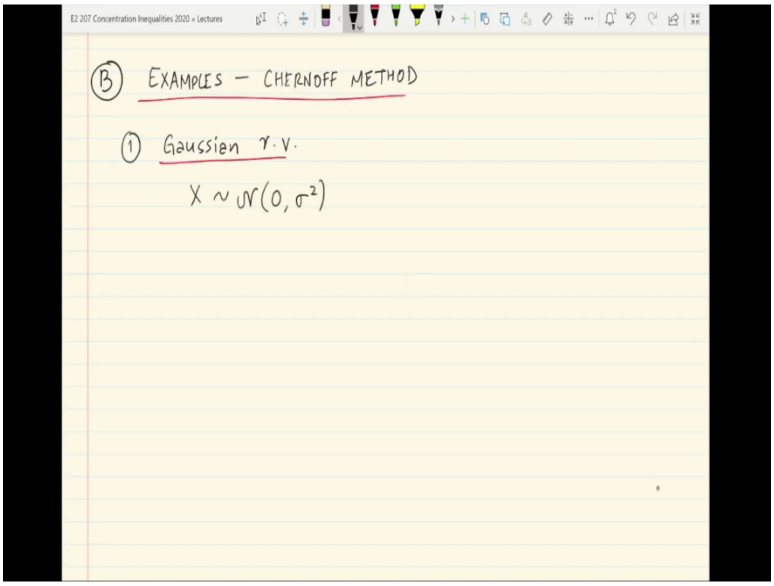
Concentration Inequalities

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

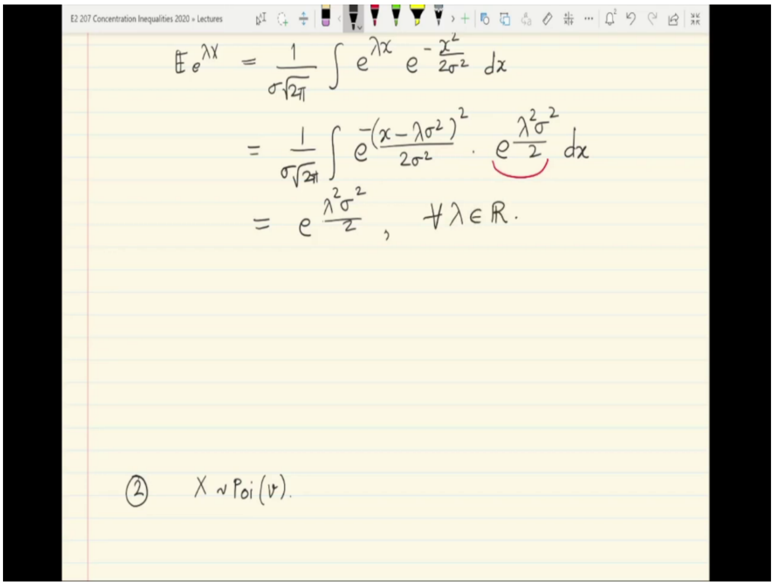
Lecture - 03
Examples of Chernoff Bound for Common Distributions

(Refer Slide Time: 00:21)



The first example of a random variable to which we will apply the Chernoff method is the Gaussian random variable. So, let us assume that X is a Gaussian random variable, with mean 0 and variance σ^2.

(Refer Slide Time: 00:38)



Let us compute the log the moment generating function of X. So, expected value of e raised to λ X / definition is 1 over σ 2 $\sqrt{2}$ π and the integral of e raised to λ X against the Gaussian density. So, this evaluates to. So, what one can do is in the exponent one can complete the 2 here to make it x - λ σ^2 the whole 2 - 1 /2 σ^2 into e raised to λ^2 σ^2 /2 dx.

Thus, this second term is a constant; does not depend on x and comes out and the rest of the expression is just the integral of the density of a Gaussian that is shifted and so, that integrates to 1. So, we have that the moment, the moment generating function of a Gaussian is e raised to λ^2 σ^2 /2 ok and this is for every λ.

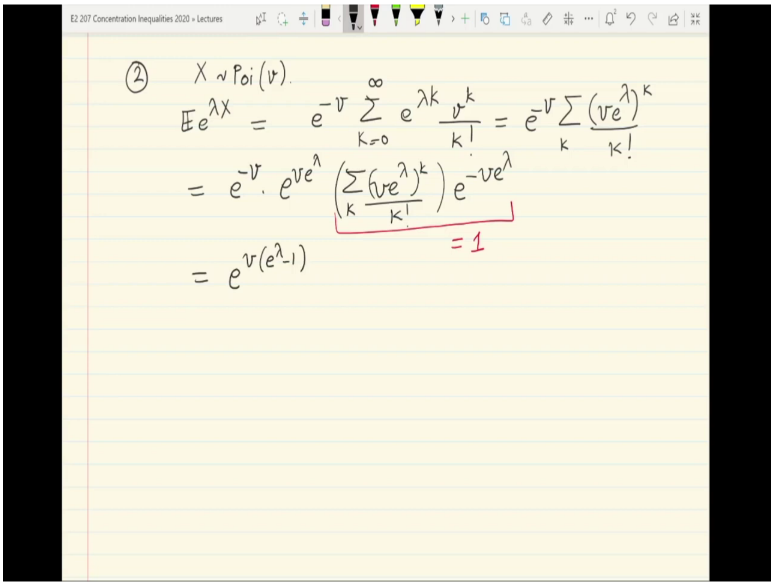
(Refer Slide Time: 02:00)

which finally, gives you t /2 σ^2, as an explicit closed form expression for the Cramer transform of Gaussian.

So, in fact, even if t is negative, you can apply the same argument to the random variable - x and finally, conclude that Ψ x * of t is this. Basically a standard quadratic function given / t 2 /2 σ^2. So, the Chernoff bound here for a Gaussian reads probability that X $\geq$ t is $\leq$ e raised to - t^2 /2 σ^2, whenever t is $\geq$0 ok.

So, you can think of this as a bound on the tail or the complementary cumulative distribution function of a Gaussian on the right side. With this, we will go to our next example which is that of a Poisson distributed random variable. So, let us say that X is distributed as a Poisson random variable, with the parameter v which is the same as its mean.

(Refer Slide Time: 04:03)

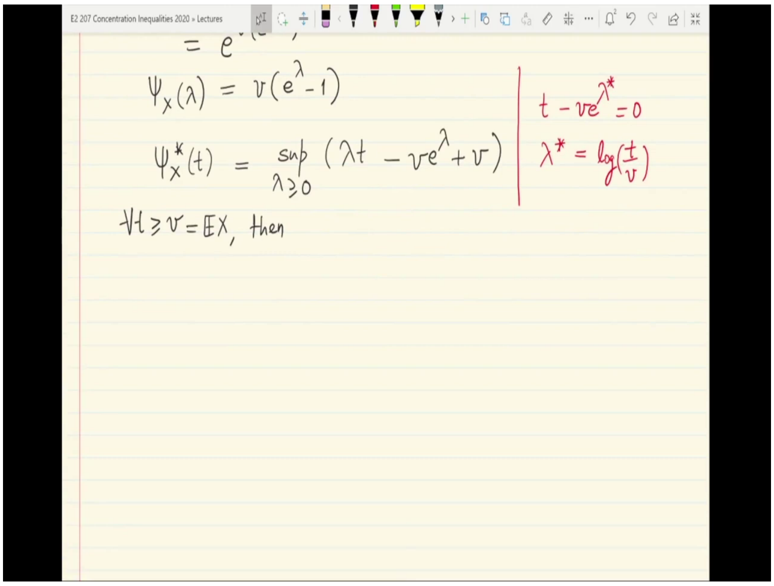


So, in this case, the expected value, the moment generating function of X is e raised expected value of $e^{\lambda X}$ which is given by e raised to - v the sum of the possible values that a Poisson random variable can take which is discrete values from 0 to ∞, e raised to λ K v raised to K / K factorial ok.

Now, one can evaluate this as follows. One can write this as e raised to - v sum over K. Let us collect v into e raised to λ K; v into e raised to λ and raise it to a common power K divided / K factorial and what one what one can do is we can multiply and divide by the quantity e raised to v e raised to λ.

And then, you have the sum over K v e raised to λ raised to K / K factorial and we have to divide / e raised to v e raised to λ and notice that / the definition of the Poisson probability mass function, this sum is = 1. It is just the sum of the Poisson PMF with the parameter v e raised to λ instead of v. And so, you get that the moment generating function is finally, e raised to v into e raised to λ - 1, all of it in the exponent.

(Refer Slide Time: 05:45)

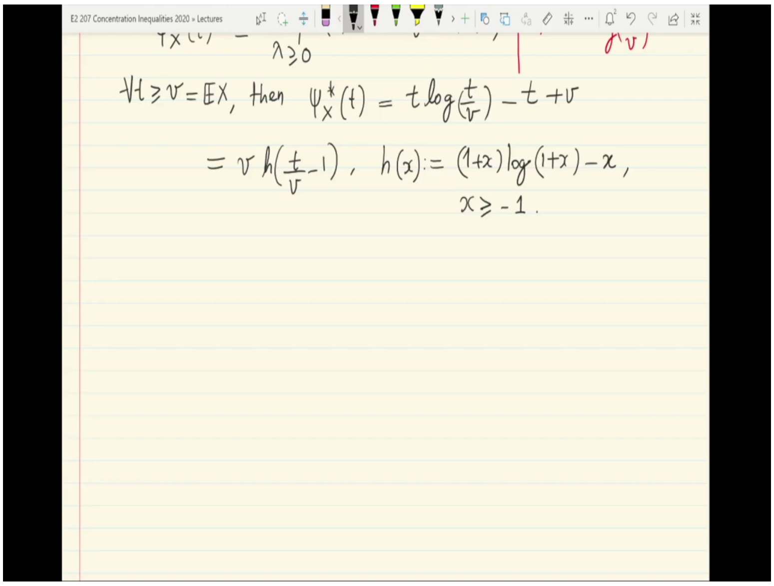


So, this means that the cumulant generating function of X is v × e raised to λ - 1 ok and in turn the Cramer transform $\Psi_X{}^*$ of t is / definition. The supremum over λ ≥0 of λ t - Ψ_X of λ which we just evaluated as v e raised to λ - v. So, there is a + v here ok.

So, if t is > = v which is the mean of the Poisson random variable then; so let us try to just naively optimize this objective function pretending that it is an unconstrained problem. So, the natural thing to do is to take the derivative of the objective function and equate it to 0 and solve for λ.

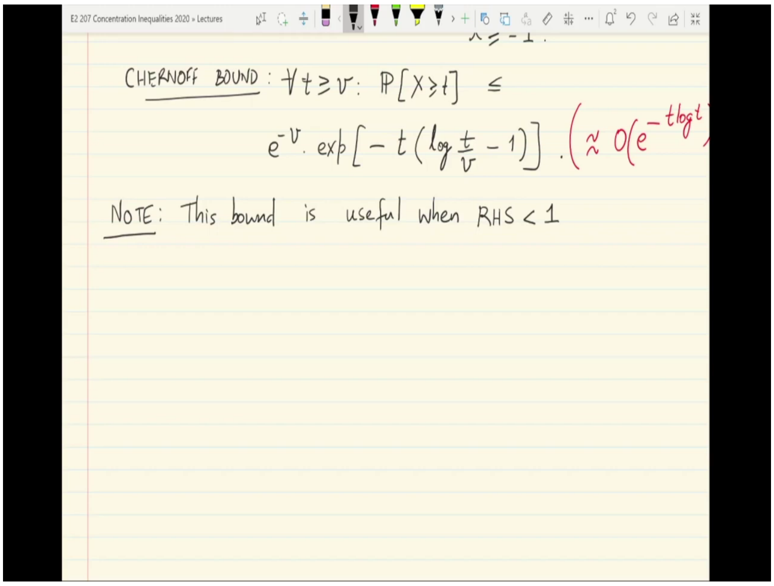
So, what you get when you do that is that t - v e raised to λ * is 0 at the optimum λ *. So, that just means that λ * is = log t / v ok. And if t is larger than v, then log of t / v is a quantity larger than log 1 which is at least 0. So, λ * indeed turns out to be ≥0, if t is ≥ v that is the solution of the unconstrained maximum is the same as the solution of the constrained maximum.

(Refer Slide Time: 07:33)



So, one has that $\Psi_X{}^*$ of t / explicit solution is nothing but t × the optimal λ which is log t / v - t + v and let us write this as let us denote this as v × h of t / v - 1 for reasons that we will see later ok. With h of x the function h of x being defined to be 1 + x log 1 + x - x, for defined for every x ≥ - 1 so that the log is well-defined. So, you can see that this is $\Psi_X{}^*$ of t is exactly v × h of t of t / v - 1 ok.

(Refer Slide Time: 08:38)

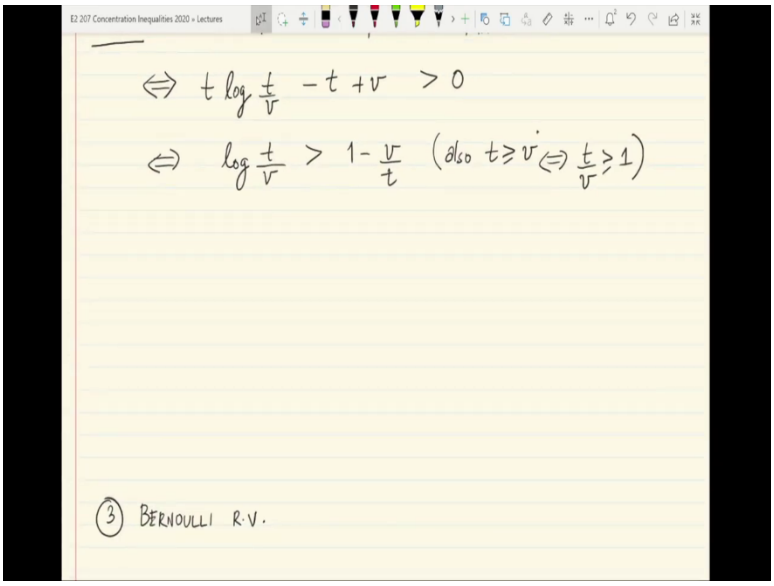


So, the Chernoff bound for a Poisson reads as follows. We have for all t ≥v probability x ≥ t is ≤e raised to - v into e raised to - t × log t / v - 1. So, as t becomes larger and larger, this bound essentially goes down as e raised to - t log t ok.

So, this is roughly order wise. So, this is order wise e raised to - t log t. So, you can think of it essentially as having an exponential type t ok. Slightly heavier than an exponent, slightly lighter than an exponential tail because of the presence of log t that multiplies the t in the exponent.

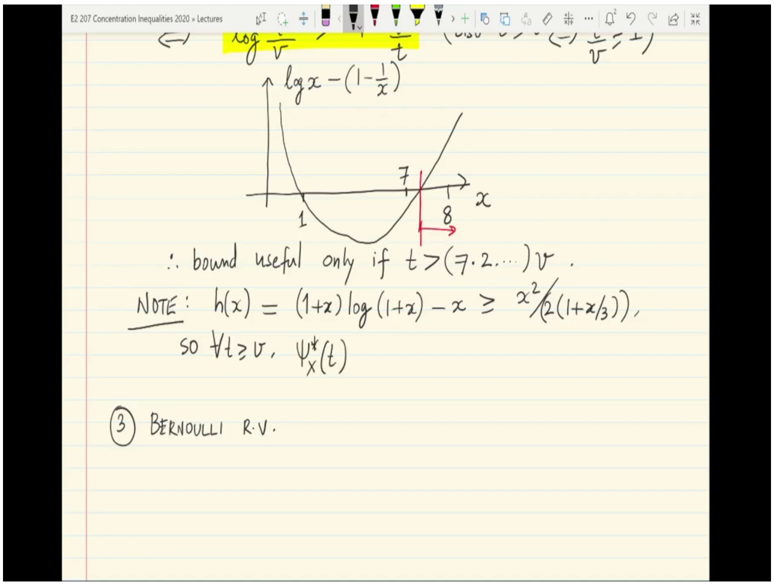
A note here is that this bound is useful ok, is practically useful naturally only when the right hand side which is supposed to be a probability is at most 1 ok. So, when the right hand side is strictly less than 1 is when it becomes useful. If the right hand side is larger than 1, anyway we know that the probability is has to be bounded by 1.

(Refer Slide Time: 10:29)



So, let us evaluate when for what values of t this right hand side becomes starts becoming lesser than 1. This is the same as saying that t × log t / v - t + v is larger than 0 which in turn means that log t / v is larger than 1 - v / t ok. So, note that, we have also assumed that t is larger than = v which is the same as saying that t over v is at least 1. So, let us think of t over v as some number x, this ratio x.

(Refer Slide Time: 11:23)



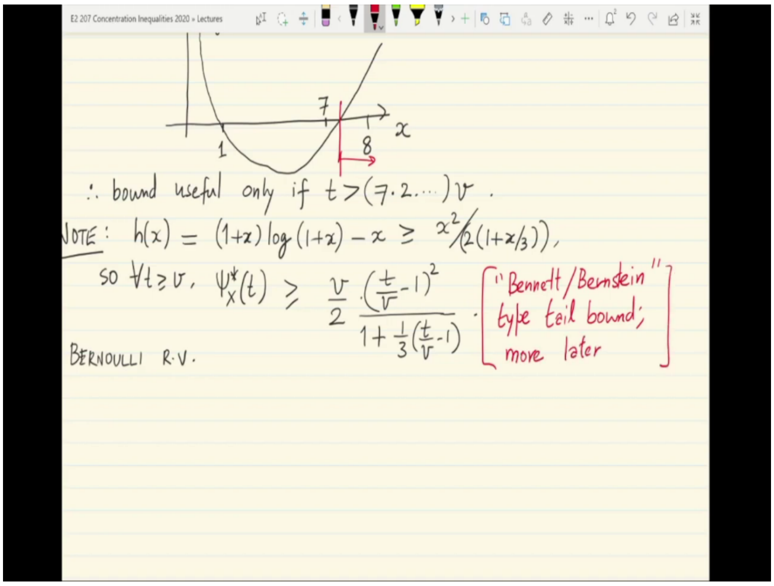
If you plot this, let us try to plot on the x axis, the quantity x; on the y axis, what we will do is we will plot the quantity log x. We will plot the difference between the left and right hand sides of this expression. So, log x - 1 - 1 / x and if you plot this function, it will it will look as follows. At 1, it will become 0 and it will rise only after some time. So, before the rise comes the number 7 and after the rise comes the number 8. So, at about 7.2 or something, this function again starts to rise above 0.

So, this is how log x - 1 - 1 /x looks and if we want this inequality to be useful, we know that we can only live with x to the right of this point of intersection with the x axis. So, this roughly means that the bound is useful only if t is larger than let us say some number called 7.2 something some fraction × v ok.

So, when you have t exceeding this number, this bound starts giving you something non trivial. Another note here is that we do not need to be you know we do not need to necessarily work with such a complicated form for the Cramer transform. We can also try to lower bound the Cramer transform Ψ_X * of t to get looser, but perhaps more insightful bound.

So, one way of doing that for the Poisson distribution is as follows. We have that h of x, this function h of x that we wrote to write down the Ψ_X * of their ok right here. This function, it is a its it is an easy exercise to show that this function has a nice lower bound given / x 2 divided /2 × 1 + x / 3 ok; all of this in the denominator and so, this means that for all t ≥ v, let me make some more space. So, for all t ≥v, $\Psi_X{}^*$ of t which was just h of t / v -1 into v.

(Refer Slide Time: 14:25)

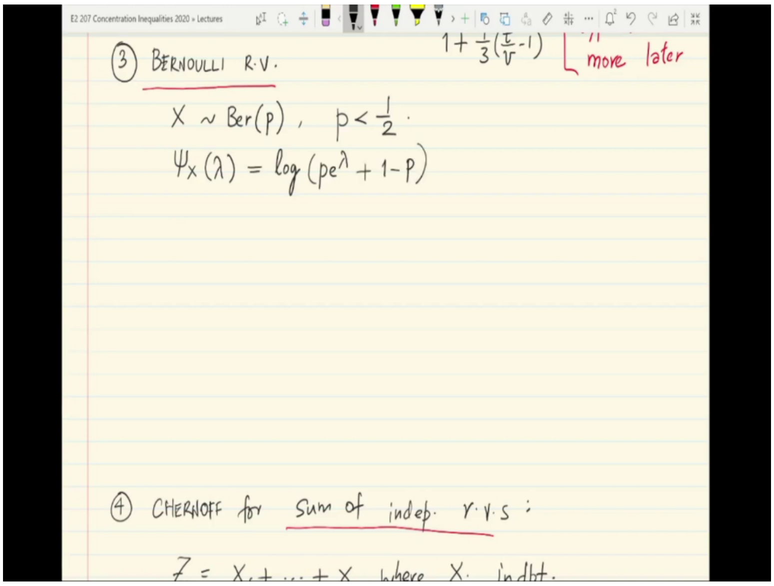


You can lower bound this as v /2 into t / v - 1 the whole 2 divided / 1 + one-third of t / v - 1 ok. So, we actually have a lower bound for $\Psi_X{}^*$ of t the Cramer transform in the exponent of the Chernoff bound in terms of as a rational function ok. So, as a ratio of two polynomials depending on t or t / v.

So, it is a quadratic form on the top, on the numerator and a linear form on the denominator. So, depending on whether t / v is close to 1 or not, you have either quadratic behavior for $\Psi_X{}^*$ of t for large t or for relatively smaller t, you basically have linear type behavior ok.

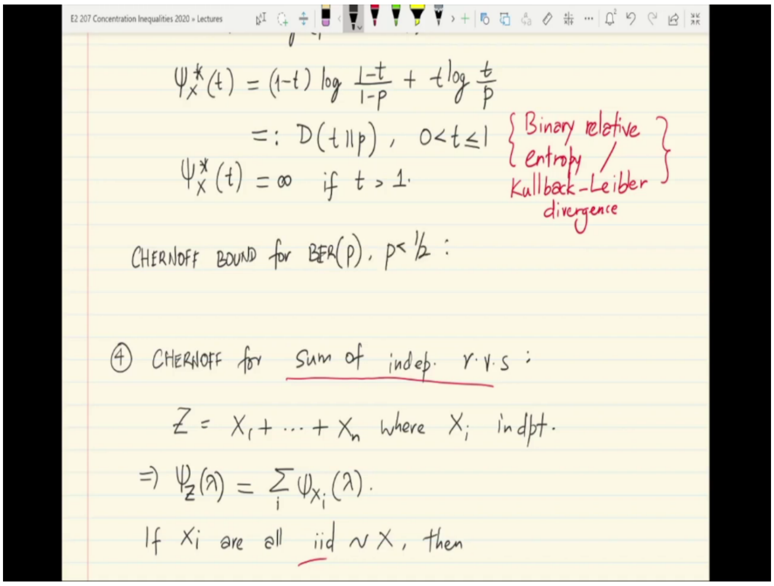
So, this is what is called a Bennett or a Bernstein type tail inequality tail bound ok. We will see this in more detail later. This example was just to show that you can get non trivial lower bounds for the Cramer transform of random variables. The third example is a Bernoulli random variable, which is probably one of the simplest random variables.

(Refer Slide Time: 16:12)



So, let us take X as being Bernoulli distributed with parameter p and without loss of generality, we will take p to be less than 1/2, the too much loss of generality. So, you can do the calculations. To do, to see that Ψ X the log moment generating function of X is simply log p (e $^{\lambda}$ + 1 – P).

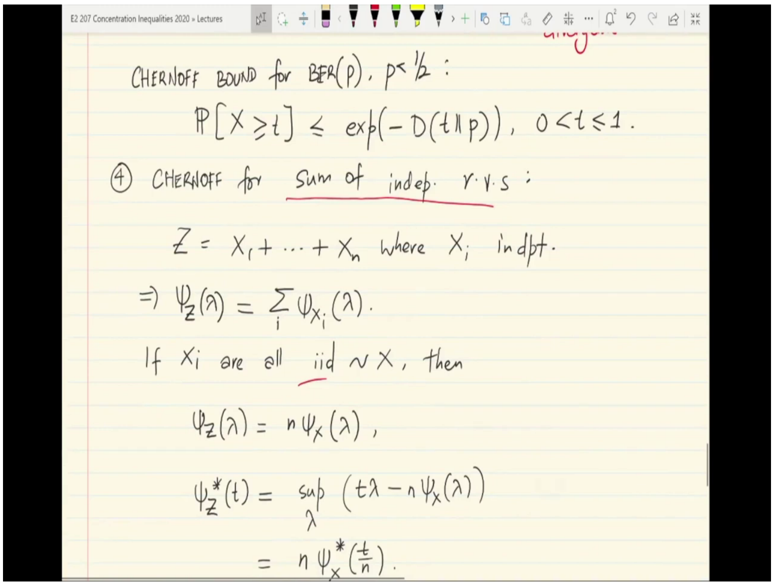
(Refer Slide Time: 16:45)



And if one continues further, one can also calculate Ψ X * of t as 1 - t log 1 - t / 1 - p + t log t / p which to some of you who have seen some amount of information theory or related areas is exactly what is called the binary relative entropy of t versus p for t lying between 0 and 1 ok. So, this is called the, this is a very important object in information theory and statistics. It is called the Binary relative entropy function or also the Kullback-Leibler divergence between Bernoulli distributions with parameters t and p.

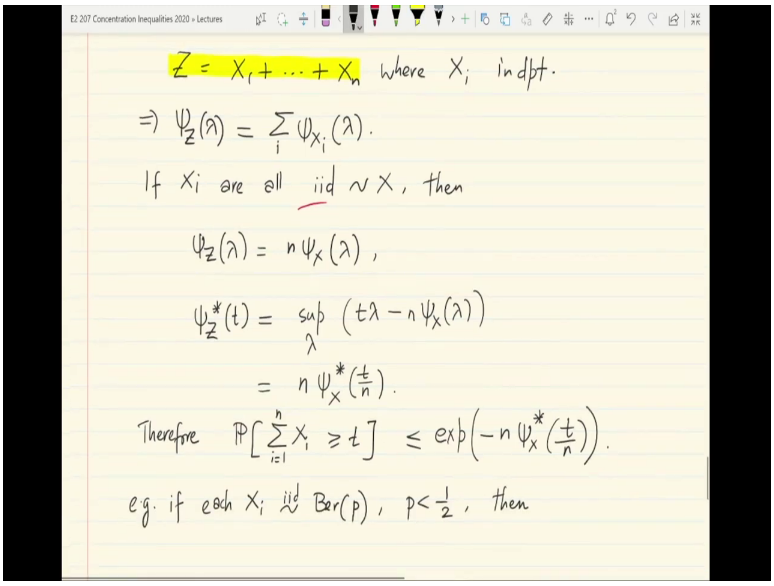
By the way if t is larger than 1, then it is easy to show that Ψ X * of t is = ∞, if t is >1 or > yeah t is larger than 1 ok. So, one can put an = sign here.

(Refer Slide Time: 18:40)



And hence, the Chernoff Bound for a Bernoulli random variable with p less than 1/2 basically gives you that probability that x is larger than = t less than = e raised to - D t p for 0 less than t less than 1 ok.

(Refer Slide Time: 19:03)

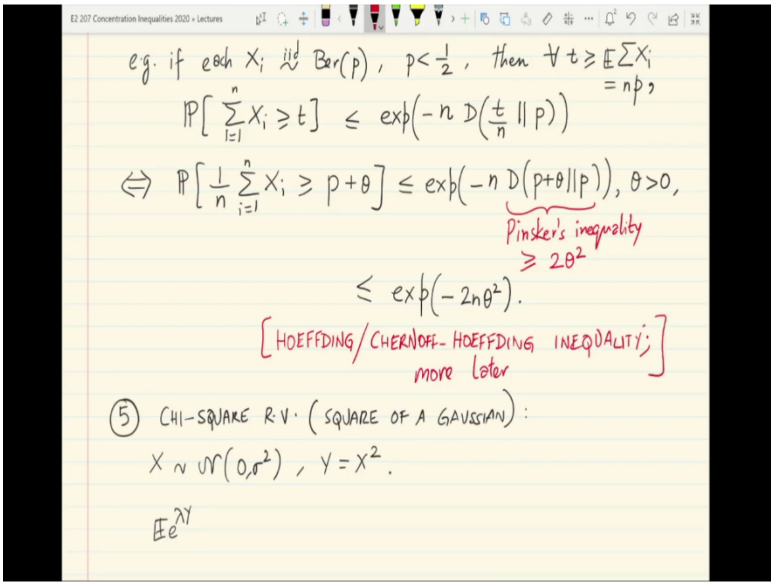


The next application of the Chernoff bound will be for a sum of independent random variables. So, think of a random variable z expressed as the sum of n independent random variables X i; X 1 through X n. So, if you start computing the cumulant generating function of z.

So, Ψ_Z of λ just ends up being the sum of the individual cumulant generating functions of all the Xi's evaluated at λ. And moreover, now if all the X i's are also identically distributed in addition to being independent, then this Ψ_Z of λ just becomes n × of each Ψ_X of λ.

And consequently, the Cramer transform Ψ_Z * of t just ends up being n × n × the Cramer transform of each random variable evaluated at t over n ok. This is just by algebra and using the definition of the Cramer transform. Therefore, we have the Chernoff bound probability that the sum exceeds t is ≤ e raised to - n × Ψ_X^* of t / n, where X is X represents any of the X i's.

(Refer Slide Time: 20:30)



So, for instance, if you apply this to the sum of Bernoulli random variables which is actually a binomial random variable with parameter n parameters n and p, where p is less than 1/2 . Then, what we get is that the probability that their sum exceeds t is bounded / e raised to - n × the Cramer transform of a Bernoulli which is D the relative entropy with parameter t over n relative to P.

So, in other words, if you normalize, so this isby the way for this is by the way for t ≥ the mean; ≥ expected value X i which is = np in this case. We have the following one and one can just divide / n throughout and reparameterize to get the following result.

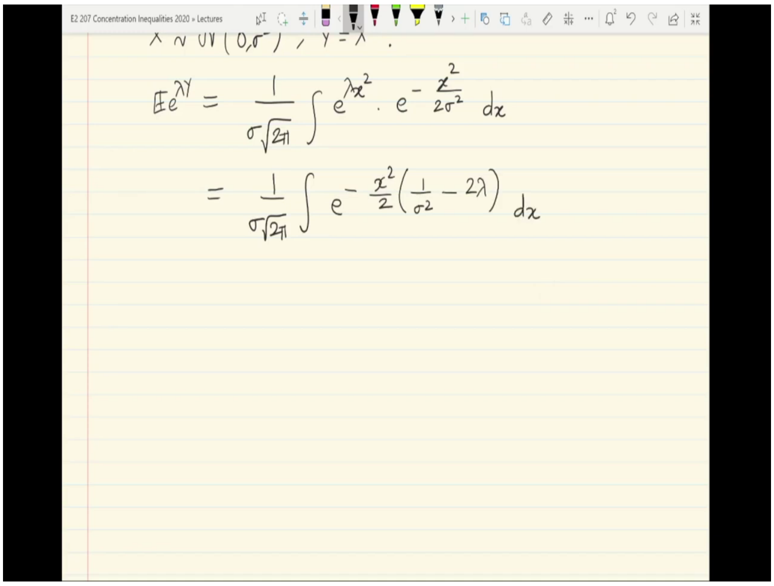
So, the probability that the sample mean of n Bernoulli’s exceeds the true mean p + a positive number θ is at most e raised to - n D of p + θ relative to p ok for I guess $\theta > 0$ ok. So, you get this kind of tail inequality and note that the nice thing here is that as n increases the right hand side goes down exponentially with them. Because D of p + θ relative top is just a constant, it does not depend on it.

Moreover, there is an inequality called Θnsker’s inequality, well-known information theory in the convex analysis that helps to lower bound the relative entropy / what is called the total variation distance or the L1 norm. And in this case, this just means that D of p + θ relative to p can be bounded below by 2 θ^2 ok, irrespective of the p.

So, finally, you have that the probability of the sample mean exceeding the true mean / an amount θ is at most e raised to - 2 n θ^2 ok. So, we will see this later, but this is also what is called Hoeffdings inequality or Chernoff Hoeffding equality for Bernoulli random variable ok. So, more later.

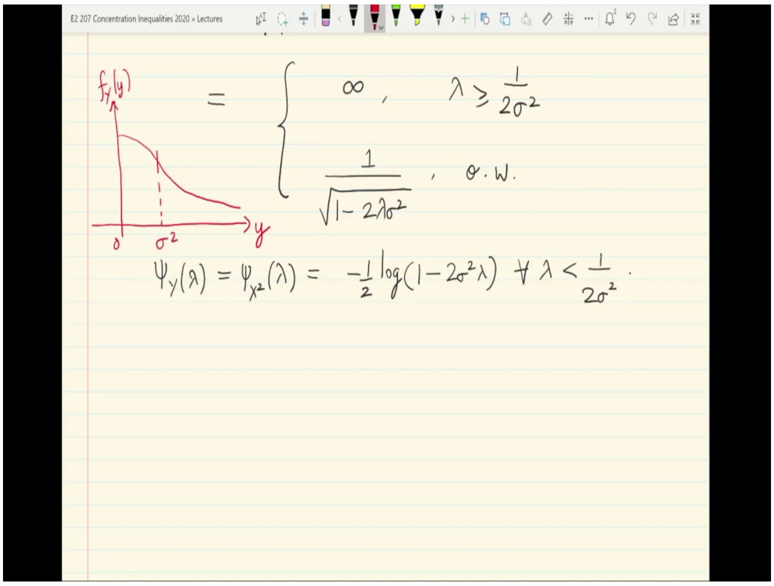
The last example is going to be the derivation of the Chernoff bound for what is called the chi-2 random variable. In simple terms, this is just the 2 of a Gaussian random variable. So, think of this think of squaring a Gaussian random variable and that becomes a one special type of χ-2 random variable. So, let us say that X is normally distributed and you 2 X to get Y.

(Refer Slide Time: 24:19)



So, the moment generating function of Y will be one over σ 2 $\sqrt{}$ 2 π integral of e raised to x^2 λ x^2 because that is λ Y into e raised to - x 2 /2 σ 2 d x. This is just = 1 over σ $\sqrt{}$ 2 π integral e raised to - x^2 /2 collecting terms 1 over σ 2 - 2 λ dx outside.

(Refer Slide Time: 25:01)

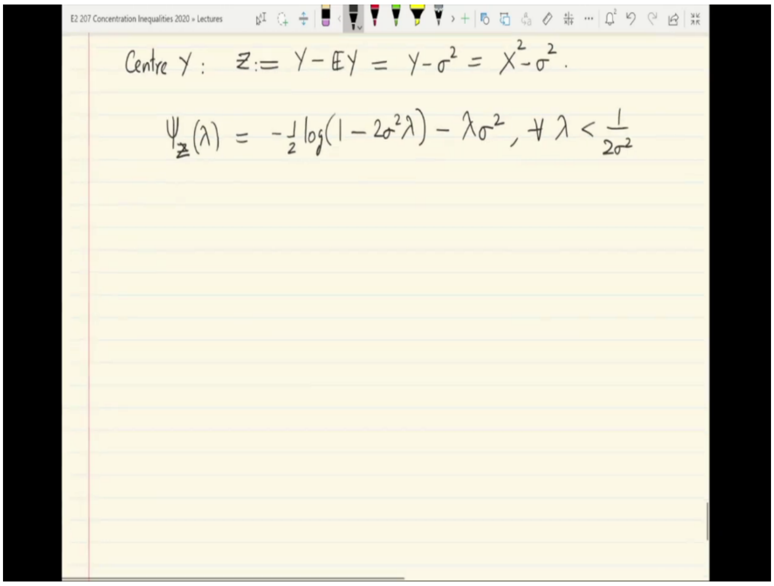


And you can convince yourself that if the exponent of the if the exponent of the integrand turns out to be positive, then the answer is ∞. So, this is when λ exceeds 1 over 2 σ^2 and otherwise, you can integrate this explicitly to get 1 over 2 √ 1 - 2 λ σ^2 otherwise ok, so if λ is smaller than this critical value.

So, this means that Ψ Y of λ which is Ψ X 2 of λ is = - 1/2 log 1 - 2 σ 2 λ, for λ less than 1 /2 σ 2. So, we will find it convenient to proceed with our derivation, if we centre Y ok. So, so recall that Y is X 2 and let us center Y ok.

So, so just to paint a πcture of how things look Y is X^2. So, if you plot the density of y, it will basically look something like see, like this and we know that the expected value of Y which is the expected value of X 2 is σ^2. So, this is just going to centre Y / σ^2.

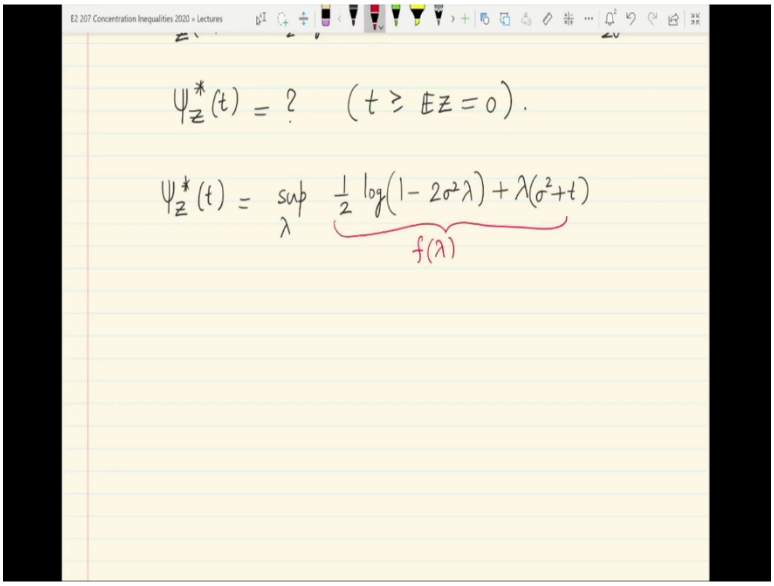
(Refer Slide Time: 26:51)



So, let us centre Y and define Z as Y - expected value of Y which is $Y- \sigma^2$ which is x - $X^2 - \sigma^2$. So, that Z is a 0 mean random variable. By centering a random variable, we will always mean subtracting its mean to get a zero-mean random variable.

So, we have Ψ Z of any λ as a - 1/2 log 1 - 2 σ 2 λ which is the cumulant generating function of λ and from that we have to just subtract $\lambda \times \sigma^2$ which is the mean of y, this holds for all λ less than 1 /2 σ 2 as usual.

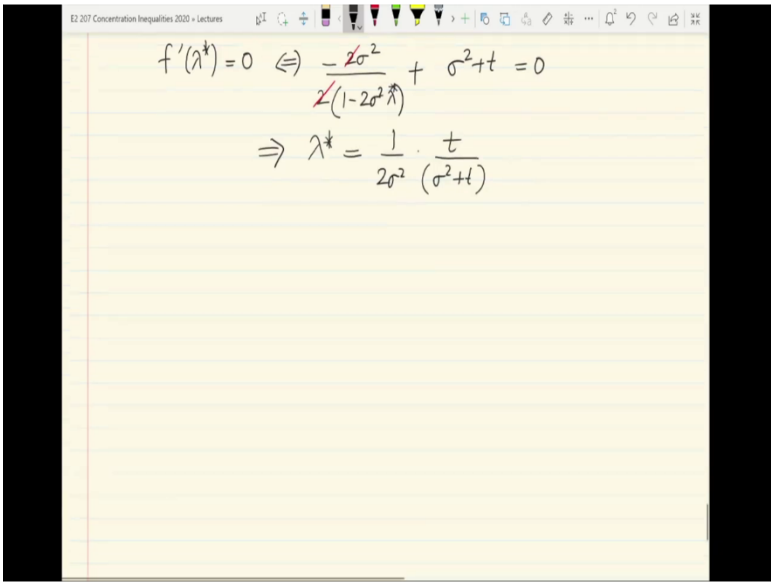
(Refer Slide Time: 27:50)



So, what about the main object of interest which is the Cramer transform of the centered Y which is Z ok? Let us say assuming t is > = the right of expected value of Z which happens to be 0 ok. So, t is any positive number and we would like to compute Ψ Z * of t to get an idea of the tail behavior or the Chernoff bound.

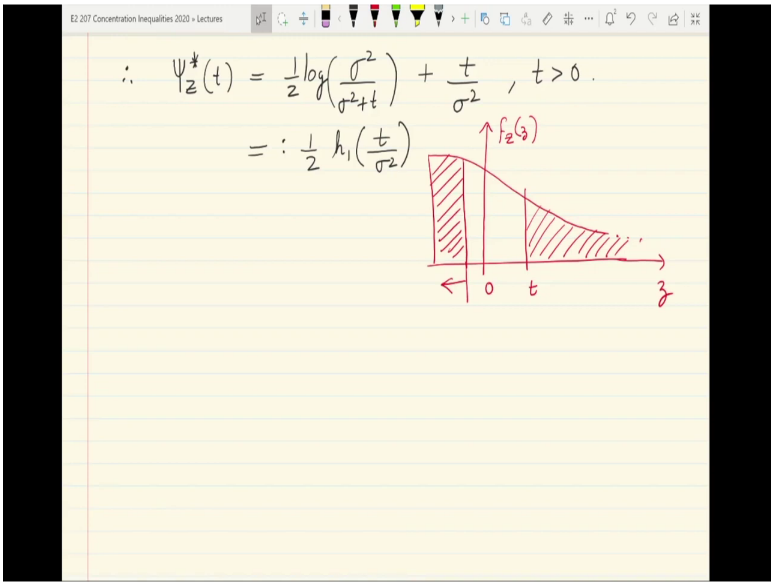
So, to do that, let us first *t / writing Ψ Z * of t. This is / definition the sup over λ of 1/2 log 1- 2 σ 2 λ + λ σ ^{2} + t. This is by the definition of the Cramer transform. Let us call this as f λ. We know that if t is ≥ the mean, you can always perform the unconstrained minimization ok. So, λ belonging to R. So, how do we do that? We will differentiate f and set it to 0.

(Refer Slide Time: 29:00)



So, f prime at $\lambda^* = 0$ is equivalent to saying $-2\sigma^2$ / twice $1 - 2\sigma^2\lambda + \sigma^2 + t = 0$. So, you have a cancellation here and finally, what you will get is that λ^*. So, this is really λ^* here; λ^* is just going to be 1 over $2\sigma^2$ t over $\sigma^2 + t$.

(Refer Slide Time: 29:43)



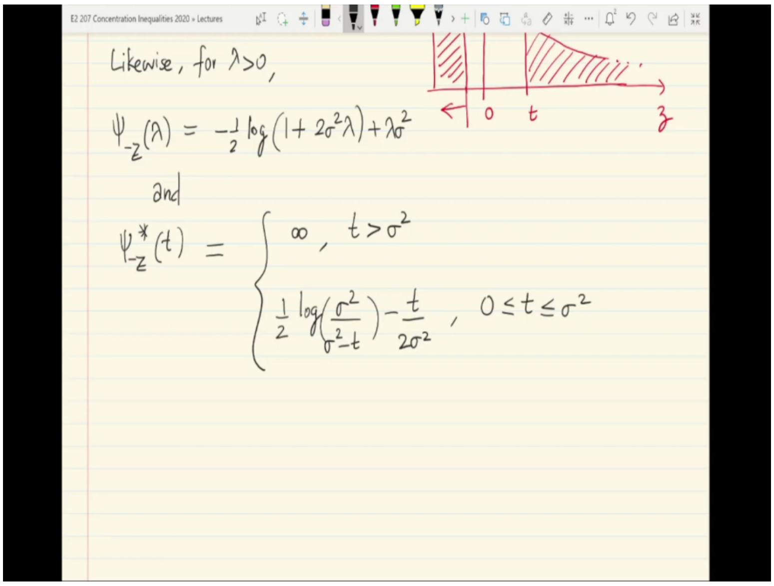
So, substituting this value of λ back, we get that Ψ Z * of t $= 1/2 \log \sigma^2 / \sigma^2 + t + t / \sigma^2$, for whenever t is >0. So, let us let us define this to be a new function which is 1/2 of h 1 of the quantity t / σ^2.

So, notice that the only thing that matters is a ratio of t over σ^2 ok. So, we have evaluated Ψ Z * of t up to some reasonable level of level of clarity. So, it is worth pausing to examine what we have done in a more πctorial sense. So, recall that we have basically taken.

So, Z was essentially the centered version of Y and so, its pdf will look something like this f Z of z ok with mean 0. So, we have essentially found the Chernoff bound essentially allows us to bound the tail probability of the random variables at going to the right of this number t.

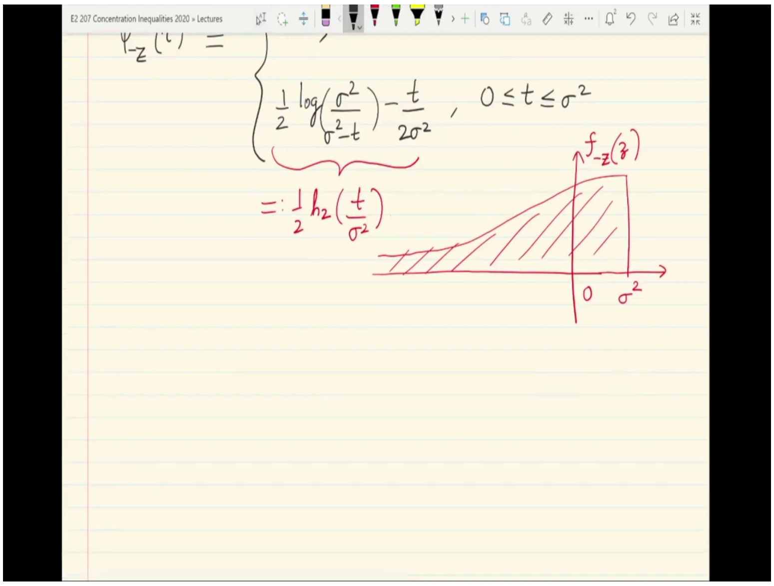
Now, what about the left tail? Ok; how does the left tail look? What about if we wanted to investigate, how this probability went ok? So, in order to do that, we can one way to do that is to just invert to negate z. So, consider - z ok. So, let us do that. So, let us consider Ψ - Z ok.

(Refer Slide Time: 31:55)



So, so, likewise, we can do similar computations for λ >0, if we *t with - Z, so Ψ - Z of λ turns out to be $- 1/2 \log 1 + 2 \sigma^2 \lambda + \lambda \sigma 2$ and Ψ - Z. For the random variable - Z, it is Cramer transform turns out to be ∞ if t is larger than σ^2 and it turns out to be this quantity 1/2 $\log \sigma^2 / \sigma^2 - t - t / 2 \sigma^2$, if 0 is less than = t less than = σ^2. Now, what is happening here?

(Refer Slide Time: 33:06)

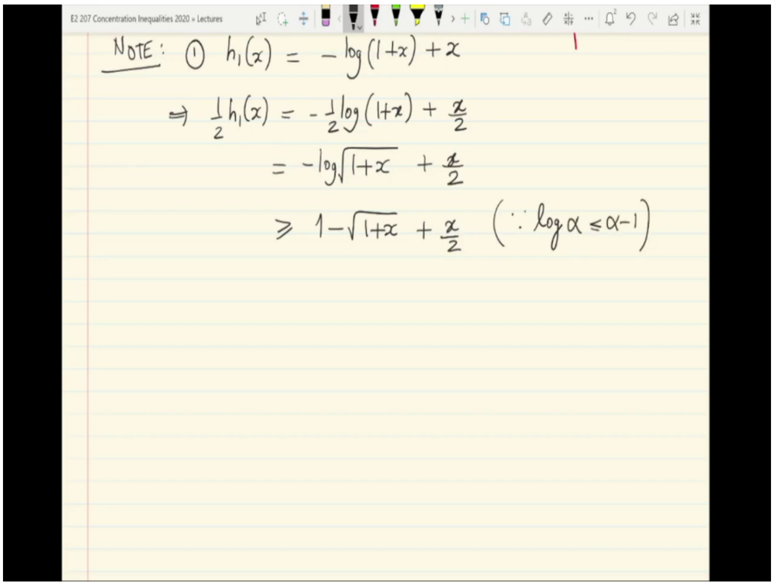


So, it will be useful to draw a probability this density function for the random variable - Z. So, on the x axis, if I plot z and on the y axis, if I plot f - Z of z. So, what is - Z? - Z is just so let me extend this a little bit; - Z is having 0 mean and it is a negation of Z ok. So, it is just going to be having a pdf that fall something like this ok. With this quantity being σ^2 ok, because you have shifted y / an amount σ^2 to get z ok.

So, its density is between - ∞ and σ^2. So, which is the reason why if t is larger than σ^2, then the Cramer transform of - Z is simply ∞ and its otherwise finite. So, let us denote this finite function / 1/2 h^2 of t / σ^2 ok.

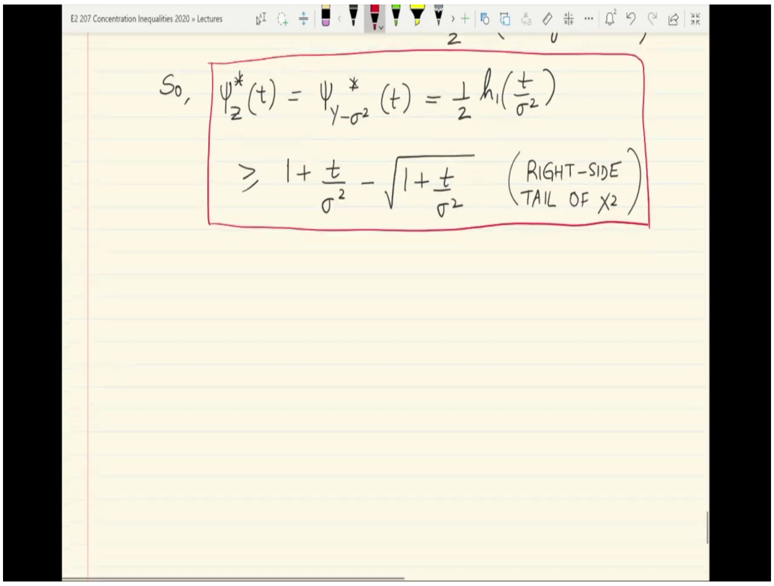
So, we have essentially found the right tail decay and the left tail decay of a chi-2 random variable which is a Gaussian 2 ok; but the Cramer transforms are rather complicated functions, they are expressed as logarithms of rational functions of t and so on. So, let s try to understand these functions, these Cramer transforms of Z and - Z, a little better / trying to lower bound them and one way in which we can do this is as follows.

(Refer Slide Time: 35:09)



So, let us *t with h 1 ok. So, this is our first note h 1 of x recall that it was - log 1 + x + x. So, this means that 1/2 h 1 x = - 1/2 log 1 + x + x /2. This is just this the same as - log 2 √ 1 + x + x /2 and this can be lower bounded by 1 - 2 √ 1 + x + x / 2. This is because for any number α which is >0 log α is less than = α - 1 universally.

(Refer Slide Time: 36:16)

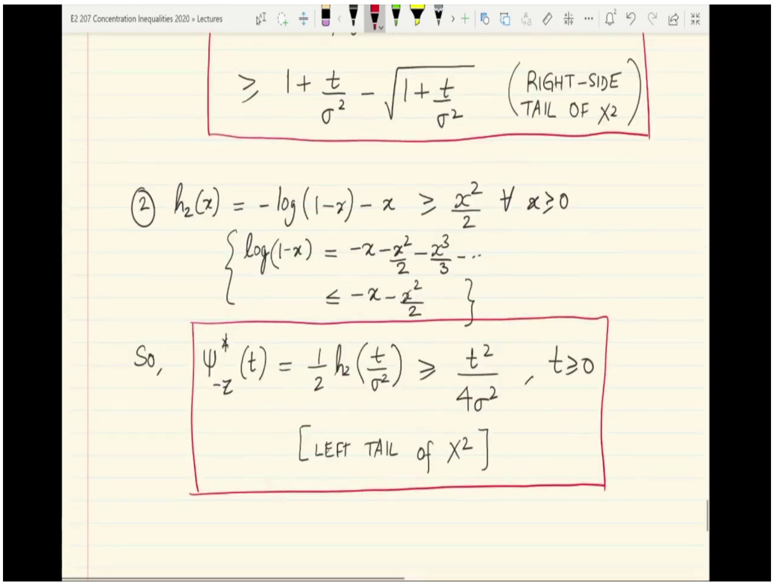


So, that is what gives rise to this form. So, using this, we get that the Cramer transform of Z at t which is the same as Ψ Y - σ^{2} * at t, which was earlier found to be 1/2 h 1 t / σ^2; using this bound is at least 1 + t / σ^{2} - 2 √ of 1 + t / σ^{2} ok.

So, this is what is the right tail bound; the right Right-Side Tail bound for Y or for X^2 really, ok, which gives you basically some insight about how this bound looks as a function of t ok. So, it is essentially a linear function in t - the 2 √ of 1 + t 1 + constant × t.

On the other hand, the left side tail of X^2 which is really the right hand side tail of - X^2 or - Z can be bounded in the following way.

(Refer Slide Time: 37:33)



So, the second point here is that h^2 of x which is - log 1 - x - x, one can easily show by an exercise that this is > = x^{2} /2, for all non-negative x ok. This is this is basically because log 1- x, if you think of its Taylor expansion. This is basically - x - x^{2} /2 – x^3 / 3 - on and since we are subtracting off, this is at most - x - x^{2} /2 ok and that is basically the inequality above ok in sπrit.

So, what this finally gives us is that the left tail of Z which is really the right tail of - Z for any t > = 0 which is just 1/2 h^2 t / σ^{2} is lower bounded by t^{2}/ 4 σ^{2}; just a quadratic function. So, ok so, this is the Left Tail of X^2 ok centered suitably.

So, if you compare, if we compare the left and right tails, the tail decay, the exponential tail decay suggested by the Chernoff bound, what we find is that the right side tail essentially decays as e raised to - constant × t e raised to - t like an exponential; whereas, the left tail decays as e raised to - x^2 ok, e raised to - t^2.

So, the right tail is much more heavier than the left tail ok. So, this is to be expected because the left tail is really bounded, it does not go beyond a certain value. So, we will probably discuss more of this when we come to Sub-Gaussian random variables and variables that are heavier than Sub-Gaussian. That concludes this lecture.

Thank you.