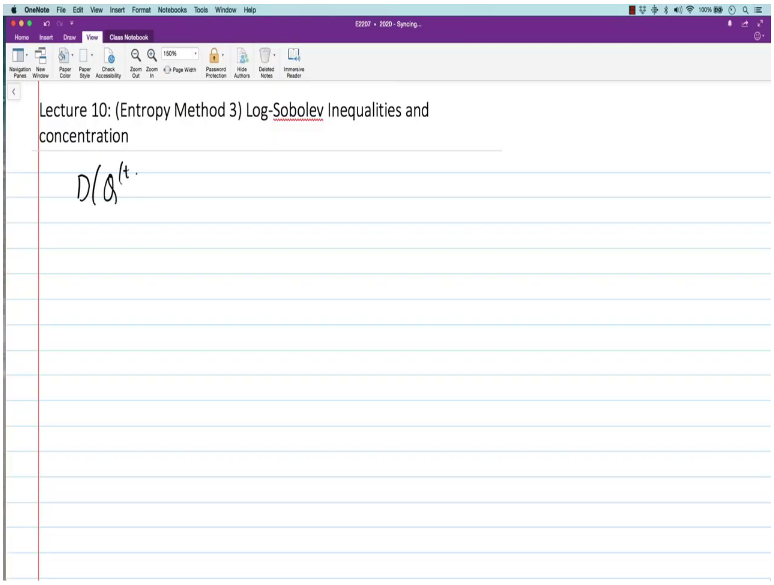
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

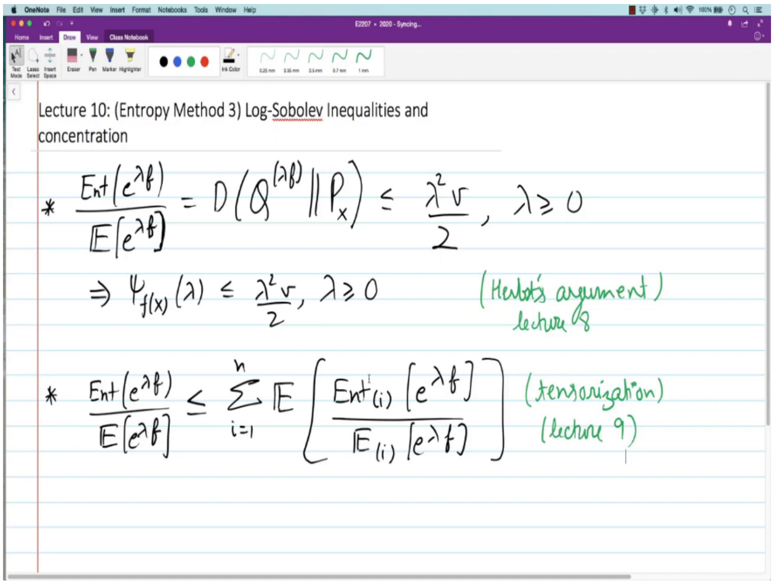
Lecture - 11
Binary and Gaussian Log-Sobolev inequalities and concentration

(Refer Slide Time: 00:20)



We will continue our discussion on Entropy Method and present the so, called Log Sobolev Inequalities. We have been seen that if one can establish bounds for this quantity the divergence between Q of.

(Refer Slide Time: 00:43)



So, if we can establish bound for entropy of E to the power λ f / the moment generating function and this quantity we saw is actually a divergence between a tilted measure λ f and then and the original distribution P.

So, if we can establish bound for this guy, if we can show that this is ≤ say some $\lambda^2 v$ / 2 then the log moment generating function for this random variable x here is less than for this random variable f of x is ≤ again λ^2 v / 2 ok.

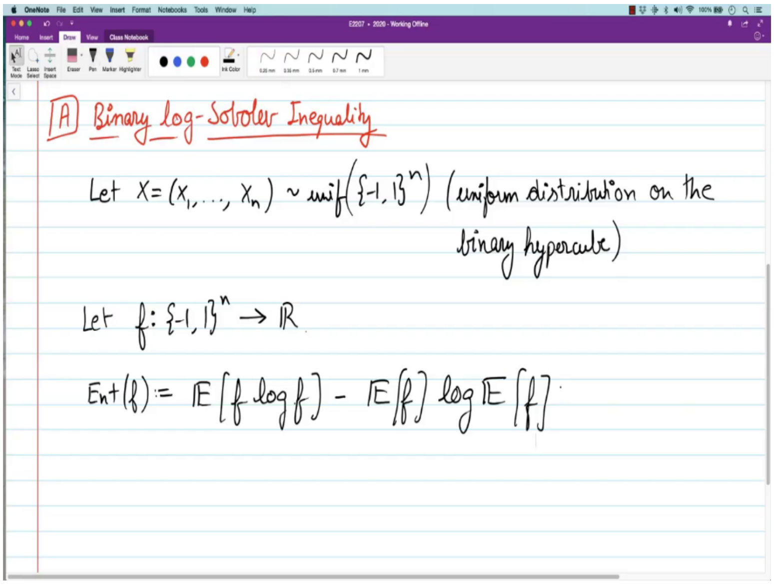
And yeah, this let us do it for only λ ≥0. So, we will get one sided Gaussian tail bound alright. If we can show this and then that was the first part of the entropy method which is called a Herbst argument and second part is a tensorization property.

Here we saw that actually this ratio this entropy of e to the power λ f by the moment generating function, when you have iid random variables can be bounded by something very similar to fano's inequality expected value of similar quantities, but now described by fixing all, but the ith coordinate ok that is the second this is the tensorization part. This was done two lectures ago right lecture 8 and this is what we saw in the last lecture ok alright.

So, this is just essentially a property of this divergence alright. So, now, so, what this allows us to do is to in order to establish a bound like this, you can establish a bound on individual quantities, but how do we establish bound on these quantities and that is where log Sobolev inequalities enter.

Now, log Sobolev inequalities had existence of very rich and rich and successful existence before they were applied to concentration also as well and so, we can first present those inequalities without showing how they have how they can be used for concentration.

(Refer Slide Time: 03:41)

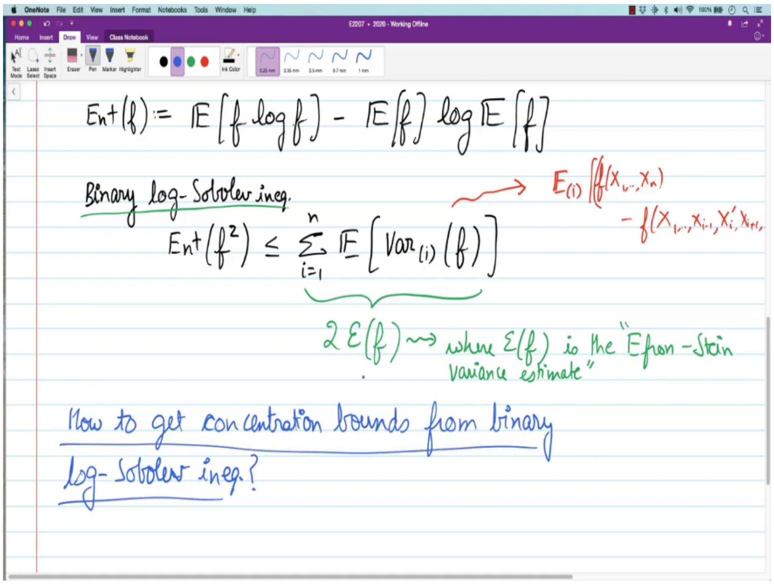


So, I will start with this so, called binary log Sobolev inequality and then we will use it later to derive the so, called Gaussian log Sobolev inequality. Just like we derived Gaussian Poincare inequality from a binary version of that Poincare inequality ok that is the plan. So, let us start with this. So, this is a very specialized inequality let X equal to X 1, X n be uniform on - 1 1 to the power ok.

So, this is the uniform distribution on the binary hypercube so, called this set is called the binary hypercube, uniform distribution on the binary hypercube ok. So, in particular X 1 to X n are independent and they are all random marker.

Next let f be a function on this binary hypercube. So, it is a function from - 1 1 n to R. So, recall that entropy of any function is defined as expected value of f log f - expected value of f log expected value of f ok that is what entropy of this function is. So, the binary log Sobolev inequality that we plan to show looks like this.

(Refer Slide Time: 06:03)



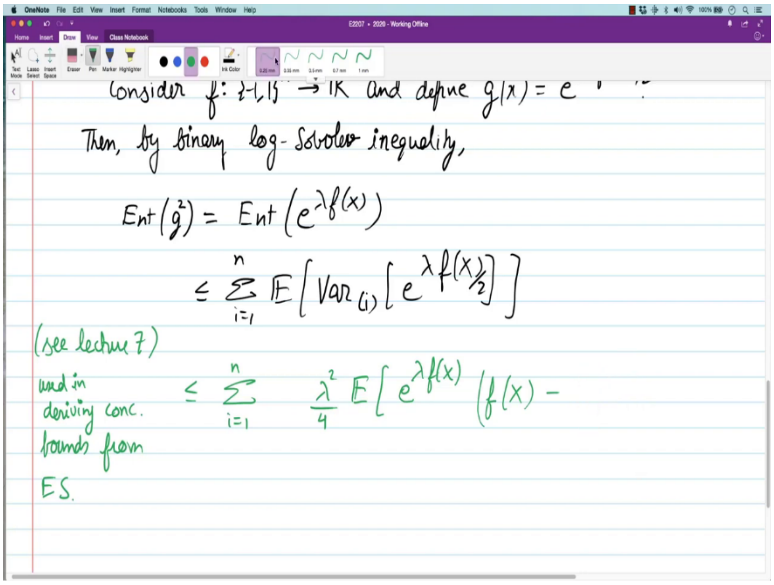
Binary log Sobolev inequality. There are other similar related inequalities, but the one you want to show is the following. Entropy of f^2 is $\leq$ entropy of $f^2$is $\leq$ Σ i equal to 1 to n expected value variance i f ok.

So, we can name this guy this is 2 to 2 times E f where E f is sort of the energy if, but we will call it E f is the Efron Stein. I am just making this term up variance estimate not the very common. So, this is basically the right side of the Efron Stein inequality.

So, this E f. So, remember if you recall your Efron Stein inequality this guy here can be written as expected value given all the past of of X - f of X i f of X sorry may be I will write it more detail f of X 1 to X n - f of X 1 to X i - 1 the place i with the ith independent copy.

And X i + 1 to ^{2} of this ok. This is exactly what the variance is given all the other entries ok. So, that is what this binary log Sobolev inequality will show that this entropy of f^2 is less than this. So, before we prove it let us see concentration on concentration bounds let us see how to get concentration bounds from binary log Sobolev inequality. So, this of course, this is just an example and later we will extend it to other distributions as well ok.

(Refer Slide Time: 09:36)

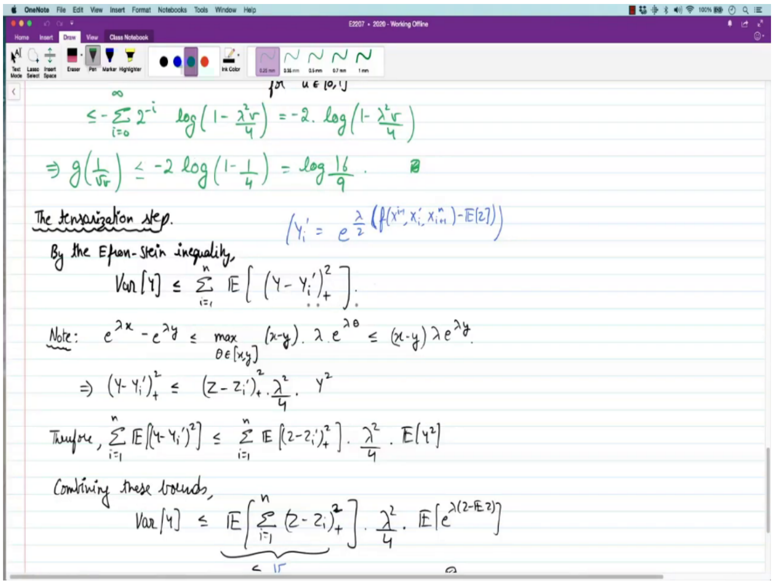


So, what we will do is we will apply this inequality to the function. So, consider f from - 1, 1 to R, but we will not apply binary log Sobolev to this function and define we will apply it to remember we have to look at entropy of e to the power λ f apply it to e to the power λ f x ok be x / 2.

Then by binary log Sobolev inequality, entropy of g which is an entropy of g^2which is an entropy of e to the power λ f x this entropy is ≤ Σ i equal to 1 to n expected value, variance given everything of e to the power λ f x ok and this quantity actually this e to the power λ f x this is the quantity we had seen earlier.

Remember when we were deriving earlier we had shown, when we were deriving concentration bound using Efron Stein inequality, we saw that this quantity here is ≤ see lecture 7, I think. So, used in deriving concentration bounds from Efron Stein you check what lecture was that; was it 7, yeah 7 right.

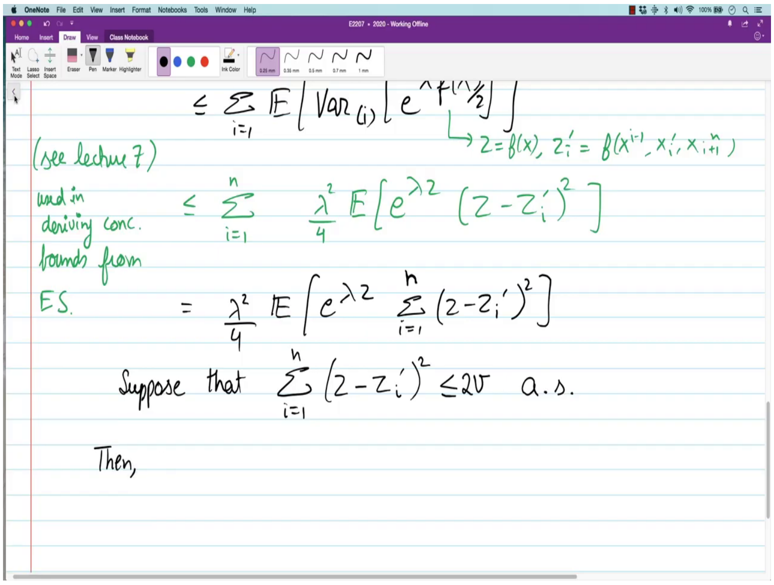
(Refer Slide Time: 11:55)



So, we see we saw this inequality this quantity here this is our variance and we saw that this guy is ≤ um. So, there you had λ / 2, but here we now only have λ is ≤ variance Z - Z i prime times this Y^2 this guy here. So, if we show we showed this guy here is actually sorry this entropy of this is ≤ variance of this by 2;/ 2 remains.

So, we have entropy of $f^2$is ≤ variance. So, / 2 remains alright. So, now, that earlier bound that we saw that one that is why this / 2 is required that gives you each guy here is less than λ^2/ 4 in the same expected guy ok e to the power λ f x this is the bond we saw earlier times f x - f - 1. So, the notation that we used there was.

(Refer Slide Time: 14:05)

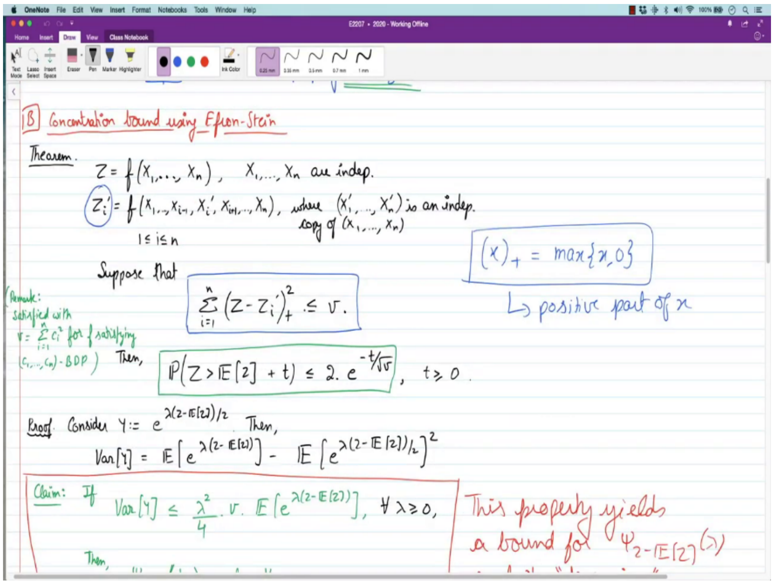


So, this Z was f x and Z i prime is equal to f of x i - 1 x i prime, x i + 1 to n that was the notation we are using for Efron Stein, but that is quite convenient. So, what we get is e to the power λ Z, Z - Z i prime^2 only the positive part of this ok.

So, that would have been a factor of half here this is what we get and therefore, this guy here this is we have seen it earlier, this is a just a Taylor series approximation bound, but very nice. So, what we get is λ^2/ 4 expected value of to the power λ Z Σ i equal to 1 to n v - sorry Z - Z i prime ^{2}.

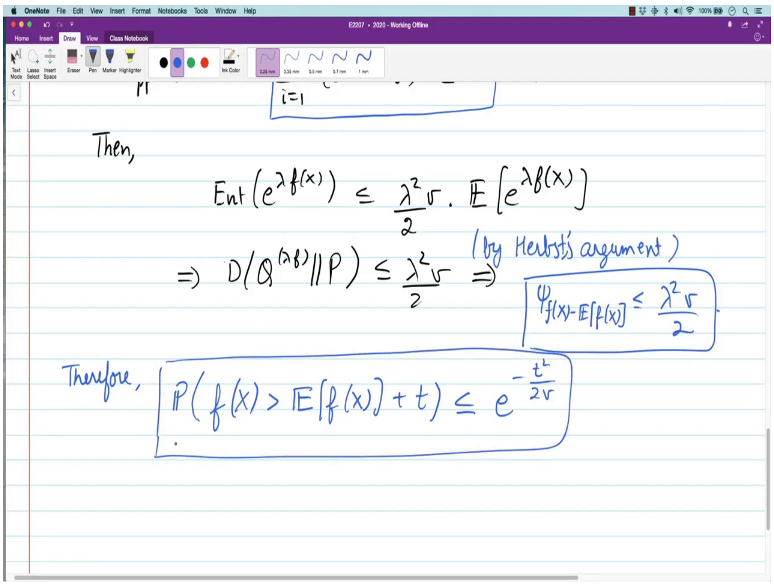
So, now suppose this is the assumption from earlier, suppose that Σ i equal to 1 to n, this v - Z - Z i prime ^{2} is ≤ v, 2 v almost sure ok then. So, this is our assumption that is what we also had at this point. So, just to refresh your memory you can go back and check. So, again here again we were assuming the assumption was that we were deriving this kind of concentration bound.

(Refer Slide Time: 15:51)



So, this guy here is ≤ v. So, now, we dont have this + sign. So, that is from symmetry that is 2 v. So, this was ≤ v. So, this is less than 2 v and we got this kind of bound. Now, we are trying we will see that this bound gets improved using entropy method ok. So, this is ≤ v 2 v suppose this is true. So, this is our assumption. So, that under this assumption this previous bound here of for e for entropy of λ f x this guy is bounded / v.

(Refer Slide Time: 16:33)



Then entropy of e to the power λ f x is ≤ λ^2 / 2 v times expected value of e to the power λ f x and this implies that the ratio of these two which is the divergence between Q λ f from P is ≤ λ^2 / v in which / Herbst's argument implies that the log moment generating function of f x is ≤.

So, you can have f x - those two log moment generating functions are the same I mean the ratio you can subtract the same quantity and then nothing changes. So, you can show this is ≤ $\lambda^2 v / 2$ ok.

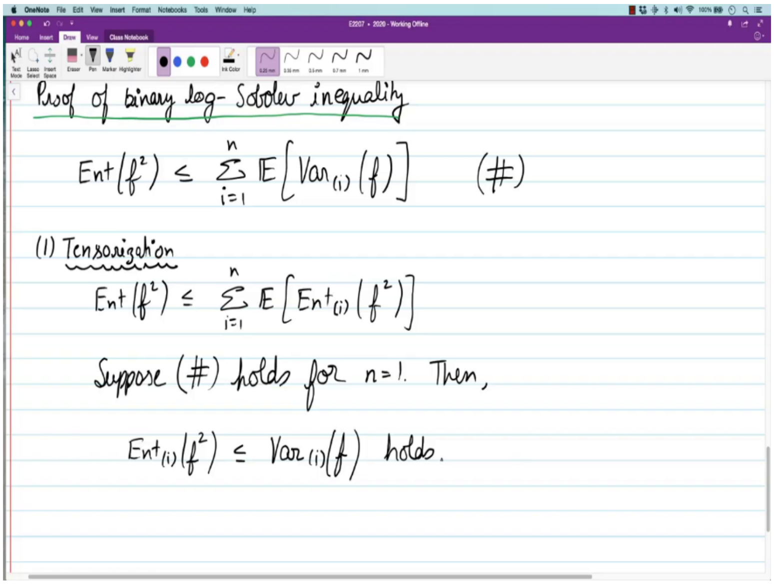
So, that is great. So, what we have shown as the sub Gaussian bound earlier we were not able to get the sub Gaussian we were we only had a sub exponential form if you are worried about this part, the only thing I am claiming is that if you subtract - expected value of x from here and here nothing changes both that sort of a homogeneous function ok. It is easier to see here, in the tilt you can subtract it from both numerator and the denominator just a constant multiply 2 with both of them.

So, f x is greater than expected value of f x + t is ≤ e to the power - t^2 / 2 v ok that is the Gaussian concentration bound that we were looking for earlier we only had t / √ v. So, that is great. So, this is we only did it for this random marker this uniform distribution on the binary hypercube, but at least this illustrates how this log Sobolev inequalities of this kind if proved in some way, note that it was not straight forward how to use it to get a concentration bound.

But, we use some elementary inequality and we were able to get this quantity here from e to the power λ f x this is a roughly Taylor series approximation ok. So, I hope I have convinced you that there is something some way to connect these inequalities to concentration bound and that is the recipe of entropy method. Now it only remains to prove this inequality the binary log Sobolev inequality, we just claim that entropy of f^2 for any function f is ≤ this guy here.

This is very similar to the this is exactly the two times the Efron's Stein variance estimate, which is sort of the sort of the energy in f or sort of the fluctuation in f this. I think this it is better to think of it as a discrete gradient norm ^{2}. So, keep Gaussian Poincare inequality in mind and this is the discrete counter part of that of the expected value of the gradient norm^{2}ok.

(Refer Slide Time: 20:08)

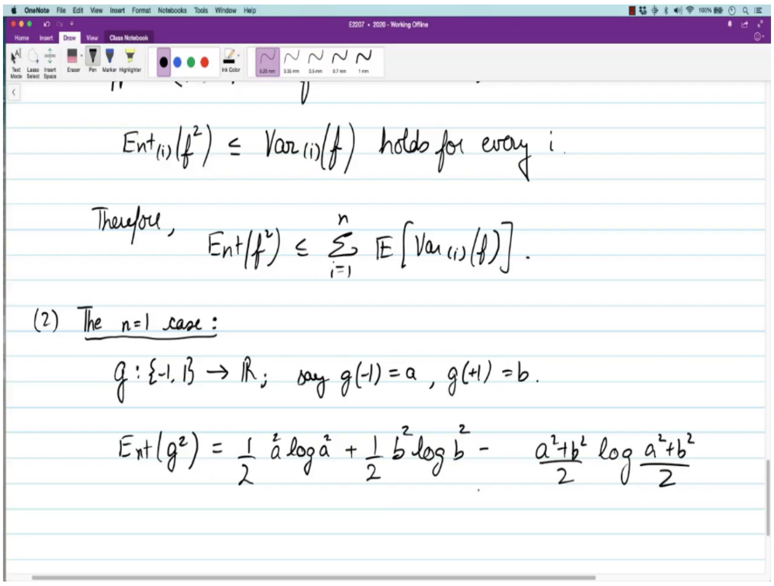


So, what remains to be done? So, we still have to prove proof of so we convinced you hopefully I convinced you this is useful the binary log Sobolev inequality is useful, but let us prove it now, binary log Sobolev inequality ok. So, how do we show this? So, first we notice that. So, we have to show that entropy of f ^{2} is ≤ Σ i = 1 to n expected value variance given everything of f ok that is what we have to show.

So, first step is tensorization step. So, / the tensorization of entropy we note that entropy of f ^{2} is ≤ Σ i equal to 1 to n expected value entropy of f given everything, but the ith quadrant that is the conditional you call it the conditional entropy of f ^{2} ok that is the first step ok. Andso, suppose hash holds for n equal to 1, then this entropy of f^2 is ≤ variance expected variance.

So, there is no nothing to take expectation over in this case because all the other guys are now fixed ok. So, for n equal to 1 there is no additional expectation, this expectation is over the remaining coordinates ok.

(Refer Slide Time: 23:11)

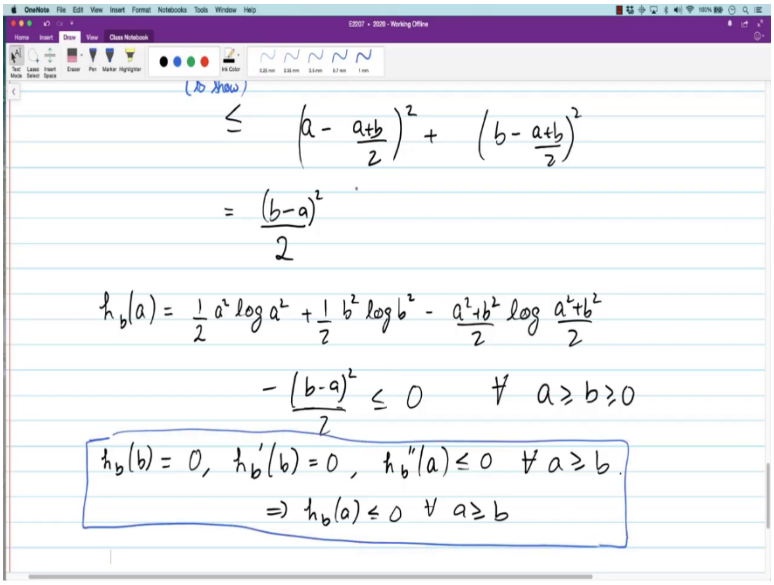


So, then this holds ok for every i. Therefore, before we get what therefore, entropy of f^2 is $\leq$ Σ i equal to 1 to n expected variance i of f ok. So, this is how we were able to use tensorization of entropy to reduce this binary logs over log Sobolev inequality to only one dimensional case.

So, how do we show this one dimensional inequality? That is the; that is the second part of the proof the n equal to 1 case this is some elementary inequality we have to show. So, what is this inequality for n equal to 1? We have to show that entropy of f^2. So, think of just a function of a single bit, now entropy at g you have function from - 1 1 to R ok. So, we can just say that say g of - 1 equals to a g of + 1 equals to b.

Then entropy of g^2 it is for n equal to 1 is equal to a^2 is expected value of g of the function. So, expected value of g log x. So, with probability half g takes the value a. So, you get a log a then there are another probability half g takes the value b. So, you get b log b - sorry g^2 sorry. So, a^2 log a^2 b^2log b^2 and then - what is the average of what is the expected value of g^2that is a square + b^2 / 2 log a^2+ b^2/ 2 that is the entropy.

(Refer Slide Time: 25:33)



And we would like to show that this guy is ≤ there is something to show it is here comma to show here. So, this guy is ≤ the variance of g. So, what is the variance of g? So, that is variance of g is it is a with it is a - a + b / 2 yeah. So, I guess you know what the answer will turn out to be, but let us just write it down half into b - a + b / 2, these are both b - a / 2 ^{2}.

So, this is just b - a /. So, the probability half this is the mean with probability half you take the value a, you get this probability you have to take the value of b. So, you get this b - a /; b - a ^{2} 4 ok. So, this / 4 looks bit wrong to me I think ok. So, maybe I am missing a factor here right. So, the binary log Sobolev inequality that yes we would like to show maybe it has a this variance is correct we may need a factor of 2 here and let us just put a factor of 2 see how it looks this is 2.

So, we can show this times this. So, that becomes / 2 yeah this one is the one I think we can show ok. So, how do we show this? So, one interesting thing here is that this inequality this inequality is an elementary inequality it just involves two numbers.

So, what we have to show is that this function h b of a which is given / say half a ^{2}log a ^{2}. So, I am thinking of it as a function of a for a fixed b that is half b ^{2} log b ^{2} - a ^{2} + b^2/ 2 log a ^{2} / b ^{2} / 2 yeah and - b - a ^{2} right.

So, what you have to show is that this function is ≤ 0 for all a and b. But we can put further restrictions, first thing you notice is that if a and b, first thing is that without loss of generality we can assume a greater than b we show it for a greater than equal to b, then / symmetry it holds for a less than b also. And the next thing is that, you can assume that both a and b are of the same sign and that sign is positive.

So, if there are the same sign then positive or negative does not matter. So, they are both of the same sign and the reason we can do that is if there are different sign then you are subtracting more and it is a smaller quantity. So, this is the this is the region in which we have to show this inequality. So, we notice that h b of b is equal to 0 and then you take this derivative and notice that its derivative also is 0 at b ok that is something you can check about h b.

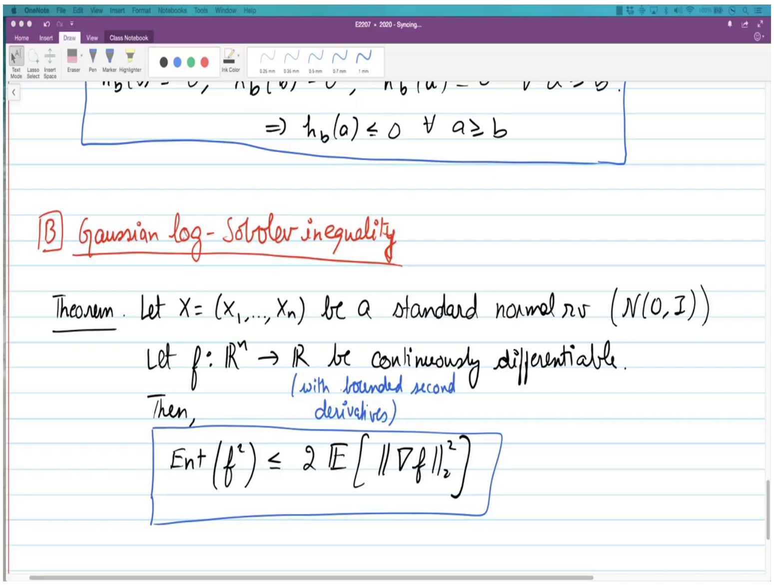
The function is 0 and its derivative is also 0 ok and then you can check its second derivative at any point and the claim is that the second derivative is actually negative ok not positive. So, this for all a greater than equal to b. So, this implies that h b of a is ≤ 0 for all a greater than equal to b.

So, I did not show this proof, but this is something this is an elementary inequality you can verify this alright. So, this is the magic of tensorization, it allows you to convert this sort of sophisticated looking inequality into an elementary inequality. So, we have seen this. So, I missed this factor of 2, but it will not change too many things here I think it is just a constant factor half.

So, when you bring in this factor of 2, I think everything we will still work out alright. So, to conclude we have shown the we have shown the binary log Sobolev inequality and we also saw how that binary log Sobolev inequality applies implies a concentration ok. Next what we look at is. So, where is this. So, in main part of this thing is this binary log Sobolev inequality.

Now, recall that earlier when we were deriving Gaussian Poincare inequality, we actually derived it using a similar poincare kind of inequality on the Boolean hypercube ok. And that is something we will do again.

(Refer Slide Time: 31:14)



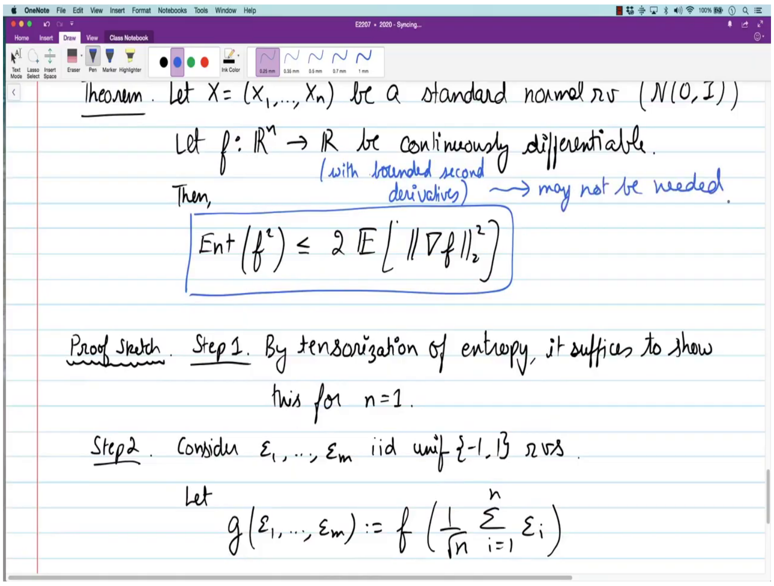
So, since we have binary log Sobolev inequality we can now use it to obtain the Gaussian log Sobolev inequality. How do we do this? So, to get the Gaussian log Sobolev inequality from the from its binary counterpart, we proceed as before. So, let ok. So, let it is the claim of the Gaussian log Sobolev inequality, Gaussian log Sobolev inequality says that be the standard normal random variable. What is that? That is N 0 I independent components all with unit variance.

So, let this be a standard Gaussian random variable and let f be any function from this R n to R such that it is continuously differentiable just like we had in Gaussian log continuously differentiable ok. Then what we can show? And we can show this / very similar steps as before maybe we need more assumptions of some boundedness of second derivative continuously differentiable with bounded second derivative maybe that is the check something to check.

That is the condition we used when we derived the Gaussian Poincare inequality. I think that may still be needed because we need a Taylor series approximation in the middle, but then the claim is that entropy of f^2 is ≤ 2 times expected value of gradient norm ^{2}ok.

This is very similar to the Gaussian Poincare inequality except that variance has been replaced / entropy ok that is the that is the inequality this is the Gaussian log Sobolev inequality. That entropy of f is bounded above / 2 times expected value of gradient of f ^{2} gradient 2 norm of gradient of f^2ok.

(Refer Slide Time: 35:05)



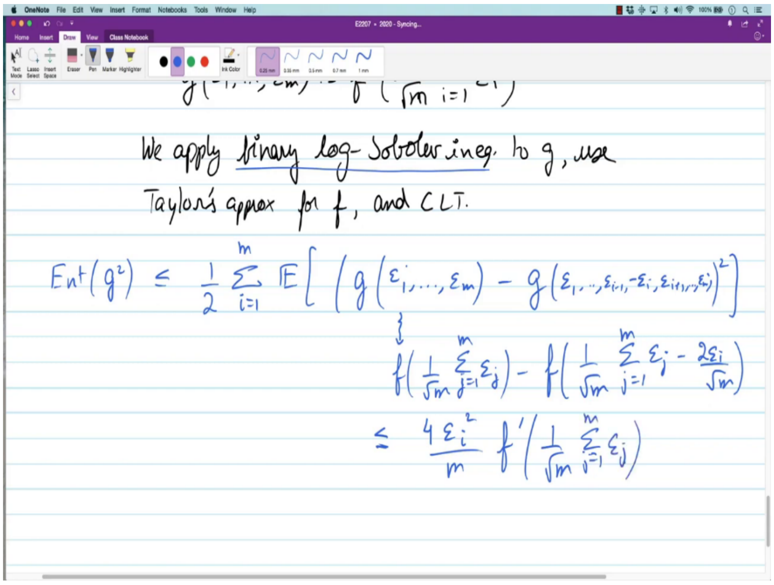
So, how do we show this? The proof is very similar to what we did earlier, I will just gave a proof sketch. So, step 1 is to first notice that / tensorization of entropy, it suffices to show this for n equal to 1. So, one dimensional case ok because this also becomes sum of derivatives and now step 2 is we consider ∈ 1 to ∈ n.

So, now we only are showing it for n equal to 1. So, we consider ∈ 1 to ∈ n as iid or maybe ∈ 1 to ∈ m, but m is not a good idea, m iid random marker random variables and let g of 1 or g of Σ∈ 1 to ∈ m we defined as f of 1 / √ n Σ i equal to 1 to n ∈ i ok.

And we apply binary log Sobolev inequality to g, use Taylor's approximation for f that is why we require this boundedness of second derivative I think that we will be needed or at least bound uniform boundedness of the first derivative will be needed it is approximation or f yeah.

I think in this case because yeah in this case we do not need the boundedness of the second derivative, in this case it is enough to have bounded first in this case I think we are fine. So, let us I will maybe elaborate a little bit may not be needed may not be needed, let us see there is approximation for f and CLT at the end.

(Refer Slide Time: 37:53)



So, let us try to do this. So, first step is / this binary log Sobolev inequality sorry entropy of g ^{2} viewed as a function of this m random variables is ≤ half Σ i equal to 1 to m expected value.

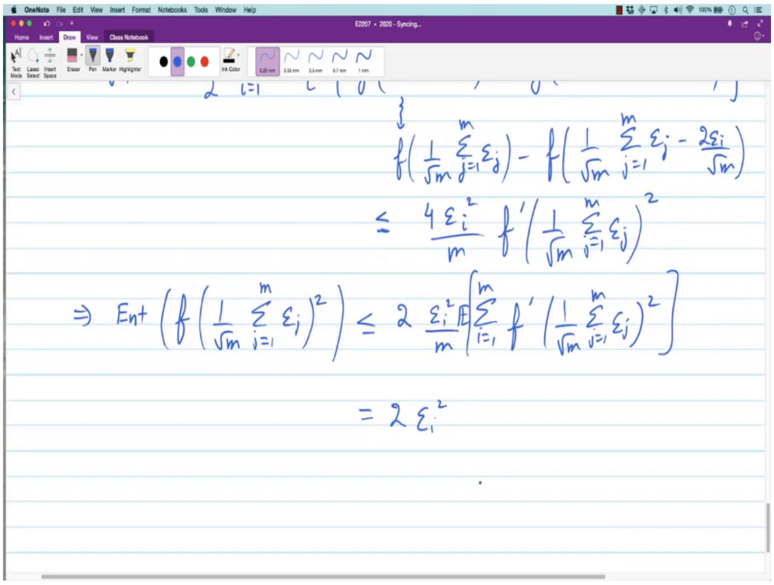
Now, since this is a uniform distribution we can write a slightly different form of the variance that is something you can verify we can write it as g of ∈ 1 to ∈ m - g of ∈ 1 to i - 1. So, these guys will make a same ∈ 1 to ∈ i - 1 you flip the sign of the ith guy and then ∈ i + 1 to ∈ m^2 ok.

So, this quantity this form that I am putting down here applies only because we have random marker distribution uniform distribution. In fact, we only proved binary log Sobolev for this distribution and so, that is the reason. So, we can expand it we can do terraces approximation around this guy.

So, if you look at these terms. So, this term here is f of 1 / √ m Σ i equal to 1 to m ∈ i, when you are proving when we were proving Gaussian Poincare inequality we did not have this form ok there was some deviation from both of them.

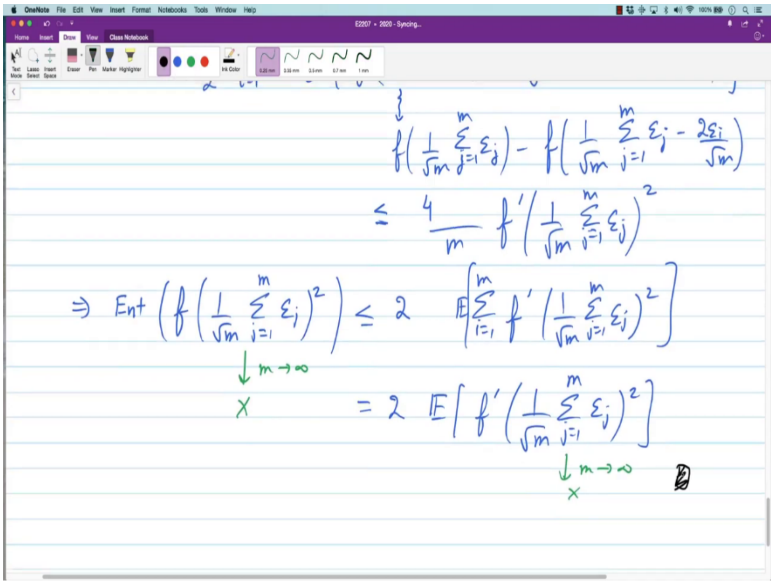
And therefore, we had to assume second order boundedness, but here I think we do not need to assume that. So, this is f of $1 / \sqrt{m}$, we do not have to assume that the second order derivative is probably $- 2 \in i / \sqrt{m}$. So, when we do Taylor series approximation this is $\leq 4 \in i^2 / m$. So, this part the derivative here f prime $1 / \sqrt{m}$ j equal to 1 to m $\in j$ ok great ok.

(Refer Slide Time: 40:39)



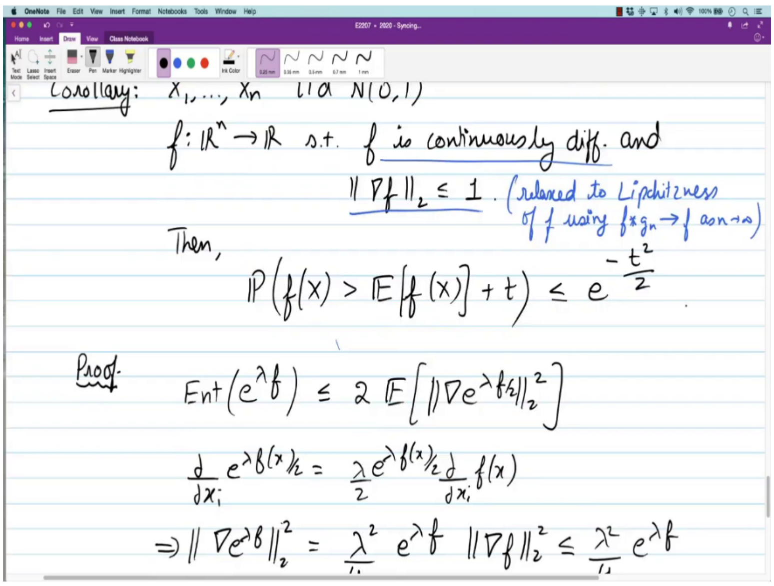
So, this guy this implies that entropy of f of $1 / \sqrt{m}$, I am writing it in this form. So, that you understand what are the random variables we are taking average over is $\leq$ half. So, this 4 comes out. So, you get this also get square. So, this 4 comes out.

So, you get $2 \in i^2 / m \, \Sigma$ i equal to 1 to m f prime $1 / \sqrt{m}$ j equal to 1 to m e j^2expected value of that. Now there is no dependence on m on i here. So, this is just exactly equal to $2 \in i^2$ (Refer Slide Time: 42:08).



So, this is just 2 expected value of f prime 1 / √ m Σ j equal to 1 to m ∈ j^2. And now / central limit theorem because f prime is continuous function and this is also continuous function we can take limit of m going to infinity and so, in that limit this goes to X standard Gaussian and this goes to X ok and this is where again we use that f is continuous and its derivative is also continuous alright. So, we get this form and then we sum it over all coordinates and we get that result. So, that is the proof of the Gaussian log Sobolev inequality ok.

(Refer Slide Time: 42:57)



And then as a consequence as a corollary what we get is the following. So, again X 1 to X n iid let us say N 0, 1 and then f is a function from R n to R such that f is continuously differentiable and gradient of f let us say norm of gradient of f is ≤ 1. Then you can prove a concentration form, then probability that f x is greater than expected value of x + t is ≤ e^{λ} - t^2ok.

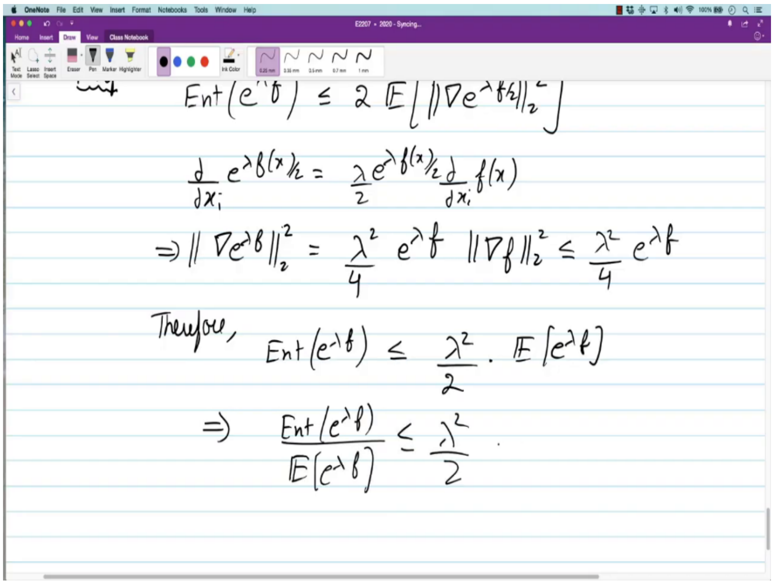
So, let us see. I think that is the that the claim we can show maybe a constant is, but let us try to do this proof. So, proof. So, how do we show this? So, by the way this condition here continuously differentiable this I think it can be relaxed to Lipchitzness of f using some Kernel to take a Gaussian Kernel and convolve f with Gaussian Kernel and you check that this goes to f as n goes to infinity.

So, basically you start with the lipchitzness function f and then sbyow that for a lipchitzness function f this guy here becomes nice and contextually differentiable and this guy will have its gradients bounded / 1 and so, you apply this bound for this function and take the limit.

Some approximation like that can be done, but let us just show it for this case that is the main result here ok. So, how do we show this? So, since f is continuously differentiable we can apply Gaussian log Sobolev inequality to f and by applying Gaussian log Sobolev inequality to f we can see.

So, let us look at the entropy of e^{λ} f ok. Since f is continuously differentiable, e^{λ} f is also continuously differentiable it is ≤ 2 times expected value norm gradient of e to the power λ f ^{2} So, what is the derivative of e^{λ} f x that is λ e^{λ} f x times right.

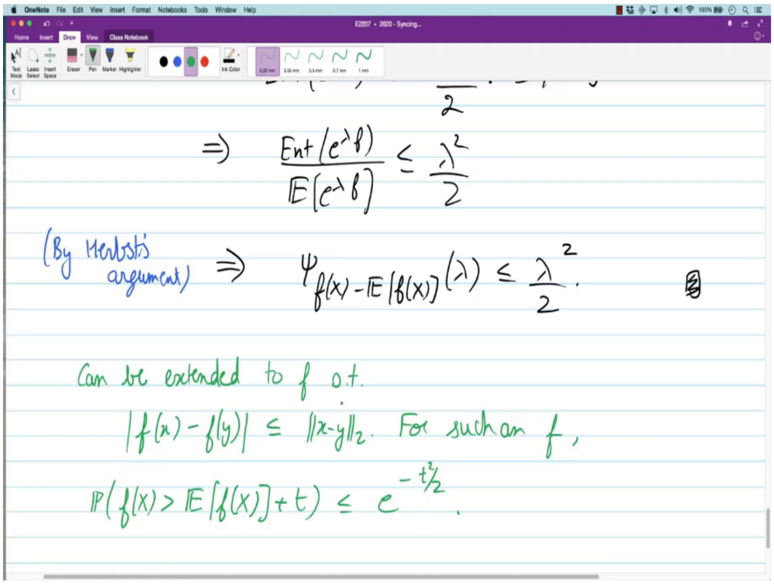
(Refer Slide Time: 46:35)



So, what we notice is that the gradient of e to the power λ f norm ^{2}is equal to λ^2 e^λ f this should be / 2 because this is entropy of f^2 yeah. So, this / 2 / 2 / 2 that is λ^2/ 4 e^λ f because you are squared and then the norm of gradient of f ^{2} ok and so, this guy is $\leq \lambda^2 / 4$ e^λ f ok.

Therefore, entropy of e to the power λ f is $\leq \lambda^2$/ 2 times expected value of e to the power λ f and this implies that entropy to the power λ f / expected value of e to the power λ f is $\leq \lambda^2 / 2$ which / Herbst's argument this guy here implies that of the log moment generating function of in fact, f x - e of f x you can always multiply and divide / e to the power - expected value of x this at λ is $\leq \lambda^2 / 2$ and that gives you a Gaussian concentration bound.

(Refer Slide Time: 48:28)

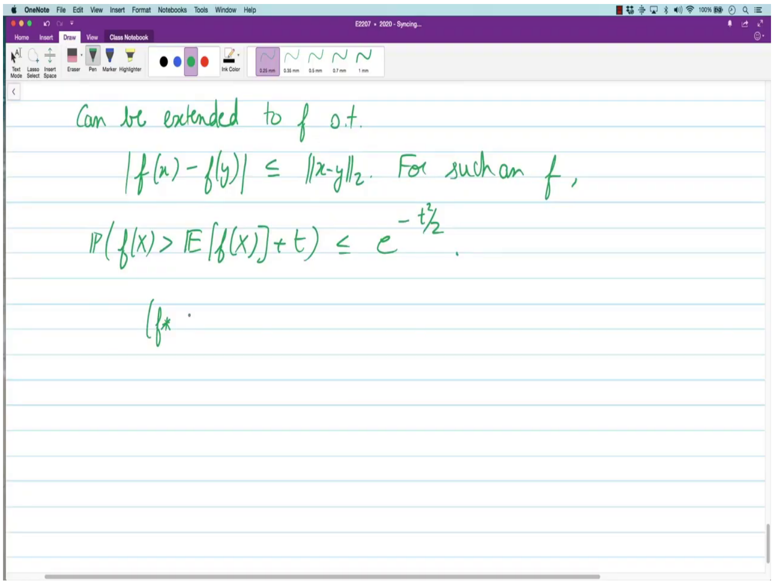


So, this is sub Gaussian with parameter 1 actually we just showed it for one side. So, that is why one sided bound. So, you should have 2 here ok and if you have v here you will get a v here. So, that bound the Lipchitz constant become the sub Gaussianity parameter that is the result here it is a very interesting bound.

So, it says that if you have a continuously differentiable function with gradient norm bounded / 1, then it has a Gaussian concentration bound with variance parameter 1 and can be extended to as I said / using some approximation f x to f such that f x - f y is ≤ norm x - y.

So, if you have if you have such an f we have probability f x for such an f probability f x is greater than expected value of f x + t we assume that this exists ≤ e to the power - t^2 / 2 ok. So, just as I said you can approximate this Lipchitz functions / a sequence of contextual differentiable functions with bounded gradient norms and then apply this bound to those functions.*Type equation here.*

(Refer Slide Time: 50:47)



In fact, the sequence that I was talking about is just a convolution of as I said f and the Gaussian kernel with different means how about different variances alright. So, that is all I wanted to say in this lecture to summarize we saw a binary log Sobolev inequality and extended it to obtain a Gaussian log Sobolev inequality which is very similar to the Gaussian Poincare inequality except that variance gets replaced with entropy of f^2 and with that inequality we were able to using Herbst's argument show a concentration bound for Lipchitz functions of Gaussian random variables.

So, Lipchitz functions of Gaussian random variables are have a sub Gaussian tail bound one sided sub Gaussian tail bound at least with actually Lipchitzness is a symmetric property.

So, you can / symmetry also show the other direction. So, they are indeed they have sub Guassian tail bounds in both sides with sub Gaussainity parameter 1 or Lipchitz constant if the Lipchitz constant that is the claim alright that is all I wanted to say about say in this lecture.

In the next week, I will give two more lectures final lectures on this entropy method where I will use entropy method to derive the where I will extend now to beyond just this Gaussian random variable or binary random variable to arbitrary random variables and establish the so, called modified log Sobolev inequality which will give sub Gaussain which will give similar bound for general random variables and not necessarily just smooth functions or Lipchitz functions of Gaussian random variables that because that is what we will see in the next lecture. See you next week.