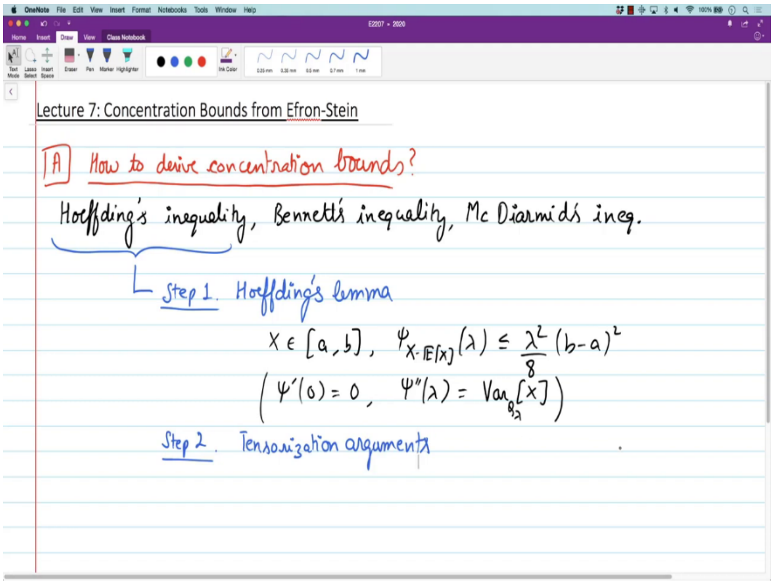
Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 08
Tail bounds using the Effron-Stein inequality

(Refer Slide Time: 00:21)



Hi. In the previous week, starting from lecture number 5, we saw this Efron-Stein bound which controls the variance of random variables. It allows us to divide the variance of a random variable which is a function of multiple independent random variables into variances associated with individual components.

And that bound can be viewed as some bound for controlling the variance itself, but it has other implications. For in for example, last week we saw the Gaussian Poincare inequality where this Efron-Stein bound was used to complete the so called tensorization argument. Namely, it was used to boost the argument for n = 1 case, one-dimensional functions, functions with one-dimensional input to the general n case namely, functions with n dimensional input.

And today in this lecture, what we will do is we will use Efron-Stein for a similar tensorization argument, but this time towards deriving concentration bounds for random variables. So, we have already seen several concentration bounds and before I proceed with what I want to say for Efron-Stein inequality, let me quickly review the general recipe, sort of a general recipe for deriving concentration bounds which we have been following till now.

So, how do we derive, how to derive concentration bounds ok? So, till now, we have seen the Hoeffding bound, Hoeffding's inequality and we have seen several others, we have seen Bennett's inequality and then, we have seen McDiarmid's inequality ok. In fact, McDiarmid's inequality the proof it does not really require the random variables to be independent, it requires them to be it does not require any assumption about the random variable as such.

It requires that the that we have a function of that the function satisfies certain bounded difference property and then, we use the multiplicative family formed by conditioning and we derive some inequality, but if you look at the so, this is slightly different proof, slightly different not too far from the recipe that will describe, but these first two bounds basically follow a general recipe.

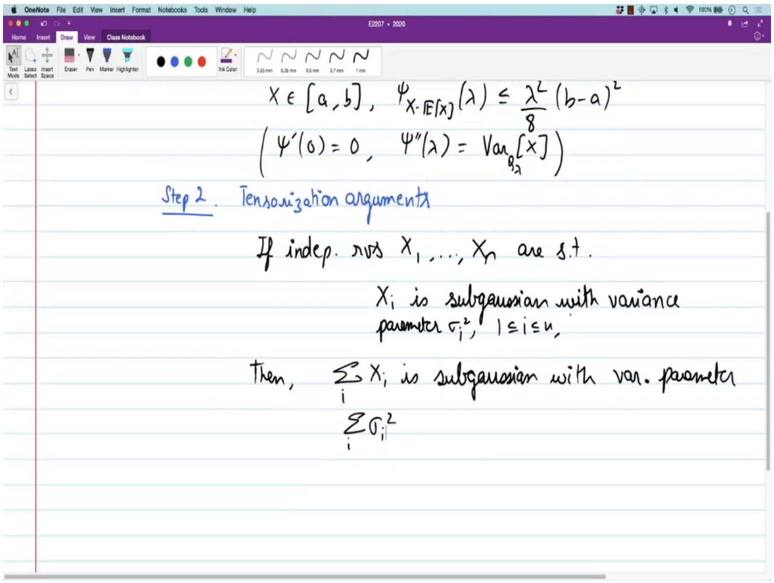
So, let us look at what we did just list this recall what we did for deriving Hoeffding bound. So, there were two steps involved in deriving Hoeffding's bound, step 1 was what we called Hoeffding's lemma.

So, here what we saw was that if you have a random variable X that takes values in a and b, then its log moment generating function ok let us look at the centralized log moment generating function. So, Ψ of X - expected value of X of λ we saw is $\leq \lambda^2 / 8$ b - a ^{2} ok that is what we showed.

And in fact in showing this, we you can just assume X is 0 mean here, but in showing this, what we noticed was that this function ∅; this function Ψ the log moment generating function its derivative at 0 is 0, I am just omitting the subscript here because it is fixed throughout. And its second derivative at λ can be written as a variance of X under some appropriate measure, some other thing some let us call it Q λ something depending on λ ok.

And we know that variance of bounded random variables are bounded and we use that there are other proofs of this as well, but we use sub we somehow examined some analytical properties of this function and we got a bound on this and the great thing about this bound is that if you have it for one-dimension then it extends to n dimension in some way.

(Refer Slide Time: 05:49)



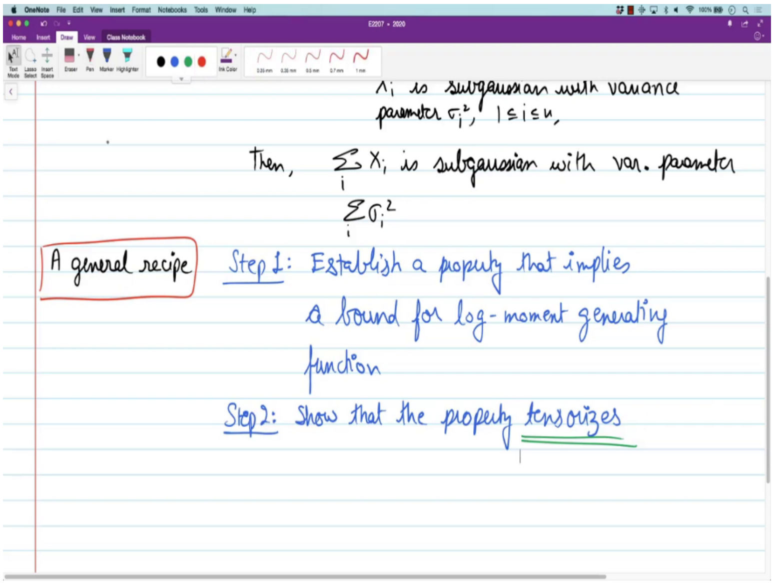
So, step 2 was what we can called a tensorization argument which is the argument tensorization arguments which is the arguments we made to extend this bound to general n dimensional random variables. In the case of Hoeffding bound; in the case of Hoeffding's inequality, what we did for tensorization was use sub additivity of sub Gaussianity parameter, the variance parameter so, here is the claim we were using.

If independent random variables X 1 to X n are such that X i is σ i ^{2} ok sorry, I think the language we have been using is sub-Gaussian with variance parameter σ i ^{2} and this is true for i between 1 and n, then Σ i X i is sub-Gaussian with variance parameter Σ over i σ i ^{2} ok. This is some property of sub-Gaussian random variables.

So, this argument allowed us to extend the bound that we have for one-dimension to bound that we have for n dimension, we only have to add this b i - a i is now ok. So, all the proofs that we have seen even Bennett's inequality has this two-step approach. First, we should show some bound for the log moment generating function in some way and the way that bound is established the proper for instance it is a sub-Gaussianity bound here, that property can be tensorized ok, this the second part is some tensorization argument.

So, if you establish a sub-Gaussian boundary tensorizes, if you establish a sub (Refer Time: 07:38) bound or sub-exponential bound or sub-gamma bound, it again tensorizes. So, I think for Bennett's inequality we saw a sub-gamma bound with appropriate parameter and that bound also tensorizes. This is a general two-step approach that we have been showing.

(Refer Slide Time: 08:00)

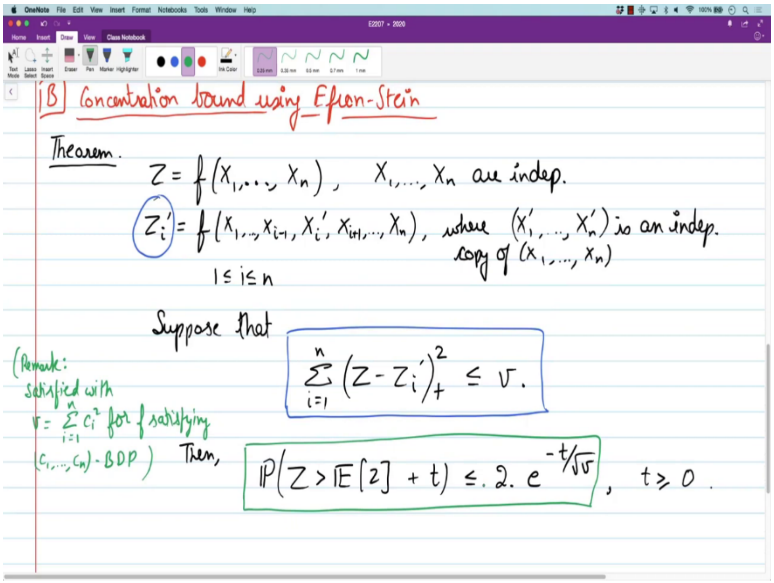


So, to summarize the general recipe that we have been following sorry for deriving concentration bounds is the following: the step 1, establish a property that implies a bound for the log moment generating function ok. So, for example, for Hoeffding's inequality that property was boundedness property of the random variable.

Step 2, show that the property tensorizes ok and how do we show that this tensorization? Yeah, that is not obvious. In fact, when you are trying to identify this property, you will like look for properties which tensorize and that is how these proofs are completed.

So, I just want to highlight this general recipe because throughout the course, we will essentially follow this general recipe and what we will see is a lot of very cleverly crafted properties so that they tensorize well right that will be a general recipe. So, with this general recipe in our mind, we can now what I will now present is how you can use the Efron-Stein inequality to establish a concentration bound ok.

(Refer Slide Time: 09:47)



This is the main topic of this lecture; concentration bound using Efron-Stein. So, we will have to do two things as we said, we first have to identify this property and then, we have to see that the bound tensorizes. The tensorization part we will see, we will follow from Efron-Stein inequality ok, but before that, let us look at the bound that is maybe yeah, the I will state the theorem and then, I will show how it can be proved using Efron-Stein and how it can be proved using this recipe.

But if you want to deconstruct the proof, first you have to identify the property that is the main observation and then, you tensorize and then, this theorem comes out, verifying is easy, but coming up with this theorem is very difficult ok. So, what is the theorem? Well, verify verifying is not so easy itself, but at least it can be structured ok.

So, once again you are given this random variable Z which is a function of X 1 to X n and these guys are X 1 to X n are independent and what we assume about these guys is some sort of bounded difference property, but we will write it in the compact way. So, let us the Z i prime be f of X 1 to X i - 1 and then, you replace X i with independent copy and X i + 1 to X n where X 1 prime to X n prime is an independent copy of X 1 to X n ok.

So, you define these guys and you define this for every i between 1 and n ok that is what you define yeah. So, if you can recognize these guys from our Efron-Stein inequality where we had a similar independent copies coming in.

So, what we assume is that suppose that Σ i = 1 to n Z - Z i prime ^{2} and let us just take the positive part of it. Remember from symmetry, we can have this term showing on the right side of Efron-Stein inequality, this guy is $\leq$v. So, this is sort of the proxy for the variance that we have. So, this is the assumption.

So, aside remark here satisfied with v = C i ^{2} for f that satisfies; for f satisfying the bounded the C 1 to C n boundary difference property ok. So, if f satisfies the C 1 to C n boundary difference property, then this property this assumption here holds with v = Σ C i ^{2}.

This is essentially equivalent to this point nothing more ok, this is essentially equivalent it is not more general or less general. So, we just, but we just keep it in this compact way and this will help us later on as well, we will use this compact notation here to summarize this boundary difference property.

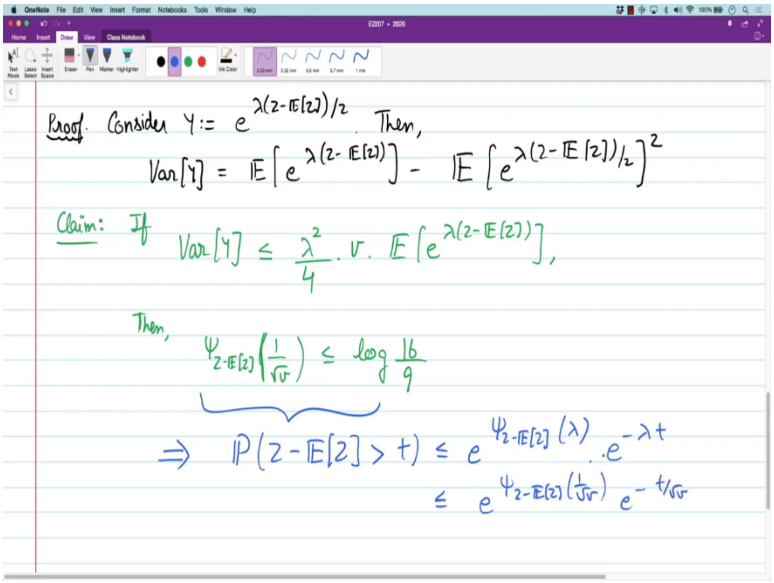
So, suppose this holds, then probability that Z exceed expected value of Z + some t too many +es and t's here + some t is $\leq$2 $\times$ e to the power - t / $\sqrt{}$ v yeah. So, this is a slightly different bound from what we had seen earlier. Earlier, using McDiarmid's inequality, we had seen a bound which looks like I think t ^{2} / Σ C i ^{2} so, t ^{2} /v; t ^{2} / 2v something like that.

Now, we see t / $\sqrt{}$ v here and depending on where the t; where t is, this can be better or worse, but in general the McDiarmid bound is a better bound than this one, but we will give a different proof of this bound, we will use we will use Efron-Stein inequality to come up with this one and we will illustrate the general recipe we pointed out ok.

So, what is so, this is interesting. We bound the variance and somehow from that, we can bound the log moment generating function, this is sort of a bound-on log moment generating function and the question is how do we translate this variance bound to a log moment generating function bound? With roughly the same variance parameter.

By the way note that this is only e to the power - t / √ v. So, it is the so called sub-exponential bound rather than sub-Gaussian tail bound, but that is ok, we will we are getting a weaker bound, but the purpose here is to illustrate the recipe ok.

(Refer Slide Time: 15:52)



So, let us try to prove this. The proof follows our general recipe which is been outlined above where we start by defining a property which will leave the bound for log moment generating function and that property will be such that it tensorizes.

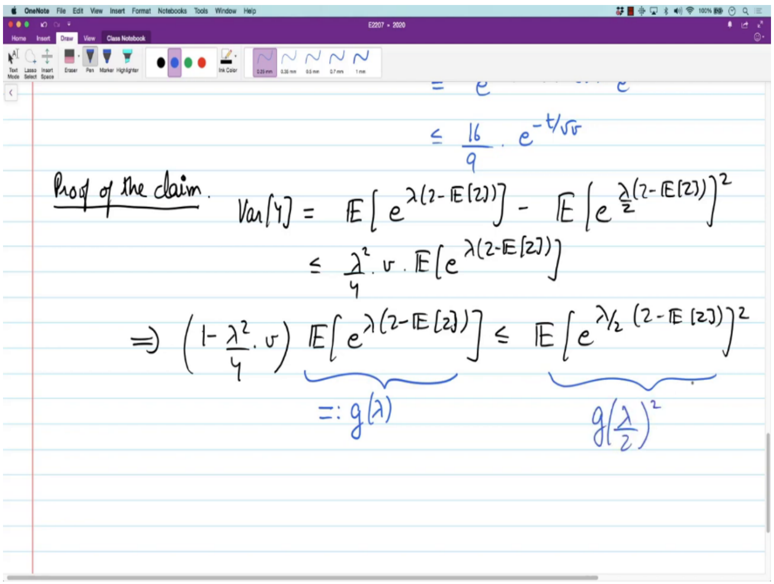
So, we consider this random variable Y defined as the one which we use for defining log moment generating function just normalized / 2 that is just for convenience, we will see how it plays a role. Then, variance of Y is = the expected value of this guy, this is the moment generating function ok square of this so, that is just this guy here - the expected value of this guy ^{2} ok that is what variance of Y is ok.

And the property is the property we want to prove, this is sort of a claim I am putting it down here. If variance of Y is ≤λ^2 / some constant, we will get 4 v and then into let us say expected value of e to the power λ. Suppose you can show that then we can derive a nice bound for the log moment generating function, then is ≤maybe we write simpler implication of this bound, then is ≤a constant that is the claim.

So, if you have this bound, then you get a bound like this. By the way so why do we want the bound like this? So, this in particular implies that if you look at the probability that c - E Z is >t just using Chernoff boundary and this guy is ≤we have this bound for log moment generating function.

So this is ≤expected sorry e to the power λ × e to the power - λ t and if we choose λ = 1 /√ v here; then if we choose λ = 1 /√ v here, then what do we will get? This is ≤e to the power, but this guy is a constant.

(Refer Slide Time: 20:11)



And this guy is ≤16 / 9 e to the power - t / v ok. 16 / 9 this is this log is the log that we have here yeah so, that is correct so, we get this bound. So, any constant here you will get the same constant λ and you get the same e to the power - t× that constant bound. This is also called the sub-exponential bound. So, its (Refer Time: 20:35) find a constant λ in this case which is 1 / √ v for which the log moment generating function is constant.

If you have that bound, then my Chernoff bound you will get a sub-exponential bound with the same constant. So, this part is easy. So, it will give our desired result, I put down 2 there, you can get 16 / 9, but that constant is not important for us ok. So, this property is essentially what we want to show. So, if you have this property, then you get the desired concentration that is the claim.

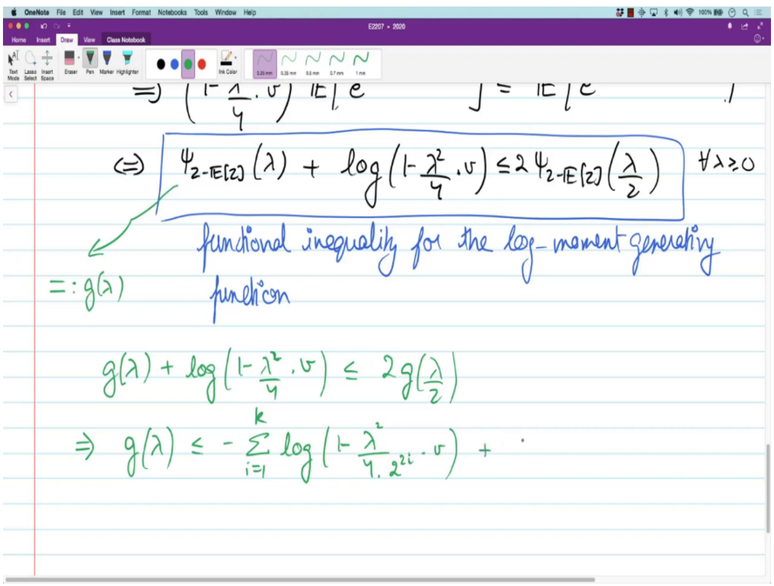
So, first part that how do we; how do we show this property where v remember is the bound here that is the first part, that is the first that is the first question that may come to your mind, we will come back to this question later. In fact, this is where we will use tensorization. We will show that this property tensorizes because of Efron-Stein inequality, but before we show that, let us first convince ourselves that indeed this property is the one we want to show ok.

So, what is the proof of the claim? So, proof of the claim. So, to prove this claim, what we note is that this variance of Y is = expected value of e to the power λ this term - expected value of e to the power λ/ 2 whole square, square outside ok this is just for variance of this and this we are claiming is $\leq \lambda^2/4 \times v \times$ the same function here ok.

If you have this, then you can take this on the one side, take this on this side, this guy is $\leq$ (expected value $e^{t\lambda}$ / 2) ^{2} ok that is what that property implies.

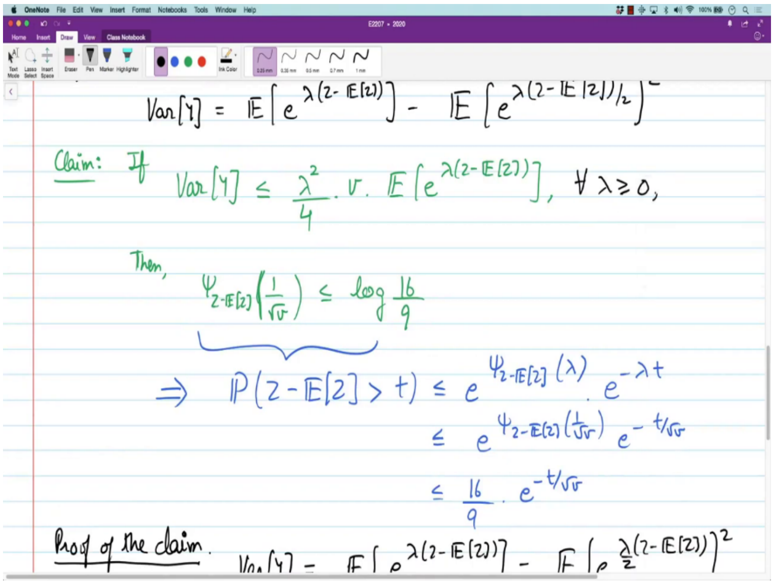
We define this guy here to be something let us call this maybe f λ maybe I should call it just g λ sure I have used this f, f is already taken f λ and this guy if you notice is g λ/ 2 /actually, it is better to first take the log and then, define this function ok. So, let us take a log here.

(Refer Slide Time: 23:54)



So, this is if and only if the log moment generating function + log of 1 - $\lambda^2/4$ v is $\leq$the same log moment generating function at λ/ 2, but there is a factor of 2 outside ok. So, that is what that the property that we are assuming here, this if part here, this property is equivalent to having this condition here for the log moment generating function. So, this is a functional inequality for the log moment generating function ok.

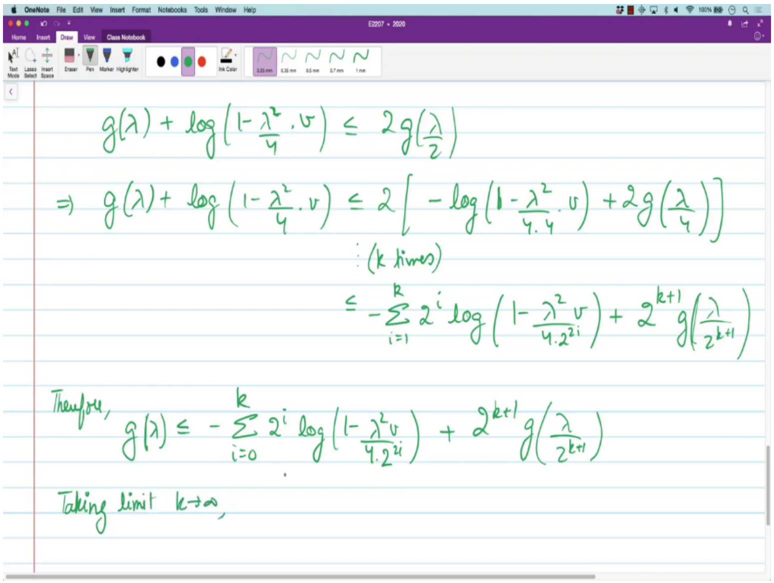
(Refer Slide Time: 24:56)



And then, this property by the way we are assuming, this one holds for all $\lambda > 0$ ok. So, this one here is for all $\lambda > 0$; or ≥ 0 ok. So, once you have this property, what we are claiming is this sort of implies some bound for this function itself. So, let us just define this guy here, this guy here for convenience we define it as g λ.

So, suppose you have g λ is + some log 1 - λ ^{2}/ 4 v, v some constant is $\leq$2g λ/ 2 from that can we get a bound for g λ. So, we keep on repeating this implies if you repeat it k $\times$ this implies g λ is $\leq$- Σ i = 1 to k log of 1 - λ ^{2} / 4 into 2 to the power 2 i $\times$ v + ok so, I remove this let me just try to think of this (Refer Time: 26:42) here yeah, let us do this slowly.

(Refer Slide Time: 26:47)

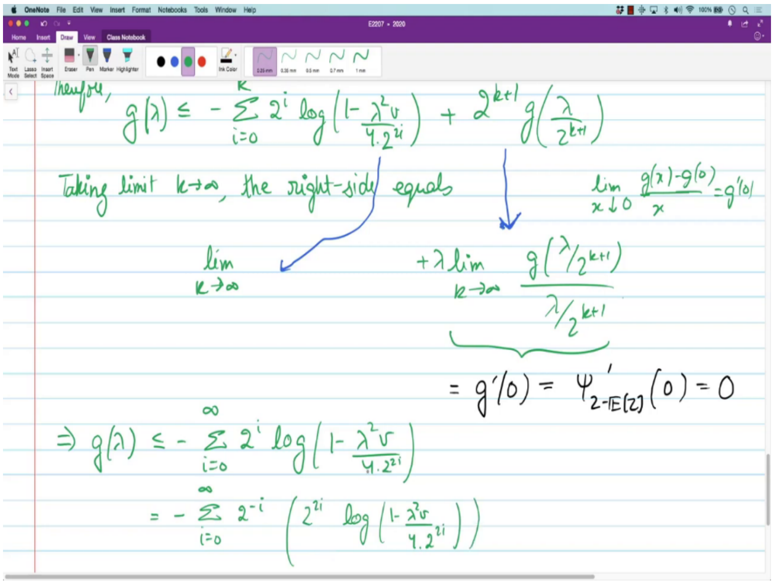


So, g λ is ≤this. So, g λ then is also ≤now, I will apply it 1 once again to this function here because it holds for all λ >0 so, ≤2 × I will substitute the same thing here g - log 1 - λ^2 / 4 into 4 because now, it is λ/ 2 × v and + 2 into g λ / 4 ok and then, you can apply it again.

Every time you apply this inequality, you get an extra factor of t; you get an extra factor of 2 with the log term and you get an extra λ / 2 here and this extra factor of 2 outside.

So, when you apply it k × , what you get is this is ≤i = 1 to k 2 to the power i log of 1 - λ^2 / 4 2 to the power 2i × v + 2 to the power k + 1 λ /g of λ / 2 to the power k + 1 ok that is the one ok. Therefore, what we have is that g λ is ≤Σ i = 0 to k 2 to the power i log of 1 - λ^2 / 4 2 to the power 2i v + 2 to the power k + 1 g λ / 2 to the power k + 1.

(Refer Slide Time: 29:52)

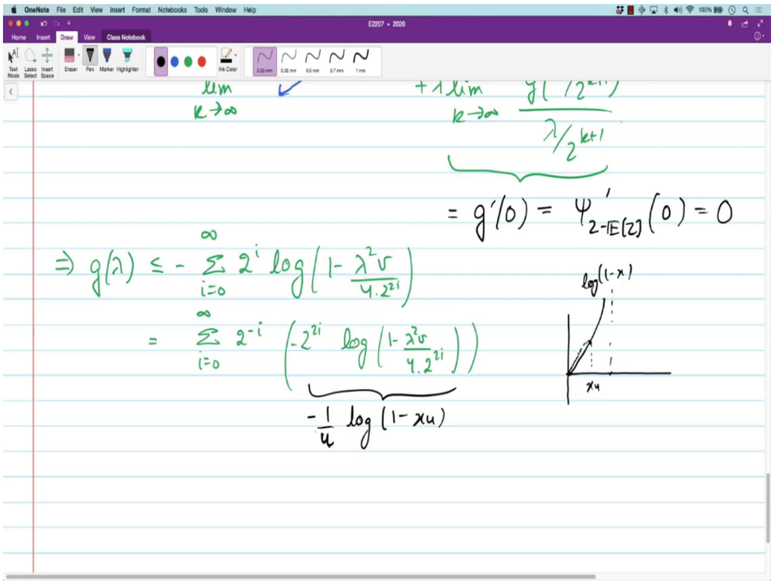


So, when we take this k to be very large as we take the limit k going to ∝ , what happens to this guy, this guy here? So, taking limit k going to ∝ , right side equals so, there are two terms, the first term I will come to it later, limit k going to ∝ of this guy here + limit k going to ∝ of the second guy. So, this second guy is 2 to the power so, I will take a λ out so, it looks like this λ / this /λ / 2 the power k + 1.

So, what is this guy? So, as k goes to ∝ , this goes to 0 from the right side and this function g so, this is g of sort of g of x - g of 0 / x in the limit as x goes to 0 from the right side. So, this guy here is just g prime of 0 ok. So, this guy here is just that g prime of 0. So, this limit here is = g prime of 0.

But remember the function g was this guy here and we are taking its derivative with respect to λ and putting it as 0 and this is a center this is a 0 mean random variable so; this guy is also 0 ok. So, this term here is 0. For what we have is the bound that we have obtained is that g λ is ≤- Σ i = 0 to ∝ 2 to the power i log of (1 - λ^2 v/ 4 $)^2$ i ok that is what you get.

(Refer Slide Time: 32:43)

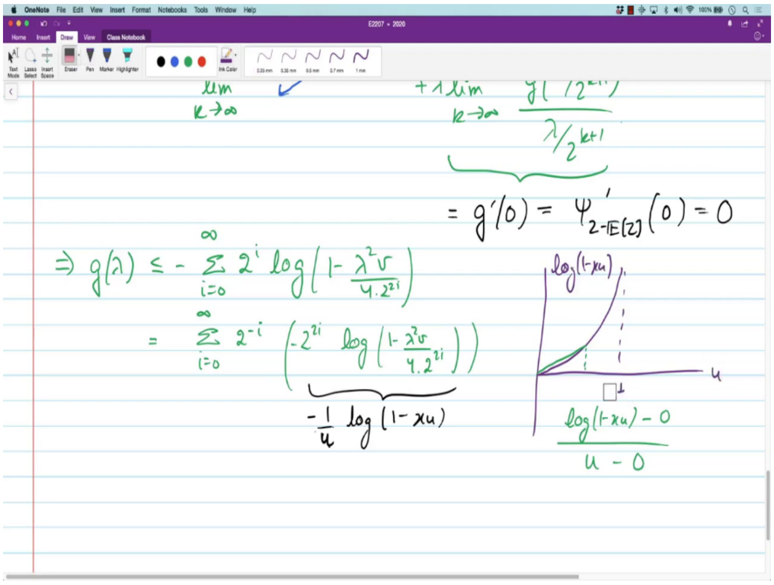


So, we can express this sum here as - Σ i = 0 to ∝ that is how you can express it as and the reason, I am doing in this way is the following. So, now, if you look at this function, here this has this form of 1 b/u log 1 - xu and where u is a number between 0 and 1.

So, let us plot 1 - log of 1 - x ok for x between 0 and 1, this log of 1 - x how does it look for x between 0 and 1. So, this guy add actually - log, there is a - sign here, I will take that - inside right. So, this I am just writing as u, u is some number between 0 and 1.

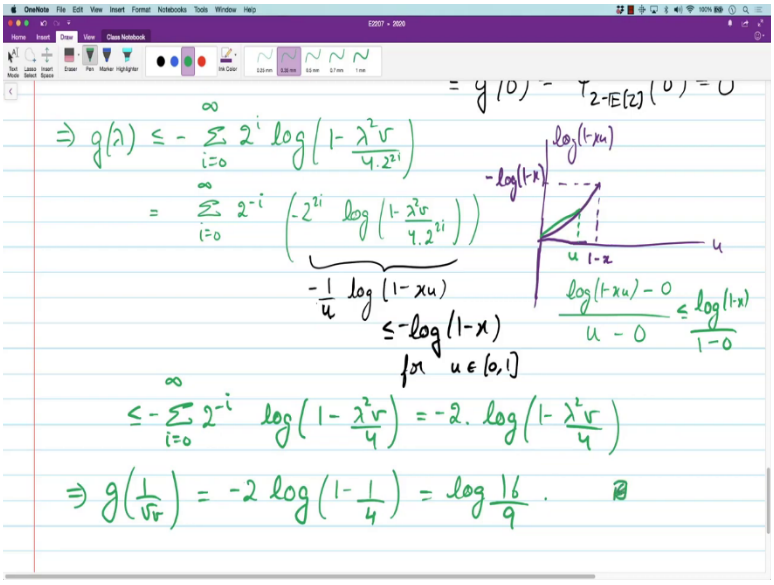
So, at x = 0, this is 0 and that x = 1, this is ∝ so, - of that so, it looks like this that is what this function looks like. So, as you take this u from this is log of 1 - x. Now, what we are doing is we are parameterizing a point xu and we are looking at this line here, we are using bigger graph and different colours.

(Refer Slide Time: 34:20)



So, we plot this function. Just want to notice a very simple inequality about this function. 1 and then, I take two points, I will take this point here xu and draw this chord, it lies above the function because as you can see and verify this is a convex function. So, if this lies above this function, then this guy here this log of 1 - xu - log of 1 - 0 which is just 0 so, - 0 / u - 0 actually, what I would like to do is I would like to plot log of 1 - xu as a function of u and so, for u between 0 and 1.

(Refer Slide Time: 35:38)



So, this function now will truncate at a particular point, let us see where it truncates. So, again it looks like this, but it stops here which is at 1 - x and at this point, its value is log of 1 - x. Now, you take any point. This is the point x, this is the point some arbitrary u and draw this line, this will also lie above it and the slope of this chord is this much and because this function is concave as you increase u; as you increase u, what happens as you increase u?

As you increase u, this chord its slope increases ok and so, this slope is ≤the slope at u = 1. So, this guy here, this function here, the inequality I have roughly argued is this is ≤= log 1 - x for all u between 0 and 1 ok that is the bound you have for this function.

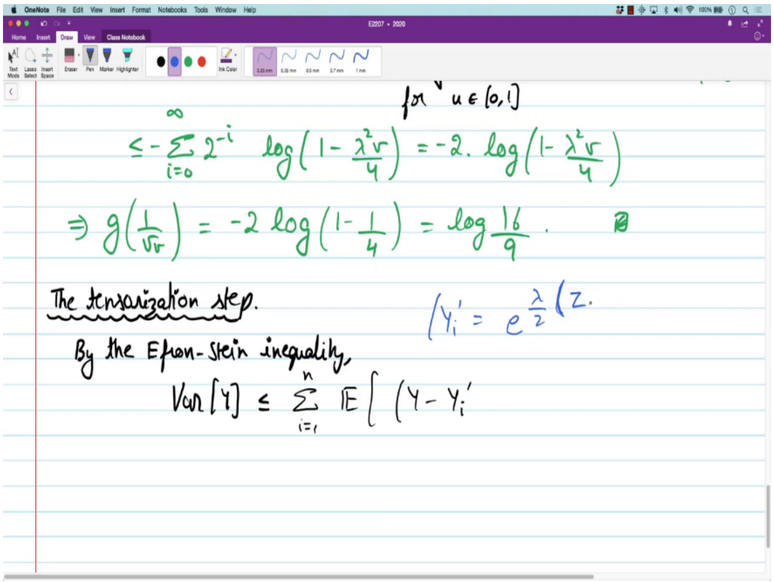
Therefore, this guy here is ≤Σ i = 0 to ∝ 2 to the power - i and what you have here is log of the - outside log of 1 - x, x for me is λ^2 v / 4 ok and this Σ here now is easy to sum, this just sums to 2. So, this is = 2 × log of 1 - λ^2 v/ 4.

So, in particular, g of 1 / √ v is = - 2 × log of 1 - 1 / 4 right. So, that is log of 16 / 9 as I said ok that was the claim. So, this is just the proof of the claim. So, let us go back to a claim ok.

So, we have this nice intermediate step that if you can show this bound, then this function is ≤log bound generating function is ≤log 16 / 9 and it uses a functional inequality in the middle ok and this is some property. So, this is like replacing the Hoeffding lemma with something else. So, this is a replacement of Hoeffding lemma.

And now comes the second part of the proof namely the tensorization step which says that; which says that this property proving this property for one-dimension is the same as proving this property for n dimensional things and that is what we will do through you we will show this property through the tensorization step will be provided through the Efron-Stein inequality.

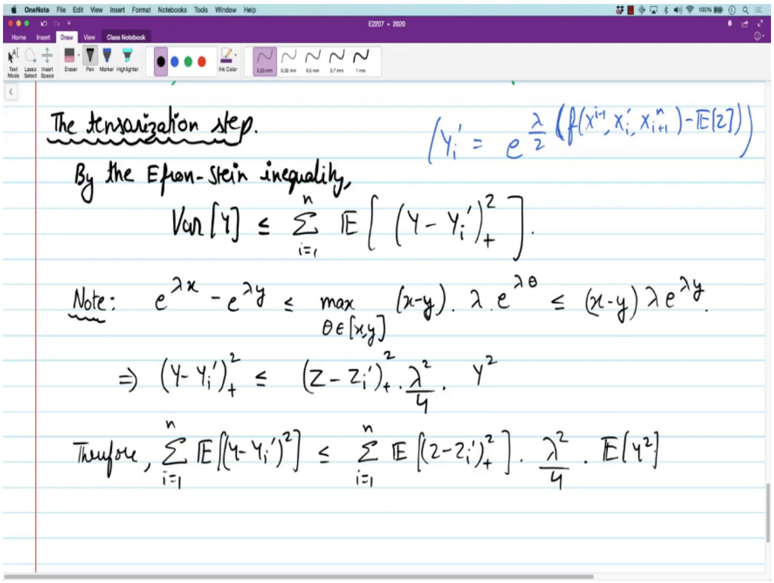
(Refer Slide Time: 38:44)



So, let me do the next step now, the tensorization step. So, by the Efron-Stein inequality, variance of Y is ≤variance of so, Y again is a function; Y again is a function of those x 1 to x n, the independent random variables.

So, we can apply that inequality to this function as well and what you get is its ≤i = 1 to n expected value of Y - Y i prime. So, now, remember that what is Y i prime? Y i prime is = e to the power λ/ 2, the function was Z - expected value of Z.

(Refer Slide Time: 40:11)



And so, this will become f of X i - 1, X i prime, X i + 1 to n - expected value of say expected value of Z yeah, that is what this quantity here is and Y - Y prime^2 and there is a factor of half and if you use symmetric property, you can just write it as the positive part.

This is just the positive part of a number ok, this is by now Efron-Stein ok. So, this is still not bringing in the v part, but at least the tensorization property holds. Now, we have to handle just a one-dimensional function here which has a similar form, you can think of this entire thing as that as that function for a fixed X 1 to X i ok.

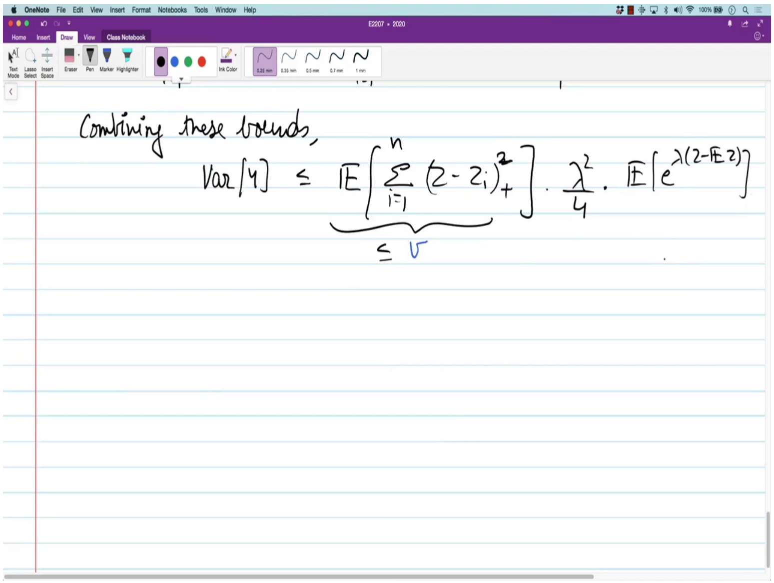
Now, we notice an elementary property it is a simple bound for e to the power λ x small x is ≤So, if you look at e to the power λ x - e to the power λ y by Taylor series approximation, this is ≤the first order Taylor series approximation max over theta in let us say x comma y, we are just assuming x is smaller than y ok.

So, max over theta in x comma y, if x is smaller than y, derivative of this guy × x - y, but what is the derivative of this guy? The derivative of this guy is λ e to the power λ theta right and again, e to the power λ theta is increasing function of theta so, this is ≤x - y into λ into e to the power λ y because I am assuming x is; I am assuming x is > y sorry y is >x right.

So we can use this inequality here and this inequality here implies that Y - Y i prime^2 is ≤so, now, I will just have x - y, this exponential part will go away Z - Z i prime +^{2} and I have a I am working with $\lambda / 2$ so, this becomes $\times \lambda^2 / 4$ and this becomes, this remains the same guy, this remains this remains Y ok. So, this is that magical inequality we were looking for this becomes Y^2 actually because I^2 this inequality into this ok.

Therefore, expected value of Y - Y i prime ^{2} is less, this is the fundamental inequality which tensorizes using Efron-Stein inequality, the fundamental inequalities for each coordinate is $\leq$ Σ i = 1 to n expected value of Z - Z i prime ^{2} $\lambda^2 / 4$ and this is just expected value of Y^2.

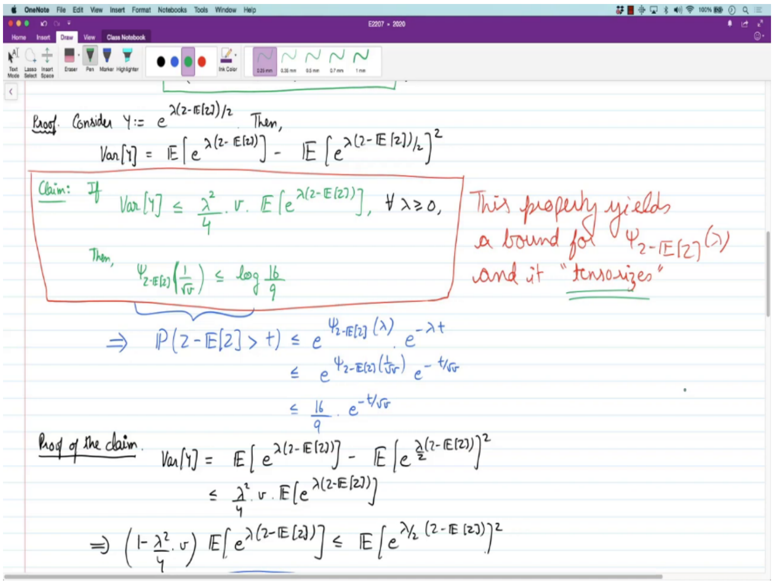
(Refer Slide Time: 44:13)



Combining both the bounds, what we get is variance of Y is $\leq$expected value, I am just taking this expectation out $\times \lambda^2 / 4$ and then, this guy this is by definition this is the function expected value of λ Z - expected value of Z ok that is and by our assumption, this guy here is $\leq$v, the variance factor.

Therefore, what we have established here is this assumption of this claim that variance of Y is $< \lambda^2 / 4 \, v \times$ this function and in establishing this, we essentially had to establish this for one-dimension part here. If this holds a one-dimension part ok, then it automatically holds for n dimension by the Efron-Stein inequality. So, we just had to show this bound for one-dimension. So, this is the tensorization step which can be shown for n dimensional.

(Refer Slide Time: 45:54)



So, this completes the proof and the important point to notice here is how we prove this, the with recipe itself. We the magic was in identifying this property, this particular property that sorry the property in the claim this one here and it is this property that we identified. This property yields a bound for this that is that was the good thing about this property and it tensorizes ok.

So, a large part of literature of concentration bounds is around identification of such properties which will yield a bound-on log moment generating function and tensorize. We have already seen that sub-Gaussianity is one such property and that is a bound which tensorizes and this version is more abstract.

Here, we ask for a function inequality and when it holds, it implies a bound for the log moment generating function and it tensorizes ok and this tensorization for this particular bound, this sub-exponential bound that we finally, attained this one here, this tensorization relied on the Efron-Stein inequality.

So, that is what we wanted to show that the purpose of introducing this part was actually to as a segue to the next part where we will talk about the entropy method where this particular property here, the variance of Y bounded in this way. This property will be replaced with some other property of the log moment generating function, a function inequality which we assume holds and when that function inequality will hold, we will get another bound for the log moment generating function and this is the so called Herbst argument and it will lead to the entropy method that is what we will do in the next lecture ok. See you in the next lecture.