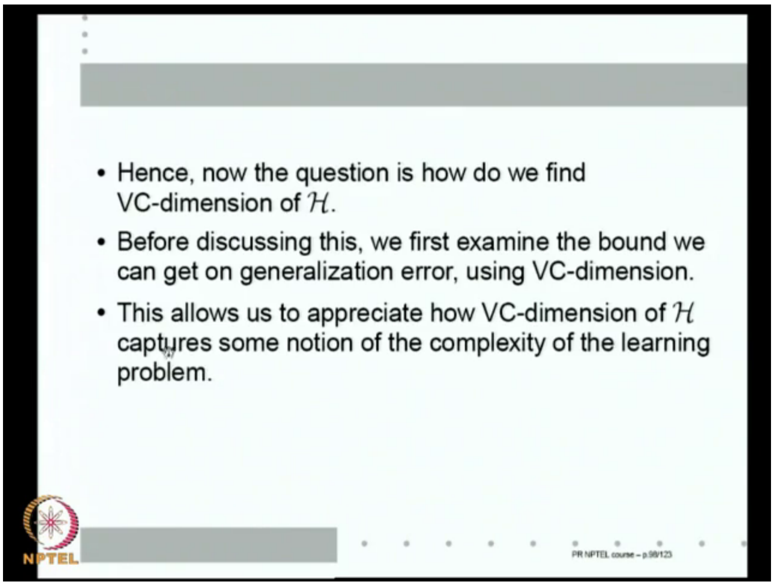
So, this is all we have defined V C-dimension to be, so at this point of time except that because of the Vapnik Chervonenkis result. We know it exists we still have not really conceptualized, how to calculate V C-dimension of given h that we will do later on.

(Refer Slide Time: 54:46)



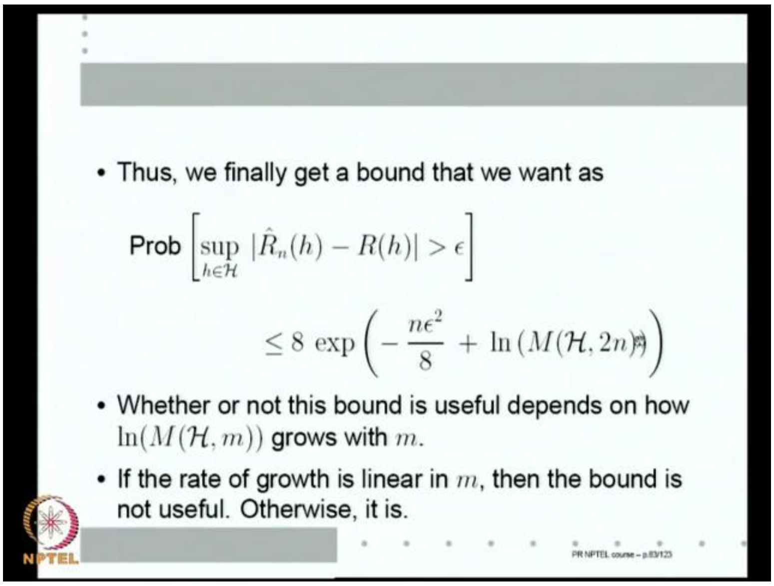
But, let us understand our summary empirical risk minimization to be effective, we want R h hat star n to converge to R h star. We seen that this will happen if the law of large numbers convergence of R hat n h to R h is uniform over the family H. And the uniform convergence will happen, if we h has finite V C-dimension.

(Refer Slide Time: 55:18)



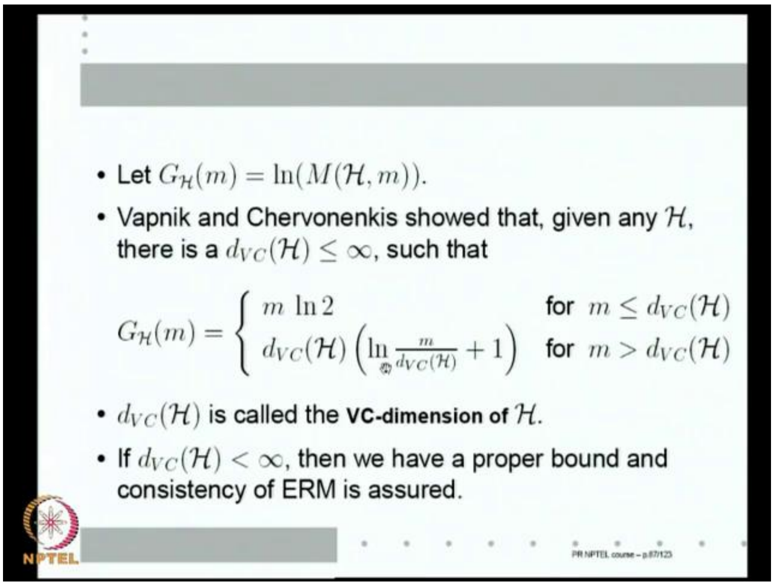
Hence, the question now is to find the V C-dimension of H, how do we find V C-dimension of h, but before getting into V C-dimension of H which we will do in the next class. We will we will very quickly take a look at what is the kind of bound we obtain on the generalization error. Actually, what we will do is the following, see we have this bound that we obtained on the generalization error.

(Refer Slide Time: 56:05)



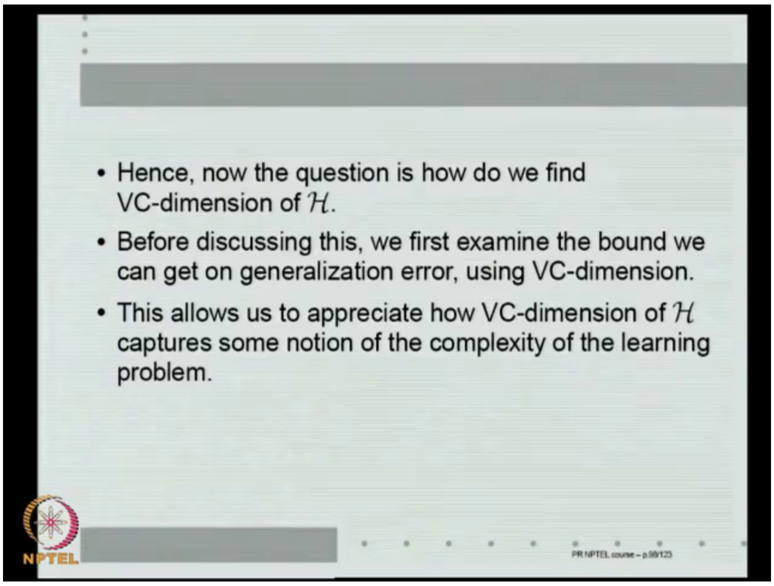
This is essentially generalization for us, this is how many examples see if I need to calculate n, so that this is less than delta. That will tell me the uniform convergence hold, that will tell me how many examples I need before R hat n h is close to R h for every H in my family.

(Refer Slide Time: 56:33)



Now, this depends on the V C-dimension, so by by looking at this bound and what we know about how this l n M H m can grow.

(Refer Slide Time: 56:46)



We can actually, so we can actually get some interesting idea about the generalization error. So, what we will first do this also we will do next class, is we we will examine this bound, so that we we understand what V C-dimension is giving us. V C-dimension not only tells us whether or not the needed uniform convergence holds, and hence whether or not the empirical risk minimization is effective. It will also give us as we shall see some idea of the complexity of the learning problem, how complex is it to a learn a particular H.

So, what we will do next class, is we will understand how this bound that we derived brings out the issue of complexity of H. And after that we will see how starting from the definition that we gave here we can actually conceptualize a procedure for calculating V C-dimension of a given set of classifiers. Then we will go back and calculate, it for some of the examples we have seen earlier.

Thank you.