

**Concentration Inequalities**  
**Prof. Aditya Gopalan**  
**Prof. Himanshu Tyagi**  
**Department of Electrical Communication Engineering**  
**Indian Institute of Science, Bengaluru**

**Lecture - 04**  
**Hoeffding and Bernstein inequalities**

(Refer Slide Time: 00:20)

The screenshot shows a presentation slide titled "Lecture 3" dated "12 October 2020" at "11:38". The slide contains handwritten notes in black ink on a yellow background. The notes are as follows:

- \* Agenda: Sub-Gaussian random variables, Hoeffding's inequality, Bennett/Bernstein inequality.
- A) SUB-GAUSSIAN R.V.s
- Def: A r.v.  $X$  with  $EX=0$  is said to be SUB-GAUSSIAN if  $\exists v > 0$  s.t.  $\forall \lambda$   $\psi_X(\lambda) \leq \frac{\lambda^2 v}{2}$
- Below the inequality, there are two red annotations:  $\log E e^{\lambda X}$  with an arrow pointing to  $\psi_X(\lambda)$ , and  $\log E e^{\frac{\lambda^2 v}{2}}$  with an arrow pointing to  $\frac{\lambda^2 v}{2}$ .
- Below the definition, it says: "We denote this by  $X \in \mathcal{G}(v)$ ."

Hi all. In this lecture, we will study sub Gaussian random variables, which is a convenient and common class of random variables for which we can prove nice concentration inequalities. Then, we will study two classical concentration inequalities derived using the Chernoff method that we have seen earlier. One is called Hoeffding's inequality and the other is like an improvement on Hoeffding called the Bennett or the Bernstein inequality.

So, let us start out with defining sub Gaussian random variables. So, here random variable  $X$  with 0 mean is said to be sub Gaussian random variable. If, there is some number  $v$ , a positive number  $v$ , such that it is log moment generating function is upper bounded  $\lambda^2 v / 2$  ok, for every value of  $\lambda$ .

Now, just to refresh your memory  $\psi_X(\lambda)$  is the log moment generating function of  $X$  at  $\lambda$ . And, if you recall  $\lambda^2 v / 2$  is actually the log moment generating function of a Gaussian

random variable. So,  $e$  raise to  $\lambda G$ , where  $G$  is distributed as normal with mean 0 and variance  $v$  ok.

So, this is the reason why this random variable  $X$  is called sub Gaussian, it is because its log moment generating function is upper bounded by the log moment generating function of a normal random variable, with the appropriate variance, which is  $v$  in this case.

(Refer Slide Time: 02:07)

$\hookrightarrow$  all random variables sub Gaussian w/ (variance) parameter  $v$ .

NOTES: ① If  $X_i \in \mathcal{G}(v_i)$ ,  $X_1, \dots, X_n$  independent, then  

$$\sum_{i=1}^n X_i \in \mathcal{G}\left(\sum_{i=1}^n v_i\right).$$

②  $X \in \mathcal{G}(v) \Rightarrow \forall t \geq 0: \begin{cases} \mathbb{P}(X \geq t) \leq \exp\left(-\frac{t^2}{2v}\right), \\ \mathbb{P}(-X \geq t) \leq \exp\left(-\frac{t^2}{2v}\right). \end{cases} \left. \vphantom{\begin{matrix} \mathbb{P}(X \geq t) \\ \mathbb{P}(-X \geq t) \end{matrix}} \right\} \text{by Chernoff.}$

③  $X \in \mathcal{G}(v) \Rightarrow \text{Var}[X] \leq v$ .

\* EQUIVALENT CHARACTERIZATIONS of SUB-GAUSSIAN R.V.s:

So, we use the notation  $\mathcal{G}$  of  $v$  to basically denote all random variables, which are sub Gaussian with a parameter, we will also call it the variance parameter, for reasons that will become clear later  $v$  ok.

So,  $\mathcal{G}$  of  $v$  is the class of all random variables, which are sub Gaussian with variance parameter  $v$ . And, so, we use the notation  $X$  belongs to script  $\mathcal{G}$  of  $v$ , calligraphic  $\mathcal{G}$  of  $v$ , if  $X$  is 0 mean and it is sub Gaussian with this variance parameter  $v$ . So, moving on there are couple of remarks about this definition of sub Gaussianity.

So, it is easy to check that, if you have random variables which are independent called  $X_i$ . And, each of them is sub Gaussian with an appropriate variance parameter  $v_i$ , then their sum is also sub Gaussian, where the sub Gaussianity parameter of the sum is simply the sum of the sub Gaussianity of each random variable.

The second point here is that if  $X$  is a sub Gaussian random variable, then it is / definition 0 mean. And, you have concentration inequalities that bound the probability mass of it exceeding a level  $t$  a positive level  $t$ . So, you immediately have / the moment generating function bound and the Chernoff method, the probability that  $X$  exceeds  $t$  or  $-X$  exceeds  $t$  is at most  $e$  raised to  $-t^2 / 2v$ .

So, these are just applications of the Chernoff method ok. So, / the Chernoff technique ok. Because, you have an upper bound and the log moment generating function, which is that of a Gaussian random variable. This will give you a lower bound on the dual of  $\Psi_X \lambda$  which is  $\Psi_X^*$  of  $t$  and that is precisely what comes  $\times$  this exponent here ok.

(Refer Slide Time: 04:21)

\* EQUIVALENT CHARACTERIZATIONS of SUB-GAUSSIAN R.V.s:

Theorem: Let  $EX=0$ . The following are equivalent for suitably chosen  $v, b, c, d$ :

- ①  $X$  is  $v$ -subgaussian
- ②  $P[|X| \geq t] \leq 2e^{-bt^2}$
- ③  $E[e^{cx^2}] \leq 2$
- ④  $\forall q \in \mathbb{N} : E[X^{2q}] \leq (4d)^q \cdot q!$

stated without proof.

The third remark here is that it is it is easy to show that if  $X$  is sub Gaussian with parameter  $v$ , then its variance is actually bounded by  $v$  ok.

So, this is the reason we call this parameter  $v$  as also as the variance parameter. Now, the definition for sub Gaussian random variables that we stated here, is not the only one that allows you to specify a tail behavior for the distribution of such random variables. There are in fact, several equivalent ways of being able to enforce a sub Gaussian tail for random variables.

So, this is the following result. So, if you have a random variable whose mean is 0. Then, the following statements are essentially equivalent for suitably chosen numbers  $v$ ,  $b$ ,  $c$  and  $d$  ok. So, essentially up to constant scale factors all these definitions give you the same property of random variable  $X$ , which is essentially the fact that it has a sub gaussian tail.

So, the first property is our usual definition that we just made about  $X$  being  $v$  sub Gaussian. The second property is the resultant Chernoff bound that you get for its tail, which essentially decays at rate  $E$  raised to  $- \text{constant} \times t^2$ . The third equivalent characterization of a sub Gaussian random variable says that, if you take the expected value of  $E$  raised to  $c X^2$  for some positive number  $c$ , for an appropriate positive number  $c$  then it is bounded it is finite ok.

So, such a random variable  $X$  enjoys finite expected values of its 2 ok, constant time 2. The fourth properties is essentially, the fourth property is essentially controlling the rate of growth of the moments of the random variable. So, if you take each even moment, which is expected value of  $X$  raised to  $2q$  then it is basically bounded by the rate of growth of the moments of a Gaussian random variable.

So, it turns out that if you do the computations for a standard Gaussian random variable, then the  $2q^{\text{th}}$  moment will essentially grow at this rate. You know by  $q$  factorial  $\times$  let us say some constant 4 times  $d$  raised to  $q$  ok. We will not prove this theorem we will just state this result without proof ok, because it will be a bit of a diversion from the flow of this lecture.

So, with this definition of sub Gaussian random variables in hand, it is useful to study what types of random variables can be sub Gaussian. It turns out that a large and common commonly known classes of random variables that we have seen, actually turn out to be sub Gaussian.

(Refer Slide Time: 07:20)

④  $\forall q \in \mathbb{N} : E[X^q] \leq (4d)^q \cdot q!$

B) HOEFFDING'S INEQUALITY  
Hoeffding's Lemma (Lemma 3.3 in book) (Hoeffding, '63)

Suppose a r.v.  $X$  takes values in  $[a, b]$ . Then,  $X \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$ ,  
i.e., it is subgaussian w/ (variance) parameter  $\frac{(b-a)^2}{4}$ .

So, the first step in our agenda to examine to study sub Gaussian random variables is what is called Hoeffding's inequality, which finally, says that bounded random variables are actually sub Gaussian.

So, in order to prove Hoeffding's inequality, which we will show later. We first need an auxiliary result called Hoeffding's lemma. So, this is also lemma 3.3 in the concentration inequalities book and it takes back to the work of Hoeffding in 1963. So, the lemma is rather simple to state.

So, it says that if you have a bounded random variable; that means, a random variable that takes values in an interval  $a$  comma  $b$  with probability 1. Then it is actually sub Gaussian with the following sub Gaussianity parameter, the following variance parameter which is  $(b - a)^2 / 4$  ok. So, it helps you conclude that bounded random variables are sub Gaussian provided you adjust the scale of the sub Gaussianity appropriately ok.

(Refer Slide Time: 08:25)

PROOF: • If  $X \in [a, b]$ , then

$$\text{Var}[X] = \text{Var}\left[X - \frac{a+b}{2}\right] \leq \mathbb{E}\left(X - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}.$$

Also, with  $\Psi_X(\lambda) = \log \mathbb{E} e^{\lambda X}$ , observe that

So, let us do the proof of this lemma due to Hoeffding. So, there are several ways in which one can prove this lemma. We will present one of them, which essentially uses the language of moment generating functions. So, the first observation is to say that, if  $X$  is a bounded random variable between  $a$  and  $b$ , then one can upper bound its variance as follows.

So, the variance of  $X$  is the same as the variance of  $X - \text{the midpoint of } a \text{ and } b$  ok. Subtracting a constant does not change the variance. And, the variance is upper bounded / the second moment, because the variance is = the second moment - the 2 of the first moment.

So, this is at most expected value of  $X - a + b / 2$  the whole 2 ok. And, since  $X$  is bounded between  $a$  and  $b$  and  $a + b / 2$  is the midpoint, the largest that  $X$  can deviate from the midpoint  $a + b / 2$  is again  $/ a + b / 2$ . And, so, this gives you a bound of  $(b - a)^2 / 4$  ok. /By the way this bound is tight and you can easily convince yourself that you can design an appropriate random variable  $X$  between  $a$  and  $b$  that meets this inequality with equality.

The second observation is to say that so, let so, we have the notation  $\Psi_X$  of  $\lambda$  is the log moment generating function of  $X$  at  $\lambda$ . So, let us make a couple of observations.

(Refer Slide Time: 10:20)

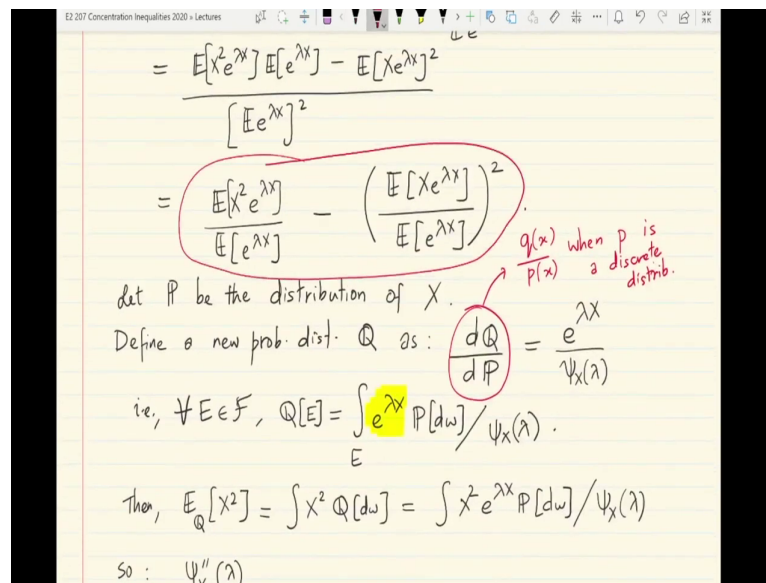
$$\begin{aligned}\Psi_X(0) &= \Psi_X'(0) = 0. \\ \Psi_X''(\lambda) &= \frac{d}{d\lambda} \Psi_X'(\lambda) = \frac{d}{d\lambda} \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \\ &= \frac{\mathbb{E}[X^2 e^{\lambda X}] \mathbb{E}[e^{\lambda X}] - \mathbb{E}[X e^{\lambda X}]^2}{(\mathbb{E}[e^{\lambda X}])^2}\end{aligned}$$

The first is to observe that, if you set  $\lambda = 0$ , then you get log of 1 which is 0. And, the same holds if you differentiate  $\Psi_X$  of  $\lambda$  with respect to  $\lambda$  and set  $\lambda = 0$  ok. So,  $\Psi_X$  of 0 is going to be = the derivative of  $\Psi_X$  at 0, which is 0 ok. So, this is easy to check.

Now, let us move on to the second derivative of  $\Psi_X$ . So, let us compute the second derivative  $\Psi_X$  double dash of  $\lambda$  at any  $\lambda$  is / definition  $d/d\lambda$  of  $\Psi_X$  prime at  $\lambda$  which is just. So, if you differentiate  $\Psi_X$  once you will basically get expected value of  $X e^{\lambda X}$  divided /  $E$  to the  $\lambda X$  expected value of  $e^{\lambda X}$ .

And, so, we just differentiate this using basic rules of calculus. So, on the numerator you will have expected value of  $X^2 e^{\lambda X} \times$  expected value of  $e^{\lambda X}$  — expected value of  $X e^{\lambda X}$  raised to  $(\lambda X)^2$ . And, at the denominator you will just have the original denominator the whole 2 expected  $e^{\lambda X}$ .

(Refer Slide Time: 11:53)



$$= \frac{\mathbb{E}[X^2 e^{\lambda X}] \mathbb{E}[e^{\lambda X}] - \mathbb{E}[X e^{\lambda X}]^2}{(\mathbb{E}[e^{\lambda X}])^2}$$

$$= \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left( \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \right)^2$$

let  $P$  be the distribution of  $X$ .  
 Define a new prob. dist.  $Q$  as:  $\frac{dQ}{dP} = \frac{e^{\lambda X}}{\psi_X(\lambda)}$  (q(x)/p(x) when p is a discrete distrib.)

i.e.,  $\forall E \in \mathcal{F}, Q(E) = \int_E e^{\lambda X} P(d\omega) / \psi_X(\lambda)$ .

Then,  $\mathbb{E}_Q[X^2] = \int X^2 Q(d\omega) = \int X^2 e^{\lambda X} P(d\omega) / \psi_X(\lambda)$

So:  $\psi_X''(\lambda)$

So, one can rearrange this as follows you can write this as distributing the numerator you just have  $X^2 e^{\lambda X}$  divided by expected  $e^{\lambda X}$  - the 2d of expected  $X e^{\lambda X}$  divided by expected  $e^{\lambda X}$  ok. So, this is just basic algebra and calculus. So, now, we will do something nontrivial we will define so, let  $X$  let  $P$  be the distribution or probability measure of  $X$  of the original random variable  $X$  ok. Define, let us define a new probability distribution.

So, denote so, let us define a new probability distribution  $Q$  on the same space as of values that  $X$  takes. As, so, you will use the following formula, we will say the ratio of  $Q / P$  at every sample point is  $e^{\lambda X}$ ;  $X$  is the value of the random variable divided by  $\psi_X(\lambda)$  to make it a valid probability distribution that is to make it sum to 1 ok.

So, the notation  $dQ / dP$  if you are not familiar with this. If, so, you can just for concreteness imagine that  $X$  is a discrete random variable. And, so, this if  $X$  is a discrete random variable, this just reduces to  $q$  of  $x$  /  $p$  of  $x$ . So, when  $P$  is a discrete distribution, if  $P$  has a probability density function then  $dQ / dP$  is taken to mean the ratio of probability densities of  $Q / P$ . And, this can be extended even for general random variables ok.

So, in other words an equivalent way of defining this new probability distribution is that for every event  $E$  in the original  $\sigma$  algebra of  $X$  the measure the probability of the set  $E$  as



measured /  $Q$  is defined to be the integral over this event  $E$ , of  $e$  raised to  $\lambda X$  with the original distribution  $P$ ,  $P d\omega$ . And, you have to finally, divide /  $\Psi_X$  of  $\lambda$  ok, to make it a valid probability measure.

So, this is basically how the distribution  $Q$  is defined ok. It is also called an exponentially tilted measure, because what you are doing is basically you are multiplying the original  $P$  / , you are hitting this original  $P$  /  $e$  raised to  $\lambda x$ . So, you are up weighting you are weighting up the probability distributions of values of  $X$  that are very high, if  $\lambda$  is positive and down weighting the corresponding probabilities of the random variable when the values are low, when the values  $X$  of  $\omega$  are low.

So, it is easy to see / a simple calculation that, if you compute expected value under this new distribution  $Q$  of the same random variable  $X^2$ . So, this turns out to be integral, by definition this is the integral over the entire sample space of  $X^2 Q d\omega$ , instead of  $P d\omega$ , and this is = / definition  $X^2 e$  raised to  $\lambda X P d\omega$ . Finally, divided /  $\Psi_X$  so, divided /  $\Psi_X \lambda$  ok.

(Refer Slide Time: 16:09)

The image shows a handwritten derivation in a OneNote application. The text is as follows:

Then,  $E_Q[X^2] = \int X^2 Q[d\omega] = \int X^2 e^{\lambda X} P[d\omega] / \Psi_X(\lambda)$

So:  $\Psi_X''(\lambda) = E_Q[X^2] - (E_Q[X])^2 = \text{Var}_Q[X] \leq \frac{(b-a)^2}{4}$

Finally, applying Taylor's theorem gives:

$$\forall \lambda, \Psi_X(\lambda) = \underbrace{\Psi_X(0)}_{=0} + \lambda \underbrace{\Psi_X'(0)}_{=0} + \frac{\lambda^2}{2} \underbrace{\Psi_X''(\theta)}_{\leq 1/4}, \text{ for some } \theta \in [0, \lambda]$$

$$\leq \frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$$

$$= \frac{\lambda^2 (b-a)^2}{8}$$

A red square symbol  $\square$  is at the bottom right of the derivation.

And, so, you will easily see that,  $\Psi_X$  double prime of  $\lambda$ , which we were computing earlier here, which is this quantity here, in fact, can just be written as  $E_Q$ . So, the first term here simply becomes  $E_Q[X^2]$ , -  $E_Q[X]^2$ . The second term is exactly  $(E_Q[X])^2$  ok.

So, this is  $(E_Q[X])^2$  the second term here. And, this is simply the variance of the same random variable  $X$ , but evaluated under a different probability distribution  $Q$ . And, since  $X$  has not changed the random variable  $X$  has not really changed. So, it is still bounded between  $a$  and  $b$ .

And, so, by our variance result its variance under any distribution, whether it is  $P$  or  $Q$  or any other distribution can be no more than  $(b - a)^2 / 4$  ok. So, what we have shown is that the second derivative of  $\psi_X$ , which is defined with respect to the original distribution  $P$  is actually bounded ok. So, you have a function  $\psi_X$  as a function of  $\lambda$ , with some bounds on its curvature or second derivative.

So, the last step here is to use so, finally, applying the Taylor series formula. So, Taylor's Theorem in the remainder form will give us the following, it gives that for any  $\lambda$ , we have  $\psi_X$  of  $\lambda$  is = so, let us expand this as a Taylor series about  $\lambda = 0$ . So, we have  $\psi_X$  of  $0$ , +  $\lambda$  times  $\psi_X$  prime at  $0$ , + the second order term is going to be  $\lambda^2 / 2 \psi_X$  double prime at some number  $\theta$  between  $0$  and  $\lambda$  ok.

So, this is Taylor's theorem up to second order with the remainder  $\delta$ . And we know that this term is  $0$ , this first derivative is also  $0$  at  $0$ . And, we know that the second derivative is bounded / one-fourth ok. And, so, using this we finally, get that this is upper bounded /  $\lambda^2 / 2 \times (b - a)^2 / 4$ , which is exactly what we wanted in the theorem  $\lambda^2 / 8 (b - a)^2$  ok, that concludes the proof of Hoeffding's lemma.

So, basically what we have shown is that for a random variable  $X$  bounded between the values  $a$  and  $b$ , its log moment generating function is always bounded by this quadratic function of  $\lambda$  with a bound and the coefficient being  $(b - a)^2 / 8$  ok.

(Refer Slide Time: 19:31)

NOTE: Hoeffding's lemma is TIGHT.

\* Corollary - Hoeffding's Inequality (Corollary 3.4, book)

Consider independent random variables  $X_1, \dots, X_n$  such that  $X_i \in [a_i, b_i] \forall i$ . Then, (All  $X_i$  are 0-mean)

$$P\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

PROOF: By Hoeffding's lemma,

$$\sum_{i=1}^n X_i \in \mathcal{G}\left(\sum_{i=1}^n \frac{(a_i - b_i)^2}{4}\right). \quad \square$$

c) BENNETT'S INEQUALITY :

So, an additional note here that I would like to add is that this lemma is actually TIGHT is TIGHT ok. Meaning that, you can actually find a non-trivial random variable  $X$  given any interval  $a$  and  $b$ , which for which the log moment generating function is exactly  $= \lambda^2 / a (b - a)^2$  ok.

So, once we have Hoeffding's lemma in hand, it gives you what is called Hoeffding's inequality, which is Corollary 3.4 in the concentration inequalities book ok. So, this is a concentration inequality that is obtained / appropriately using Hoeffding's lemma. And, it says that if you have independent random variables  $X_1$  through  $X_n$  which are each bounded in the respective interval  $a_i b_i$ , then the probability that their sum exceeds a level  $t$  ok.

So, I should also say that, I should also say that all  $X_i$  are 0 mean ok. Then, the probability that their sum exceeds  $t$  is at most  $e$  raised to  $-2t^2 / \sum_i (a_i - b_i)^2$  ok. So, this gives you a way to control the deviations of bounded random of a sum of bounded random independent and bounded random variables.

The proof is basically an application of Hoeffding's Lemmsa so, it is just one line, so, by Hoeffding's lemma. We have that the sum of these independent random variables must be sub Gaussian with the variance parameter being  $=$  the sum of all their individual variance

parameters, which / Hoeffding's lemma is  $(a_i - b_i)^2 / 4$  ok. So, that is the, that is the proof once you have Hoeffding's lemma.

(Refer Slide Time: 21:53)

c) BENNETT'S INEQUALITY :

A Shortcoming of Hoeffding's ineq: It doesn't bring in the actual variance of  $\sum_i X_i$  (i.e., it is designed for "worst-case" rvs lying in  $[a_i, b_i], \forall i$ ).

Bennett's inequality addresses this gap.

BENNETT'S INEQUALITY (Lemma 3.5, book) :

Consider indep. and zero-mean r.v.s  $X_1, \dots, X_n$  s.t.  $|X_i| \leq c$   $\forall i$ . Let  $\sigma^2 := \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i)$ . Then,

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp(-n \sigma^2 h(t/c))$$

So, we will now move on to proving what is called Bennett's inequality, which is which can be motivated / the observation that, if you stare a little closely  $\times$  Hoeffding's inequality. There is one issue or a weak point in Hoeffding's inequality, it is that it does not bring in the actual variance of the sum of all these  $X$  is or the variance of let us say each of these random variables.

So, there is no dependence here, if you look at Hoeffding's inequality. There is really no dependence here on the actual variance of each  $X_i$ , there is only a dependence in the worst case sense on the complete range of  $a_i$  and  $b_i$ ,  $a$  and  $b$ , which are very far apart, then you  $X_i$  were within very very loose bounds  $a_i$  and  $b_i$ ,  $a$  and  $b$ , which are very far apart, then you would have a fairly weak bound. Because, you know that with high probability they are actually confined to a much smaller range ok, effective range.

So, how do we bring in that measure of spread of each random variable or the spread of the sum? Ok. So, this is one thing that Bennett's inequality will help to address.

(Refer Slide Time: 23:01)

for "worst-case" r.v.s lying in  $[a_i, b_i], \forall i$ .

Bennett's inequality addresses this gap.

BENNETT'S INEQUALITY (Lemma 3.5, book) :

Consider indep. and zero-mean r.v.s  $X_1, \dots, X_n$  s.t.  $|X_i| \leq c$   $\forall i$ . Let  $\sigma^2 := \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i)$ . Then,

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{tc}{\sigma^2 n}\right)\right),$$

where  $h(x) := (1+x)\log(1+x) - x^2$ ,  $x \geq -1$ .

PROOF :

And, so, one version of Bennett's inequality is as follows it is also lemma 3.5 in the book. So, let us take independent and 0 mean random variables  $X_1$  through  $X_n$ , where each of them is bounded, it is the same setting as Hoeffding's inequality.

So, a bunch of independent 0 mean bounded random variables. And, let  $\sigma^2$  be their average variance ok. So, this is not  $n$  here I should say it is this is not  $\eta$  it is  $n$ . So,  $\sigma^2$  is basically the average variance, then we can control the fluctuation of the sum of these random variable independent random variables, by a quantity that depends on the variance the average variance here ok.

So, the  $\sigma^2$  appears here, as well as here ok. The function  $h$  here is something that we have seen earlier, which you should probably recall as related to the Chernoff method or for the tail probability bound for a Poisson random variable ok. So, this appears and we will make the connection to a Poisson random variable in the proof ok.

So, the upshot of this inequality Bennett's inequality is that, it gives you actually a way to bring in the actual variance of the sum of these random variables  $X_1$  through  $X_n$ . And, one can stand to benefit if the actual random variables enjoy a much smaller variance than their worst case bounds  $a_i$  and  $b_i$  s. So, let us go on to prove this inequality. So, the first step of the

proof is to write out the is to bound the moment generating function of each random variable  $X_i$ .

(Refer Slide Time: 24:52)

$$\begin{aligned}
 \text{PROOF: } E[e^{\lambda X_i}] &= \sum_{k=0}^{\infty} \frac{\lambda^k E X_i^k}{k!} = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k E X_i^k}{k!} \\
 &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} E[X_i^2 | X_i|^{k-2}] \\
 &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} C^{k-2} \text{Var}[X_i^2] \\
 &\quad \quad \quad = \sigma_i^2
 \end{aligned}$$

So, let us begin by writing the expected value of the moment generating function of any  $X_i$  ok. So, this is by definition if you just expand it out as infinite power series, this is  $k = 0$  to infinity  $\lambda$  to the  $k$  expected  $X_i$  to the  $k / k$  factorial ok.

So, since the first moment is 0 / assumption, let us take out the first two terms; the first 2 terms are  $1 + 0$  and let us write this sum from  $k = 2$  onwards ok, the  $k = 1$  term is missing, because the expected value of  $X$  is assumed to be 0. So, this is  $\lambda$  to the  $k$ , expected  $X_i$  to the  $k / k$  factorial. And, we will now perform the following bound bounding operation on the summation on the right.

So, this is  $1 +$  this is at most  $1 +$  the sum over  $k = 2$  to infinity. So, we will have  $\lambda k / k$  factorial. So, what we will do is from expected value of  $X_i$  raised to  $k$  since  $k$  is at least 2, we can take out  $X_i^2$  and then we can keep the remaining  $X_i$  raised to  $k - 2$ . So, we have expected value of  $X_i^2$  and the absolute value of  $X_i$  raised to  $k - 2$  ok.

So, / putting the absolute value we can make an upper bound. And, let us now invoke the bound that the absolute value of  $X_i$  is at most  $C$  ok. So, that whatever is remaining is the expected value of  $X_i^2$ , which is just it is variance. So, this is at most  $1 +$  the sum  $\lambda k / k$

factorial. So, the absolute value of  $X_i$  is bounded by  $c$  and then what remains is just the variance of  $X_i$  ok. Let us denote this  $\sigma_i^2$ , this is variance of  $X_i$ .

(Refer Slide Time: 27:05)

$$= 1 + \frac{\sigma_i^2}{c^2} \sum_{k=2}^{\infty} \frac{(\lambda c)^k}{k!}$$

$$= 1 + \frac{\sigma_i^2}{c^2} (e^{\lambda c} - 1 - \lambda c)$$

So, finally, what we have is the expression  $1 + \sigma_i^2 / c^2$ , sum over  $k = 2$  to infinity,  $\lambda c / \lambda c$  raised to  $k / k$  factorial. So, I have taken out the  $c$  to the  $-2$  outside. And, this in turn is just  $1 + \sigma_i^2 / c^2$  ok. So, this is just the power series expansion of  $e$  raised to  $\lambda c$  - the first 2 terms -  $1 - \lambda c$ .

(Refer Slide Time: 27:46)

$$\begin{aligned}
 &= 1 + \frac{\sigma_i^2}{c^2} (e^{\lambda c} - 1 - \lambda c) \\
 &\leq \exp \left\{ \frac{\sigma_i^2}{c^2} (e^{\lambda c} - 1 - \lambda c) \right\} \quad \left\{ \because 1 + \alpha \leq e^\alpha \right\} \\
 \therefore \Psi_{\sum_{i=1}^n X_i}(\lambda) &\leq \left( \sum_{i=1}^n \sigma_i^2 \right) \frac{1}{c^2} (e^{\lambda c} - 1 - \lambda c) \\
 \text{RECALL: If } X \sim \text{Poi}(v), \text{ then } \Psi_{X-v}(\lambda) &= v(e^\lambda - 1 - \lambda), \\
 \Psi_{X-v}^*(t) &= v h\left(\frac{t}{v}\right). \\
 \Psi_{\sum_{i=1}^n X_i}(\lambda) &\leq n \sigma^2
 \end{aligned}$$

And, we will use  $1 + X$  is less than  $= e$  raised to  $X$ , now to make an upper bound here,  $\sigma_i^2 / c^2 e$  raised to  $\lambda c - 1 - \lambda c$  ok.

So, this is because  $1 + \alpha$  is at most  $e$  raised to  $\alpha$  for any value ok. So, finally, we have.

So, therefore, we have we will be able to extend this to the sum of  $n$  independent random variables to show that  $\Psi$  of summation  $X_i$   $i = 1$  to  $n$  at  $\lambda$ , is just  $=$  the sum over all the is well, I think it is it should be an upper bound ok. From  $i = 1$  to  $n$   $\sigma_i^2 / c^2$  and  $e$  raised to  $\lambda c - 1 - \lambda c$  this is just by summing, this inequality above ok. For all  $i = 1$  to  $n$ .

Now, this kind of log moment generating function bound is something that we have actually seen before and just to refresh our memory, let us recall the following fact. That, if  $X$  is a Poisson distributed random variable with some parameter  $v$ , then it is log moment generating function, when you center it by  $v /$  it is mean  $v$ , is  $v \times e$  raised to  $\lambda - 1 - \lambda$ .

And, it is dual function  $\Psi_X^*$  it is Kramer transform of the centered Poisson random variable is  $v$  times  $h$  of  $t / v$  ok, where  $h$  of  $x$  is basically the same function  $h(1 + x \log 1 + x - x)$  ok. So, this is a fact about the tail of a Poisson random variable or the moments of a Poisson random variable. And, you can we can it is we can now do some pattern matching here. So, this expression here can somehow be thought as this here ok.



(Refer Slide Time: 30:26)

$$\Psi_{\sum_{i=1}^n X_i}(\lambda) \leq \left(\frac{n\sigma^2}{c^2}\right)(e^{\lambda c} - 1 - \lambda c)$$

$$\therefore \Psi_{\sum_{i=1}^n X_i}^*(t) \geq \left(\frac{n\sigma^2}{c^2}\right) h\left(\frac{tc^2}{cn\sigma^2}\right)$$

$$= \left(\frac{n\sigma^2}{c^2}\right) h\left(\frac{tc}{n\sigma^2}\right).$$

\* BERNSTEIN'S INEQUALITY (Cor. 3.6, book)

In the setting of the previous lemma,

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(-\frac{t^2}{cn\sigma^2}\right).$$

So, we have now that  $\Psi$  of the sum of  $i X_i$  at  $\lambda$  is at most  $n \sigma^2 / c^2 \times (e^{\lambda c} - 1 - \lambda c)$ . So, this summation of all  $i \sigma_i^2$  is just  $n \sigma^2$  / definition  $n \sigma^2 / c^2 \times e^{\lambda c} - 1 - \lambda c$  ok.

So, let us compare this moment generating function for a centered Poisson with this expression, which we have obtained for our sum of  $X_i$  is ok. So, it follows that one can set  $v$  as  $n \sigma^2 / c^2$ . And, / doing this / making this connection you can find the dual function of this right hand side expression here, to get the bound  $\Psi^*$  of  $\sum_{i=1}^n X_i$  at  $t$  is lower bounded.

Because the primal function  $c$  is upper bounded it will imply essentially a lower bound for the dual function, / the definition of  $\Psi^*$ . So, you get finally, after some algebra  $n \sigma^2 / c^2 h$  of  $tc^2 / cn \sigma^2$  ok. And, this is the same as  $n \sigma^2 / c^2, h$  of  $tc / n \sigma^2$  ok.

So, this is this should be what we want. So, that is  $n \sigma^2 / c^2 h$  of  $tc / n \sigma^2$ . So, that is exactly the expression that we wanted. And, this implies the tail bounded that we see ok. So, that is the proof. So, the proof essentially involved not succumbing to using the standard sub Gaussian bound from Hoeffding's lemma for these random variables, but actually keeping a dependence on their variances.

And, directly relating the moment generating function of each random variable to its variance and finally, connecting it to the moment generating function of a Poisson random variable.

So, that is Bennett's inequality. Now, Bennett's inequality is phrased in terms of this  $h$  function, which has essentially an  $X \log X$  type dependence.

And, often Bennett's inequality is simplified to get a more transparent insight  $\times$  tails of random variables depending on their variances. So, here's one such consequence of Bennett's inequality called Bernstein's inequality, it is given in corollary 3.6 of the book. So, Bernstein's inequality basically says that in the same setting that is if all these  $X_i$ 's are 0 mean and bounded an absolute value  $\leq c$ , then the probability that their sum exceeds the level  $t$  is at most  $e^{-t^2}$ .

So, this is basically in the exponent is a ratio of polynomials, there is a quadratic polynomial on  $t$  on top and there is a linear polynomial depending on  $c$  in the bottom ok. So, this form is what you get as a consequence of Bernstein of Bennett's inequality ok.

(Refer Slide Time: 34:06)

\* BERNSTEIN'S INEQUALITY (Cor. 3.6, book)

In the setting of the previous lemma,

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}ct}\right).$$

PROOF: Use Bennett + the bound

$$h(x) \geq \frac{x^2}{2 + \frac{2}{3}x}.$$

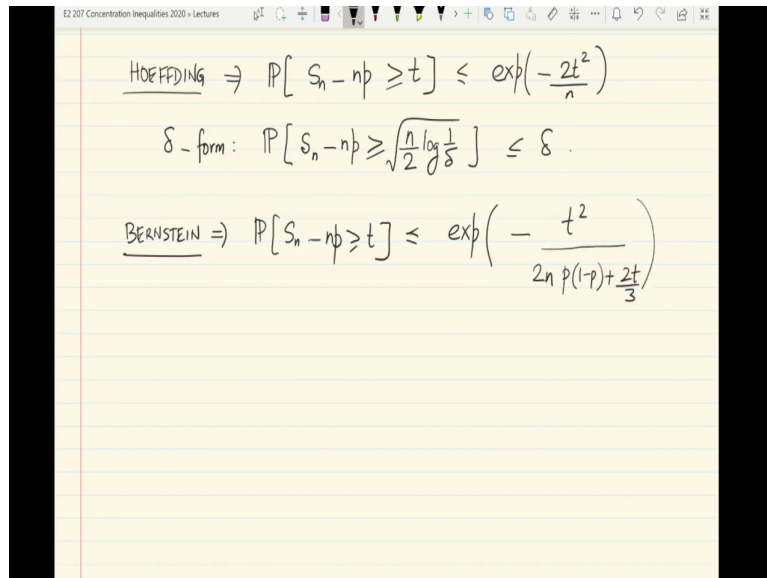
\* EXAMPLE:  $X_i \sim \text{Ber}(p)$ ,  $S_n = \sum_{i=1}^n X_i$ .

The proof is / using Bernstein + the elementary bound so, by using Bennett + the bound. So, the  $h$  function can be bounded below uniformly. So,  $x^2 / 2 + 2/3 X$  sorry this is  $X^2$  yeah ok.

So, this is a universal bound on the  $h$  function and using that in Bennett just gives you Bernstein's inequality. So, let us do a quick example here to understand what Bernstein gives over and above what Hoeffding could have given you. So, we will take the simple example where  $X_i$ 's are all Bernoulli random variables they are iid Bernoulli random variables with

parameter  $P$  and  $S_n$  is their sum. So, in other words  $S_n$  is a binomially distributed random variable with mean  $np$ .

(Refer Slide Time: 35:17)



Handwritten notes on a yellow background showing Hoeffding and Bernstein inequalities. The text is written in black ink on lined paper. The top line shows the Hoeffding inequality:  $\text{HOEFFDING} \Rightarrow \mathbb{P}[S_n - np \geq t] \leq \exp(-\frac{2t^2}{n})$ . The second line shows the  $\delta$ -form of Hoeffding's inequality:  $\delta\text{-form: } \mathbb{P}[S_n - np \geq \sqrt{\frac{n}{2} \log \frac{1}{\delta}}] \leq \delta$ . The third line shows the Bernstein inequality:  $\text{BERNSTEIN} \Rightarrow \mathbb{P}[S_n - np \geq t] \leq \exp\left(-\frac{t^2}{2np(1-p) + \frac{2t}{3}}\right)$ .

So, Hoeffding would give you the following. So, just applying Hoeffding on these would give you that, the probability that,  $S_n$  - its mean which is  $np$  exceeds a positive number  $t$  is at most  $e$  raised to  $-2t^2 / n$ . So,  $a$  and  $b$  are set to 1 here, because that is the range of every Bernoulli random variable.

So, equivalently the  $\delta$  form of Hoeffding's inequality, says that with probability at most  $\delta$  this event occurs  $S_n - np$  exceeds so in fact, you can put  $\geq \sqrt{\frac{n}{2} \log \frac{1}{\delta}}$ , so, it is basically  $\sqrt{n \times \log 1 / \delta}$  ok, at most  $\delta$ .

Now, contrast this with Bernstein ok, which would give you the following inequality. Probability that  $S_n - np$  exceeding  $t$  is at most  $e$  raised to  $-t^2 / 2$  times  $n$  times the variance of each Bernoulli, which is in this case  $p \times (1 - p) + 2t / 3$  ok,  $c$  is set to 1 here which is the range of each random variable ok. It is a little more transparent if you try to find the  $\delta$  form of Bernstein's inequality in somewhat more tedious manner.

(Refer Slide Time: 36:52)

Handwritten notes on a digital notepad showing the derivation of the  $\delta$ -form. The top line shows the equation  $t^2 = (2np(1-p) + \frac{2}{3}t) \log\left(\frac{1}{\delta}\right)$ . The bottom line shows the solution  $t = \frac{1}{3} \log\left(\frac{1}{\delta}\right) + \frac{1}{3} \sqrt{\left(\log\left(\frac{1}{\delta}\right)\right)^2 + 72np(1-p) \log\left(\frac{1}{\delta}\right)}$ .

So, let us try to find the  $\delta$  form. So, to find the  $\delta$  form we need to equate the right hand side of this inequality to  $\delta$  and solve for  $t$ . So; that means, that  $t^2 = (2np(1-p) + \frac{2}{3}t) \log 1 / \delta$  ok, that is the only way that the right hand side will become  $e$  raised to  $-\log 1 / \delta$ , which is  $\delta$ .

And, it turns out so, this is a quadratic equation in  $t$  and it turns out that if you actually solve it in closed form you will get  $t = \frac{1}{3} \log 1 / \delta, + \frac{1}{3} \sqrt{\log 1 / \delta \text{ whole } 2, + \text{some constant } 72np(1-p) \log 1 / \delta}$  ok all under the 2 root.

(Refer Slide Time: 38:13)

$$\begin{aligned} & \leq \log \frac{1}{\delta} + 4\sqrt{np(1-p)\log\left(\frac{1}{\delta}\right)} \\ \therefore \mathbb{P}\left[S_n - np \geq \log \frac{1}{\delta} + 4\sqrt{np(1-p)\log\left(\frac{1}{\delta}\right)}\right] & \leq \delta \end{aligned}$$

And, you can just upper bound this if you like by  $\log 1 / \delta + 4$  times  $2 \sqrt{n p (1 - p) \log 1 / \delta}$  ok. For  $\delta$  upper bounded by some number  $\delta$  naught for instance ok. So, for small enough  $\delta$  this certainly holds.

(Refer Slide Time: 38:35)

$$\begin{aligned} & \therefore \mathbb{P}\left[S_n - np \geq \log \frac{1}{\delta} + 4\sqrt{np(1-p)\log\left(\frac{1}{\delta}\right)}\right] \leq \delta \\ & \text{compare w/ Hoeffding: } \sqrt{\frac{n \log \frac{1}{\delta}}{2}} \\ & \Rightarrow \text{Gain when } p(1-p) \text{ is "small".} \\ & \text{e.g. } X_i \stackrel{\text{iid}}{\sim} \text{Ber}(p), \quad i=1, \dots, n, \quad \text{where} \\ & \text{Var}(X_i) = \frac{1}{n} \Leftarrow p \approx \frac{1}{n}. \end{aligned}$$

And, so, finally, you will get that the probability that  $S_n - np$  exceeds this number which is  $\log 1/\delta + 4\sqrt{2np(1-p)\log 1/\delta}$ , this event occurs with small probability ok. So, just to remind you what Hoeffding would have given for the same  $\delta$  target probability of error.

So, let us compare with Hoeffding. So, what would the typical deviation from Hoeffding be, it would be basically  $2\sqrt{n \log 1/\delta}$  ok. So, the way to compare this so, one way to compare this is let us say that  $\delta$  is set to some small number,  $\delta$  is fixed at a constant. So,  $\log 1/\delta$  is a constant and  $n$  is a large number  $1/n$  is a fairly large number.

So, here this quantity goes basically as  $\sqrt{n}$  whereas, this quantity actually goes as  $\sqrt{n \times p(1-p)}$  ok. So, this is very useful if  $p \times (1-p)$  is a fairly small number. So, if each of the variables the Bernoulli variables  $X_i$  have very small variance, then one could actually gain by reducing the deviation of from Hoeffding which is order  $\sqrt{n}$  to something which is of order  $\sqrt{n}$  times the variance of each Bernoulli ok.

So, this implies that we gain when  $p \times 1-p$  is small ok. So, for example, here is just a case just as an example there could be  $n$  Bernoulli random variables. So,  $X_i$  iid Bernoulli  $p_i = 1/n$  to  $n$  where each so, let us say the variance of  $X_i$  is for instance  $1/n$  itself ok. So, in that case, so, one way in which the variance could be  $1/n$  is basically if  $p$  is itself of order  $1/n$  ok.

So, if that is the case then you basically get a constant deviation, constant scale deviation here, because you have inside the  $\sqrt{n \times 1/n}$ , which is much better than what Hoeffding could give you ok. So, basically you gain when the variance of the  $X_i$ 's is fairly small.

So, that is where Bernstein is often used to get you the additional gain from exploiting random variables that are effectively quite small as measured by their values, that concludes this lecture.

Thank you.