Computational Geometry

Prof. Sandeep Sen

Department of Computer Science and Engineering

Indian Institute of Technology, Delhi

Module No. # 13

E-nets, VC Dimension and Applications

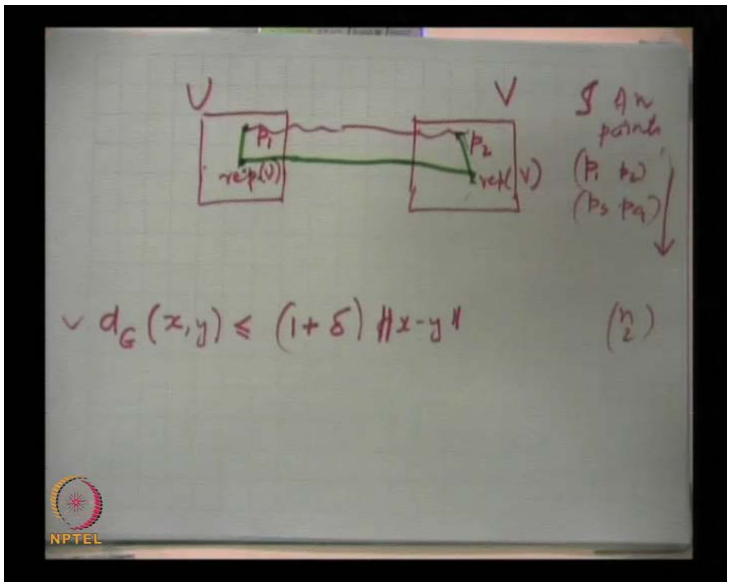
Lecture No. # 01

E-nets & VC Dimensions

So, we will start with a new topic and I left off in the last lecture, where we are I was trying to write down a proof for the construction of the epsilon 1 plus epsilon spanner. So, I do not want go to through the proof in details. I just give you an idea and leave it as an exercise. It is rather simple actually. You just have to fill up the details before we start in this new topic. I just open up the page and tell you what needs to be done ok.

So, we are trying to prove this by induction right. So, this these 2 point sets, these are well separated point sets right. These two sets are well separated. We are trying to prove by induction, induction on the pair wise distances. So, we have a set of n points. So, a set S of n points, the the pair wise distances could be ordered right in terms of increasing distances.

(Refer Slide Time: 01:11)



So, let say the pairs are some like p 1, p 2 and p 3, p 4. So, these are the distances in in order to increase in distances, the n choose 2 distances in order of increasing distances. So, we do the proof by induction on this sequence right. So, suppose we have proved that the spanner property, that is distance in the graph between points x and y is no more than 1 plus using a delta spanners, I think yeah delta spanner 1 plus delta times the Euclidian distance between x and y ok

So, we need to prove this for every pair of points. So, suppose we have done it in theis sequence for some of the pairs, so suppose that now we are dealing with pairs of points, p 1 and p 2 and for all distances less than p 1 and p 2, this holds. That is my induction hypothesis ok.

Now, what we do the construction is nothing but, for each of this well separated pairs here, we have a representative. We just pick a arbitrary representative. So, suppose this set is, did I name this set? Suppose, let us call as capital U and capital V or something like that. So, this is representative capital U and this is some representative capital V. This edge is, what we add to this spanner, the spanner that we are constructing. This is explicitly added and we claim that everything is taken care of by the sets of such edges, and the induction proof will use this fact.

Now, since we are assuming that all distances less than p 1, p 2 is taken care of because these are well separated sets subsets. It means that the distances within the subset, they are already taken care of here because these are well separated pairs or distances within the subsets are much less than the distances. So, at least epsilon fraction, less than theses distances, so everything within a subset is already taken care of.

In particular, this distance in the graph is already less than or equal to 1 plus epsilon times the Euclidian distances, for this connection. We can also argue that the distance between that Re, the representative of two subsets is not more than some constant fraction times the distance between p 1 p 2. That is the Euclidian distance right because it is well separated. So, Euclidian distances between the representatives of these two sets are not very different from basically any of the pairs chosen from across the two subsets. So, that is one property we are going to use. Other property we are going to use is induction hypothesis that this distance in the graph is already captured in the spanner ok.

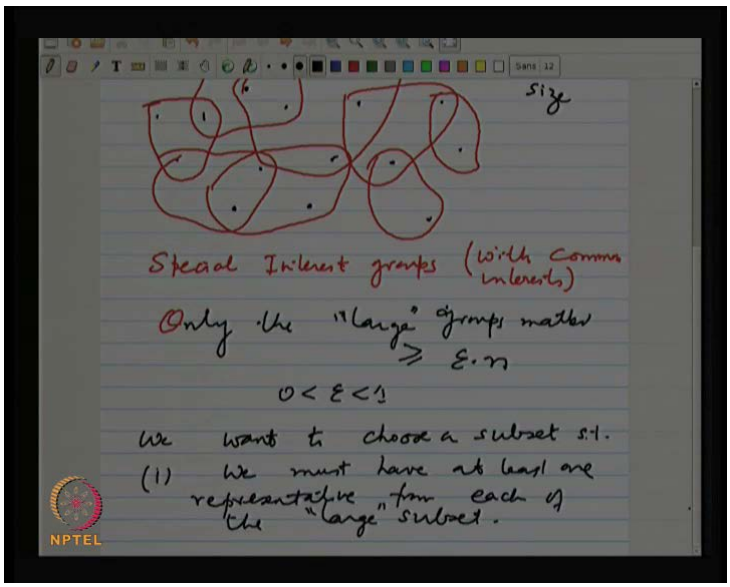
You just have to put the put the triangle inequality to figure to argue that in the graph, this distance is not going to more than 1 plus delta times Euclidian distance right. So, that I just leave it as an exercise, so you can complete that.

So, with that let us move over to the next topic. What are the epsilon nets and what are V C dimensions? I think you have seen you have heard the first term before. Now, I write so at least I believe that when you are talking about these cutting partition trees and cuttings. Did this term come up at all when Professor Aggarwal was lecturing? Epsilon nets or you do not have any collection.

So, this epsilon net is conceptually is no different from the definition of the one by R cutting use to construct the partition tress ok. So, but, we will actually explore in depth about what these things are. So, these are very deep you know combinatorial consequences.

So, let me introduce with a problem and introduce the problem and give you the motivation in what started off this. Then, this has this was first discovered in the context of what is called hyper graph theory. It was back in like in early 70s or so VC denotes you know the actually names of the two authors, two Russians mathematicians. Vapnik-Chevronenkesin I think. So, VC dimension is a short form of this.

(Refer Slide Time: 06:26)



So, they discovered some some nice combinatorial properties for certain subset systems ok and which later, very soon found huge amount of applications in in all kinds of fields, you know especially, you know the people in learning theory make heavy use of this. Then, even in geometric searching, people make lots of use of this and essentially, in any subset system that satisfied a certain kind of property, so which is basically this V C dimension is.

So, VC dimensions is some property, if a subset system satisfies, then you can find, let us say, I mean I will just discuss the problem. Find covers of small size essentially right. So, let me introduce with a generic problem. So, we have some population, which are nothing, but let us say vertices of the hyper graph and there some groups of people and the groups of people are nothing, but subsets of these set of people of the population.

Ok so let me mark this out. This could be one subset, this could be another subset and this could be another subset. These subset systems are hyper graph basically, something like this. So they are all kinds of subsets or you can say may be interest groups of population. Special interest groups may be with wasted interests. One very common problem is that we want to do some kind of survey on this population. We have a certain opinion poll that may be we are trying to conduct, like was the common wealth games successful.

Now, different groups of people may have different opinions. So, I have said wasted group, the people organize it you know that they will certainly say, that yes it was successful. The people who participated, you know they may have a split opinion on that and the people who you know bore the brunt of all these constructions would certainly say that nothing was successful. Now, for the investigating agencies got into it and they have to start out with an assumption that everything was corrupt ok.

So, suppose there is a opinion poll. Now, in a in a democratic society what happens is that you know we do not tend to go by you know how large the size of the group right. We do not really care what kind of wasted interest group has. Somehow, the democracy is supposed to take care of everything. No doubt that you know it was said that it is a system government for the fools, by the fools, of the fools right. That is Aristotle thing. Anyway, so we only go by you know the size of the group. The bigger the group, you believe that you know they have the right to the authority to be right. Ok.

So, these subsets basically represents those groups and then, what we want to make sure is that when we do a opinion poll, we do not actually go and ask every person right. If you could ask the entire population, of course, then we get a very accurate picture of ok, so here is a group of people who think this way, here is an another group of people who think that way. So, that is not possible. All the opinion polls that are conducted by the media is essentially on the basis on the sample right

Some subset is small subset. They can realistically actual go and spend few minutes with and get the feedback. It can be with a phone call, could be a letter, you know could be a with a stick your microphone up your nose, but then, when the sample is selected, if the sample is not selected thoughtfully and properly, then of course, the finally, the opinion that comes out of it is completely could be completely fallacious. That is what often happens when you see that opinion polls go wrong. Actually, the basic theory of opinion poll is quite statistically sound. It is very sound. The problem happens is either the sample has not been chosen properly or that a large enough sample may not have been chosen. Both ways ok.

The sample is chosen properly, may be it should be a proper random sample, so maybe it was really not a random sample or that not enough people were consulted. The sample size was inadequate and if both these things are right and I mean there is some formula, so if you satisfy these two criteria, large enough sample size and sampling was run properly, then there is a mathematical formula that, let us say that opinion poll is going to be correct with you know any improbability tending to almost one. So, this is fairly sound to you.

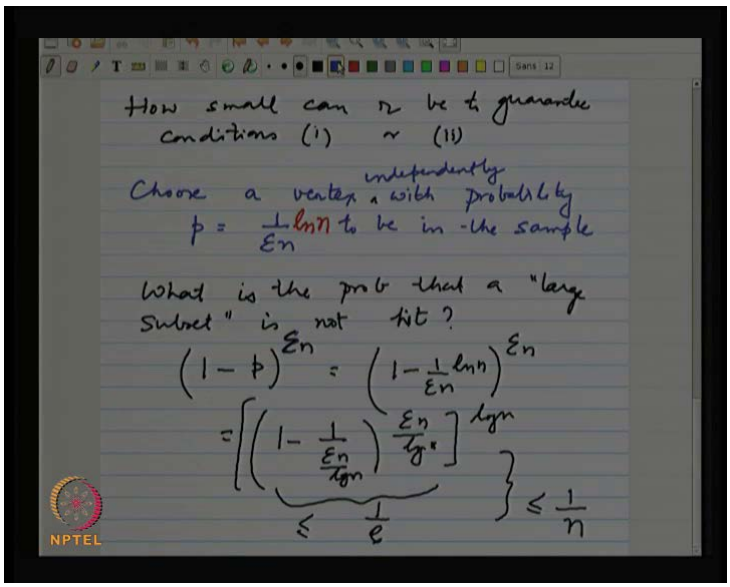
Now, here you know it is slightly different. What we you know intend to do is that the groups that are large, only the large group matters. Let us say, only the large groups matter. How large is large? Let us choose a fraction. Let us say, larger than some epsilon times n. N is a size of population. So, some fraction epsilon, let say epsilon is between 0 and 1. So, may be epsilon 10 percent, 5 percent, 1 percent. So, we are saying that unless the group has at least 1 percent size of the population, we do not count. Then, what you do? You basically, then you want to sample in a way, so we want to find or chose a subset such that, number 1, we must have at least one representative from each of the large subset. By large, I mean that you know at least epsilon in size.

Of course, the objective is not just to have at least one representative, the subset that we choose, we want to choose as small as subset as possible to satisfy this condition right. I want to choose as small subset as possible and satisfy the condition that every large subset is represented right.

So, this is somewhat weaker condition. Then, perhaps someone who is interested in the fact that just choosing one may not be adequate. We really want a representative fraction of the subset. One possibility is this one that I want to, at least 1 percent to be picked in the subset of the sample. The other one may be most stringent saying that I mean the two groups of people. One is let us say, about 0.20 percent, another is let say 5 percent, then the subset that I pick up must capture the derivative sizes.

So, the subset that I pick up, you know the one that has larger subset should have more representatives, of proportionally more representatives than the smaller subsets right. So, that is another condition that you may specify that say, if m is the size of a subset. Let us assume that again, we are only talking about the large subsets, not arbitrary subsets, not arbitrary small subsets. M is larger than epsilon n and then, the number of representatives R of m should be such that, R of m is roughly about R of m over r.

(Refer Slide Time: 15:32)



The fraction of sample chosen from that group divided by the actual size of sample, that is R is the size of the sample is roughly, can you complete that, m by n right. So, whatever is the fraction that population that, sorry that group represents of the entire

population, my subset should kind of capture that. So, this is the more stringent condition and probably is the most fair condition in the context that I post a problem ok.

So, the question really here is that how small can r be to guarantee these things. I mean there will be two different bounds, usually preferred condition 1 or condition 2. In fact, you can see that condition 1 is kind of a subset of a condition 2. If you can satisfy condition 2, then condition 1 will be also satisfied. So any idea how to we go about this? What does it look like, this problem that I posed right now? It should be something familiar, actually not such a new problem. So, what we are trying to do? We are trying to hit some of the subsets right by choosing. See what is it? Look at this figure. We are trying to, let us say color some of these points which is basically, I am picking them in the sample such that I hit all the subsets right.

So, it is hitting set problem right. It said that I am interested to hit only the large subsets. So, other subsets do not even count. So, I can simplify forget about those and pose it as a hitting set problem. So, if I pose it as a hitting set problem, then what is obviously we are solving it. Yeah, greedy great. So, if you use greedy, you can solve the problem, but let us say that I am not even looking at the any algorithm right now. I am only thinking terms of the bound of r, but greedy gives a reasonable solution in terms of that bound also. If you assume that the greedy bound is something that we can achieve because it is a constrictive solution, then do you remember what kind of bound you get. It is some logarithmic factor, somewhere right all right. So, since I am not, I am talking about a bound, let me do the following. I will do a calculation. It is a very elementary calculation I will say that choose, finally the bound will be very similar to the (()) ok.

So, choose a vertex. Now, I am talking about that graph, the hyper graph is the vertex and the subsets are the hyper graphs right. Choose a vertex with probability p equal to 1 by epsilon n to be in the sample. So, for every vertex, it is not just this quantity, there is another quantity I am going to add on it ok. So, I will actually multiply this by so I just use a, you see why I am doing it with a different color. So, I have I am going to choose every vertex to be in the sample with probability 1 over epsilon n and this is the size of the large subset, but with a slight twiddle factor with the extra log in. So, if I do that, so what is the probability that a large set, a large subset is not hit? That we do not pick any representatives. So, I am trying to answer to question 1 here, not the second question ok.

So, each vertex will be chosen with its probability, something a large set subset is not hit that probability is that none of these chosen vertices will be going to hit it. So, that is 1 minus p and they have been chosen independently.

(No audio available: 22:54-23:07)

So, 1 minus p is the probability is not going to be hit and all these n vertices, basically sorry, these epsilon n vertices are not going to be chosen right. So, look at that particular subset, the large subsets. Each of those vertices are going to be picked independently with the probability p and because you have not picked any of them, it means that all those epsilon n vertices have not been picked right. So, to the power epsilon n. So, which is now if it is a larger subset, the probability will be even smaller than this ok.

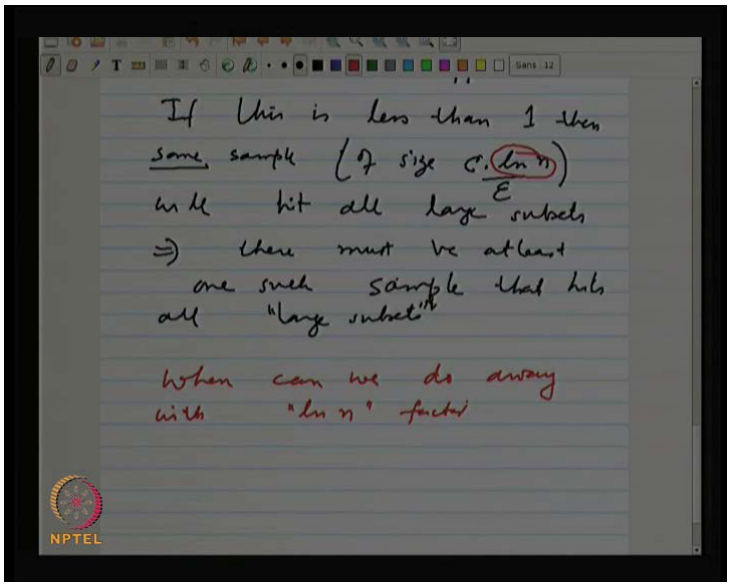
So, let me work with just you know the epsilon n. Let subset, so that is 1 minus 1 over epsilon n log n to the power epsilon n. So, we can rewrite this as 1 minus 1 over epsilon n over log n to the power epsilon n over log n to the power log n. Why did we write like this? Because this term is less than or equal to 1 over e right. So, then, this is basically 1 over, this whole thing is basically less than 1 over n right.

Now, this 1 over n, we can make it even smaller by multiplying things by a constant c or things like that. We can make this even smaller. Instead of choosing a sample probability 1 over epsilon n log n, we can multiply that. We can increase piece piece slightly to make this even smaller. Why do want to make this even smaller? Because we want to now argue this is for a specific subset, large subset right. There can be many large subsets right. So, we want to add those failure probabilities. So, you want to multiply by that and therefore, you want to make that even smaller, so that when we add all of them, the probability is still less than 1 ok.

So, all I am saying is that the probability that we miss even one among all the large subsets is less than or equal to 1 over n to the power. Now, I am writing 1 over n, but as I said we can argue that become make it even 1 over into the power c by a corresponding c there ok, but let me not just add c. If I assume that the number of large subsets is not that large, so is basically going to be multiplied by number of large subsets right. Certainly my probability of failure is no more than the sum of these probabilities ok. So, as long as I can ensure, so this quantity I want to ensure is less than 1. If I can ensure that this is

less than 1, it means there is a positive probability that some sample will hit all the large subsets. Are you with me right?

(Refer Slide Time: 25:53)



So, if this is less than 1, then some sample of size, how much? Then, the probability, if the probability, you also get a size right. Probability is 1 over epsilon n log n, so multiply that by n, that is your expected sample size right. Some sample size of size n over log n over epsilon, sorry. So, there is some constant factor. Let say, some I am just adding c some, c times log n over epsilon will hit all large subsets implying there must be at least one such subset. This is an existential proof right.

(Audio not available: 28:20-28:38)

Is just for each of the large subsets, the probability is 1 over n. So, I just add up the probability for all these subsets, then they are not, they are correlated, but I can always add it up right. The union of the events yeah. I am just, I just yeah some yeah I am just using the sample some two's bound is union of the elements that is all.

So, there must be at least one subset that hits all the sorry, one such sample that hits all large subsets. So, this is an existential proof that there must be some at least one subset, otherwise that the probability cannot be strictly less than 1. There has to be something by which we can succeed to hit all the large subsets. So, that is an existential proof. So, the thing here to note is, till now it is a very elementary calculation fact. I can draw your

attention to something that we did a while ago when we were discussing this you know things like randomized incremental construction etcetera. I did not actually complete the proof in the class. I showed the slide, but you did not go through it. Basically, this kind of a proof was not exactly this proof, but it was this kind of a proof.

Now, doing the trapezoidal maps, I said that you know I want to make sure that I choose the sample and a sample defines some trapezoids and then, the other segments are going to hit those the unsampled segments inside a trapezoid I should be able to bound. It was a calculation basically of this form. You can go back and refer to the slides.

So, one such sample is it will it all the it will hit all the large subsets. So, it gives me this number of c log n, n over epsilon. Now, may be the actual, there may be a smaller subset, who knows. There could be even small subsets by which we can manage to hit all of them, but it turns out that the people know actually looked at this problem and they have not been able to prove much better bounds. I mean it is believed to be kind of tight. You know modulo the constant factor c. So, you need about so essentially we are looking at a figure where epsilon is a parameter to the problem right. My large population is described in terms of epsilon and even if epsilon n is a not a constant, you know I said it has to be constant between 0 and 1.

What do I mean by constant? It means some fraction basically 0 1. So, all these calculations that I did, does not depend on the fact that epsilon has to be 0.1. It can be must smaller than that. Whatever epsilon is, this is parameterized on the basis of epsilon. So, this subset that we are picking is a function of epsilon and that is to be expected. The smaller the epsilon, the larger the sample has to be right because if you want to hit smaller subsets, I need more and more sample, but the thing the sort of kind of subset is a sour point is this thing ok. If it were not that, then we should be happy. In fact, you can prove that this is completely tight. I cannot do anything better than picking a sample of size 1 over epsilon. That you can argue even with a linear and just look at the points on a line and I want to hit every subset of size epsilon. I cannot do with less than so many samples. It cannot be done right but this log n is the only disturbing factor. So, that is what let too much of the development that we are talking about.

So, when can we essentially do away with the log n factor or in other words, when can I have a hitting set of size roughly about one over order of 1 over epsilon? So, now this is

the way we did the calculations, did not make any assumption about the subset system. It can be any kind of subsets, accept that we are hitting the large subsets. Now, if we you know constraint the subset system to have some properties, then we can actually do away with this and that is what this theory is about ok.

So, with that introduction, let me lead you into some definitions. Essentially, we are exploring some you know subset systems with certain kind of properties, where we can do this hitting set and finally, you know I will not today, but you know sometime next week I will end with a lecture, where I will show you that under what conditions, especially in the context of you know sets subset system in the Euclidian space, most of them actually satisfies the properties, so that we can actually you know get a small hitting set. That is what the connection to the geometry is.

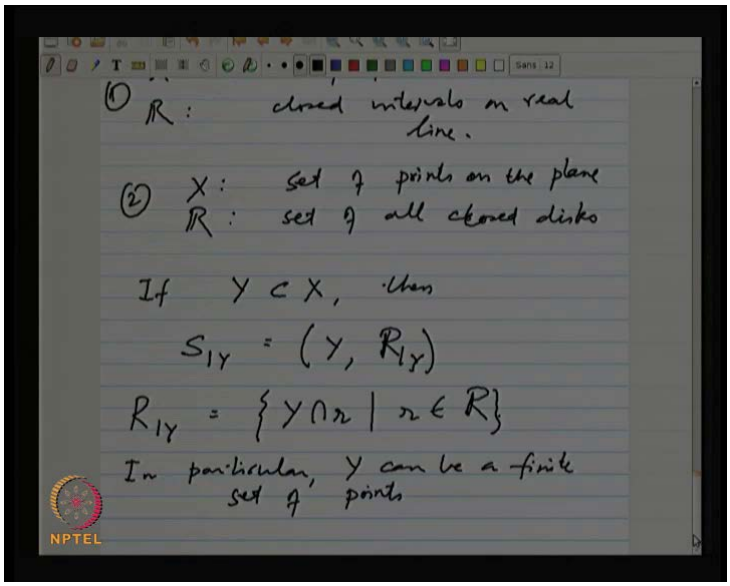
(Audio not available: 33:49-33:55)

See those are things that you know people will talk about when you are doing a linear programming formulation etcetera. So, this has got nothing do with it. It is purely a powerful thing. It is quite surprising. It is very powerful thing. What you are saying is a sufficient condition.

(Audio not available: 34:07-34:17)

So, for that, I will define you a notion about what is called a range space. Now, range spaces have already been introduced in the context of range searching right. So, I will kind of again recap those definitions.

(Refer Slide Time: 34:31)



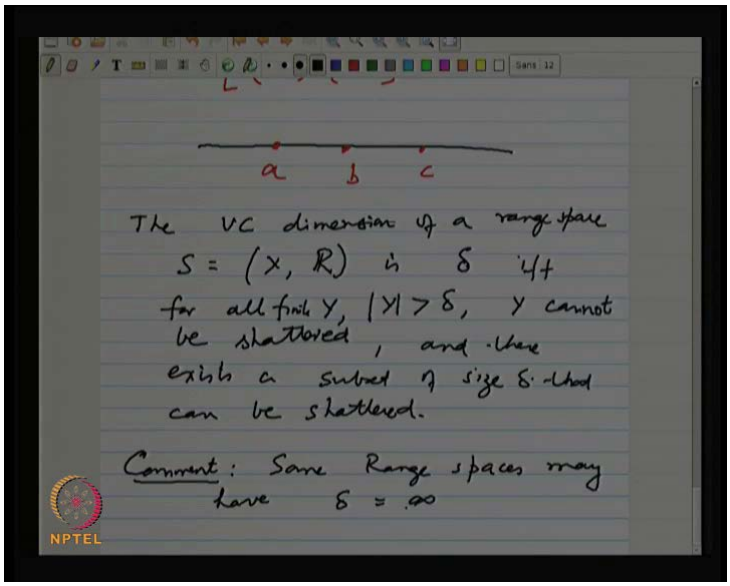
So, range space, let us say S is a topple X, R, where X is set of points may be infinite and R is a family subsets X. Normally, we will be talking about you know finite set of points. So, one way to specialize that is to talk about restriction. So, just to give an example as what I mean by this. So, my X could be set of points in the line and R could be closed intervals. That is the range space or this is one example. X could be let us say set of points on the plane. So, these are clearly examples of infinite systems.

One way of restricting or making things finite is talking about restricted range spaces. So, if you say if Y is a subset of X, then the range space restricted to Y is defined as Y and the range is restricted to Y. What is the range restricted to Y?

(Audio not available: 37:23-37:40)

So, in particular Y can be finite. So, if Y is a finite set of points, so everything is finite right. Even in the family of subsets is also finite. This is an absolutely you know general and useless definition right. What we are really looking at is a property of this range spaces and that property is the following.

(Refer Slide Time: 38:26)



We say that a finite subset of point Y is shattered if the range space restricted to Y. So, the range is the number of the ranges restricted to Y is a power set of Y. Take example.

Let say example 1, points on a line and intervals right. Consider Y is a set of 2 points. Here is my line; my Y is these 2 points. So, what is the size of the range is restricted to Y. I could have an interval containing this, I can have an interval containing this, I can have an interval containing this and I can have an interval not containing any other points. So, number of ranges restricted to Y is the power set of Y. Maybe I should actually use both sides ok.

Now, let us see how about 3 points. Can this be shattered? Can we have all possible subsets, where my subsets are defined by intervals? Let us call them a, b, c. These are 3 points. Yes, we cannot help also picking up b right. So, it cannot be shattered. So, not just these 3 points, but no 3 points can be shattered by that range space ok.

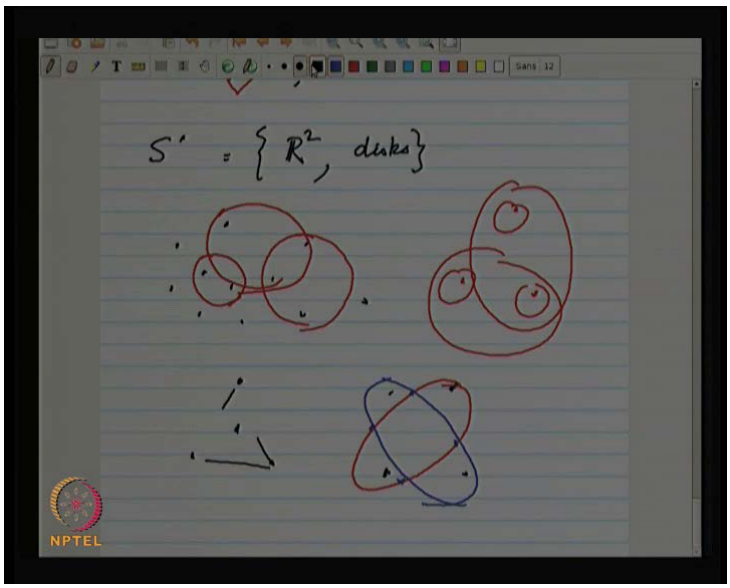
So, that is what we are looking at. So, this is basically which gives us the definition of VC dimension. So, the VC dimension of a range space S is equal to X, R is, let us say delta if and only if for all Y, such that Y finite Y is greater than delta Y can be shattered. I mean to complete the definition one should write that you know if for a size and there exits the subsets of size delta, sorry this cannot be shattered. So, it is a largest cardinality subset that can be shattered, that is the VC dimension.

So, for example that we just had the VC dimension is 2. So, some essentially you know comment. Some range spaces may have infinite VC dimension or it is a delta equal to infinity, that is, it is not a bounded VC dimension. Think about an example. Here is the example.

(Audio not available: 43:57-44:05)

So, consider this range space S conf R. All points on the plane and the ranges are all convex subsets. So, I take, let us say so I take an arbitrary set of points, sorry not an arbitrary. I take a set of points that were on the convex. So, I have to argue that the VC dimension at least n, for any n let say ok.

(Refer Slide Time: 44:08)



So, I take all the points in convex position. The points of R square, they said that I am looking at they all are in a convex position and my ranges are convex subsets all right. I claim that I can capture exactly any subset of this is in convex subset. Why? Because any subset of points in convex position are convex.

I want these 3, I can get these 3 whatever you want right. So, this is one example of a range space with infinite subset. Now, how do you actually go about proving finite VC dimensions? I mean how do you prove? Here is I am giving a subset system, sorry range space and how do you actually argue what is the VC dimension, may or may not be easy right. So, for example, just consider this one. So, here is a range space S prime, which is

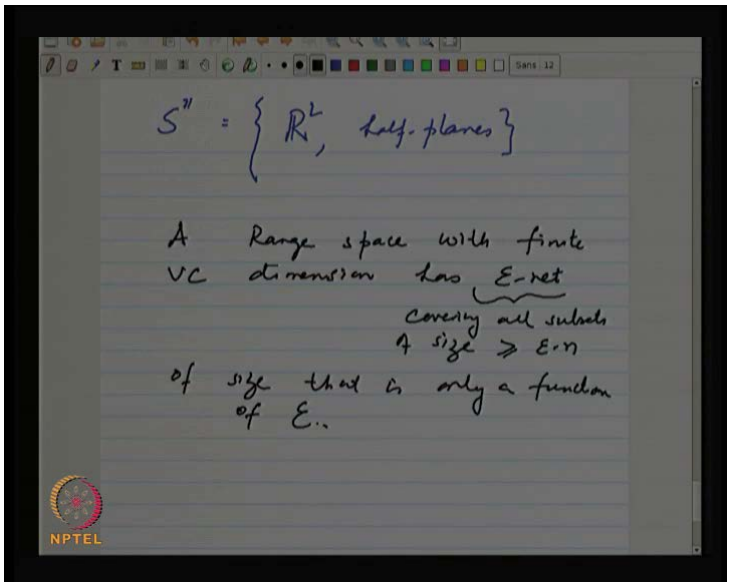
suppose S points on the plane and disks. What I mean is that here points and my subset have to be busy. Disks of any size will match. That is it.

What do you think is a VC dimension of this? Is it, let us try to argue is it at least two? At least 2 ok. Is it at least 3? So, see as long as there is one subset that you can shatter, as long as there is one subset that you can shatter it is good enough. You do not have to shatter every subset of size 3. So, if I take 3 points, I should be able to draw disk containing these, you know these or these. I should be able to find subset of points. So, I claim here at least 3 and to say that, it is no more than 3, then we have to argue that if there are 4 points, there is no way that it can be shattered right.

So, here is a quick proof, hand going proof little bit. So, the 4 points may not be in convex positions. So, one point could be inside the triangle. Can you shatter this? No way right because if I have these 3 without excluding this one, I cannot do it. So, then no point is in the convex combination of the other 3 points.

So, then if you take now these 4 points, then you should be somehow able to argue the following that I want to include these 2 points and not the other 2 points and in other cases, I may want to include these 2 points, but not the other 2 points right. If they were so, then I will have 4 intersections which is not possible between them there cannot be 4 intersections right. So, it is not possible. So, the VC dimension of this range space is 3 right and you know things can become quite quite difficult. So, one one very interesting VC dimension is points, so this is what you can try. If do not succeed, we can talk about next time.

(Refer Slide Time: 49:10)



So, this range space let us say S2 S is double prime points in the plane and half planes. Try that later. Anyway, so I should probably end today with just this remark that, it turns out that VC dimension is a very crucial property of many subset systems and what will eventually end up proving or that is what our goal is to show that this is the result that we will basically prove ok.

About the course in the next 2 lectures, that a range space with finite VC dimension has epsilon net. So, the epsilon net is essentially the cover that I was talking about. I want to cover every subset with size at least epsilon times n. So, that is called epsilon net. This is basically covering all subsets of size greater or equal to epsilon times n. So, range space specified has epsilon net of size that is only a function of epsilon. So, we do not have any log n factors with that. If my fraction is you know whatever 10 percent, then it is purely, that fraction is not 1 over epsilon. So, that would be an ideal. So, that what turned out to be close, not very far away from that. You know it will be some polynomial in 1 over epsilon. That is what we end up proving. It will take some work. It will take me a couple of lectures through that.