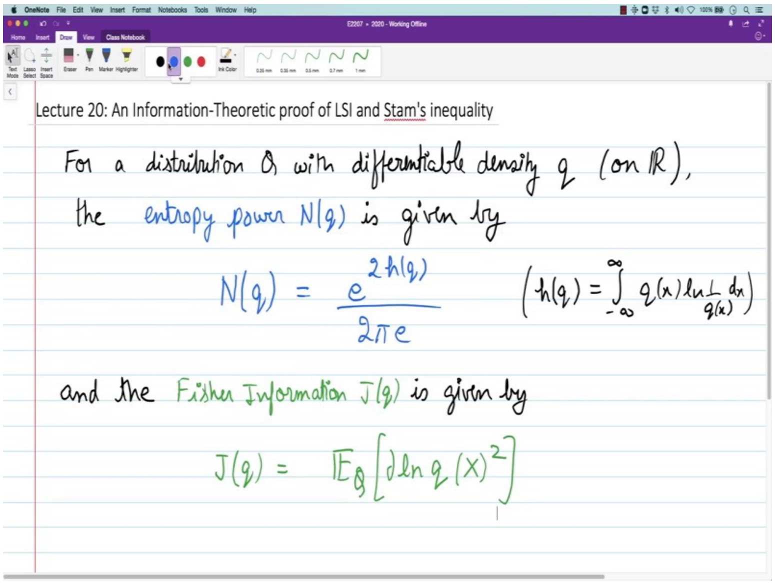
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 20
An information theoretic proof of log Sobolev inequality

(Refer Slide Time: 00:21)



In the last lecture we saw how log Sobolev inequality is equivalent to this inequality called Stam's inequality and here therefore, by showing log Sobolev inequality which we have already done in the past we have also obtained a proof of Stam's inequality. Now, what we will do in this lecture is present a different proof of Stam's inequality and in turn that will imply the log Sobolev inequality because these two are Gaussian log Sobolev inequality, in fact these two are exactly equivalent.

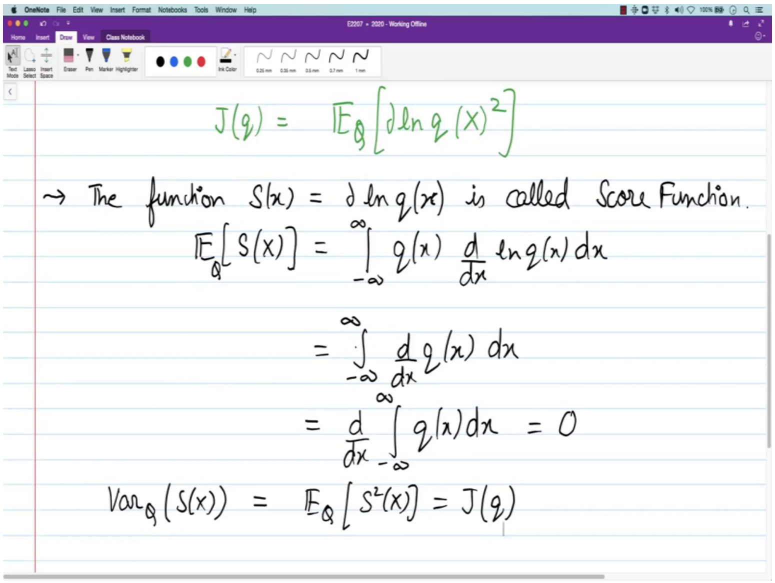
By the way the presentation of this lecture and the previous lecture are based on the now article by now article on measure concentration by Igal Sason and Maxim Raginsky, its one of the references mentioned for the course ok so let us begin. So, recall that for a distribution Q with differentiable density q ok.

The, we consider two quantities; the first quantity was the entropy power N of q is given by N q equals to e to the power 2 by differential entropy of q divided by 2 πe ok. Here this h q was if you remember this was just integral - infinity to infinity q x ln 1 by q x dx ok that is hq.

And the fisher information fisher information on J q this quantity is given by. So, since we assume that the density is differentiable we can take its derivative think of better notation as this one this is the derivative of we can take the density is derivative.

Therefore, we can also take the derivative of its log this is that function the derivative of log of q evaluate that function at X and take the square of it and take the expected value under q of this kind ok this is the fisher information ok.

(Refer Slide Time: 03:35)



So, just a quick discussion on this fisher information, quantity that the random variable or actually the function may be, the function S x equals to is called the score function. Well, why is it called the score function? Because when you are computing the, likelihoods and you are looking for maximum likelihood estimates you often set this derivative to 0 and so you compute the score function and make it small.

So, this score function is used in maximum likelihood procedures and from theoretical perspective this score function is used for analysis ok. So, this is called score function one thing about this score function is that it I mentioned this in the last lecture also the expected value under Q of this score function = let us see its equal to integral - infinity to infinity q x d by d x log q x dx and this is integral - infinity to infinity q prime x dx.

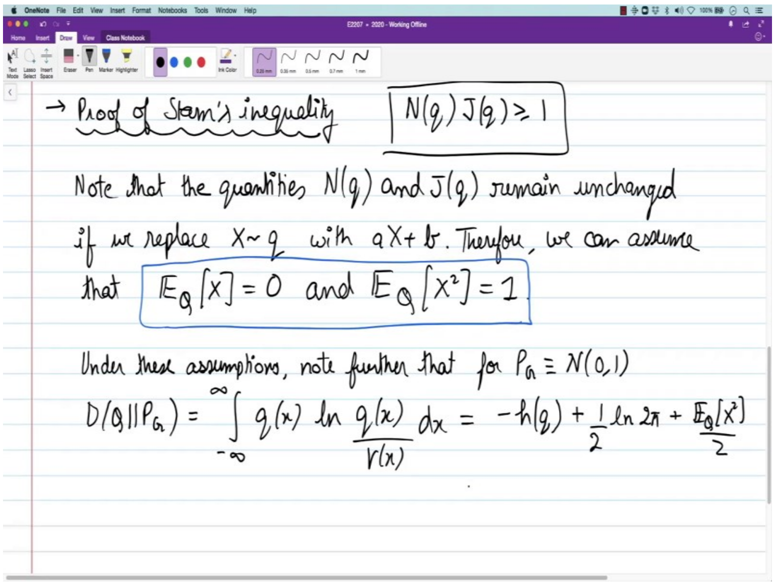
So, this is the derivative of q x here right because derivative of log is 1 by q x so that cancels in this guy, this boy this guy remains. Now, this thing here now we under some simple condition we can take this derivative outside and this integral is always 1 so this becomes 0.

So, this guy is a 0 mean random variable and therefore, its variance under Q = expected value under Q of S square X, which is exactly what we are calling the fisher information ok q it's the variance of the score function ok score function is a 0 mean random variable its variance is called fisher information ok.

There is a lot more geometric meaning that can be given to this quantity and but that is a little bit outside scope for us. This quantity entropy power on the other hand is related to the volume of a large probability set, a set which has large probability under q ok. And it comes up in as I said it comes up in handling lower bound for Gaussian coding problems ok.

So, now that we know meaning for both these quantities we will we want to show Stam's inequality directly. In fact, this inequality is quite old and it has, it has been studied thoroughly in information theory and we will give an information theoretic proof of this Stam's inequality.

(Refer Slide Time: 07:04)



So, towards that so towards proving Stam‘s inequality which is N q J q is greater than equal to 1 that is Stam’s inequality. Towards this note that the quantities N q and J q do not change, q remain unchanged; if the random variable which is generated by q, if we replace X q with some scaled version of it ok.

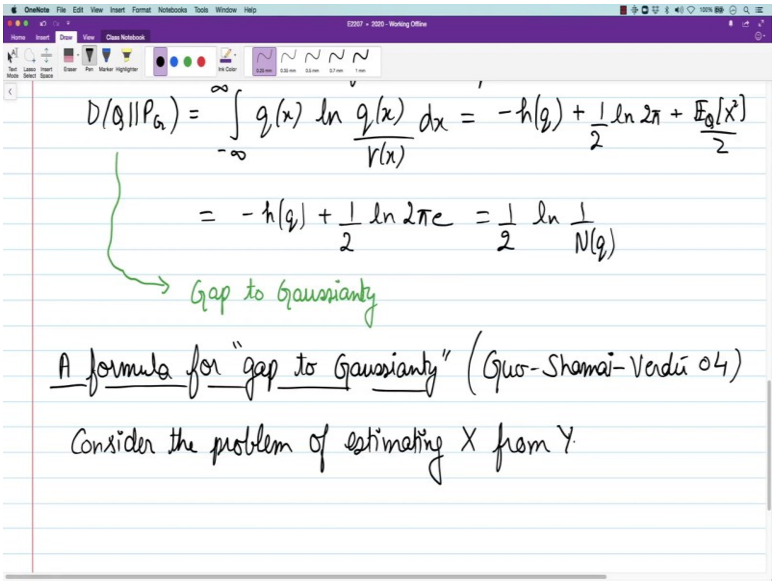
Because it can recover X back from it and these quantity only depend on the distribution such it does not depend on exact values, here they do not depend on these quantities do not depend on these exact values they are a function only of this distribution so if you scale and translate it doesn’t change.

Therefore, we can assume that the expected value under Q of X is 0 and that its variance or a second moment is 1. We can assume that because they remain because this is finite we only want to show it for the finite case that was assumption throughout and we can normalize by that second moment and make it unit second moment and also we can translate to make it 0 mean ok, we can assume that. So, we will from here on we will proceed with this assumption, these are the assumptions we make from here right ok.

So, under this assumption note. So, if you look at the divergence between Q and a Gaussian distribution so this P G is the Gaussian distribution with 0 mean variance 1. So, basically same mean and variance as this X. So, this one this can be written as expected value under Q, maybe something more explicit.

q x ln q x by gamma x this is the Gaussian density dx and this = this first term is just - h q that’s how we defined h q. And this term here we have seen this before its half log 2 π ok half log 2 π + expected value under Q of X square by 2 ok.

(Refer Slide Time: 11:00)

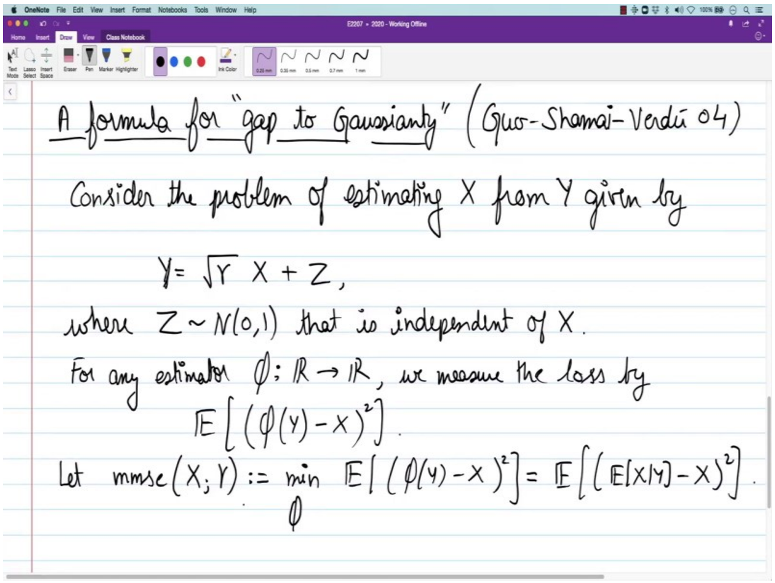


And so this is nothing, but - h q + half ln 2 π e because expected value under q of x square is 1 right. But then this is nothing, but half of ln 1 by N q. So, in other words this 1 by its half log 1 by entropy power is exactly the same as divergence between Q and P G ok. Just exactly the same as divergence between this Q and P G here ok. So, what is this guy here? So, this is basically what we can think of as gap to gaussianity.

So, think of this distribution so q is a distribution we want to figure out how far it is from a distribution a Gaussian distribution with same mean and variance is Q. And one way to measure the distance is divergence between Q and P G, and this is this divergence is exactly what we used to define entropy power is 2 to the power - 2 times this divergence that is what entropy power is. So, gap to Gaussianity related to entropy power ok.

So, with this observation we actually recall now a formula for this gap to gaussianity. So, we recall a formula for gap to Gaussianity ok and this is the formula due to Guo Shamai and Verdui 2004 it says that.

(Refer Slide Time: 13:44)



So, we consider a specific channel so consider the problem of estimating X from by using observation Y given by Y equal to square root gamma X + Z, where Z is a standard Gaussian random variable with, the standard Gaussian random variable that is independent of X ok.

So, X can have any distribution X is a real valued random variable and we look at this observation Y which is some scaled version of x + some noise, Gaussian noise. We want to estimate X from this noisy measurement about X and we want to do this, we want to do this under a mean square error criterion.

So, for any estimator phi which looks at we will use phi with Y as input and we will get some estimate X hat. We measure the loss expected loss by the mean square error ok, by expected value so we use mean square error phi y - x square ok. Let mmse minimum mean square error for X, when this strength here is gamma.

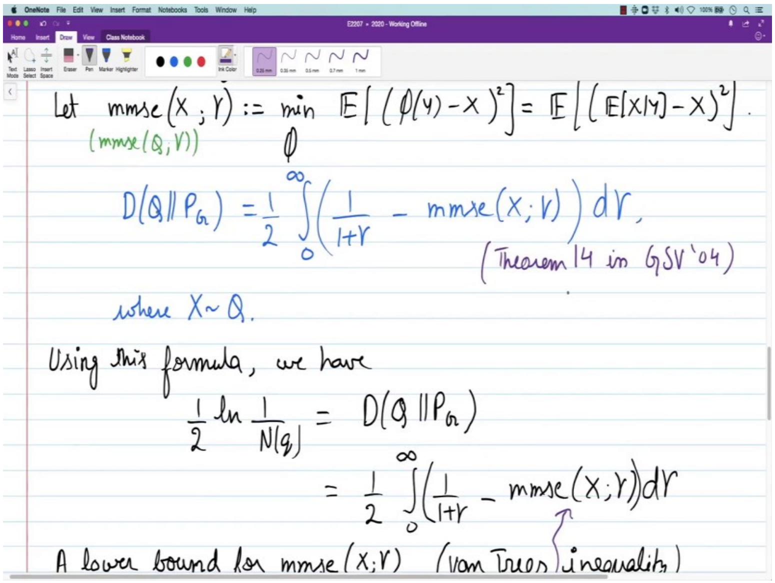
This is this gamma denotes the signal to noise ratio because the noise variance is 1, this is the signal variance is gamma we can assume X is 0 mean ok. So, this is defined as this minimum mean square error is defined as minimum over all phi of this guy.

And indeed it is well known which function phi attains this minimum right that is something that you learn in any basic statistics course the detection estimation course in the ec department that function which attains this is the conditional expectation ok.

And once again I emphasize that this is valid for any joint distribution here in this case we are putting this specific distribution where the distribution of x is arbitrary, but distribution of z is the Gaussian distribution ok.

So, we define this minimum mean square error this way, it turns out that there is a very nice sort of beautiful formula that characterizes gap to Gaussianity in terms of this mmse function and this formula is due to Guo Shamai and Verdui, I will recall this formula here.

(Refer Slide Time: 16:44)



So, the formula says that the divergence between Q and standard Gaussian P G = integral - infinity to sorry, integral half integral 0 to infinity 1 by 1 + gamma - mmse X gamma. So, integrate this thing or this ok where X is distributed according to Q ok so this X here has distribution q.

In fact, this there is a little bit of use of notation this mmse does not depend on X it depends only on Q so maybe I will; I do not know which one you prefer. So, you can also think of it as mmse of Q gamma it depends on the distribution not really the realization ok.

So, this is its clear what this term is what is this term this 1 by 1 + gamma, this term here represents the minimum mean square error for the case when x itself was a Gaussian with 0 mean and unit variance ok. Then we can actually see that the conditional mean is a linear function and it is the; it is the mean square error for that corresponding to the conditional mean.

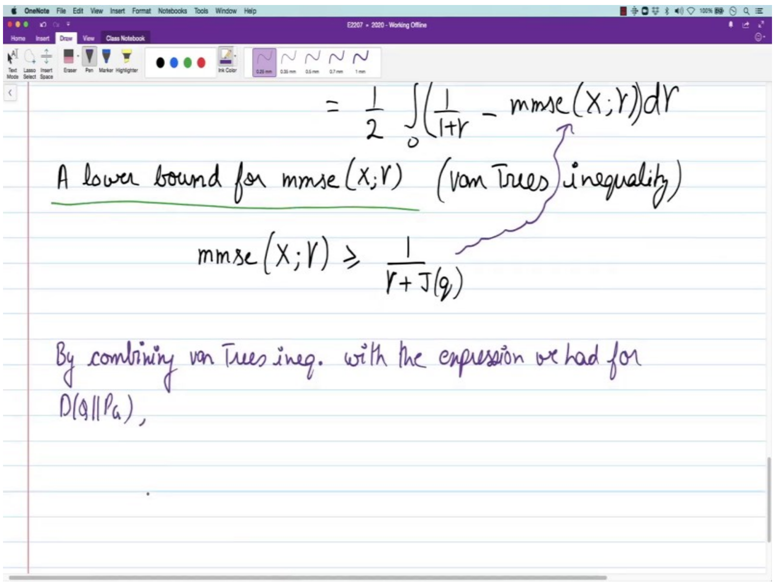
So, it is the so, it is very nice so basically the divergence between Q and Gaussian is the same as the difference between the minimum mean square error for the case when X is Gaussian and for the actual distribution Q which is the distribution of X this difference integrated from 0 to infinity ok.

So with the so this with this formula we will now use this formula to derive Stam's inequality, we had seen that half so using this formula for divergence. So, earlier we had seen that half log 1 by entropy power of Q equals to divergence between Q and P G, but this guy = half integral 0 to infinity 1 by 1 + gamma - mmse is d gamma ok.

So, now, we will. So, we have a formula for this half log entropy power, but there is this term mmse X gamma which we still need to handle ok. In fact, another thing that we need another component of the proof that we need is a lower bound for mmse X gamma ok. And, the lower bound that we that we use is.

So, we this was the first component is the first formula we have and now this is the second formula that we are deriving. It is a very popular lower bound its sort of a classic lower bound called the Cramer Rao lower bound we using an extension of it which two is by now sort of classic.

(Refer Slide Time: 20:26)



And this is called van trees inequality ok we will derive this inequality later but when applied to this specific case van trees inequality says that this thing this mmse for estimating X in presence of Gaussian noise is at least 1 + gamma J q.

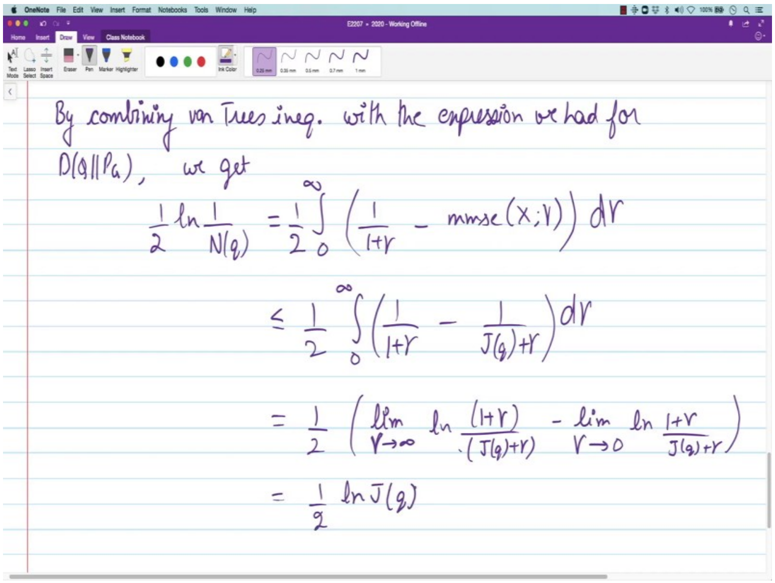
So, little bit of history for mean square error loss a classic lower bound on mean square error loss a lower bound for mmse means that no matter what estimate of phi you apply you can never improve over this mean square error.

So, classic bound for this is the so called Cramer Rao bound which holds for all unbiased estimates unbiased estimators phi, estimators for whose expected value coincides with x ok. Its extension called van trees inequality gets rid of this assumption of unbiasedness, and this inequalities Cramer Rao bound tyπcally entails fisher information, for the channel the noisy channel under which you observe X which is the Gaussian case here.

So, this gamma here represents that fisher information for the channel. This part of the fisher information this J of q is the fisher information for the distribution of the unknown X the distribution of the prior. So, van trees inequality in effect replaces bias with a prior ok. It does not have any condition of unbiasedness, but it works only with the prior. So, for our application here the prior is q so we have this fisher information of the prior that is J q.

And this part here is the fisher information of the parametric family or the noisy channel that is the Gaussian channel that we are using ok, that is what this van trees inequality is. We will prove it later, but if you substitute this van trees inequality here ok, what do we get? So, by combining van trees inequality with the expression we had for D for gap to Gaussianity we get.

(Refer Slide Time: 22:48)

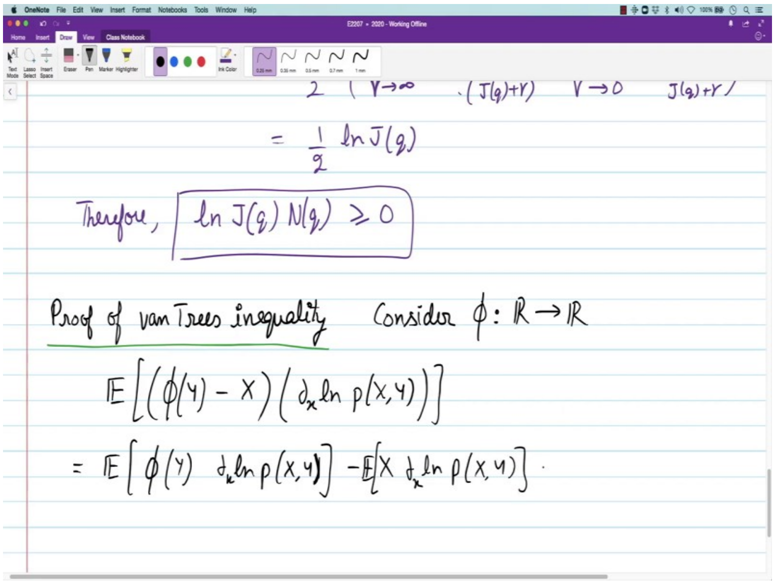


So, gap to Gaussianity was just 1 by half log entropy power and we saw that this is 1 by 1 + gamma - mmse X gamma d gamma sorry. And this is less than equal to I will lower, I will bound this from below so will I have an upper bound on the whole thing 1 by 1 + gamma - 1 by J q + gamma d gamma ok.

So, let us integrate this the first term we can integrate that is log 1 + gamma and in the limit 0 gamma going to 0 this is 0 in the limit gamma going to infinity will keep that part so limit this is an improper integral, in the sense that the in the limit these things will cancel. ln 1 + gamma and this ln J q + gamma - limit ln 1 + gamma by J q + gamma ok, that is what the integral is.

So, in the limit as gamma going to goes to infinity this thing is just 1 sorry this is the log is outside right. So, in the limit as gamma goes to infinity this thing is just 1 so the log of 1 is 0 and as gamma goes to 0 this is 1 by J q. So, that there is a - log so this log J q. So, this is exactly equal to half ln J q ok. So, what have we shown?

(Refer Slide Time: 24:56)



So, therefore, ln J q N q is greater than equal to 0 which is the same as Stam's inequality ok. So, in to in conclusion this inequality comes up quite naturally once we use this formula and van trees inequality. In the remaining lecture I will prove this van trees inequality, this formula itself I will not be proving and this is a sort of a slightly not long an involved proof, but it requires us to introduce too many terms.

So, I will just keep it, but its this paper this Guo Shamai Verdui paper on immse. That this paper is easy to find and it has a very clean derivation of this formula. I think this is theorem 14 in the paper. By the way this formula has been generalized by Verdui to divergence between Q and P ok.

So, this is divergence between Q and P G, when you look for divergence between Q and P instead of having this term here you will have mmse under the assumption that the prior of X is P, that is what you get that is the more general form. This and then it coincides with this Gaussian formula when that P is P G ok.

So, we use van trees inequality in the middle let us see how we derive van trees inequality. Van trees inequality is the sort of the classic statistics counterpart for Fano's inequality because it provides some lower bound for statistical for loss function of statistical estimation. And this particular loss function that we get lower bound for in van trees inequality is the l 2 loss function.

The proof is essentially just the Cauchy Schwarz inequality let me show the proof. So, for any function phi consider this function phi from R to R ok. And then look at expected value of phi of Y - X times the derivative. Let us look at the joint distribution of X Y, this is the; this is the joint density of X Y. Note that for our application this joint density exists ok.

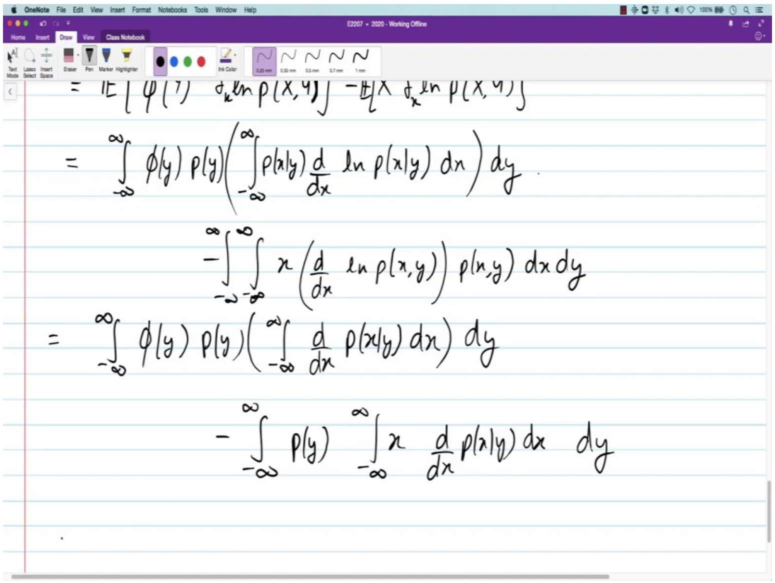
This x has the density p and then a q and then given density of x we have this Gaussian density gamma so this joint density exists for our application, but the proof will give us for more general thing. Let us look at this expected value we can divide into 2 parts, expected value of phi Y - expected value of X; I should be clear here that this derivative here is with respect to x ok.

So, it is not clear right now why I am starting with this quantity, but you can see that by Cauchy Schwarz inequality I can bring in the mean square error times something else and that will be an upper bound on this quantity and we want to understand this quantity. So, I am just computing this quantity first yeah, sorry I am not spending too much time on this proof this is just a quick presentation of van trees inequality.

But, as I said once you start with this particular function you will just you will see that this is just an elementary proof, although it is not clear why we start with this function. Now, we understand this proof quite well, the best proof is the one I have found is in lecture notes on information theme statistics by (Refer Time: 29:23).

And that proof entails a sort of a variational formula for chi square divergence ok, it just follows from that ok, but here I am giving a self contained proof using Cauchy Schwarz inequality ok.

(Refer Slide Time: 29:40)

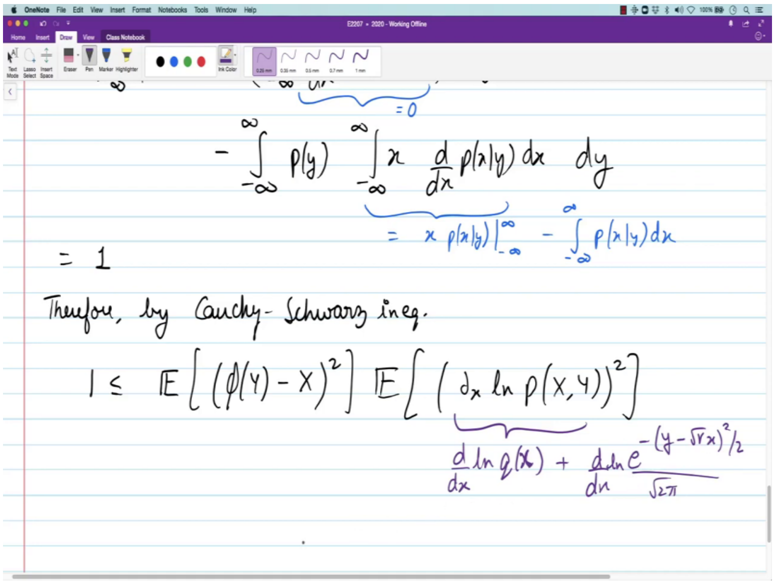


So, once you have this what do we do? So, this can be written as integral - infinity to infinity phi of y P of y, remember y is the output, but you are still doing P of y and then - infinity to infinity derivative with respect to x of log of P of x given y dx and dy, this is the first term here - integral - infinity to infinity x derivative with respect to x ln P x y. I am sorry to bring in density of x also here so this is again density of x given y d x d y ok.

So, when we look at the first term here this derivative is 1 by P x given y and then derivative of P x given y with respect to x. So, that is - infinity to infinity phi y P y, integral d by d x of P x given y for any given y, we can do this d y ok. That is the first term and then the second term is - infinity to infinity again I can do 1 by P x y which cancels this P x y and then derivative of P x y with respect to.

So, let me do the same thing P y integral - infinity to infinity x ok. I am just giving you the technical step without giving all the details you have to carefully see when how you can expand this integral, what can you take out of the integral, what can you differentiate all those things you have to take into account. Now, this thing here =, so now we look at each of these terms separately.

(Refer Slide Time: 32:38)



So, this first term here this one here you can take the derivative outside the integral and then this integral is 1 and then the derivative is 0. So, this can be formalized and it can be shown to be 0 under some conditions, some regularity conditions on this P x given y, but the only thing you want to ensure is that you can take a derivative outside. So, suppose you can do that then this is 0 we have always been assuming these kind of things.

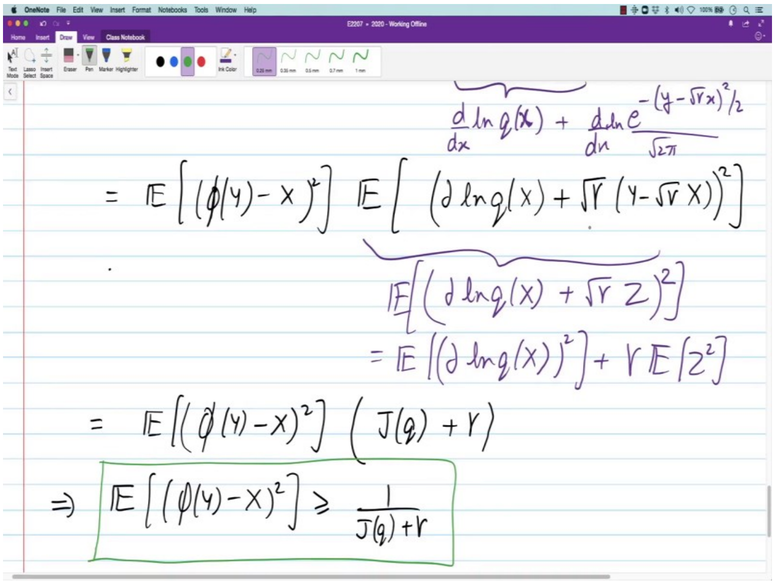
The second for the second term if you look at this kind we can apply integral by part. So, this is x P x given y - infinity to infinity, - integral - infinity to infinity P x given y d x. So, once again as we argued before in the limit as x goes to infinity and as x goes to - infinity in the both those things this will be 0, because we have assumed that x has second moment bounded, but this guy here integrates to 1 ok and so this P y also integrates to 1 so this whole thing becomes - of - 1 that is 1.

So, this whole thing becomes second term is first term is 0 second term is 1 the whole thing is 1. So, that is what have we shown we have shown that this quantity here this in this inner product here is 1. Therefore, by Cauchy Schwarz inequality by Cauchy Schwarz this 1 is less than equal to. So, now we will apply Cauchy Schwarz you can take square of this guy.

And, note what the second term here was the second term was this particular partial derivative ok. So, this is the mean square error that we were after and this is that second term that we have obtained. Let us look at this second term this derivative here is this is d by d x log q x, because the distribution of distribution of x is q the density of x is q + d by dx of p of y given x. Now, y given x is Gaussian ok so this is d by d x of gamma of y - no sorry same notation comes up here yeah.

So, it is the Gaussian of Gaussian density of y - root gamma log of this, x whole square by 2 divided by root 2 π. So, when we take this derivative we will just see that we get y - root gamma x times root gamma.

(Refer Slide Time: 36:02)



So, this whole thing here this can be shown this can be seen to equal expected value of this mean square error times expected value of this derivative +. Now, what is this guy here? This is root gamma into Y - root gamma X yeah that is what it is and whole square.

So, once again we just look at this quantity here this is gamma ln q X + root gamma Z the noise whole square. But x and the noise are independent so this expected value here = the expected value of this q X whole square + gamma expected value of z square, because these two are independent and Z is 0 mean so they are uncorrelated. So, the remaining term is 0 so you get just this ok.

But, z square was a 0 mean unit variance Gaussian random variable so, this thing is just gamma. Therefore, this whole term becomes expected value this is my first term the mean square error times 1 sorry times this term here which you can recognize as J q + gamma ok. So, to conclude this mean square error for any phi exceeds 1 by J q + gamma.

So, this is the; this is the van trees inequality ok this is the van trees inequality and. In fact, this calculations are more or less general holds for any distribution the Gaussian part was not so important here. The place where we use the Gaussian part was this last step otherwise you will have a the conditional distribution here, and that conditional distribution same kind of derivative needs to be taken for the conditional distribution with respect to x the input and squared and then taking expected value.

And, and that last thing is called the fisher information for that conditional distribution 2. So, it would the general form is this J q the fisher information of prior + the fisher information for the channel or the parametric family that you are working with, that is the classic form. In this specific case that fisher information is common so this is van trees inequality.

So, this completes the proof of last component that we use in a proof so just a quick summary. So, we started with observing that divergence between Q and P G in the gap to Gaussianity is half log N q, half log 1 by N q. And then we use this nice formula for gap to Gaussianity. And then this formula involves in minimum mean square error term which we bound using van trees inequality.

And then we come then we evaluate the formula to get half log fisher information and when we combine terms we obtain Stam's inequality namely J q times N q is greater than equal to 1 ok, and in the middle we use van trees inequality which we proved alright.

So, this is what I wanted to cover in this part, we saw that least this we saw that log Sobolev inequality is equivalent to Stam's inequality and then we presented an alternative we presented a proof of Stam's inequality which in turn gives an alternative proof of log Sobolev inequality.

In the next lecture and maybe even the lecture after that you will see connection of log Sobolev inequality to another very interesting class of inequalities namely strong data processing inequalities, which in turn are related to hyper contractivity. So, that is what we will see going forward, see you in the next.