

**Scalable Data Science**  
**Prof. Anirban Dasgupta**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Gandhinagar**

**Lecture -14 a**  
**Original**

Welcome to the course on Scalable Data Science. My name is Anirban. I am from IIT, Gandhinagar. Today's topic is going to be on random projections.

(Refer Slide Time: 00:28)

**Curse of Dimensionality**

- Refers to rise in complexity of related issues with increasing dimensions
  - E.g. failure of intuition from low-dimensional cases: if  $x \sim N(\mu, I_d)$  then  $x$  does not lie "near"  $\mu$

$x_i \sim N(\mu_i, 1)$

$\|x - \mu\|^2$

$\mu$

The slide features a normal distribution curve on the left, a sphere with a point on its surface on the right, and a small video inset of the professor in the bottom right corner. The bottom of the slide contains logos for IIT Gandhinagar and NPTEL ONLINE CERTIFICATION COURSES.

So, before we start talking about random projections, let us take a look at this a called the curse of dimensionality. Since you are taking the course ah, I do not need to sell the concept of large data to you or even the concept that data is not only large most of the data that we collect has a very high dimensionality right.

So, imagine for instance that you are creating a classifier for individual web pages or for individual documents, right. So, the first task at hand would be to represent a document or a webpage as a vector. And this vector then is inherently very high dimensional. For instance, one trivial way this vector you could obtain this vector is just by seeing, it is just by taking let us say the English dictionary, and then seeing which words are in the document or not.

So, the representation of a document is then this binary vector for instance, right whose dimension equals the number of words in let us say the English dictionary, right which itself is pretty large. And if it is a web page for instance, you might need to add in html tokens, you might need to add in by grams or trigrams which are basically tuples or mean pairs or tuples or tuples of words of consecutive words ok.

So, there is no question that most of the data that we face has dimensions that we do not really I mean have much intuition about. For instance, the problem is that when data becomes high dimensional a lot of the intuition that we bring from low dimensional sort of pictures or images or intuitions completely breakdown right. And a let us see one simple example of this right. So, imagine that you are generating data from a normal distribution ok.

So, if it was one dimensional, then you could easily draw a sort of a picture of this which is the standard bell curve right; so, which says that; this would be the bell curve in one dimension ok. And if you look at this, you can easily see that most of the data is really concentrated around the center  $\mu$  ok. Most of the data lies very close to  $\mu$  right, but now imagine that you were generating data not in one dimension, but in some dimension  $d$  right.

So, what you are generating? So, what you start with is a center  $\mu$  that is a point in  $\mathbb{R}^d$ . So, think of  $t$  as something pretty large right. But it will turn out that we will see this the non-intuitive phenomenon that we talk about even for pretty small  $d$  even for  $d$  equal to 10, 20 etcetera ok. So now, we are generating a  $d$  dimensional vectors ok. And we are generating them as per the following distribution.

So, each entry let us say that it is the  $i$ th entry  $x_i$  is nothing but a normal variable whose center is  $\mu_i$  which is the  $i$ th coordinate of the center  $\mu$ , and whose variance is 1 ok. And  $x_i$ 's are independent of each other ok. So, one way to sort of write this distribution is to say that  $x$  comes from a multivariate normal distribution right. So, the center of this multivariate normal distribution is this vector at the point  $\mu$  which is a  $d$  dimensional vector. And the covariance matrix of this happens to be  $I_d$  right. Because the variance of each of the coordinates was 1 and the covariance of 2 different coordinates is 0. Therefore, the entire covariance matrix turns out with identity matrix.

So now the question is, that if you look at this d dimensional space right, do most of the data points lie very close to mu or not right. Because that that was what the intuition that we took from the 1 dimensional case that most of the data points in the one dimensional case was very close to mu right. And if you want to calculate this if you want to sort of see an intuition for this, right we need to let us try to calculate the distance x minus mu square right. So, for a so far a let us say let us say x is a sample from this distribution, and I want to see how does this random variable behave this x minus mu square right.

(Refer Slide Time: 05:33)

$\|x - \mu\|_2 \approx \sqrt{d} + o(\sqrt{d})$  *why x*

### Curse of Dimensionality

- Refers to rise in complexity of related issues with increasing dimensions
- E.g. failure of intuition from low-dimensional cases: if  $x \sim N(\mu, I_d)$  then  $x$  does not lie "near"  $\mu$

$x_i \sim N(\mu_i, 1)$

$x - \mu \sim N(0, I_d)$

$$E\|x - \mu\|_2^2 = E \sum_{i=1}^d (x_i - \mu_i)^2 = d$$

*Diagram: A sphere in d dimensions with radius  $\sqrt{d}$  and center  $\mu$ .*

*Inset image: A man reading a book.*

IIT Gandhinagar  
Indian Institute of Technology Gandhinagar

NPTEL ONLINE  
CERTIFICATION COURSES

So, this you can imagine is then the random variable that captures the distance between the data point generated according to this distribution and the mu right. So, if so, if most data points are close to mu this particular random variable that we wrote the x minus mu square right, should be pretty small, and if the if x is more or less close to mu, but is it the case. So, you see this let us try to calculate the expectation of this random variable ok. So, this is a scalar quantity of course, because there is a distance.

So now, because x i x has x has x is of multivariate normal distribution whose center is mu, x minus mu is a multiple is also multivariate normal distribution whose center is 0, and the covariance matrix remains the same all right. So, the center is 0 the covariance matrix is id; which means that if I write down if I write this down. So, I have just used the definition of the other l 2 norm.

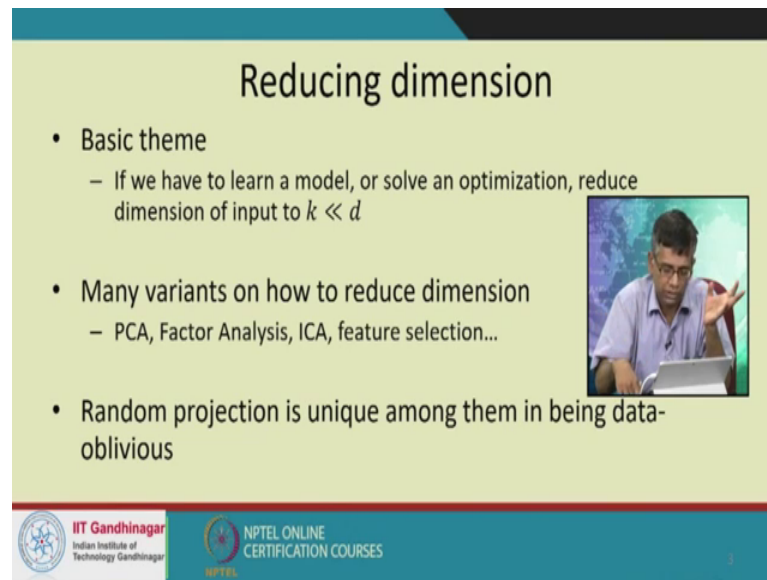
So, the  $L_2$  norm between  $x$  and  $\mu$ , the square of that is nothing but the sum over all  $i$  equal one to the  $x_i - \mu_i$  squared right. And because  $x_i - \mu_i$  by the definition that we have given out here,  $x_i - \mu_i$  again is a random variable, now with expectation 0 and with variance 1. Therefore, this is nothing but the sum of the variances right which is  $d$  because each of them is variance one right. So, what that means is that, and it is not really hard to show that not only is the expectation  $d$  right. The variable the random variable  $x - \mu$  squared is actually pretty close to its expectation; which means that what we will see is  $\|x - \mu\|_2$  will be more or less  $\sqrt{d}$  plus lower order terms or the smaller  $\sqrt{d}$  ok.

So, here we got  $d$  because we were taking the square of the distance, and when we do away with the square we get a square root  $d$  right. So, what does the statement mean? That the  $x - \mu$  that which that for most of the data points, which means that with high probability, over the choice of  $x$  right,  $\|x - \mu\|_2$  of that will be more or less  $\sqrt{d}$  plus small order of  $\sqrt{d}$ . What does that mean geometrically? Geometrically what it means is that, if you look at the ball of radius  $d$  around  $\mu$  right, most of the data points will be on the surface of this ball right and on a very small annulus around it. And the width of the annulus is small of  $\sqrt{d}$ ; it is actually something like  $d$  to the one 4th or something ok.

So, this is already pretty counterintuitive right, because if most of the ball is hollow, if you if you imagine that most of the ball is  $\mu$  lies at the center of the ball, most of the ball is hollow and almost all the data points lie on something that is very close to the surface of the ball right. That is not what we were expecting if we when we look at the one dimensional plane right.

So, since I mean and phenomena like this are fairly common right. So, the point that I am trying to make here is that, dealing with high dimension is a non-trivial thing. And it is not just a question of our intuitions breaking down.

(Refer Slide Time: 09:14)



**Reducing dimension**

- Basic theme
  - If we have to learn a model, or solve an optimization, reduce dimension of input to  $k \ll d$
- Many variants on how to reduce dimension
  - PCA, Factor Analysis, ICA, feature selection...
- Random projection is unique among them in being data-oblivious

The slide includes a small inset image of a man in a light blue shirt speaking, and logos for IIT Gandhinagar and NPTEL ONLINE CERTIFICATION COURSES at the bottom.

There are a number of other issues. For instance, if you have, if you I mean if you are trying to build a model for instance right. And the number of features is something that is close to the number of examples that you have, then your model is bound over fit unless you very strongly regularize right. So, and one method of regularization is essentially cutting down the number of dimensions. And this is what we will try to do right.

So, our basic intuition that we follow, that we will follow is that if the number of dimensions is too large for me to handle. And we have to learn a model or solve an optimization problem; which depends on the number of dimensions. Then one of the very common tricks in machine learning is to reduce the number of dimensions to a manageable quantity, something like  $k$  we will we will typically call it  $k$  which is going to be much less than the number of original dimensions that were presented with.

And then we will work with the data that is a dimension  $k$  ok. And this thing will come comes over and over and over back again in machine learning right. For instance, this is also and this is known by various names known as finding out the intrinsic dimensionality of the data, I mean being able to do feature selection effectively and so on and so forth right. And depending on the particular variant that you are interested in there are different ways of actually formalizing how to do this reduction.

You have seen singular value decomposition before or principal component analysis before that is one way of doing this. You have also seen factor and you might have also

seen factor analysis before only or independent component analysis or feature selection. These are all different formalizations of the same theme that we want to reduce the number of dimensions and  $d$  of the data.

So, random projection is a unique one among them. And it is unique in the sense that it is possibly the only one that is data oblivious. So, we have 2 properties. Number 1, is that we will have a, we will make the dimension reduction data oblivious in some sense. That is how we do the dimension reduction is completely oblivious to the data set represented with. And secondly, we will make sure that we have very nice very tight interesting guarantees on what we can say are on the structure of the original data versus a structure on the reduced data ok.

(Refer Slide Time: 11:51)

**Preserving pairwise distances**

$x_1, x_2, \dots, x_n \in R^d$ , Euclidean space

Want  $x'_1, x'_2, \dots, x'_n \in R^k$ , possibly also Euclidean,  $k \ll d$

with a guarantee that  $|x'_i - x'_j| \approx |x_i - x_j|$  for every pair  $(i, j)$ ?

The slide includes a diagram with red arrows and points. On the left, a solid red arrow connects two points  $x_i$  and  $x_j$  in a higher-dimensional space. On the right, a dashed red arrow connects their corresponding points  $x'_i$  and  $x'_j$  in a lower-dimensional space, illustrating the goal of preserving pairwise distances.

**IIT Gandhinagar**  
Indian Institute of Technology Gandhinagar

**NPTEL ONLINE**  
CERTIFICATION COURSES

So, what does this guarantee? This guarantee is going to be as follows, that suppose we have points  $x_1, x_2, \dots, x_n$  in  $R^d$  which are Euclidean space. And suppose I want a representation of these points as  $x'_1, x'_2, \dots, x'_n$  where  $x'_i$  is the representation of the point  $x_i$ . And they should lie in some dimensional space  $R^k$ . Because I started with Euclidean distance I let us say we also want to preserve we are we also want the distance function of this of this target space to be also Euclidean, and we also want  $k$  to be much smaller than  $d$  ok.

So, the guarantee the kind of guarantee that we want is that the distance between  $x_i$  and  $x_j$  should be more or less equal to the distance between  $x'_i$  and  $x'_j$ .

$j$  prime. And this should ideally hold for every pair  $ok$ , but it means is that imagine that you started with a with a kind of with a kind of tetrahedron in space right. And you want to reduce it to much smaller number of dimensions right.

And you want to make sure that the distances between all the points are preserved. Right now it is easy for it was it is not too hard for you to see that if you start with a tetrahedron in 3 dimension and you want to reduce it to 2-dimension right. And want to preserve the exact distances that are not possible right; because you have to squish something or the other right.

So, this thing so, this preservation can only be done approximately. So, what is this notion of approximation? And how does that help?

(Refer Slide Time: 13:38)

**Johnson Lindenstrauss Lemma ('84)**

$\epsilon, \delta > 0, k \geq \frac{C}{\epsilon^2} \log(n)$ . There exists a linear mapping  $A$  such that for all  $(i, j)$

$$(1 - \epsilon) \|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1 + \epsilon) \|x_i - x_j\|_2$$

So, very interestingly, this something like this is fairly easy to do in a particular sense right. And this was found by and this construction originally came from 2 mathematicians William Johnson and Durham Lindenstrauss in this lemma that they gave ah, it is (Refer Time: 13:44) called a lemma in about 1984 ok. What it says is that such a construction not only does such a construction exist; you can actually find a linear mapping that achieves. This linear mapping is the same as a matrix right.

So, what it says is that that suppose you are you fix 2 parameters, let us say you fix let us say at this point, let us forget the delta. Let us fix a parameter epsilon to think of epsilon

as my error tolerance; as in how much how approximately do I want to maintain the original distances. Suppose epsilon is greater than say something greater than 0, and you start with the k that is c which is going to be remember k is going to be my size of the target my target dimension. So, k should at least be something like c by epsilon square log n.

We will decode slowly how that came about. The result by Johnson and Lindenstrauss says is, that there exists a linear mapping A such that for all pairs i j right, for all pairs ij the distance the l 2 distance between Ax in Ax j lies within A 1 plus minus epsilon guarantee.

Now, 1 plus minus epsilon factor of the l 2 distance between x i an original x i and x j; so, supposing this was the original distance x i minus x j. This is the value of this is the real line. So, the distance between Ax i and Ax j right and this is 1 minus epsilon x i minus x j and this is 1 plus epsilon x i minus x j right. So, what it says is that? The distance between x i x j the projected points lies in this in this interval, and this happens for all pairs ij ok.

(Refer Slide Time: 15:35)

**Johnson Lindenstrauss Lemma [JL84]**

$\epsilon > 0, k \geq \frac{1}{\epsilon^2} \log(n)$ . There exists a **linear mapping**  $A$  such that for all  $(i, j)$

$$\rightarrow (1 - \epsilon)|x_i - x_j| \leq |Ax_i - Ax_j| \leq (1 + \epsilon)|x_i - x_j|$$

Can show that this actually happens with a high probability, also that such matrices  $A$  are easy to construct (randomized)

The slide includes logos for IIT Gandhinagar and NPTEL ONLINE CERTIFICATION COURSES, and a small video inset of a man speaking.

So, we can actually show that, if I choose this linear mapping if this is possible to choose this mapping A in a random manner, such that this actually happens with a high probable probability. So, what that means, is that not only is it not only does it exist such a mapping, it is also easy to find such a mapping right. Because once we give a random



construction for it for specific types of random constructions that we will see right, that we will be able to guarantee that this particular property holds with high probability ok. So, this will give a randomized algorithm for this for this particular problem, for this particular matrix construction ok.

So, the constant  $c$  that you see up here that I had forgotten to write is in practice fairly small right. It is sometime, I mean in something like 6 or 7 is what you need to prove in theorems. And in practice it comes out to be smaller than that.

(Refer Slide Time: 16:37)

So, just to sort of harp a little bit on the intuition of this problem. Why are we interested about pairwise distances right? So, this is the picture right, that we start with let us say 1 2 3 4 5 or even more dimensions in line in my original space. So, this is my original space, and these are the points this blue are the points right. For any of these points  $x$ , we multiply  $x$  by  $A$  right, and then we look at; so, this is the image of the point  $Ax$  right. And this target dimension this target dimension  $k$  is much smaller than the is potentially much smaller than the dimension of the of the original space right.

In fact, it only depends on the error epsilon and on the number of points that you are interested in right. It does not depend on the on the original dimension right. What it says is you look at all the pairwise distances here, all the pairwise distances here. They are more or less maintained in my target dimension up to this factor of  $1 \pm \epsilon$ . Actually maybe it should have been this point ok. So, why is this important? This is

important for instance imagine that the original points are clustered in some way. Then preserving the pairwise distances also implies that we are preserving a cluster structure. Similarly, if the original points are separable, right for instance we had we have class 1 and we have class 2.

And there is and they happen to be nicely separated from each other by some by some hyperplane right. Then preserving the pairwise distances also means that the resulting problem is also separable; which means that I can actually learn a classifier on this on this target dimension very easily ok.

So, the important point I want you to notice that the input dimension does not come into the bound at all right. Only the number of pairs that we want to preserve or the number of points whose distance we want to preserve comes into the picture. And we will see why I am talking about pairs right.

(Refer Slide Time: 18:48)

The slide is titled "Intuition" and features two diagrams. The left diagram shows a 2D coordinate system with several points and lines connecting them, with a handwritten red  $\binom{n}{2}$  next to it. A blue arrow points to the right diagram, which shows a 3D coordinate system with the same points, with a handwritten red  $\frac{1}{\epsilon^2} \log \binom{n}{2} \sim 2 \log n$  next to it. Below the diagrams, the text reads "Preserving pairwise distances preserves the inherent geometry of the points" and "Input dimension does not come into bound" (the latter is circled in red). At the bottom, there are logos for IIT Gandhinagar and NPTEL Online Certification Courses, and a small video inset of a man speaking.

So, if there are  $n$  choose 2 pairs right, then it should be log of  $n$  choose 2 which is basically the same as to log  $n$  right. That should be that should come into my target dimension. And of course, there is the  $1$  over epsilon square factor; where epsilon is the error parameter.

(Refer Slide Time: 19:08)

Why is it useful?

- Used in
  - learning classifiers efficiently
  - randomized numerical linear algebra: matrix factorizations and regressions
  - a number of streaming algos are essentially random projections
  - LSH for L2, other near neighbour data structures
  - compressed sensing

$y = Px$

The slide features a list of applications for randomized numerical linear algebra. A handwritten equation  $y = Px$  is shown in red, with a vertical vector  $y$  on the left, a matrix  $P$  in the middle, and a horizontal vector  $x$  on the right. The slide also includes logos for IIT Gandhinagar and NPTEL Online Certification Courses, and a small photo of a man looking at a laptop.

So, why is it useful? As I have already hinted that it is potentially one of its uses and we will see very concretely a sort of example later in the course is that it will help us learn classifiers very efficiently right. That instead of learning classifiers in my original dimension very high dimension we will do a projection and then we will learn the classifiers in this resulting dimension. We will also see a lot of examples in what we in what is known as now known as randomized numerical linear algebra right.

And here for instance, we want to we have a huge data set in which for instance we undergo matrix factorization. Why? Because maybe people (Refer Time: 19:43) recommender system. We want to solve a regression problem, and in all of this a symmetric would be to reduce the target dimension, solve the problem there, approx and thereby get an approximate solution in the original space with high confidence.

So now if you go back we have seen a lot of streaming algorithms. We have seen count min sketch we have seen count sketch and so on and so forth. If you go back and think about this, you will see that a number of these are essentially random projections right. You have also seen a bunch of algorithms of locality sensitive hashing right. And we have also discussed the variations of kd tree or rp trees and so on. And a lot of these as we will see in today also in today's class and in next class, a lot of these again the core idea behind that is this distance preservation notion, and then we are doing things on top of it right.

There is a huge literature in communications as well as in computer science called compress sensing. The idea there is that that suppose we have a signal, and we can we cannot observe the signal directly, but we can observe it through it is interaction through some through some interaction matrix; that is, given the signal  $x$  we cannot observe  $x$  directly, but we can make observations of the form  $p$  times  $x$ ; where  $p$  is called observation matrix right. And we might have assumptions that say that  $x$  is sparse right. And now the problem is that given  $A y$ . So,  $x$  potentially is very, very large dimension, but is sparse small number of non 0's.

By making a small number of observations  $y$  which is putting which is ideally only proportional to the number of non 0's of  $x$  right; if we want to reconstruct  $x$  right. So, the number of observations should be should be a function of more the number of non 0's of  $x$  and the dimensionality of  $x$  ok. This is the compress sense in literature and this very intimate connection between compress sensing and random projections ok.

(Refer Slide Time: 21:48)

The slide is titled "How to create such a matrix". It contains the following content:

- $R_{ij} \sim N(0, 1)$  independently
- $\frac{1}{\sqrt{k}} R \in \mathbb{R}^{k \times d}$  is the required matrix

Handwritten notes in red ink include:

- A diagram of a matrix with dimensions  $k$  (rows) and  $d$  (columns) indicated. A specific entry is labeled  $ij$ .
- Equations:  $E x_i^2 = 1$  and  $E \|y\|_2^2 = k$ .

The slide footer includes the logos for IIT Gandhinagar (Indian Institute of Technology Gandhinagar) and NPTEL ONLINE CERTIFICATION COURSES, along with a small video inset of a speaker.

So, we talked a lot about why this is potentially useful. But how do we create such a matrix right. So, here is a very easy way of creating matrix a bunch of such algorithms have been given by researchers, we look at a easy way about which you can actually prove that it works. So, the way we will create the matrix is imagine that we start with the empty matrix which is of size  $k$  by  $t$ , and we and we are going to fill this up. So, how do we fill this up? It is very simple, you go to the entry  $ij$ ,  $ij$ th entry and in for filling up

the  $ij$ th entry, you sample a random variable from  $N(0, 1)$  basically from the standard Gaussian. You take that value and you set that as the value  $R_{ij}$  and you keep on doing this for all the entries.

So, independently you sample  $k$  times  $d$  entries and you fill use that to fill up the matrix, that is it ok. So, this is my random projection matrix. And so, one other thing we need to do is that we need to normalize it by this factor  $1/\sqrt{k}$ . And we will see and it is easy to see why right. So, if you notice what the expected norm of a row is; so, take the row the first row right. Each of these are our samples from  $N(0, 1)$ .

And therefore, each of them have a variance 1, which means that expectation of  $x_i^2$  where  $x_i$  is the  $i$ th entry of row is 1 and therefore, the expected norm of the first row is equal to  $k$  right. And I want it to be 1, right. Because if it is going to act as a projection matrix, right all if it is going to add as a projection matrix I need it I basically if it is going to act as a norm preservation matrix. I need the norm of each column each column to be one not the row maybe I was talking about a row before. So, I need the norm of each column to be 1 right.

(Refer Slide Time: 23:58)

**Why does this work?**

JL lemma:  $k \geq \frac{C}{\epsilon^2} \log\left(\frac{1}{\delta}\right)$ . For the previous matrix  $A = \frac{1}{\sqrt{k}}R$ ,  $|x| = 1$

$\epsilon, \delta > 0$   
 $\Pr[(1 - \epsilon) \leq |Ax| \leq (1 + \epsilon)] \geq 1 - \delta$

$x_i - x_j$   
 $\binom{m}{2}$

- Enough to show this since linear mapping
- Previous claim follows by taking union bound over all pairwise distances

The slide also features a video inset of a man speaking and logos for IIT Gandhinagar and NPTEL.

So, therefore, I am dividing by  $1/\sqrt{k}$  ok. So, why does this work? In order to see why this works, we need to look at we need to prove a smaller result first. And this is, but this is really the core of the of the of the theorem right. And this is going to JL lemma. So, what does he say?

What it says is that that suppose we set  $k$  to be some at least  $c$  by  $\epsilon$  square log of  $1$  over  $\delta$ ; where  $\epsilon$  and  $\delta$  are 2 non negative quantities. Then for the previous matrix that I created let me call that  $A$ , and let us take any vector  $x$  that has 2 norm 1, then  $Ax$  the probability that  $x$  lies in the interval  $1 - \epsilon$  and  $1 + \epsilon$ . So, the 2 norm  $x$  of  $Ax$  lies in the probability that the 2 norm of  $x$  lies in the interval  $1 - \epsilon$   $1 + \epsilon$  is at least  $1 - \delta$  ok.

So, see notice a couple of things. First of all, I am stating the theorem only for the 2 norm of  $x$  equal to 1. And this is enough because this is a linear mapping because for any other  $x$  i can always normalize it by the by that norm of  $x$  right. And once I multiply by  $A$ , I can again multiply back by the norm and so on. So, having such a stating a theorem for the unit on vectors is enough right. And also notice that if I can actually show this, if I can actually show this particular lemma right. Then in order to show the previous Johnson Lindenstrauss theorem, we needed to look at all pairs  $x_i - x_j$  right, and there are  $n$  choose 2 pairs. So, I needed to take union bound over all this space and I would be done ok.

So, what we will do in this class in today's in this lecture is just see the proof of this and then finish ok. So, why does this work? Let us just do a proof sketch a brief proof sketch of this that.

(Refer Slide Time: 25:53)



### Proof Sketch

$$Ax = \frac{1}{\sqrt{k}} Rx = \frac{1}{\sqrt{k}} (Y_1, \dots, Y_k) \quad Y_i = \sum_j R_{ij} x_j$$

$$|Ax|^2 = \frac{1}{k} \sum_i Y_i^2, \text{ we are interested in the distribution of this}$$

2-stability: If  $p \sim N(\mu, \sigma)$ ,  $q \sim N(\alpha, \gamma)$  are iid,  $p + q \sim N(\mu + \alpha, \sqrt{\sigma^2 + \gamma^2})$

$$Y_i = \sum_j R_{ij} x_j, \quad Y_i \sim N(E[\sum_j R_{ij} x_j], \sigma(\sum_j R_{ij} x_j)) \sim N\left(0, \sqrt{\sum_j x_j^2}\right) \sim N(0,1)$$

Suppose we look at the vector  $Ax$  right. And the vector  $ax$  is nothing but  $1/\sqrt{k}$  times  $x$  which is nothing but  $1/\sqrt{k}$ , let me call the entries this is a  $k$  dimensional vector  $Y_1$  to  $Y_k$  and  $Y_i$  is  $\sum_j R_{ij} x_j$ . So, this is the definition of  $Y_i$ . So, what we are interested in; is the 2 norm of  $ax$  square which is  $1/k \sum_i Y_i^2$ . And we are interested in the distribution of this.

So, the thing that comes to rescue is something known as the 2 stability of the normal distribution, which is as follows. That if  $p$  and  $q$  are 2 normal random variables and  $p$  happens to be from  $N(\mu, \sigma^2)$ ; that is that is the mean is  $\mu$  and the standard deviation is  $\sigma$ .  $Q$  happens to be from  $N(\alpha, \gamma^2)$ , the mean is  $\alpha$  and the standard deviation is  $\gamma$ . Then  $p + q$  has mean  $\alpha + \mu$  plus  $\alpha$  the sum of the means. And the variance is  $\sigma^2 + \gamma^2$  which means standard deviation is square root of that and furthermore it is a normal distribution. It is normally distributed random variable with this with this particular mean and variance ok. So, this is what we will use.

So now, it is easy to see that each  $Y_i$  is really the sum of normal random variables. Because  $R_{ij} x_j$  are normal random variables. Therefore,  $R_{ij}$  times  $x_j$ 's normal random variable and therefore,  $y_i$  is also itself a normal random variable with expectation to be the sum of the expectation and the variance to be the sum of the variances of this. It takes only a little bit of calculation to see that the variance of  $\sum_j R_{ij} x_j$  is really  $\sum_j A_j^2 x_j^2$ , which is equal to 1; because I started with  $x_j$  to be norm 1 ok. So, therefore, each  $y_i$  is really distributed as  $N(0, 1)$  right.

(Refer Slide Time: 27:47)

*z*




### Proof Sketch

$E\left[\frac{1}{k}\sum_j Y_i^2\right] = 1$ , also  $kZ = \sum_i Y_i^2$  is  $\chi_k^2$  distributed

We now apply Chernoff style tail bounds to show concentration of  $kZ$  around the mean.

$$\Pr[Z > 1 + \epsilon] \leq \Pr[\exp(tkZ) > \exp(tk(1 + \epsilon))] \leq \frac{E[\exp(tkZ)]}{\exp(tk(1 + \epsilon))}$$
$$\leq \frac{\prod_i E[\exp(tY_i^2)]}{\exp(tk(1 + \epsilon))} \leq \dots \leq \exp(-k\epsilon^2 C)$$

Missing steps involve the MGF of the  $\chi_k^2$  distribution and algebra

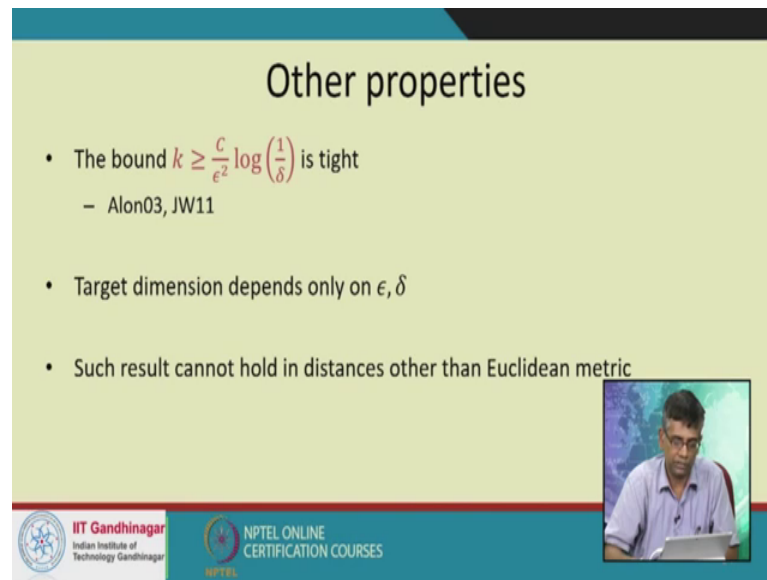


And now, therefore, the expectation of this quantity  $\frac{1}{k} \sum_j Y_i^2$  which is which I was defining to be  $z$  is also one. Furthermore, I can also say that  $k$  times  $z$  is something like a chi squared distribution with  $k$  degrees of freedom right; because the square of a normal distribution is a chi square distribution.

And beyond this we just need to apply chain of bounds. We want to show that probability of  $z$  bigger than  $1 + \epsilon$  is small we apply we apply the standard steps that we did in sort of showing chain of bounds. And it is again really a chain of style bound, but it is a little more complicated. Because it involves the moment generating function of the chi square distribution and some algebra to show that the property that  $z$  is bigger than  $1 + \epsilon$  is that most  $\exp$  of minus  $k$  epsilon square  $c$  ok.



(Refer Slide Time: 28:45)



The slide is titled "Other properties" and contains three bullet points. The first bullet point states that the bound  $k \geq \frac{c}{\epsilon^2} \log\left(\frac{1}{\delta}\right)$  is tight, with a reference to Alon03, JW11. The second bullet point states that the target dimension depends only on  $\epsilon, \delta$ . The third bullet point states that such a result cannot hold in distances other than Euclidean metric. The slide also features a small video inset of a man speaking and logos for IIT Gandhinagar and NPTEL.

### Other properties

- The bound  $k \geq \frac{c}{\epsilon^2} \log\left(\frac{1}{\delta}\right)$  is tight
  - Alon03, JW11
- Target dimension depends only on  $\epsilon, \delta$
- Such result cannot hold in distances other than Euclidean metric

IIT Gandhinagar  
Indian Institute of  
Technology Gandhinagar

NPTEL ONLINE  
CERTIFICATION COURSES

So, that is it really I mean, and the proof for the lower tail bound is also the same. In next in next lecture we will discuss some other properties of the of the random projections ok.

Thank you.