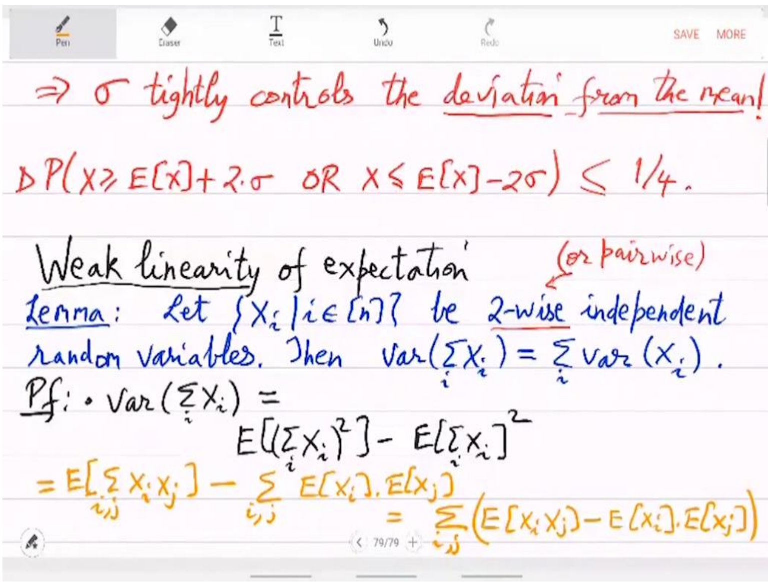
Probability for Computer Science
Prof. Nitin Saxena
Department of Computer Science and Engineering
Indian Institute of Technology - Kanpur

Module - 4
Lecture - 15
Weak Linearity of Variance. Law of Large Numbers.

Last time we did Chebyshev inequality; and before that, we did Markov inequality. And we also defined the variance from the expectation. So, this is all part of concentration inequalities, to understand how far can a random variable go beyond the expectation. So, in particular, we learnt that, even 2 sigma away from the expectation, the probability is very low of that happening.

So, of course, it can happen, it is not impossible, but it is a low probability event. Now, once you have variance, let us prove some more interesting properties of that. So, variance was expectation of discrepancy square, X minus average, X minus expectation squared. So, let us continue with that formula. What we will show is weak linearity of expectation.

(Refer Slide Time: 01:19)



So, the lemma will show what is weak linearity. So, let X i's be random variables, n of them; and say they are 2-wise independent. So, what is 2-wise? 2-wise or pairwise, it means that, if you take X 1 and X 2, then they are independent. This notion of independence that you have already seen, just that, nothing new. So, say you have 2-wise independent random variables.

Then, the variance of sum is sum of variance, which is an amazing thing, because, variance was defined by a quadratic expression in X.

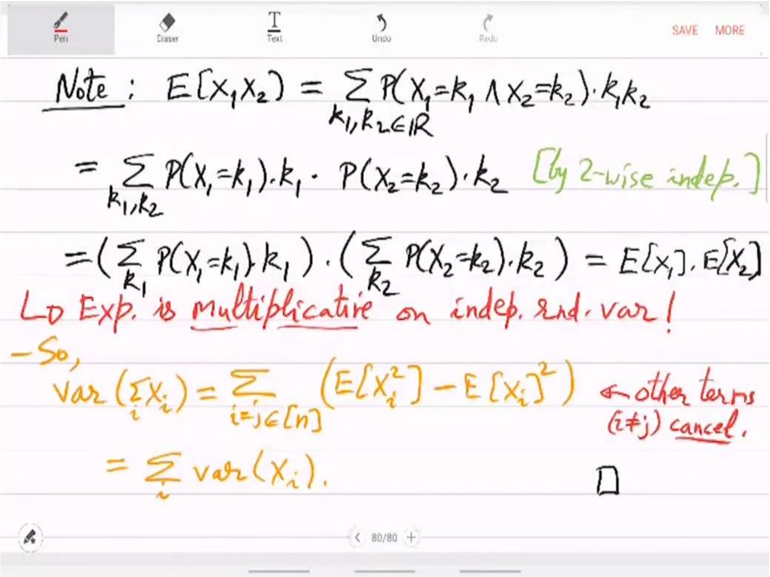
There was simply no chance that it will behave nicely under sum, but it does. And the cost; the price you pay is that, you have to have pairwise independent. If any of these X i, X j's are dependent, then this formula fails. So, I am calling it weak linearity because it is not like expectation. Expectation did not need anything, but variance does. It needs this restriction. But anyways, in many applications, this holds; anyways, this holds true; so, it is a very useful property.

So, proof is quite straightforward. So, you take variance definition. And from that, it followed that expectation of X square minus expectation of X whole square. So, let us use that. It is expectation of sum of X i. And that is what I wanted here also; sum of X i, variance of sum of X i is equal to expectation of this square minus expectation of this thing whole square. That is variance of Sigma X i, random variable.

Now, this is expectation of; so, when you square, what do you get? All possible products minus; so, this will be Sigma expectation of X i whole square, right? So, you will be multiplying expectation of X i with expectation of X j, and it goes over all the i, j's. So, i, j, when they are equal, this is square; when they are different, then there are 2 ways, i less than j or the flip. So, you get the factor 2 in that case. And this is; how do I work with this now?

So, I will write this as Sigma i, j; by linearity of expectation, I get expectation of X i X j minus the product. But what is this? How do I make sense of this? How do I simplify this? So, this is where I will use pairwise independence.

(Refer Slide Time: 06:37)



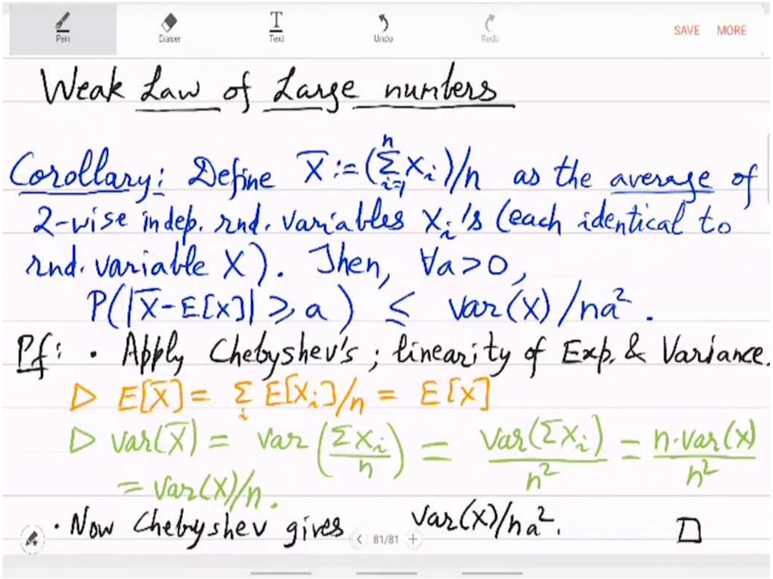
So, here is the claim. I need the claim that expectation of X 1, X 2; since X 1, X 2 are independent by definition, what I get is probability of X 1 being k 1 and X 2 being k 2 times the value k 1 k 2, over all k 2 real numbers. This is expectation of the product, is basically the probability times the value; but since they are independent, the probability factors. So, what you get is, sum X 1 k 1 times k 1 with X 2 k 2 k 2 overall k 1 k 2.

This is by 2-wise independence. And so, that gives you the product. So, you have, clearly, if you see term by term, then this product gives you all these summons of which appear in expectation definition. So, this is then expectation of X 1 times expectation of X 2. So, what you have shown is that expectation is multiplicative as long as the random variables are independent. So, expectation is multiplicative on independent variables.

That is what you have learnt. Now, with this learning, you can easily simplify the expression which we had before, because this expectation of X i X j minus expectation of X i times expectation of X j, this is 0, except when i is equal to 0. So, you get that variance of Sigma X i is equal to expectation of X i square minus expectation of X i whole square. This is the only term which survives in that sum on variance. And what is this?

This is clearly Sigma i variance of X i. That was the statement of the lemma, variance of sum is sum of variance; so, we have shown this. Other items have cancelled; that was the multiplicative property of expectation. So, that is a very nice property to have. And let me show you just 1 example which is already major.

(Refer Slide Time: 10:54)



It is called weak law of large numbers. So, what is this law of large numbers? So, it is the following statement. It is a corollary of what we just did; follow quite easily. So, define the mean of those pairwise random variables as X bar; and the random variables and sum divided by n, that is the mean; as the average of 2-wise independent random variables X i's. And each; so, recall the setting; setting of this was; no; this is something new.

So, each identical to a random variable X. So, these are just copies actually of the same X, but these experiments were done not completely independently, but just pairwise independently. So, if you look at X i and some other X j, they are independent. And then, you are looking at the average random variable X bar. So, what you can show for all positive a: The probability that the average is away from the expectation of X, this discrepancy being larger than this a is given by variance.

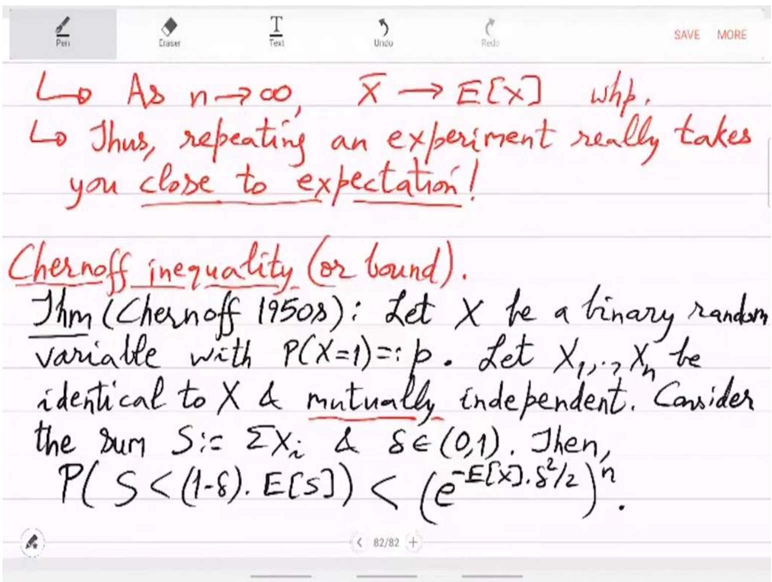
So, the variance of X decides whether the discrepancy can be more than a. So, in fact, if the variance is small; and especially, when you keep increasing the number of experiments n; as n tends to infinity, this tends to 0. So, the probability is very small that average is different from expectation. That is what you learn. That is what this corollary is saying. So, let us prove it. It is quite easy. So, just apply Chebyshev's inequality and linearity of variance.

So, what would Chebyshev say? So, Chebyshev says that you need expectation of X bar and variance of X bar. So, let us do that. So, expectation of X bar is; by linearity, it is actually expectation of X i's summed up, divided by n, which is just expectation of X, because they are all identical to X. And the second thing is variance, which we have just shown, weak linearity of variance.

So, variance of X bar is summation of variance of X i divided by n for all i, which is; let me do this in a proper way. So, this is variance of summation X i by n. Now, you use the property of variance that variance of a times X is a square times variance. So, that is, in this case, 1 by n is the multiplier. So, it comes out and it becomes variance of the sum divided by n square.

Now, variance of the sum is sum of variance, which is n times variance of X divided by n square. So, this is nothing but variance of X divided by n. So, after this Chebyshev gives you variance of X by n a square. So, that is the full proof. It is just simple application of what you did till now. And the interpretation is, as I said, if you repeat an experiment again and again, many times; look at the average; that is basically expectation with very high probability.

(Refer Slide Time: 17:07)



So, as n tends to infinity, X bar tends to the expectation, with high probability. The average random variable is just the expectation with high probability. And thus, repetition; so, repeating an experiment really takes you close to expectation. So, repetition really takes you to expectation. That is what you have learnt from this. So, this is called weak law of large numbers.

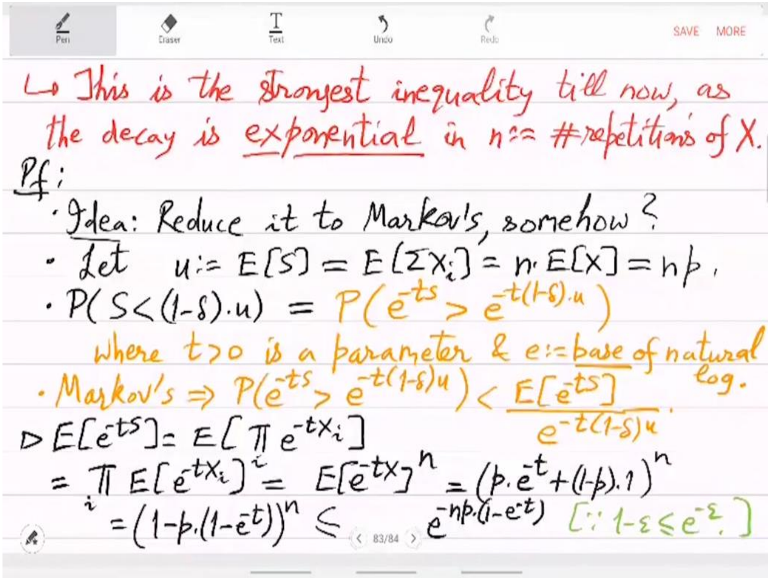
And let us now move to the third and the strongest concentration bound which is called Chernoff inequality. So, you have seen Markov inequality, Chebyshev inequality and now, Chernoff inequality. These are the 3 important concentration theorems, especially in computer science. So, the theorem is credited to Chernoff around 1950s. So, let X be a binary random variable; like a coin toss, head or tails, 0 or 1; with probability of being 1, is called p.

Let X 1 to X n be identical to X and mutually independent. This keyword is very important. It is not pairwise, it is not 2-wise, it is mutually independent, which means that any of these X i's, given any of these X i's, even a subset of them, any subset of X i's, some other X j is independent. So, it is not just between pairs, but it is for any subset. And similarly, any 2 disjoint subsets, if you partition X 1 to X n into 2 subsets, disjoint subsets, they are independent.

This is a maximum kind of independence. This is being assumed. And then, you will get a powerful conclusion. What is that? So, look at their sum; Sigma X i and a delta which is a fraction. Then, the probability that S is away from its expectation. What is the probability that the sum is significantly smaller than the expectation? By significant, I mean, 1 minus delta; it is a fraction. So, say delta is half, then this is half.

So, what is the chance that S is half of expectation? So, this is e to the minus expectation of X delta square by 2 and the whole thing raised to n. So, what is so special about this statement is; compare it with the weak law of large numbers. So, in weak law of large numbers, you were getting, n was coming in the denominator, right? This 1 by n was appearing. So, probability was falling like linearly, linear in n; but here, the probability is falling exponentially in n. That is the difference. It is a qualitative difference.

(Refer Slide Time: 22:25)



So, this is the strongest till now, as the decay in the probability is exponential in n, which is number of repetitions of X. So, the same experiment, if you repeat again and again in an independent fashion, then, even this multiplicative, the value of S being multiplicatively smaller than the expectation, this probability is decaying exponentially in n. So, it becomes extremely small; it can be made very quickly, very small, as small as you want.

So, first let us prove it, then we will see more interpretations. So, what is the proof? Idea is, reduce it to the only thing you know, which is Markov's, definition of expectation. So, reduce it to Markov. So, let u be expectation of S, which is expectation of Sigma X i, which is actually n times expectation of X. That is what you know. That is what the random variable X is up against or S is up against.

You want to compare S with n times E X; you get the probability. So, that is, probability that S is less than 1 minus delta u. So, let us change this to the Markov format, which is, you have to flip less than to greater than. How do you do that? You can do that by introducing, by multiplying it by minus; but there is a nice trick in the proof which will give you exponential decay, and that is use of exponentiation.

So, directly we will write e to the -ts greater than e to the -t 1 minus delta u, for a positive parameter t. And what is e? e is the base of natural log. So, this is around 2.781 I think. So, base, it is given by power series 1 plus 1 over 1 factorial plus 1 over 2 factorial and so on. So, why is this true? These 2 events are essentially the same; they are isomorphic, because, when S is small, then e raised to -ts is large.

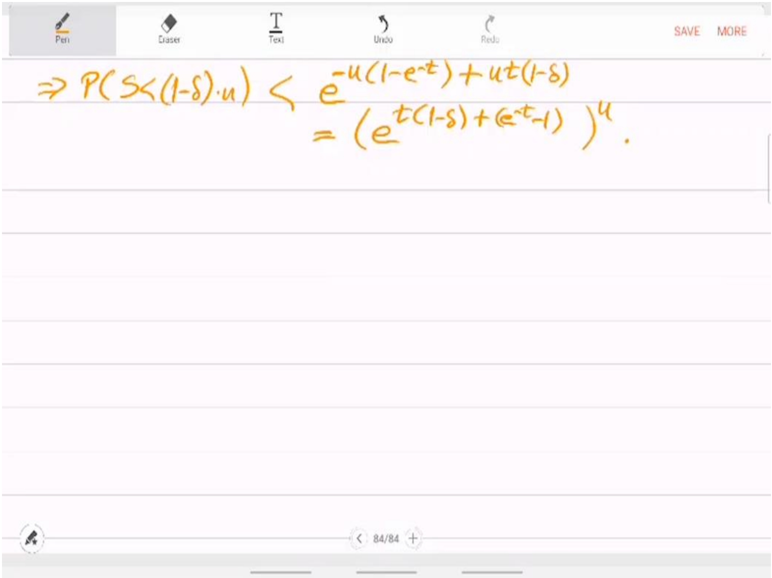
S is small means, -ts is larger; and then you get e raised to that also larger. You can check this. So, in this way, we have converted less than to greater than; and then you invoke Marcov's inequality. So, Markov's implies that probability of this random; view e raised to -t is a random variable. This being more than e raised; the value, this value is less than the expectation of the random variable divided by this.

Now, the expectation of e raised to -ts is what? So, this is expectation of e raised to -t X i; product, right? S is a product. Now, the beauty of exponentiation is that sum becomes product. So, S was a sum. And because of exponentiation, you get actually product of e raised to -t X i. Now, since X i's were completely; so, they are mutually independent. You have seen a proof that this will completely factorise. So, you will get product outside.

Well, now, all the X i's are, essentially, they are the same, right; they are say isomorphic to X. So, we just have to compute the expectation of this and then raise it to n. So, let us do that. So, X is binary random variable. So, it is 1 with probability p. So, you get p times e raised to -t; and it is 0 with opposite probability; and this thing raised to n. So, that is a neat expression. So, you know the expectation of e raised to -ts, and it will help us to simplify it further.

So, let us write it as 1 - p, like this. And then it is less than equal to; I can write it using e again, like this. This is because of 1 minus epsilon being at most e raised to minus epsilon. For a fraction epsilon, you can use the screwed upper bound. And so, ultimately, the expectation is just e raised to -np times this parameter 1 minus e raised to -t. So, let us now go back where we were, which is Markov application, right?

(Refer Slide Time: 30:43)



So, you get that probability of S being less than 1 minus delta u is less than expectation that we just calculated. Let me rewrite this. What is np? np is; I should have said here that this is equal to np. So, u is actually np, and let us use that. So, it is -u 1 minus e raised to -t plus; in the denominator you have t 1 minus delta u, which is just e raised to t 1 minus delta plus e raised to -t minus 1, whole thing raised to u.

So, that is what we have done after this long calculation. We now have a handle, a good handle on probability of; than the expectation, by a multiple of 1 minus delta; but t's are unknown parameters, so, we have to now fix t as a function of delta, so that this right-hand side is minimised. So, that is what we will do next.