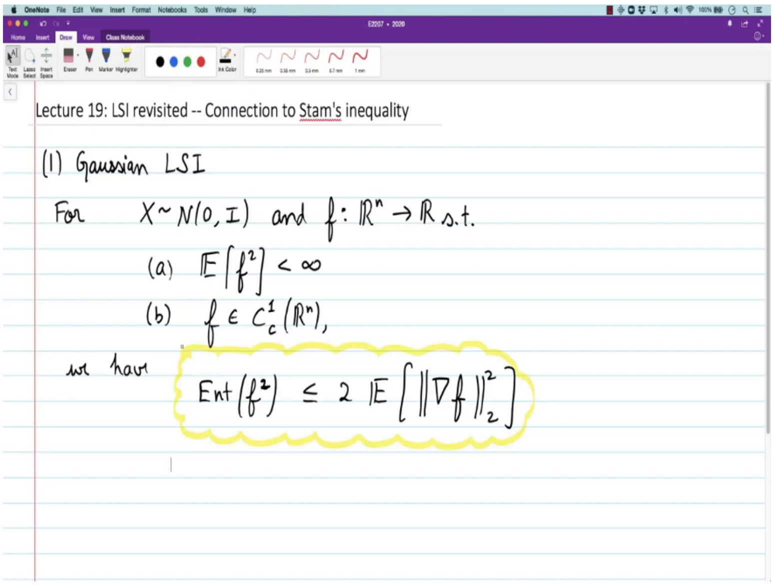
Concentration Inequalities

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

Lecture - 19
LSI revisited- Connection to Stam's inequality

(Refer Slide Time: 00:21)



Hi, so far in the course we have seen various method for showing concentration bounds, we have seen the basic churn of method where tensorization plays an important role and we saw various ways in which tensorization can be shown starting with the bounded difference property and how it leads to tensorization. And we moved on to some more abstract methods which are in some sense more powerful where we showed the entropy method and the transportation method.

In the entropy method we use this argument called Herbst argument which gives an alternative formula for log moment generating function in terms of divergence and in this method the main tool the main technical tool that is used to show concentration bounds is this so called log Sobolev inequality.

We established three different log Sobolev of inequality, we established one for the case of binary random for case of random variable distributed uniformly over + - 1 valued vectors of length n or d, then we saw the Gaussian log Sobolev inequality. And then we see a more general modified log Sobolev inequality which allowed us to derive concentration bounds for a for an arbitrary random variable. For the next 2 to 3 weeks I will be focusing on this log Sobolev inequality.

And what I will show is I will show how this log Sobolev inequality is connected to other inequalities that are that have a longer history perhaps and they are not directly similar if you look at it you will not see the similarity, but you will see very interestingly that many of these inequalities are related to this log Sobolev inequality. So, we will start with the connection between log Sobolev inequality, the Gaussian log Sobolev of inequality and so called Stam's inequality. So, let us begin.

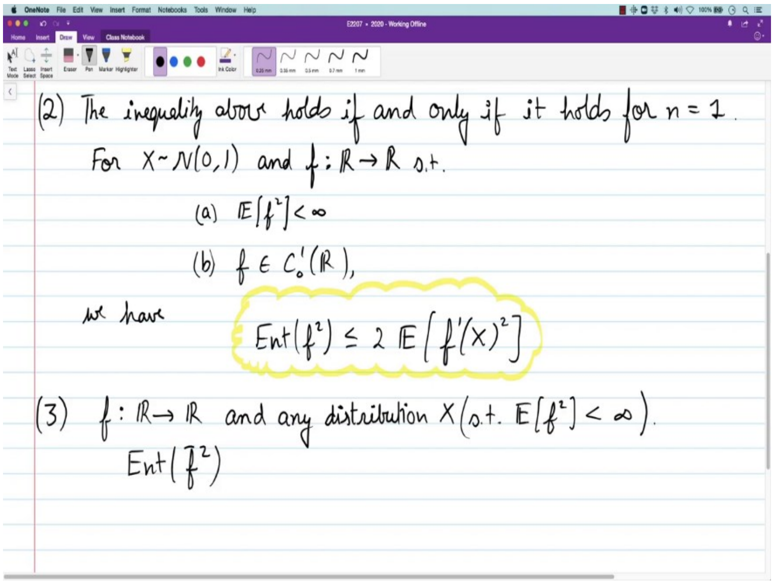
So, let us first recall the Gaussian log Sobolev inequality in its general form that we saw. So, I will just abbreviate log Sobolev inequality by this LSI these 3, this is a LSI stands for log Sobolev of inequality. So, the Gaussian LSI that we saw was the following suppose X is a Gaussian random variable with mean 0.

And identity covariance matrix and let f and let f be a function from R n to R such that it satisfies two properties. This function it satisfies two properties first is that this function has finite second moment and second that, the entropy of f square is sorry second that this function is continuously differentiable ok and let us say its derivative is and also let us say it has a compact support ok.

So, suppose you have such a function and given this function we have a bound on its entropy, entropy of f square remember entropy can be defined only for non negative function. So, we just take a square here. If this entropy is less than equal to expected value of this function is differentiable. So, we can talk about its gradient. So, this gradient is a vector of length n and we take the two norm of this vector and this is the inequality this is the log Sobolev inequality ok.

This is the Gaussian log Sobolev inequality. We presented a proof of this inequality where we first derived the binary log Sobolev inequality and then we use central limit theorem to derive the Gaussian log Sobolev inequality.

(Refer Slide Time: 04:57)



Now, one important property that we used in all these proof is the tensorization property which essentially its just that, it suffices to show this inequality for n equal to 1 for the one dimensional case.

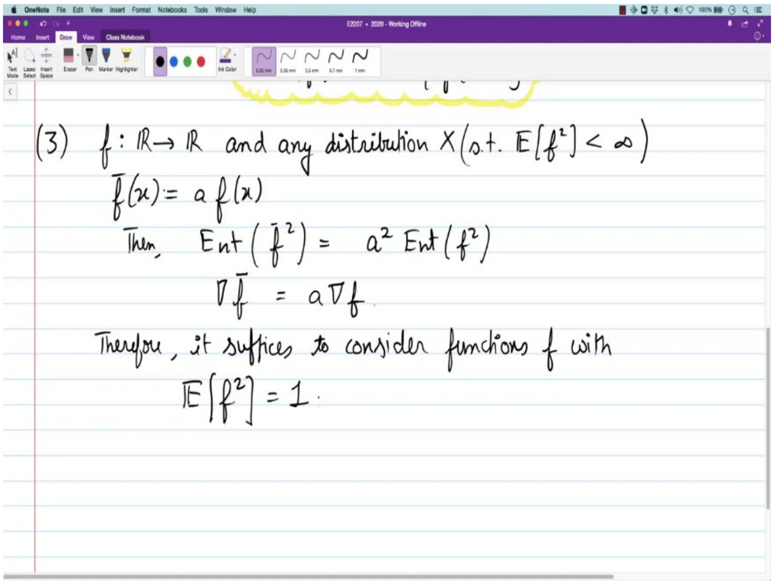
So, the inequality above holds if and only if it holds for n equal to 1. So, if this inequality holds for all functions for n equal to 1, then it must hold for all functions for larger n ok that is the that is the that is the result. This to show this result we use the tensorization property of entropy in other words the inequality that we showed was the following.

So, we can now think of just a Gaussian a single Gaussian random variable with mean 0 and variance 1 and consider a function f from R to R such that it satisfies two properties, one that it has finite variance and second that this function f belongs to again same as before except that the domain is now R instead of R n ok. Suppose, we have these two properties then we have again this is the one dimensional version.

So, entropy f square is less than equal to 2 expected value. So, we can just talk about the derivative because now we are in, we are talking about functions on R this was the second form that we this was the simplified inequality that we actually showed and it is very important that we could reduce our original inequality to this one dimensional version and this is what we said was that this is an elementary inequality, it involves few random variables few parameters strictly for the binary case and therefore, it can be directly verified ok.

Next, what we observe is that we can even further relax the requirement, we make the following observation that for a function f from R to R and any distribution X such that let us say the entropy exist. So, function is let us say the function is non negative we will take f square such that entropy such that this is finite second moment is finite the entropy of.

(Refer Slide Time: 08:24)

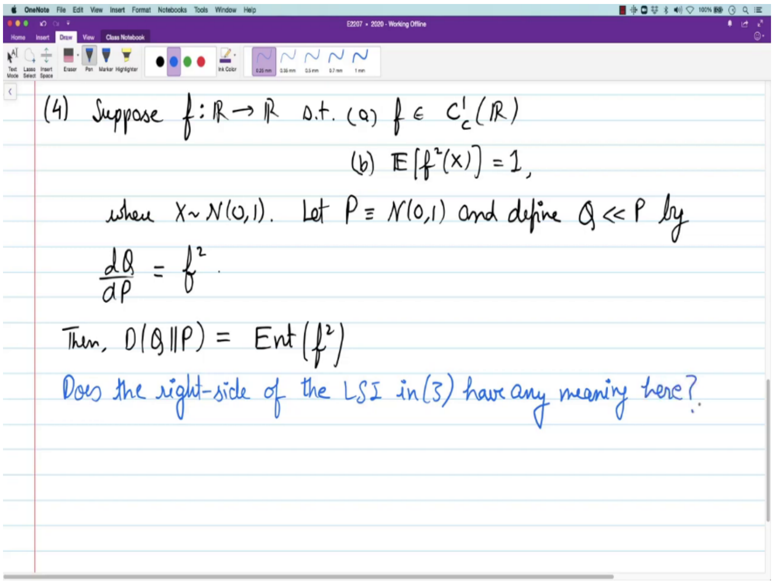


So, consider this function f and now consider another function let us call it f bar of X which is just a times f(x). This is the definition, then entropy of f bar square = this is something you can easily check = a square times entropy of f square and clearly the gradient of f bar = a times gradient of f ok and therefore, it suffices to consider functions f with we can normalize them expected value of f square equal to 1 ok.

We just had expected earlier we were saying expected value of f square is finite but we can always normalize a function the inequality does not change ok. So, we will that is another equivalent form. So, what I am saying here is that in the second form here we can we might as well just require expected value of f square equal to 1 ok. So, basically the form we have is that for every yeah need not write it again.

So, its the same as this previous one except that we have expected value of f square = 1 that is the third form that one suffices ok. So, now, that we have this simplification we will provide slightly different looking form ok.

(Refer Slide Time: 10:16)



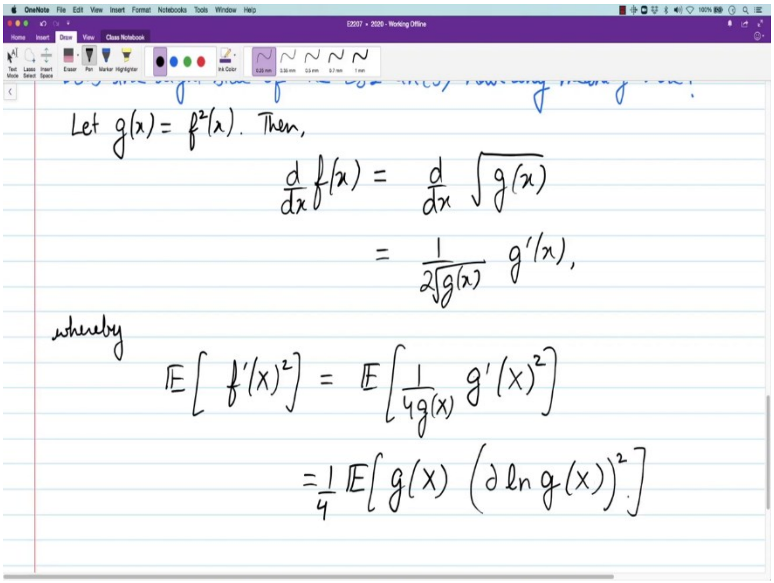
So, suppose you have a function f from R to R such that f is continuously differentiable and expected value of f square is 1 ok where X is the standard Gaussian random variable standard normal random variable. Suppose you have such a function. So, we will define a new probability measure let P be the let P be the Gaussian measure. So, let P means this Gaussian measure and define Q that is that has a density with respect to P by. So, its given by we can just since it has a density with respect to P we can just describe the density.

The density is given by let us say f square ok that is the definition of Q. The Q is absolutely context with respect to P which means it has a density with respect to P and that density is the same as f square. So, if you look at such a Q then is something we had seen earlier then the divergence between Q and P is exactly equal to the entropy of f square ok right ok.

So, this is saying that the left side of log Sobolev inequality that we want to prove can be express as entropy of can be express as a divergence between Q and P, but what about the right side? So, does the right side of LSI in 3 have any meaning here. So, we saw that the left side is just the divergence here the left side is just the divergence.

What about the right side? So, it turns out that in fact, right side also is a rather interesting quantity ok. And we will tell what that is. So, here is the idea.

(Refer Slide Time: 13:03)



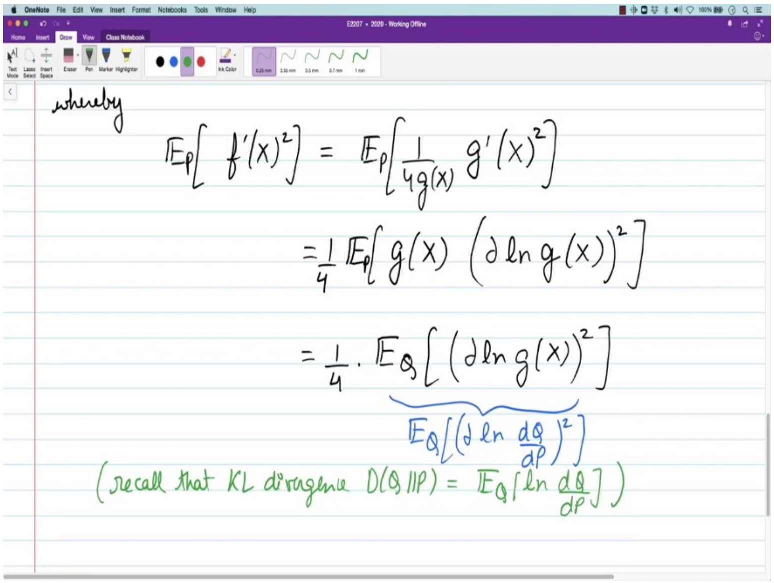
So, what we do is let us say let g x be just f square, then if you look at the derivative of f x we might as well assume that f is non negative because we all only work with we only work with f square here ok and then we take a mod of the we take the mod of the gradient. So, we can we might as well just assume that f square is non negative, f is non negative.

So, this guy is the same as derivative of square √ of g x which is the same as 1 by 2 g x g' x right and therefore, the right side of LSI which was expected value of the derivative of f square. So, this term which appears on the right side is a factor of 2 appears that appears on the right side this = expected value. So, we take this square here.

So, this becomes g' X square and we get 1 by 4 g X, but this can also be written as expected value of let me take this 1 by 4 outside expected value of g X times. So, you get g' X square by g X square because there was this g X I multiply and divide it by g X. So, what is g' by g X whole square? Where that is just the derivative of log of g X ok.

So, I am using this notation so derivative of log of g X whole square ok and in fact, this g is the density of Q with respect to P and all these expectations are with respect to the Gaussian measure P.

(Refer Slide Time: 15:17)

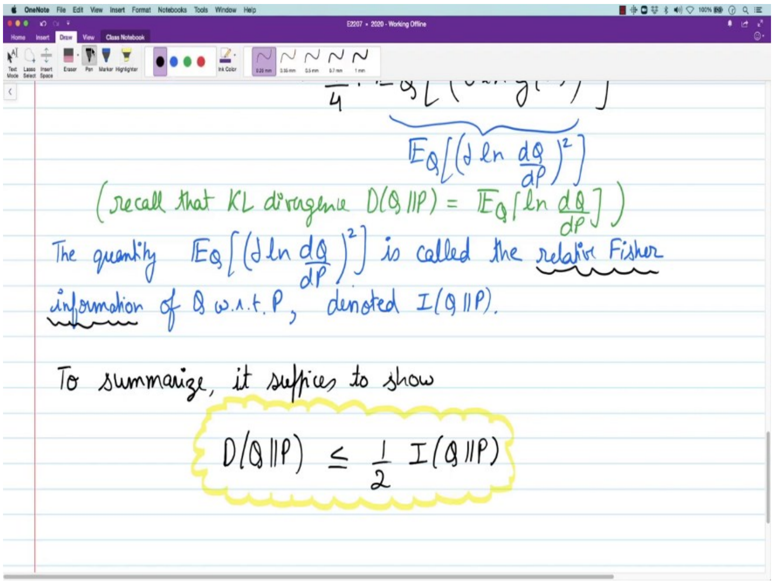


And when you multiply it with density which is something we discussed earlier about it on Radon-Nikodym derivatives or densities this makes it expectation with respect to Q of this guy, ok right.

So, we can view it as follows, this guy here is expected value with respect to Q of the derivative of log density of Q with respect to P whole square. Just for comparison recall that Kullback Leibler divergence of Q with respect to P is the expected value with respect to Q of log of density ok and this guy now looks like the derivative of log of density whole square.

So, this in that sense this looks also like some measure of divergence between Q and P and in fact, this is also a well studied measure and it has a name ok.

(Refer Slide Time: 16:33)



So, we will call this thing the quantity E Q expected value with respect to Q of derivative of log likelihood ratio of Q with respect to P square is called the relative Fisher information and of Q with respect to P.

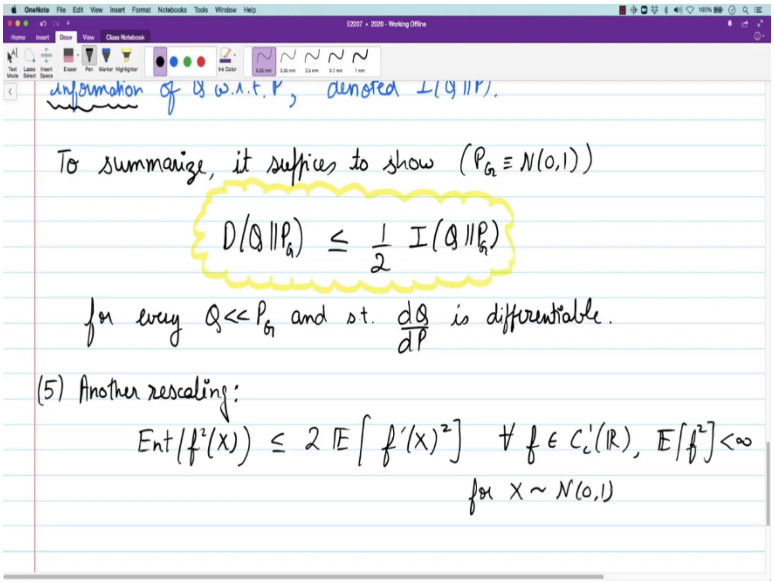
So, to even talk about this relative Fisher information we must Q must have a density with respect to P and that density must be differentiable only then we can talk about relative Fisher information and perhaps a better name would be Fisher divergence then it would, then you can easily draw a parallel between Kullback Leibler divergence and the Fisher divergence ok.

So, it turns out that this quantity on the right also that quantity that we saw on the right also is a also looks like a standard quantity. So, to summarize we summarize to have the general Gaussian log Sobolev inequality, it suffices to show. This is equivalent exactly equivalent to the general log Sobolev inequality it suffices to show that D Q P is less than equal to.

So, there is a factor of 2 but there is a 1 by 4 here half ok sorry one notation here this relative Fisher information is denoted as I of Q P. So, it suffices. So, basically what we the log Sobolev inequality where after is exactly equivalent to saying D Q P is less than equal to half of I Q P it is similar to very similar to Пnsker's inequality ok.

Except that now, you are showing it for divergence and showing that divergence is in turn less than equal to half of some other notion of distance this relative Fisher information distance.

(Refer Slide Time: 19:21)



And we need to show this for every Q that is that has a density with respect to P that is absolutely context with respect to P and such that its density is differentiable ok.

So, any Q that has a continuous differentiable continuously differentiable entropy with respect to P, if we show that it satisfies D Q P less than equal to half I Q P and the here P is the Gaussian measure, then we will have the Gaussian log Sobolev inequality ok. Just to insist that P is the Gaussian measure, I will just put P G instead of this, P G standard Gaussian measure ok P G ok. By the way just a quick fact having density with respect to Gaussian is equivalent to having the standard density ok and that is equivalent you can show that.

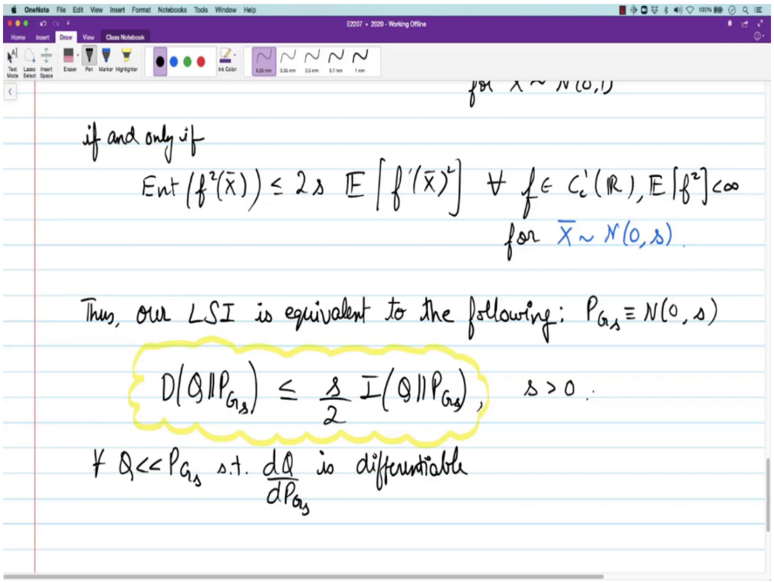
So, a measure is absolutely context with respect to the Gaussian measure if and only if its absolutely context with respect to the Lebesgue measure the standard measure that we use for integration. And in fact, the density with respect to Gaussian will be given by the standard density / the Gaussian density ok.

So, this log likelihood ratio this density with respect to Gaussian is this density with respect to Gaussian is just the likelihood ratio alright. So, now, we see a different form of log Sobolev inequality, the amazing thing here is that it was a dry inequality not having so, much meaning and now suddenly it looks like we are talking about sort of relation between two different notion of distances between distribution.

So, you have a Gaussian distribution and this Q which is slightly away from Gaussian distribution, then you want to claim that the divergence between Q and P G is less than equal to half times relative Fisher information between Q and P G ok that is what we wanted to show. Indeed we have already shown this; this lecture is about showing different equivalent forms of the same inequality alright ok.

In fact, we can. So, we have seen now four different equivalent forms we are continuously simplifying the original inequality so in fact, there is another equivalent form we can do another rescaling. So, observe that entropy of f square X is less than equal to 2 times expected value of f prime X square for all f that are continuously differentiable with expected value of f square equal to 1 or let us say finite its a bit more convenient form where for X distributed as standard normal this inequality holds if and only if it holds.

(Refer Slide Time: 22:57)



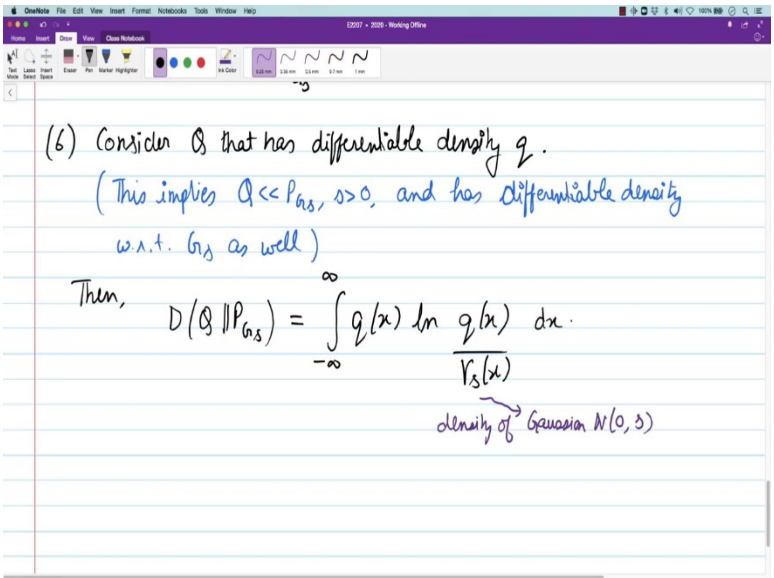
So, now, I am trying a different random variable here, if this guy holds for all f ok. I am changing X to X bar for now we will use a different distribution for X bar that has Gaussian 0 mean and variance s ok. This is easy to see you just rescale X to X by to multiply and divide X with s and use s X as X bar.

But the function changes from f(x) to f(x) by s ok that is the new function, but it holds for all functions so, that is why you get this equivalence just a simple rescaling argument. So, it suffices to verify this inequality for s equal to 1, but once you get it for s equal to 1 you get it for all s ok. So, our log Sobolev inequality is equivalent to the following thus our S I is equivalent to the following.

D Q P G is less than equal to s by 2 I Q P G for all Q that have density with respect to Gaussian which that the density is this is differentiable sorry this is not density with respect to Gaussian this is density with respect to a Gaussian with variances ok. So, P G s is a Gaussian with 0 mean and variances ok. This is the; this is the this is the log Sobolev inequality that we have already alright ok so far so good. So, if it holds for s equal to one it holds for every s greater than 0 ok.

Now, what we will do is we will see another equivalent form we have seen so, many equivalent form, now we are at this form 5 now we will see a final equivalent form which is form 6 ok I will just call it form 6.

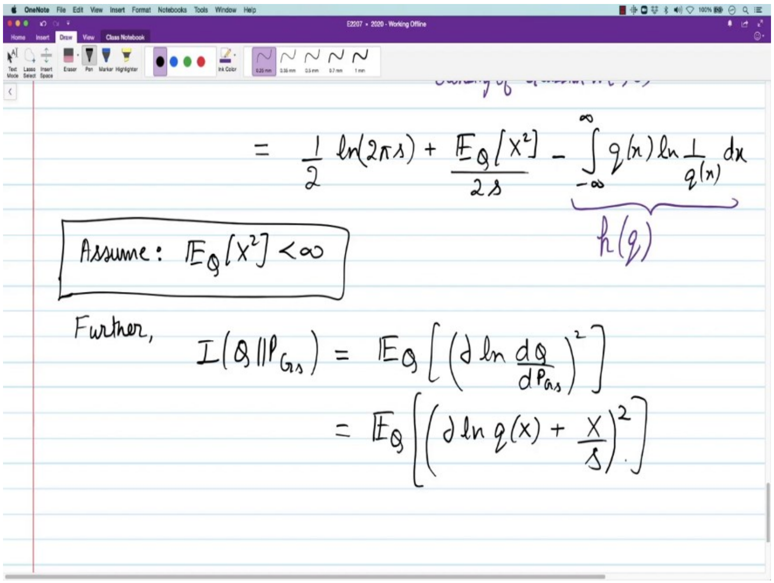
(Refer Slide Time: 26:15)



So, what is the 6? So, let us consider Q ok let us say the Q let us consider Q that has differentiable density Q ok.

So, then it if you assume this about Q, this implies Q is absolutely continuous with respect to g s for any s greater than 0 and as differentiable density with respect to G s as well ok. In fact, this density with respect to G s is the ratio of Q X by Gaussian density ok. So, what is D Q P G s? This = q of x. So, the integral is from - infinity to infinity over the real line log q of x by this γ s x ok. This γ s x is the Gaussian density this is density of Gaussian N 0, s ok.

(Refer Slide Time: 28:32)



So, that is what this is and we can expand the log. So, this Gaussian density we can write as we know what this is E to the power - X square by 2 s / √ 2 π√2 π s correct. So, when we substitute that and take this log here what we get is half log 2 π s + expected value under Q of X square by 2 s this is for this denominator part and then - .

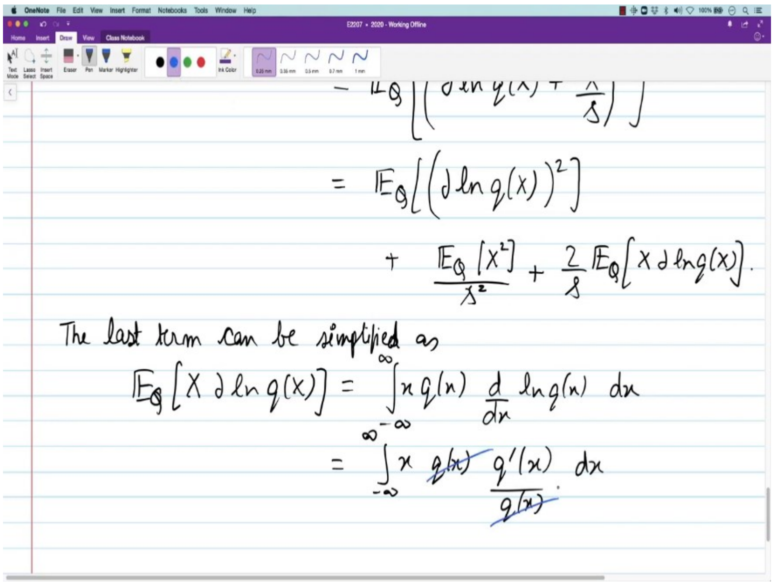
So, this quantity that we see here if you have not seen it before is sort of an entropy just like Shannon entropy captures cardinalities of large probability sets, this is what is called differential entropy and it sort of captures the volumes of large probability sets ok. So, you may not even worry about this operational significance of this differential entropy and just view it as some quantity ok this formula here differential entropy ok.

So, you have this differential entropy and let us also assume its an assumption that expected value under Q of X square which appears here is finite ok that is an assumption that we make ok. So, this guy just this therefore, ok. So, we have already seen this. So, D Q so D Q P that we have here is half log 2 π s + E Q X square by 2 s - differential entropy of Q ok.

That is what d Q that is what the left side of our log Sobolev inequality is what about the right side? Let us see further if you look at the right side I of Q P G s which we said was equal to expected value with respect to Q of the derivative of the density whole square the density here is Q X by γ s X its derivative of its log is yeah.

So, let us do the derivative of this log. So, that is the derivative of q X right + the derivative of log of Gaussian. So, log of Gaussian is just - X square sorry - the derivative log of Gaussian. So, log of Gaussian is - X square by 2 s here and its derivative is 2 X by 2 s - 2 X by 2 s and so, that two cancels and so, all you get here is X by s ok this whole square expected value with respect to Q ok.

(Refer Slide Time: 32:16)

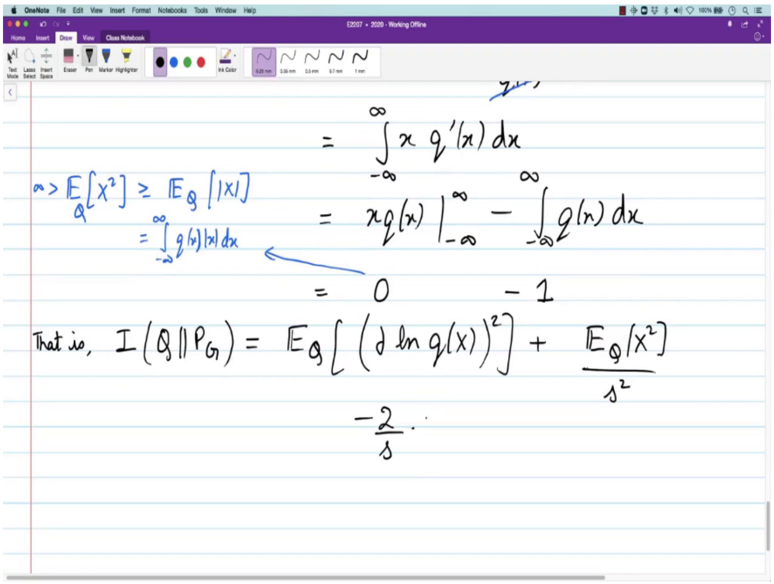


So, we can expand this whole thing, this can be written as expected value over Q of this quantity. This is log of density the Gaussian part of the density has now disappeared this is log of the standard Leibler density + this guy here expected value under Q of X square by s and + 2 by s into expected value under Q of X times this guy ok.

So, let us look at this last term here, this is the expression for information note that we will multiply this with s by 2 for our inequality. So, this term here sorry there is no by s square by s square. So, when we multiply this with s by 2, we see what we see on this side ok. So, this term will cancel on both sides of inequality ok let us look at the last what is this term? The last term it can be simplified further can be simplified as expected value over Q of X times this = I will do I will write.

Let us say for simplicity let us just write the whole integral q x x d by d x ln q x d x now we can do integration by part ok actually even before that this we can write this derivative of log q X as Q prime X by Q x. So, this looks like this to infinity x q x q prime x because q is differentiable by q x d x.

(Refer Slide Time: 34:48)



So, this term goes away and then this becomes. Now when we use integral by part we take this as the first part and this is the second part.

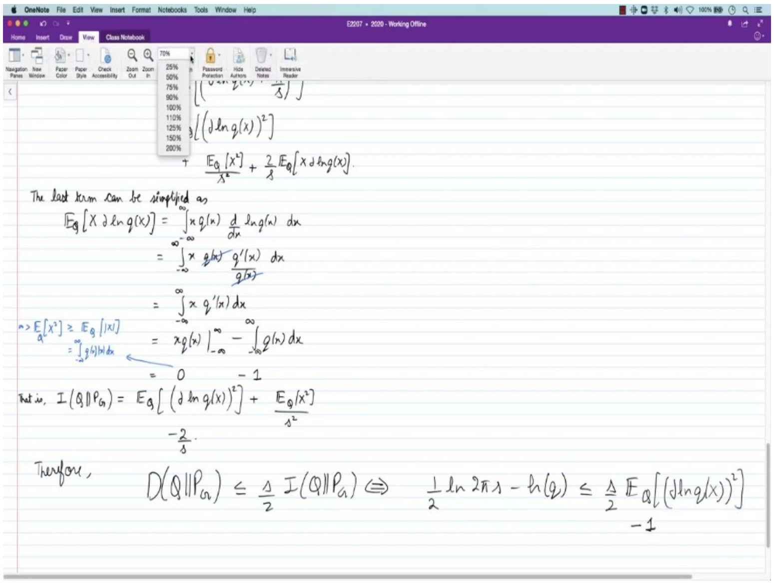
So, this becomes x q x - infinity two limits we have to take - q x d x. So, now, what is this second term? This is a density and we integrate the density to its just a - 1. So, this term is just - 1 what about this first term? This term is 0 ok and the reason why this term can be seen to be 0 is because of an assumption that we made the fact that this first term is 0 can be seen as follows.

So, we have assumed that expected value under Q of X square we have assumed that this guy is finite, but this second moment always exceeds first moment and therefore, this is finite and therefore, if you take this is by the way the integral - infinity to infinity q x mod x d x and therefore, in the limit as x goes to infinity both q x x and q both q x in both the limits as x goes to infinity and - infinity q x mod x must be 0 and therefore, both these terms are 0.

So, this becomes 0 ok. So, basically this term this X derivative of log q X this looks so formidable, but this is just - 1 ok. So, this last term here in this expression is just - 1. So, to summarize what we have shown is that this information this relative Fisher information = this + expected value under Q of X square by s square - 2 s I think its - sorry - 2 over s - 2 over s ok.

So, this is this term and the first term divergence was this one here let me just zoom out. So, that we can see both of them at the same time hopefully right ok. So, you see s by 2 here you see log s by 2 here this term here and we see this term here right. So, this implies.

(Refer Slide Time: 38:13)

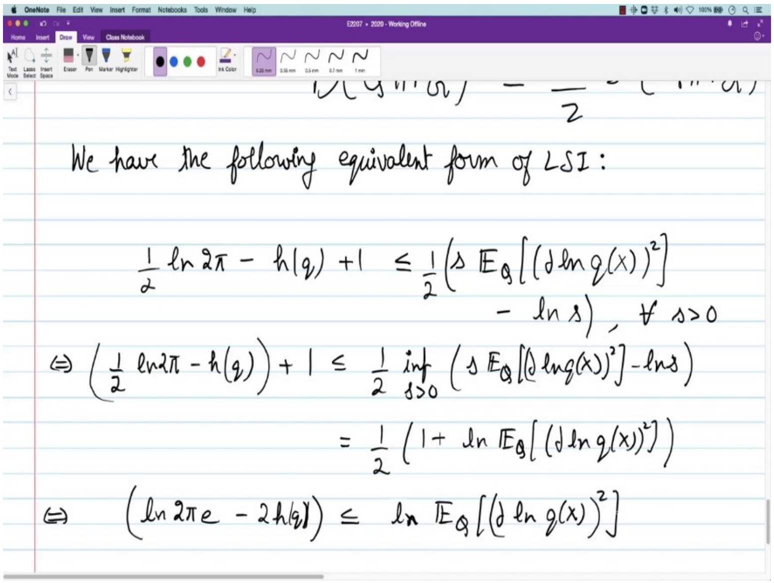


Therefore, D Q P G is less than equal to s by 2 I Q P G if and only if this time thing cancels here if you do s by 2 this is same as this.

So, this guy cancels this is - h q this. So, what we get is when only if half ln 2 π s - h q is less than equal to s by 2 this guy here E Q of - 1 ok that is that is the equivalent form that we can have ok. So, this is the equivalent form that we can have is exactly equal to log Sobolev inequality alright.

So, we now what we will do is we will take all the s related term on the same side and get the other terms from the other side. So, what do we get by the way this is - s by 2 right that is - 1 right. So, this is the same as we have the following equivalent form of LSI ok.

(Refer Slide Time: 40:05)

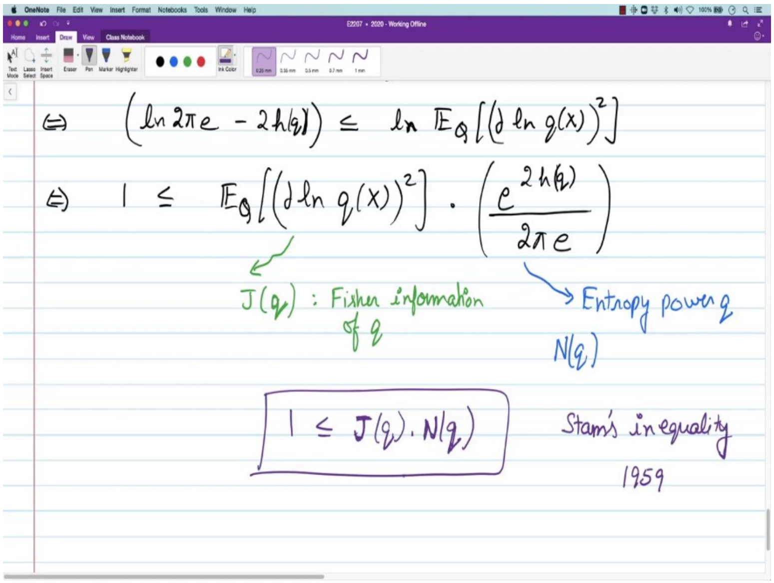


So, what is this equivalent form? It says half ln 2 π - h of q + 1. Because there was a - 1 on that side that I brought here is less than equal to s take this half out s expected value under Q of this expression here l n q X whole square - l n s this is this l n s here ok and this must hold for all s greater than 0. So, this is true if and only if half l n 2 π - h q + 1 is less than equal to half smallest value the infimum over s of this guy - lns.

So, what is the infimum over s we can just take the derivative. So, when you take the derivatives, this is 1 by s and this is this so, that its attained for s equal to 1 by this. So, when you put s equal to 1 by this what you get is equals to half of 1 + ln expected value over Q ok that is the inequality which we get right and so, now, we rearrange the terms we get half on this side and what we get.

So, we first subtract half from this one and then we take this factor of half here. So, what it says is that half. So, ln 2 π e - 2 h q is less than equal to ln expected value over Q this derivative q X whole square ok that is.

(Refer Slide Time: 43:36)



So, we will change some things and take it on this side. So, that is the same as saying that 1 is less than equal to expected value of this is this term and the second term is e to the power 2 h q / 2 π e ok that is the second term ok.

So this is these are two terms which look sort of or defined in fact, these two terms have a rich history this is the so called entropy power of q ok entropy power of the beam of q sorry of the density q ok and its denoted by N q its used quite a bit in quite a bit in information theory especially for Gaussian channels.

For example, a basic inequality used to show converse bound for Gaussian channel is called the entropy power inequality which says that N of q 1 N of X 1 + X 2 exceeds N of X 1 + N of X 2 ok that is the entropy power inequality and this term here actually also has a name this is the derivative of the log likely of the log likelihoods this thing is called the score function.

And so, we are looking at the first thing to check is that this expected value of the score function is actually 0 ok under some regularity conditions expected value of the score function is 0. Actually we already show use some regularity conditions here I think no so yeah. The expected value of the score function is 0 and therefore, this is just the variance of the score function and so, this thing is called the Fisher information of the distribution q ok.

So, Fisher information of Fisher information of q. So, unlike the tyπcal Fisher information that we use in sort of these inequalities called Cramer Rao bound or van trees inequality, this is a Fisher information of a prior which also comes up in van trees inequality actually and that prior is q here. So, this is a Fisher information of q.

So, you can associate this special information with any probability measure q with a differentiable density ok. So, what this inequality here is saying is that one is less than J q times N q ok. So, J q times N q is greater than 1 or this is sort of an uncertainty principle says that both Fisher information for q and entropy power of q cannot be small at the same time.

And in fact, this as we saw that this inequality is exactly equivalent to this is exactly equivalent to log Sobolev inequality and this is called Stam's inequality ok and its from actually its from 1959 and the LSI this Gaussian LSI is due to Gross from 1975, but Gross did not know most very likely that this inequality existed this terms inequality exists which was exactly equivalent to log Sobolev inequality. I think this equivalence was shown only in early nineties ok.

So, this was. So, we have shown that log Sobolev inequality is equivalent to this Stam's inequality, we have already given a proof for log Sobolev for Gaussian log Sobolev inequality and therefore, we have already established this Stam's inequality ok. So, yeah Stam the Stam's inequality holds because log Sobolev inequality hold.

So, now that we have seen this equivalence between Stam's inequality and the log Sobolev inequality for N equal to 1 which we already saw was equivalent to the general log sobolov inequality, in the next lecture we will prove this Stam's inequality direct not using sort of central limit theorem and binary log Sobolev inequality which we did earlier.

But directly using some information theoretic method in fact, we will use the operational significance of these quantities J q and N q and hopefully we will be able to give more heuristic meaning to this inequality. Earlier the proof of log Sobolev inequality in some sense was a technical proof just showing some mechanics techniques involved in the proof. But now hopefully we will be able to give more meaning to it and that is what we will do in the next lecture we will prove Stam's inequality. See you in the next lecture.