GILBERT STRANG: OK, so basically probability ideas today, because that's a part of the subject, part of deep learning as we get there. And it's probably a good topic for the day before spring break, because lots of you will have seen-- of course, you will have seen the sample mean, the average of the data. And you'll know about the expected mean.

Let me complete that. What's the expected mean? So this is the expectation of the value x where we get x1 with probability P1 along to xn with probability Pn. So we just want to say, what's our average output-- average outcome? And we weight it by their probabilities. So it's P1 x1 plus so on plus Pn xn. So that's the expected value of x.

Are you comfortable with that symbol E? Because that's like everywhere. It gives a handy shorthand.

For example, the variance is the expected value of what? The variance is an expected value based on these probabilities of the square of the distance. So everybody remembers it involves square. And it's the distance from the mean.

Let me call this, say, m, maybe, just to have a smaller letter. So it's the distance from the mean minus m. It's the expected value, the average value of x minus m squared.

And in general, of course, the expected-- this covariance matrix I could express with that E notation. But let me just stretch it as far as what would be the expected value of any function of x?

Well, we've got n possible outputs, x1 to xn. We look at f of x1 up to f of xn. We weight those by the probabilities that they happen.

So this would be-- let me make just a little corner for this E letter-- so this would be the probability that that is f of x1 times the value of f of x1. So this is the contribution from the x1 possibility.

And now, we include them all. So it will be output f of xn with probability Pn. And if that f of x is x minus m squared, then we get what we expect. And let me remember that.

So I just want to keep going with variance. So it's the sum. It's the first probability times the first output minus the mean squared the last probability times that last output xn minus m squared.

And everybody should know a second expression, a second way, to do that the sum. If I just write out those squares and combine them a little differently, I get a second expression which is really useful, often a little faster to compute. So can I just do that?

So that's x1 squared minus 2 x1 m1 plus m1 squared. And then same thing here, Pn times xn squared minus 2xn m plus m squared. Good with that?

AUDIENCE: On that m--

GILBERT STRANG: Sorry?

AUDIENCE: On that m--

GILBERT STRANG: Plus n, oh, sorry. No, I mean-- am I--

AUDIENCE: So for P1 it's x1 squared minus--

GILBERT STRANG: Oh, it's just an m. Correct. Thank you. Thank you. Just an m. Good. OK.

Can we take that sum? So I get P1 x1 squared if I take these, the first guys. So I've accounted for this and this. Now, I'll take minus 2 P1 x1 m. So P1 x1 m plus Pn xn m. I'm just writing it all out, and I'm going to recombine it.

So now, I have P1 m squared plus P2 m squared plus Pn m squared. So what do I have from the P1 m squared all the way up to Pn m squared? Are you with me? So m squared is in every term. So I'm going to have an m squared.

And what's it multiplied by? P1 here, P2 here, Pn here. I add those up and I get?

AUDIENCE: 1.

GILBERT STRANG: 1. So that's it. OK, now, I'll just simplify this thing. So this is really the expected value of what? What am I seeing in this term?

AUDIENCE: x squared.

GILBERT STRANG: The expected value of x squared, right. Different from the expected value of x minus m squared, of course. This is just a first term from here.

But, now, what do I get for this second term? Well, the point is that m comes out. So this is minus an m and a 2.

And what do I have left? So I've used up the m. I've used up the 2. P1 x1 dot dot dot Pn xn, what's that? Everybody should just pay attention to this. Trivial, I mean, we're just doing high school algebra here. But P1 x1 up to Pn xn is m.

So I have another m, m squared there. And I have a plus m squared from the n. So you see that it is another expression, the expected value of x squared minus m squared. It's just algebra.

That is the same as this. So that if you happen to have a handy way to compute the expected value of x squared, you would just subtract m squared. And you'd have the same as this. Yeah, it's just algebra.

OK, let's go a little deeper with something here-- if I can find it. There are two great inequalities in statistics. And the first one is due to Markov. And I don't know if you know Markov's inequality. It comes out easily, in fact, too easily.

I'm kind of happy to discuss him. And now I've jumped to Section 5 of the book. So I'll need to post Section 5, which is probability and statistics. And you'll see this Markov inequality.

So it just involves this stuff. So that's why I'll go do it now. Markov's inequality.

He was a great Russian mathematician, oh, probably about 1900. And we will see Markov chains and Markov processes, that's beautiful linear algebra. But this little inequality is not matrices. It's just playing with these.

And it applies to non-negative events. So shall I say applies when all the x, all the outputs, are greater than or equal to zero. So I'm going to use that fact.

So it doesn't apply to something like a Gaussian, because there, the Gaussian, the outputs go all the way from minus infinity to infinity. It does apply to a lot of important ones and simple ones.

I'll give you the proof for this finite probability. And there will be a similar proof, similar discussion everywhere here for continuous probability.

So what does Markov say? Let me be sure I get it right, because I'm not a pro at this. It's natural to want to estimate the probability that x is greater or equal to some number a. Get some idea of what's the probability of x being greater or equal to a.

So what do we know? This is certainly a number between 0 and 1. That number is going to get smaller as a increases, because we're going to be asking for more. If I take a to b, say, twice the mean, can I estimate what that probability could be. And that's what Markov has done.

He says the probability of that is at least-- at most-- sorry. Let's see I used to have an eraser-- at least-- sorry, at most-- yes, got it, got it-- is less or equal to the mean-- x bar is another way to write the mean-- divided by a. And this is the mean over a or it's the expected value of x over a. We could see any of those notations. OK.

And as we expect, as a increases, the probability, this number, goes down, the probability goes down of exceeding a. So that's a pretty simple estimate to get this probability just in terms of the number a, which has to come in, because it's part of the question, and the mean, x.

So let me take an example as a equals 3. For example, suppose a with 3. I want to show that the probability of x being greater than or equal to 3. Yeah, OK. We don't have many facts to work with. So if we write those down, we should see the reason.

So I know that the mean is E of x. So let's see, am I going to take-- yeah, for example, let's take the mean to be 1. So I'm going to imagine that the mean is 1 and that I'm asking for what's the chance that x will be bigger than 3. And I'll get an estimate of 1/3.

So I'm trying to show that the probability of x greater or equal 3 is less or equal to-- the mean, I'm saying is 1. a is 3. So it is less than or equal 1/3.

Now, why is that true? That's what I have to show. I think that if I write down what I know, I'll see it.

So let me just raise that a little so that I have room to write.

So what do I know? I know the definition of the mean. So I know that x1 times P1 plus x2 P2 plus x3 P3-- allow me to get carried away here-- x5 P5, say, is what? So what I've written down there is the mean. And I'm assuming that to be 1. So this is the fact that I know is 1.

And what is it that I want to prove? I want to know the probability of being greater or equal 3. So what's the probability that the result will be greater or equal 3? It's P3.

So this is saying that P3 plus P4-- these are the probabilities. These are the different ways that I might be greater or equal 3. And I'm claiming that that's less than or equal 1/3.

What I liked about this elementary approach is that I've stated these facts, these probability assumptions and conclusions directly in terms of numbers. So I just want to show that if this is true, then that's true.

Let's see, I guess I'm thinking that the-- I'm sorry, I even took a more special case. I'm taking the case where x1 is 1, x2 is 2, x3 is 3, x4 is 4, and x5 is 5. So that satisfies my condition that the outputs-- 1, 2, 3, 4, or 5-- are all-- Markov only applies when they're all greater or equal 0.

So I'm just imagining the special case where that possible outputs are 1, 2, 3, 4, 5. Their probabilities are P1, P2, P3, P4, P5. The mean is 1. And what I want to show is that the probability of being greater than 3 is less than or equal 1/3. And can you put together these two? Given this, we want to conclude that.

Let me just step back a minute. So what do we know here? We know this, the first line. And we want to prove the second.

We know one more thing. All the probability-- well, we know the probabilities add to 1. And we know they're all greater equal 0. So let me put those facts in here too.

We know that P1 plus P2 plus P3 plus P4 plus P5 is 1. That we know. And we also know that all the P's are greater than or equal to 0. OK.

So here we go. My idea is this is looking at 3 times P3 plus P4 plus P5. So I'm going to take 3 P3 plus 3 P4 away from this. So this we'll say P1 plus 2 P2 plus 3 of P3 plus P4 plus P5. I'm just picking out three of those guys. Plus I have one more P4 to account for and two more P5s gives 1.

Good?

Now, this is what I'm trying to prove. So that is here. I'm trying to prove that this thing is-- what am I trying to prove about that number? Sorry, I'm talking a lot. But now, I've really come to the point. What is Markov telling me about that number? That's--

AUDIENCE: Less than or equal to 1.

GILBERT STRANG: That is less than or equal to?

AUDIENCE: 1.

GILBERT STRANG: Thanks. OK, I'm trying to prove that this is less than or equal 1. That's what Markov tells me. But suppose it was greater than 1? Do you see the problem? Do you see why it can't be greater than 1? Because why?

AUDIENCE: All are the other terms--

GILBERT STRANG: All the other terms are greater equal 0. Probabilities or greater equal 0. These are all greater equal 0. And the total things adds to 1. So that this piece has to be less or equal to 1. That's right. That's it.

So a lot of talking there. Simple idea. And you'll see exactly this example written down in the notes. And then you'll see a more conventional proof of Markov's inequality by taking simple inequality steps. But they're somehow more mysterious. For me, this was explicit. OK, so that's Markov.

Chebyshev is the other great Russian probabilist of the time. And he gets his inequality. So there are the two. There's Markov's equality. Let me write it down again what it was. Here was Markov's inequality and Markov's assumption.

Chebyshev doesn't make that assumption. So now, no assumption of that the outputs are greater equal 0. Doesn't come in.

Now what is Chebyshev trying to estimate? OK, let's move to Chebyshev. And that's the last guy.

So Chebyshev was interested in the probability that x minus the mean, m-- can I use that for

mean-- is probably greater equal to a-- the probability of being sort of a distance a away from mean. So again, as a increases, I'm asking more, I'm asking it to be further away from the mean, and the probability will drop. And then the question is can we estimate this?

So this is a different estimate. But it's similar question. And what Chebyshev's answer for this?

So this is the probability of this. I have to put off big-- that's all one mouthful-- the probability that this x minus m is greater equal to a. And again, we're going to have is less than or equal to sigma squared now comes in over a squared.

So that's Chebyshev. And I just take time today to do these two because they involve analysis. They're basic tools. They're sort of the first thing you think of if you're trying to estimate a probability.

Does it fit Markov? And Markov only applies-- so I'll put only applies-- when the x's are all greater or equal 0. Here, does it fit Chebyshev? And now we're taking absolute values. So we're not concerned about the size of x. And we're taking a distance from m. So we're obviously in the world of variances. We're distances from m.

And the proof of Chebyshev comes directly from Markov. So I'm going to apply Markov-- so good thing that Markov came first-- to-- now let me just say this right-- to a new, let me call it, y-- this will be a new output. And it will be x minus m squared. Of course, with the same probability. So yi is xi minus m, the mean, squared. And the same probability, same probabilities Pi.

So I guess if I'm going to apply-- I'm just going to take the y's here instead of the x's here and then apply Markov.

So what is x bar? So if I want to apply Markov, I have to figure out the mean of x. Over here, I have to figure out the mean of y. What is the mean of y? The mean value, the sum of probabilities times y's.

You're supposed to recognize it. This is the sum of probabilities. And my y's are the xi minus the mean squared. So this is the mean for this y thing that I've brought in has that formula. And we recognize what that quantity is.

That is? That's sigma squared, sigma squared for the original x's. So that's great. So the mean is sigma-- is the old sigma squared. Those are exclamation marks.

Do see that now Chebyshev is looking like Markov? Over here will be the x-- now I want the expected value of y over the-- let's see, yeah, so the expected y is going to be that.

And now what do I have to divide by? I want to know probability of this thing being bigger than a. But now I'm looking at the y's. So the probability of if x minus m is greater than or equal to a, then x minus xi minus m squared is greater equal a squared.

So my a over here for x is now turning, in this problem where I'm looking at probability greater equal a but squaring it, this is the a squared. So that's Markov applied to y.

Here is Markov applied to x. And x had to be greater equal 0. So over here, Chebyshev took a y, which was greater equal to 0 than just applied Markov and recognize that mean of his variable, x minus m squared was exactly sigma squared. And it fell out.

So again, here is a very simple proof for Markov. And then everybody agrees that Chebyshev follows right away from Markov. So those are two basic inequalities.

Now, the other topic that I wanted to deal with was covariance, covariance matrix. You have to get comfortable with what's the covariance.

So covariance, covariance matrix, and it will be m by m when I have m experiments at once. And let me take m equal to 2. You'll see everything for m equal to 2. So we're expecting to get a 2 by 2 matrix.

And what are we starting with? We start we're doing two experiments at once. So we have two outputs, an x and a y. So the x's are the outputs from the x experiment. The y's are the output from a second experiment. We're flipping two coins.

So let's take that example, two coins. Coin 1 gets 0 or 1 with P equal to the probability 1/2. Coin 2, 0 or 1 with probability 1/2. So they're fair coins.

But what I haven't said is, is there a connection between the output-- this is the x. This is the y-- if I glue the coins together, then the two outputs are the same. I think for me this is a model question that brings out the main point of covariance. If I flipped two coins separately, quite independently, then I don't know more about y from knowing x. If I know the answer to one flip, it doesn't tell me anything about the second if they're independent, uncorrelated.

But if the two coins are glued together, then heads will come up for both coins. I'll only have

two possibilities. It'll be heads heads or tails tails. Let me let me write down those two different scenarios.

So unglued-- I never expected to write that word in a math class-- unglued. And what am I going to write down? I'm going to write down a matrix with heads, tails for coin 1, and heads and tails for coin 2.

So the possibilities are coin one get heads and coin 2 gets heads. What's the probability of that? This is the unglued case. So I'm going to create a little probability matrix of joint probabilities. That's really the key word that I'm discussing here-- joint probability.

So let's complete that matrix. So I have unglued coins, independent coins. I flip them both. What is the chances of getting heads on both? 1/4.

What are the chances of-- what do I put in here? This means heads on the first coin and tails on this second coin. And the probability of that is? 1/4. And 1/4 here and 1/4 here.

So I've got four possibilities, which I put into a 2 by 2 matrix, instead of a long vector. My four possibilities are heads heads, heads tails, tails heads, and tails tails. And they have equal probability-- 1/4.

But now, if they're glued, heads and tails on the first coin, heads and tails on the second coin, now what do I put in there? So the two coins are glued together. What is the chance that they both come up heads? 1/2. Because if one comes up heads, the other one is glued to it. It will also.

What's the probability of heads tails, heads on one, tails on the other, is of course?

AUDIENCE: Zero.

GILBERT STRANG: Zero, thanks. And here, zero. And here, 1/2. So what I've created are those two setups, two different scenarios of unglued and glued. But each experimental setup has its matrix of joint probabilities.

That's the thing that there are four numbers here, four numbers. We have all possibilities. We have any possible x and at the same time any possible y.

So suppose we were running three experiments. So what would be the what would be the situation if I was running three experiments with three independent, fair coins. I'd be in this

unglued picture. But I would have three different experiments that I'm running. Then what would I be looking at then? Just the whole idea is to see what is this like joint probability.

So suppose I have three coins unglued. Then I want to know like the probability of getting heads on the first, heads on the second, heads on the third. Will be what? Just give me that number. What will be the probability that all three of them independently come up heads?

AUDIENCE: 1/8.

GILBERT STRANG: 1/8, OK. But now my question is-- so what do I have? Then I have probability of heads, of, say, tails heads heads. I've got three indices here and eventually down the probability of tails tails tails. Everybody sees that the numbers are going to be 1/8.

But where do those fit in? They don't fit in a matrix, because I've got 3 indices here. So I guess what we're seeing, I sort of realized today, we're seeing for the first time a tensor.

A tensor is a three-way structure, three-way matrix you could say. So I guess I think of this instead of a square like that, an ordinary matrix, I have to think of a cube, right? I have a cube with two rows, two columns, and two whatevers. Layer, somebody might say layers for that.

You see that the matrix has become a three-way thing, a tensor. And the entries in that tensor-- so it's 2 by 2 by 2. But instead of m by n for a matrix, I have to give you the number of rows, the number of columns, and the number of layers going into the board. So rows going one way-- you know, columns, rows, and then layers are going in deep.

So it will have eight entries. And, of course, in this simple case each will be 1/8 in that unglued, totally independent way.

But then you can imagine some dependence. So what would happen if I glued coins 1 and 3? I would still have a tensor, still have a 2 by 2 by 2 tensor of all the possibilities. But some of those are going to have probability zero, the joint probability. If I've glued coin 1 to coin 3, then the probability of jointly seeing heads on one, whatever on two, tails on 3 will be 0, right? Because that can't happen if I've glued coins 1 and 3.

So I'll have eight entries in here. This is the unglued case. And then I could have a case where two coins are glued. And as I say, I think I'd 1/4 four times probably.

And then if I had any spare glue, I glue all three coins. I flip that stuck together thing, and I

never get heads tails heads. Only possibilities I get our heads heads heads and tails, tails, tails, because they're glued together.

So what would be the situation for three coins glued? What will be the entries in the in the matrix of joint probabilities? What will be the joint probability? So the probability of heads heads heads, of seeing heads from all three will be?

AUDIENCE: 1/2.

GILBERT STRANG: 1/2. 1/2. Because I'm flipping this heavy mix of three coins together, I get 1/2 twice. Actually, this is a good introduction to tensors in a way, because the first step in understanding tensor is to think of three-way matrices, think of three-way things. We just haven't done that. And now we have to do it.

And, of course, four-way or n-way tensors are now understood. So these are tensors with very special, simple things.

So now, I have to say, what is the covariance matrix? What's the covariance matrix? So now I'm ready for that, and I'll put it here. That's the final topic for today.

So the covariance matrix. So I'm saying matrix, because I'm just going to have-- yeah, the covariance matrix. Yeah, it's going to be 2 by 2 for two coins. So it is a matrix. For three coins, it will be 3 by 3. But it is a matrix.

So how is it defined? And what am I going to call it? I think I'll call v, because really the key ideas variance. Covariance is telling us that we're also interested in the joint outcome, an x and y. So it's a variance.

So I'm going to add up over all plausible i and j-- sorry, all possible outcomes. Yeah, that's not right. All possible xi and yj. So I'm running these two experiments at the same time. From experiment 1, the output's an x. From experiment 2, the output's a y.

Then what is Pij? What does that symbol mean? That's the guy in our 2 by 2 matrix, like that one or that one, depending on the gluing or not gluing. So Pij, let me say what it is. This is the probability that x is xi and that the second output that this is yj.

Let me give you a second example to keep in mind. Suppose I'm looking at age and height. So suppose x is the age of the sample, the person. And y is the height. I want to know so what

fraction have a certain age and a certain height. I'm looking at every pair, age and height. Age 11, height 4 feet. Age 12, height 5 feet. Age 11, height 5 feet. Each combination. So Pij is the probability that these will both happen, both.

I'm going to add more to it here. But that joint probability is really important. So I'm going to ask you more about that.

Suppose that I take these Pijs and that I add up P1j plus P2j plus P3j. In other words, I sum the Pijs over i. So I'm looking at a row, row i, of my matrix.

So let me ask the question. Maybe I have to put it somewhere else. What's the meaning of the sum of Pij over the i's? What does that quantity look like?

So that's a probability. Pij is the probability of getting a certain i and a certain j. But now I'm including all the i's. So what am I seeing there?

AUDIENCE: Pj.

GILBERT STRANG: Pj. Thanks. Pj. This is Pj. That's the probability of seeing j in the second guy, because I had to see something. If I see j in the second one, I'm allowed to see anything here in the first one. But I'm adding those all up. So that's the point.

Those would be called the marginals. In my matrices, I would be adding up along a row or adding up down my column. Those are called the marginals of the joint probabilities. So the marginals would be the individual probabilities, Pi and Pj, in the case of two experiments going on at the same time.

Yeah, it's just like new ideas. Everything today has been sort of straightforward. But it's different.

OK, now, I'm going to complete the definition of this covariance matrix. So it's going to be-- I want to have a square. So it's going to be-- and it should be this is between x and the mean of x. Mean 1 I could call it or mean of x. And y, the distance from the mean of y times-- so it's going to be column times row-- same x minus the mean of x, y minus the mean of y. xi. yj.

Can you look at this formula? So this is with two experiments, two coins, two experiments. I get a 2 by 2 matrix. Everybody sees that. A column times a row, 2 by 2 matrix.

And let's just see what would be the 1, 1 entry in that matrix. This is the covariance matrix. So

what is the 1, 1 entry in that matrix? So the 1, 1 entry is coming from that times that, which is that thing squared, times all the Pijs. Add them up. What do you think I get for that? I get the variance of the x experiment, the standard variance of the x experiment, so v-- I have to tell you what v is now.

V, this is V. This is a 2 by 2 matrix. Up here, I get the variance of the x experiment. What do I get down here? The variance of the y experiment by itself. Because it's y's times y's, it gives me that 2, 2 entry. So this is just sigma y squared,

But the novelty is the 1, 2, and it will be symmetric. So it's a symmetric matrix. This is V transpose. I just have to see here.

So I've Pij times the distance of this guy times this guy. That's what's going to show up in the 1, 2 position. It'll be in row 1. It'll be in column 2. It's the distances.

So what will it be in the case of unglued coins, independent coins? Zero. I mean, it's just feeling like 0. I haven't done the computation. But I know that when I have independent experiments, then this covariance, which everybody would write as sigma xy-- and it's the same here. It's symmetric, sigma yx if you like.

So those subscripts are telling me that the sum of the P's, joint probabilities, times the distance of x from its means, the distance of y from its main, added up over all the possibilities. So the case of unglued coins, the case of independent ones, in that case, those are 0. Maybe worth just writing that out. You would get 0. So you have a diagonal matrix.

The diagonal matrix is just separate variances, because that's all the two independent-- the experiments are independent. So all you can really expect-- information is sigma x squared and sigma y squared.

But if the two coins are glued together, then what? If the two coins are glued together-- well, let me just say because time is up. This matrix will be singular. If the two coins were glued together, the determinant would be 0 here. The sigma xy in the glued case would be-- squared-- would be the same as sigma x squared sigma y squared.

Actually, we're probably getting all these 1/4s. And that would make sense.

I'll just end with this statement. This matrix is positive semidefinite always. Positive semidefinite always. Because it's column times row, we know that's positive semidefinite. And it's multiplied

by numbers greater or equal 0. So it's a combination of rank 1 positive semidefinite definite. So it's positive semidefinite definite. Or positive definite. It's certainly positive definite in the independent case when it's diagonal.

And the totally dependent case, when the coins are completely stuck together, that will be the semidefinite case when these entries would all be the same actually. So that's a first look at covariance matrices. It brought in tensors. It brought in joint probabilities. It brought in column times row. It kept symmetry. And we recognized positive definite or positive semidefinite definite.

So in between, coins that were partly glued, partly independent, but not completely independent experiments, then this number would be smaller. This wouldn't be 0, but it would be smaller than these numbers. I've run four minutes over. You're very kind to stay. So have a wonderful break. And I'll see you a week from Monday. Good. Thanks.