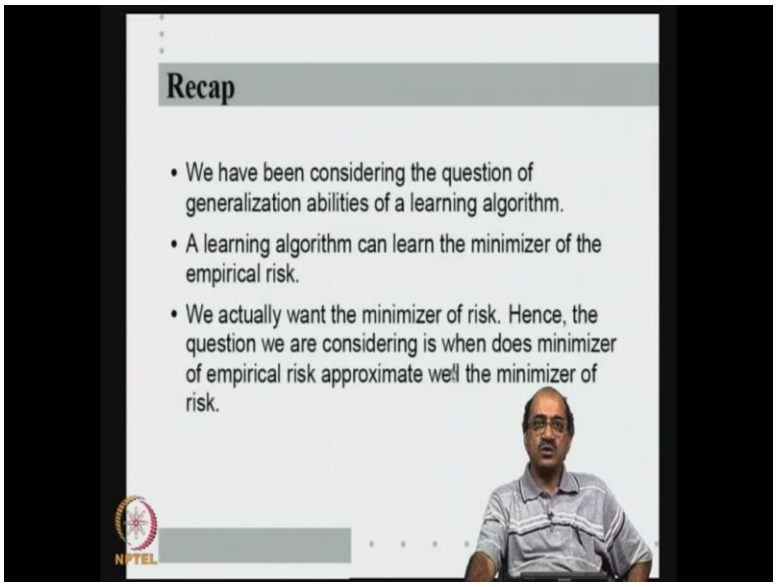
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 24
Complexity of Learning Problems and VC Dimension

Hello and welcome to this next lecture in the course and pattern technician. We have been discussing issues of statistical learning theory. Specifically, we were looking at this so called (( )). We have been looking at the problem of, when is empirical risk minimization consistent. So, we looked at the empirical risk, the discrimination frame work. We have concerned the actual formulism and then, we were discussing how can one put some bound on the generalization errors that learning algorithms make or which is same as, see if the algorithm minimizes empirical risk, is it good enough for us can we have confidence that what we learnt is good in the sense that it is close to the global minimized of true risk. Right.

(Refer Slide Time: 01:21)

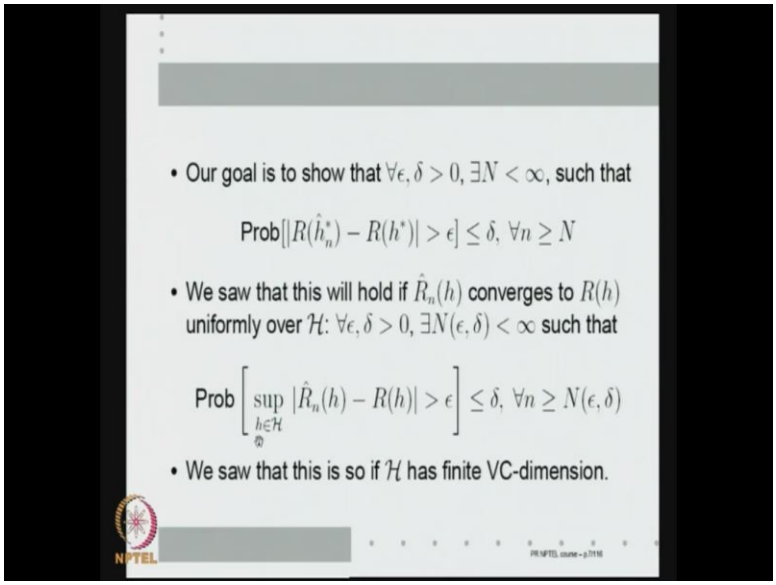


So, let us we have been considering the question of how to address the generalization abilities of a learning algorithm? How does one quantify it? So, we have discussed the formulism of risk minimization. So, essentially we are given an input space X to the feature space and we are looking for example, in simplest case 2 class classifiers on X that is, 0 1 valued function defined on X. And the way we evaluate any learn function is through loss through a loss function. And we have defined the risk at the expectation of

loss with respect to the same probability distribution that the examples are given to us. And any learning algorithm that we saw will essentially minimize the empirical risk. But, what we actually want is to minimize the true risk. The true risk of any function h is accept a value of all of h x comma y where, expectation with respect to the joint distribution of x, y; the same distribution with respect to which that are examples. We do not know the distribution hence; we cannot minimize the true risk.

Since, we have IID examples; we can approximate the true risk by its sample mean estimator which is called the empirical risk. And empirical risk is what most learning algorithm can minimize. And hence the question is, when does the minimizing of empirical risk well approximate the minimize of true risk. That is what we have been discussing. And when we say approximate, we do not want the functions to be closes functions the classifier is to be closes functions and x. But, there true risk should be close because a classifier for us is evaluated only in terms of its risk.

(Refer Slide Time: 03:18)



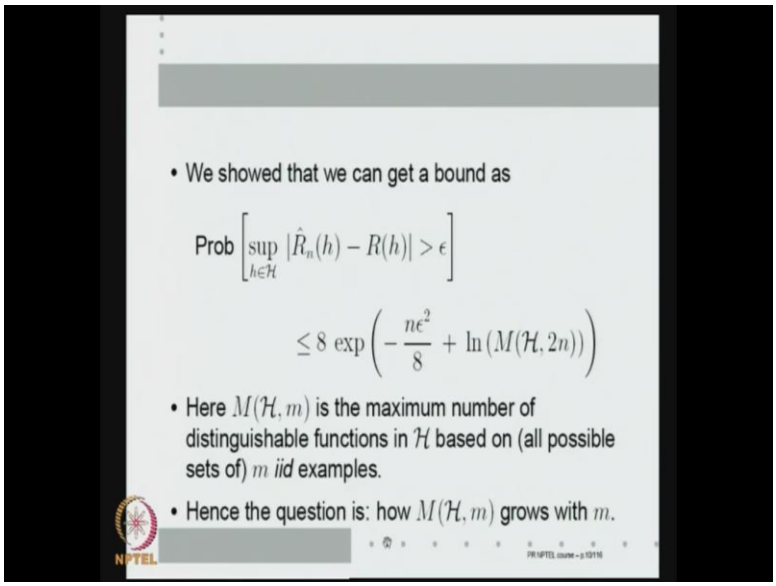
So, what we actually want is if h star is a global minimizer of R then, the R h star is the global minimum of true risk. And h R star n is the minimize of the empirical risk. Given n examples this is the minimizer of the empirical risk, the global minimizer of the empirical risk. So, we are asking is if the risk of what I can learn, what I can learn is the minimize of the empirical risk. So, what you are asking is the risk of what I can learn, is it close to the global minimum of risk? So, what I learn as a classifier is close to the best.

So, what we want to achieve is that given any epsilon or delta. We can put a bound on number of examples we need. So, if you see in that many examples that many iid examples and minimize the empirical risk then, the true risk of the minimize of the empirical risk is close to the global minimum achievable on true risk. This is what we want to show. And we have been for the last 2 classes; we looked at the analysis of this. And we saw that this is possible if this R hat n h; the empirical risk of any function any classifier h converges to R h, the true risk of the classifiers uniformly over h which is the family of classifier. So, which were searching for the minimize of the empirical risk.

As you seen such a uniform conversion is sufficient to ensure this and we also stated that it is enhancer to ensure this. So, ultimately the question. See we know that R hat n h which is the sample mean estimate of R h. And hence, R hat n h converges to R h has intense to infinity that is the law of class numbers. But, as you saw that is not enough. What we actually need is that, this conversions given by the law of large numbers that R hat n h the samples mean estimated converges to R h the expected value, uniformly over h. And we also the uniform conversions means the following: That given any epsilon delta, there is 1 number in epsilon delta and if I have that many examples then, for every h in a family, R hat n h is close to R h.

Normally, the law of conversion only ensures that given any specific h, I can tell you how many examples I need so that R hat n h is close to R h. The uniform conversion means, the same number of examples would work for every single h in this family h. right. So, probability supremium of the difference R hat n h minus R h. So, the supremium is taken or the family of classifier h yes script h. This supremium difference greater than epsilon, should be less than delta. This is what uniform conversions means. We also saw that this is the uniform conversions holds, if the class of classifiers has a finite VC-dimension.

(Refer Slide Time: 06:33)



Now, we actually showed that, to recall what we showed is that the probability of supreme difference between R hat n h minus R h greater than epsilon where, the supreme is taken or the family of classifiers can be bounded above by this. Right. And where, this M H 2n is the maximum number of distinguishable functions in H based on all possible samples sets.

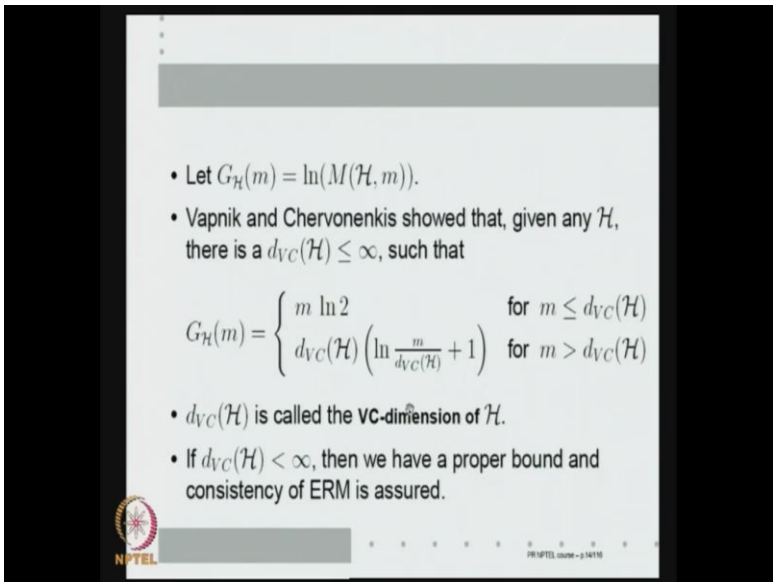
What we actual have for a R hat n h? Here, we actually have M H comma 2 n. But, for any integer M, M H comma m is the maximum number of distinguishable functions in h based on all possible m iid example. Given, m iid examples; any 2 functions that take the same values on those examples are distinguishable as far as the sample is concerned. So, considering that way we are asking if I choose every possible set of m iid examples, every possible set of m examples because we are choosing every possible set of iid makes no difference. So, for every possible set of iid examples, every possible set of m examples; among all if you if you concern all possible sets of m examples, we are asking what is the maximum number of distinguishable functions in h.

So, there will be some set of m examples on which this maximum will be reached. So, we are asking, if I look at every possible m set of examples; what is the maximum possible number of functions within h that I can distinguish? Looking at this bound; essentially, we want this bound to be made as small as we want as n terms to infinity. This term is decreasing linearly in n. So, the question is how does this grow with n?

right. if with ahh if M H comma m grows exponentially with m then, the logarithm will grow linearly with n here. Then, obviously inside a exponent there is one term that decreases linearly with n, another term increases linearly with n. So, we can do nothing about whether it will go to 0.

On the other hand, if the algorithm of M H comma 2n does not grows less than linearly or it does not linearly or grows less than linearly with n then, with sufficiently large n we can make this as small as we want and hence, we get the uniform conversion. So, ultimately whether or not uniform conversions holds depends on this particular quantity, how it grows with m.

(Refer Slide Time: 09:25)



As we stated in the last class we can bound this quantity. Basically, in this we want to bound l n H comma 2n. So, if we call that quantity some G H of m then, one can show or Vapnik Chervonenkis has actually shown in a very very important Vapkin machine learning; has shown that, this function G H m has only 2 possible growth patterns. There will be some integer which we call d V C H standing for VC-dimensional of H, which could be for example infinity. As long as m is less than or equals to d V C of H this grows as m l n 2.

That means, capital M that H comma m, what is inside this algorithm that goes as to for m right exponentially. But, after this d V C H is reached, when m goes beyond that l n of capital M grows only. Which means, that capital M H comma m at that the maximum

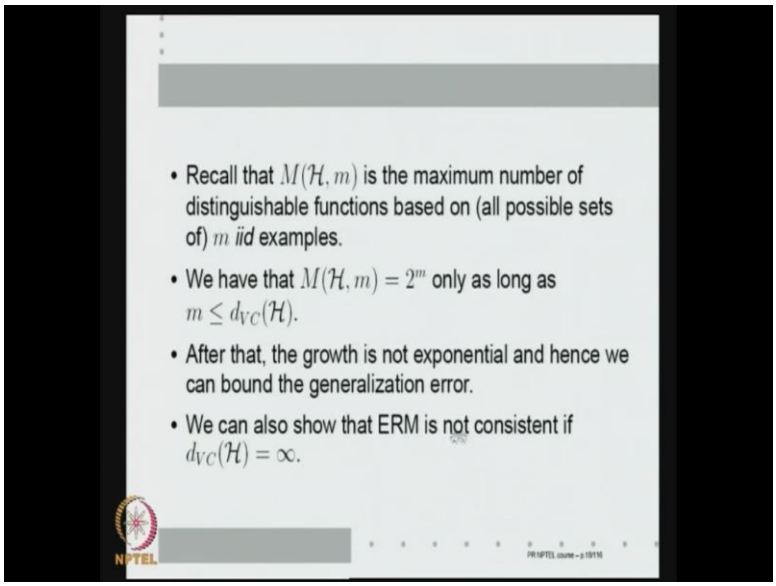
number of distinguishable function that grows that does not grows exponentially; that grows less than exponentially. It essentially grows as So, this is l n of m that grows essentially as l n of m right with some factor here. So, it grows as l n of m.

What it means is that here, the dependence with respect to n will come as l n of 2 n. So, this can grow only larger logarithmically in n whereas, this term decreases linearly in n. So, ultimately this dominates. So, as n increases, what is inside the exponent? Increases and hence, the right hand side can be made as small as you want. Right.

So, because this G H m has this 2 possible growth patterns; what it means is that, if this number d V C of H which is character characterizing number for the family of classifiers H, if that number is less than infinity then, there is some finite m beyond which growth of G H m is only logarithmic and hence, uniform conversions holds. Right. This d V C of H is called VC-dimensional of H. And hence, if VC-dimensional of H is less than infinity then, we have a proper bound. Right. If VC-dimensional of H is less than infinity then, after some m, l n of capital M is only grows logarithmically. Which means, as I just told you; the second term this exponent grows only as l n of m, l n of n whereas, this decreases as n linearly in n. So, as n goes to infinity, this exponent goes to 0.

So, if the VC-dimensional of H if the VC-dimensional of h is less than infinity then, we have a proper bound and consistency of ERM is assured. So, we started the question of consistency of empirical risk. We said that if there is uniform conversions of empirical averages to that true expectations uniformly over h then, ERM is consistence. And this inturn, this uniform conversions inturn holds, if the VC-dimensional of H is less than infinity.

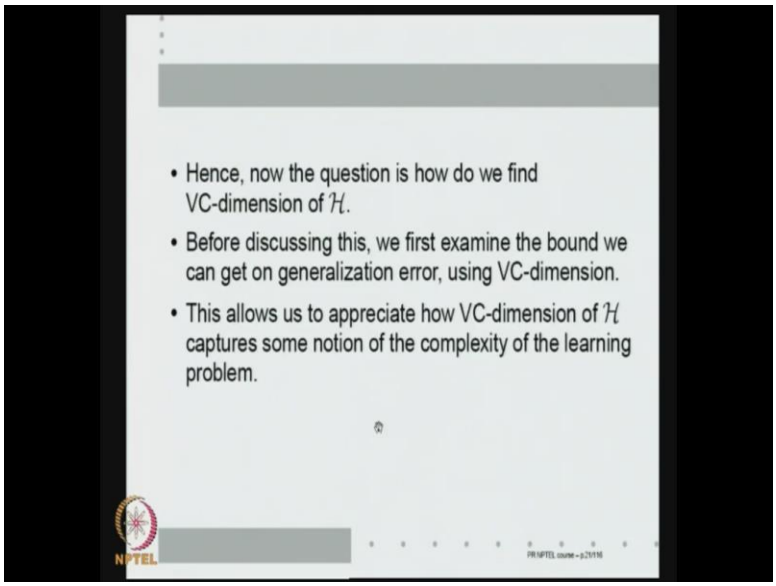
(Refer Slide Time: 13:22)



Now, to get good appreciation of what we mean by this VC-dimensional. M H comma capital M of h comma small m, is the maximum number of distinguishable functions based on all possible m sets of examples, all possible examples sets of size m. What the previous result that we just now saw says is that, this number is equal to 2 power m, only as long as m is less than the d VC-dimensional of H. Once m is more than VC-dimensional of H this grows less than exponentially and hence, we can bound with generalization error. That is how the VC-dimensional works.

So, till VC-dimension, the maximum number of possible number of distinguish function will grow exponentially but, after that the growth is less than exponential and hence, that is how a generalization bound works. Of course, this only shows that if VC-dimensional is less than infinity, we have a mis-consistency because this only a bound, if it is equal to infinity we do not know we could bound or some other way but, it can be shown that if VC-dimensional is infinity then, we cannot bound the generalization error.

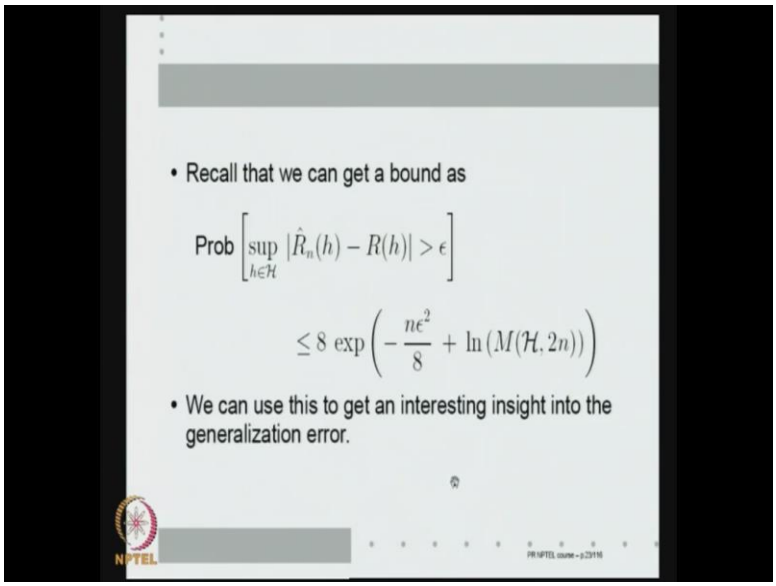
(Refer Slide Time: 14:40)



Now, the question is how do we find the VC-dimensional of H. So, ultimately given a family of classifiers of which you want to set let us say (( )) classifiers, let us say access parallel rectangles whatever, particular class of classifiers or which were minimizing empirical risk. Given that H, first thing we want know is the VC-dimension is less than infinity.

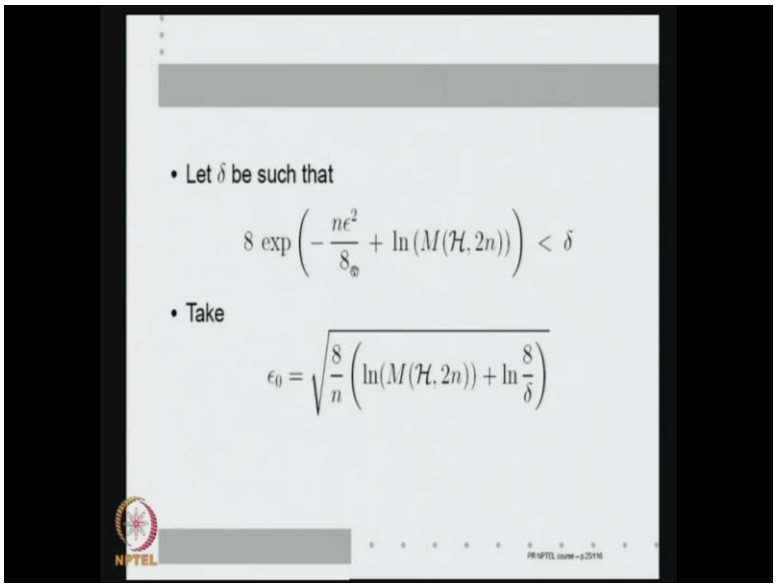
So, one question we would like to know is given a H, how does one find VC-dimensional? Will see that; but, before we go there we first examine given that there is a finite VC-dimension exactly, what will be the generalization error bound or rather the uniform conversion bound using which we derive some any a very interesting bound on the risk of any classifier which is also useful in its own guide. So, what we going to first see is how we can actually bound the risk of any classifier using VC-dimension and that allows us to appreciate, how at the complexity of the family of classifier grows, we need more examples. So, essentially this bound allows us to appreciate that VC-dimension is a measure of complexity of the learning problem.

(Refer Slide Time: 15:58)



So let us once again get back to the bounds for the uniform conversions. This is the bound we have. Right. Now, basically we want to show that if n is large enough this can be made less than any delta we want.

(Refer Slide Time: 16:15)



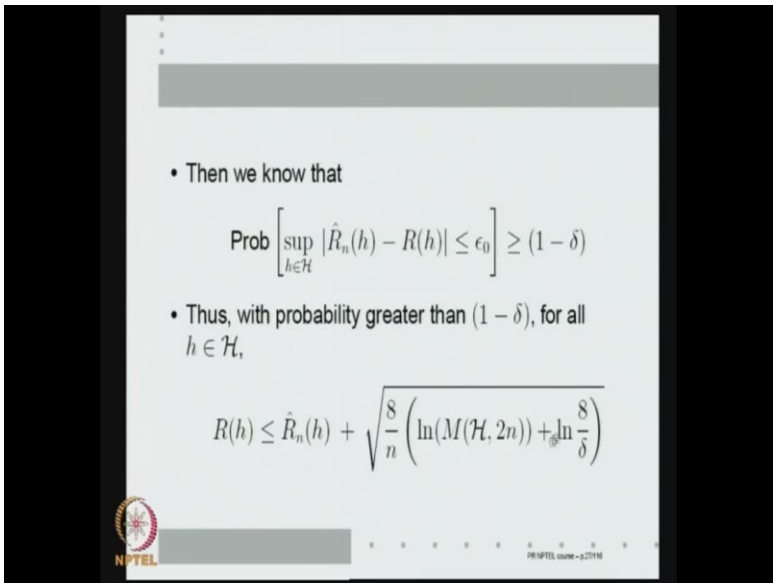
So, let us say, given a delta we want to make that less than delta. Right. This is our out bound. So, we want to make this less than delta. So, given a particular delta we want to make this less than delta. Of course, using this equation for a given epsilon we can say how large n should be so it is less than delta. Or for a given n, very often does what we

have we know how many examples we have; I can say what is the what is the vast case epsilon that I get. So, I can solve this equation for epsilon that you tell me if I have n examples and I want to make some statement with probability greater than 1 minus delta, what is the epsilon. Right.

So, this comes from. So, if this is less than delta; what it means is that with probability greater than 1 minus delta, I can say that R hat n H is epsilon close to R h. the difference between them is at most epsilon.

So, using this, I can now compute a given an n I can compute an epsilon so that I can make the statement. So, let us invert this for solving for epsilon. So, I can take 8 on this error becomes delta by 8, take logs on both sides then, I get minus n epsilon square by 8 plus l n of this is less than l n of delta by 8, bring that this side and n of square that side minus l n delta by 8 is l n 8 by delta. So, I get l n 8 by delta plus logarithmic term that should be less than n epsilon square by 8. So, the worst case epsilon would be 8 by n into this square root of. So, if I am given n examples with probability 1 minus delta, R hat n H is epsilon close to R h. It cannot be more than epsilon away from R h. That statement I can make with probability 1 minus delta.

(Refer Slide Time: 18:12)

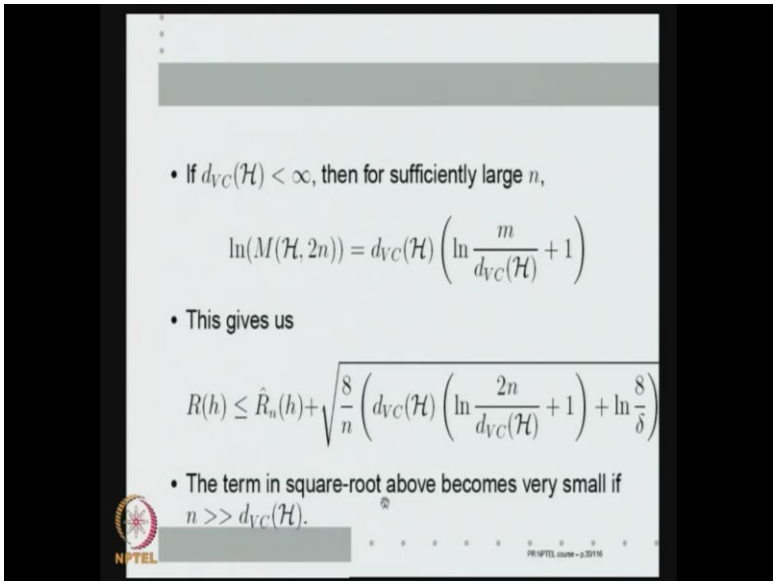


So, let us make that statement and see what it means. So, the way we calculated this epsilon naught, so we can say that given that delta with probability atleast 1 minus delta, I know that for every h in my family, R hat n H is at most epsilon away from R h. Which

means, with probability greater than 1 minus delta; for every single h, R hat n H, R of h the true risk of h is at most R hat n H plus this epsilon. I know R hat n H is epsilon plus 2 R h. So, what is the worst case stories I can have? For every h, the true risk of h is bounded above by the empirical risk I actually calculated using examples plus this term.

Because I can i can say with probability 1 minus delta atleast 1 minus delta for every h, R hat n H is epsilon naught close to R h which means, R h can be at most R hat n H plus epsilon. It cannot be more than that. So, R h is less than or equal to R hat n H plus epsilon naught and this is epsilon naught. Now, let us assume that we have finite VC-dimension and hence, substitute for this logarithmic term here in terms of VC-dimensional.

(Refer Slide Time: 19:27)



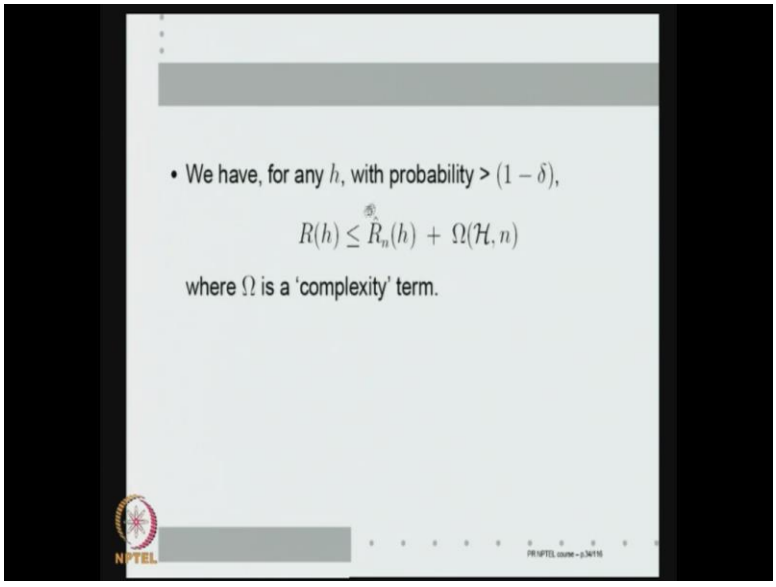
So, when VC-dimension is finite and n is sufficiently large. Right. This is this. I am sorry this term should be 2 n here. I am sorry about that. So, this is the bound that we saw earlier in that Vapnik Chervonenkis theorem. So, let us substitute that. So, I can say R hat R h the true risk of any classifier, h is bounded above by its empirical risk which I can calculate from my samples plus this term.

In this term, if I take this n inside the brackets it becomes d V C H by n. So, d V C H by n decreases with n, this increases with n. But, this increases only logarithmically in n by d V C H, and this decreases linearly in terms of d V C by H n. So, as n increases because logarithm grows slowly, this growth is more than compensated by decrease of this term.

So, the term in the square root can be made very small if n is sufficiently larger than d V C of H. So, if n is much larger than d V C of H, the d V C H by n term is more compensates for this term. In particular as n terms to infinity, this term goes to 0. So, R hat n H will becomes same as R h. That is what uniform conversances. But, given n how small this term is depends on how large is d V C by H. What does that mean? Firstly, before we can ask what that means, we can make many conclusions.

But, before we get carried away let us understand that the bound is very loose. The bound we obtain is very loose. If we remember, we started first with a half ding bound which is a distribution independent which has to hold for every possible kind of distribution and hence, it can be very tight and any specific distribution. Then, we used a union bound. Union bound means, we have many sets: a 1, a 2, a n probability of union, we bound a (( )) some of the individual probabilities. That is that is also a loose bound very often and top of that we needed to find the expected value of the number of distinguishable functions because we cannot find they expect a value we bounded that by the maximum number of the distinguishable functions. So, every step of the way we through in very loose bounds so, the final bound we get must be very loose. Right. So, numerically the bound may not be very useful. But, what is interesting is the form of the bound. The form of the bound we can put on R h gives us a very interesting ideas.

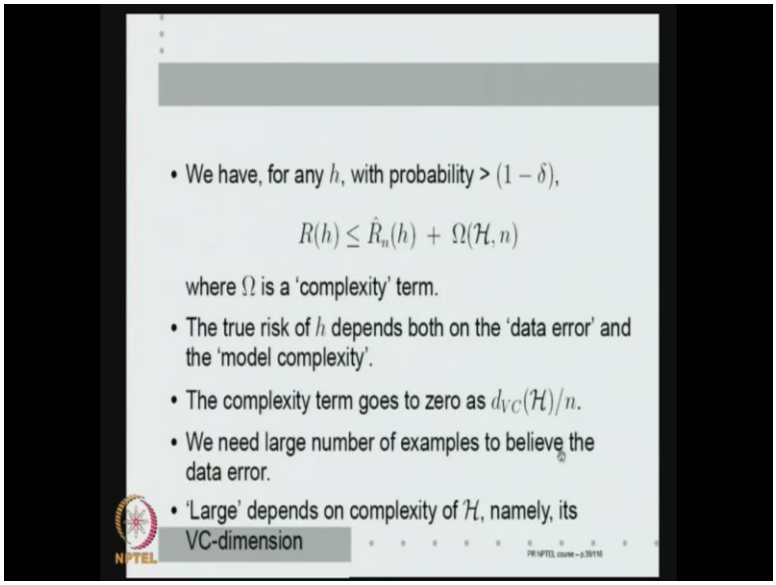
(Refer Slide Time: 22:30)



We can write the bound as R h is less than or equal to R hat n H plus some term, we call it omega H comma n.

What distinguishes this omega H comma n is that, this term as you can see does not depends on my sample. See, R hat n H depends on the sample, this actual sample on which what is the error or what is the average loose that my h makes. But, this second term does not depend on sample, it only depends on the classifier class of family over which I am searching for minimizer empirical risk and the number of examples I have. So, something that characterizers the complexity of the class of classifiers from which I am learning. Right. So, I can write this as R h is less than R hat n H plus some term that depends on n, on the family of classifiers but, not on the specific samples set that I have.

(Refer Slide Time: 23:56)



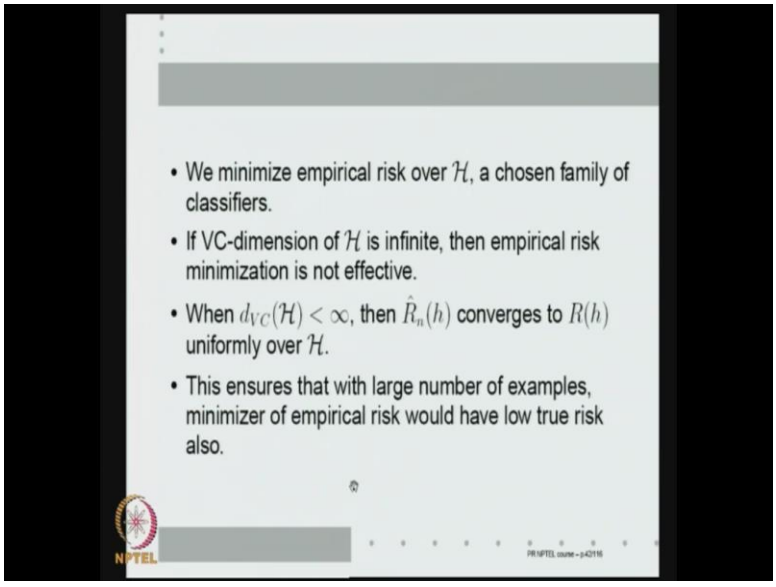
So, I can write it like this; where, omega is a complexity term. So, this is the data error as we already seen when we looked at regular as least squares. This is the data error, this is the actual error where or loose that h suffers on the data. This tells me lack of it between h of x i, x i and y i average roar, all x i y i which are my samples. This is a complexity term. Right. It only depends on the class of classifier on which I am searching and the number of examples I have.

So, true risk h depends both on the data error right. So, even if my data error is very small it does not mean that my true risk is very small because this term can be large right. So, true risk depends both on the data error on the data complex. This depends on

how complex it is the model is. So, of course I know that this term goes to 0 as term of infinity but, how large n should be for this term to be sufficiently small depends on what is the complexity of h. right. So, the complexity term goes to 0 as d V C H divided by m. right. So, if n is much larger than the VC-dimensional of H that is, when the constant goes to 0; Right means, we need large number of examples to believe the data error. When this term is sufficiently small then, small error on the examples will also mean that small true error. So, it is real generalization error which is the true risk will also be small if its empirical risk is small.

So, the data error is a good indicator of what it is generalization error would be, if this term is small and this term will be small if n is sufficiently large. How large is large? How large is large is depends on the VC-dimensional. So, n should be large relatively dimension of h. So, it is in this sense that VC-dimension tells us about the complexity of the class of classifier from which I am learning. So, how many examples I need? I need examples which are much larger then VC-dimensional of H. So, for different H, how large n should before I can believe my data error depends on how complex H. H is complex when H is very high VC-dimension. Then, the same number of examples will not give me the same amount of confidence that my data error is the true refraction of the true error of the classifiers.

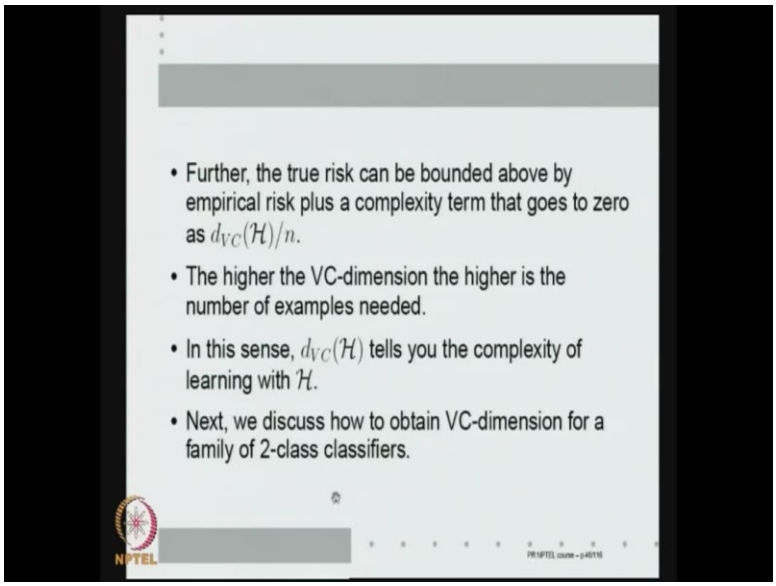
(Refer Slide Time: 25:51)



So, let us sum this up. We minimize empirical risk over H, a chosen family of classifiers. We chose a family of classifiers H and then, on that we minimizing empirical risk that is my essentially minimizing the data error. If the VC-dimensional of H is infinite of course all (( )) are of, we can say nothing about whether the minimize of the empirical risk will get us anywhere. Minimize of empirical risk this is not effective if VC-dimensional which is infinite.

So, that is one thing that we have to any way ensure. So, unless VC-dimensional which is finite; minimizes the empirical risk is not an effective learning strategy. When VC-dimensional of H is less than infinity we choose in a class H that is a VC-dimensional of H is less than a infinity then R hat n converges to R h uniformly or h. And this means that with a large with with with giving sufficiently large number of examples minimize of empirical risk would have low true risk also. Right. So, the global minimize the empirical risk will be close to the global minimum of true risk. So, if VC-dimension is less than infinity given sufficiently large number of examples atleast learning is possible.

(Refer Slide Time: 26:59)



The second point is, how large for number of examples? True risk is bounded above by empirical risk plus a complex data as we just saw. And the complexity term goes to 0 as VC-dimensional of H divided by n. So, higher the VC-dimension, higher is the number of examples we did. Before I can say good performance on examples means good

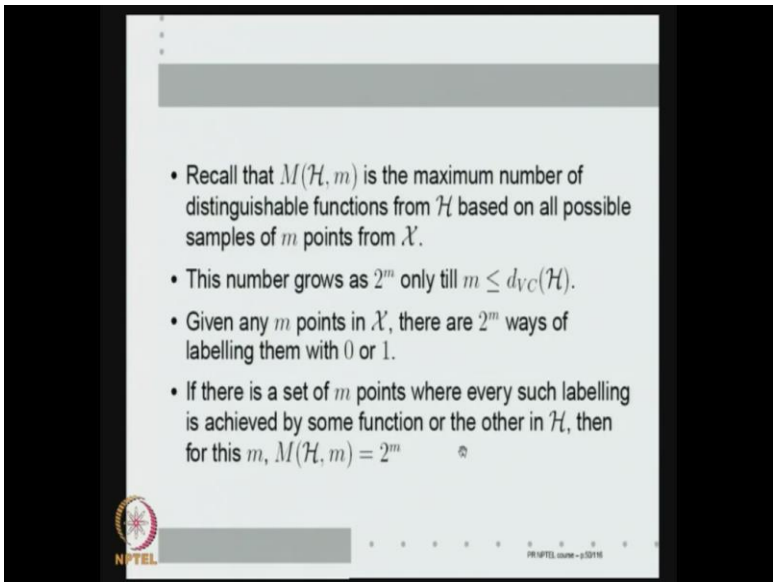
performance on unseen data. Our saying is that, if I solve only 1 exercise and got right, I cannot have much confidence that I will do well in the test.

But, if I solve 100s of exercise problems and got them right then, I think I can have sufficient confidence that I will do well in the test. Right. That is all it means. And off course, how many examples I have to solve depends on how complex is the concept on trying to learn. So, here we are learning a classifier. So, higher the VC-dimension higher is the number of examples needed and it is in this sense that VC-dimension tells the complexity of learning the things.

So, not only VC-dimension tells me whether empirical risk minimization will be effective or not, let us say binary question. So, VC-dimension is less than infinity empirical risk minimization would be effective. That is that is good to know and that is important. But, that is not the whole story. VC-dimensional of H as a number will also tell me how complex is the class of classifier from which I am trying to learn is, it tells me in the sense that for me to learn well; that is, for me to be confident low error on the data set, will mean low error on test also. What I mean is that the number of examples should be large relative to the VC-dimension. While this bound is rather loose. A rule of thumb most often used in must learning community is that the number of examples should be atleast 10 times VC-dimension because I have VC-dimension H by n multiplied by log of n by VC-dimension H.

So, if there is a fact of 10 the the the the increase because of log term will be more than compensated by the decrease in VC-dimensional of H by n. So, roughly 10 times VC-dimension is a good thumb rule and it is in this sense that higher the VC-dimension more complex in the learning problem. Have been appreciated VC-dimensional for what it is, let us ask how do we find VC-dimension. Let us remember that we are done all these only for 2 class classifiers as this entire analysis we have done for where the hypnosis space H contents functions feature by the valued functions all my input space.

(Refer Slide Time: 29:48)

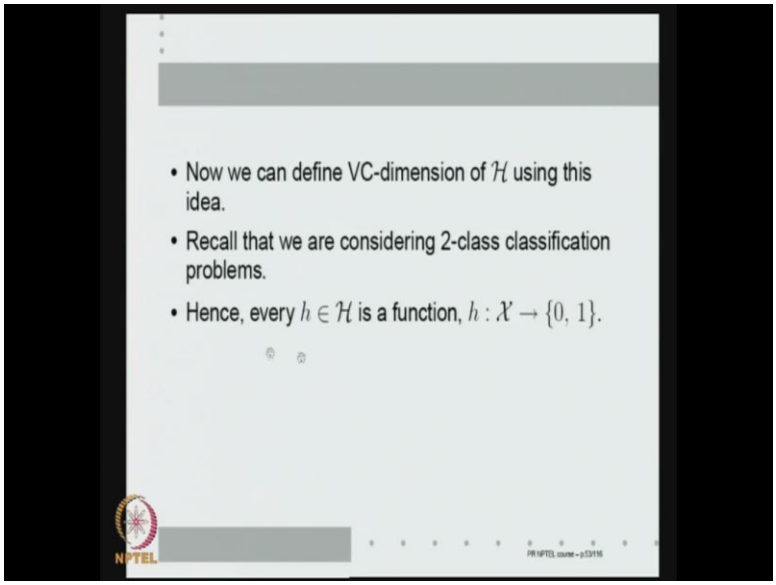


So, VC-dimension is tied to the growth rate of this quantity. Now, what is this quantity? This quantity is the maximum number of distinguishable functions that I can pick from my class of functions H based on all possible samples m points from input set, from my input space, right from my feature space. We know that this number, the maximum possible number of distinguishable functions from H grows exponentially as 2 power m only as long as m is less or equal to VC-dimensional of H. So, given a set of m points in X, any specific set of m points in x, how many different ways I can assign class label to them? I can assign 2 power m base of assigning labels 0 or 1 to them. Right. Because I am considering all the functions in my hypnosis space H are binary valued functions, each function can take only value 0 or 1.

Given any m points in X there are 2 power m different ways of assigning 0's and 1's to those m points. That means, anyways the maximum number of distinguishable functions is always 2 power m, there cannot be more than 2 power m distinguishable functions any way given m points. Right. So, the question is, the maximum number of distinguishable functions will be 2 power m, if every possible label in is realized by some or the other function. Right. So, given a set of m points, I can label it into 2 power m ways, for each labeling there is a function H which exactly realizes that labeling. Meaning, that function H takes those values 0's and 1's and those m points; then, I know the maximum number of possible functions will be 2 power m. right.

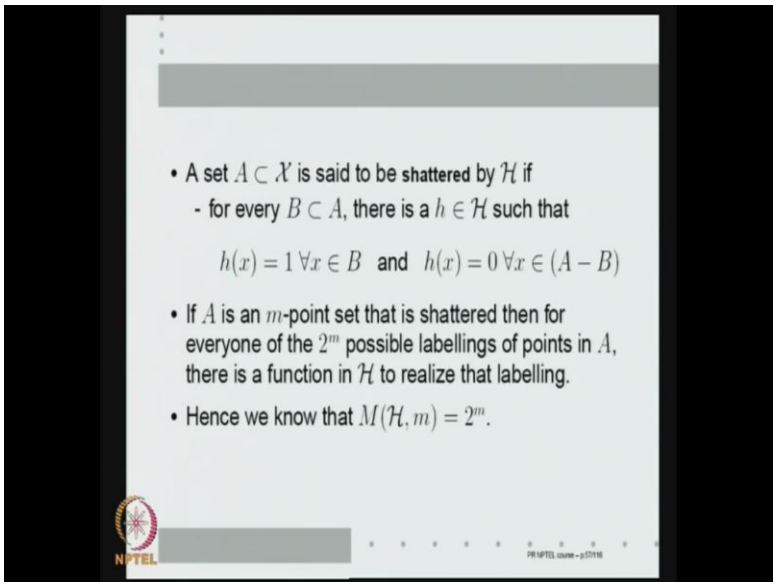
So, maximum number of distinguishable will be 2 power m. If given m points, if I label them on all possible ways, for every such labeling there is some function in H that with respect to which that labeling is consistence and that is that is what this quantity is. The maximum number of distinguishable functions is nothing but if I take m points and keep labeling them 0's and 1's, there are 2 power m base m for doing it. For every such labeling if there is one function in H that realizes that labeling then, they will be total of atleast 2 power m distinguished, they they there will be atleast 2 power m distinguishable functions. So, we will use this to get a more useable definition of VC-dimension.

(Refer Slide Time: 32:28)



So, how do we go about this? Before we do that let me re-emphasis, we are defining all this only for 2 class classification problems that is, for us the hypnosis space h transits of set of function H set where each function maps my input space script type is the feature space to 0, 1. So, we are considering only 2 class classification problems where, what we call the action space is same as the class labels. We are only considering bind the valued functions on X. If my h is that kind of family we are trying to ask, what is the VC-dimension?

(Refer Slide Time: 33:21)



To define VC-dimension we define the concept of what is called shattering. We will say a set A which is a subset of input space X is said to be shattered by the family of functions H. Let me emphasis, the concept of shattering is a property of subsets of my input space. It has nothing to do with labels, nothing to do with underline problem distribution nothing. It is a property of subsets of X and also of the family of classifiers that I am considering. So, given a particular subset of X it is said to be shattered by the family of classifiers H in the following holes.

So, given any set A which is subset of X, I will say it is this shattered by H. If, if you give me any subset B of A, for every subset B of A there is a function in my family of functions, which of course it will depend on B. Such that, function H will take value 1 for all points in B and take value 0 for all points in A minus B. I do not get what it takes outside A but, this set A is shattered if for every subset B of A there is a function in my family such that, that function takes the value 1 for all points inside that set inside set B and x value 0 in the rest of the points. That is, I say point A, not in B. This is what is done; I say A shattered by H.

If B is an m point set that is shattered, what does it mean? See, every possible subset, how many possible subsets of B subset of A are there, if A is m possible subsets. Why? Choosing the subset of A is same as putting a particular labeling of 0's and 1's on elements of A. right. So, choosing a giving the subset of A simply says that ok, these

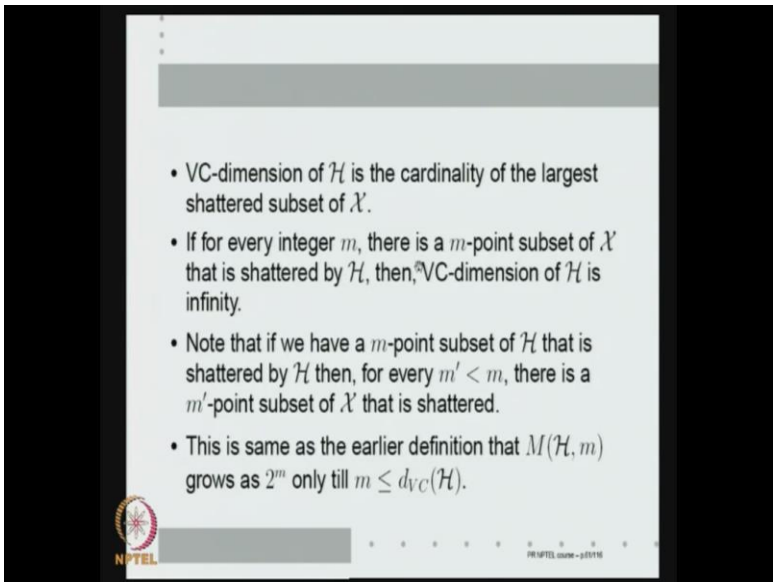
elements of A are not in the subset, these elements of A are not in the subset. Which means, a subset is characterized by the so called the characteristic function which is simply a way of labeling the elements of A by 0's and 1's. Those that are label by 1 are in the subset, those are label by 0 are not in the subset.

Now, for every possible subset there is a function that exactly realize the 1's and 0's. Which means, if A is an m point set and I label the points of A in all possible ways with 0's and 1's that 2 power m possible labeling. For every single labeling, every single labeling will be a subset of A. Every single labeling, there is a H that realizes that labeling because that H takes value 1 for points in B and 0 for points A minus B.

So, if m point set is shattered then, for every 1 of the 2 possible labeling of points A there is a function to realizing that labeling. That means, here is m point set for which the number distinguishable functions is 2 power m. Right. So, which means because I can find atleast one m point set such that, the distinct number decimal function 2 power m and I know number distinguishable function cannot be more than 2 power m. We know the maximum number of distinguishable function is 2 power m. So, if an m point set is shattered then, we know for that m the maximum number distinguishable functions will be 2 power m. All right.

So, this tells us how we can define the VC-dimension. What is VC-dimension? The maximum value of m till which this is true. The maximum number of maximum possible number of functions is 2 power m. Right. VC-dimension is that value of this integer such that, the maximum number of possible functions is 2 power m which means, VC-dimension is nothing but the cardinality of the largest shattered subset of X.

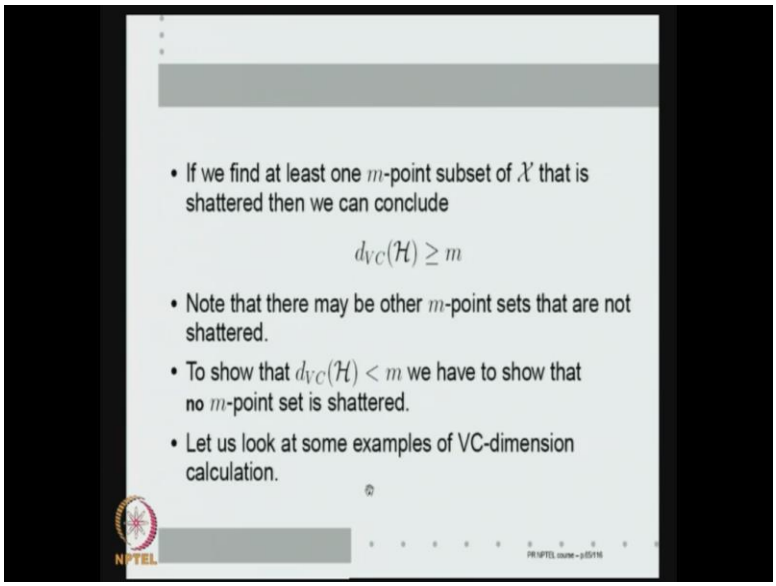
(Refer Slide Time: 37:14)



So, we can define VC-dimensional of H as the cardinality of the largest shattered subset of X. Of course, when I say largest; essentially, we can say rink finite sets. So, if for every integer m there is a m point subset of X that is shattered then, VC-dimension is infinite. So, for every integer m there is a m point subset that is shattered then, VC-dimension is infinite. Otherwise, VC-dimension is the cardinality of the largest shattered subset of X. fine. I also remember that if I found one m point subset for which that is shattered then; if I take any subset or that set that is also shattered. Obviously, because if every possible labeling of A can be realize the function in H then, every possible labeling of any subset of A can also be realized by a functions in H.

This means, if I have an m point subset of H that is shattered. Then, for every m prime which is less than m there is a m prime point subset atleast one namely, an m prime point subset of that particular shattered set which is shattered. So, that is why we need only look at the m, the largest m for, the highest cardinality set that is shattered. So, hence this definition VC-dimension is the cardinality as the largest shattered subset of X is the same as what we saw earlier that it is the value of m, the largest value of m for which the maximum number of distinguishable functions grows as 2 power m. So, now we can work with this. VC-dimension is the cardinality of the largest shattered subset because it is easier to find whether given is shattered by H or not.

(Refer Slide Time: 39:01)

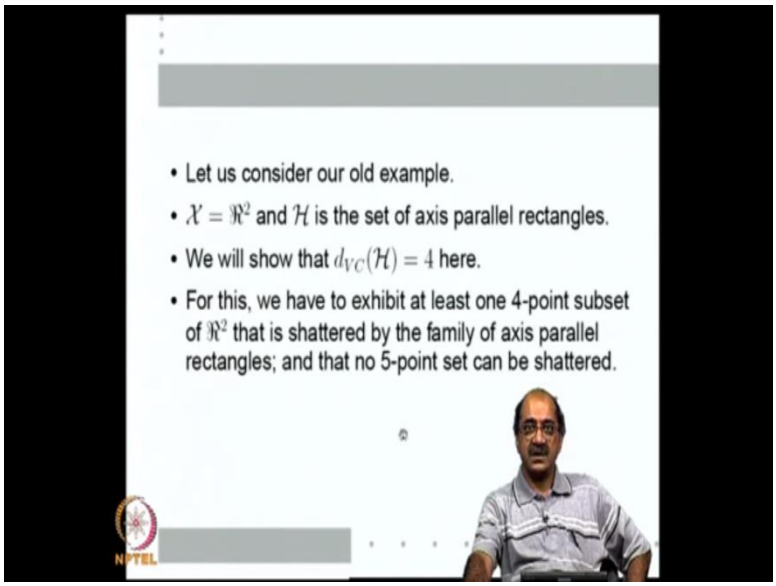


So, let us sum the state this again so that we fully understand. If we find atleast 1 m point subset of x that is shattered then, we know that VC-dimension is atleast m. right I can exhibit an m point subset that is shattered then, I know VC-dimension can be less than that m. So, if I exhibit 1 m point subset of x that is shattered then, we can conclude that VC-dimension is atleast m.

On the other hand, when I say this we should also understand that not every m point subset will be shattered because VC-dimension is the overall possible m samples what is the maximum number of distinguishable function. So, if I can find atleast 1 m point subset of x that is shattered then, I know VC-dimension is atleast m. But, this does not mean that not that the the there can be the settling does not preclude, they being other m points of that are not shattered that does not concern us. So, to show VC-dimension is atleast m, we need to just exhibit atleast 1 m point subset. Of course, not every m point subset can be shattered but, there might be 1 m point subset that will be shattered.

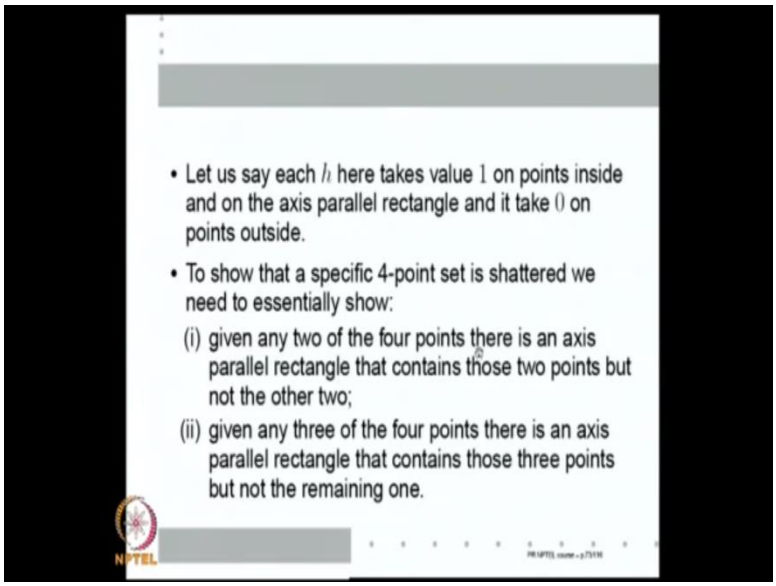
On the other hand, to show that VC-dimension is strictly less than m we have to show that no m point subset is shattered. Ultimately, to show that VC-dimension is m, we should show atleast 1 m point subset that is shattered and also have to show that no set containing m plus 1 points is shattered. That is how we can conclude that VC-dimension is a particular integer.

(Refer Slide Time: 40:42)



So, let us go back to our old example. Our X is R 2 and H is the family of axis parallel rectangle. Let us calculate, what is VC-dimension of the family of axis parallel rectangles? We are going to show that the VC-dimension of axis parallel rectangles happens to be 4. How can I show just from that what I have just shown, to show that VC-dimension is 4. We have to exhibit atleast one 4-point subset of R 2 that is shattered by the family of axis parallel rectangles that no 5-point set can be shattered. Right. These both of them I have to show before I can conclude that VC-dimension if 4. To conclude that VC-dimension is 4, I have to exhibits atleast one 4 points of set of or 2 that is shattered by the family of X by rectangles and that no 5 point set can be shattered by this family. So, this is what we are going to do.

(Refer Slide Time: 41:35)



So, we need to have some convention of, when h takes value 1 and when it takes value 0, 1 and 0 are arbitrary. So, let us say our convention is takes value one on points inside on the access files rectangle and takes the 0 on points outside.

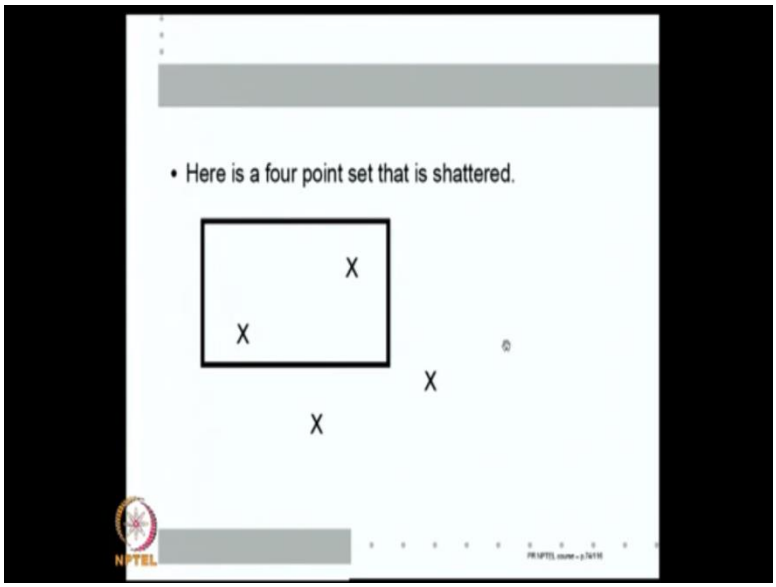
Now, a given a particular 4 point subset that is shattered, what does we have to show? For given any subset of these 4 points, on those subset it has to take value 1 and the rest of the set it has to take value 0, that is what shattered means; for every mean subset have to do this. When will the axis parallel take value 1, if the points are inside. So, what it means is for this particular 4 point to be shattered give me any subset of these 4 points, I can draw an axis parallel rectangle such that, only that subset of points are inside the axis parallel rectangle and rest of them are outside the axis parallel rectangle.

Now, what is the subset that can take 0, that is very easy give me 4 points I can put axis parallel rectangle somewhere totally outside. And similarly, this subset contains all of them also is fine; I can certainly put axis parallel rectangle over them. So, you should consist subsets of 1, 2 and 3. So, any one point can be separate from the other 3, any 2 points can be separate from the other 2 and any 3 can be separate from the other 1. Separating 1 from 3 is same as 3 from 1.

So, essentially we need to show 2 things. Given any 2 of the 4 points, there is an axis parallel rectangle that contains those 2 points but, not the other 2. And similarly, given any 3 of the 4 points, there is an n axis parallel rectangle that contains those 3 points but,

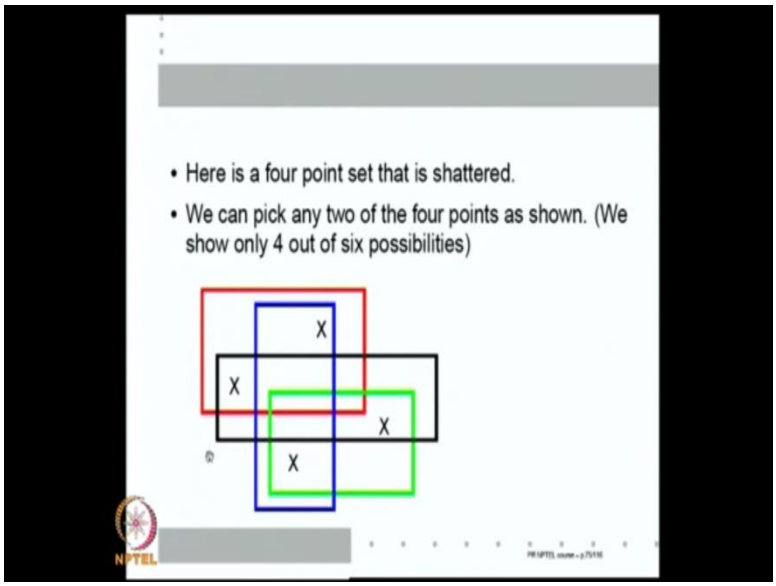
not the remaining ones. these these This is what I have to show, to show that a specific 4 point set is shattered. So, this is an example. So, we just have to exhibit one 4 point set. So, let us exhibit on such set.

(Refer Slide Time: 43:22)



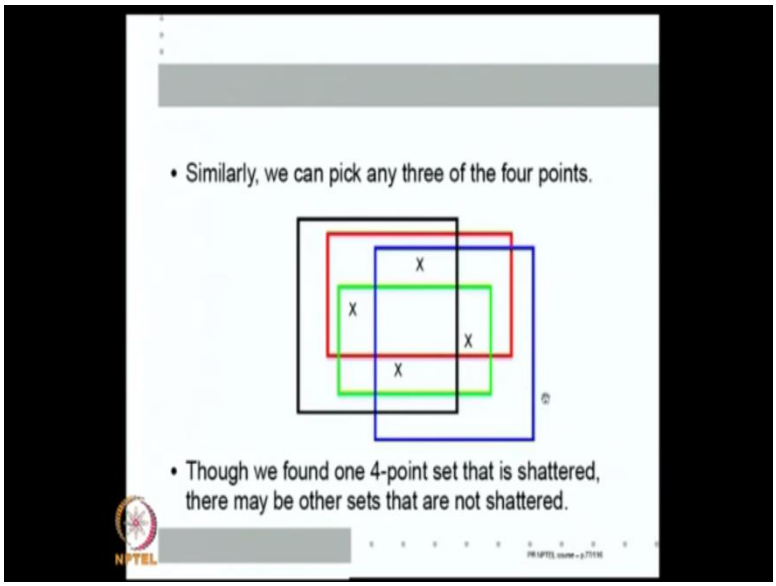
Here, is a 4 point set that is shattered that means, R 2 I did not draw the axis, just any 4 points on the plane. So, will show will currently see that 4 point set is shattered, what I am saying is essentially give me any subset be that means, I am going to label these 2 sets as 1 and the other 2 as 0 that what given me a subset means. Then, an element of my family of X parallel rectangle that realizes that labeling is accepted. Because this element of my family takes value 1 on these 2 points and takes value 0 on these 2 points. To really see that it is shattered.

(Refer Slide Time: 43:59)



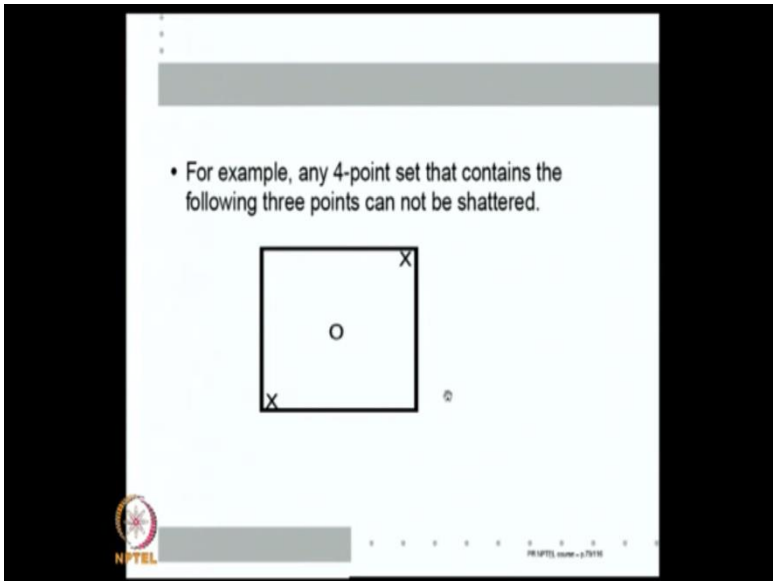
Here is the full glow. So, if I want any 2 of course, 6 possible ways of choosing 2 from 4 m I put 4 because already the diagram is scattered. So, for example, I want to choose this and this point then, the blue axis parallel rectangle may keep those 2 inset. On the other hand, if I want to take be other 2 this and this, the black one (( )). Similarly, the red one picks this to the green one picks this 2 on so on. Similarly, we can draw the remaining. Essentially, it is very easy to see now once you draw this, I have put this 4 points so that each one has a unique x and y coordinate. So, at this x or y there is no other point at this x or y there is no other point to inform. Once it is there I can write axis parallel rectangle enclosing any subset. right is is is It is in totally very clear.

(Refer Slide Time: 44:57)



Here, is how I can do it for 3. There are only 4 ways of picking any 3 points and all the 4 are exhibited here; all the different, all the 4 size 3 subset of these 4 points are enclosed in each of this 4 axis parallel rectangle. So, this shows that this particular set of 4 points is shattered by the family of axis parallel rectangles. Of course, that this particular subset is shattered does not mean that every 4 point set is shattered, there can be either 4 point set which is not shattered. Right.

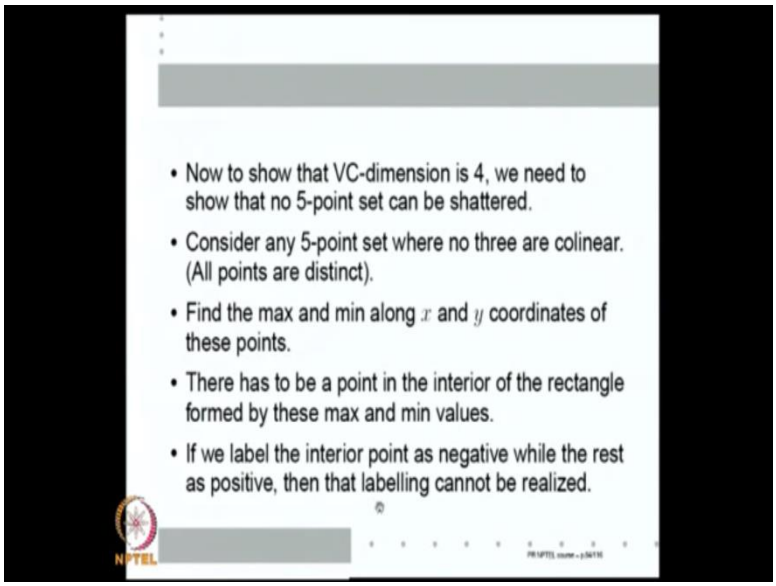
(Refer Slide Time: 45:36)



For example, if you take these 3 points and put the fourth point anywhere it really does not matter. Then, that is not shattered. Just to show is not shattered I put a particular label here, I want this to be outside the rectangle and this not inside the rectangle and this 3 in a line. Because this 3 are in a line, it is impossible to put these 2 points inside an axis parallel rectangle which does not contain this.

For example, any axis parallel rectangle contains these 2 has to contain this. So, here is a 3 points of said that is not shattered. Right. A matter of fact any 3 collinear points cannot be shattered by X parallel rectangle because this particular labeling cannot be realized by any axis parallel rectangle. I want this and this to be inside the rectangle but, this to be outside the rectangle as not possible. So, in the other 4 points subset we shook it is shattered because no 3 are on a line. But, otherwise they can be 4 point subset or even 3 point subset that cannot be shattered, as we said same that VC-dimension as piece m only means that there is 1 m point set that is shattered not every m point that is shattered. So, now we know that axis parallel rectangles of VC-dimension atleast m.

(Refer Slide Time: 46:49)

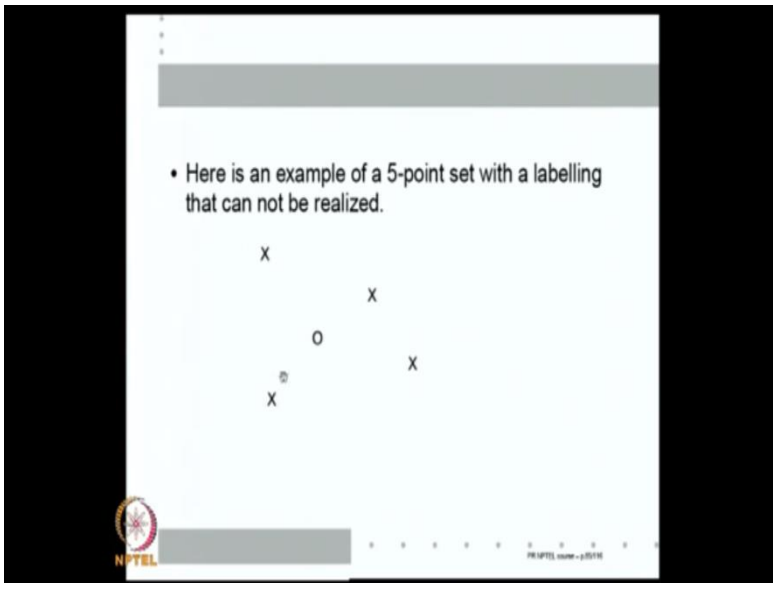


Now, to show that is actually m, we need to show that is no-5 point set can be shattered. How do you argue this? Now, we cannot argue with an example, I have to say no-5 point set is. So, consider any 5 point set where no 3 are collinear. Why? If any 3 of them are collinear because I have to show this for every point set. So, how can I put restriction? Take any 5 point set if any 3 of them are collinear, I already know it cannot be shattered.

Because once 3 are collinear it will be like this. No matter where you put the other 2 points, this particular labeling itself cannot be realize so that 5 point set is not shattered. So, if any 3 points are collinear any way it cannot be shattered. So, take a 5 point set where no 3 are collinear. No 3 are collinear then, you find the maximum and minimum x and y coordinates among this 5 points. Because no 3 are collinear, a little thought will show you that they can be at most, they has to be atleast one point in the interior of the rectangle formed by these max and min values. Once I find max x coordinate, min x co-ordinate, max 5 coordinate, min y co-ordinate, I can learn axis parallel rectangle because all points are distinct and no 3 or collinear. At most 4 of them can be on the rectangle, the other one its x and y coordinate at atleast one of them has to be less than the max and hence it has to be inside the rectangle.

So, if I put the interior point as negative and the rest of them as positive that means, the rest of them should be inside the rectangle but, interior point should not be. Such a labeling cannot be realized just like the 3 collinear things. So, which means there is a 4 point subset that is shattered and no 5 subset is shattered. And hence, VC-dimension of x parallel rectangles is 4.

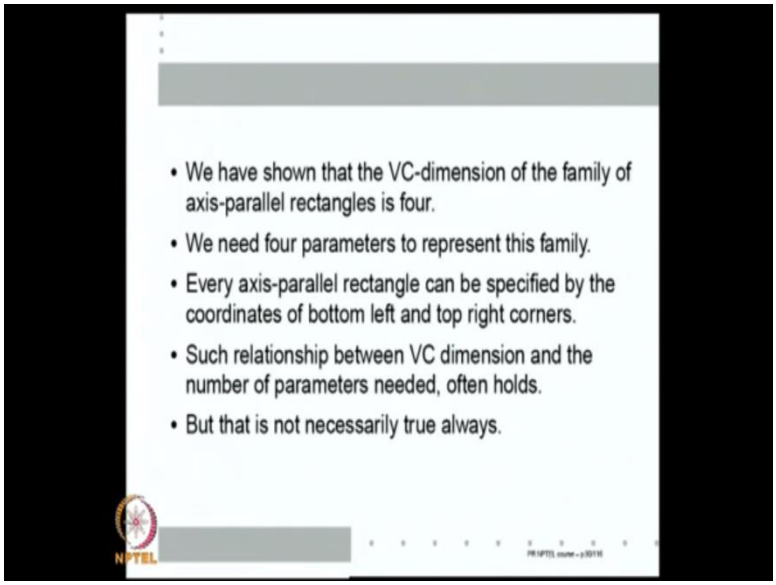
(Refer Slide Time: 48:42)



Here, is a 5 point set that is and the labeling that can. So, if I take the max min x max x, min y max y. So, is very easy to see that is the axis parallel rectangle I will draw and

hence, there will be an interior point. If I label this as 0 and less of them as one such a label cannot be realized.

(Refer Slide Time: 49:09)

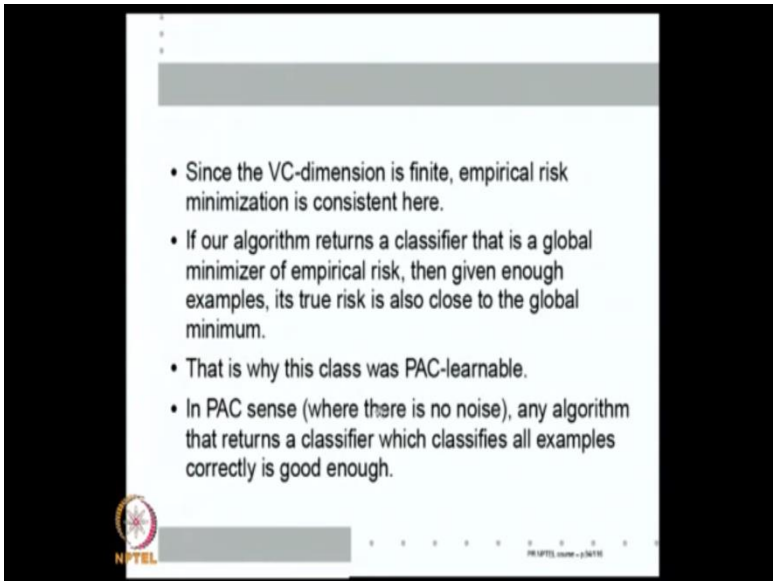


So, we showed that the VC-dimension family of X rectangles is 4. Because it is also interesting that we need 4 parameters to represent this family. What do you mean 4 parameters to represent a family? Any axis parallel rectangle can be specified by the coordinates of the bottom left and top right corners because access parallel if you give me the x, y coordinates of the bottom left corners and the top right corner axis parallel rectangle is completely specified. So, I just need 4 numbers to specify any one access parallel rectangle. So, the entire family of access parallel rectangles can be represented with 4 parameters.

So, to learn a element of this family I essentially need to learn this specific values of 4 parameters and the VC-dimension happens to be 4. In general such a relation between VC-dimension, the number of parameters needed often holds. Very often VC-dimension is roughly same as the number of parameters needed to represent that class of functions. Of course, this is not always true. We can have one parameter family of function which as VC-dimension infinity. For example, if I essentially cancel all possible sign curve. Essentially, saying no if I put a sense on the sign curve on one side sign curve is positive other side the sign curve is negative then, a one parameter family give me sign curves of all possible frequencies.

So, in (( )) if I put some points, I can essentially draw some sign curve where the other keeping some points on one side and other points on the other. So, we would not consider that example right now but, just remember thus necessarily true that the VC-dimension is equal to the number of parameters needed, as I am telling you that is a one parameter family of functions whose VC-dimension is infinite. But, very often the number of parameters we have to learn, to learn an element of the family is equal to the VC-dimension and in that sense, once again it is very satisfying that VC-dimension is the complex. But, we have to learn many more parameters. Obviously; the running problem is more complex.

(Refer Slide Time: 51:11)

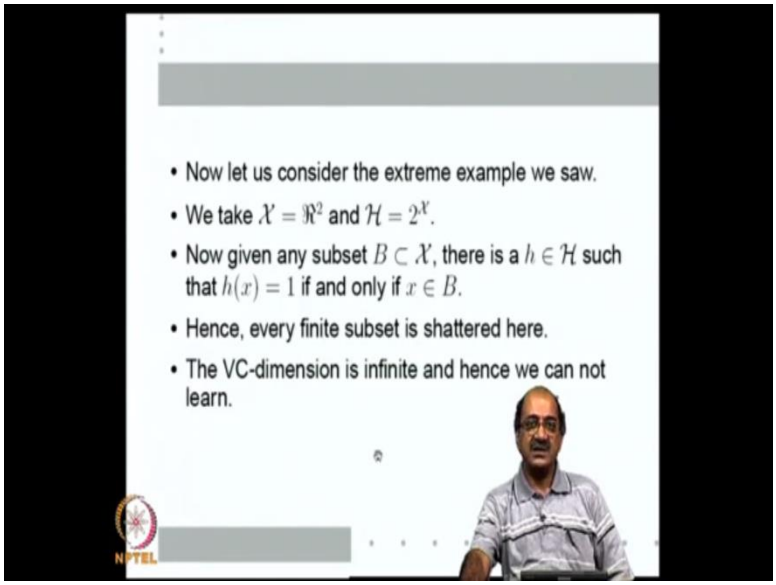


Since, VC-dimension is finite, we know for this access parallel rectangle problem empirical risk minimization is consistent. So, if our algorithm finds the global minimizer of empirical risk then, given enough examples the true risk of the classifier will also be close to the global minimum of true risk. In particular, that is why it is PAC-learnable. In the PAC frame work there is no noise. So, the global minimum possible for empirical risk will be 0 because there is always some function in my family that consistent with all examples. So, if you give me any global minimizer of empirical risk that means, any classifier that classifiers all examples correctly that is good enough. Actually, earlier when we concern we said give me the smallest access called rectangle that encloses all the positive examples.

We need have said that now, we know once all the. An algorithm that is gives me either this smallest access parallel rectangle they enclose all the positive point examples as the largest access parallel rectangle that does not enclose any negative examples or anything in between. Essentially, as long as my algorithm returns a access parallel rectangle that classifies all the examples correctly, that algorithm PAC-learn because given n of given n of examples it will be close to the global minimize of true risk. That is what VC-dimension finite means.

VC-dimension is finite and like in PAC sense there is no noise then, anything that does well on data will do well. Anything that class class classifies all the examples correctly will also have very small text error. In particular, if there is no noise of course, if there is noise then, also it is true in the sense anything goes empirical risk is sufficiently small will have its true risk also sufficiently small, if the number of examples is large enough. How large is large? Large is relative to the VC-dimension.

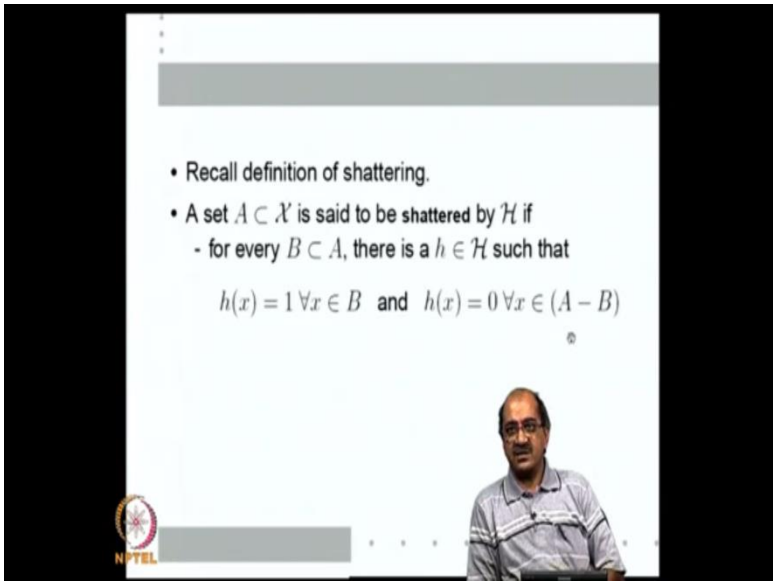
(Refer Slide Time: 53:09)



Now, let us consider the other examples. So, we when we consider that we had the other extreme example where instead of taking X parallel rectangles we taken all possible 2 class classifier on X. So, every possible subset of X is in h. When I say H is equals to 2 for x essentially, means every by the valued function on X is in my family H. As you seen this is true flexible we cannot learn. So, we should expect VC-dimension to be infinity here. Will show that infinity. What does H equals 2 for x means? You give me

any subset B of X or 2; so, give me any subset of or 2 because H contains every subset. Essentially, what means is there is a function H which takes value 1 only for points in B and 0 or self. So, for every subset B of X there is a function H. I can call it H B which is the characteristic function of that said B actually which takes value 1 for all x in B and value 0 or set.

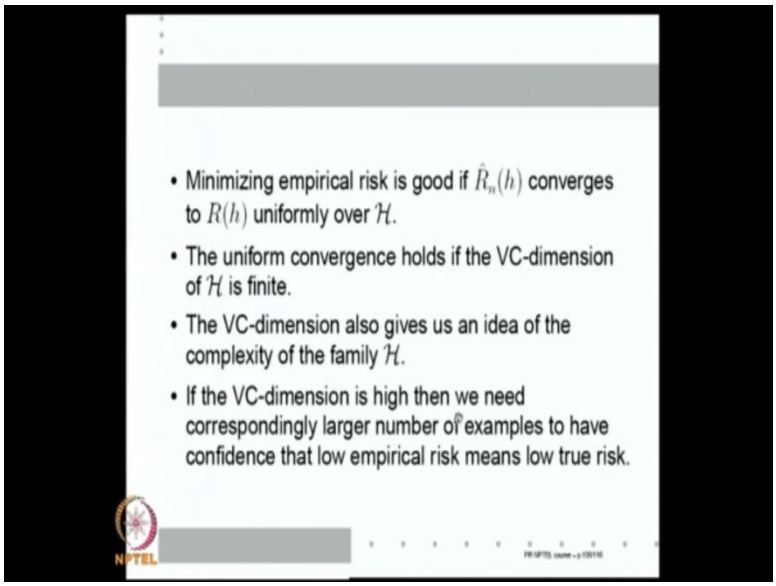
(Refer Slide Time: 54:11)



What is my shattering definition? Given a set A that is shattered, if for every subset B of A there is a h that takes value 1 on the B, 0 in A minus B and outside A, I do not care. Now, given any subset B whether subset of A or not there is a function that takes 1 only for points in B.

That is what we have just now seen. So, which means because we have this, give me any finite set it is shattered by this family. Because give me any finite set A, give me any subset B of A then, there is a function that takes value exactly 1 and be in 0 elsewhere. So, 0 is A minus B which means, every finite subset is shattered here. So, because every finite subset is shattered, the VC-dimension is infinite and obviously we cannot learn. Right. As you already seen there is no generalization possible if I try to learn with this thing and that also comes through because this particular H is VC-dimension infinity.

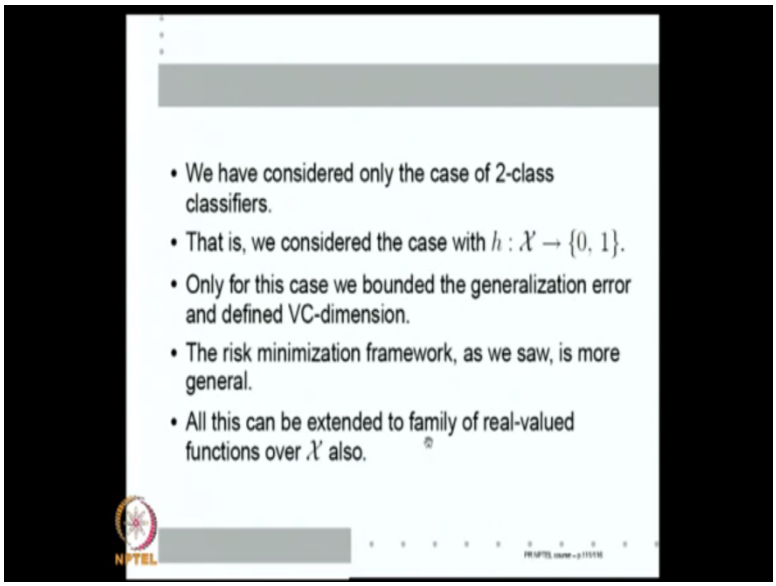
(Refer Slide Time: 55:16)



So, essentially what we have done is; what the essence of the theory is that we have presented so far is. Minimizing empirical risk is good if R hat n h converges to R h uniformly or h. The uniform conversion holds, if VC-dimensional of H is finite. And VC-dimension also gives us an idea of the complexity of the family h. We have seen some examples. Essentially, complexity means if the VC-dimension is high then, we need correspondingly large number of examples to have confidence that low empirical risk means low true risk.

Because the true risk of any h can be bounded above by its empirical risk plus a model complexity term which goes to 0 as VC-dimension by the number of examples. So, if the number of examples is sufficiently larger L 2 to VC-dimension then, we can have sufficient confidence the low empirical risk means low true risk right. In this sense it also tells us gives an idea of the complexity of the learning problem.

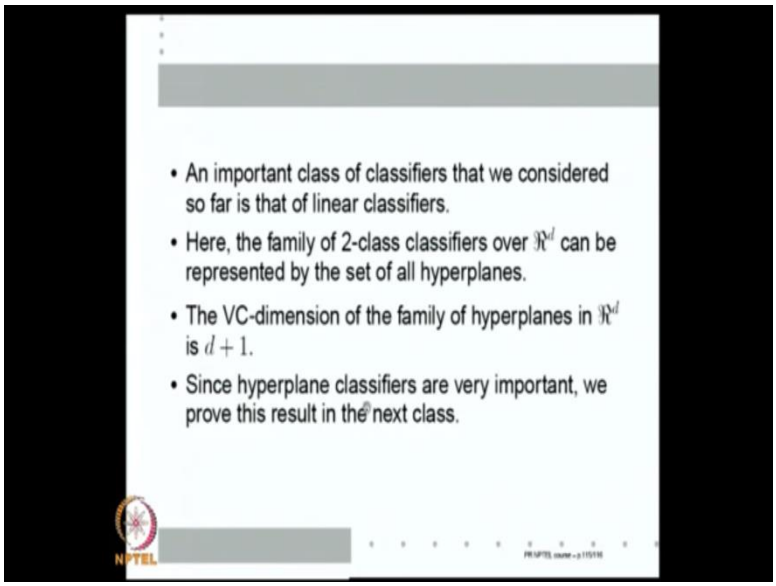
(Refer Slide Time: 56:17)



We have considered only the 2 class classifiers so far within that define. All the functions we considered by the value of functions on X. Only for this case we form bounding, we bound the generalization error and defined VC-dimension. Right. For example, the shattering that we have defined does not even make sense if h is for example, real valued function on X. Then, what label what labeling is realize or not realize how can we say. Will just state that as we already know the risk minimization frame work is more general, we do not have to take function h to be X to 0, 1 or X even to class label, we can take h to R also. We have seen that even discriminate function can be improved in the risk minimization frame work.

So, as it terms out, everything that we did about of the VC theory about VC-dimension, about this need for empirical risk to converse to true risk uniformly or H for empirical minimization to be consistent on the notation VC-dimension. All this can be extended to family of real-valued functions over X because the extension is both conceptually as well as notationally complex, it needs lot more knowledge of real analysis and so on. So, we are not even attempting it. So, will just state that this can be extended to general function on X also.

(Refer Slide Time: 57:41)



And been said that we will still look at 2 class classifiers functions. One important of 2 class of classifiers that we considered hyperplane classifiers so called linear classifier made out of linear discriminate functions. And this is important for us for later on also in the course when we look at what are called kernel base methods. So, here each 2 class classifier is essentially hyperplane in r d, one side of the hyperplane is one, the other side of the hyperplane with the other. As transferred in d dimension, the VC-dimensional of hyperplane is d plus 1. Since, the hyperplane classifier are very important to us we will also learn how to actually show that the VC-dimensional of hyperplane classifier is d plus 1 in the next class.

Thank you.