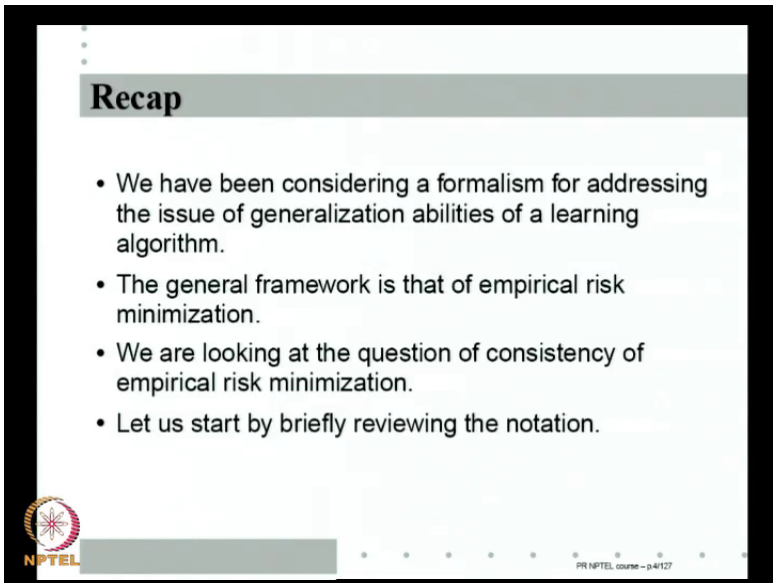
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 22
Consistency of Empirical Risk Minimization

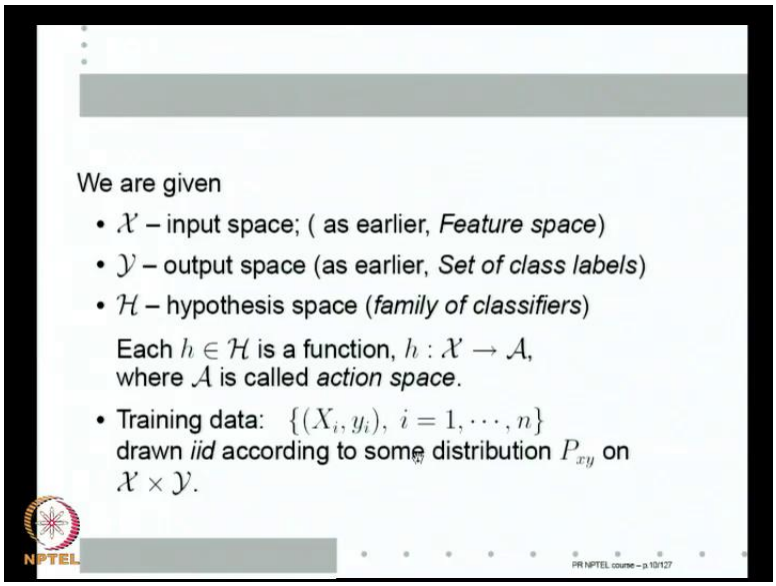
(Refer Slide Time: 00:43)



Hello and welcome to the next lecture in this pattern recognition course. We have been considering some basics of statistical learning theory specifically we are looking at a formalism that will allow us to address the issue of whether a learning algorithm learns correctly or what we call the generalisation abilities of a learning algorithm. So, the we have, we are basically considering a formalism for addressing the issue of generalisation abilities of a learning algorithm. What we have looking at is the general framework of empirical risk minimization.

We, are looking at approaches that define a suitable loss function, and then try to minimise the risk that is the general framework we are following. And the I, the question we are looking at is what I told you last class it is called the consistency of empirical risk minimization; that is that the minimizer of empirical risk, is it close in a proper sense to the minimizer of the true risk. So, let us briefly review the formalism, once again to get our notation, and then go ahead and see what is needed.

(Refer Slide Time: 01:40)



So, this is the formalism that we started with the learning problem. We are given the following, we are given this script X which is the input space which for us is the future space very often it is a d dimensional Euclidian space. All the input feature vectors come from X, but because it can be classification regression anything we, we use an abstract symbol X for the input space. Similarly, Y is the output space for the classification problem this sort of class labels; if I considering two class problem Y will be 0 1 or plus 1 minus 1, m class problem; it could be 1 to 3 m or it could be m unit vectors depending on what representation you want to use, for a regression problem it will real line and so on.

So, the target values are in the set Y that is called the output space. Then we had what is called the hypothesis space, this essentially is the family of classifiers or family of models or family of functions over which you are searching to find the one during the examples. As you know in, in our simpler earlier formalism we called it hypothesis space. And that allows only binary valued functions, but in general it can be a arbitrary set of functions from functions with domain X.

So, specifically we defined elements of this hypothesis space to be functions that map the input space X into some other set a which we call the action space. So A could for example, be Y if we think that each h is a classifier and say for example, Y could be 0 1. Then each h each little h which is a element of hypothesis space could be a binary valued

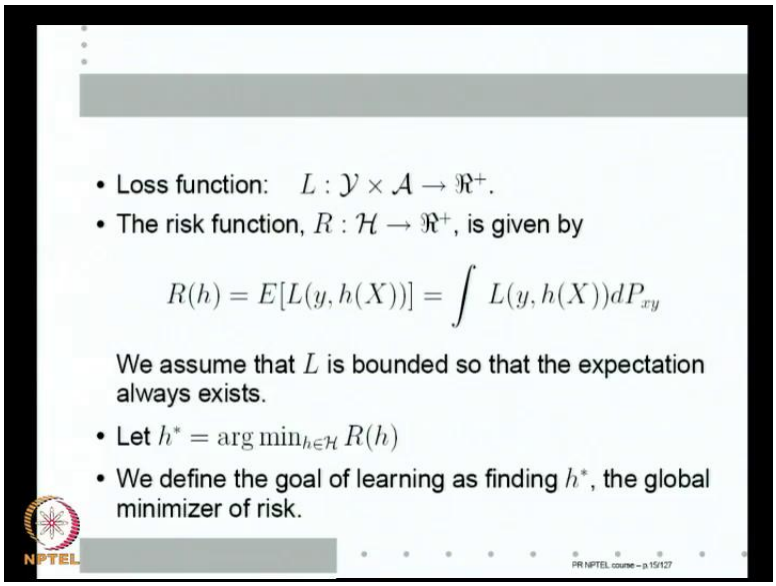
function on X or even when Y is equal to 0 1 as we saw A could be the real line then h assigns a real number to each feature vector so it is like a discriminant function.

So, you know this, this is a generic kind of formulation whereby we can accommodate many different learning situations. So, hypothesis space consists of functions that map the input space or the feature space to some other set called action space. Mostly it will be either be class labels or some real numbers. Then the training data is as usual sets X i pi is X i y i via X i in script X y i in script Y which are drawn independently according to some distribution P x y on X cross Y.

So, we are not talking of any target concept as I explained last class. So, there is no one particular hypothesis in the hypothesis space according to which all training data are classified, training data are simply drawn using some distribution P x y on script X cross script Y. So which in particular means as we explained last class given a particular X more than one Y can have positive probability. So, because we can always factorise P x y as the marginal of X multiplied by the conditional of Y given X.

So, the conditional of Y given X is a proper distribution then same X can come from with different with different probabilities. This is like overlapping class conditional densities and so on. There is most often we have such situations of overlapping class conditional densities or other similar source of noise, this that the examples are drawn i i d according to distribution X Y as I explained last class allows us to handle many realistic situations. But now that there is no target hypothesis as such, because given these examples I am asking which h is the best we have to define some goal for learning. So, we have used the loss functions for defining the goal of learning.

(Refer Slide Time: 05:31)



So, a loss function maps Y cross A to R that is given a a target value at a class label and a h X given a Y and a h X. It assigns a real number by a convention we will assume loss values are always non negative that is why we said R plus A. So, the idea is L of Y comma h X tells you on a random example X Y, if I use h to predict Y what is the loses of R. Then we define the risk function so on a random example Y X comma Y l of Y comma h X tells me the loss I suffer.

So, if I take the expectation of this with respect to the P x y distribution. So, I have written the expectation as a integral with respect to P x y distribution. So, if I take expectation of L of Y comma h X with respect to the P x y distribution that is what we call the risk of the function h. So, risk of h is the expectation of loss, it tells you on the average, the average is defined by on samples drawn according to P x y from X cross Y how well does h do. And doing well is in terms of the loss function which we specify we can decide how to measure the loss we have seen many different loss functions and so on last class.

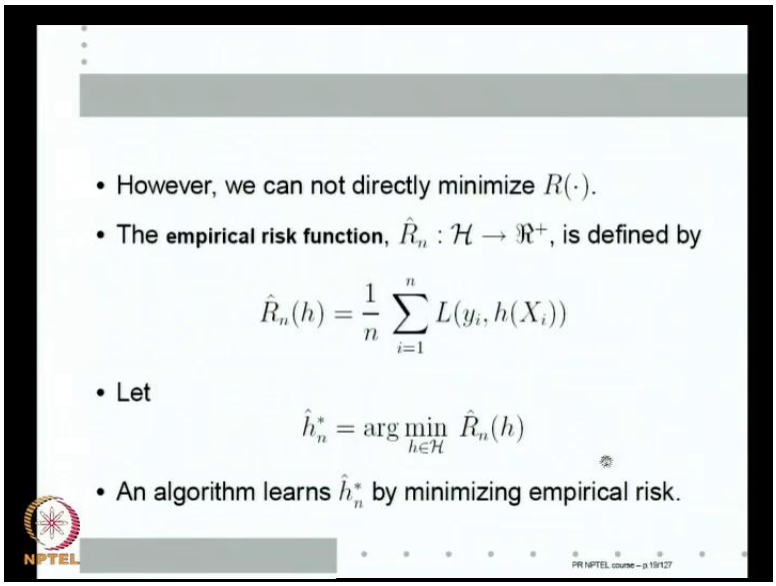
So, this is the risk function and just for mathematical matchness, if you are defining this as the risk function for every h in my hypothesis space, this expectation integral should exist, L is anyway bounded from one side we are assuming it to be non negative. So, we will simply assume L is bounded on the other side also, it is the expectation always exists that is takes away any, any more technical questions on whether for every h R h X and so

on. Most of what we say quite a bit of it can also be said for a loss function that are not really bounded, but satisfy some other criterion, but it really, because we are only looking at a generic flavour of this statistical learning theory results. It does not take away anything if you make a slightly more simplifying assumption actually later on we will make even more simplifying assumptions.

So, we are simply assuming that L is bounded, so that the expectation always exists, we h star is defined to be the global minimizer of h. So, it is the argument over the script h of R of h, so h star is that function in my hypothesis state which achieves the global minimum value of risk. Of course, as we have seen h star may not be unique that really does not matter we will compare different functions only with respect to their risk value. So, the idea is that we, our the goal of learning is to find h star, essentially h star is not unique by h star we mean we want to find something whose risk is same as R of h star that is the global minima of risk.

So, we are essentially interested in the global minima of risk, because we do not have a target concept now we, we, we specified a loss function that tells how to measure, how good the prediction by a particular function is and the expectation of loss is risk. And the goal of learning is to find a function that achieves the global minimum value of risk. The problem of course, is that we cannot directly minimise R given a particular h I cannot calculate R of h, because I do not know P x. So, as a matter of fact as we seen earlier our whole idea of the formalism is we should our algorithm should work no matter what P x y is. So, given a particular h I cannot even calculate R h.

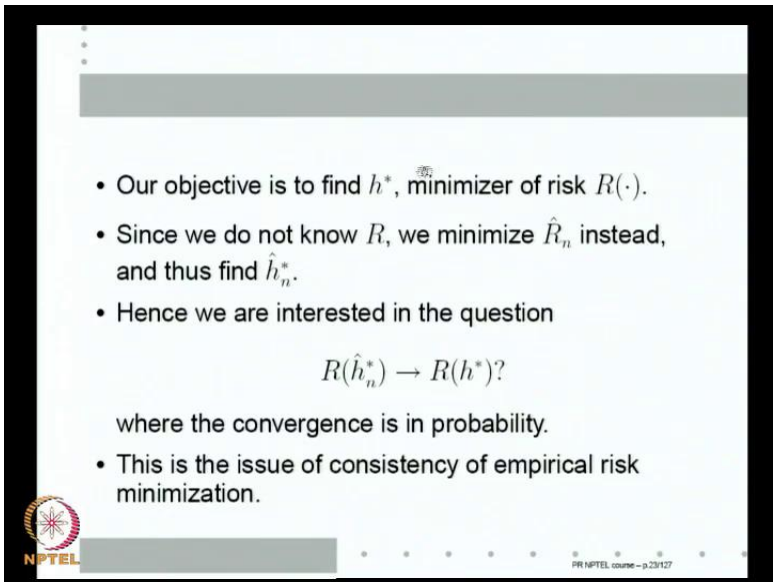
(Refer Slide Time: 09:35)



So, minimising R directly is is not feasible. So, what is our idea? We define an empirical risk function which is actually the sample mean estimator of of R, R is the expectation of L Y comma h X, we have X i y i i i d samples So, instead of the expectation I take the sample mean that is 1 by n L y i h X i summation and that is what we call the empirical risk function.

Given our training data set X i y i I can calculate the empirical risk function for every h and we define the global minimizer of the empirical risk function as h hat star n. So, this is a estimate of h star that is why a hat and estimate obtained through n samples that is why the subscript n. So, we call this h hat star n and what does n e learning running algorithm do? It learns h R star n by minimising the empirical risk, this is our overall framework. So, we are interested; we specify some loss function; we are interested in minimizing risk. So, we want to find a h that h is global minimum risk we cannot minimise risk directly. So, we minimize empirical risk instead.

(Refer Slide Time: 10:53)



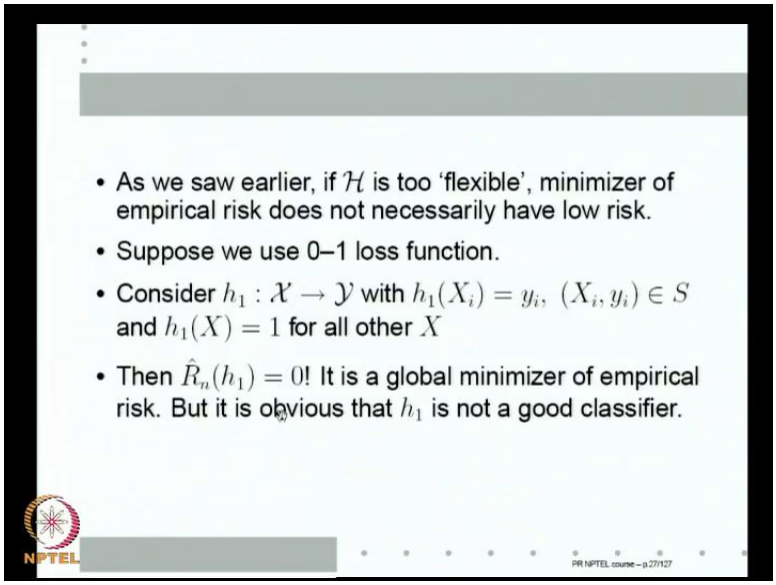
So, let us sum this up, our objective is to find h star which is the global minimiser of risk. Since we do not know R, we cannot minimize R we minimize R hat n you, you cannot do this so, whatever we can do, we do and that is find h star n h hat star. So, the question is is h hat star n a good, good approximation to h star. And as I said we compared 2 functions only in terms of their risk, otherwise the, the whether the functions are same or not does not matter to us. If risk of 2 functions are same as far as we are concerned the functions are same.

So, essentially we are interested in the true risk of h R star n that is R of h hat star n will be same as R of h star, same in the sense of as number of examples goes to infinity. So, we are asking does R of h hat star n converts to R of h star as n goes to infinity. If it does given sufficiently many samples I can be sure that what I learn has a true risk which is not true for away from the global minimum minimizer of risk. As we have already seen h hat star n depends on the random sample. So, it is a random variable so this is sequence of random variable.

So, I have to decide what this sense of this convergence is and the convergence is in probability. So, this is the issue of consistency of empirical risk minimization that is what we want to address now. So, the issue of consistency of empirical risk minimization is I am learning h hat star n while I, I am interested in h star. Of course, functions are distinguishable for me only, based on there the values of their risk. So, is the true risk of

h hat star n close to the global minimum of risk which is R of h star, this is what I want to ask.

(Refer Slide Time: 12:46)



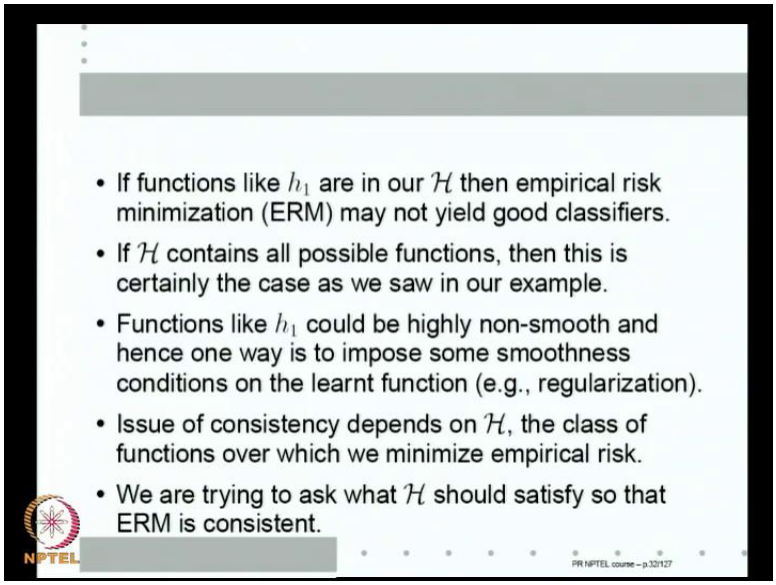
As we have saw earlier the reason the, the question needs to be ask is that is not always; obvious, we can have h S such that if I minimise empirical risk I do not get a good function, we have seen an example where the minimizer of empirical risk gives me nothing. So, as we saw if essentially if the hypothesis space is too flexible it has it has all kinds of functions. Then the minimizer of the empirical risk does not necessarily have a low true risk as a matter of fact. The example we considered we of course, made a particular R 2 classification problem under all that that is to make a example more interesting, but the essence of the example can be stated in just 2 lines.

Let us say we are considering the 0 1 loss function then consider a particular function h 1 that maps X 2 Y, let us assume a is equal to Y. So, all our hypothesis in the hypothesis space are functioned at map X 2 Y. So, let us consider a particular element of hypothesis space which is defined as h 1 of X i is equal to y i, where X i y i are samples that is on each of the training examples h X i h 1 takes the proper value the y i value. And h 1 X is equal to 1 for all other X , what kind of a classifier is h 1? I just memorise blindly all the examples I have seen. If I seen see exactly the same example again I will recall it is classifier, every other example I will simply close my eyes and say it is class one, that is what h i h 1 is doing h 1 of X i is y i for all the training examples X i y i for all other X h

1 of X is 1; obviously, a useless classifier it learns nothing it does not generalise it cannot do anything about unseen term, but its empirical risk is 0.

And hence it is a global minimum of empirical risk while it is obvious that h 1 is not a good classifier is as a matter of fact not a classifier at all that is not what we want. But the empirical risk of h 1 is 0 and hence is a global minimizer of empirical risk. This was the problem we had in our example, we just created an entire two dimensional pattern recognition problem where this happens. But essentially this is the issue if I have functions like h 1 then the there, there will be global minimizers of empirical is there might be other global minimizers of empirical risk. But how can I stop my algorithm from picking functions h 1, because they are also global minimizers of empirical risk there will be many global minimizers of empirical risk. And the algorithm picks up something based on some general (( )) it is difficult to ensure that it will not pick up this kind of things.

(Refer Slide Time: 15:40)



So, the issue is if function like h 1 or in our hypothesis space then empirical risk minimization may not yield good classifiers we may say take simple functions. But as we have seen in our examples, the function like h 1 is in our hypothesis space and that happens to be simple by our definition of simple function. A differential simple function meaning; the smallest set, look like a very nice idea. It is a nice idea if I chose at the H, but if I chosen a two flexible a H as we saw it is a bad choice.

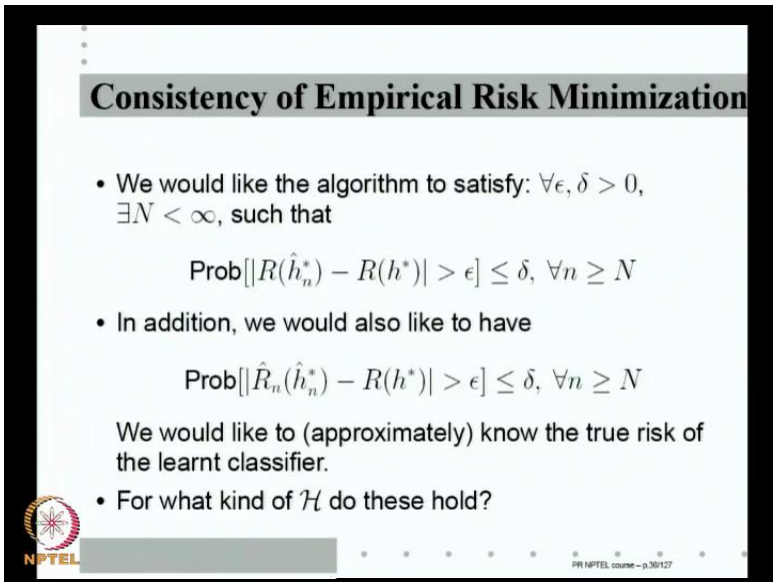
So, it is in general difficult to avoid empirical risk minimization to pullout to converts to functions like h 1, if such functions are there in our H. Obviously, if H contains all possible functions say Y is equal to 0 comma 1, the two class classifier. And if I take H to be 2 power X all possible two class classifiers on H. Then obviously h 1 will be in it and that is what we saw in our previous example as to why we can never learn. Of course, functions like h 1 could be highly non smooth, because at specific points it has to take proper y i values everywhere else it has to take one. So, there might be many discontinuities the function may look very non smooth.

So, one way of imposing conditions is somehow look for nice functions. Of course, it is not always easy, we in our example, we put some criteria for niceness which did not turn out to be good. But we can certainly think of looking for some nice criteria to say the learn function should have some sufficiently smoothness properties that is what regularization was in the in the linear least squares algorithms. We saw regularization that is an algorithmic issue, we will we will come back to that many algorithm that we are going to consider from now on will impose something other than just empirical risk. We minimise empirical risk plus something which will ensure that highly non non smooth functions even though they have very low empirical risk will come over with a very high we are minimizing.

So that is one way to avoid this, but at this point for the next few lectures, we are not looking at how to go about algorithmically ensuring; we get very good functions; we are asking more theoretical questions. We are asking when would empirical risk minimization be consistent. So, we now know that consistency depends on the class of functions over which we are minimizing the empirical risk. If we choose proper class of functions it is all we choose a bad class of functions it may not be all.

So, this is a very important insight that is not just the criteria that you are minimising that is important, but over what class of functions you are minimising that criteria is equally important in deciding whether or not you learn well. So, what we are going to do now is we are going to ask what condition should H satisfy? So, that empirical risk minimization over H would be consistent, so request your addressing under what is known as statistical learning theory. As I said already we are we are only looking at a simple version of this, so to get a flavour of how such results look like.

(Refer Slide Time: 19:08)



So, let us state more formally what is that we want, this is the requirement of consistency of empirical risk minimization. We want our algorithm to satisfy the following, we wanted R of h hat star n to converts to R of h star in probability what does that mean? You give me any epsilon delta positive, strictly positive epsilon delta. Then I can give you a number capital N such that probability that R of h hat star n minus R of h star greater than epsilon is less than delta if I see if the number of examples I have seen is greater than capital N.

So, no matter what accuracy and confidence with which you want me to learn if you give me sufficiently many examples I can learn. So, this is this means that R of h hat star n converges to R h star in probability. So, essentially I want given any epsilon delta to find a N as you already seen N is often N will be a function of epsilon delta and is often called the sample complexity. So the N should satisfy that probability the difference between R of h star h hat star and R of h star being greater than epsilon is less than delta if I have seen at least n examples.

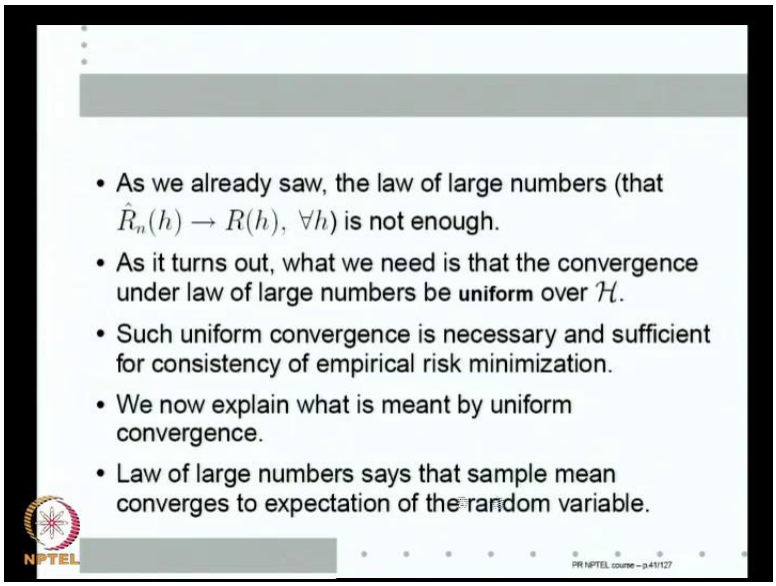
We may also like for under consistency one more thing, not only R of h hat star n should be close to R of h star. But we may want R hat n of h hat star n also to be close to R h star why, why would I need this? Because I, we, we I am learning h hat star n and h star is our symbol for what I want to learn, and any two functions have to be distinguished only based on their risk values. So, this is what I want to be less why should I want R hat

n of h hat star n to be close to R of h star the reason is I cannot calculate, so h hat star n is what I have learnt. Now, I want to know how well h R star n will perform how well h R star n will perform is given by R of h hat star n. But I cannot calculate R of h hat star n but I can calculate R hat n of h hat star n. So, if I actually look at the error that I am getting on my training data with h hat star n does it tell me how well h hat star n in general will perform; obviously, we like to have some knowledge. So, we like to at least approximately know the true risk of what we learnt then we can estimate how well our classifier will do.

So, in addition to satisfying this which is the strict consistency requirement for empirical risk minimization we would also like to satisfy this as it turns out. If I can satisfy this I will be satisfying this also, I am not really asking much we will we will see at least some simple proofs for this. So, this is what we want, so we are asking for what kind of families of functions h do these conditions hold. Now, we are not considering algorithmic issues, we are somehow assuming that given some loss function I have an algorithm that can minimise empirical risk. That can actually find the global minimization of empirical risk which is not a simple issue, we have already seen that the 2 issues in learning; one is optimization and one is statistics.

So, the optimisation part is how for a given loss function how do I minimise empirical risk, we still need to get clever algorithms, we sometimes we may not be able to go to global minimizer of risk and all that, but we will come to that later on. Now, we are assuming that we are somehow finding the global minimizer of empirical risk that is what h hat star n is I am asking is that good enough if I can find the global minimizer of empirical risk. Then will, will the global minimizer of empirical risk will have risk that is close to the global minimum risk that is achievable.

(Refer Slide Time: 22:56)



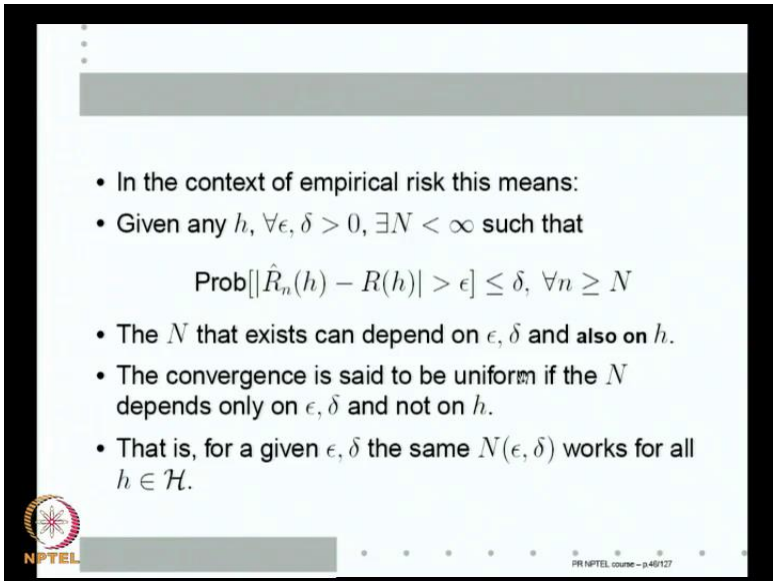
We have already seen that the law of large numbers is not enough for this though our intuitive idea for minimizing empirical risk the intuitive idea of a minimizing empirical risk is that for any h R hat n h converges to R h if there are sufficiently remain examples. So, for every hypothesis the empirical risk is a good approximation to the true risk and hence minimizer of empirical risk should be able to approximation to minimizer of true risk that is the intuition.

But we have already saw in our example in our example problem earlier that this is not enough. As it turns out what is enough is this convergence law of large numbers ensure assures us that for any h R hat n h converges to R of h. It simply says that the sample mean estimator converges to the expected value that is what this means. What we need is that this convergence should be uniform over the family of classifiers H.

The family of classifiers or family of functions we are considering should be such that the convergence assured by a law of large number should be uniform over H. Some of you may not know what uniform convergence means we are we are going to explain that shortly. The uniform convergence is both necessary and sufficient for consistency of empirical risk minimization that is a very very strong result that if a family of classifiers is such that there is uniform convergence then empirical risk minimization is consistent. And if empirical risk minimization over a family of functions is consistent that means

that this convergence will be uniform over that H. So, uniform convergence is both necessary and sufficient for consistency of empirical risk.

(Refer Slide Time: 25:07)



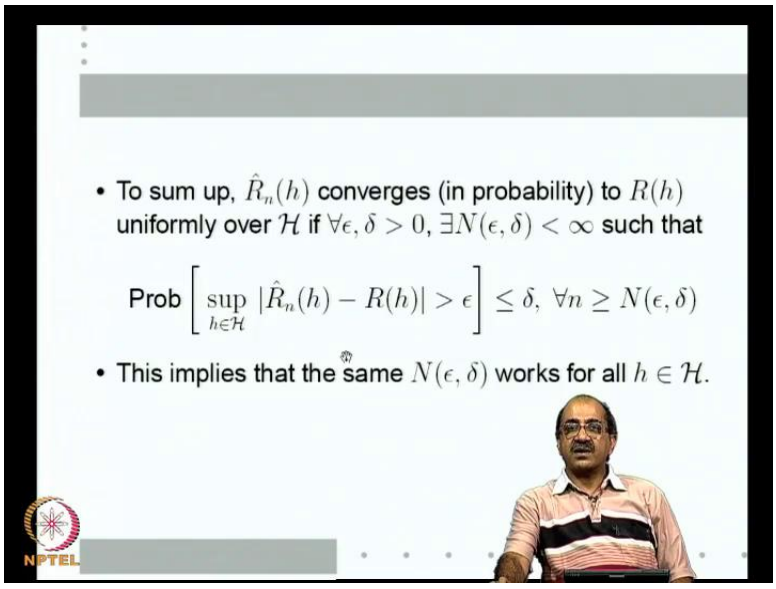
So, first we will explain what is meant by this uniform convergence? What does law of large numbers? Say in in words the sample mean converges to expectation of the random variable that is what law of large number says converges in probability converges in almost truly and so on, but we only want weak law. So, we know sample mean converges to expectation of a random variable in probability which in the context of empirical risk. What does that mean? R hat n h is the sample mean estimator of the expectation R of h for any given h. So R hat n h minus R h greater than epsilon will be a less than delta if I, if I have sufficiently many samples. So, in the context of empirical risk minimization law of large number says that for any for any specific h given any epsilon delta greater than 0. There exists a N less than infinity such that probability that the difference between R hat n h minus R h is greater than epsilon is less than delta.

So, you give me any epsilon delta I will tell you a sample size such that if I have seen that many samples that many i i d samples. Then the, the difference between the sample mean and the true mean be greater than epsilon would be less than delta this is what convergence and probability means this is what weak law of large numbers is. Because when we write epsilon delta definitions, we sometimes do not pay attention to all the necessities even though we may implicitly know them. And we say this n exists the idea

is I fixed a h then you give me an epsilon delta I give you a n. So, what does that mean? The n that exists can depend on epsilon delta and can also depend on h for a particular h for a given epsilon and delta. No matter what epsilon delta you give me I have to exhibit such a N. But you are telling me for this h I want to know whether this converges in probability or not. So to exhibit that I have to say give me any epsilon delta I will find you a N once I can show that. Then the convergence is there which means the n that I have to give you can depend on epsilon delta and also on H.

The convergence is said to be uniform if the N that exists depends on epsilon delta, but not on h. So, for any h and any epsilon delta the N is a function of only epsilon delta naught h, what does that mean? That if I you give some epsilon delta and I give you an a epsilon delta and now this n epsilon N of epsilon delta I wrote n of epsilon delta to explicitly remember that N is a function of the epsilon and delta that we are giving the same number of samples works for all h that is you give me an epsilon delta Then I will give you a number n of epsilon delta if I see that many samples. Then the sample mean estimate of any h will be close to its true expectation that is what uniform convergence means.

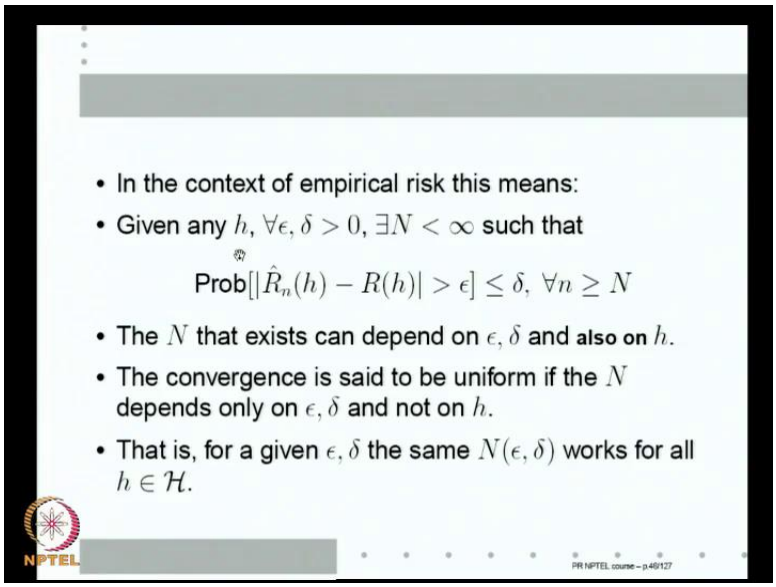
(Refer Slide Time: 28:03)



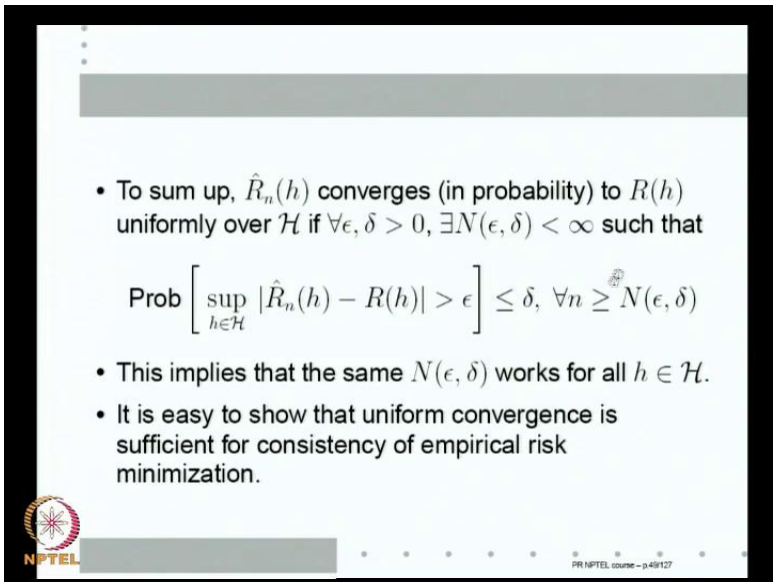
So, we can write it like that R hat n h converges to R h uniformly over the family h if for any epsilon delta you give me any epsilon delta. Then I can give you N of epsilon delta suchthat probability of this difference R hat n h minus R h greater than epsilon. But this

difference is R hat n h minus R h supremum over all h at the among all the elements of the hypothesis space the maximum difference between R hat n h minus R h being greater than epsilon is less than delta. That means if I see N epsilon delta samples then for every single h probability is that the reference is greater than epsilon less than delta, because the probability of the maximum difference is greater than epsilon it is less than delta. So, this definition implies that the same N epsilon delta works for all h belonging to H. So, this is the definition of uniform convergence.

(Refer Slide Time: 29:13)



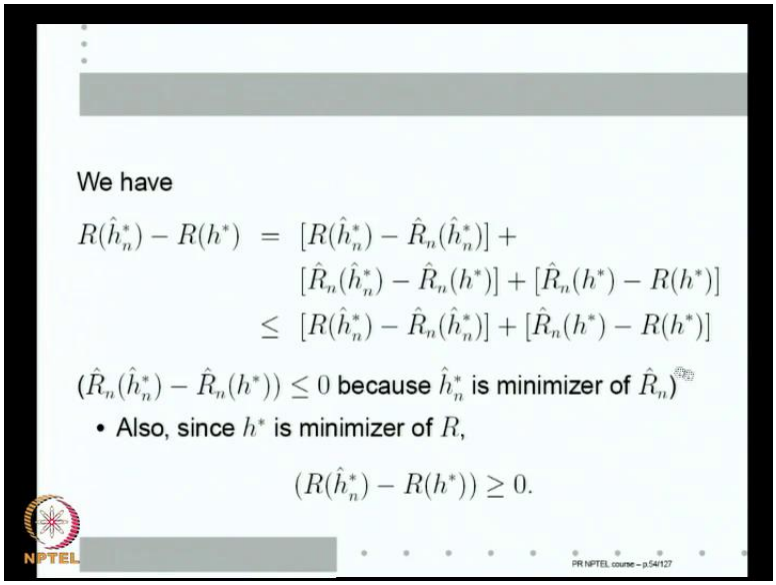
(Refer Slide Time: 29:26)



So, convergence given by the law of large numbers only means this given a h epsilon delta there can be N which can be a function of epsilon delta h to satisfy this for that H. Whereas, uniform means the supremum over all elements of the hypothesis space of the difference between R hat n h minus R h that supremum value being greater than epsilon should be less than delta. We have to put supremum greater than maximum, because h may be infinite if h is infinite maximum may not be defined generally maximum you, you, you say something is maximum if it is attained in the set.

So, for infinite sets maximum may or may not be attained that is why we call it supremum. So, it is easy to, so if we what we said earlier is that uniform convergence is necessary and sufficient for consistency of empirical risk minimization. So, what, what we will first do? We show that if uniform convergence is there then empirical risk minimization is consistent it is because showing the sufficiency of uniform convergence is much easier. So, let us show that at least then we understand where we are using the uniform convergence. So, we will show that if the convergence given by law of large numbers is uniform then empirical risk minimization would be consistent.

(Refer Slide Time: 30:45)



So, what we have to do consistent of empirical risk minimization? We have to essentially show that R of h R star n minus R of h star. Of course, the absolute value can be controlled can be controlled means if you by taking N sufficiently large the difference can be made as small as I want as larger probability as I want. So, let us first handle this

number R h hat star n minus R h star I can write it like this, what did I do? I added a R hat n h hat star n and I subtracted R hat n h hat star n and added a R hat n h hat star n.

Once again subtracted a R hat n h star and added a R hat n h star. So, this I just added and subtracted 2 numbers; one is R hat n h h hat star n another is R hat n h star it makes no difference. Then I grouped them differently what the, what is the idea? See in the first bracketed term I have the same element of h something that is called h hat star n. And I have R of that minus R hat n of that this can be controlled using my law of large numbers. Similarly, here I have R hat n and R on the same element of h some element that is called h star. So, the first and last terms are easily bounded so to say by using law of large numbers.

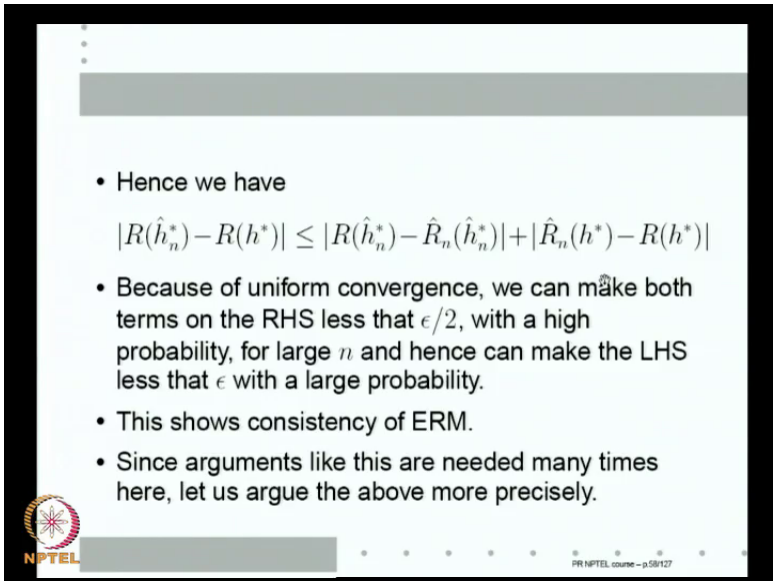
What about the middle term? I can get rid of the middle term, why can I get rid of the middle term? The middle term R hat n h hat star n minus R hat n h star is always negative, it is non positive, why is it non positive? Because by definition h hat star n is the global minimiser of R hat n, so R hat n of h hat star n is less than or equal to R hat n of anything. And hence in R hat n of h star, because h hat star n is the global minimzier of R hat n R hat n of h hat star n minus R hat n of anything. So, in particular h star has to be less than or equal to 0, because this term is negative if I drop this term the R h S value can only increase.

So, this is less than or equal to this, now I got it in a form I want, because both the terms can now be controlled using law of large numbers. Of course, I still need to put absolute values then if I put absolute values on both sides of inequality the inequality may turn around. But in this case it does not, because this is also I have drawn negative quantity. This is non negative quantity why? Once again h star by a definition is one that h is globally minimum of risk.

So R of h star is less than or equal to R of anything, so R of h hat star n minus R of h star has to be greater than or equal to 0. That means this term is greater than or equal to 0, because it is less than or equal to term this term is also greater than or equal to 0 and hence if I put absolute value on both sides it will not change. Now, if I put absolute value on both sides this side is absolute value of what I want R of h hat star minus R h star. This side I have absolute value of sum of 2 numbers absolute value of a plus B as you

know is by triangular inequality less than or equal to absolute value of a plus absolute value of b.

(Refer Slide Time: 34:13)



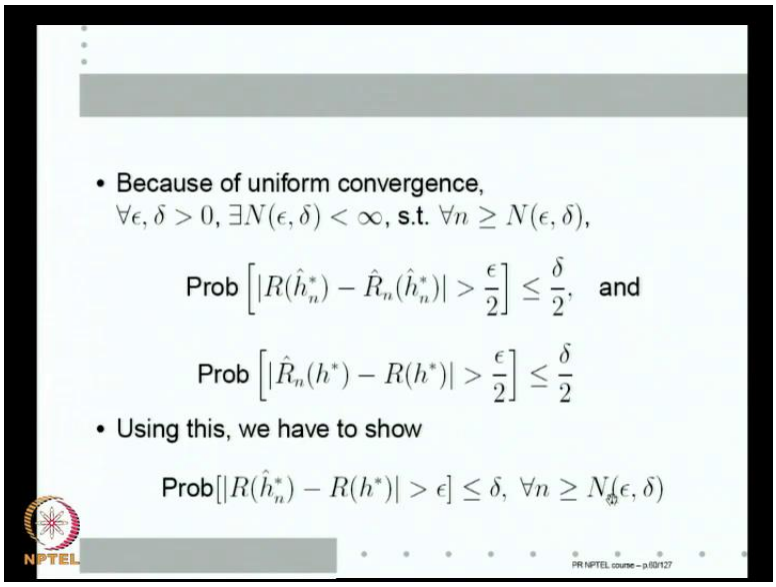
So, what have we got R of h hat star n minus R h star absolute value is less than or equal to absolute the difference between R of h hat star n minus R hat n h hat star n and R hat n of h star minus R of h star. Now, because I have uniform convergence given a particular n the sample mean estimator of every element of h is close to its true value with the same level of accuracy. What does that mean? Because it is uniform convergence I can find a N such that if I have that many samples then both terms on the R h S can be made less than epsilon by 2 with a high probability.

Because this is of course, the law of large numbers for h R star n, this is law of large numbers of a h star. But because I have uniform convergence the same n will work for any 2 elements of h. So, with the same number of I can find a number of sample such that both the terms are less than epsilon by 2, because both of them are less than epsilon by 2. With the same this will also; this will be less than epsilon that is the basic idea of the algorithm, because with the high probability I can make both the terms less than epsilon by 2 with a high probability I can make this term less than epsilon that shows consistency of empirical risk minimization.

In while for the rest of this series of talks and series of lectures and statistical learning theory, we will be using this kind of argument many times, we will have a inequality like

this among random variables. And I say if we can control the hand side you can control the left hand side, if you can make this substitution this small with the high probability. You can make this substitution small since people who come across argument for the first time you know may not really see it immediately, so for one time we will we will make this argument very precise. So, so that we also know how to actually argue it using epsilon deltas.

(Refer Slide Time: 36:30)

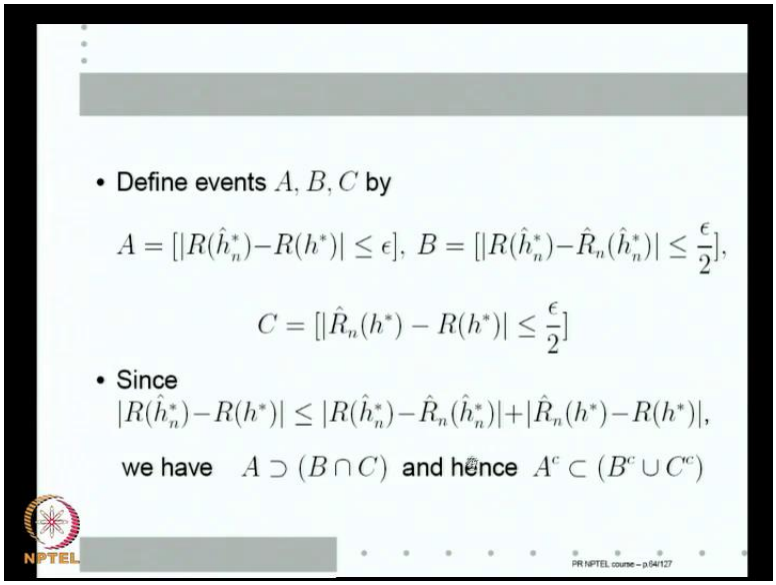


What is that we are saying, because of uniform convergence, you give me any epsilon delta there exists a N epsilon delta such that if I have seen N epsilon delta examples. So, if the number of examples n is greater than capital N of epsilon delta. Then for both h R star and n h star which are 2 arbitrary elements of h the, the, the uniform convergence of law of large numbers whole. So, probability absolute value of R h R star n minus R hat n h R star and greater than epsilon by 2 is less than delta by 2. And similarly, of course, I can do it for any epsilon deltas.

So, I can you give me epsilon delta I can put an epsilon two by delta by 2 and find corresponding n here. So, I can achieve this, the uniform, because we are assuming uniform convergence, uniform convergence guarantees that given any epsilon delta. There exists a n to satisfy this, this is what is given, given this, this is what we want to show, for the same N epsilon delta R h R star n minus R h star greater than epsilon

probability is less than delta, this is what you have to show given this we want to show this.

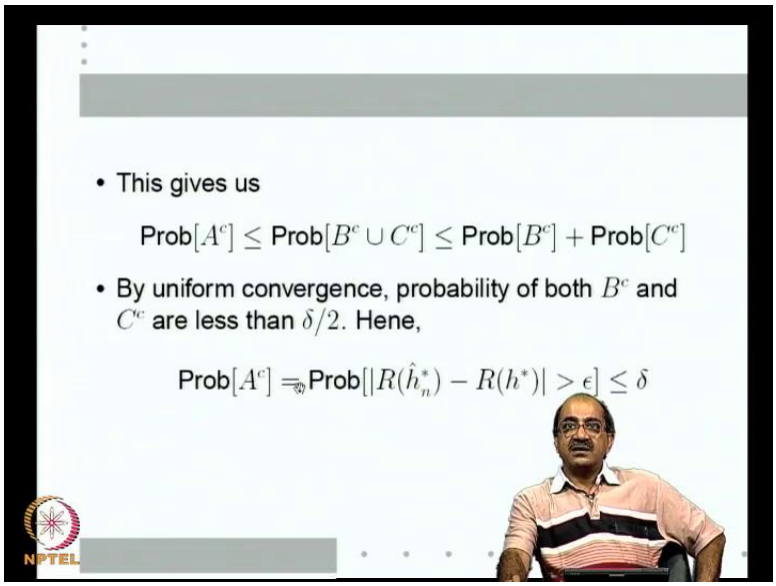
(Refer Slide Time: 37:45)



So, to do this let us define 3 events A B C a is R h R star n minus R h star less than epsilon this is what we want, B and C are the 2 events that we we already know. We can control using law of large numbers this is R h R star n minus R hat n h R star n less than epsilon delta. This is R h star minus R hat n h star less than; this is the inequality that we already proved, what is that mean? If both B and C are true, if the event both B and C happened, if both B and C are true then this is less than epsilon by 2 this is less than epsilon by 2.

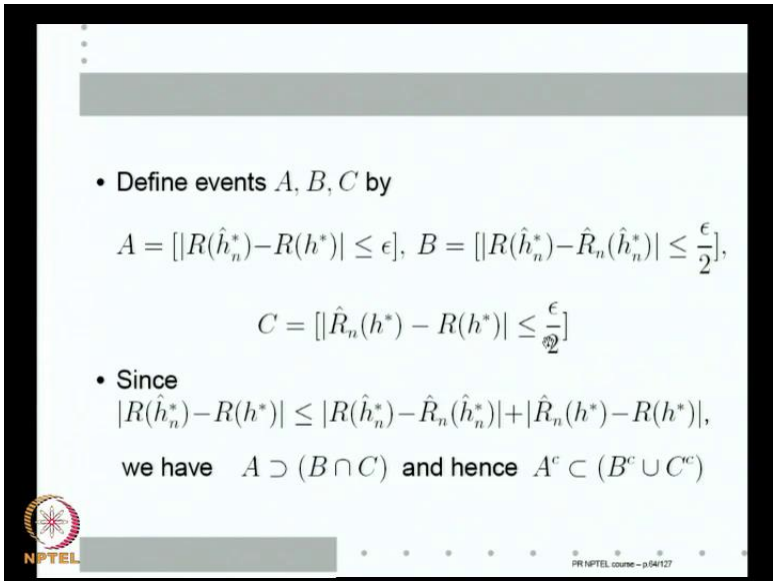
So, the sum is less than epsilon, so if both B and C happened then A is also true, what does that tell me given the events like this what it tells me is A is a super set of B n. If B and C occur then A occurs of course, A and B occurs without B and C, because it can be less than epsilon even though one of them is greater than epsilon by 2. If other is sufficiently small, so a may be bigger than B intersection C, but if both B and C happen then A will happen, so B intersection C is a subset of A. If I take complements on both sides which means A complement is a subset of B compliment union C compliment.

(Refer Slide Time: 39:16)

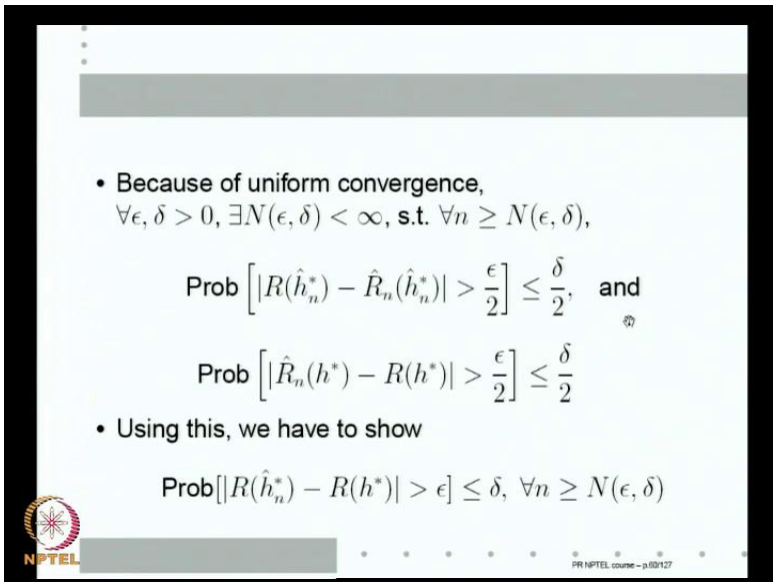


Now, as we know if A is A for two side given A and C if A is subset of B probability of A is less than or equal to probability of B. So, we know probability of A complement is less than equal to probability of B complement union C complement. And we know probability of union of 2 sets is less than or equal to sum of the probabilities, this is called the union bond. It is very useful bond in statistical learning theory, we use it again, and again, probability of union is always less than or equal to sum of the probabilities.

(Refer Slide Time: 37:45)

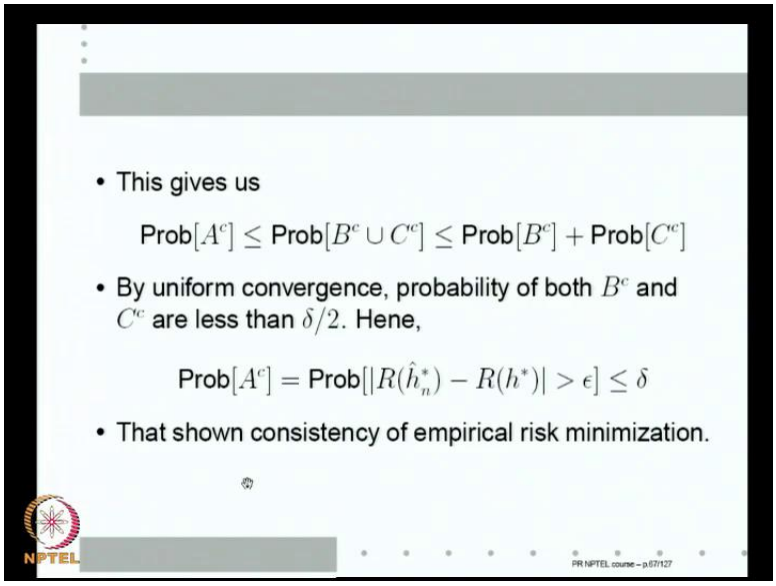


(Refer Slide Time: 39:54)

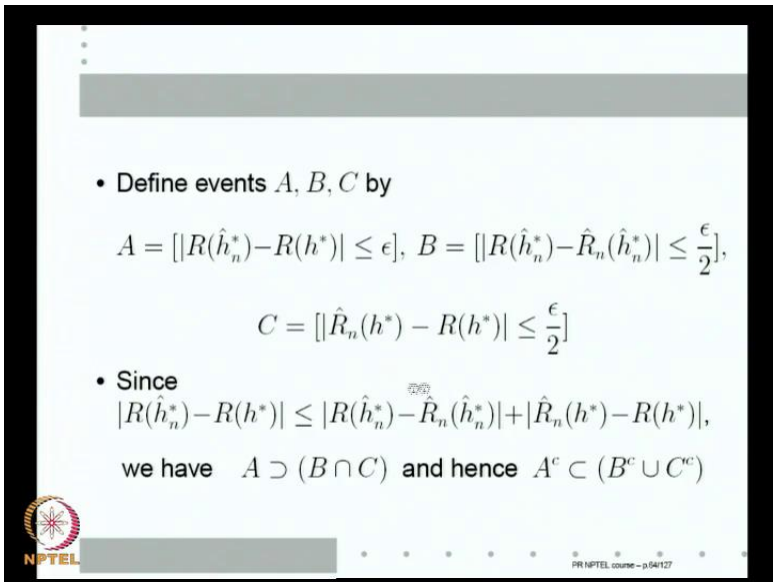


Now, we know probability B complement and probability C complement both are less than delta by 2, what is B complement? This is greater than epsilon one by 2 C complement which is greater than epsilon by 2, that I know is less than delta by 2. So, we know that both B complement C complement of probability less than delta by 2. So, A complement as probability less than delta and A complement is nothing but what we want.

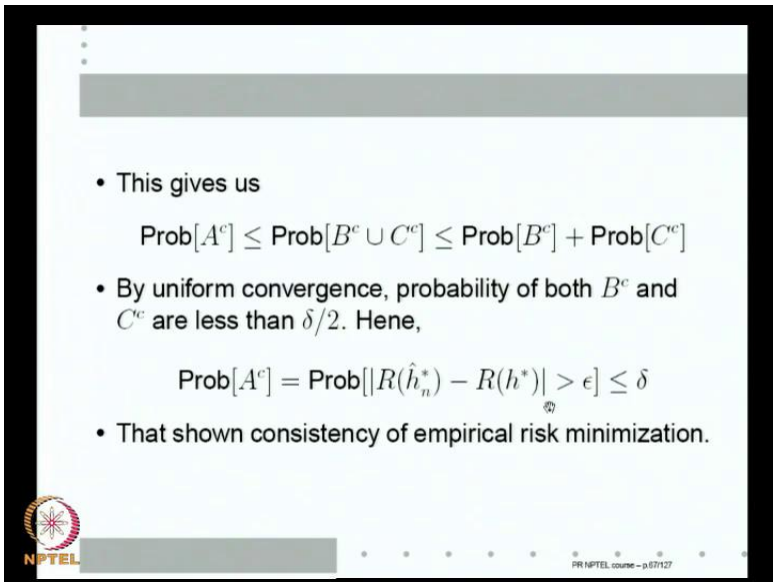
(Refer Slide Time: 40:03)



(Refer Slide Time: 37:45)

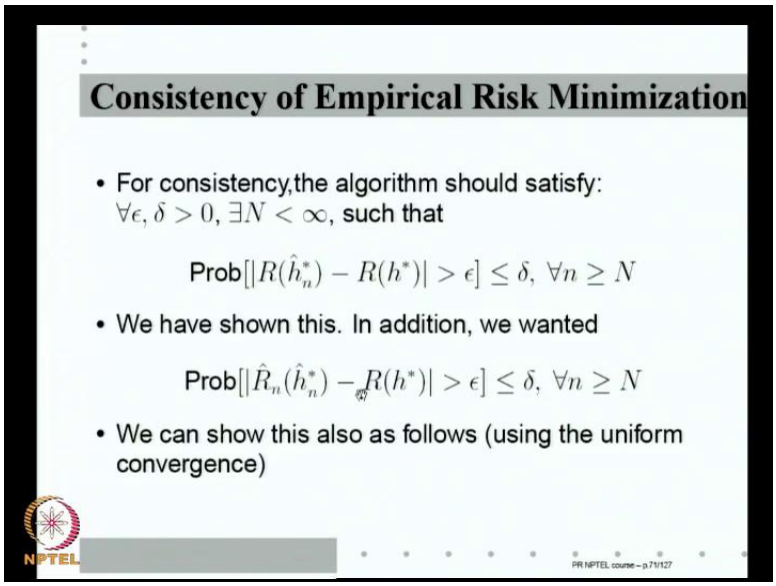


(Refer Slide Time: 40:03)



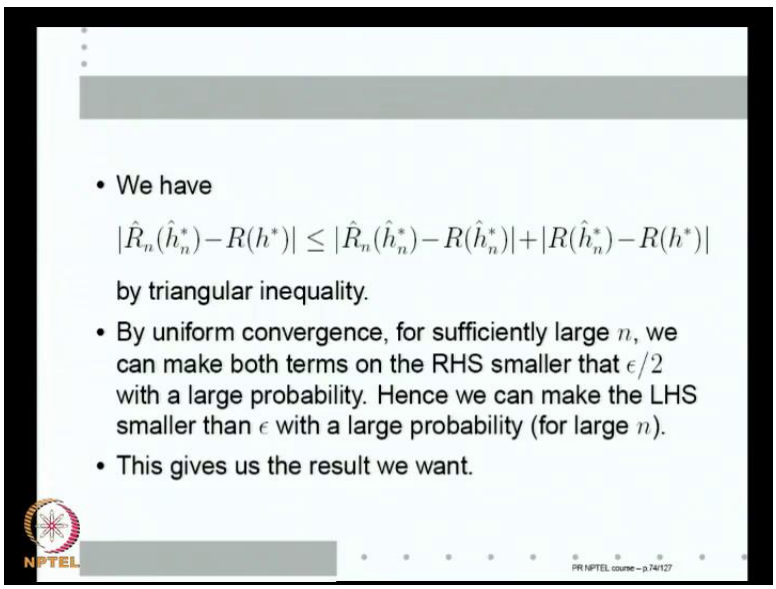
So, that is how we argued, so basically we should understand that given any inequality like this. If I know that I can make this sufficiently small with high probability and sufficiently small with high probability then I can make this also sufficiently small with high probability any such inequality implies this and this is the, we have showing this. So, that that shows the consistency of empirical risk minimization.

(Refer Slide Time: 40:41)



So, to sum up consistency the algorithm should satisfy this R of h R star n minus R of h star greater than epsilon will be less than delta for all greater than n given any epsilon delta you should be able to find N and we have shown this. We also said that we like to have this. Now, let us see whether uniform convergence will give me this also it, it is possible to show that this also follows uniform convergence, this is actually a little simple than the other one, what is it that we want to show? Probability of R hat n of h R star n minus R of h star that difference will no, will not be too much large probability, so R hat n h R star n minus R of h star I add and subtract R of h R star n.

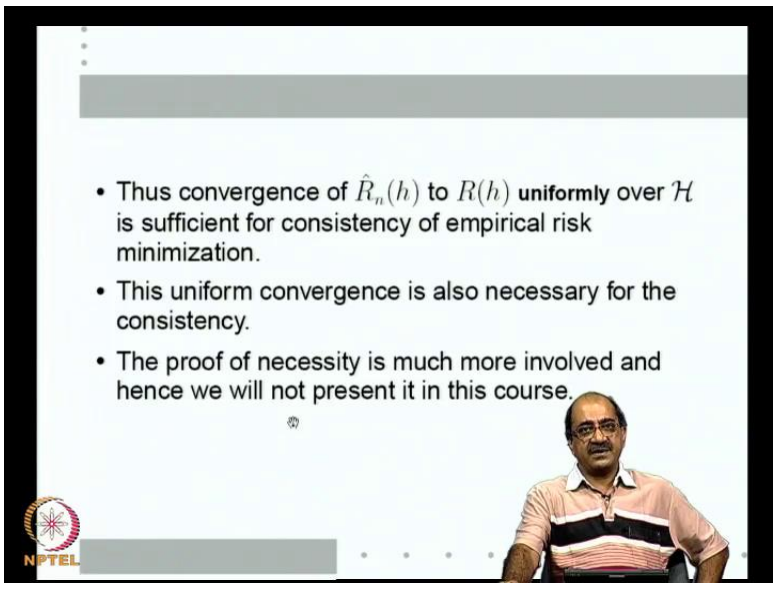
(Refer Slide Time: 40:30)



So, I can write this as R hat n of h R star n minus R of h star n plus R of h R star n minus R of h star I just add and subtract R of h R star n. And by triangle inequality absolute value of this will be S n is equal to sum of the absolute values. Now, the rest is easy this is controlled by law of large numbers, and this is controlled by what we just proved. So, by uniform convergence for sufficiently large N, we can make both terms in the R h S smaller than epsilon, but the same argument we are using we already know that we can make this as small as we want that is given any epsilon delta there is this a n. So, at this is true and this is also true by the uniform convergence and hence, because both of them can be made small with a high probability. This can be made small with a high probability that gives the result we want.

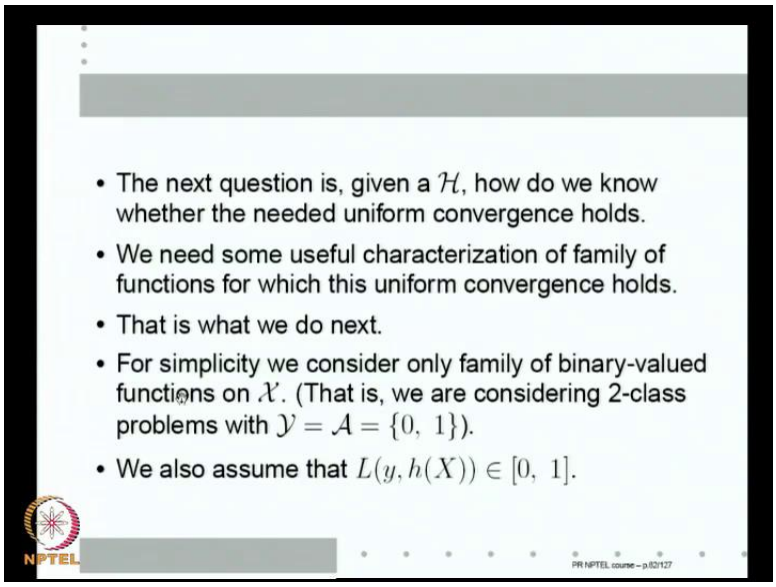
So, we get consistency as well as the, the empirical risk we find of the minimize of the empirical risk that is the global minimizer of the empirical risk where actually be closed to the global minimum of true risk. So, if we just minimise empirical risk we do find something that also minimises globe true risk that is what we want. And the actual empirical we calculate for the global minimizer that the global minimizer of the empirical risk is a good estimate of the true risk of the classifier, say everything is fine if we have uniform convergence. So, if, if the, if R hat n h converges to R h uniformly over the class of function h over then it consistency is true.

(Refer Slide Time: 43:13)



Of course, uniform convergence is also necessary for consistency. And unfortunately the, the, the proof of necessity is much more involved and hence it cannot be presented it is a little beyond the mathematics that we are pegging this course on. So, we would not prove the necessity, but we will just take that uniform convergence in both necessary and sufficient, we have seen the sufficiency proof in sufficiency proof we just take for granted.

(Refer Slide Time: 43:58)

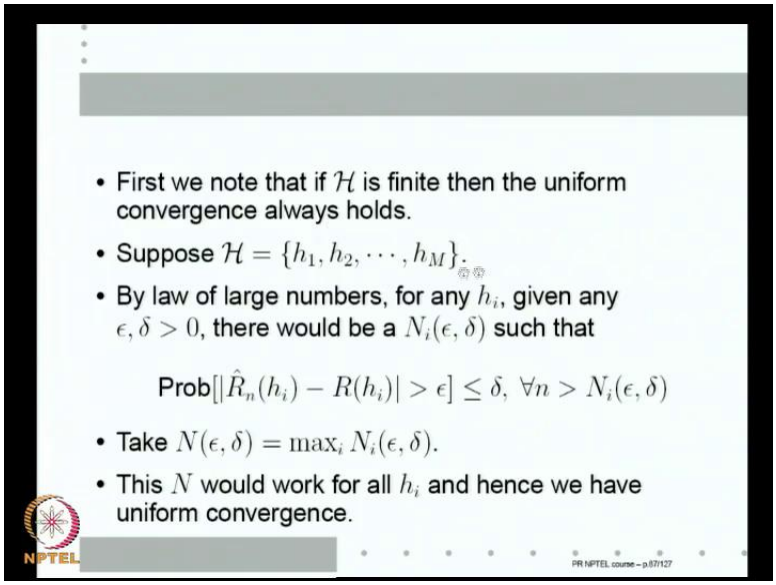


The next question is given a H how do we know whether the needed uniform convergence holds or not. So, I wanted consistency you said consistency holds if law of large numbers convergence is uniform over H. Now, how do I check given a H whether the, the convergence is uniform over H, or not I need some useful and easily calculable characterisation of family of functions for which this holds. So, given a family there must be some simple way for me to check whether or not uniform convergence holds this is what you are going to do next. For this we, we consider only family of binary valued functions on X, we do not consider any hypothesis space that is we are considering Y is equal to A is equal to 0 1.

So, we will only do this for two class classifiers and assuming that I am searching over only classifiers so not any real valued function, because during this for real valued functions is this is this will be complicate enough during it for real valued functions is

much more complicated. So, we will consider only this class and we will also assume for simplicity that we already anyway assumed that loss is bounded.

(Refer Slide Time: 45:27)



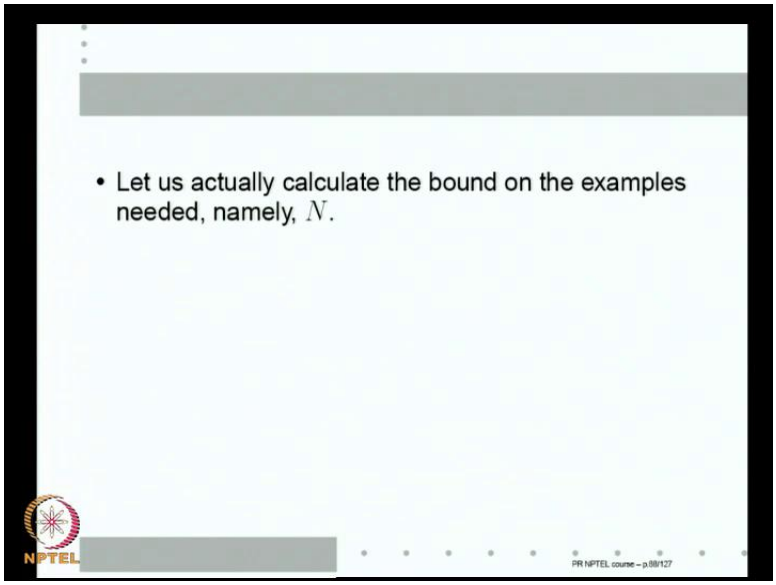
So, we assume is bounded between 0 onward and is bound on other side making it between 0 and 1 is not particularly more distinctive, but this makes the the rest of the mathematical statement little bit simpler. So, now, the question is for what kind of H is the uniform convergence holds, let us look at some simple cases. So, first we know that if H is finite then the uniform convergence always holds by suppose H is some m function h 1 h 2 h M. It is for some finite number m the law of large numbers any ways say for each h the convergence holds.

So, give me any h i and an epsilon and delta then I will have a N that is a function of h i epsilon delta I will write it as N i epsilon delta so give me h i epsilon and delta then they will be N i of epsilon delta. So, that probability R hat n hi minus R of h i greater than epsilon or less than delta if N is greater than N i epsilon delta. Now, I take n epsilon delta to be maximum over i N i epsilon delta I can always do this because I am taking maximum over finite numbers give me any finitely many numbers the maximum always exist.

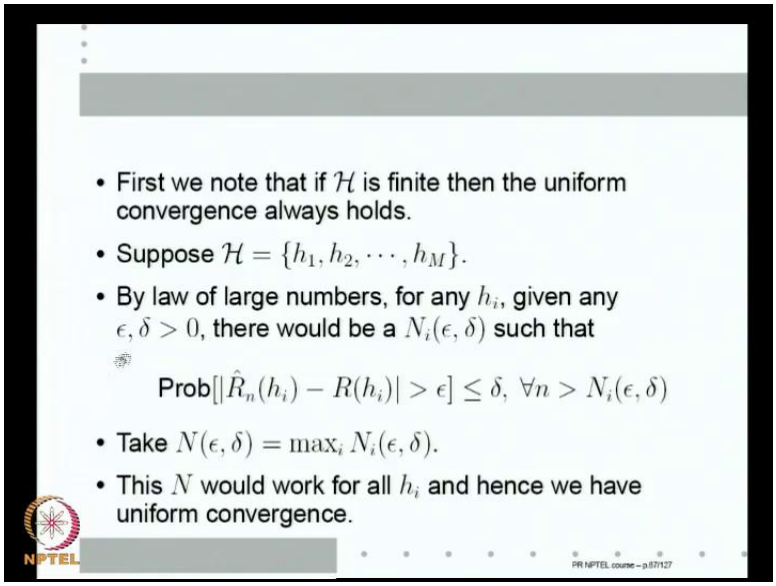
The will be finite, if we give me finitely many finite numbers each of these N i's are finite, if you give me finitely many finite numbers the maximum is always finite. But if you give me infinitely many finite numbers, it is it if you give me 1 comma 2 comma 3

so on. What is the maximum? For any finite subset of it there exist a maximum, but for the that does not exist a max, but because here I only finitely many functions the maximum exist. So, what does this mean? If you now give me enough epsilon delta examples this will hold for every h i. So, this, this N will work for all h i and hence uniform convergence holds, this is very straight forward there are only finitely many functions for each h there is an n. And if I take the maximum of all those n's for all the h S and hence for finite h the uniform convergence always holds. Is of course, is not particularly great insight, but let us actually calculate the bond on the examples needed.

(Refer Slide Time: 47:26)

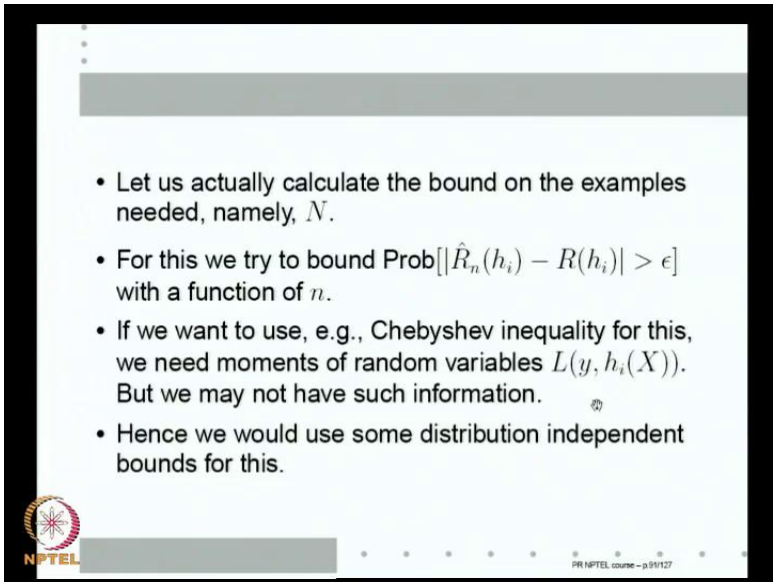


(Refer Slide Time: 47:35)



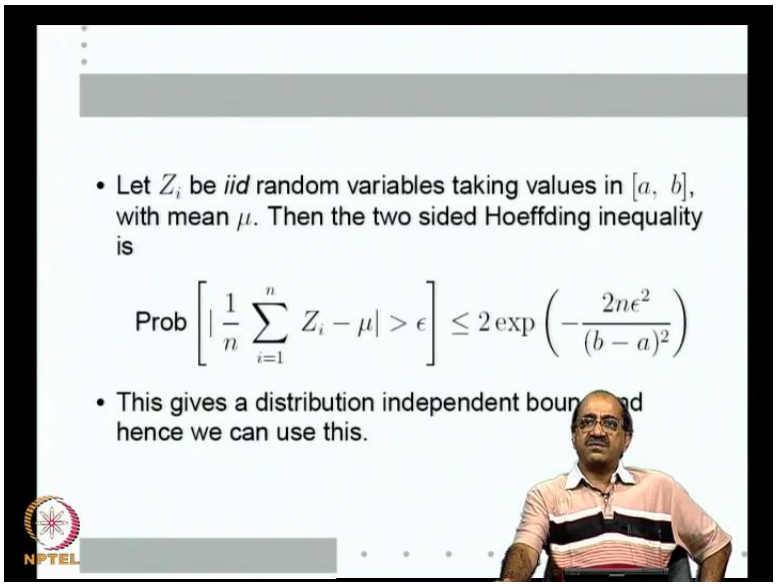
Here we have not calculate an epsilon do you saying it axis let us ay can I calculate an epsilon delta. One way I can calculate is if I can bound this probability the LHS probability by a function that depends on little r if I can find it depending on little n. Then I can say that function should be less than delta and then do some algebra to say that function has to be less than delta how large N should be. So, basically the idea is I should bond this probability the probability at left hand side of this inequality by some function of n.

(Refer Slide Time: 48:22)



Then this is a sample mean; this is the expectation I want probability of the difference being greater than epsilon then many bonds for example, I can put a Chebyshev, bond, but all such bonds. So, what is that I want to do? We want to bond this probability with a function of n. Of course, I can use many bonds, but if I use any bonds for example, if you want to use Chebyshev bond. Then you need any variance of this random variable which is like knowing moments of the random variable L Y comma h X for X Y random I got into P x y as I do not know P x y I cannot calculate those moments.

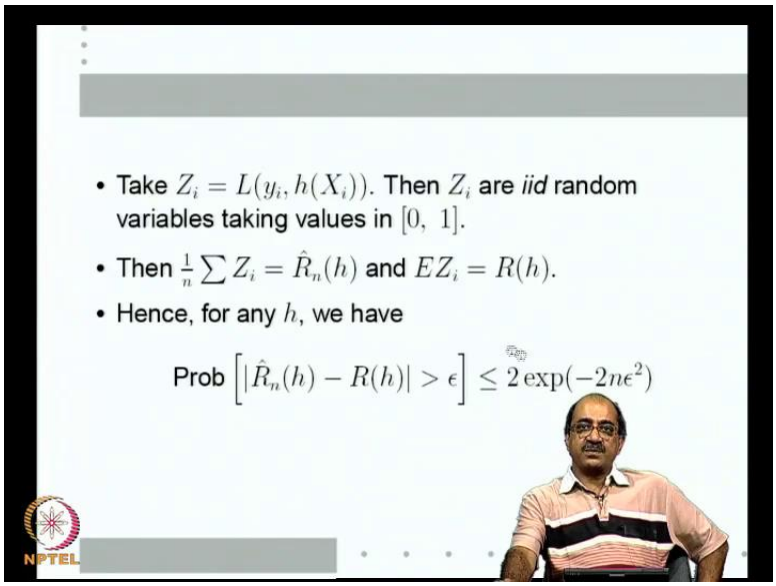
(Refer Slide Time: 49:07)



So, if when I want to bound it actually I need to find some proper in equal bonds, the standard inequality is which may need the moments of the random variables or not good enough for me. So, we essentially need some distribution independent bonds there are such distribution dependent bonds. Let us say Z i is or a Z i meaning Z 1 Z 2 some Z n or i id random variables. All of them take values in some interval a comma b and let us say there mean is mu, because there i i d all of them have same distribution so same mean.

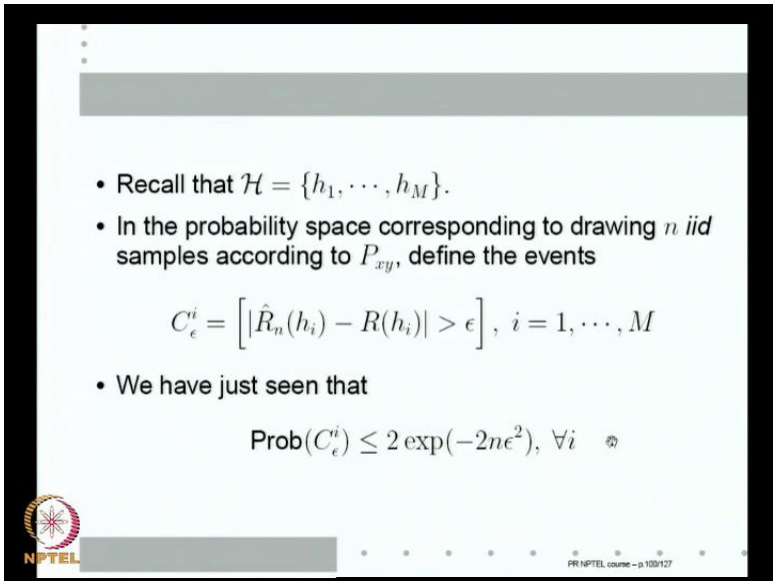
So, Z 1 Z 2 Z n are independent random variables. So, identically distributed taking values in a comma B and having mean mu then there is what is called a two sided Hoeffding inequality which says that the difference between the sample mean and the expectation being greater than epsilon is bounded above by 2 times exponential minus 2 n epsilon square by b minus a whole square. Say it only depends on n the ranger the random variables are nothing less it does not depend on any moments.

(Refer Slide Time: 50:08)



So, this is the very useful inequality to use as this is a distinguished independent bond and we can use this for our case two data bond in the finite h case. So, let us calculate the bond essentially what we are saying is Z i to be L of y i h X i then Z i R i i d random variables taking values in 0 1 is very nice for us b is 1, a is 0. So, b minus a is 1, so even that term will go away that the reason why I have taken 0 1. Then 1 by n summation Z i is same as R hat n X and expected value of Z i is nothing, but R h by definition.

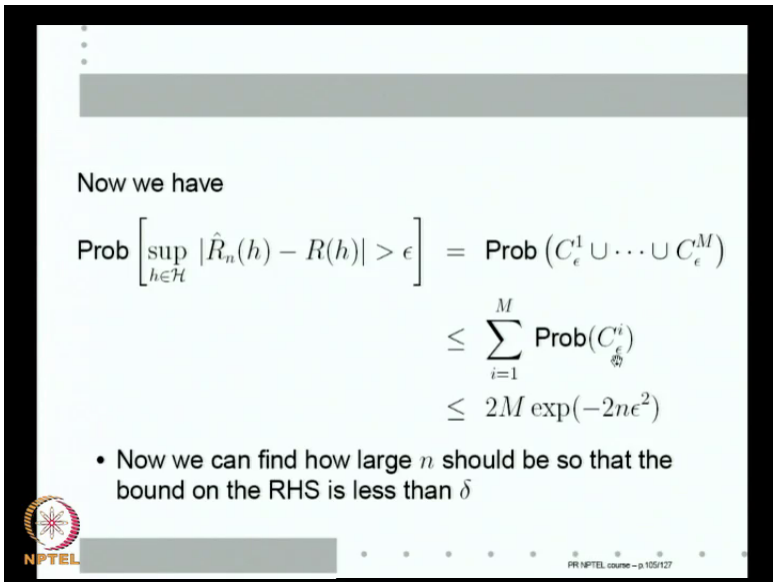
(Refer Slide Time: 51:03)



So, what Hoeffding inequality gives me is probability R hat n h minus R is greater than epsilon is less than 2 times exponential minus 2 n epsilon square. Of course, this is not the bond I want I want to put a supremum inside here I want to put supremum over h of this. So, I have to find a bond for that, so that bond can also be obtained, so this is a bond that shows me that as n tends to infinity it goes to 0. So, I can make it as small as I want for example, smaller than delta by taking n sufficiently large, but before that let us first put the supremum inside this probability.

So, H is h 1 to h M, so basically all our probabilities with respect to drawing antiple of samples. So, in the same probability space let us define an event C i epsilon over drawing antiples of this, we C i epsilon set C i epsilon or R hat n h i that is what i minus R h i. It is expectation being greater than epsilon that is the event C i epsilon, C i epsilon says that for the ith hypothesis the empirical risk, and the true risk differ by more than epsilon. Then what we have just shown is probability of the event C i epsilon is less than 2 exponential minus 2 n epsilon square, this is true for every i.

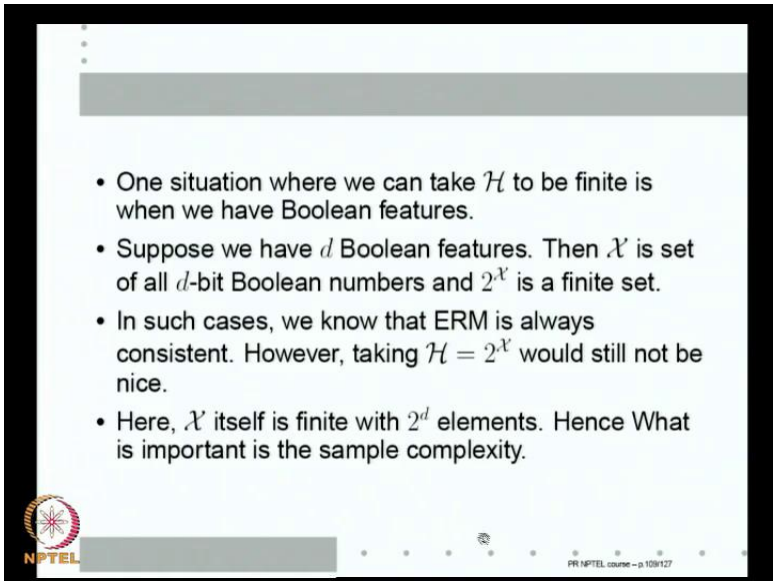
(Refer Slide Time: 51:53)



Now, if I want to put supremum here. So, I am saying for some what is the probability for some h or the other this difference is greater than epsilon that is same as probability of the union of all the C i epsilon sets. C i epsilon tells me that it is true for h i i a asking at least one of the I have for some h i or the other this is greater, but is simply given by the union of these events. Now, we have already seen union bonds probability of union is

less than or equal to sum of their probabilities and each individual probability we know what it is each individual probability 2 exponential minus 2 n epsilon square.

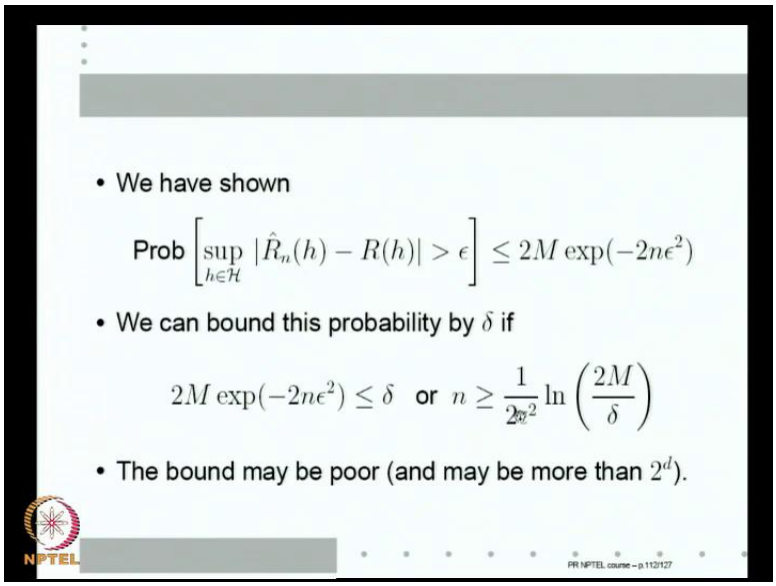
(Refer Slide Time: 52:53)



So, this is less than or equal to 2 m times exponential minus 2 epsilon square. So, this tells me that this can be made as delta. So, we can now find n how large is n should be so that this can be made less than delta. So, this is how I can actually calculate n if you give me epsilon and delta. Before we go forward where can where is finite h i useful thing if you have Boolean features, if you have Boolean features. Then the X itself is set of d-bit Boolean numbers S itself is finite and hence 2 power X is also finite set.

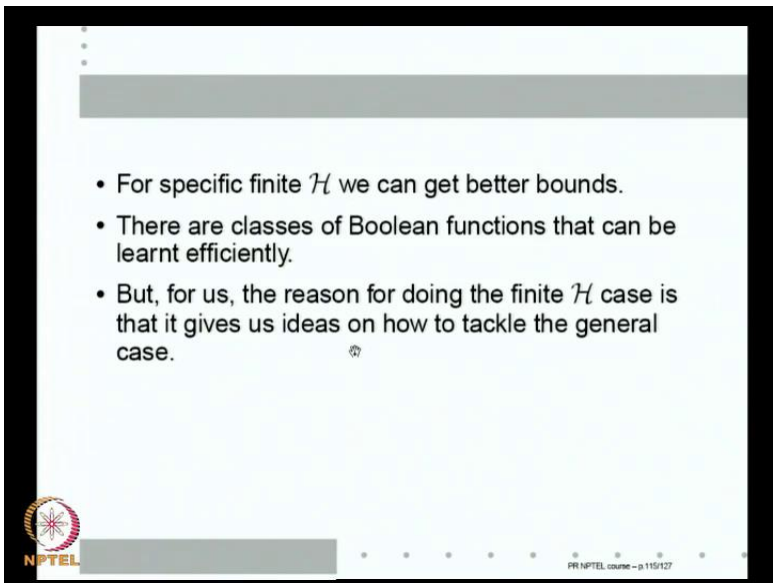
If 2 power X is finite set every subset of 2 power X is also finite set. So, if I am learning Boolean features two class classifiers. Then obviously even multiclass classifier for that matter; obviously, every h I can take will be finite and for all finite h uniform convergence holds, because it does not mean even in finite case taking 2 power X should be a a nice choice. As we already seen two complete over all possible classifiers never learns well that we can still see why basically, because X itself is finite saying given sufficiently many examples I can learn means nothing. Because there are only 2 power X examples if you see all the 2 power X examples then what more do you want to see there is nothing more to see nothing more to predict.

(Refer Slide Time: 54:01)



So, what is important is given an epsilon delta how many samples to be made, it is this my M is a total number of elements. If X has 2 power d elements and I am considering all boundary valued functions are on X M will be 2 power 2 power d. So, this may not be a very nice bond for example, if I want is less than delta and is greater than 1 by 2 epsilon square l n 2 m by delta.

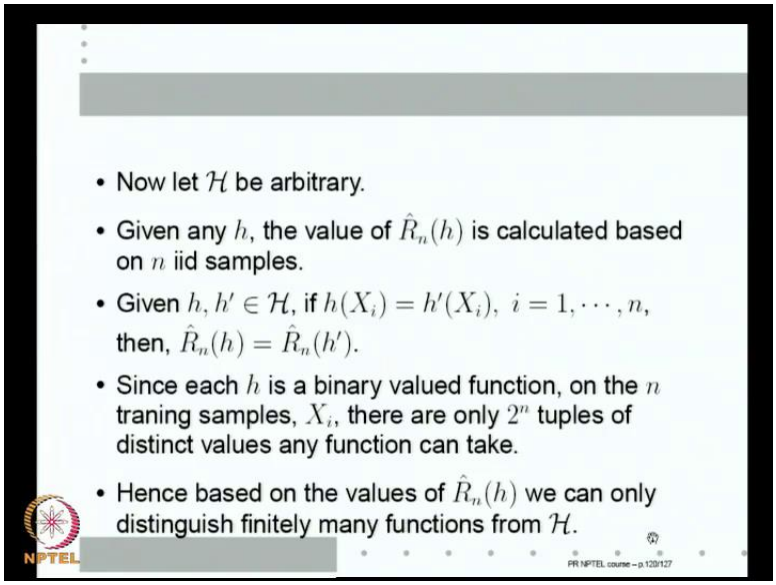
(Refer Slide Time: 54:50)



Now, M is 2 power 2 power d, so this might be this bond might be even greater than 2 power d. So, just because we found a bond it does not mean that it is a very nice way to

learn I mean even though it is consistent the number of examples I am asking means show me every single example then I can predict all of them.

(Refer Slide Time: 55:16)



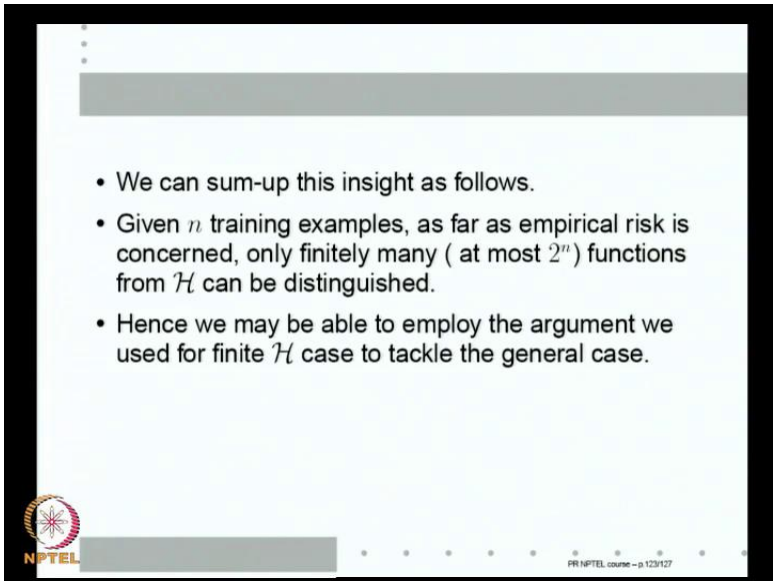
So, infinite sets actually we want better bonds we are much more interested in how fast is number of examples grows with epsilon and delta. There are class of Boolean function be learnt efficient, we can show that we do not need exponentially many examples, but we can do with polynomially many examples, But that is not what we are going now I might come back to that. Later the reason why we actually did the finite case in detail if that it gives us ideas on how to tackle the general case the idea is the as follows.

Now, consider h is arbitrary it may have infinite uncountable infinite items, given any h the empirical risk is calculated based on n i i d samples. What is that mean? If I have two functions h and h prime in my hypothesis space there such that h of X i is equal to h prime of X i on all the n training examples. Then the empirical risk of h on h prime is same the h n h prime may take many vastly different values on other axis. But let us say they just agree on the training samples I have with they agree on training samples I have then there empirical risk are same.

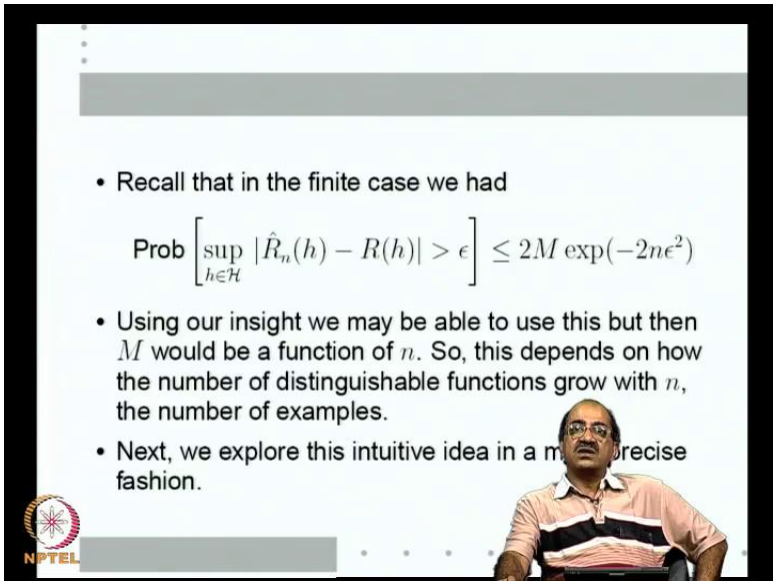
So I can only distinguish between function while minimising empirical risk only those functions use values differ on the training samples each h is bounded valued function. And I have n training samples given n training samples, how many different values can binding valued functions taken n type training samples they are only 2 n some different

possible values that any function can take on which means at most I can distinguish through an possible elements 2 in h. Even if h contain uncountable infinite things given any n training examples I cannot distinguish between more than 2 power n different h's hence based on the values of R hat n we can only distinguish between finitely many functions from h.

(Refer Slide Time: 56:40)



(Refer Slide Time: 57:03)



So, what is the insight we got? Given N training examples as far as empirical risk is concerned only finitely many at most 2 power n functions form h can be distinguished

which means we may be able to employ the same argument that we have used in this finite h case to tackle the general case that is the reason why looked at it. So, how do I apply the finite h case, in the finite h case I have this where M is the total number of functions unfortunately. If I put M is equal to 2 power n, this inequality is useless for me, because this goes exponential with n this falls exponential with n what do I get.

So, using our insight may be able to bond this, but then M would be a function of n. So, this whether or not I can bond this properly depends on how the number of distinguishable functions grow with n the number of examples. We know the given n examples and given a family of functions h they can at most be 2 power n function that from h that can be distinguished based on the example, but they may be less.

So, for a particular h the question is how this number of functions grows with n. This is the insight as a matter of fact, we will explore this intuitive idea in more precise fashion next class. But essentially looking at this idea and knowing that this M may become a function of n, we want to know, what is the kind of growth function I should put here? And then I may be able to characterise the h for which rather the distinguishable functions grows with 2 power n or not this is what we do next.

Thank you.