

**Algorithms for Big Data**  
**Prof. John Ebenezer Augustine**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Lecture - 28**  
**Estimating Expectation of Continuous Function**

So far we have seen that using truly random bits that is bits that are uniformly at random and independent of each other, we can actually generate a lot of uniformly random bits, but bits are pairwise independent of each other. And this seems like promising for unless we can make use of these pairwise independent random bits in some way in some meaningful way that they are not of much value.

So, in this current segment, we are going to talk about how we can take advantage of these pairwise independent random bits.

(Refer Slide Time: 01:00)

*Tail Bounds for Pairwise Indep R.V.'s*

*Bad news: Chernoff bounds crucially  
rely on independence.  
Thus won't work!*

*Good news: Variance is sufficiently  
well behaved.  
Thus Chebyshev's inequality will work!*

NPTEL 25

The first thing that we are going to look into is whether the tail bounds that we have studied so far make any sense in when we consider pairwise random bits. So, let us look into that a little bit more carefully. Let us begin with some bad news.

Well, if you recall Chernoff bounds crucially rely on the independence of the individual random variables, and so this does not much hope to get Chernoff bounds to work. But all is not lost because we do have a Chebyshev's inequality and that one does not require independence as much as it requires the knowledge of the variance of the random variable. So, in this case, with pairwise independence variance is sufficiently well behaved and so we should be able to use Chebyshev's inequality.

(Refer Slide Time: 01:57)

*Chebyshev's Inequality*

Let  $X_1, X_2, \dots, X_n$  be pairwise independent random variables.

$$X = \sum_{i=1}^n X_i$$

$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$


NPTEL

So, let us see how that can work. So, we are going to try to use Chebyshev's inequality. And for that let us let us define  $X_1, X_2$  and so on up to  $X_n$  to be pairwise independent random variables. As usual we are interested in this sum of these random variables, so that is denoted by  $x$  we want to establish that the variance is clean and easily use.


So, when we want to now let us look at how to compute the variance of  $X$ , and this variance is given by this formula where we first sum over  $i$  puts 1 through  $n$  variance of each individual  $X_i$ , but these random variables are not independent of each other. And so we need to actually also add their covariance's; so in addition to the initial summation, we will also have to have a second summation over all pairs  $i$  and  $j$  the covariance of  $X_i$  comma  $X_j$ .

(Refer Slide Time: 03:02)

Chebyshev's Inequality


$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$
$$= E[(X_i - E[X_i])(X_j - E[X_j])]$$
$$= E[X_i X_j - X_i E[X_j] - X_j E[X_i] + E[X_i] E[X_j]]$$
$$= E[X_i X_j] - \cancel{E[X_i] E[X_j]} - \cancel{E[X_j] E[X_i]} + \cancel{E[X_i] E[X_j]}$$

(by linearity of expectation)



Now, let us look at the covariance term a little bit carefully. What is the covariance of  $X_i$  comma  $X_j$  formula is the expectation of  $X_i$  minus the expectation of  $X_i$  times  $X_j$  minus the expectation of  $X_j$ . So, let us expand that out we are going to get this long expression. And if we further I mean using linearity of expectation, if we takes the expectation in to each of the terms individually there is a few cancellations that take place. And we will be left with  $E$  of  $X_i$  times  $X_j$  minus  $E$  of  $X_i$  times  $E$  of  $X_j$ .

(Refer Slide Time: 03:59)

*Chebyshev's Inequality*




---


$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[x_i] + 2 \sum_{i < j} \text{Cov}(x_i, x_j)$$

$$= E[(x_i - E[x_i])(x_j - E[x_j])] = E[x_i x_j - x_i E[x_j] - x_j E[x_i] + E[x_i]E[x_j]]$$

$$= \underbrace{E[x_i x_j]} - \underbrace{E[x_i]E[x_j]} = 0$$

Equal  
(Pairwise Independence)

But notice that these random variables  $X_i$  and  $X_j$  are pairwise independent and therefore,  $E$  of  $X_i$  times  $X_j$  will be equal to  $E$  of  $X_i$  times  $E$  of  $X_j$ , so this expression turns out to equal to 0.

(Refer Slide Time: 04:18)



*Chebyshev's Inequality*

---


$$\text{Var}[X] = \sum_{i=1}^n \text{Var}[x_i] + 2 \sum_{i < j} \text{Cov}(x_i, x_j)$$

$$\Pr(|X - E[X]| \geq a) \leq \frac{\text{Var}[X]}{a^2}$$

Example

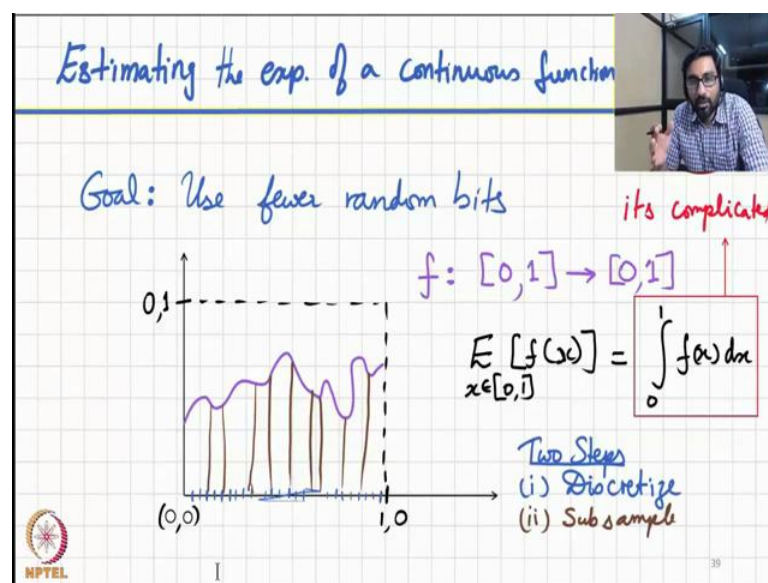
$$= \frac{\sum_{i=1}^n \text{Var}[x_i]}{a^2}$$



So, going back to the formula we had for the variance, we have summation over the

individual variances plus the summation pair wise covariance's, but these pair wise covariance's vanish we just saw that. So what we are left with is that the variance of  $x$  is simply the summation of the individual variances. And so this is very nice, because now it can be easily applied into the Chebyshev's inequality, so to apply Chebyshev's inequality, we just have to look at probability that the absolute value of  $x$  minus the expectation of  $x$  is greater than or equal to sum  $a$ .

Chebyshev's simply tells us that at most variance of  $x$  divided by a square. And we have nice simple formula for the variance of  $x$ , so we simply plug that in. So, it is nice to see how the Chebyshev's inequality can be applied and that is nice, and theoretically elegant, but let us look at an example where it is actually applied.

(Refer Slide Time: 05:38)



Let us try to estimate the expectation of a continuous function; yes, this is going to be our application. We are going to use fewer random bits; and by that I mean that we will be using fewer independent random bits than normally need it in order to be able to successfully estimate the expectation of a continuous function within reasonable epsilon delta (Refer Time: 06:13). Here is the function  $f$ ; we are interested in the function in the range 0 to 1, domains also 0 to 1. And we are shown that the function is continuous, and differentiable in the range 0 comma 1.

What we want to do is compute the expectation of  $f$  of  $x$ , where  $x$  is arranging from 0 to 1. And we all know the formula for that is simply the integrating  $f$  of  $x$   $dx$  from 0 to 1, but we do not know anything about  $f$  of  $x$ ,  $f$  of  $x$  can be a fairly complicated function. All we know is that it is differentiable and is continuous. So, this expression might not be easy to work with. So, we need to estimate this quantity. So, here is how we do it. We apply two important steps here.

Firstly, we discretized the region 0 to 1, and then even when we discretize it, we get a lot of discrete points so we really cannot compute  $f$  of  $x$  in all of those discrete points. Instead, what we do is we sub sample from those discrete points; and only compute  $f$  of  $x$  in those sub sampled points and then use that to estimate this inter graph.

(Refer Slide Time: 07:45)

The slide is titled "Discretize" in blue handwriting. It features a number line from 0 to 1 with discrete points marked. A point is labeled  $i/2^n$ . To the right of the number line, the inequality  $0 \leq i \leq 2^n - 1$  is written. Below this, a formula for the expectation of  $f(x)$  is shown in a red box:  $E[f(x)] \approx \frac{1}{2^n} \sum_{i=0}^{2^n-1} f(i/2^n)$ . A note with an arrow pointing to the formula states: "Requires the derivative of  $f$  to be bounded by a small const." The NPTEL logo is visible in the bottom left corner.

So, let us look at it a bit more carefully. So, first of all what we are doing is the first step is discretization. So, we have the domain of the function ranging from 0 to 1, we are going to break that up into small pieces; and these pieces can be thought of as being indexed by this variable  $i$ ; each discrete point therefore is  $i$  divided by  $2$  power  $n$  because there are  $2$  power  $n$  discrete locations. And when we discretized this domain in this fashion, what we can see is that the expectation of  $f$  of  $x$  is approximately, the average value that  $f$  takes in all those discrete points and that is what this formula says, but there

is a little bit of care that we need to take.

if the function  $f$  was to even though it is differentiable, if it were to fluctuate widely then this formula may not work. So, what we need to ensure is that the derivative of  $f$  which kind of tells you about how much it how steeped  $f$  can go if we want to ensure that derivative is bounded by small constant. So, this is an extra assumption that we are making, which is reasonable for most applications.

(Refer Slide Time: 09:24)

Subsample (This is our focus)

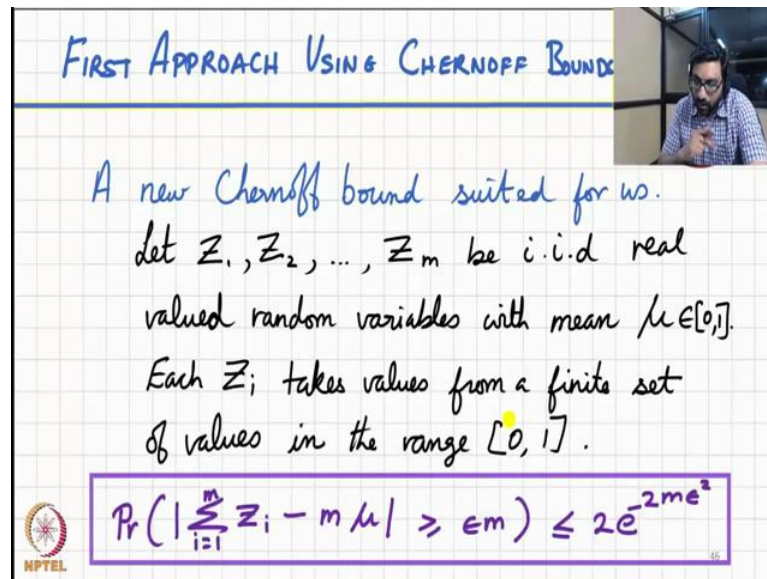
Evaluating  $f$  at all  $2^n$  points in  $[0,1]$  is too expensive.

Trick: subsample a few of the discrete points and estimate  $\bar{f}$

NPTEL 42

So under that assumption, we get this approximation, but notice that to get this good approximation, we needed to have  $2^n$  discrete points and that is just way too much, and way too expensive we cannot compute  $f$  of  $x$  at all of those  $2^n$  discrete points. So, as we mentioned earlier, the trick is to sub sample a few of the discrete points and then estimate  $f$  in those locations, compute  $f$  in those locations, and use that to estimate  $E$  of  $f$  of  $x$ , which we denote by this  $\bar{f}$ . So, the question is how many samples we need, and more importantly how many independent and uniform random bits do we need in order to be able to get those samples.

(Refer Slide Time: 10:24)



**FIRST APPROACH USING CHERNOFF BOUND**

A new Chernoff bound suited for us.  
Let  $Z_1, Z_2, \dots, Z_m$  be i.i.d real valued random variables with mean  $\mu \in [0, 1]$ .  
Each  $Z_i$  takes values from a finite set of values in the range  $[0, 1]$ .

$$\Pr\left(\left|\sum_{i=1}^m Z_i - m\mu\right| \geq \epsilon m\right) \leq 2e^{-2m\epsilon^2}$$

NPTEL

Let us first approach this using the traditional technique that we know when that is using Chernoff bounds. And in a context like this, we need to use a slight variation of the usual Chernoff bound that we normally know. So, let us take this new Chernoff bound. Here let us  $Z_1, Z_2$  up to  $Z_m$  be independent and identically distributed real valued random variables.

So this for example, will be used to model the samples that we draw from each of the 2 power n discrete points, but this bound itself is more general, so  $Z_1$  to  $Z_m$  are just i.i.d real valued variables and their mean is in the range 0 comma 1. And each of the  $Z_i$ 's take value from a finite set of values in the range 0 comma 1. They do not even need to be equally spaced, but in our case, they are equally spaced. We do not even need that, but we do have that if in the as far as we have this freedom to allow that  $Z_i$  is used to take any finite set of values.

And now let us look at the probability that the sum of these  $Z_i$ 's deviates far from its mean. In particular, let us look at the difference between the sum of the  $Z_i$ 's minus  $m$  times the  $\mu$ , the  $\mu$  being the mean of each  $Z_i$ . What is the probability that the absolute difference between these two terms is more than  $\epsilon m$ ? Well, we get a Chernoff bound here and that is at most  $2e^{-2m\epsilon^2}$ . So, this



is very nice.

(Refer Slide Time: 12:26)

**FIRST APPROACH USING CHERNOFF BOUND**

A new Chernoff bound suited for us.  
Let  $Z_1, Z_2, \dots, Z_m$  be i.i.d. real  
valued random variables with mean  $\mu \in [0, 1]$ .  
Each  $Z_i$  takes values from a finite set  
values in  $\mathcal{I}$ .

*n bits each* ←

$$\Pr\left(\left|\sum_{i=1}^m Z_i - m\mu\right| \geq \epsilon m\right) \leq 2e^{-2m\epsilon^2}$$

$\leq \delta$

$\Rightarrow m \in \Omega\left(\frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$

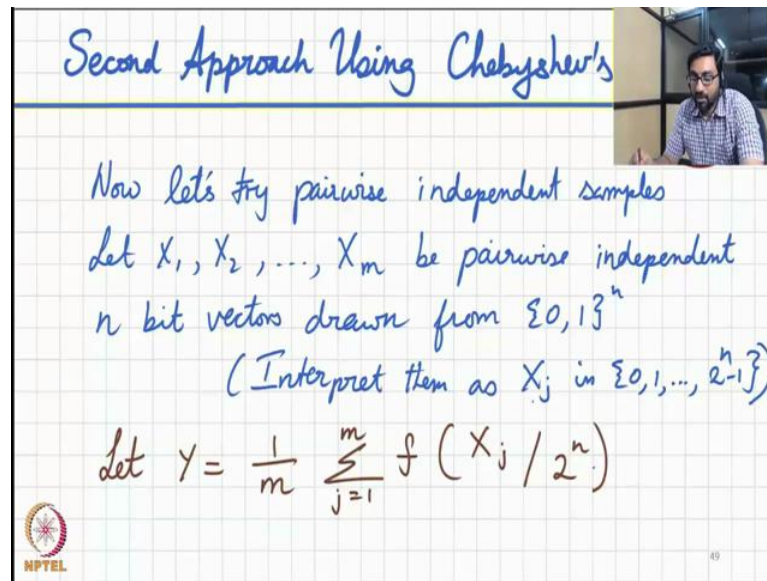
Total # of random bits  
 $\in \Omega\left(\frac{n}{\epsilon^2} \ln \frac{1}{\delta}\right)$

NPTEL

So, let us it make slight easy, we can apply this directly to our context. We can in order to get a nice epsilon approximation within a confidence interval of 1 minus delta, what we need to ensure is that the right hand side of this Chernoff inequality is at most delta. And in order to ensure that the right hand side is at most delta, all we need ensure is that the m basically m is the number of random variables that we considered.

In other words, the number of samples that we need must be in big omega of 1 over epsilon square ln 1 over delta. And each of these sample members, these are real valued variables; in our context, they are going to be 1 of the 2 power n discreet locations. So, each of those samples requires n bits. And so the total number of random bits in this approach would be big omega of n over epsilon squared ln 1 over delta, which is quite a bit. So, of course, we are interested in improving this; so we ask ourselves, can we do better.


(Refer Slide Time: 13:56)



Second Approach Using Chebyshev's

Now let's try pairwise independent samples  
let  $X_1, X_2, \dots, X_m$  be pairwise independent  
 $n$  bit vectors drawn from  $\{0, 1\}^n$   
(Interpret them as  $X_j$  in  $\{0, 1, \dots, 2^n - 1\}$ )

$$\text{let } \gamma = \frac{1}{m} \sum_{j=1}^m f\left(\frac{X_j}{2^n}\right)$$

 49

So our approach tells second approach, we will try to improve this and that is using Chebyshev's inequality. And here we try to exploit the pairwise independence. So, let us try you know pairwise independence sample so our now we call them  $X_1, X_2$  and so on up to  $X_m$  these are pairwise independent  $n$  bit vectors drawn from  $\{0, 1\}^n$ . Now, you can interpret these as basically indices in the range  $0$  to  $2^n - 1$ .

Now let  $\gamma$  be average value of the function  $f$  evaluated at each of the discrete locations indicated by these pairwise independent  $n$  bit random variables. So, remember these  $X_j$ 's we are interpreting them as  $n$  bits basically a location and index into the two power  $n$  discrete locations. So, this  $X_j$  divided by  $2^n$  will give the real value, the value between  $0$  and  $1$  that at which we need to evaluate  $f$  and that and we are evaluating  $f$  at that point.

(Refer Slide Time: 15:24)

Second Approach Using Chebyshev's

$$Y = \frac{1}{m} \sum_{j=1}^m f\left(\frac{x_j}{2^n}\right)$$
$$E[Y] = \bar{f} \quad \text{Var}[Y] = \text{Var}\left(\frac{1}{m} \sum_{j=1}^m f\left(\frac{x_j}{2^n}\right)\right)$$

$\text{Var}[X] = E[X^2] - E[X]^2$

$$= \frac{1}{m^2} \text{Var}\left(\sum_{j=1}^m f\left(\frac{x_j}{2^n}\right)\right)$$
$$\leq \frac{1}{m^2} \cdot m \cdot \text{Var}\left(f\left(\frac{x_j}{2^n}\right)\right) \quad \text{For any } j \in [m]$$
$$\leq \frac{1}{m} E\left[f\left(\frac{x_j}{2^n}\right)^2\right]$$

Since, it is  $x_j$  is drawn uniformly at random from 0 to  $2^n - 1$ , the expectation of  $y$  is exactly equal to the  $\bar{f}$  which is the estimate of the expectation that we want, so that is correct. But how we are doing on the variance front, well, variance of  $y$  is simply the variance of the right hand side which is  $\frac{1}{m} \sum_{j=1}^m f\left(\frac{x_j}{2^n}\right)$ . And as you all know when we get a constant from inside the variance, and we pull it outside of the variance, it has to get squared so that is equal to  $\frac{1}{m^2}$  variance whatever inside at the (Refer Time: 16:26).


Now, the variance of the summation we have already seen can be rewritten where the variance is taken into the summation. So, this variance that is outside can be taken into the summation and so we will get  $\sum_{j=1}^m \text{Var}\left(f\left(\frac{x_j}{2^n}\right)\right)$ . But each of these  $x_j$  values is essentially of the same variance, so we simply for any fixed  $j$ ; we multiply the variance  $m$ , with by  $m$ , and as what we get over here.

And so with some cancellations, we get  $\frac{1}{m}$ , and then notice that the variance of a variable  $x$  for example, is equal to  $E[x^2] - E[x]^2$ . So, if we just want an upper bound, we can just use the first term in this formula for the variance. So, what we do is we replace this variance term with the expectation of the

square of the random variable. And we by doing so, we will get an upper bound which is what we care about.

(Refer Slide Time: 17:51)

Second Approach Using Chebyshev's



$$Y = \frac{1}{m} \sum_{j=1}^m f\left(\frac{X_j}{2^n}\right)$$

$$E[Y] = \bar{f} \quad \text{Var}[Y] = \text{Var}\left(\frac{1}{m} \sum_{j=1}^m f\left(\frac{X_j}{2^n}\right)\right)$$

$$= \frac{1}{m^2} \text{Var}\left(\sum_{j=1}^m f\left(\frac{X_j}{2^n}\right)\right)$$

$$\leq \frac{1}{m^2} \cdot m \cdot \text{Var}\left(f\left(\frac{X_j}{2^n}\right)\right) \quad \text{For any } j \in [m]$$

$$\leq \frac{1}{m} E\left[f\left(\frac{X_j}{2^n}\right)^2\right] \leq \frac{1}{m}$$

$\leq 1$

But it is clear that this expectation is going to be at most 1. Well, why is that because  $f$  member is bounded from above by 1 throughout its domain which is 0 to 1. So this expectation is going to be at most 1, and so we can get an upper bound in the variance which is just 1 over  $m$ .

(Refer Slide Time: 18:20)

The slide is titled "Applying Chebyshev's Inequality" and features a small video inset of a speaker in the top right corner. The main content is written on a grid background. At the top, the title is written in blue cursive. Below it, the Chebyshev inequality is presented as  $\Pr(|Y - \bar{f}| \geq \epsilon) \leq \frac{\text{Var}[Y]}{\epsilon^2}$ . A yellow sticky note on the left says "Significantly better". The inequality is then simplified to  $\leq \frac{1}{m\epsilon^2}$ , which is boxed. A red arrow points from this box to the expression  $\leq \delta \Rightarrow m \geq 1/\delta\epsilon^2$ . Below this, a box contains a recall note: "Recall: With  $2n$  UAR and independent random bits, we can get  $2^n$  samples of  $n$  bits each that are pairwise independent." To the right of this box, the text "More samples than before" is written with a sad face emoji  $\text{:(}$ .

So, let us plug that into our Chebyshev's inequality, what is the probability that our random variable  $y$  is close to the estimate that we want. In particular, what is the probability that the absolute value of  $y$  minus  $\bar{f}$  is greater than  $\epsilon$  and that Chebyshev's inequality tells us that it is bounded by variance of  $y$  divided by  $\epsilon$  square. So, if we apply that member variance of  $y$  is  $1/m$ , and so we get  $1/m\epsilon^2$  as the upper bound on this bad event. And we want this bad event to be at most  $\delta$ .

And would this will imply that  $m$  has to be more than  $1/\delta\epsilon^2$ . And if you recall that is more samples than the approach that we took earlier you saying Chernoff bounds. So in the Chernoff bounds approach, we only needed  $\log(1/\delta\epsilon^2)$  over  $\epsilon^2$ , but here the log fact log is replaced by just  $1/\delta\epsilon^2$ , so that is seems like bit of negative use.

So, how do we interpret this situation? Well we have to recall that we can generate a lot more samples using fewer bits, when all we want is pairwise independence samples. More precisely if we had  $2n$  uniformly at random independent random bits then we saw that we can get  $2^n$  samples each of  $n$  bits each  $n$  bit long and these  $2^n$  samples are guaranteed to be pairwise independent, and if the number of

samples that we need here one over delta epsilon square in that constants, so we should be able to generate that with very few truly random bits.

So, in that sense, when our goal is to minimize the number of truly random bits that is we need in order to do the sampling this approach tends to be much better. So, let us conclude what we saw in this segment is that pairwise independent random bits are actually useful. First of all we saw that tail bounds like Chebyshev's inequality can be applied although Chernoff bounds cannot be applied, so (Refer Time: 21:19) lose some.

And moreover we also looked at an application where the number of truly random bits that we needed was far fewer, even though the number of samples that we took was that we need to take is a little bit more and that is a significant improvement. In particular, keep in mind that there is some context where the number of truly random bits must indeed be kept low. So, one reason why this might this is required is when you need when you run some sort of a scientific experiment, when you need to be able to replicate the sampling.

If the number of samples is a lot, you may need to store all them for application, but if the actual truly random bits that control these experiments was small, we only need to store those random bits and then we can since the pairwise independent random bits were generated by an algorithm. You can actually generate the required pairwise independent random bits and replicate the experiment. And all we needed to do was store very few truly random bits, so that is just one of many reasons why we go through all this (Refer Time: 22:47) to ensure that the number of truly random bits is as small as possible. So with that we conclude this lecture.

In the next lecture, which is an extension of the ideas that we have discussed in this lecture, we will extend our ideas of pairwise and k-wise independence to hash functions, so stay tuned for that.