[SQUEAKING]

[RUSTLING]

[CLICKING]

ANKUR MOITRA: So today, we're going to be continuing our probability unit. So I just want to remind you where we were. So when I was last lecturing, I introduced the basics of probability. I told you what a sample space was. And then it led to things like events, where we talked about what was the probability of different events. We talked about independence. We talked about random variables that are functions from our sample space, the real values.

And once we had random variables, that allowed us to do a lot of exciting things, like we could talk about the expectation and the variance of the random variables. It allowed us to do things talk about conditioning, like when we condition on some event in the sample space occurring-- how does that change the probability of other events, how does that change the expectation of random variables.

And then we talked about this powerful tool of inclusion-exclusion. And then after that, Peter gave you a bunch of lectures on counting methods, on generating functions. And the truth is, for the introductory lectures in this class, there are a lot of tight connections between probability and counting. So we really need both perspectives. And we're going to go back and forth between them.

But today, we're going to cover one of the other important topics in probability, something that we're going to spend a bunch of time on, which are called tail bounds. So the idea for a tail bound is to get a handle on how likely it is for random variables to be far away from their expectation.

So remember that we take a random variable and we look at its expectation. And that tells us, on average, what we expect the random variable to be. But of course, some random variables might have large averages just because they have some very tiny probability of having a gigantic value.

So then we talked about notions variation and variance that told us how well a random variable concentrates around its expectation. So today, we're going to make that precise. We're actually going to get bounds on what's the probability that a random variable is very far away from its expectation.

So let me tell you one simple tale bound in action. And really, the theme of today and Thursday's lecture will be we're going to use more and more information about random variables to get tighter and tighter control. So let's start off with a very basic example of what theorem needs very little information about my random variable, but gives me a rudimentary baseline of what's the probability it's far away from its expectation.

So this is a very simple to state and prove bound. It's called Markov's inequality. And here, we're going to assume that x is not just a random variable, but it's a non-negative random variable. Then what I claim is that the probability that x is larger than some constant C is bounded by the expectation of x all divided by C.

So this makes intuitive sense because, if I have this non-negative random variable, as C gets larger and larger, I'm going way out into the tails of my random variable. The probability that the random variable is actually that large should be getting smaller and smaller.

Now, the important thing is that the expectation shows up on the right-hand side, because that gives us some baseline of how large we expect x to be, on average. So the larger its expectation is, the weaker my bound on this tail probability is. But as C increases, I also get some width.

So we're going to prove this bound. And then we'll state another corollary, an equivalent way to rephrase Markov's bound. And you can go back and forth between them, depending on what's more convenient. But one thing I want to emphasize is that, here, I'm crucially going to need the fact that x is non-negative.

So just as a spoiler, I'm going to ask you where in this proof I'm really appealing to non-negativity. And after I give you the proof, I'm going to expect you guys to tell me where this fact was used. So let's prove this. And the proof for this is really just breaking up the expectation into different pieces. So let's start off with our random variable x. We know its expectation, E of x. And I'm going to break this up using conditional expectation.

So I'm going to break it up into two terms. I look at the conditional expectation of x given that x is at least some value C times the probability that x is at least C. And then I'm going to add the other contribution, which is the conditional expectation of x if that event does not happen, if x is strictly less than C, times the probability that x is less than C.

So this should look familiar. This is basically a random variable version of something we covered way back when, which is we talked about the law of total probability. We had things, the probability of some event B. And we talked about it as the probability that B happens conditioned on event A happening times the probability that event A happens plus the probability that B happens conditioned on A not happening times the probability A does not happen.

So when we talked about conditional probability, this was one of our first applications, once we had Bayes' rule, was we used it to reason using the law of total probability. And this is just the same thing, but expectations of random variables. Sometimes, you'll hear this referred to as the law of total expectation. But it's just breaking up my space into two different things-- either this event A happens, that x is at least C, or it does not happen. And then I average the conditional expectation in those two different possibilities.

And so now, what I claim is that I can look at each of these pieces. I claim that the first piece right here is at least C. And then I'm going to pick up this probability of x at least C. And then I claim that the second piece right here is non-negative. So that I'm just going to drop. And I'll get a lower bound.

So I've really just used the fact that the expectation of x, given that x is at least C, had better be a value at least C because x always takes on a value that's at least C. And then I'm still going to pick up the probability that that event happens. And for this term, I'm just going to drop it, because this term is non-negative. And so now, what I can do is I can just rearrange to finish the proof.

That's it. So that's our first tail bound. Our tail bounds will get a bit more sophisticated from this. But this already gives us some rudimentary understanding of the probability that x is very, very large. So now, let me ask you the question that I promised I was going to ask you. Where in this proof did I use the fact that x is non-negative? Yeah?

STUDENT: When you were counting x [INAUDIBLE] x given, x less than c, [INAUDIBLE].

ANKUR MOITRA: That's right. That's exactly right, because if x were not a non-negative random variable, I couldn't say that this term is non-negative. This term is definitely non-negative. But otherwise, if x took on negative values, this could be negative. And dropping it, I wouldn't get an inequality like this. Perfect.

So we definitely have to use this. And that's crucial for Markov's bound, at least the intuition behind it. I mean, this is the algebraic proof. But the intuition is just that if x had a very large probability of being at least C, then its expectation could not be what I promised it was, because the expectation would actually be too large. Any questions? Make sense? All right.

So let me restate this in another form. I'll state it as a corollary. This is just a restatement of Markov's bound. But sometimes, you'll find it convenient to invoke it in one form versus the other. So I claim, under the same conditions, if x is non-negative, then the probability that x is at least C times its expectation of x is at most 1 over C.

So let me make sure we're all on the same page. So why is this a corollary? How do I get from here, the thing I proved, to here? In fact, let me make it a bit easier. Let me just call this a different constant, C-prime. Why is this a corollary? They seem related. Yeah?

STUDENT: You just substitute c prime [INAUDIBLE] c times e. [INAUDIBLE].

ANKUR MOITRA: Yeah, so let me do it the other way. Let me substitute for C is equal to C-prime times Ex. Then I get exactly this bound right here. And then on the right-hand side, I cancel out the E of x. And I get 1 over C-prime. So this is very intuitive because, if I have a non-negative random variable, the probability that it's twice as large as its expectation is at most 0.5. Because otherwise, if that probability were larger than 0.5, the expectation would be wrong. So any questions about this? All right.

So just to illustrate my point that, in tail bounds, we're going to be asking for tighter and tighter bounds that use more and more information about our random variable, let me do a running example, which we'll use as a way to check how much progress we're making. So let's do the following running example. And we'll come back to it with our next tail bound.

So in the US, the average male height is about 69 inches. So we can use this fact to reason about the probability that someone is super duper tall, because x being the height of a random male in the US is a random variable. So we know its expectation. And we can use Markov's bound and just see what Markov's bound tells us.

It tells us the probability that x, which, remember, is our random variable-- this would be the height of a randomly chosen US male. The probability that it's at least 2 times its expectation, which is equal to 138 inches, or 11 foot 6 inches, that probability is at most 0.5. So we've proven this amazing fact.

You're walking down the street in Boston. The probability you see someone who's above 11 foot 6 is at most 0.5. So pretty good or maybe not. So see, the problem is that Markov's bound-- you can actually check-- really is tight. It's just that real random variables, a lot of times, there's more information we know about it-- things like the variation of the random variable. The more information you tell me about the random variable, the tighter my tail bound will be. And this Markov is just the get me over first warm-up for a tail bound.

So we'll come back to this example. I want to make sure we still have the right intuition for this proof. So let me ask you a diagnostic question, just to make sure we're on the same page. I think probably you can guess the answer to this. But what I really care about is to find an example. So does Markov's bound hold without non-negativity?

So in the statement of the theorem, I definitely assumed that the random variable was non-negative. In the proof I gave you, I definitely used the fact that x was non-negative. We talked about where we used that. But could I hope for some bound that looks like this, even when x is not non-negative? Maybe there's some other proof out there that doesn't use non-negativity that still bounds the probability that x is twice its expectation by 0.5 or something like this.

So I claim that this is spectacularly false, but this is good intuition. So how would I come up with some random variable that has negative values and completely breaks any kind of statement like this, that the expectation gives you a handle on the tails? Anyone have any intuition? Yeah?

STUDENT: [INAUDIBLE]

ANKUR MOITRA: Perfect. I love it. So this is definitely not true. It's spectacularly false. So let's just consider some x where the right-hand side doesn't make any sense because the expectation is 0. So what if x is 1 with probability 0.5 and it's minus 1 otherwise? The trouble is that the expectation of x is equal to 0. But the probability that, say, x is at least 3 times its expectation, which is still 0, is definitely not less than 1/3.

So non-negativity really is essential. And sometimes, these examples are very useful for both building intuition and conveying a proof to your reader. So these kinds of things, when you start thinking about your term paper, even if I didn't literally need to show this to you, this is the good kind of thing to keep in mind as a pedagogical device.

All right, so let's prove our next tail bound. This one will be fairly simple as well, but it's much more powerful. And we're going to get some awesome applications of it. And this is called Chebyshev's bound. Actually, I'm not even sure about the history because I think Chebyshev's bound was proofed before Markov's bound, even though it's going to use the proof of Markov's bound. So I don't know. Maybe Markov just looked at his proof and called the thing inside it his own bound.

So in any case, now, we're going to start off with any random variable x. So this is already very different than Markov's bound because I'm not going to assume non-negativity. I don't even need it for this proof. So for any random variable x, what I claim is that the probability that the absolute value of x minus the expectation of x is larger than C-- so it deviates not by a multiplicative factor of C, but by some additive factor of C from its expectation-- is bounded by the variance of x all over C squared.

So this will be my standard form of Chebyshev's bound. And the same way for Markov's bound-- I stated it in one form, and then I stated a corollary-- I'll do the same after we prove this bound. But this is what I promised because now we're going to use extra information. See, Markov's bound did not have the variance showing up. So you have to actually be able to compute the variance in order to appeal to the Chebyshev bound.

But as we'll see, it's going to give us much tighter bounds and some really cool qualitative corollaries. So let's prove Chebyshev's bound. And then I'll state the other corollary. The proof will also be short and sweet. But it's going to have a really cool trick to it, which is a hint about the tricks to come later when we do much fancier tail bounds.

So the first thing I'm going to do-- I can rewrite the quantity I care about, the probability that the absolute value of x minus E of x is at least C I claim is equal to the probability that x minus E of X quantity squared is at least C squared. So this is trick number 1 for Chebyshev's bound. It's not a hard trick to explain. But this is a trick. Why are these two probabilities equal? Why am I allowed to say that these two probabilities are equal?

It's because they're the same event. So at the end of the day, what's happening in here is I have an event. Once you fix the random variable, I know its expectation and you fix C. I have this event, whether or not x is far from its expectation. And I claim that these two probabilities-- what's going on in here are two different events. But these events coincide. The only way that one event triggers is if the other one does too, just because we're taking the absolute value. Make sense? So that's trick number one for Chebyshev.

And now, here's the really cool trick. What we're going to do is we're going to define a new random variable. So I'm going to call this random variable y. And it's going to be equal to just the thing that shows up in here on the right-hand side-- x minus the expectation of x quantity squared.

See, I told you, in Chebyshev, we're going to appeal to Markov's bound in order to prove it. But if you remember, for Markov's bound, we needed that the random variable we are looking at is non-negative. For Chebyshev, I'm not assuming that x is non-negative to begin with. So what I have to do is I have to manufacture my own non-negative random variable.

And that's this new random variable y that I just constructed right here. This is definitely non-negative because I'm taking the square of a value. And now, I can appeal to Markov's bound. So using Markov on y, we know that this quantity right here, which I'll call star-- star is equal to the probability that y is at least C squared. That's literally just using the definition of what my random variable y was.

And now, appealing to Markov's bound, which is allowed because y is non-negative, this is at most the expectation of y all over C squared. Notice that my threshold here is C squared, not C, as in the original statement of Markov's bound. And now, what I can do is I can just remember what y is.

So y is just the expectation-- y itself is x minus the expectation of x quantity squared. I still have my C squared. And when I take the expectation of y, this is literally the definition I gave you earlier for the variance of x. So that's my proof. Any questions? Good? All right.

So as before, I'm going to state another equivalent way to think about Chebyshev, which is sometimes a bit more convenient. So feel free to use whichever is easier in a proof. So again, for any random variable x, the probability that the absolute value of x minus its expectation is larger than C times sigma of x-- I'll define what sigma of x is-- is at most 1 over C squared.

And sigma of x is called the standard deviation. And it's just defined to be the square root of the variance of x. So the same way I had this simple version of Markov that allowed me to remember it-- the probability that a non-negative random variable is twice its expectation patient is at most 0.5-- the probability that a random variable is 2 times away from its expectation in terms of its standard deviation is at most a 0.25. All right, so let's go back to our running example. And let's see why Chebyshev is a lot more useful for us.

So in particular, I need to tell you more information about this random variable, about the height of a random male, US male. And what I claim is that the standard deviation of male height is 2.7 inches. So now, we're in good shape because instead of using Markov, we can instead use Chebyshev.

We have the probability that x, where x is our same random variable as before-- the probability that you're 2 standard deviations above average, which is the same thing as 6 foot 2.4, is at most 0.25. So this is still not a great bound. But it's at least not a nonsensical bound, like our 11 foot 6 inches version before. And so you can do the same types of exercise. When you tell me not just the variance of x but more information about the moments of x, you would get tighter and tighter bounds until you would actually get something pretty reasonable.

All right. So now, I want to tell you-- now that we have Chebyshev, it may not be obvious, but it turns out to be very qualitatively powerful, not just numerically in these examples like height. So let me tell you about one really cool corollary. I'll state it as a theorem. And then we'll talk about some applications to gambling.

So Chebyshev allows you to prove one of the most important results in tail bounds, which is called the weak law of large numbers. Now, the weak law of large numbers, we're going to strengthen it at various times and get better quantitative bounds. But already, we can see some of the power of tail bounds. So let me state what this weak law of large numbers is. And then we're going to prove it just using the tail bound technology we've developed so far.

Now, for the weak law of large numbers, you have to be very careful with the quantifiers. So let me write it down. And then we'll talk about it. So first, I'm going to fix some positive epsilon. That'll be some measure of how close I want a random variable to be to its expectation.

And I'm going to let x1 all the way up to xn be independent copies of some fixed random variable x. And moreover, let's assume that the expectation of x is equal to some value C. And I'm also going to assume-- and this is critical for the weak law of large numbers-- that its variance is not infinity-- so just some basic bounds that my random variable is somewhat well-behaved.

Well, what I'm going to do is I'm going to look at averages of these random variables as a way to try and approximate this expectation C. So I'm going to define a new random variable, Sn, that's just going to be the average. It'll be 1 over n sum from i equals 1 to n of xi.

So just to make sure that we're on the same page, let's keep some mental model of this in mind. So one of the most important applications of the weak law of large numbers is to something like polling. So maybe you have some random variable, which is the probability that a random person selected from the electorate votes for one candidate, candidate A over candidate B.

Now, you want to know what's the expectation of x because you want to know who's going to win the election. So you want to do this forecasting before election day. The natural thing you could try is you can try and get independent samples from x just by calling random people and asking them who they're going to vote for.

Now, there are a lot of issues with that. That won't really give you independent random variables. But if we pretend for the moment that it did, each of these xi's is a realization of these random variables. What would my polling scheme do? I would just choose some large denominator n for the number of people I'm going to poll. I'm going to average up all of their votes, which way they're voting. Are they voting for candidate A or candidate B?

And what I'm hoping is that the empirical average of their votes of these random variables converges very well to my true expectation so that I can forecast who's going to win the election. And so that's exactly what the weak law of large numbers tells us. You just have to be careful with all the quantifiers.

Then what I claim is that the probability that Sn is epsilon away from the true value C that this quantity, as you take the limit, as the number of people you're sampling in your poll goes to infinity, is going to go to 0. So there are a ton of quantifiers here.

So let's say that I fix epsilon at the very outset. I want to figure out the probability people are voting for candidate A within some additive tolerance, 0.01. And so once I've fixed my epsilon, this is my success criteria, is that my empirical average of the people I poll-- that empirical average should be within an additive 0.01 of the truth.

And the claim is whatever you fix for epsilon, there's an n that's large enough so that the probability that you're far away by more than this 0.01 is going to go to 0. So the weak law of large numbers is deliberately non-quantitative. And we will get quantitative versions out of it, even just from the proof and from strengthening of our tail bounds. But it just tells you some basic conditions that, if you're given enough samples, you really can approximate the expectation to any desired accuracy. And in the proof, we'll see how large little n needs to be.

So this is quite a mouthful of a statement. Are there any questions about the weak law of large numbers? Give me a thumbs up if this makes sense, at least as a statement. OK, good. All right. So the order of quantifiers is super important for this. Now, let's prove the weak law of large numbers. The good thing is that the proof will be pretty straightforward now that we have Chebyshev's bound.

So first, I claim that the expectation of Sn-- remember, that's a random variable itself because each of the xi's is a random variable. And I could first care about-- is its expectation correct? So now, I can use linearity of expectation to write it out as the sum of the expectations of each of the individual xi's that form it.

Now, because each of these xis are assumed to be independent copies of this fixed random variable x, whose expectation is C, I know that each of their expectations is C. So that means, when I average their expectations, I get the same thing back again. I get C. So this is just a basic sanity check.

If I'm trying to approximate the expectation of my random variable x with another random variable Sn, I better have that the expectation of Sn is equal to x. Otherwise, I'm going to have trouble. So this makes sense. This just means that I have an unbiased estimator of the thing that I really care about.

Now, the crucial thing is computing the variance. We're not going to be able to get by by using Markov's bound. We're really going to have to use Chebyshev. And that's where independence is going to save us. So remember, for Chebyshev's bound, I need to know what the variance of my random variable is in order to quantify how close a random variable will be to its expectation, because that's the thing that shows up in my statement of Chebyshev.

The main action for the weak law of large numbers is just computing what this variance of Sn is. So I can do the same thing as before. I can replace Sn with what it is defined as, in terms of the xi's. And now, what happens when I scale a random variable by some scalar C-- in this case 1 over n? Well, the variance is going to scale by a C squared, because I'm taking the expectation of some quantity squared in there.

So this is the same thing as 1 over n squared times the variance of the sum of the xi's from i equals 1 to n. And now, the main trick is that I claim that this is now equal to-- I'll carry through my 1 over n squared. I claim that this is equal to n times the variance of each of these random variables x. So this is the key step in the weak law of large numbers. This is the part that I want to dwell on. So why is this step OK? So how did I go from here to here? We have to remember way back when with what I told you when we did the probability unit. Yes?

STUDENT: Because they're independent.

ANKUR MOITRA: Because they're independent. That's right. So remember, we had things like the linearity of expectation. We add two random variables. Their expectation of their sum is the sum of their expectations. That assumed nothing about the random variables. It was always true.

But when I talked about the variance of two random variables, the variance of the sum of two random variables is equal to the sum of their variances, but only if I have independence. So the crucial assumption that these are independent draws I need in order to get this step. And that's where we're going to be winning, because you can see that I pulled out a 1 over n squared. And it's only canceling out one of these. It'll cancel out this n. But I'll still have a 1 over n left over.

So that's why, as I take more and more samples, I'm going to get better and better concentration. Just to completely drive home the point, we needed independence in order to get this last step. If I didn't have independence here, what would my polling methodology be? I'm trying to get a hint about who's going to win the election. And I call the same guy over and over again and ask him his opinion. That would be a pretty ridiculous way to get an idea about who's going to win the election. So you crucially need independence.

But now, we can write this out. This is the variance of x over n. And now, we can just appeal to Chebyshev. And Chebyshev is going to get us our weak law of large numbers. So by Chebyshev, we know that the probability that Sn is off from its expectation by an additive epsilon is at most the variance of Sn over epsilon squared, which our right-hand side is the same thing as the variance of x over n times epsilon squared.

So this tells me immediately that, as I want some better and better target accuracy for how accurate my poll is-- maybe I don't want like 0.05 error. I want 0.01-- what's going to happen is that my epsilon is getting 5 times as small. So what do I need? I need my n to be 25 times as big in order to get the same bound on the tail probability.

So actually, in the course of proving the weak law of large numbers, we even got quantitative bounds that tell us exactly how large our n needs to be instead of hiding it all within this limit. So this is one of our cool applications. Any questions? Good? All right.

So let's talk about an application to gambling. I hope none of you all are big gamblers. Actually, it turned out-- so when we bought our house in Belmont, we kept getting all these weird packages and fruits in the mail, like very fancy things. And around Christmas time, people would send us these very expensive wreaths.

And it took us a while to figure out that that was because the person who owned our place before was the CTO of DraftKings. So now, we get a lot of free fruits in the mail because of gambling. So that's good. But let's talk about an application of the weak law of large numbers to gambling, or maybe a caution about why it's not a great idea to gamble too much, especially in Vegas.

So there's an important example, which is called gambler's ruin. So the way that it's colloquially phrased is that playing a game with negative expected value, you'll eventually go broke. So this makes sense. When you gamble in Vegas, definitely all the games, aside from maybe blackjack, if you're counting cards, have this property that they have negative expected value.

And intuitively, it makes sense that if you have negative expected value, on average, you're giving up some money every time you play. But maybe you get really lucky. Maybe there's some probability that you just keep on winning. Well, gambler's ruin is going to tell us that that's not the case, that the probability that you go broke is going to converge to 1 as soon as the expected value is negative.

So let's do this, for example, with a concrete thing. So if you ignore multiplayer betting games, like blackjack, if you just stick to the standard ones, all of the games in Vegas have negative expected value. But the one that has the least negative expected value is craps. Has anyone played craps before? No? OK, I'm talking to the wrong audience. That's good.

So craps is one where you roll dice. And then depending on whether it lands on black or red, you make different kinds of wagers. And there's basically a bet you can make that's very close to 50/50. So you have to get to that point in the game first. But the house edge, if you play that strategy in craps, is just 0.0141.

So we can think about a simplified version of making that bet in craps, that with some probability you're going to win $1. And what is that probability going to be? Well, it'll be 0.4929 because you want that the probability that you win-- the probability you lose minus the probability you win is equal to this house edge here. And then otherwise, you're going to lose $1.

And so what I claim is now that, by weak law of large numbers, well, you can look at Sn, which is just your cumulative winnings-- so you can add up in each game whether you win $1 or you lose $1. Now, the point is that that average is negative. So the probability that you're epsilon away from your average is actually going to 0 as you play more and more games. So on average, you really will be losing some amount of money per step. And you will not get that lucky with probability converging to 0.

So by the weak law of large numbers, except with vanishing probability-- something that gets swallowed up by the limit-- I claim the winning per round will go negative, which, of course, since it's an average over your rounds, if you play infinitely long, you will go broke. So this is called gambler's ruin.

And I want to give one more motivating example before we take a step back and I tell you about another probabilistic tool. So for us, so far, we've been talking about tail bounds as a way to get a handle on how likely or unlikely it is for a random variable to be far away from their expectation.

Now, it turns out that, really, the reason we care about these tail bounds isn't just for their own sake. But it really comes up in all kinds of applications in discrete math. So we'll see probability throughout some of the later units in class. But I want to give you a little bit of a hint about why we care about tail bounds, especially when it comes to algorithm design.

So let's talk about an algorithmic application. And this is a real application, something that gets used all the time. So let's say that we're given a randomized algorithm. And let's call it Aq. So A is the algorithm. q is the input.

And let's say this algorithm, this randomized algorithm, has the property that when we run on A on q, well, its answer isn't deterministic. It's a randomized algorithm. And let's assume that it gets the answer right more often than it gets it wrong. So let's say it outputs the correct answer with some probability. Let's just say 2/3. And otherwise, it's going to output the incorrect answer.

And the basic question we can ask is how many times we have to repeat this algorithm in order to get a good to boost its success probability. So one way to think about it-- our next unit after we do tail bounds-- is I'll be teaching you about modular arithmetic. And the basic thing we want to do there is-- imagine I'm given some giant number. A lot of properties of the factorization of this number are going to affect how it works when we do modular arithmetic.

And later on, much later in the class, we're going to talk about applications of modular arithmetic to things like cryptography-- so communicating securely with a public encryption scheme. These are things that get used all the time in e-commerce. But when I'm creating a scheme for communicating securely, I have to start off with some number that's the product of just two primes.

So what I can do is I can try and generate a random integer and hope that it's prime. And that turns out to work pretty well. You're reasonably likely to get a prime. But the trouble is that our best algorithms for determining whether or not a given integer is prime, they're randomized.

So maybe I'm running my cryptographic scheme. And I have some integer q. And I want to figure out if it's prime so that I can safely use it in the context of my secure communication. So you have an algorithm that, with 2/3 probability, tells you the right answer and otherwise tells you the wrong answer. But if you are wrong, there's a huge penalty to pay.

If you're wrong and the thing wasn't prime and you thought it was, then maybe you've just leaked all of the sensitive information, like your bank account, your social security number. So maybe this failure probability of 1/3 is just not good enough. What kind of failure probability would you be OK with your social security number leaking?

Just give me a number. 10 to the minus 30. OK. I was going to say 10 to the minus 10, but that's even better. You're even more paranoid than I am, which is a good thing. So let's say I want to get a really tiny failure probability, like 10 to the minus 10. So how many times should I repeat this algorithm to make its failure probability less than 10 to the minus 10 or maybe even 10 to the negative 30?

So if you think about it, this is really the setting of the weak law of large numbers. It's the same kind of thing. Each of my runs of this random variable is a random variable xi that tells me whether or not q is prime. And what I'm guaranteed is that, as I take the limit, as the number of repetitions goes to infinity, the average of the number of outputs that say that it's prime is going to converge to 2/3, if it really is prime. And the number that are going to say it's not prime is going to converge to 2/3 if it isn't.

So this is the setting of the weak law of large numbers. I'm just voting with runs of my algorithm. And we've already seen some quantitative bounds from the weak law of large numbers about how large n needs to be in order to get some valid epsilon. And if we think about it, if we just think about applying the weak law of large numbers, we're actually going to get very pessimistic bounds even here. So let's think about what the weak law of large numbers and Chebyshev would tell us.

So the weak law of large numbers slash our Chebyshev proof that gave us a quantitative bound-- well, it's going to tell us that the right-hand side, the probability of failure is bounded by the variance of x n over epsilon squared. And in our case, we can take x to be just whether or not the output of our random-- of our randomized algorithm is correct or not.

And you can compute that the variance of x is going to be 2/3 minus 4/9. So it's some constant. That's a reasonable constant. And what we can do is we can set epsilon to be something like 1/9. So x is supposed to tell us whether or not the thing is prime or not. We only have to set our tolerance really to split the difference between 2/3 and 1/3. That's not where any of the action is happening. These are all just constants.

But what's going on is that the probability that our average is off from the truth, we can bound it by something like 18 over n. Now, none of the other computations really matter too much. You can compute the variance. You can compute what epsilon makes sense.

But the crucial thing is really the fact that the n shows up in the denominator. So if I want my failure probability to be absolutely tiny, what kind of n would I have to take? I'd have to take it to be like 10 to the 10 or 10 to the 30. That would be ridiculous. So I want to make sure that I have vanishing probability of using something that violates security in my crypto scheme. And I just have to repeat this so many times that I can't really do a transaction online because I would just clog up the computer.

So this takes me to the last main topic. This is something which I want to introduce now. You're going to start to cover it in recitation. But this tail bound is not at all obvious how to prove it. And it'll be basically the full lecture on Thursday. But this was really just the motivation for it.

So you see that, like I told you, the theme for today's lecture was the more information you tell me about the random variables, the better I get for my tail bounds. So let's take a minute to think about the proof of this weak law of large numbers. So the setting for this was that I assumed that the xi's are independent draws from x.

But I claim my proof didn't really use full independence. I claim all it used was that they're pairwise independent, that if you look at any pair of them, they're independent. But I didn't use the fact that any triple or any quadruple of them are independent. So this requires a little bit of thinking. So where in my proof of the weak, law of large numbers did I actually use independence? It was only in this step right here.

And I didn't actually need that all of these random variables are mutually independent. I just needed that they're pairwise independent. That's exactly what got me the quantitative bound. But that's the weakness in applying Chebyshev, is that I'm not using as much information as I could hope to. Does that make sense?

So that's what leads me to the next bound, which is called the Chernoff bound, which is not just going to use pairwise independence, but it's going to use mutual independence. So let's state the Chernoff bound. This was actually proven at MIT, presumably in this building even, which is cool. So this was proven by our former colleague, Herman Chernoff.

And let's assume that x1 up to xn, let them be mutually independent. So this is a much stronger notion of independence, because I need all subsets of them to be independent from all other subsets. So let's say they're mutually independent random variables.

And let's just say that xi is equal to 1 or 0 with probability pi. And otherwise, it's equal to 0. Now, I'm going to do the same type of thing I did before. I'm going to look at some kind of average or sum of these random variables. So let me look at capital x, which is the sum from i equals 1 to little n of xi. And let me look at the expectation of x. Let me call that quantity mu.

But remember, that's the same thing as just the sum from i equals 1 to n of the pi's. So this is all just setup. I have mutually independent random variables, the x sub i's. In fact, they're Bernoulli random variables because they just take on the value 1 or 0. And I'm telling you that the probability that they're 1 is pi.

So this is already a bit of a departure from the weak law of large numbers because, in the weak law of large numbers, I assumed that all of these xi's were identical. They were all draws of this random variable x. Here, I'm actually allowing these to be different random variables. So that's even one extra layer of power.

But the real power just comes from the right-hand side. So let me write down one form of the Chernoff bound. So the probability that x is very far away from its average mu by some delta is at most 2 times e to the minus mu delta squared over 3. I think I need the average here. OK. No. OK.

All right, so this is the conclusion for the Chernoff bound. So the important thing is really that what I had on the right-hand side when I had Markov's bound was a kind of weak bound. It looked like 1 over C. So as you increase C, you just improve the right-hand side by the same factor. When I had Chebyshev, I had 1 over C squared on the other side. So the probability that I'm C away was degrading still at an inverse poly rate.

The crucial thing is that, here, I have an exponential rate. And this is what mutual independence buys me, is that the probability that you're far away from the average is just plummeting to 0 once you have this extra information about independence. So the Chernoff bound, there are many different forms for it. We're going to prove this next time.

But it turns out that, if we were to use Chernoff in our numerical example about testing out primes, well, we would only have to take n, our number of repetitions, to behave like some constant times the natural log of whatever we want to be in our failure probability over here on the right-hand side. So if we had 10 to the 10, really, the number of repetitions would be much tamer. It would just be the natural log of 10 to the 10. So even if I put a 10 or a 30 in the exponent, it's not such a giant overhead to actually drive the failure probability basically to 0.

So that's what we'll do next time. We'll do the Chernoff bound. But I want to cover one last topic in probability before we get to the Chernoff bound on Thursday. This will be something that's going to help you a lot for this week's assignment, which is a writing problem that involves using the probabilistic method. So it's something we've talked a little bit about on some of the P sets. But I just want to be explicit and show you some neat examples. So that will be our last topic for today, is the probabilistic method.

So you should think of this as like one more important application of probability tools. So far, we've talked about probability for its own sake, like looking at the deviation of random variables. We talked about probability for the sake of analyzing the failure probability of randomized algorithms. And we'll also be interested in using probability in the context of combinatorics.

It turns out that, in many applications, there are exotic types of combinatorial objects that the only way we know how to reason about why they exist is by using randomness. So this leads to a very important paradigm in discrete math, which is called the probabilistic method. You can take entire classes on using sophisticated tools to do things like this. But let me give you the high level bumper sticker version first before we talk about some concrete applications.

So what we'll be interested in is we'll be interested in constructing certain kinds of random objects or structures, which are exotic and have some pretty wild properties. And we're going to show that these things exist not by building them with our own hands, but by arguing that a random object succeeds.

So that's the high level idea. And there are many instantiations of this probabilistic method. It depends on what kinds of structures I want to prove exist. We'll talk about some concrete applications. And we'll see how the probabilistic method helps. But we're going to apply this to another important branch of applied mathematics, which is called Ramsey theory. So I'll introduce both of these topics right now.

So what's Ramsey theory? Well, let me give you a famous example of it. So in every party with at least six people, I claim that either we have the situation that there are three people all of whom are friends with each other, all pairs of these three people are friends with each other, or there are three people all of which are strangers, none of which knows the other person.

So Ramsey theory was actually first invented or realized in Hungary, where there was a Hungarian sociologist who realized that he was studying social groups of kids. And he realized that in every group of, I think, 20 kids, there were always four, a group of four people he could find who all knew each other or who all didn't know each other. And you might think that this is a sociological phenomenon. Maybe it has to do with the way kids play with each other, that they form some of interconnected groups and they don't know other groups.

But he consulted with his mathematician friends, including Erdos and others. And he realized that it wasn't a sociological phenomenon. It was just the property of graphs. So what this is really saying, like the same way that we can think about representing friendship relations with a graph, what we're really saying is that, on any graph with six nodes, no matter how I draw on the edges-- and the edges represent who's friends with who-- there must either be a triangle where all the edges are present. Or there's an anti triangle, where there's a set of three and none of the edges are present.

So this is the basics of Ramsey theory, is that we want to show that, in large enough objects, certain kinds of ordered substructures must appear because they always have to happen. There's no way of avoiding it. So in fact, there's a pretty famous and interesting example from pop culture. This happened like when I was younger. But in the mid-'90s, there was a lot of interest in something called the Bible codes. Has anyone heard of the Bible codes? No? OK. Yeah? You want to explain it?

STUDENT: I think it's like maybe you skip letters in the Bible. You make sentences or something.

ANKUR MOITRA: That's right. So basically, what happens, if you take the Hebrew translation-- if you take the Hebrew version of the Torah, what you can do is you look at evenly spaced characters in them. And if you choose the step size in the right way, you can find all these instances in the Torah of examples of words that cross with other words, which have historical significance. There would be examples of famous rabbis names with the dates in which they were later assassinated, like after the Torah was written.

So this was quite amazing because people let their imaginations run wild. They thought that maybe the Torah was predicting the future in arithmetic progressions in Hebrew. You could certainly write some bad movies about it. But it turned out that it was really just Ramsey theory in disguise. And the things that people found weren't statistically significant, that, in fact, you could find the same types of things in the Hebrew translation of *War and Peace.* So that was an interesting rejoinder to it.

But that's the kind of thing we're going to do with it today. Let me tell you about one example of structure that has to do with arithmetic progressions. And we're going to use the probabilistic method to prove types of lower bounds on these quantities, or at least very strong lower bounds.

So let me tell you the last topic for today is something called van der Waerden's theorem. So let me first give you a definition. So a k term, arithmetic progression, is just a very particular type of structured subset. So it's a set of integers of the following form. And I'll call it Q sub ab because I'm going to use it later.

But it's just the set of integers that looks like a, a plus b, a plus 2b, B all the way up to a plus k minus 1 b. So this is an arithmetic progression. You start off with some base number a. And then your steps are of size b. And then you go for k steps. So that's an arithmetic progression. This was the kind of thing that they were using in the Bible codes, because you would look at arithmetic progressions and concatenate the symbols you get at those positions together.

And so now, I can tell you this van der Waerden's theorem. We're not going to prove it. The proof is in the notes. But we're going to prove lower bounds for it. So here's a really cool statement. So for every k you give me, what I claim is that there is an n such that any two coloring of the integers 1, 2 up to n-- so before I get to the theorem statement, what's going on is I want to find a k term arithmetic progression.

And what I'm going to do is you're going to fix some n. And then once you look at the integers from 1, 2, all the way up to n, you're going to color these integers with either red or blue. And what I'm looking for is what's called a monochromatic arithmetic progression. So I want some arithmetic progression of this form, like Qab, but with the property that I look at the colors of the integers along the sequence, that they all should be the same color.

So this necessarily contains a monochromatic k term AP, arithmetic progression. So this is the statement. Again, there are a lot of quantifiers in this. So let's make sure we understand it. I want to find some k term arithmetic progression, one that's monochromatic. For whatever k I choose, there are some large enough n so that no matter how you two-color the integers from 1, 2 up to n, there's no way to avoid it. Somewhere in that set is sitting some kind of entirely red or entirely blue arithmetic progression.

So this is a bit of a mouthful. But this is one example of a much more interesting version of Ramsey theory, is that any large enough object, which, in this case, is a two-coloring of the integers from 1, 2 up to n, must necessarily contain this ordered substructure, which, in our case, is this monochromatic k term arithmetic progression. So this has a lot of quantifiers. Are there any questions about this? Make sense?

So that's van der Waerden's theorem. The proof of it's not too hard. But it just doesn't involve any probability. So we might cover it later. But it turns out that probability, especially using the probabilistic method, is a great way to prove lower bounds for this kind of theorem. So let's ask for quantitative things. Let's let W of k be the smallest n such that the property holds.

Now, you have to be careful and unpack this. W of k is the smallest n so that which property holds? The property that every two coloring of the integers from 1 up to n necessarily has a k term arithmetic progression. So this property has an extra quantifier because I'm not telling you what the two-coloring is.

Now, one of the things you can show that-- you can reason about small values of these types of quantities. So one of the concrete things you can do is you can ask what's the smallest n so that every two coloring has a three-term arithmetic progression. And it turns out to not be super easy to show this. But one side is easy.

So the easy part is just to show a lower bound. So we're going to prove that W of 3 is at least 9. And the way that you do this is you just construct a valid two-coloring that avoids having a three-term arithmetic progression. So for example, maybe all of these things I circle are red. And everything I don't circle is blue.

So you can check that this is a coloring of the integers from 1, 2 all the way up to 8. And you can just check, by trying out different choices of a and b, that there is no three-term arithmetic progression. The fact that everything that's larger than this is a lot harder to prove, because I would have to argue that when I add another number to this list, not just this coloring has a three-term arithmetic progression, but any other two coloring has a three-term arithmetic progression. So it involves a huge amount of brute force computer search to do these kinds of things.

But what we're going to do instead is we're going to prove the following theorem, which is a beautiful theorem in its own right, which is we're going to show nice lower bounds on Wk that are going to kick in for large k. So they won't be tight for small values. But they'll give us a good handle on how large the n needs to be in this van der Waerden's theorem.

So here's the theorem we're going to prove in what remains. What I claim is that, for any k, W of k, the smallest n such that this property holds is at least square root of k minus 1 times 2 to the k minus 1 over 2. So that's the thing we're going to prove.

And the crucial thing about this, which I really want to explain, is the strategy. What we're going to show is that a random coloring for the choice of n equal to this has positive probability, probability greater than 0, of having no k term AP.

So now, this is making good on some of the promises I made you. See, the whole point of the probabilistic method is that we want to show that certain things exist, even if it's hard to construct them by hand. So in order to prove this theorem, really, I have to choose some n that's equal to the right-hand side, that's equal to the square root of k minus 1 times 2 to the k minus 1 over 2.

If I choose this as my n, then the thing I have to construct and build is what is the valid two-color. That's very difficult. I could try some things. They might or might not work. And then I might get stuck. But it turns out that the way we're going to show that there is a coloring that avoids k term arithmetic progressions is just by showing that a random one works with positive probability. So that's what the probabilistic method is.

And let's execute this strategy. So just to make sure we're on the same page, what I'm going to do is I'm going to consider a random variable x that's going to be equal to the number of monochromatic k term APs. I'm just going to count. So you give me a random coloring. The dumbest thing we could do is we randomly decide independently for each integer whether it's red or blue. And we're going to show that that works. And we're going to be interested in this random variable x that's the number of monochromatic k term APs in my random coloring.

Now, the basis for making this strategy make sense is really this very simple lemma that's going to remind you of things we did earlier in lecture. What claim-- and this is just a general statement, but it'll explain why this strategy makes sense in the first place. Suppose x is a non-negative random variable. And let's say it's an integer valued random variable. And let's say that the expectation of x is less than 1.

Then what I claim is that the probability that x equals 0 is strictly positive. So this is the basis for our strategy, because the way we're going to show that a good two-coloring exists is we're going to look at this random variable x that's non-negative. It counts the number of monochromatic k term APs. It's integer valued.

And as soon as we can show its expectation is less than 1, that means that the probability that it's actually equal to 0 and there's no k term apes is strictly positive. So this is the basis for our strategy. So let's prove this lemma. This lemma is super simple. But then let's execute using the strategy in the case of van der Waerden's theorem.

So we can just write out the expectation. So the expectation of x, because it's integer valued, is the sum from i equals 0 to infinity of the probability that x equals i times i. This is lower bounded by the sum from i equals 1 to infinity. So I've dropped the first term. And then I'm going to drop the i because each of these i's is at least 1. And this quantity right here on the right-hand side is just 1 minus the probability that x equals 0 by definition.

So now, if I rearrange this, I'm going to get that the probability of x being equal to 0 is lower bounded by 1 minus the expectation of x. And the expectation of x is less than 1. So that simple proof is the basis for our strategy, because all we have to do now is we just have to compute what the expectation of x is.

I told you what x represents, what random variable it is. I told you the fact that we're doing a random coloring. And I told you what the strategy is of how to reason about the probability that x is 0. So all that's left is really just the compute this expectation of x and hope that it's less than 1.

So let's just compute expectation of x. So I'm going to let Iab denote the relevant indicator variables. It'll be 1 if Qab is monochromatic. And otherwise, it'll be 0 because these are just the events that are going to help me get a handle on what my random variable x is. x is just summing up all of these indicator variables.

So in particular, then x, the thing I want to keep track of, is the sum over all ab pairs such that Qab is a subset of 1, 2 up to n-- those are the only arithmetic progressions I'll consider-- of this quantity Iab. And now, we're good because what I claim is that the expectation of Iab, the probability that this event happens, that this arithmetic progression Qab is monochromatic-- I claim that this is very simple to compute. It's equal to 1 over 2 to the k minus 1.

So just to make sure we're on the same page, so Iab is the indicator event for Qab being monochromatic. Remember, Q is a k term arithmetic progression. So why is this probability true? What's the one line proof of this? I'm fixing a particular arithmetic progression. Why is the chance that it's monochromatic going down exponentially with the length of the arithmetic progression? Who can help me out? Yeah?

STUDENT: Choose the color of each term independently with probability 1/2, and then multiply by 2 because you have 2 [INAUDIBLE].

ANKUR MOITRA: Perfect. That's exactly right. So the proof of this is exactly what he said, which is that once we pick the color, the first color-- I'm just rephrasing your proof slightly. So we just pick the first color, namely the color of a, the first term that shows up in our arithmetic progression. The only choice-- let's say a was red. Then a plus b had better be red. a plus 2b had better be red, a plus 3b, and so on.

Then all the other colors must match. All the other colors in Qab must match. And that's the proof because there are k minus 1 other things. So now, we're in good shape. We just have one very last ingredient. And then we'll get our bound to follow, which is I have to tell you how many Qabs there are.

Lastly, I claim. the number of valid ab pairs which induce an arithmetic progression, which stays in the integers from 1 to n, is at most n times n over k minus 1. So the way to think about it-- remember, our arithmetic progression starts at a. And then it goes a plus b and a plus 2b all the way up to a plus k minus 1 times B.

So in order for this arithmetic progression to stay in the integers from 1 to n, I better choose my a to be at most n. Otherwise, it's not going to be in the set. And then my b would better be at most n over k minus 1 because I'm going to add b k minus 1 times. So if it were larger than this, I would fall outside the set.

So putting this all together, we know that the expectation of x is at most n squared over k minus 1-- that's the number of valid things-- times the expectation of each of these Iab's, which is 1 over 2 to the k minus 1. And now, you can check that that implies that if n is less than this quantity I told you, k minus 1 times 2 to the k minus 1 over 2, then it implies that the expectation of x is less than 1. So that's our proof.

So this is a very beautiful but subtle proof because of the way we're using probability. I want to show that W of k is large because I want to construct a valid two-coloring that avoids all arithmetic progressions. What I'm going to do is I'm going to reason about the random variable that counts the number of k term APs as a way to argue that there is a good two-coloring to begin with. And the rest of it is a computation using linearity of expectation. But that allows us to prove that these exotic objects, which are hard to actually build and construct, must exist even if we can't find them.

So that's one of the cool things about the probabilistic method. Sometimes, you can show very exotic things exist, even though we don't know how to build them from scratch. And we'll do some more applications of this later, too, and next time. And we'll also cover the Chernoff bound. So we'll stop there. And I'll see you guys on Thursday.