[SQUEAKING]

[RUSTLING]

[CLICKING]

ANKUR MOITRA: So remember, last lecture, I introduced you to the idea of tail bounds, where we want to get a rough handle on how likely it is for a random variable to be far away from its expectation. So we talked about Markov's bound, Chebyshev's bound. And each of these used more and more information about the random variables as a way to get a better bound on the tail probabilities.

So Markov just used that the random variable was non-negative. And then we already could bound the probability that it was double its expectation. Chebyshev, we used its variance, which is a measure of how spread out the random variable is. And today, we're going to do the most important tool, the highest power tool, which is called the Chernoff bound.

So this is going to use that not only do we know properties about x's expectation and its variance, but that x itself is composed of the sum of mutually independent random variables. So we'll see that implicitly in the proof of this. We're going to be using not just the first moment, which is the mean, and the second moment, but we're going to be using all of the higher moments at once.

But the basic setup for the Chernoff bound, at least the plain vanilla version, is-- remember from last time, we have these random variables, xi, that are mutually independent. So this is critical. Chernoff bound only works when you have mutual independence, not just pairwise. It's completely false otherwise.

And these xi's, we're going to assume for simplicity that they're Bernoulli random variables. So each of the xi's has probability pi of being 1 and is otherwise 0. And just to have some notation, we'll let mu denote the expectation of x, which is the sum of all the xi's. But of course, that's the same thing as the sum of all of the pi's.

So the Chernoff bound comes in a lot of different flavors. Let me state a few different versions you might care about. First, there's what's called the upper tail, which this is the probability that the random variable is way larger than its expectation. So for all delta greater than 0, we can look at the probability that x is larger than its mean by a 1 plus delta factor.

So you should think about delta here as being something like 0.1. The chance that this happens is exponentially small. And it depends on this delta parameter, of course. But then it also depends on mu. So this is a complicated algebraic expression. But it's important to have some intuition for what's going on here.

So if I take a fair coin and I flip it, I know the probability that it's heads is 50. It's 50%. And intuitively, if I flip my coin a whole bunch of times and keep track of how many heads I get, I should expect that the empirical average of the number of heads is very quickly going to 50/50. Maybe if I flip it three times, I'll get two heads out of three. And that'll be far away from 50/50. But the larger the number of flips, well, the more it should be driven to going back to the average. That's what things like the weak law of large numbers are telling us.

And the way that this shows up is really in this mean mu because, as the number of flips n gets larger and larger, my mean is larger and larger. So the chance that I have 10% too many heads is going down exponentially with the number of flips I make. So this is the critical distinction between the Chernoff bound and things like the weak law of large numbers, is that, here, our failure probability, the chance that we're outside this range where we think it lands, is exponentially small in n, whereas for things like the weak law of large numbers, it was only inverse polynomially small in n.

So that's called the upper tail because all I care about is the probability that x is too big compared to mu. There's a corresponding lower tail. The only reason we separate these two things out is because they have a slightly different dependence on delta. So here, it really only makes sense to have a delta that's between 0 and 1.

And now, the probability that x is at most 1 minus delta times mu-- that I get to few heads instead-- is, again, something that's exponentially small. It just doesn't have this 2 plus delta in the base. So it looks like this instead. So this is the lower tail. And there's also a way that you can combine them, if you don't have to worry that much about what the dependence on delta is.

So what I claim is that when you just put these bounds together-- so for all delta between 0 and 1, you can just write it instead, the probability that x minus mu, in absolute value, is larger than delta mu-- so it's more than delta mu away from its expectation-- goes down like 2 times e to the minus mu delta squared over 3.

So Chernoff bounds aren't really just one bound. There's a whole family of bounds. The ones I'm stating here, these are for Bernoulli random variables. And we have an upper tail, a lower tail, a combined bound. In fact, there are versions of Chernoff bounds where the xi's aren't Bernoulli random variables. There are other kinds of random variables with bounded second moments.

So the way you should think about Chernoff bounds is I'm going to give you one proof of the Chernoff bound. It'll be in this restricted case of Bernoulli random variables. The same basic proof works for a lot of other exotic variants of the Chernoff bound. And when you need a different kind of Chernoff bound, like your xi's are not Bernoulli, you just look it up and quote the bound. So the key thing is really just to understand one of these bounds, how the proof works, and intuitively why the proof should extend beyond literally the assumptions I'm making. So are there any questions about it?

So this is probably the most complicated proof we've done up until this point. And it's something which we're going to spend a bit of time in recitation thinking about. Yeah?

STUDENT: Does mu have to be positive?

ANKUR MOITRA: So mu is positive here just by assumption because my xi's are non-negative. I mean, if mu were 0, it would be kind of trivial. So you have to be a bit more careful when you have a Chernoff bound with random variables which are not non-negative. Yeah, you have to shift and scale it. And you'll see through the proof that a lot of these things can be done.

It's just that it changes subtleties like what exactly happens in the exponent here. So there might be other quantities that show up. But the important thing is really the mu and the delta dependence. And then the other scalars might change in different ways. But good question.

All right. So we're going to cover this in more detail in the recitation. But I'm warning you. So this will be one of the most complicated proofs we've done up until this point. But what I'm going to try and do is I'm going to explain the high level strategy for the proof. And that way, we can then try and fill in some of the technical details, the actual estimates to get the overall strategy to work.

So what I really want to focus on is how we're going to prove this bound at a high level because, in all proofs, there are the hard parts of how you're going to fit the pieces together, and then there are the tedious parts of how do you do the literal computations to execute the different steps of your strategy.

So let's start with the main idea. And this will be a very similar trick to what happened when we proved the Chebyshev bound. So let's say I have my random variable x. And for a that I'm going to set later, because I'm going to set it to be equal to this 1 plus delta mu or 1 minus delta, that kind of thing, let's say I'm interested in the probability that x is very large, that it's at least some value a.

I claim that this is the same thing as the probability that e to the bx is at least e to the ba. And this is true for all b that are non-negative. So this is the first key trick. So remember, when we had the proof of the Chebyshev bound, Chebyshev doesn't assume any kind of non-negativity about the random variable. And I had this trick where when I wanted to look at the probability that x is too large, I squared both sides. And then I defined a new random variable y. And I said that that random variable y is non-negative. And I just said those two probabilities are the same because they're literally the same event.

So that's exactly what's happening here. You can take b to be 1, just for simplicity. I'm just saying that these two events are the same. x being at least a is the same thing as e to the x being at least e to the a. And that's just because e is a nice, monotone function. And this is true now even if I put in a scalar b in there, as long as b is non-negative. So is everyone on board with this? This is fine? Yeah?

STUDENT: Why does b need to be strictly greater than 0?

ANKUR MOITRA: I don't know. Let's just make it strict. I'm not going to use the equality anyways. And we'll get into a technicality. But thanks for the catch. Any other questions? All right. So that's the first idea. So just to belabor the point right, so why is this true? It's true because e to the y is an increasing function.

And so that really means that these two events are the same-- the set of all x's where x is at least a are the same as the set of all x's where e to the bx is at least e to the ba. So they are literally the same event. So this is the first thing in our trick. And so now, exactly like in the proof of the Chebyshev bound, what I'm going to do is I'm going to define a new random variable y which is e to the bx.

So this is a really interesting trick. Just like for Chebyshev, I defined a new random variable y, which was the square compared to the expectation. Now, I'm doing this funny thing where I take the exponential of my random variable. And I'm using that to define this new random variable y. So now, just by construction, y is non-negative. It's a non-negative random variable, which means we can apply Markov.

So let's apply Markov and see what we get. So the probability that x is at least a is at most e expectation of e to the bx over e to the ba. So what this is doing right here is I'm just using two facts. So first of all, this one right here is just the probability that y is at least e to the ba. And I'm using the fact that, if I had probability of y at least e to the ba, this would be my expectation of y. This would be my threshold.

So this is just Markov. And now, I'm using this fact that these two probabilities are the same to replace my left-hand side with the probability x is at least a. So this is subtle. Does everyone follow that? Any questions? Give me a thumbs up if it makes sense. OK, good.

All right. So now, it turns out that the quantity on the right-hand side, this thing right here, this is important for its own right. It actually has a very special name and a special role. So before we go on with the proof, let me explain what this quantity is and why it's so important, because otherwise it will seem a bit too magical.

So this is called the moment-generating function. So the moment-generating function-- we've seen generating functions in the context of combinatorics. And here, it's a generating function in a different disguise because it'll be keeping track of other things. So when we talked about generating functions in the context of combinatorics, the coefficients that showed up in different powers of a-- different powers of x, that kept track of things we wanted to count.

So it turns out that this quantity right here keeps track of important quantities of x. But the way it keeps track of it is that there's a free parameter b. So what I really care about is how this term depends on b. And so it turns out that the series expansion I get in terms of b keeps track a lot of interesting properties of the random variable. That's why we call this the moment-generating function.

So the moment-generating function of x is literally just the thing on the right-hand side. We denote it by mx of b. And it's equal to the expectation of e to the bx. So even though x is our random variable, that's not the thing that shows up in the moment-generating function because I'm taking the expectation over x.

The important thing is really just how that expectation depends on b. So let's tease out these properties of the moment-generating function, just to get some intuition for what properties of the random variable this proof of the Chernoff bound is using. So let's do some intuition.

So first of all, mx of b is a function of b. In fact, one way that we could think about it is we could take the Taylor expansion of e to the bx. So if we take the Taylor expansion of e to the bx because, inside the expectation, x is just some particular realization. And it could look like 1 plus bx plus 1 over 2 factorial b squared x squared plus 1 over 3 factorial b cubed x cubed and so on.

So if we just take the Taylor expansion of our exponential function, this is what we get. And now, we can use linearity of expectation. And this is the same thing as the sum from i equals 0 to infinity of 1 over i factorial b to the i the expectation of x to the i.

So this is what I promised you. So the moment-generating function is this generating function that depends on b. And what is the first term of this? Well, it's trivial when I equals 0 because I just get 1 out of it. But now, when I do the next term corresponding to i equals 1, what am I getting? Well, the coefficient of b is the expectation of x.

And when I look at i equals 2, it's not the variance of x But it's the second moment of x, and so on, and the third moments. So the moment-generating function keeps track of the moments of x. And in fact, what I claim is that you can extract the moments from the moment-generating function. So we'll see how this works just in a simple case.

And this will parallel what Peter taught you about generating functions. So how could I go about extracting a particular moment of x from the moment-generating function? Well, in particular, the way that I can recover the first moment of x is just by taking my moment-generating function, taking the partial derivative with respect to b, and then evaluating it at b equals 0.

All that would do is it would kill the first term that's just a constant. And I just pick up the expectation. And then when I evaluate it at b equals 0, all of the higher order terms would disappear. So you can check that this really does give you what you want right. In particular, the partial with respect to b of mx of b, we can just apply it to this expression that I wrote out right here. And that's the same thing as the partial with respect to b of 1 plus b bx and so on.

And then we'll get this expression that looks like ex plus bex squared and so on. And we evaluate at b equals 0. We get exactly what we wanted. So for example, how could I extract the second moment of x from the moment-generating function, just to make sure we're on the same page? Yeah?

STUDENT: You take the second derivative.

ANKUR MOITRA: Second derivative. You just have to be careful with the constants. But they all work out nicely here. So this gives us some intuition about-- I told you that the Chernoff bound would be using higher order moments of my random variable. I'm not just using the first and second moment, like I am with Markov and Chebyshev. And the way that I'm using it is really what's happening on the right-hand side. That moment-generating function contains all of the information about x that we'll want.

So now, let's get back to the proof and figure out why this is a productive thing to do. So here's our first key lemma. So in the setting of the Chernoff bound, where we have x equals the sum of these xi's and all these Xi's are mutually independent, then what I claim is that the moment-generating function of x evaluated at b is the same thing as the product of all of the constituent moment-generating functions for each one of the xi's all evaluated at the same thing, b.

So that's the first key statement. This is why the moment-generating function plays nicely with mutual independence, is because it decomposes as a product of a lot of things. Was there a question? No? OK. All right. So the proof for this is very simple. It's just about unpacking what's happening in this definition.

So by definition and our assumption, we just have to remember what the moment generating function is. So mx of b is defined to be the expectation of e to the bx. That's where it came from. And then we can remember what exactly x is because x is the sum from i equals 1 to n of the xi's. And now, what should I do in my last step right here? So how am I going to proceed? Yeah?

STUDENT: You can break the [INAUDIBLE].

ANKUR MOITRA: That's right. So just to belabor this, let me write it out in two steps first. I can write this as the product from i equals 1 to n of each of these eb xi's. But now, the crucial thing is that, by mutual independence of the xi's, I actually know that not only are the xi's independent, but so are the e to the b xi's. As well.

And so by mutual independence, we can pull the product outside. This is where I crucially need not just pairwise independence. And I'll get the product of all of these expectations of e to the b xi. And that's literally the product of their moment-generating functions exactly as I promised. So that's the first lemma, the first key thing we need for the ingredient for the Chernoff bound. Any questions?

So this is the basic architecture of the proof for the Chernoff bound, is that we define this new random variable y that's the exponential of x. And then when we applied Markov, we ended up with the moment-generating function the right-hand side. And the moment generating function turns into the product of all of the constituent moment-generating functions.

So now, what we need to do is we need to actually estimate the individual terms that go into the right-hand side. So now, the whole name of the game is executing the strategy and figuring out what the right-hand side looks like in this one bound that I wrote down right there. So now, we're actually going to use the properties of like the pi's.

So let me tell you the second key lemma, which will be our second main ingredient in the proof. So let's figure out what these moment-generating functions are for each of these x sub i's. So here's what I claim. This is our lemma 2. So if x sub i is 1 or 0-- again, with probability pi and 0 otherwise, so same as the assumptions of the Chernoff bound-- then what I claim is that the moment-generating function of xi evaluated at b is at most e to the pi e to the b minus 1.

So this is now just the numerical value for it. Obviously, it has to depend on what the probability pi is of xi being 1 for that Bernoulli random variable. And then it also depends on where we're evaluating the moment-generating function at because that shows up as a double exponential bound.

So what we're going to do to prove this is we'll use the fact that 1 plus y is at most e to the y. And this is true for all y. And so this is a very simple thing to prove just with calculus. Or you can just plot it to check yourself. You plot 1 plus y and look at e to the y. And you'll see that the line is below the exponential curve. So this inequality is going to play a crucial role in getting a convenient form for the right-hand side.

So now, let's prove this lemma. Well, we're going to prove this just by direct computation. And eventually, we're going to appeal to this fact in order to write it in a more convenient form. So what is the moment-generating function of xi? Well, we can write it out because there are only two possibilities.

So either xi equals 1 or xi equals 0. In the case where it's 0, our quantity contributes a 1 because we get e to the 0. In the case where it's a 1, we get an e to the b. So this is just literally using the definition of the expectation. And now, we can plug- in for what the probability of xi being 1 is. So this gives us, on the right-hand side, pi e to the b plus 1 minus pi. And we can write this in a slightly more convenient form. We can write this out instead as 1 plus pi times e to the b minus 1. So I'm just collecting the two terms.

And now, I'm almost done because I should use this fact I told you I needed about the exponential. So I'm going to think about my y variable as just being this pi quantity on the right-hand side. I have something that looks like 1 plus y. And I'm going to upper bound it by e to the y.

So using our fact about exponentials, we get that the moment-generating function is now, at most, e to the pi times e to the B minus 1. So literally, all I'm doing is I'm appealing to the fact where this is my y here. And that's the end of the proof for lemma 2.

So right now, it's not totally obvious why we want an expression like this. Why did we not just stop at the expression I had above? Let me give you a little foreshadowing of where we're going. See, at the end of the day, the way that the Chernoff bound works is that the right-hand side has this quantity mu in it, which is the average of your random variable. And that quantity mu is the sum of the pi's.

So in our lemma 1, we're taking the product of all these moment-generating functions. And so what's going to happen is we take the product of terms that look like this, where the pi is happening upstairs. And then we're going to get some of the pi's. And that's where our mu is going to fall out. So if I kept that other bound over there, that would have been still true and just an even sharper bound. But it doesn't give me a convenient expression for actually getting the Chernoff bound that I want.

So now, we can prove the Chernoff bound. And I'm just going to prove the upper tail. But you can check the notes for the proof of the lower tail. And it works very similarly. What I'm going to do is-- at the end of the day, I have to figure out how to execute this proof strategy because I have a whole bunch of parameters that I now need to set.

So I have to set my threshold a in that bound that I wrote up there. I have to choose my parameter b, which is my helper parameter, so that I get a tight enough right-hand side that I get the upper tail for the Chernoff bound popping out. So let me just tell you what I'm going to choose for these a and b.

So a is the obvious thing because it corresponds to the event we care about. So it's 1 plus delta times mu because that's our upper tail condition. And b we're just going to set to be equal to the natural log of 1 plus delta. So this might look a little strange. But the way that you would prove this yourself or come up with this b is you would just work through the calculations I'm about to do, but keeping B as a free parameter.

You would get a right-hand side that depends on b. And then you would optimize over what is the best choice of b that gives you the tightest upper bound. So all I'm doing here is I'm doing the work for you. And I'm telling you that this will be the answer for b. So let's just work with this b. And then we'll check what happens for the upper tail.

So now, we can put this all together. I told you this would be a long proof. All right, so let's put this all together.

Well, we care about the probability that x is at least a. And remember, a is 1 plus delta times mu. We know that this is the same as the probability that e to the bx is at least e to the ab, where a is, again, my 1 plus delta mu. This uses that first fact that these two events are the same.

We know by Markov that this is at most the expectation of e to the bx. That's where we first saw the moment-generating function. Now, we can appeal to lemma 1, which told us how to write the moment-generating function as the product of the moment-generating functions of its constituents.

So we get the product from i equals 1 to n of all of these individual moment-generating functions still over this denominator eab. And finally, we can use lemma 2, which gave us a good quantitative upper bound for the setting of the Chernoff bound. We get the product from i equals 1 to n of e to the pi e to the b minus 1 all over e to the ab.

And I'm going to call this right-hand side star. And the name of the game is really just to figure out what star is so that the Chernoff bound pops out. So our bound star right here, we can plug in for all of our values of a and b. We'll get the product of i equals 1 to n of e to the pi times delta because b is natural log of 1 plus delta. So when I plug this in, I'm going to get delta sitting out right here.

And then my denominator, well, that's e to the ab. So I claim it's equal to 1 plus delta to the 1 plus delta times mu. Because again, when I have e to the ab, I can take the b downstairs. And I'll get 1 plus delta instead of my e. And that'll be being raised to the a-th power, which is exactly this 1 plus delta times mu.

So you see, for the Chernoff bound, it's very easy to potentially get lost in some of the computations. But the highest level point is really what's happening over here, the architecture of how the pi's of the proof fit together. The rest of it now, from now on, is just calculus because all we have to do is figure out how to actually bound the right-hand side so that the upper tail for the Chernoff pops out.

So now, we have this expression right here. And we can write this out. Of course, the sum of the pi's is mu. So we're going to get e to the delta mu all over 1 plus delta to the 1 plus delta times mu. And what I claim-- this will be my lemma 3, The Last piece that we need for the Chernoff bound. I claim that this right-hand side is at most whatever shows up in the upper tail, this e to the minus delta squared mu type of term. So once I prove that last lemma 3, we'll be done.

So let's prove that last part. And it's not too bad. So the way that we're going to do this is like this is what our lemma 3 is. We want to show that this is at most this quantity right here. So what I'm going to do is I'm going to take the natural log of both sides.

So the natural log of the left-hand side of my expression is equal to delta mu minus 1 plus delta mu natural log of 1 plus delta. And I can rewrite this expression and pull out the mu. So I'll get mu times delta minus 1 plus delta natural log of 1 plus delta. Nothing fancy is happening here.

And now, we need our last basic fact that we're going to take advantage of, which is that I claim, for all x that are strictly positive, we have the following bound. So the log of 1 plus x is lower bounded by x over 1 plus x over 2. So this is all very natural. I mean, I told you these things follow from the Chernoff bound, from calculus types of proofs.

And the idea is that if I took the Taylor expansion of natural log of 1 plus x, my linear term would be x. And now, what I want to do is I just want to make sure that the way that I get rid of the other terms, I always have a valid lower bound. So here, what I'm going to do is you have to choose the right expression to put in the denominator. But this is, again, something you can check either with elementary calculus or just by plotting these things to convince yourself it's true.

But now, we're in good shape because what this tells us is that-- oh, boy. Sorry, I'll have to do it over here. So using this fact and our expression for the left-hand side, this is at most mu delta minus mu 1 plus delta delta over 1 plus delta over 2.

So you can see what's going on here is that I'm just plugging in this natural log of 1 plus x in order to get rid of this term right here, because I'm replacing this with delta over 1 plus delta over 2. And then I'll get exactly this expression that I've written down right here. And now, when we collect terms, we'll be home free.

So this is equal to mu times a whole bunch of things. So I'm going to multiply this part by 1 plus delta over 2, just so that I can put everything to have this denominator of 1 plus delta over 2. And then I'll get delta plus delta squared minus delta minus delta squared. And then life is good because I get minus mu delta squared over 2 plus delta. And that's the proof.

So one of the points that I want to make-- so we'll talk more about this in recitation. So the Chernoff bound is definitely one of the most complicated proofs we've done up until this point. But when you're presenting things that are this complicated or things like this in your term paper, one of the things to keep in mind is that it's not really just begin proof end proof. So the way that you break up the proof makes it a lot easier to understand the overall strategy.

So for example, the way that I broke it up into these lemma 1, lemma 2, and then lemma 3, it allows me to give an overview of the Chernoff bound, even though it was a full lecture-long proof, that fits very compactly. So a lot of times, when you're first doing proof-based math, you have this temptation to prove everything linearly, where you just start at the beginning and you keep proving things until you're at the end.

When I was first starting out, that was also my temptation, too. I think my first paper, when I wrote it, it was a 15-page begin proof end proof. And then my advisor told me afterwards that he didn't understand the proof, which was a little bit worrisome. In fact, it wasn't even so obvious from the way I'd written it that, when you look at the dependencies between the different pieces, that they were actually acyclic.

So to a certain extent, sometimes, you can refactor a proof so that it's not just linear so that you can give people a preview of how exactly it fits together because there's a difference between what are the parts where the action is happening, where you're telling me about what's the strategy you're going to use, where you're going to use mutual independence, and the parts which are mechanical and are more technical to actually execute, because a lot of these things are algebraic computations.

I can follow these things and check them. And I can work them out myself. But the truth is that I don't learn a whole heck of a lot from the proof of lemma 3, except for the fact that, if I massage things algebraically, I'll get the right expression that I wanted. So you should think about a lot of these pedagogical things for your own writing, because are there ways to assert some pieces of the lemma so that you can give people a blueprint for how the pieces fit together, and then maybe fill in the parts which are the most interesting first and table the parts which are more mechanical and algebraic for later?

So another thing I did to try and present this proof was that there were things that I explained that I didn't literally need to explain. So can you give me any examples of things in the proof, the way that I presented it, that aren't strictly necessary for getting the upper tail for the Chernoff bound? Just out of curiosity. Most of it's on the board. And also, are there any questions about the Chernoff bound proof? Yeah?

STUDENT: So at the beginning, you said if you want to adapt it to a different variable, we could use bounds on the moments? Where exactly in the proof-- let's say we want to use Bernoulli inequality. Where would we incorporate that information?

ANKUR MOITRA: Great question. So where in this proof did I use the fact that the xi's were Bernoulli? So now, when I have this architecture of the proof just sitting on the board, it becomes a lot easier to think about each individual piece. That's the advantage of presenting it this way.

Did I use Bernoulli here? No. Did I use Bernoulli here? No. What about here? This was the place I used it. So that's what I was saying when I meant that the Chernoff bound can be adapted to many different settings. It's really just a blanket statement that sums up independent random variables have nice tail-bound properties.

And the only thing that would change was, if you gave me non Bernoulli random variables, I would have other explicit estimates. Now here, I just used Bernoulli to figure out what exactly the moment-generating function was literally. This is an equality. But then I had to do a bit of massaging to get the pi's to be in the exponent.

So that's the part that would change, is that if you gave me some other random variables, I could plug in and compute literally what the moment-generating functions are for them. But then I would have the same problem about how to massage that into an expression where, when I take the product of all of them, I get the mu sitting out there.

So those are things which can be done. But for the most part, they're not very instructive. So you can just cite whichever Chernoff bounds you want when you have to use it. And you should know that the architecture of the proof is not very surprising. Are there any other questions about it? So the other thing I'll ask, which I asked before, is what part of my exposition was not literally needed to prove the Chernoff bound? Yeah?

STUDENT: [INAUDIBLE]

ANKUR MOITRA: Yeah. Actually, I didn't need to talk about moment-generating functions at all. So I could have just said that the right-hand side is the right-hand side. It is whatever it is, expectation of e to the bx. And then I could have just said, because x is the sum of xi's, it's equal to the product of their expectations. And we could have skipped straight to Lemma 1.

That's one way I could have proved the thing. And there would be nothing mathematically wrong with that. But this class is a CIM. So it's not just about writing mathematically correct statements. But it's also about thinking about how to do the exposition to get the audience to better understand and appreciate it. Why do you think I talked about the moment-generating function, even though I didn't literally need to? Sorry to put you guys on the spot, but this is the way I designed it, was I want to interrogate you guys about why I presented it the way I did. Just for fun? Stalling?

STUDENT: It's more modular [INAUDIBLE].

ANKUR MOITRA: So it's more modular? Yeah, that's one good thing. Any other ideas why I did it that way? Modularity was definitely very important for being able to fit together a blueprint of the proof on just one section of the board. But there are other reasons I used the moment-generating function. Yeah?

STUDENT: So [INAUDIBLE].

ANKUR MOITRA: Yeah, that's a big reason. So I used it to demystify because otherwise this would look like black magic. We proved the Chebyshev bound, where we did that same trick of creating a new random variable. And that was very natural because we had a random variable before, which wasn't non-negative. So we squared it to create a non-negative random variable.

But here, where would you come up with the idea of taking the random variable and exponentiating it? That's a very clever idea. And it's used throughout probability. We're not going to see much more exotic examples. But otherwise, this would seem totally exotic. Why wouldn't I try other functions? It's really because the exponential is the thing that keeps track of all of the relevant information about the random variable. I could care about.

So I not only told you that this is called the moment-generating function, but I explained how you could extract properties of the random variable from that moment-generating function. That way, we can talk about what properties are we literally using about the random variable. There's another reason I introduced the moment generating function. I introduced it for several reasons. Yeah?

STUDENT: Maybe [INAUDIBLE].

ANKUR MOITRA: Yeah, that's a good thought. That's a good idea. But it also has to do with you, the audience. So if I were teaching a different class, I might not have said the things I said about the moment-generating function. What did I use about the moment-generating function that was specific to you and what else you've learned in this class?

STUDENT: We just covered generating--

ANKUR MOITRA: That's right. We just covered generating functions. So the way that I chose my exposition was also based on what I konw you know and how to make connections between it because, when we started the probability unit, we then switched gears and talked about counting. And there are a lot of connections between probability and counting, things like inclusion exclusion formulas. We already saw binomial coefficients coming up. And the same way, I wanted to further strengthen the connections between probability and counting because, usually, they're just different languages for talking about the same thing.

All right. So these are all things you should think about when you're doing your own exposition. I'll tell you one last thing. And then we'll call it a day. So the Chernoff bound is this amazingly powerful tool. It was invented by Herman Chernoff, who was here for a long time. It's one of his most famous discoveries.

But what's kind of funny is he-- I mean, he actually just turned, I think, 100. So he had his 100-year birthday recently. In fact, another fun fact about him is that his wife is alive, too. And they're believed to be the oldest living couple in the state of Massachusetts. So that's quite something.

But the Chernoff bound also gets misused quite a lot because it sounds like this amazing fact that, when I first introduced it, I talked about how we wanted to drive the failure probability of some randomized algorithm down. And if we repeat it a bunch of times, things like the weak law of large numbers would tell us, to get our failure probability down to 10 to the minus 30, we'd have to repeat it like 10 to the 30 times.

The Chernoff bound would tell us that, because the right-hand side is this exponentially small decaying probability, we only have to repeat it natural log of that many times. So it tells us that random variables concentrate at extremely fast rates. But what's really critical about it is the assumption of mutual independence.

All of this was predicated on it. The moment-generating function keeps track of not just pairwise information, but all higher order moments. And there are a lot of famous examples that affect all of us of people thinking about using the Chernoff bound, even though it doesn't really apply. So this is probably pre your time. But I'll tell you a story about the financial markets, which was certainly relevant when I was finishing up college.

So right around that time, there was this big financial crisis in 2008. And so here, you really have to be careful when you try and use this reasoning and when you do and don't have independence. So it's not just important-- I want to say that it's not just important to know how to use the Chernoff bound. But I want to claim that it's just as important to know when tail bounds don't apply.

So just a little bit of personal background-- so both my parents are computer scientists. My whole family is computer scientists. So I really didn't want to be a computer scientist. I really showed them. But I tried out all kinds of things when I was at your age. So around that time, I worked in Wall Street. And I worked on some quantitative finance problems. And this was pre the financial crisis. But I remember one of the things that they were big into were very fancy assets.

So there are things called collateralized mortgage obligations. Has anyone heard of collateralized mortgage operations? CMOs? No? So the way that this works is that you have a whole bunch of mortgages that you sell to people for them to buy their house. And then implicitly, when you're giving someone a mortgage, you have to assess their credit worthiness to figure out the likelihood that they're going to pay back the loan.

Now, for people who have great credit ratings, maybe that's not such a big risk. But what happens if you're making riskier mortgages? So those are things that you can charge more for. You can have a higher interest rate. But you have to deal with the fact that the probability that they pay it back might be less than you would hope.

So you had all of these issues in 2008 that there was a huge amount of construction going on. People were buying a lot of these houses. Some people had multiple houses. And they had very poor credit ratings. But the banks didn't want to hold on to these mortgages because they're risky products. So how exactly-- what do you do with these financial products?

So there was this great idea, which was to create synthetic products instead. So what you do is you take a whole bunch of what are called subprime mortgages. And you bundle them together because the hope is that all of these small probability events of the person failing and defaulting on their loan, maybe if you combine them together, you can appeal to things like the Chernoff bound and say that things concentrate more around their expectation.

So you have the issue that these subprime mortgages still have a reasonably large failure probability. So no one would buy them. So what you do is you create what's called a waterfall payoff structure. So you do something called creating tranches. So you have this baseline of all of these subprime mortgages. And then the first people who default on their loans, the people who are going to eat the losses are the people at the bottom of the waterfall. That's where the water dries up the first.

And the people at the very top of the waterfall, they only get losses if basically everything collapses or a large fraction of the mortgages collapse. So maybe each mortgage has a 10% chance of defaulting. And in order for this person at the very top to eat any losses, maybe 40% of the mortgages have to default.

So if you thought about these probabilities as being independent, this would be an extremely safe asset. What's the chance that I take a coin that is heads with 10% probability and I flip it 1,000 times and I end up with 40% heads? It's ridiculously small because we just proofed it with the upper tail.

So that allowed people to get credit rating agencies to say that these are very safe assets. So when you have a retirement plan and you put money into your retirement account, there are certain rules and regulations about what the financial risk and the gradation for the asset needs to be in order for these retirement companies to even purchase the asset because these things that are managing your retirement shouldn't be purchasing risky accounts.

So all the credit rating agencies rated these things as if the failure of the mortgages were independent events. And that made them say that, actually, these are extremely safe assets. Maybe they're even safer than things like the US currency. The trouble is that Chernoff bounds don't really apply. So what's the problem with this? Why don't tail bounds apply? Yeah?

STUDENT: They're not independent.

ANKUR MOITRA: They're not independent. That's right. What causes someone to default is a lot of economic strife. And when that happens on a massive scale, all of a sudden, you have this domino effect. And these things that were previously viewed to be as very safe things are not really safe, even though at the time I remember people were using things like Gaussian approximations to figure out their probability of default.

So I'll leave you with that last nugget, which is probability is very important for your lives. You should use the Chernoff bounds in the classes we give you in problem sets. But in real life, you should be critical about whether or not it applies because the dependence between these random variables is a nasty thing. And there are whole courses you can teach on dealing with dependent random variables. But to be continued. So see you next time. And we'll be starting our modular arithmetic unit on Tuesday.