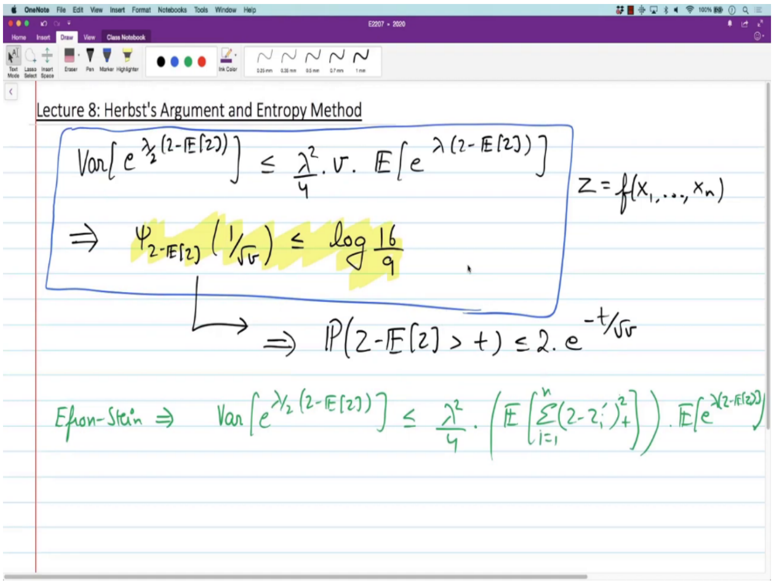
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 09
Herbst's argument and the entropy method

(Refer Slide Time: 00:21)



In the previous lecture, we saw the following bound for establishing concentration bound. We saw that if you can establish that variance of $e^{\lambda / 2}$ Z - E Z is $\leq \lambda^{2} / 4$ times some constant times expected value of e^{λ}. If we can establish this implied that the log moment generating function of Z - E Z is evaluated at 1 / root v is less than a constant log 9 something we call 16 / 9 ok, that is something we saw.

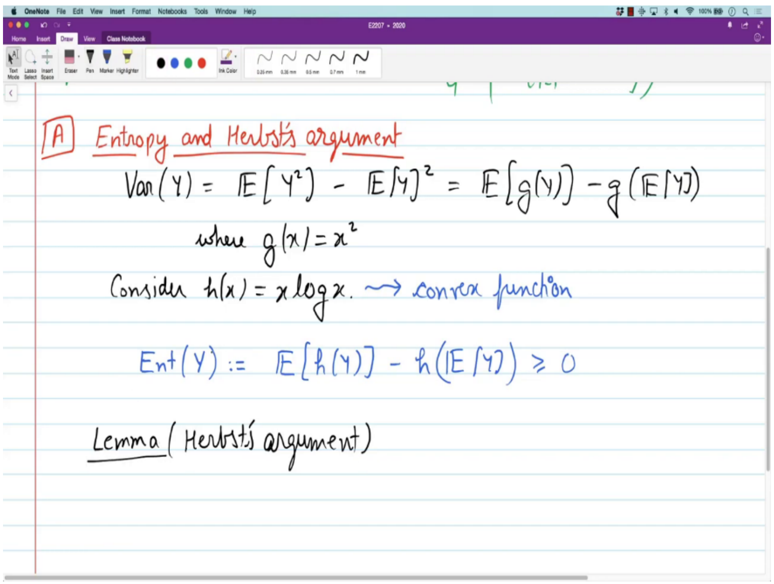
And, using this bound this is sort of a functional inequality and using this bound we can obviously, establish a concentration bound. So, this part here yeah, this bound here implies that the random variable Z- E Z $\geq$ t is 16 / 9. I just upper bounded / 2 e to the power - t / root v that is something we saw in last time.

And, what we notice was that this particular part this first inequality the functional inequality that we need this can be established / Efron-Stein inequality. So, we have Efron-Stein implies variance of e^{λ} / 2 Z - Z is less than = λ^2 / 4 times what we were looking for constant v summation i = 1 to n this is my constant v and times the same expected value of e^{λ} Z - E Z ok.

So, that is what we show that is what we saw last time and / the way here this throughout this random variable sorry, the random variable Z that we are looking for is this guy. It is a function of independent random variables X 1 to X n, ok. So, this particular part of the argument this is through Efron-Stein inequality and this uses the tensorization property which the Efron-Stein inequality gives ok.

Today, what we will do is we will basically derive another inequality like this and what this inequality will do is instead of variance it will have some other quantity here which looks very much like variance and that functional inequality will also imply another inequality like this some bound for log moment generating function ok. And this will lead to a very different approach for proving concentration bound the so called entropy method. So, let us introduce that.

(Refer Slide Time: 03:59)



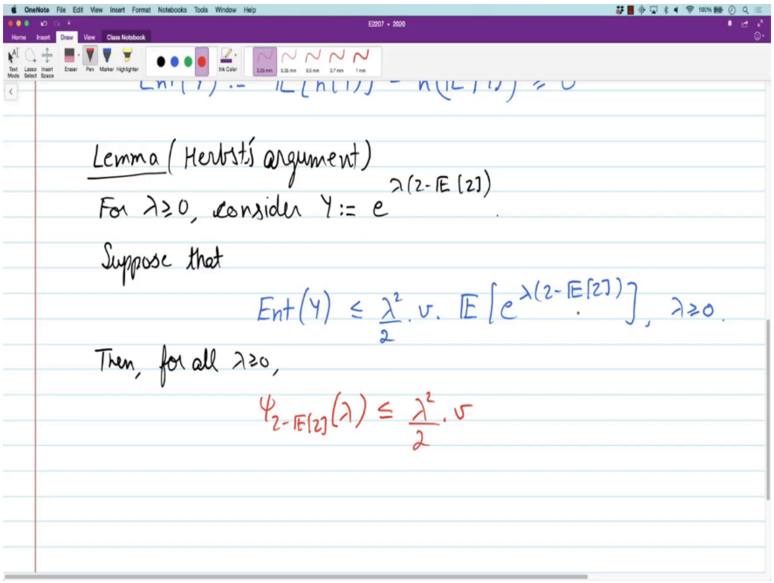
So, let us introduce entropy and Herbst's argument which is the replacement of the functional inequality that we saw last time. So, remember that variance of Y is expected value of Y^2 -

expected value of Y whole square. So, you can write it as expected value of g of Y - g of expected value of Y, where g is the convex function g of x = x ^{2}.

A similar quantity can be defined using another function consider this function h of x equals to x log x. Then the quantity we are after the entropy of a random variable x or let us say Y for us is defined as expected value of h of Y - h of expected value of Y. Note that this function h this guy here is it is a convex function ok and therefore, this guy here is non-negative ok, alright.

So, let us try to express this guy here. So, this is the entropy and what we will show is something similar to what we had earlier. So, let me write this lemma and call this Herbst's argument.

(Refer Slide Time: 06:37)



It says consider let us fix first for λ ≥≥ = 0, consider Y the random variable Y defined as e $^{\lambda}$ Z - expected value of Z. So, we can consider this random variable then and suppose that the entropy of this random variable Y is less than = λ^2 / 2 into some constant v into expected value of e to the power λ Z - expected value of Z. Suppose this is true.

So, note that this condition is very similar to what you had here that variance was less than this. We were looking at variance of that the way the random variable Y has a factor half here.

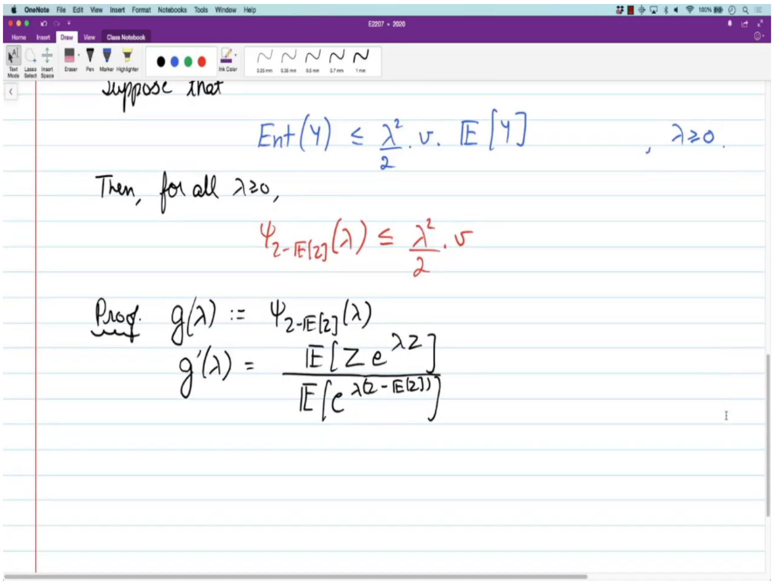
So, you could actually you can actually think of this function g as x and √ x if you like that would also look similar ok yeah. So, to this mod so, if you do not; if you do not worry about this factor of half this variance looks very much like entropy and now, the condition we are asking very much looks like this condition here except that λ / 2 has been replaced with λ .

So, suppose this bound holds ok suppose that this bound holds then very interestingly for every $\lambda \geq 0$ this holds for all $\lambda \geq 0$, then what follows is that for every $\lambda \geq 0$ you have the same bound you have the sub Gaussian bound $\lambda^2 / 2$ v ok. So, this is Herbst's argument essentially the generalization of the claim that we have seen earlier.

But, what is very interesting is that it has given us a way of establishing if this claim is true which show now that indeed it is true then this will give us a way of deriving a sub a sub Gaussianity bound, ok. So, the random variable Z will be sub Gaussian if you can show that it is if the entropy of e to the power λ Z - expected value of Z the central version of Z, if it is entropy is $\leq$ $\lambda^2 / 2$ v times the log moment generating function ok.

The expected value of the same random variable entropy of this random variable is less than = expected value of the same random variable, ok maybe a better way to write it as follows yeah ok.

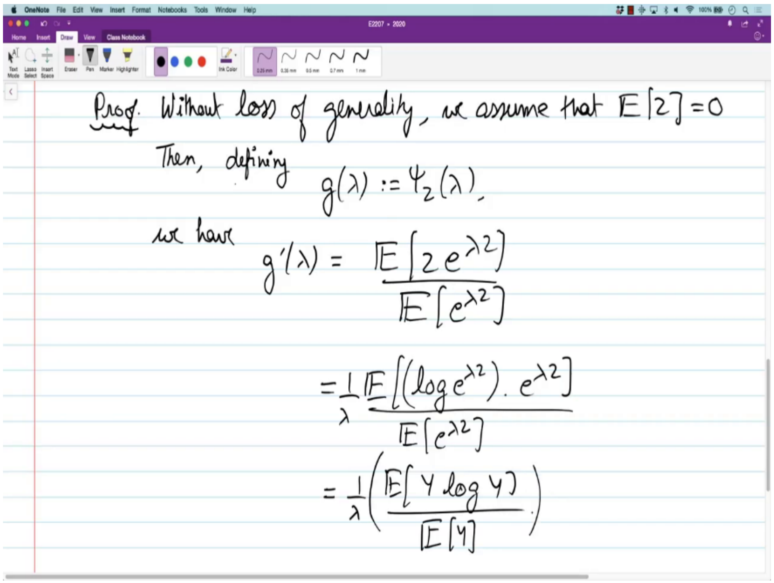
(Refer Slide Time: 10:09)



So, how do we show this? Proof is again we have to show a functional inequality that this one implies this one, it is all. So, we make the following observation. If you look at let us look at this function g λ, then if you look at g prime λ what do we know about it? Ok.

So, this is log of is something we have seen before. So, this is derivative of the log moment generating function it is 1 / the moment generating function and you can essentially take the expectation inside the derivative inside the expectation for the smooth function. So, you get this e^{λ} Z, ok.

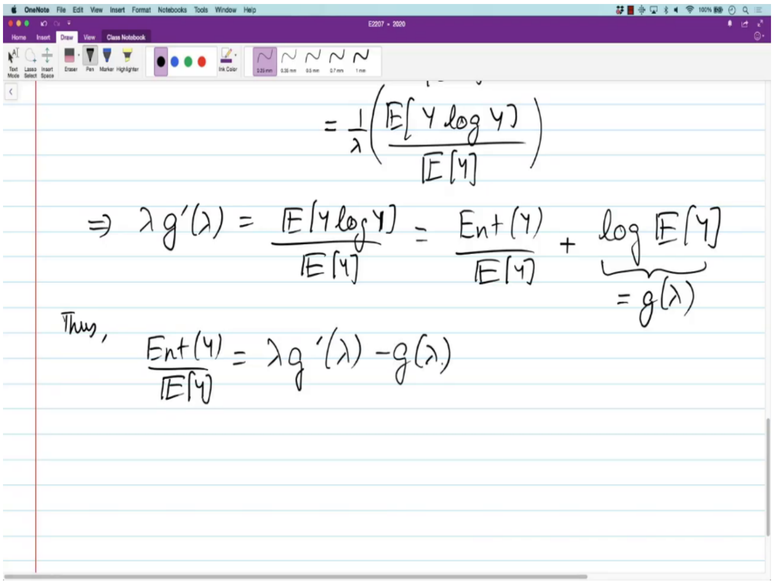
(Refer Slide Time: 11:35)



And, this guy here if you look at this guy here and everything need actually just to just to make my job easy so that I forget about this part. What I first notice is without loss of generality we assume that expected value of Z = 0. Note that since we are centering this random variable anyway we can just assume this then defining g λ as the log moment generating function of Z of Z we have what do we have about this derivative it is the expected value of Z e^{λ} Z / expected value of e^{λ} λ Z ok.

So, what is this guy? So, note that this guy here is expected value of log of Y divide / λ because π was e^{λ} Z. So, maybe I just write it this way ok this guy here is 1 / λ ok. So, we get this inequality.

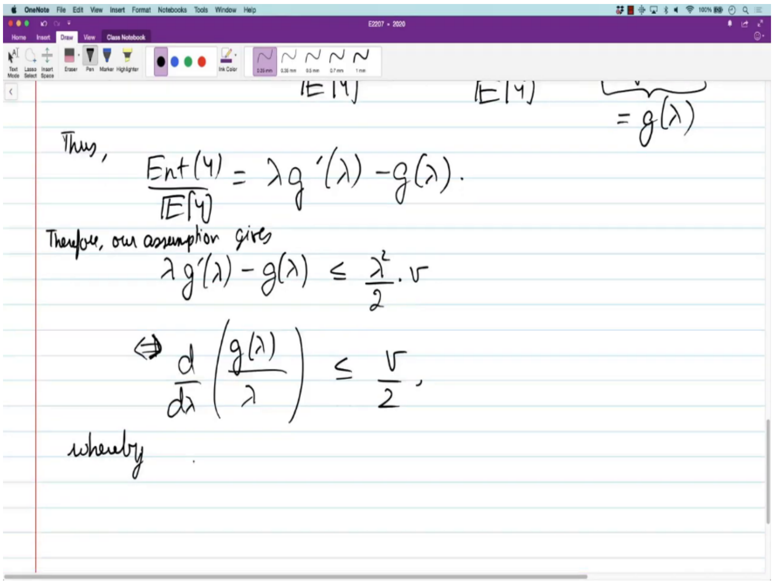
(Refer Slide Time: 13:57)



So, this implies λ times g prime λ is expected value of Y log Y / expected value of Y which is = the entropy of Y / expected value of Y + log of expected value of Y ok. And, and this guy if you if you this is how I defined things this is exactly = g of λ, ok. So, rearranging term what have we obtained? We obtained that this entropy of Y / expected value of Y is exactly = λ g prime λ - g λ .

But, now let us look at our assumption. Our assumption for this random variable just says that this ratio is $\leq \lambda^2 / 2$ v that is the exactly the assumption we are making.

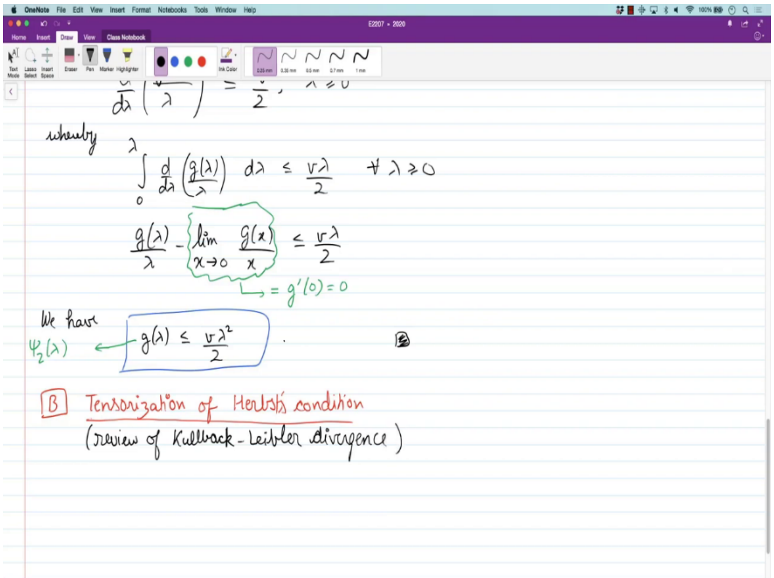
(Refer Slide Time: 15:05)



So, this implies that λ g prime λ - g λ is l$\leq\lambda^2$ / 2 v, that is ok. So, we have this and therefore, our assumption gives this part. So, we divide / λ^2 and we rearrange term and this can be seen to in the following.

Take the function g λ / λ take it is derivative with respect to λ the derivative is g prime λ / λ - g λ / λ^2 ok. So, that is this derivative this derivative is $\leq$v / 2. This statement these 2 statements are equivalent.

(Refer Slide Time: 16:17)



So, what do we get finally, now this is true; this is true for all $\lambda \geq 0$. So, you can integrate from 0 to λ d / d λ of g λ / λ is $\leq$v / 2 v λ / 2 and this is true for all $\lambda \geq 0$. But, this guy here, if you see this guy it is just g λ / λ - g of - limit x going to 0 g of 0 / g of x / x ok and this is $\leq$ v λ / 2. That is the bound we get.

So, the only thing that remains to verify is this guy here. Now, this guy here because of properties of the log moment generating function this guy here is exactly = g prime of 0 ok the derivative of g at 0 which for the 0 mean random variable is 0, this is 0. So, we have obtained g (λ) /λ is $\leq$ v λ / 2. So, v λ^2 / 2 ok. That is what we have to show / the way because this g λ remember was just our abbreviation for phi Z λ and we had just entered it ok.

So, we have the sub Gaussian form ok. So, this is the very very elegant method for establishing sub Gaussianity bound and it gives you it says that if you can establish this functional inequality for this random variable then.

So, this is just the log moment generating function, this is just the moment generating function. It show that the entropy of the moment generating function is entropy of this let us call it the moment random variable; entropy of the moment random variable is $\leq \lambda^2$ v / 2 the moment generating function.

If you show that then you have a sub Gaussianity bound. So, we note that this is exactly I insist this part again this is exactly what we did in the last class, except that instead of variance now we have now we have a different quantity which looks very much like variance instead of function you are looking at this entropy function and you use that to define some counterpart of variance.

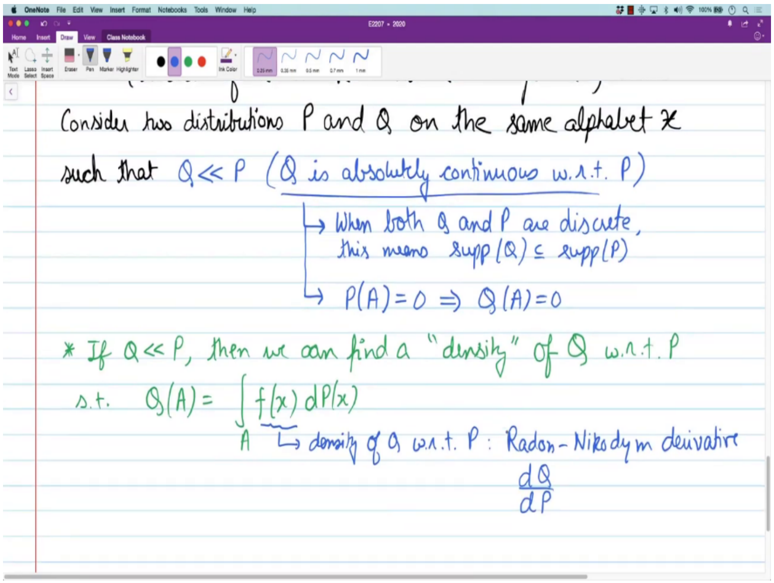
So, what Herbst's argument tells us is that if we can handle this kind of inequality for entropy of random variables then we can establish sub Gaussianity bounds and. So, this part is fine this is the first step remember the in our. So, this gives a new way of handling log moment generating function, but this is only the first step in our general recipe for deriving concentration bound.

The second step was a tensorization bound and in particular we use the Efron-Stein inequality to establish the corresponding tensorization bound for this for this functional inequality and now, what we need is some argument that can tensorize this condition here that entropy of Y is $\leq \lambda^2 / 2$ times v expected value of y.

And, in fact, that tensorization argument will follow from some information theoretic properties because this function is very much like divert is reminiscent of quantities like KL divergence that we observe that we see in the information theory course I am assuming some of you have taken that course before.

So, towards that tensorization argument, so, this is the step B. So, now, tensorization of Herbst's condition ok and to do this we need to do a quick review of the KL divergence the Kullback the notion of Kullback-Leibler divergence from information theory ok which we will do now. The Kullback-Leibler divergence is actually a measure of sort of a; sort of a measure of distances between 2 distributions.

(Refer Slide Time: 21:27)



So, consider 2 distributions P and Q on the same alphabet and such that Q is absolutely continuous with respect to P that is the notation for it. Q is absolutely continuous with respect to P. This is sort of a heavy language, but we just borrowing it so that we can keep a bounce and so, what does this mean? Q is absolutely continuous with respect to P.

So, it can mean different things, it is a very general conditioned. So, when both Q and P are discrete this means that if you look at the support of Q the set of point where Q puts mass that is contained in the support of P. So, if P of x equals to 0 for any point Q of x must be = 0. This is for the discrete case.

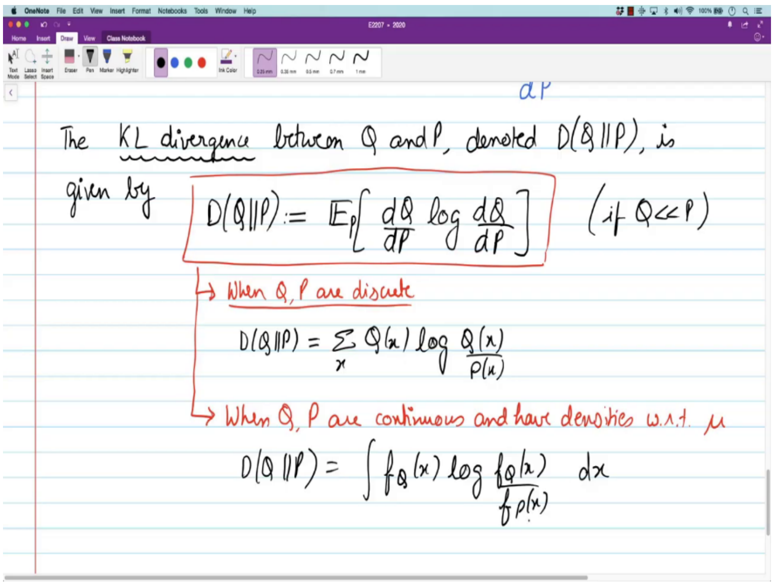
In general, it means if P assigns probability 0 to a set and that implies that Q also assigns the probability to a set. This is in general this is in general the definition of this, but why does this thing enter. The reason this thing enters is for the following result if P is absolutely if Q is this is something quick review from something from probability theory.

Then we can find a density of Q with respect to P such that, so, if you want to measure any set under Q with probability inside under Q you can integrate over this density. So, you can take measure P except that you have to integrate over this density. So, this is that density of Q with respect to P and this density has a name it is called the Radon-Nikodym derivative of Q with respect to P. So, writ10 as dQ / dP ok.

So, for discrete case this is just the ratio of the 2 pmfs and for continuous case with density if both P and Q density with respect to let us say the lebesgue measure then this is the ratio of those densities and in general also this guy exists ok. So, we can define this Kullback-Leibler divergence for any 2 distributions Q and P such that Q is absolutely context with respect to P.

If you are very confused about all these abstraction and not very comfortable with it you can just think of discrete distributions and assume that support of Q is contained in support of P. That is what this condition is, ok.

(Refer Slide Time: 25:31)



So, we have such 2 distribution where Q is absolutely continuous with respect to P then we can define their Kullback-Leibler divergence the Kullback KL divergence. This is KL is short for Kullback-Leibler between Q and P, denoted dQ P it is asymmetric in Q and P. So, we start with Q and this base measure P is kept here ok.

P is the bigger measure here and Q has a density with respect to P is given / dQ P is = expected value of dQ / dP. This is some random variable the density random variable log of dQ / dP. That is the definition ok that is the definition of Kullback-Leibler divergence. So, when actually let me elaborate this.

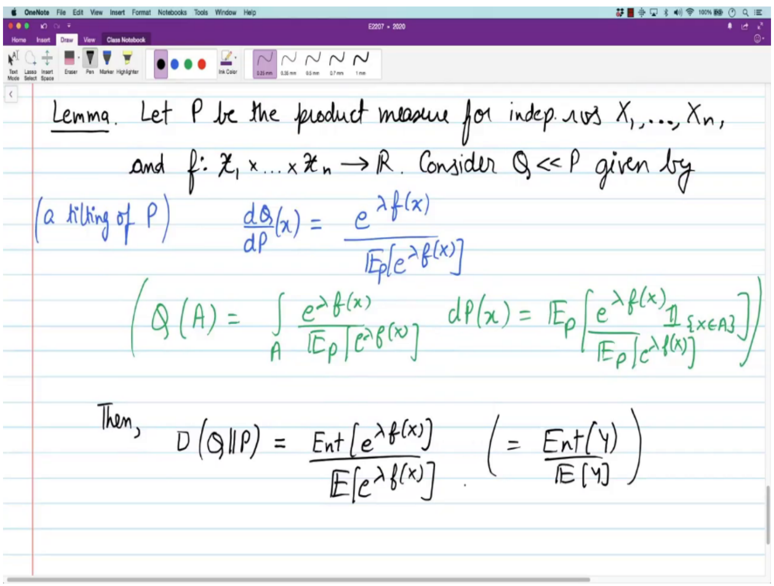
This is true if; this is true if Q is absolutely contrast with respect to P, otherwise it is just infinite this Kullback-Leibler divergence ok, but let us just put it for this case. Now, special cases when Q and P are discrete. In this case the simple formula we have is D Q P some of you may have already seen it. Summation over x Q x log Q x / P x.

/ the way, if you what there are 2 measures floating around here Q and P. What do I mean / expectation here? When I do not put any when I do not clarify any distribution it is always with respect to the larger reference measure P ok. So, this is with respect to P and for discrete case this dQ / dP is just Q x / P x, but then you take expectation with respect to P. So, you multiply with P x that goes away and you get this.

Another interesting case is when Q P are continuous and have densities with respect to a Lebesgue measure ok mu that is the Lebesgue measure then just like Gaussian or some other 2 2 distribution then this guy is = the density of Q log density of Q / density of P ok dx and this integration here is with respect to the Lebesgue measure ok the standard integration. That is the other example where this has a very concrete form ok.

So, this Kullback-Leibler divergence why did we introduce this Kullback-Leibler divergence it is well defined here it turns out that we can view our entropy function or in fact, the ratio of that entropy to expected value of Y that the thing that appeared it Herbst's condition, this thing here the ratio of entropy of Y / this ratio entropy of Y / this. This can be viewed as a divergence a Kullback-Leibler divergence and that is what we will see.

(Refer Slide Time: 29:29)



So, lemma let P be the product measure for independent random variables or underlying independent random variables X 1 to X. So, P is this base like measure and f be this function f from x 1 to x n to R. So, we consider this measure consider Q given /. So, to define this function Q this measure Q, what we will do is we will just define the density of Q with respect to P we can always do that and I will just use this notation x for the vector x. So, this is the function of x. What is the density for this x?

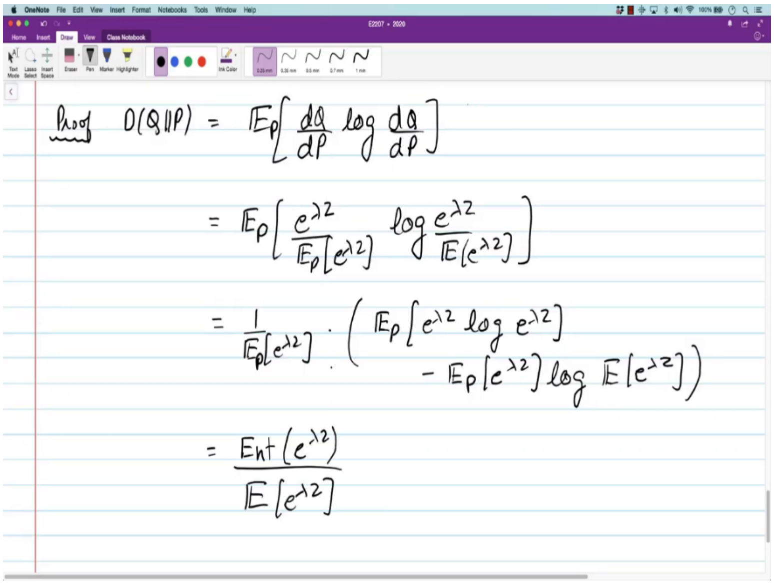
Suppose this density is we use a different color dQ / dP of x is e to the power λ f of of x, the vector x this whole vector x this is with respect to P and yeah. So, suppose this is my this is my measure. So, what does this definition mean? It means that if you want to compute the probability of any set A that is given / integral of this guy e to the power λ f x expected value of under P of E to the power λ f X, but you have to tell me which measure you use to integrate and that measure is P ok.

So, it is the expected ok. So, this is the same as just to be very concrete expected value with respect to P indicate a function of X belonging to A / expected value with respect to P e λ . So, you can define some measure Q a this way and that is this measure. So, this guy has a name it is the it is a tilting of P ok. So, you it is like you take this measure P and you tilt it with using this function and it is an exponential tilting of P because you have this e to the power λ f x here ok.

So, suppose you have this measure then if you look at the divergence between Q and P that divergence is = entropy of e to the power λ f x / expected value of e to the power λ f x. So, namely what we had earlier the entropy of Y / expected value of Y ok. So, this was the quantity that we were bounding in Herbst's argument this guy here. We were saying that this guy is $\leq\lambda^2$ time some constant / 2 and that constant turns out to be the sub Gaussianity parameter ok.

So, this actually should not call this tensorization. I will not be able to prove tensorization in this lecture alone, ok. Let us see how far we go we want to prove tensorization, but I may stop maybe let us see, ok. So, what is the proof here?

(Refer Slide Time: 34:17)

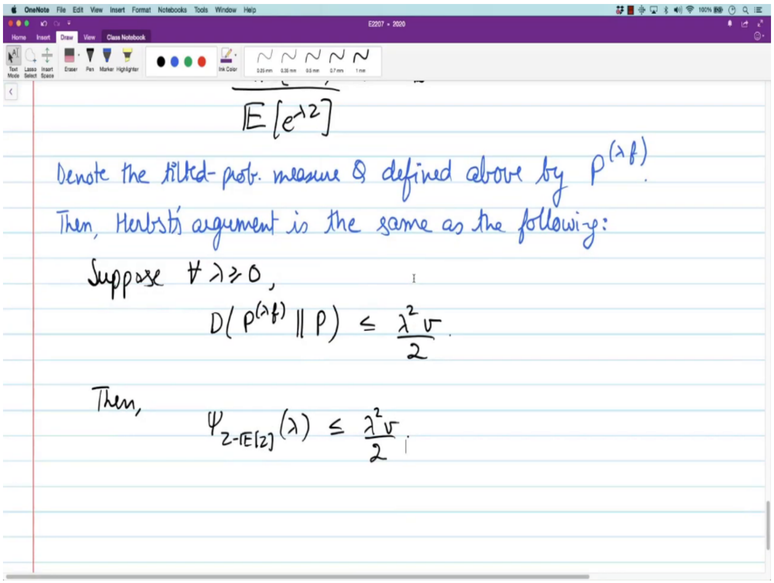


Proof D Q P is the expected value under P of e to the power the this is just the definition dQ / dP, since we have directly define density you can verify that Q is absolutely continuous with respect to P. If P of some set is 0, then this guy is over the integration over is over a set of measure 0. So, this must be 0 ok.

So, then therefore, it make sense this is = expected value under P. This is just / definition e^{λ} say that is f of x expected value e^{λ} Z log of e^{λ} Z expected value e^{λ} Z. And this is = you can take this expected value outside into the first term is and then since log a / b is log a - log b.

So, second term here is this part log of so, we this is exactly what we had entropy of e^{λ} Z / expected value e^{λ} Z, alright. So, the proof is straight forward it is just the expression. So, very nice. So, this Herbst's condition that we have seen earlier can be re expressed in terms of divergence then.

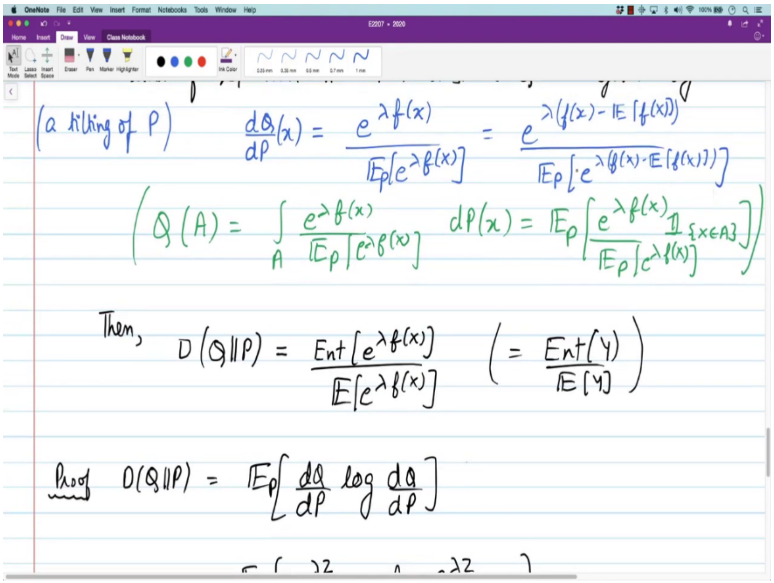
(Refer Slide Time: 36:31)



Denote the tilted probability measure Q defined above / so, it is P and we tilt P / this is the well defined thing. So, you take an take any function here and this tilting is defined as just taking this density e^{λ} f x / expected value of that, ok. So, if you denote it this way then Herbst's argument is the same as the following. Suppose for all $\lambda \geq 0$, the divergence between λ f and P is $\leq \lambda^2$ v / 2.

Then, the log moment generating function one small catch here I showed this proof with e^{λ} f, but what happens to the so we have to subtract from it - e^{λ} f x.

(Refer Slide Time: 38:31)

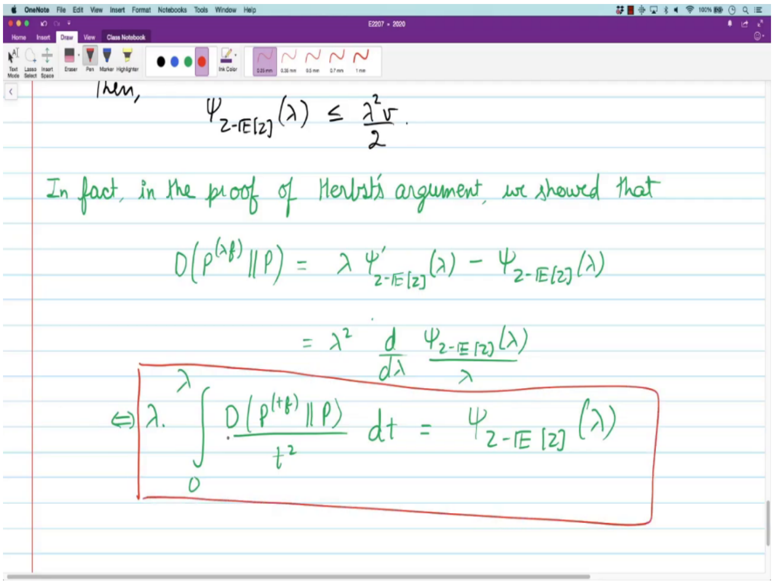


But, note that it does not matter this is exactly = e $^{\lambda}$ f x - expected value of f X that centering comes for free because there is a ratio here. So, if you were worried about the centering it is the same quantity because this cancels from the numerator and denominator just the constant which shows up in both of them ok.

So, then therefore, Herbst's argument is the same as this bound this is the assumption and under this assumption what you get is that the log moment generating function of the centered random variable is $\leq\lambda^2$ v / 2 ok that is the; that is the very interesting thing here ok.

In fact, if we go to the proof of Herbst's argument, so, we observe this guy here ok we observe that entropy, but this entropy is exactly = this guy and so, we knew we know that this is the divergence and divergence is exactly = this guy which is this derivative here.

(Refer Slide Time: 40:11)



So, this can be writen as let me summarize that bound here. In fact, in the proof of Herbst's argument we showed that this ratio of entropy that is the divergence. This divergence is exactly = d / dλ . So, which implies we can integrate over this divergence here. One second, sorry about that.

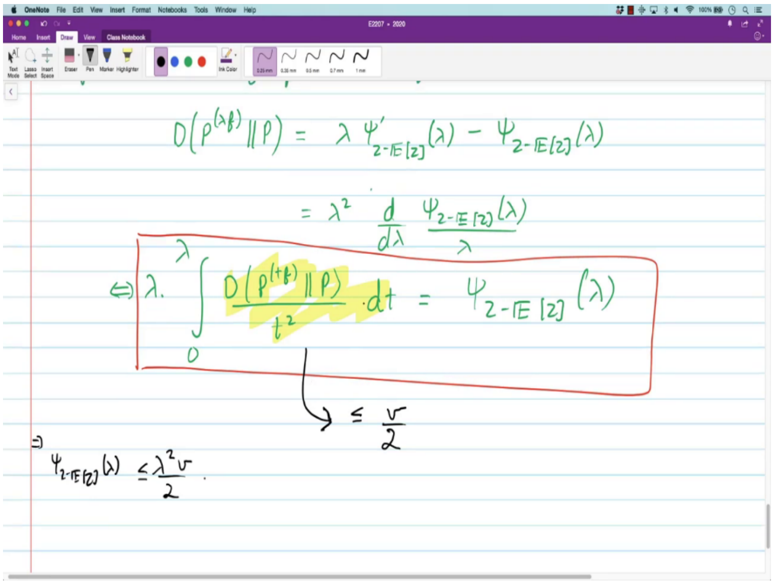
So, actually so, what we saw was this is = λ *times* the derivative of the log moment generating function - the log moment generating function. This is what we were calling g there in that proof. So, you can take out λ^2 and what we can observe is that this is just this is something we observed in that proof, the derivative of this ratio λ / λ , right.

So, you can see that this is the same as saying that if you have this tilting along f take this divergence from P divide / t^2 and take this integral from 0 to λ this is exactly = this / λ . So, I can take this λ here that is a very nice formula for the log moment generating function. So, it says that the log moment generating function is I should I should switch things around, but that is fine.

So, it says that log moment generating function is this integral from 0 to λ of the tilted divergence from P / t^2dt ok tilted divergence / t^2 dt. So, you take P you change it a little bit to get t to get P tf this is that divide the distance between the 2 and when you divide / t^2 you can think of it as sort of the second order term in the Taylor series approximation of this distance.

So, expand this distance and look at the second order term that second order term if you integrate from 0 to λ and multiply with λ is exactly = the log moment generating function. So, this is the gist of this is the gist of Herbst's argument.

(Refer Slide Time: 44:07)



So, Herbst's argument assumes that this guy here this guy here this term here is so ≤ v / 2 and therefore, this integral just becomes λ^2 v / 2 ok. If this integral is ≤ λ^2 v / 2. So, log moment generating function is ≤ λ^2 v / 2 that serves the argument.

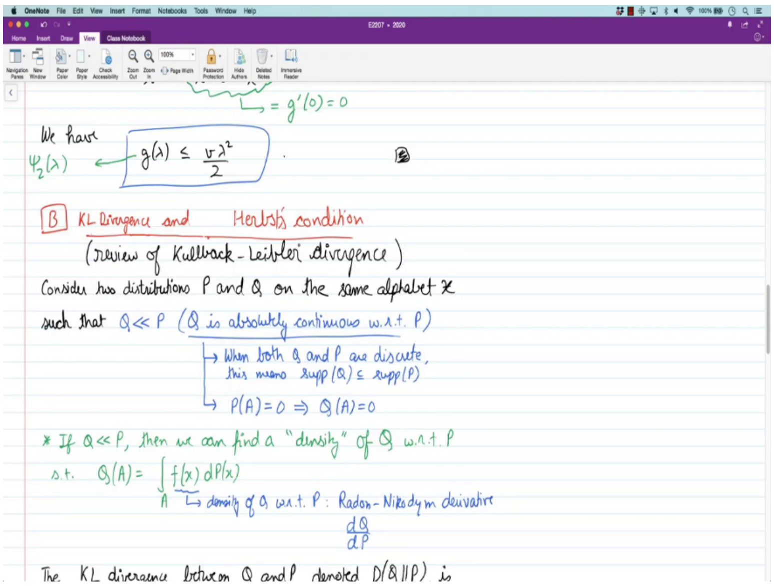
So, to conclude, we have to establish this bound and where this is very nice formula and so, why were we doing this alternative expression? Now, the point we want to make is that this particular divergence here, it can be expressed this divergence here which is the quantity we would like to control even for an enfold random variables we can somehow tensorize this guy ok because Herbst's argument gives us the counterpart.

Remember this lecture we started / saying that Herbst's argument gives us the counterpart for this inequality. But, now when we prove this inequality we have to establish this condition for which we use Efron-Stein's inequality. So, what is the counterpart of Efron-Stein's inequality and / observing that / observing that this divergence here, it looks like thethis divergence sorry, this ratio here looks like a divergence.

We are essentially asking some kind of sub additivity of divergence like Efron-Stein inequality and that is the kind of inequality which we has which we regularly seen information theory, they are called chain rules. And, we in this case need a specific chain rule for divergence sort of a chain rule it is a sub additivity part for divergence which is what we will show in the next lecture.

So, for now I did not actually complete the proof of tensorization. I think this lecture is becoming longer than what I wanted to be.

(Refer Slide Time: 46:07)



But, so, I will change the title. It is not about tensorization, it is KL divergence and Herbst's argument. So, we saw the saw the connection between Herbst's argument and KL divergence.

In the next class, we will use this connection to establish a tensorization for Herbst's condition, ok and just to summarize finally, this Herbst's condition is saying just this that the divergence if the divergence of the tilted guy is $\leq \lambda^2 v / 2$ for all λ in the log moment generating function is also $\leq \lambda^2 v / 2$ ok. Next time we will show that this condition tensorizes.

See you in the next class.