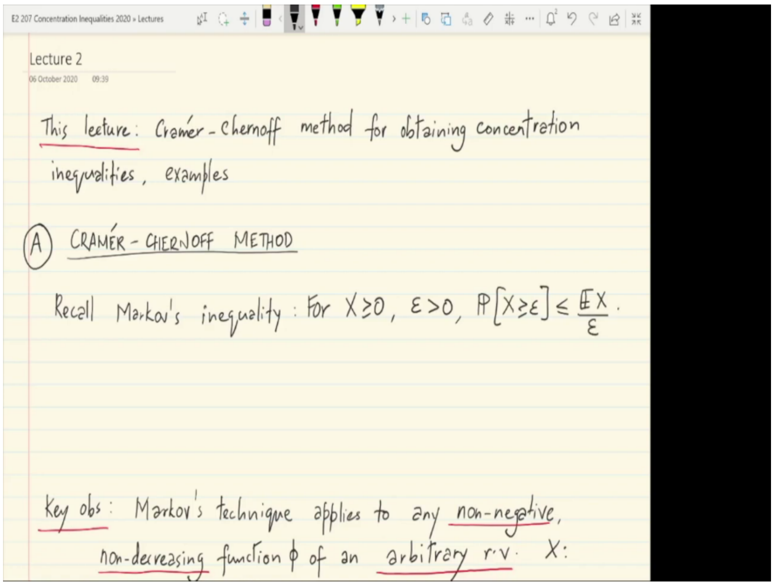
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture – 02
Chernoff bound

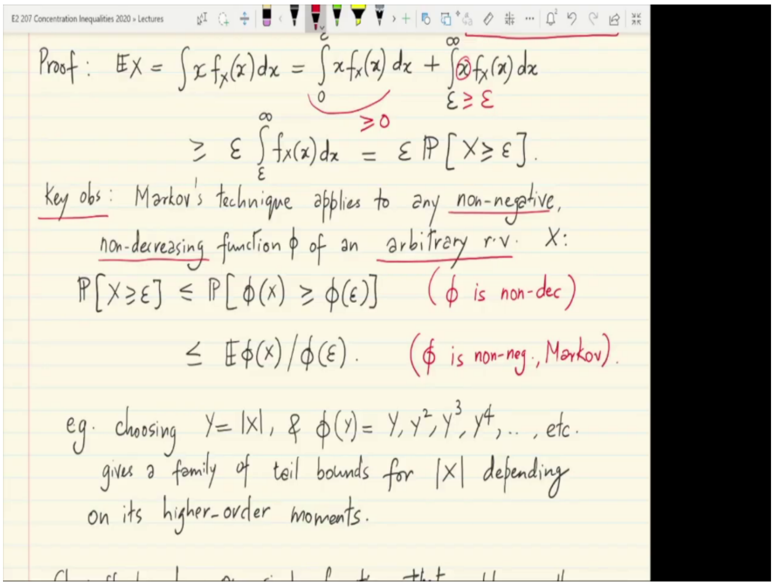
(Refer Slide Time: 00:20)



Hi all welcome to this next lecture of the course Concentration Inequalities. This lecture will be about describing the Cramer-Chernoff method for obtaining concentration inequalities which is probably the simplest and most well-known one of the most well-known methods for controlling the tail of random variables. And we will also see some examples to illustrate concrete applications of this method.

So, let us *t / recalling Markov's inequality which you may have seen in a basic probability class. So, this inequality basically says that if you have a non-negative random variable X and any number let say ϵ greater than 0, then the probability that X is > ϵ is bounded / its expected value over ϵ ok. So, this is Markov's inequality ok.

(Refer Slide Time: 01:30)



Just before we proceed it its worth looking at a quick proof of this inequality. So, this just comes about / writing down the formula for the expectation. So, let us just do the proof in the special case that X has a density function the general case is simple enough simple in analogy. So, this is the integral of x versus the density function of x d x on the entire real line.

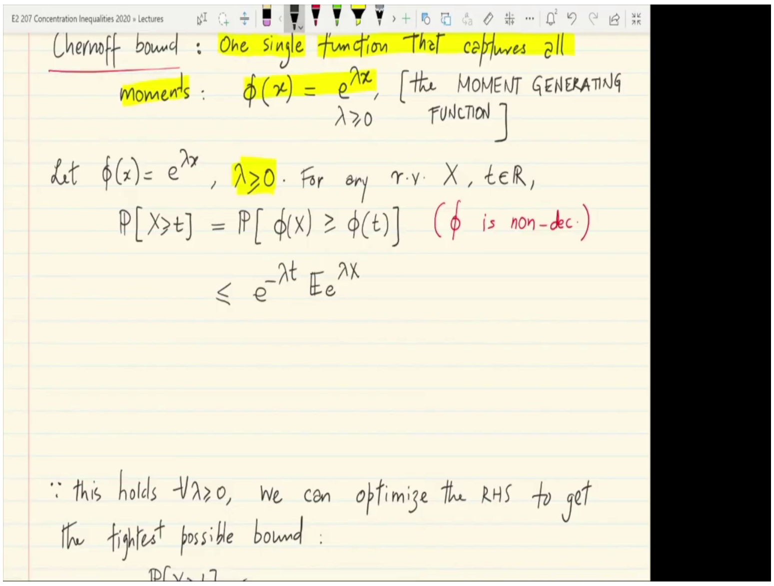
We split this in two parts. Since x is non-negative, we can first integrate from 0 to ϵ, and then integrate from to ∞ ok. So, this part is bounded below / 0 trivially. And this part x is always > ϵ. So, we have that you can bound this sum from below / ϵ × the integral from ϵ to ∞ of f x of x d x which is simply ϵ × the probability that x is > ϵ ok.

So, having seen Markov's inequality, a very simple but insightful observation is that the technique used to derive Markov's inequality can be applied to any non-negative, non-decreasing function φ of an arbitrary random variable X. So, X here does not need to be non-negative, whereas, in Markov's inequality you needed X to be non-negative and this happens the following way.

So, if you are interested in finding the probability that X exceedsϵ, this event implies that φ of X is at least φ of ϵ ok. This is because φ is non-decreasing and further Markov's inequality can be applied here to give an upper bound of expected value φ of X divided / φ of ϵ ok.

So, this holds for any random variable X provided you take an appropriate function φ which is non-negative and non-decreasing. So, the second inequality holds because φ is non-negative, and then / Markov's inequality ok.

(Refer Slide Time: 04:21)



So, for example, if we choose the if we choose y to be modulus of x ok, and φ of y =y, or y^2, or y^3, or y raised to 4, notice that φ is non-negative and non-decreasing on y which is non-negative ok. So, as long as y is non-negative which it is in this case, φ of y is indeed a non-negative, non-decreasing function. And you could think of applying this method to y to give you a family of tail bounds for the random variable mod of X depending on its higher order moments.

So, you would get on the right hand side, you would get things like expected value of Y, or expected value of Y^2, or expected value of Y^3, or expected value of Y raised to 4 which can be thought of as the family of higher order moments of X ok. So, you can get an entire collection of inequalities for the same event which is the probability of X exceeding ϵ ok.

So, one could presumably think of trying to find the tightest such inequality. And this is where the Chernoff bound comes handy. One way to think about the Chernoff bound its is that it basically does this whole family of inequalities depending on different moments / taking a single function that captures all the moments ok.

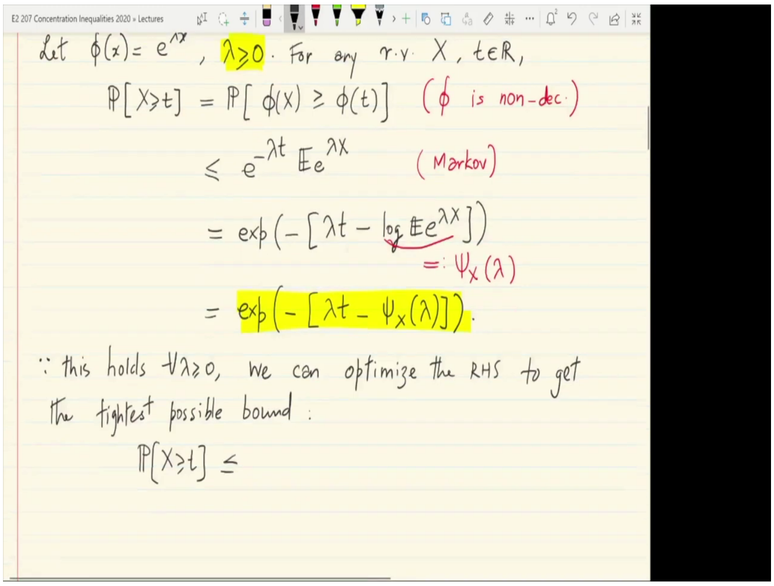
So, you may have seen this concept of a moment generating function which is this function φ x =e raised to λ x, let say for any λ ≥ 0. This function φ of x when λ is ≥ 0 is a non-decreasing and non-negative function. And when you substitute a random variable for x and take the expectation it becomes what is called the moment generating function ok.

So, really the spirit of the Chernoff bound is to try and replicate what one would do / taking / applying repeatedly Markov's inequality with different polynomial functions φ. And instead of that take a single function which is like the mixture of several polynomials with infinite degree which is the exponential function ok, so that is where Chernoff's bound usually comes in. This is one way to approach the derivation of the Chernoff bound.

So, as such the derivation is very simple. So, let us take φ of x =e raised to λ x where λ is any non-negative number. So, that φ of x becomes a valid non-negative, non-increasing function. And for any random variable X and any number t, we can have the following derivation the probability that X exceeds t is the same as the probability that φ of X exceeds φ of t.

This is because φ is as we just saw φ is non-degree non-decreasing ok with an increase in its argument. And once we have φ of φ X ≥ φ of t / Markov's inequality, we can upper bound this by expected value of φ of X divided / φ of t which we can write as follows e raised to - λ t which is the denominator into expected value of e raised to λ x ok. This is / Markov's inequality.

(Refer Slide Time: 07:57)

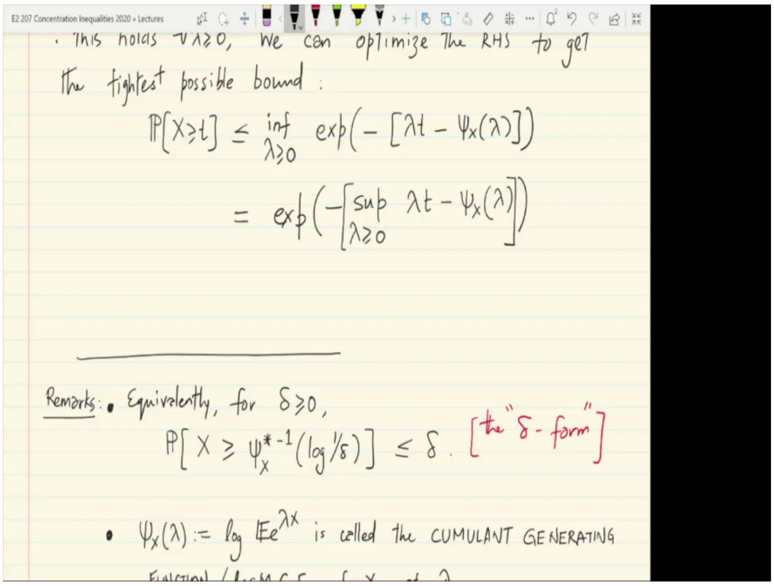


Now, let us bring all terms into an exponent form / writing as follows. This is =the exponent of negative of λ t - the logarithm of expected value of e raised to λ x. And this new function which is the logarithm of the moment generating function, we will find it convenient to denote it / φ x Ψ X of λ ok. And so finally, we get the bound e raised to negative of λ t - Ψ X of λ ok.

Now, notice that when we did this derivation, all that we needed was to choose a λ which is greater than or =0. And you get for every favorite choice of λ, you get a bound on the right hand side ok, for the same event x ≥ d.

So, since this holds for all λ non-negatives, we can optimize the right hand side above here ok which is this thing highlighted. You can think of optimizing this over all possible choices of λ ≥ 0 to get the tightest possible bound ok.

(Refer Slide Time: 09:41)

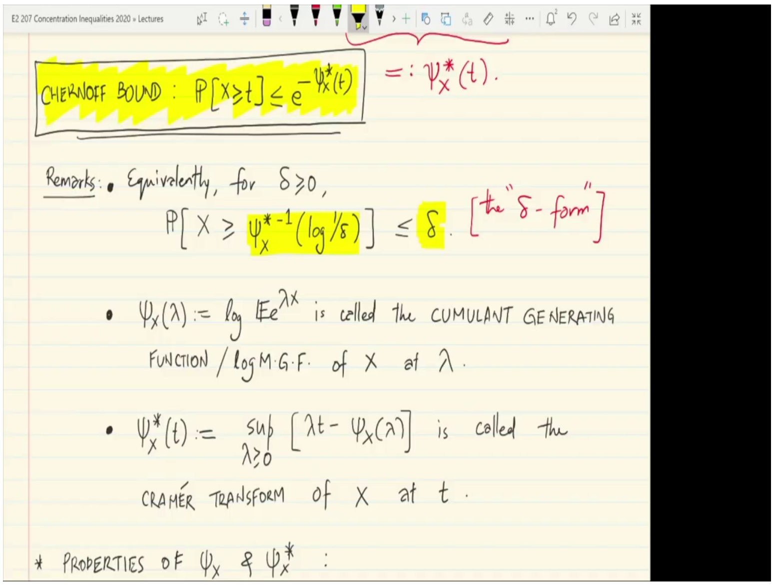


So, / this what I mean is that it also holds that probability X ≥ t is ≤ the infimum over all λ ≥ 0 of e raised to - λ t - Ψ X of λ . And you can easily move the infimum think of infimum if you have not seen this before as a minimum modulo or technicality ok.

So, infimum is really like a minimum. It allows you to take minimums over sets where the minimum may not be exactly attained. So, it follows that this is exactly the same as the x. And when you move the infimum inside the exponent with a - sign, you basically get a supremum over all λ ≥ 0 of the function λ t - Ψ X of λ ok.

So, you get what looks like an exponentially decaying tail bound for the probability of x exceeding a level t ok which has itself the interesting property that the bound itself is the solution to an optimization problem ok.

(Refer Slide Time: 10:56)



So, what is in the exponent is the solution to an optimization problem which we will denote as Ψ X * as a function of t. Note that λ is no longer an argument of Ψ X * because it has been optimized out ok. So, e raised to - Ψ X * of t is / this method the best possible the tightest possible bound for the probability of x exceeding t. And this is what is called the Chernoff bound ok.

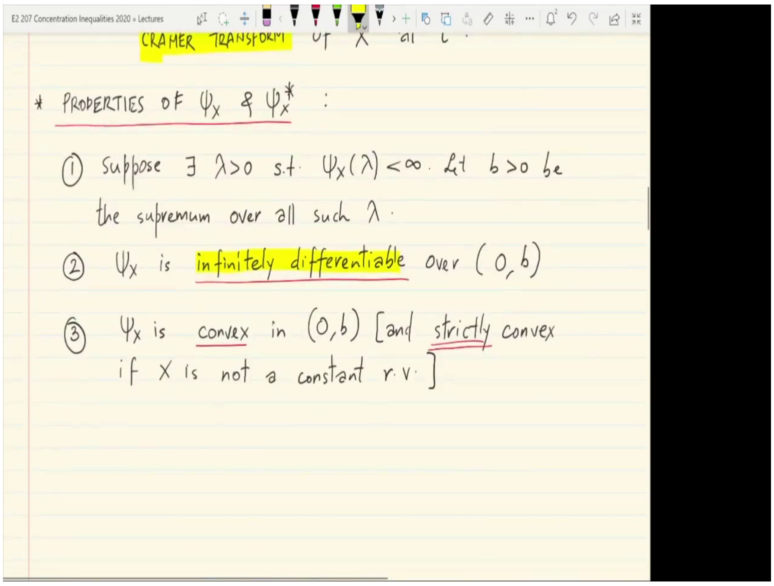
So, the Chernoff bound to summarize is that the probability that X exceeds t is at most e raised to - Ψ X * of t ok. Let me highlight it here, so that it is easy to remember ok. So, the moment you can find depending on the random variable x its associated function Ψ X * as a function of any argument t.

You can immediately enjoy the liberty of putting down bounds upper bounds on the tail probability of X exceeding t. So, a couple of remarks here, equivalent there is an equivalent form for the same Chernoff bound which is obtained / setting the right hand side of the Chernoff bound =a number δ ok.

So, if you set the right hand side =δ and invert this expression, you get that the right t that you have to put is something like this. So, the probability that X is > Ψ X * inverse of log 1 / δ is ≤ δ. So, we will often use this form this equivalent form for concentration inequalities where the tail probability is constrained to a certain δ which is this δ on the right.

And we want to equivalently find the level at which the tail probability of X falls to δ or below ok. And this you can also call this as the δ quantile of X or the one - δ quantile of X. This name is common in statistics and probability.

(Refer Slide Time: 13:36)



The second remark here is that Ψ X of λ which we defined in our derivation of the Chernoff bound to be the log of expected value of e raised to λ X is often called the cumulant generating function ok, it is of course obviously, the log moment generating function of X evaluated at λ .

The third remark is that this function which is derived from Ψ X which is called Ψ X * / solving an optimization problem involving Ψ X is often called the Cramer transform of X ok. So, if X is a random variable and you happen to evaluate Ψ X *, this function is called the Cramer transform of the entire random variable X ok.

So, if you have a random variable X on a probability space or equivalently if you have a measure on a probability space, then you automatically have a Cramer transform ok. And we will see that this Cramer transform function Ψ X * for any random variable X is a very very important object to study. And think about if we want to prove clean or if we want to give clean or usable concentration inequalities for tail probabilities of random variables ok.

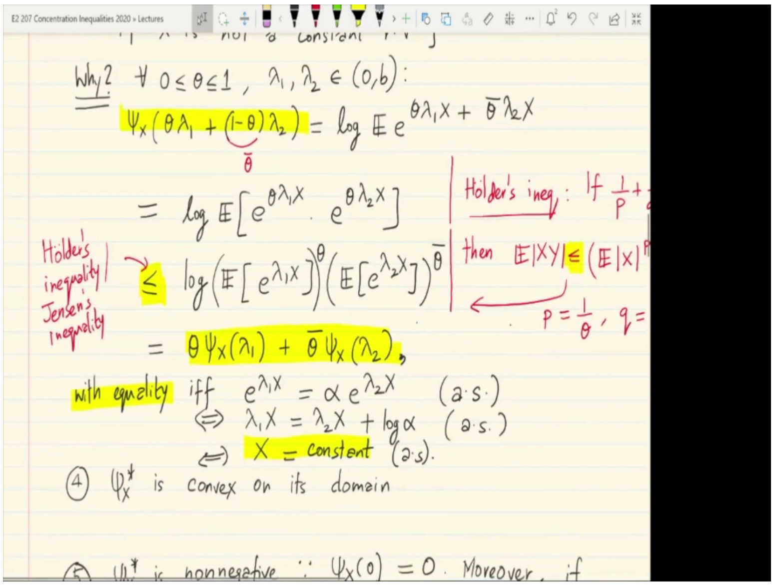
We will move forward to list down some properties of these functions that we have introduced, the cumulant generating function as well as the Cramer transform function of X. So, we will assume we will *t by making the assumption that suppose there exists a positive λ such that the cumulant generating function is finite ok of X ok.

So, that just means that the cumulant generating function of X is not trivially ∞ ∞everywhere depending on the λ is that is not the case. So, there exists at least one λ for which Ψ X of λ is finite. Let b be the largest possible or the supremum of all such λ for which the function Ψ X is finite ok.

We have then that Ψ X is infinitely differentiable over the interval 0 to b. This is not very surprising because if you look at the definition of Ψ X, it is the log of the expected value of e raised to λ X. And wherever this is finite we often know that the exponential is a very well behaved function in its domain of convergence. And it can be differentiated and integrated infinitely often. So, Ψ X of λ just inherits the smoothness properties of the exponential function ok. You can think of it as in that way.

The last property is that Ψ X is actually a convex function in 0 to b ok. So, Ψ X is always a convex function. An in fact, it is strictly convex if X is not a constant random variable. So, if X is anything other than a trivial or an unchanging random variable, then you get a strictly convex function Ψ X ok. So, why is this? Let us quickly show why this convexity property holds.

(Refer Slide Time: 17:03)

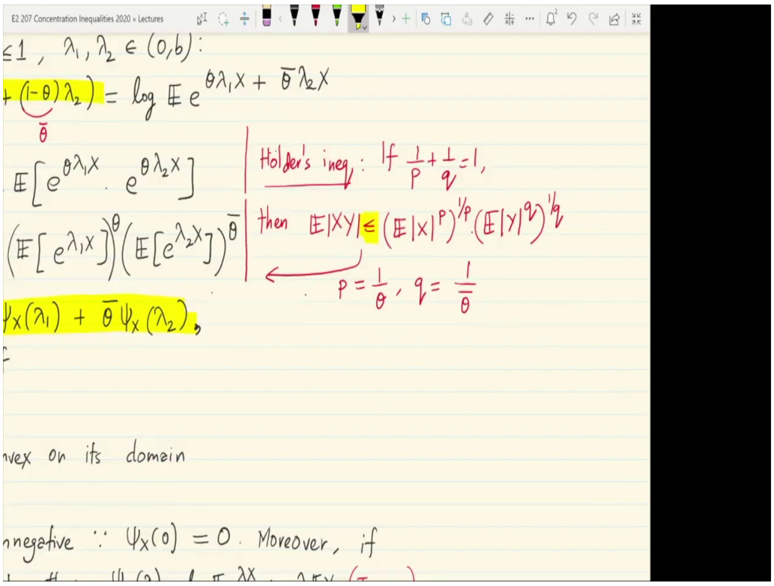


So, if you write down, so let say let us take θ between 0 and 1, let us take any number θ between 0 and 1, and λ 1 and λ 2 in this valid interval 0 to b. Let us write down Ψ X of $\theta \times \lambda$ 1 + 1 - $\theta \times \lambda$ 2. We will often find it convenient to write 1 - θ with the symbol θ bar. So, by definition, this is the log moment generating function of X evaluated at $\theta\ \lambda$ 1 + 1 - $\theta\ \lambda$ 2 ok.

So, one can write this as log of the expected value of a product of e raised to $\theta\ \lambda$ 1 X into e raised to $\theta\ \lambda$ 2 X ok. Now, we will apply a very standard result, a standard inequality which you may have seen before called Holder's inequality to get, so the log is the same outside, Holder's inequality helps us to bound an expected value of a product like this.

And in this case, you get expected value of e raised to λ 1 X ok the whole raised to θ. So, this is the expectation being raised to θ. And on the other hand, you get e raised to expected value of e raised to λ 2 X the whole thing being raised to θ bar which is 1 - θ ok. So, this is by using what is called holder's inequality ok. You can also interpret it as coming from a much simpler inequality called Jensen's inequality which is really the definition of convexity.

(Refer Slide Time: 19:26)



So, just to give you a quick idea, you can look this up separately, but Holder's inequality says the following. So, there are several ways of writing Holder's inequality. If you have two numbers p and q such that, so let say if 1 / p + 1 / q =1. And you have any two random variables X and Y; the expected value of X Y is upper bounded / the expected value of so in fact I can put a modulus signs here to make all of these quantities and random variables non-negative.

So, expected value of X raised to p the entire thing raised to 1 over p, and similarly the expected value of Y raised to q, 1 / q ok. So, apply this here with p being =1 over with p being =1 over θ, and q being =1 over θ bar ok, so that the conditions of the inequality are satisfied. And you will see that this is Holder's inequality.

/ the way if p and q are =1/2 in Holder's inequality you get what is called the very well known Cauchy-Schwarz inequality. So, Holder's inequality is just a generalization of Cauchy-Schwarz which you may be much more familiar with. And so the moment you have Holder's inequality here, this just simplifies to θ Ψ X of λ 1 + θ bar Ψ X of λ 2 ok.

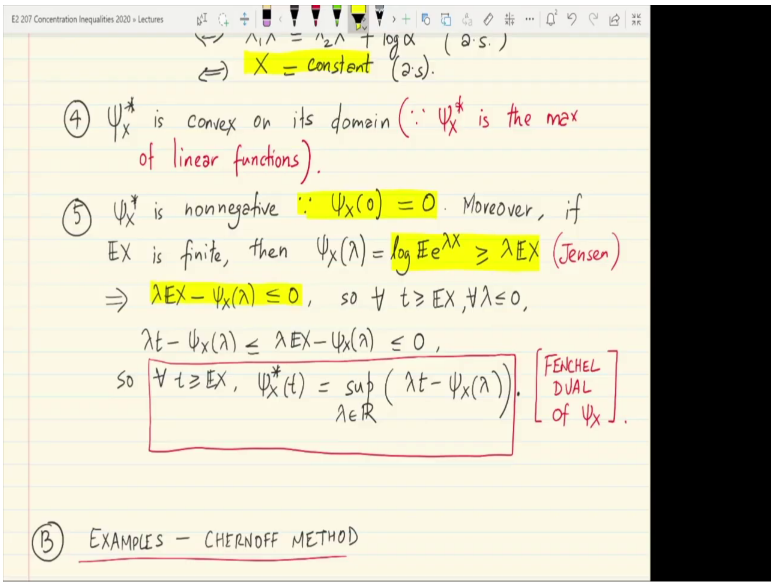
And so this is why Ψ X is convex ok. This is exactly this left hand side here less than being less than this right hand side here is Y is the definition of Ψ X being convex ok. Now, we want to argue that there is strict convexity which means that the inequality here is always strict ok in if X is non-trivial. So, this is also easy to show.

So, with equality if and only if so it turns out that in Holder's inequality which I have written down here the inequality is an equality. So, this inequality is an equality if and only if X and Y are scalar multiples of each other ok. So, the random variables X and Y are exactly related / scalar multiplication.

So, if you translate that requirement here, it means that the random variable e raised to $\lambda_1 X$ has to be a scalar multiple of e raised to $\lambda_2 X$ ok with probability 1 almost surely ok. And this is if and only if you take logs on both sides; $\lambda_1 X$ is $= \lambda_2 X$ + some log alpha which is another constant almost. Surely and this is equivalent to X being a constant ok.

So, the only way that only way that equality holds is if X =constant. So, if X is not a constant random variable, then you will always have strict inequality ok inherited from Holder's inequality. The so let me make some space here.

(Refer Slide Time: 23:38)



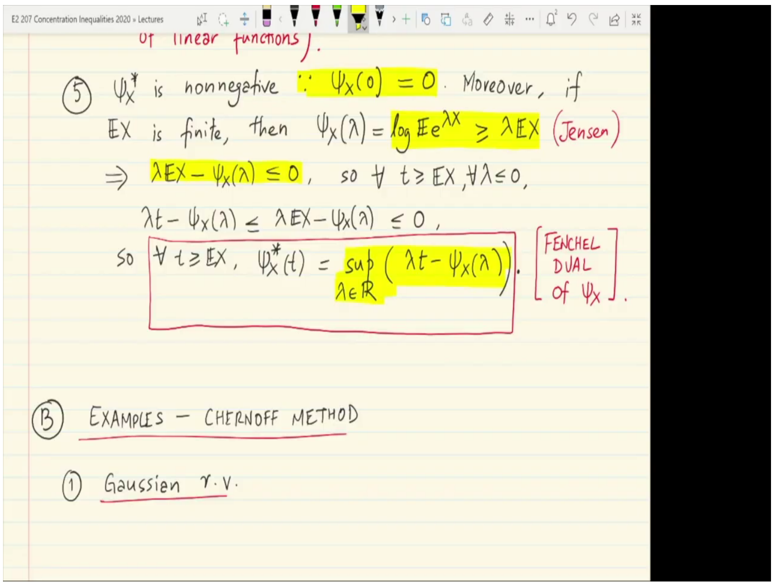
The next point here is that Ψ_X is always Ψ_X^* is always convex on its domain ok, domain is wherever it is it is finite. If it is infinite, we do not bother with talking about convexity ok. Why is this, why is Ψ_X^* convex? The reason is actually very simple. So, if you look at the definition of Ψ_X^* here ok right here, it basically means so how is Ψ_X^* of t computed? You basically take several possible λ s, and you compute this quantity and then you take the maximum of all these ok.

So, for each fixed λ, the function here as a function of t is actually a just a linear function ok, so the dependence on t is just linear. So, you can think of Ψ X * as basically taking the upper envelope or the maximum of a bunch of linear functions each linear function being indexed / λ ≥ 0.

So, Ψ X * is just the; is just the max of linear functions ok, Ψ X * is the max or the supremum technically of linear functions ok. And it is very easy to show that if you have two or more linear functions, if you take the max of them and form the new form a new function that is always going to be convex ok. You can just convince yourself about this.

The next property about Ψ X * which is the Cramer transform is that it is always non-negative. It cannot be negative at all. This is because; this is because Ψ X the log moment generating function is 0 at 0 ok. So, if you go to the definition of Ψ X * again, if you plug in λ =0 here. Then one feasible value for this optimization problem is =0 - Ψ X of 0 which is 0. So, Ψ X * can only be > 0. It can never be smaller than 0 ok.

(Refer Slide Time: 26:01)



So, Ψ X * is a non-negative function. Moreover if the expected value of X exists if X has finite mean, then Ψ X of λ the log moment generating function of λ / Jensen's inequality for the convex function e raised to λ X can easily be shown to be lower bounded / λ $\times$expected value of X. This is just by Jensen's inequality applied to the convex function taking X to e raised to λ X.

And so this means that λ e raised to X - Ψ X λ is always upper bounded / 0. And so for every t that is $\geq$expected value of X and for every λ 0 ok. So, recall that in the definition of Ψ X *, we only took $\lambda \geq 0$ ok.

But what happens if you take a $\lambda \leq 0$? So, if you take t > the mean and λ negative, what happens is that λ t - Ψ X of λ turns out to be upper bounded / λ expected value of X - Ψ X of λ, and we know that this is $\leq$ 0 by Jensen.

So, the upshot of this is that for any t > the mean the Cramer transform of the random variable at t might can might can as well be written as the supremum of the same objective function over all λ rather than only $\lambda \geq 0$ ok.

So, this is nice to remember. You do not need to take you do not need to restrict λ to be ≥ 0 when you are evaluating the Cramer transform for t > E of X to the right of its mean, but you can actually just consider the computation of Ψ X * of t to be an unconstrained optimization problem ok. So, this allows you to do much more cleaner calculations and solving of the optimization problem ok rather than operating with a constrained optimization problem.

So, by the way this function so if you give me any function Ψ X, if I define the following function which is the sup over λ in R λ t - Ψ X of λ, this function Ψ X * of t has a special meaning and a name in convex analysis. It is called what is it is what is called the Fenchel dual of Ψ X or the Fenchel Legendre dual of Ψ X.

So, Ψ X and Ψ X * have enjoy some very nice duality properties ok. So, there is a whole rich theory of duality built on these foundations of functions and their Fenchel Legendre duals ok. So, the Cramer transform is essentially the Fenchel dual function of Ψ X is what we have shown here. So, we will stop here. And then we will continue with applying this Chernoff bound technique to various random variables that we are familiar with.