GILBERT STRANG: So I've mentioned randomized linear algebra a few times, and I thought, OK, I'm going to jump in and describe randomized matrix multiplication. It's a pretty cool idea, it seems to me. So this is a topic within randomized linear algebra.

And when would we be doing any of this? It would be for matrices that are just really, really large. So we plan to sample the columns of A and sample the corresponding rows of B, so actually that when we decide on a column, we've also decided on a row.

So we're taking those pieces, which do correctly add up to AB, but we're not going to take them all. We're going to take different ones, randomly sampled with probabilities-- we have to decide probabilities-- and then we'll add up our samples, and we hope that the result is close to AB. That's the idea.

OK, so this lecture then, so I wrote these pages about six months ago. So I've been desperately trying to remember what I wrote, because it's not a subject I have ever spoken about before, but it's so neat. So here are some of the things that come into it.

We have to decide on probabilities. Then we want to compute the mean. So it's our first day with some of these key ideas from statistics and probability. So we're going to take probabilities that add to 1. We're going to figure out what's the mean value of our random AB. We hope, and we will see that the mean value of the random AB is correct AB.

But there will be a variance. Every sample won't be correct. In fact, no samples will be correct. Only when we add them up on the average, they're correct, and we get to correct AB. So the mean will come out right. It will come out as AB. You'll see it happen.

Correct. But there's a big variance, not zero. We'll be all over the place with our samples. They'll just average out right, but they'll be all over. No particular sample will be right at all. So then we want to pick the best probabilities. So our job will be, once we see how the system works, we're going to assign probabilities. And we're going to choose the probabilities that minimize the variance.

So this is a typical situation where the mean is pretty straightforward and does what you want, but having the correct mean does not mean you've got good answers at all. And the average of, you know, like minus 100 and 100 might be the correct answer is zero, but you're way off. And this measures how far you are.

So I don't know if you know these words. It's unfortunate, but I guess 18.065 can't change it now, that the variance is written sigma squared. And we already have a good use for the Greek letter sigma, but today it has a different use for variance. And this-- Lagrange multipliers will come in near the end.

So basically, let me do a practice example to recall what mean and variance and what those are about. So let me take a matrix that's 1 by 2. So my matrix is just going to be a b. OK, I'm going to sample that twice, and my rule for the two samples will be the same. They will be a identically distributed, totally identical. And what's my rule going to be? So this is like practice.

My rule is going to be I take that column or that column with probabilities I have. And I do it twice. And I take the average. So I'm going to take probabilities are going to be 1/2, 1/2 for the two columns. And I'm going to do s equals to 2 samples.

And I'm going to add-- I'll weight them with-- and I'll take the average of the two samples. OK. And that will be my randomized matrix. OK, so could we compute the mean for the-- so I've described a randomized sampling process. I've given you the probabilities, 1/2 and 1/2, the number of times I'm going to do it, and then I divide by that number of times.

So this is really-- I have a 1 over s here, because I've got s of these. And now-- so what are the possibilities here? I want to find the mean. First of all, let's practice with the mean. OK, so here are two-- I could think of two different ways to compute the mean. Let me start with this one. What is the mean, the average value-- mean means average value-- of the first sample?

So the average value of the first sample, I would-- what is the mean, the general formula is you add up all the sample times it's-- the possible samples times their probabilities. And in this case, the probabilities are 1/2 that the sample is a, 0 and 1/2 that the sample is 0, b. So those are my two samples.

And computing the mean of the total, I get-- but then mean for each sample, but then I have to multiply, so let's put what I got here 1/2 of a, b. That was the meaning of each sample, because my probabilities were equal, 1/2 and 1/2.

And now, I've got s equal 2 samples. So I multiply by 2, and I get a, b as the mean. Mean is correct. Good. We did the easy one, the mean. Now, practice with a variance, or else quit here. Maybe I should quit while I'm ahead. I've got the mean exactly right, but of course, the samples might not be right.

So now for the variance. So what is variance? Do you remember that? There are actually two ways to compute variance. Let me just remember those over here and push that board up. So the variance sigma squared. Forgive me, if you're a statistician, this is like you were born knowing this. But the rest of us, we're not.

So the variance is the sum-- one way to do it is add up the different probabilities of different things that could happen of output minus the mean squared. So it's the average-- it's the average distance squared from the mean.

So it takes whatever output that came with output number i, minus the mean, which is the average output. I square those, and I get a number. And that sort of tells me how-- it tells me like in the famous Gaussian, if I had a Gaussian distribution here, I have a distribution of 1/2 and 1/2. So like that maybe even has a name, like binomial or something or Bernoulli or whatever.

But here on this Gaussian that we all remember, can I mark what in that figure where the mean is? Right in the center. OK. Mean. And what is the variance? Just to recall what everybody in the first time may even hear that word variance, what is the variance kind of measuring? You're summing squares, so whether you're on the right of the mean or the left of the mean, no difference, because you're squaring it. And it's the distance.

The variance would be sort of like a typical width. Maybe I overdid it. But that would be a sort of typical sigma. I'm really just-- since the words statistics, mean, and variance haven't been mentioned in 18.065 until today, I'm just kind of recalling.

OK, so now I'm prepared to compute this example. OK, maybe I'll-- maybe I'll compute it over here. OK, so shall I compute the variance for each sample, and then I'll multiply by 2, because I have two samples. So what are they-- so this is the sigma squared sample.

Obviously, I could write down all the possibilities. Yeah, let me just do the sigma. So the sample could either have picked out a, 0 or 0, b. And the probabilities were a 1/2. So I have 1/2 times the probability times the output. Let's say the output is a, 0 minus the mean, which was a over 2, b over 2. And I want to square that.

So that was one possibility when I picked a, 0, and the other one, which I'm also doing with probability 1/2, is in case I picked 0, b, what was-- you see, I'm not getting 0 for the variance, because I'm making an error every time. I'm never getting the correct a, 0 or the correct 0, b, because I'm always doing this one in the middle.

Now, if I compute all that, I get a quantity, and maybe I'll just, to be on the safe side, ask your forgiveness. If I write the answer. And we could even try to get the answer, but-- so this is from two samples. So this is double that one. I guess I'm bold enough to try it.

So a, 0, so that would be minus a over 2 and a b over 2. I think we got here 1/2 of-- I think-- looks to me like a over 2 squared plus b over-- I'm missing my plus or minus, but when I'm squaring them, that's the whole point of variance. Doesn't matter. And the b over 2.

And here, I think I'm wrong by a over 2, and I'm wrong by b over 2 or minus a over 2. But when I square them again, doesn't matter. So I think I get another 1/2 of a over 2 squared plus b over 2 squared. Forgive me for this simple computation, but just to practice.

So what have I got? I've got a 1/2 of that and 1/2 of that. So that adds up to this thing a squared over 4 plus b squared over 4, but then I'm doing two samples. I have to multiply by the number of samples. So I think so times 2 for two samples. I think I'm getting-- it was 1/4, but now it will be 1/2 of a squared b squared. Yeah, I didn't-- yeah, yeah. I think that's right, but forgive me while I just ask myself. Yeah.

This will be-- actually, you already have these notes. This is section 2.4. So I think it's there on Stellar. So what's the point of this? First point was to like remember some of the steps that go into the variance. Oh, there's another formula for variance and I want to tell you. And the second point is to bring in a new idea.

Suppose we want to make this-- suppose this variance is bigger than we want. Suppose, for example, that b is a lot bigger than a. Suppose b is a lot bigger than a. Then what should we have done differently in this randomized linear algebra? If I'm trying to get this thing close, get close to that thing, and if b is a lot bigger than a, then what should I do differently?

I don't know what b is exactly, but I have the information that it's bigger than a. Then I should increase the probability-- I shouldn't do half and half. So here was randomized sampling taking the average. My probabilities were a 1/2 and a 1/2.

I believe that I could keep the mean correct. Of course, that's fundamental to get the mean right. And get a better answer, you get a smaller variance than that b squared over there by picking that thing with higher probability. So that's where the randomized-- it turns out to be called norm squared probability.

The decision on what the probability should be goes-- it turns out to be the optimal one, goes with the square of the size. So if b is twice as big as a, and I want to get the variance down, then the probability-- I should use probabilities that are four times-- four times as often I will choose b than a. That's going to be the conclusion at 2 o'clock, hopefully.

OK, so that's one point. And just another little point while we are reviewing variance, this is the standard formula for the variance, sum of all the possible outcomes with their probabilities, the distance from the mean squared. Do you know a second formula, which is very close to this and very similar, and it comes from substituting the meaning of the mean, substituting what the mean is?

So yeah, I just want to mention a second formula. And I don't know which one we'll actually use. But the second formula for the same quantity, sigma squared, is the sum of probabilities times output squared. So I haven't subtracted off the mean in this second formula. I have to do it now. And I'll do the mean-- I'll do the mean all at once, mean squared.

Of course, the mean involves-- remember that the mean is the sum of the probability times the outcome. And it's just playing with a little algebra to show that you can either-- you have a choice of whatever is more convenient, subtract the mean of from each output or do all the outputs, but then you haven't accounted for the fact that you really want the distances from the mean, and then you subtract off the mean squared.

Two ways to do it, two ways, equal ways to do it. Yeah, we will review the basic ideas of mean and variance in the section on probability. Here, yes, question? Yeah.

AUDIENCE: Is the mean a part of [INAUDIBLE]?

GILBERT STRANG: The mean? Oh, in here? That's separate. Yeah, that's the whole point. Yeah, so this was like, do this, and then subtract off the mean squared. Or keep the mean in every term, and do it that way. Yeah, you could verify that the two are the same.

OK, so when we go now to the bigger question, I've forgotten which way I do it, but I'm free to choose. OK, is that like small sample reasonable, and you get the idea that if the-- if we know that if we look at our matrix, first of all, and find out which columns are large, large norm, and which columns are smaller, then that might be useful information to weight our probabilities to pick the larger one more often. OK. OK.

In fact, let me just tell you what are the two possibilities there. One is what I just said, weight your probabilities by the square of the norm, this norm squared weighting that we'll see and take the columns as they come, but with higher probability on the big columns, or you could say another way would be mix the columns, so that they more or less have similar sizes, and then, keep track of what you've done, and then just the probabilities can all be equal.

So that would be the other way. Take your matrix, mix it up, take combinations of the columns with random numbers. It's a random world here. Do a mixing, and then operate on the mixed matrix. OK, I'm going to do it the first way. I'm going to pick these probabilities to-- they'll turn out to be proportional to norm squared.

OK, ready for that? Here it comes. So let me bring that down. Yeah. OK. Actually, I could leave it up for now, because it told us what we're up to. OK. So what have I got? Let me just see if I can-- so we're multiplying a times b, and we're going to use these probabilities. Pj is going to be the length of that column times the length of that row, norm squared.

Well, norm squared, if I was multiplying a by a transpose, then I really would be squaring. That would be the same as that. So I'm going to use the word norm squared or length squared. Also, here, where the two-- I'm not assuming that b is a transpose. OK, so that will be the probabilities will be proportional to that.

But now, those that don't add up to 1, so how do I make the probabilities add up to 1? This is the probability of choosing column j of a times times row j of b. That's what Pj refers to. OK, so what is my plan?

Oh, I have to make the probabilities add to 1, or I'm really breaking the fundamental law here. So if I have a bunch of probabilities, and I kind of know what I want, but they don't add up to 1, what do I do? Divide by their sum. Let me call c their sum.

So the probability is going to be that over c, and c is going to be the sum of however many rows and columns. I guess maybe I had r in my picture of aj bj transpose. OK, so all I did was scale the probability so that they now add to 1. Good.

OK, so now I'm ready to go to work. I'm ready to choose-- oh, yes, so here's my rule. I will choose column row j with this probability, but then I'm going to multiply it, and I'm free to do that if I want to. So my approximation, my approximate AB will be-- I'll take this, whichever comes out, I'll take the aj bj transpose that comes out.

It comes out with probability Pj. But I'm going to divide this by-- and I think I'm, this is the right one-- s, the number of samples, times Pj. So I thought, at first, that's weird. Went to all the trouble to pick these Pj's, claiming that these are the good ones. So my claim to eventually prove at the end-- first, I'll have to understand how the sampling is done. That's like the most important.

But then when I go to compute the mean, I'll get the correct mean, and when I go to compute the variance, I'll get some expression for the variance, and then the plan will be choose these Pj's to minimize that total variance. So this is what-- that's a typical sample.

With probability Pj, pick that that matrix, that rank 1 matrix. So then my approximate AB is the sum of all these over s samples. Are you with me? Let me just repeat. I'm trying to multiply AB. Each sample is just a single column times row.

So it's way wrong, way wrong. It's just a tiny piece of AB. But I take that sample with probability Pj, and I divide it by S Pj, so that the Pj's cancel here. Oh, yes. OK, right. So I would like to see that the mean is correct. I would like to see that the mean is correct.

I'm going to compute the mean of my process. So like it's falling into my lap here. I made it that way. These Pj's cancel. I divided by s. So the mean of a typical sample will be-- so the mean of one sample is the probability of getting it times what I take. So it's just the sum of aj bj transpose over s.

You're going to say, OK, you're wasting our time. But we got-- I would just want to show that I'm getting the correct mean out of this plan. So do you see that if that's a mean of one sample, so what's the mean of the sum of all the samples? Well, multiply by s, because it was the same mean.

Every sample had the same mean, just as it did in our Little League practice example. So that's the mean of one sample. So the mean of all samples added together, multiplies this by s. The s's cancel, and I get AB. Remembering my-- however way I defined AB there, yeah. Yeah.

All I'm saying here is that I did something reasonable in the sampling process, so that the mean came out right. And now is the hard part, the variance. What's the variance? OK, so what do I have to compute-- and I may-- it will depend on the p's, p1 to pr, I guess. We had r different rows, different column row pairs to choose, and we chose probabilities, these guys, that depended on this size.

And now I'm going to compute the variance, and it won't be 0, because every sample is wrong. I'm never getting from a sample. A sample is just giving me a column times a row, a rank 1 guy, and they averaged out to give the correct product. But each one is certainly wrong, because it's just a rank 1.

So when I compute variance, I'm going to definitely not get 0, right? In other words, of course, when would the variance be 0? Yeah, if AB were rank 1, I guess I'd get it right every time. Thanks. That was a better answer than I had in mind. Yeah, yeah. The variance would only be 0 if every sample was right. And that would be true if the rank was 1, and there was only one thing to choose. But that's not the problem we want.

OK, so the variance is there. My instinct is to tell you what this calculation produces, since you and I can read. Would you allow me to do that? So here, the variance for a sample turned out to equal, so we will figure it out, turns out to equal the sum over-- as it was up there of the aj bj transpose, probably squared. Let me just check. Yes, squared.

Yeah, why don't I help myself here? So these are squared because variances are squared. And then when I look to see what-- I think there is an s there, and there's a Pj, so why is there a Pj there, when it canceled here? So here, the Pj, when I multiply by that, canceled. Why doesn't it cancel over there? Because it's squared over there.

Over there, this thing is squared. So it was Pj twice. Here, I have Pj, its probability once. So I've still got the Pj in the denominator, one factor of Pj in the denominator. And then-- so that is-- I guess what I'm doing is I'm computing the variance this way.

So what I've computed now is this first bit, and then I said should subtract the mean squared. And this is for one sample. So the mean squared is-- I think it turns out to be 1 over s times AB squared in this Frobenius norm. It's a squared plus b squared stuff that I saw before.

OK, so this-- I've jumped a serious step to get from the sum-- the formula for the variance. I've plugged in this problem and got that. OK, and now I'm going to sample. Let's see where-- yeah. I would like to simplify this. I would like to simplify that.

So I have to plug in the Pj's. OK, so after plug in for that Pj, and we decided what Pj was going to be here. OK, so when I plug that in in the denominator, it will cancel one of these. And I'll just have a sum of of aj Pj bj norms. And what that is C.

So let me say this again just. It's something you can just check when you have a minute. When I plug in that value for Pj here, it cancels the squares and just leaves the first power. So then I'm adding up the first power, and I get C. But the Pj had a factor C in the denominator, and it's in the denominator over there, so that C is up there.

So it's C squared coming here, a constant squared, minus the other term. There's a 1 over s. That will eventually go away. And this other term is 1 over s norm AB the Fromenius norm squared. Or maybe 1 over s's are-- so you're seeing-- and I apologize, a little bit messy bit of algebra. A little bit messy bit of algebra. But that's what we ended up with.

And when we take s samples and combine them, that will cancel the s, and I think it'll knock that out when we combine the s samples. OK. OK. Now what? Now, we get to choose those probabilities. And how are we going to choose them?

What will be the best choice? Here is the expression for the variance. Yeah, this is good. This is good. Stay with me for now, and you will be saying to yourself, there's some steps there that I didn't see fully, and I want to check. And I agree. But let me say that we get to that point, and this is a fixed number.

So it's C that we would like to make small, and that's our final job. This was true for any choice of the probabilities P. Well, oh, yeah, sorry. Yeah, yeah. So I want to-- this still had in it a probability. Yeah. What do I want to do? I want to show that that was the best choice, that this was the best choice.

Yeah, yeah. I want to show that that's the best choice, that the choice of weights of probabilities, based on length of a times the length of b-- of course, it sounds reasonable, doesn't it? We want to-- for big columns and big rows, we want to have a higher probability to choose those. But is the probability proportional to the length of both, or should it be proportional to the 10th power or the square root or what?

That's what our final step of optimizing the P. So this is the final step. Optimize the probabilities, P1 to P2, I guess, no, P1 to Pr, for the r rows, r columns of a and r rows of b, subject to-- they have to add up to 1. And what do I mean by optimize? I mean minimize. This optimize means minimizing this expression, C. So aj bj transpose.

Where is-- over Pj. Oh yeah, wait a minute. Help. Help. So let me just see. Yeah, my variance has got a Pj in it. Yeah, my variance-- sorry-- my variance-- oh, OK. This is my variance. This is the result if I make the right choice for the-- if I make this choice for the probabilities. But I'm backing up a minute.

This is if-- this is the with optimal Pj's, then we got that answer. Great. That was our answer. But I'm backing up to this and saying, what are the optimal Pj's to make this variance small? So really, I'm just doing this. Let me write the problem simpler.

Minimize with the sum of the P's equal 1, some quantity Q squared over Qj over Pj. Yeah, that's it. How do you-- so these Qj's that I just introduced that letter for are the aj bj's. They're given. Maybe I'll just put back aj Pj.

So to repeat, this is the calculation of the variance for any choice of Pj's. This is what I get if I make the best choice, but over here, I'm going to show that it is the best choice, that it's the choice that makes this result as small as possible. So that's the Lagrange multiplier aspect.

So the statistics has been done. I'm getting this answer. And instead of putting in some weird Q, let me put in what these are. They're whatever. They're a bunch of numbers. But I'm dividing by the Pj, and how do you find the best Pj? Do you know about that optimization question?

They have to add to 1. And the Lagrange had the great idea. So this is maybe the first time we've used his idea. So do you remember what his idea is? He takes this constraint, and he builds it into the function. He multiplies it by some unknown mysterious number, often called lambda, but nothing to do with eigenvalues, of the constraints that the Pi's should add to 1. So he had 0. He had 0, but with a variable lambda.

This is Lagrange's idea. So it's pretty neat that this problem-- I've left randomized sampling. I've arrived at this final sub problem, optimizing the probabilities under the condition that they add to 1, and Lagrange's idea was build that equation into the function. Then you can take derivatives, but you also take derivatives with respect to lambda, because that's now an unknown.

And you solve-- you set the derivatives to 0, and you get the answer. It's like a miracle. But if you've seen Lagrange, it's a confusing miracle. That's what it is. Yeah. OK. So if I take the derivatives with respect to the P's, set them to 0, I think I'm going to get the recommended P's.

So I've computed the final answer with a recommended P's, but now I'm going to show that they really are recommended. So can you take the derivative of that with respect to P? Can I-- I'll just raise this a little, raise it a little more. OK, take the derivative with respect to P, each P, because I've got n unknowns there, or however many, maybe r unknowns.

And I've got lambda, so I've got r plus 1 things. So what's the derivative with respect to P. OK, calculus. Take the derivative of that. It's aj bj transpose over-- with a minus Pj squared, right? And the derivative of that with respect to Pj is? Minus lambda.

So that derivative with respect to Pj is 0, and the derivative-- so this was a derivative with respect to Pj has to be 0. And then the derivative with respect to lambda-- the derivative with respect to lambda is that, on call them j's-- j's minus equals 1.

Lagrange confused the whole world, but he gave us a break that in the derivative with respect to lambda, it just brings back that constraint, because he just built it in with the factor of lambda, then he took the derivative, and it brought back that constraint. But this part is the beautiful part.

Now, what do I learn from that? And sometimes this would be a plus. Why don't I make it a plus just to make my life easier? Lagrange is dead now, and he don't care anyway, whether it's plus or a minus. OK. So this is telling me this. So this is tell me what its multiplier is.

He's telling me that-- this equation is telling me that the multiplier is aj bj transpose over Pj squared. Or put it another way, he's telling me that Pj squared is-- I guess, I'm hoping that after the pretty confusing steps that we took, this is a separate little bit of math, using the Lagrange multiplier idea, and I hope that your thought will be, boy, that was pretty simple. So I'm going to put the Pj squareds here and the lambda there.

What does this tell me? I've taken the derivative with respect to the Pj's, and I got this equation for each j because I took the derivative, the partial derivative with respect to each of the Pj's.

And it tells me that Pj squared-- wait a minute. What's the square in there for? Help. I've only got two minutes.

And oh, they have to add to 1. Oh yeah, lambda is going to save us. Right, lambda is going to save us, because the total probabilities-- so Pj will be the square root of this stuff. And then I-- the number lambda, I haven't decided. Lagrange's multiplier, I haven't decided.

So what is it? It's the correct number to make this equal to 1. So that is the C. Oh god. Why have I got square root there? Shoot.

AUDIENCE: I think you're supposed to start off with squares.

GILBERT STRANG: I should have started with squares?

AUDIENCE: [INAUDIBLE]

GILBERT STRANG: So these should be squares? Ah, thank you. You could have told me earlier. When you see a professor in trouble, don't just let him hang there. OK, all right. OK, and this is aj bj transpose. So apart from the kerfuffle here, and the notes get it right, because I had time to think there, it turns out that this optimum gave the formula for the Pj's that I used earlier.

So when I introduced this formula, I said, let's choose those probabilities, but then I came back at the very end and showed that they are the probabilities that minimize the variance. So that's like today's lecture. Can you just think a minute, but please do go back through the notes, because there is some messy steps in the variance there that I had to go by quickly.

But you understand the principle, that we set up a randomized system. We choose probabilities, aiming to get the smallest variance. And it turns out that the good probabilities are bigger when the column is a larger column, so that to use this, you have to go through the matrix and find the length of the columns, because that's what's telling you the probabilities.

So that's like a first pass through. Before you do the randomized sampling, you must decide on the probabilities, and they depend on the sizes of the different columns. Thank you for getting me through that. I'll come back to a little more about randomized things next time, and then later, not much later, but a little bit later, we'll be seeing probability much more seriously OK, thank you.