GILBERT STRANG: OK, so kind of a few things in mind for today. One is to answer those two questions on the second line. We found those two formulas on the first line last time, the derivative of a inverse. So the derivative of A squared ought to be easy. But if we can't do that, we need to be sure we can.

And then this was the derivative of an eigenvalue. And then it's natural to ask about the derivative of the singular value. And I had a happy day yesterday in the snow, realizing that that has a nice formula too. Of course, I'm not the first. I'm sure that Wikipedia already knows this formula. But it was new to me.

And I should say Professor Edelman has carried it to the second derivative. Again, not new, but it's more difficult to find second derivatives, and interesting. But we'll just stay with first derivatives.

OK, so that's my first item of sort of business from last time. And then I'd like to say something about the lab homeworks and ask your advice and begin to say something about a project. And then I will move to these topics in Section 4.4 that you have already.

And you might notice I skipped 4.3. And the reason is that on Friday, actually arriving at MIT tomorrow is Professor Townsend, 4.3 is all about his work. And he's the best lecturer I know. He was here as an instructor and did 18.06 and was a big success. Actually, he's also just won a prize for the SIAG/LA, international prize for young investigators, young faculty in applied linear algebra. So he goes to Hong Kong to get that prize too. Anyway, he will be on the videos and in here in class Friday, if all goes well.

OK, so in order then, the first thing is the derivative of A squared. And you might think it's 2A dA dt, but it's not. And if you realize that it's not, then you realize what it is, you will get these things right in the future.

So the answer to the derivative of A squared is not 2A dA dt. And why isn't it? And what is the

right answer?

So I do that maybe just below here. Well, I could ask you to guess the right answer, but why don't we do it systematically. So how do you find the derivative? It's a limit. First you have a delta A, right. And then you take a limit.

So I look at A plus delta A squared minus A squared. So that's the change in A squared. And I divide it by delta t. And then delta t goes to 0. So that's the derivative I'm looking for, the derivative of A squared.

And now, if I write that out, you'll see why this is wrong, but something very close to it, of course-- can't be far away-- is right. So what happens if I write this out? The A squared will cancel the A squared. What will I have? Will I have 2A delta A? Why don't I write 2A delta A next?

Because when you're squaring a sum of two matrices, one term is A delta A, and another term is delta A A. And those are different in general. And then plus delta A squared. And now I divide it all by delta t.

So you're now seeing my point that now I let delta t go to 0. So I'm just doing matrix calculus. And it's not altogether simple, but if you follow the rules, it comes out right.

So now what answer do I get as delta t goes to 0? I get A dA dt-- that's the definition of the-- that ratio goes to dA dt. That's the whole idea of the derivative of A.

And now what's the other term? It's dA dt A. So it was simply that point that I wanted you to pick up on, that the derivative might not commute with A. Matrices don't commute in general.

And so you'll notice that we had a similar expression there. We had to pay attention to the order of things there. And now we get it right. It's not this, but A dA dt plus dA dt A. OK. Good.

Now, can I do the other one? Which is a little more serious, but it's a beautiful formula. And it's parallel to this guy. You might even guess it.

So I'm looking for the derivative of a singular value. The matrix A is changing. dA dt tells me how it's changing at the moment, at the instant. And I want to know how is sigma changing at that same instant.

And sort of in parallel with this is a nice-- the nice formula-- u transpose dA dt v of t. Boy, you

couldn't ask for a nicer formula than that, right?

You remember this is the eigenvector. And that's the eigenvector of A transpose. So this is the singular vector of A. And you could say this is a singular vector of A transpose, or it's the left singular vector of A. So that's our formula. And if we can just recall how to prove it, which is going to be parallel to the proof of that one, then I'm a happy person and we can get on with life. So let's remember this, because it will help us to remember the other one, too.

OK, so where do I start? I start with a formula for sigma. So I believe that sigma is u transpose times A times v. Everybody agree with that? Everything's depending on t in this formula. As time changes, everything changes. But I didn't write in the parentheses, t three more times.

Can we just remember about the SVD. The SVD says that A times v equals--

AUDIENCE: Sigma u.

GILBERT STRANG: Sigma u. Thanks. Av is sigma u. That's the SVD. So when I put in for Av, I put in sigma u. Sigma is just a number. So I bring it outside. And I'm left with u transpose u. And what's u transpose u? 1. So I've used these two facts.

Or I could have gone the other way and said that this is the transpose of-- this is A transpose u transpose. I could look at it that way times v. And if I look at it that way, I'm interested in what is A transpose u. And what is A transpose u? It's sigma v. And it's transpose, so sigma v transpose v.

And what is sigma v transpose v?

AUDIENCE: Sigma.

GILBERT STRANG: It's sigma again, of course. Got sigma both ways. OK.

Now, I'm ready to take the derivative. That's the formula I have for sigma, completely parallel to the formula that we started out with for lambda. The eigenvalue was y transpose Ax. And now we've got u transpose Av.

And, by the way, when would those two formulas be one and the same? When does the SVD just tell us nothing new beyond the eigenvalue stuff for what matrices are the singular values, the same as the eigenvalues, and singular vectors the same as this as the eigenvectors for--

For?

AUDIENCE: Symmetric.

GILBERT STRANG: Symmetric, good. Symmetric, square, and-- the two words that I'm always looking for in this course. If you want an A in this course, just write down positive definite in the answer to any question, because sigmas are by definition positive. And if they're going to agree totally with the lambdas, then the lambdas have to be positive. Or could be 0, so positive semidefinite definite would be the right answer. Anyway, this is our start.

And what do we do with that formula? So this was all the same, because v transpose v was 1. Here I had v transpose v. And that's 1. So it gave me sigma. Yeah, good. Everybody's with us.

OK, what do I do? Take the derivative. Takes the derivative of that equation in the box. It's exactly what I did last time with the corresponding equation for lambda. Same thing. And I'm going to get again-- it's a product rule, because I have three things multiplied on the right-hand side. So I've got three terms from the product rule.

So d sigma dt, coming from the box, is du transpose dt Av plus u transpose dA dt v plus the third guy, which will be u transpose A dv dt. Did I get the three terms there? Yep.

And which term do I want? Which term do I believe is going to survive and be the answer? Well, this is what I'm after. So it's the middle term. The middle term is just right. And the other two terms had better be zero. So that will be the proof. The other two terms will be zero. So can we just take one of those two terms and show that it's zero like this one?

OK, what have I got here? I want to know that that term is 0. So what have I got. I've got du transpose dt times Av. And everybody says, OK, in place of Av, write in sigma u. And sigma's a number, so I don't mind putting it there.

So I've got sigma, a number of times the derivative of u times u itself, the dot product-- the derivative of u with dot product with u. And that equals? 0, I hope, because of this. Because of that. This comes from the derivative of that.

But you see, now we've got dot products, ordinary dot products, and a number on the right-hand side. We're in dimension 1, you could say. So this tells me immediately that the derivative of u with u plus u transpose times the derivative of u is the derivative of 1, which is 0.

All I'm saying is that these are the same. You know, vectors, x transpose y is the same as y transpose x when I'm talking about real numbers. If I was doing complex things, which I could do, then I'd have to pay attention and take complex conjugates at the right moment. But let's not bother.

So you see, this is just two of these. And it gives me 0. So that term's gone. And similarly, totally similarly, this term is gone.

This is A transpose u, all transpose. I'm just doing the same thing times dv dt. And what is A transpose u? It's sigma v. So this is sigma v transpose dv dt. And again 0, because of this.

So in a way this was a slightly easier thing-- the last time was completely parallel computation. But the first and third terms had to cancel each other with the x's and y's. Now, they disappear separately, leaving the right answer.

You might think, how did we get into derivatives of singular values? Well, I think if we're going to understand the SVD, then the first derivative of the sigma is-- well, except that I've survived all these years without knowing it. So you could say it's not-- you can live without it, but it's a pretty nice formula.

OK, that completes that Section 3.1. And more to say about 3.2, which was the interlacing part that I introduced.

OK, so where am I? I guess I'm thinking about the neat topics about interlacing of eigenvalues. So may I pick up on that theme, interlacing of eigenvalues and say what's in the notes and what's the general idea? OK. So we're leaving the derivatives and moving to finite changes in the eigenvalues and singular values, and we are recognizing that we can't get exact formulas for the change, but we can get bounds for change. And they are pretty cool.

So let me remind you what that is, what they are. So I have a matrix-- let's see, a symmetric matrix S that has eigenvalues lambda 1, greater equal lambda 2, greater equal so on. Then I change S by some amount. I think in the notes there is a number, theta times 1 matrix.

That has eigenvalues mu 1, greater equal mu 2, greater equal something. And these are what I can't give you an exact formula for. You just would have to compute them. But I can give you bounds for them. And the bounds come from the lambdas.

So this was a positive. This is a positive change. So the eigenvalues will go up, or stay still, but they won't go down.

So the mu's will be bigger than the lambdas. But the neat thing is that mu 2 will not pass up lambda 1. So here is the interlacing. Mu 1 is greater equal lambda 1. That says that the highest eigenvalue, the top eigenvalue went up, or didn't move. But mu 2 is below lambda 1.

This is the new-- everybody's with me here? This is a new, and this is the old. New and old being old is S, new is with the change in S.

And that mu 2 is greater equal lambda 2. So the second eigenvalues went up. And then so on.

That's a great fact. And I guess that I sent out a puzzle question. Did it arrive in email? Did anybody see that puzzle question and think about it?

It worried me for a while. Suppose this is the second eigenvalue value-- eigenvector. So I'm adding on, I'm hyping up the second eigenvector, hyping up the matrix in the direction of the second eigenvector. So the second eigenvalue was lambda 2. And its mu 2, the new second eigenvalue, is going to be bigger by theta.

But then I lost a little sleep in thinking, OK, if the second eigenvalue is mu 2 plus theta-- sorry, if the second eigenvalue mu 2-- so let me write it here. If mu 2, the second eigenvalue, is the old lambda 2 plus theta then bad news, because theta can be as big as I want. It can be 20, 200, 2,000.

And if I'm just adding theta to lambda 2 to get the second-- because it's a second eigenvector that's getting pumped up, then after a while, mu 2 will pass lambda 1. This will be totally true. I have no worries about this.

The old lambda 1-- actually, the old-- I'll even have equality here, because for this particular change, it's not affecting lambda 1. So I think mu 1 would be lambda 1 in my hypothetical possibility.

What I'm trying to get you to do is to think through what this means, because it's quite easy to write that line there. But then when you think about it, you get some questions. And it looks as if it might fail, because if theta is really big, that mu 2 would pass up lambda 1. And the thing would fail. And there has to be a catch. There has to be a catch.

So does anybody-- you saw that in the email. And I'll now explain what how I understood that everything can work and I'm not reaching a contradiction. And here's my thinking.

So it's perfectly true that the eigenvalue that goes with u2-- or maybe I should be calling them x2, because usually I call the eigenvectors x2-- it's perfectly true that mu 2, that that one goes up. But what happens when it reaches lambda 1?

Actually, lambda 1, the first eigenvalue, is staying put, because it's not getting any push from this. But the second eigenvalue is getting a push of size theta. So what happens when lambda 2 plus theta, which is mu 2-- mu 2 is lambda 2 plus theta-- what happens when it comes up to lambda 1 and I start worrying that it passes lambda 1?

Do you see what's happening there? What happens when mu 2 passes-- when mu 2, which is-- I'm just going to copy here-- it's the old lambda 2 plus the theta, the number. What happens when theta gets bigger and bigger and bigger and this hits this thing and then goes beyond? Just to see the logic here.

What happens is that this lambda 2 plus theta, which was mu 2, mu 2 until they got here. But what is lambda 2 plus theta after it passes lambda 1? It's lambda 1 now. It passed up, so it's the top eigenvalue of the altered matrix. And therefore, it's just fine. It's out here. No problem.

Maybe I'll just say it again. When theta is big enough that mu 2 reaches lambda 1, if I increase theta beyond that, then this becomes not mu 2 any more, but mu 1. And then totally everybody's happy.

I won't say more on that, because that's just like a way that I found to make me think, what do these things mean? OK, enough said on that small point.

But then the main point is, why is this true? This interlacing, which is really a nice, beautiful fact. And you could imagine that we have more different perturbations than just rank 1s.

So let me tell you the inequality, so named after the discoverer, Weyl's inequality. So his inequality is for the eigenvalues of S plus T. So T is the change. S is where I start. It has eigenvalues lambda.

But now, I'm looking at the eigenvalues of S plus T. So I'm making a change. Over here, in my little puzzle question, that was T. It was a rank 1 change. Now I will allow other ranks.

So I want to estimate lambdas of S plus t in terms of lambdas of S and lambdas of T. And I want some inequality sign there. And it's supposed to be true for any symmetric matrices, symmetric S and T.

And then a totally identical Weyl inequality-- actually, Weyl was one of the people who discovered singular values. And when he did it, he asked about his inequality. And he found that it still worked the way we've found this morning earlier. I haven't completed that yet, because I haven't told you which lambdas I'm talking about. So let me do that.

So now, I'll tell you Weyl's inequality. So S and T are symmetric. And so the lambdas are real. And we want to know-- we want to get them in order.

OK, so here it goes. Weyl allowed the i-th eigenvalue of S and the j-th eigenvalue of T and figured out that this was bounded by that eigenvalue of S plus T. So that's Weyl's great inequality, which reduces to the one I wrote here, if I make the right choice-- yeah, probably, if I take j equal to 1. So you see the beauty of this. It tells you about any eigenvalues of S, eigenvalues of T.

So I'm using lambdas here. Lambda of S are the eigenvalues of S. I'm using lambda again for T and lambda again for S plus T. So you have to pay attention to which matrix I'm taking the eigenvalues out of.

So let me take j equal to 1. And this says that lambda i, because j is 1, S plus T is less or equal to lambda i of S plus lambda 1, the top eigenvalue of T. This is lambda max of T.

Do you see that that's totally reasonable, believable? That the eigenvalue when I add on T-- let's imagine in our minds that T is positive. T is like this thing. This could be the T, example of a T. It's what I'm adding on.

Then the eigenvalues go up. But they don't pass that. So that tells you how much it could go up by.

So I guess that Weyl is giving us a less than or equal here. Less or equal to lambda 1-- so I'm taking i to be 1-- plus theta. Yeah, so that any equality I've written down there-- there's some playing around to do to get practice. And it's not so essential for us to be like world grandmasters at this thing, but you should see it.

And you should also see j equal to 2. Why will j equal to 2 tell us something? I hope it will. Let's

see what it tells us.

Lambda i plus 1 now-- j is 2-- of S plus T. So it's less than or equal to lambda i of S plus lambda 2 of T. I think that's interesting. And also, I think I also could get lambda i plus i minus 1. Let me write it and see if it's correct. Plus lambda i minus 1. So those was add up to i plus 2. Yeah, I guess lambda i plus 1 plus lambda 1 of T.

That's what I got by taking-- yeah, did I do that right? I'm taking j equal to 1. No, well, I don't think I got it right.

What do I want to do here to get a bound on lambda i plus 1? I want to take j equal to 2. I should just be sensible and plug in j equal to 2 and i equal to 1.

All I want to say is that Weyl's inequality is the great fact out of which all this interlacing falls and more and more, because the interlacing is telling me about neighbors. And actually if I use Weyl for i and j, different i's and j's, I even learn about ones that are not neighbors. And I could tell you a proof of Weyl's inequality. But I'll save that for the notes.

So I think maybe that's what I want to do about interfacing, just to say what the notes have, but not repeat it all in class. So the notes have actually two ways to prove this interlacing. The standard way that every mathematician would use would be Weyl's inequality.

But last year, Professor Rao, visiting, found a nice argument that's also in the notes. It ends up with a graph. And on that graph, you can see that this is true. So for what it's worth, two approaches to this interlacing and some examples. But I really don't want to spend our lives on this eigenvalue topic. It's a beautiful fact about symmetric matrices and the corresponding fact is true for singular values of any matrix, but let's think of leaving it there.

So now, I'm moving on to the new section. The new section involves something called compressed sensing. I don't know if you've heard those words.

So these are all topics in Section 4.4, which you have. I think we sent it out 10 days ago probably.

OK, so first let me remember what the nuclear norm is of a matrix. The nuclear norm a matrix is the sum of the singular values, the sum of the singular values. So it's like the L1 norm for a vector. That's a right way to think about it.

And do you remember what was special? We've talked about using the L1 norm. It has this special property that the ordinary L2 norm absolutely does not have. What was it special about the L1 norm? If I minimize the L1 norm with some constraint, like ab equal b, what's special about the solution, the minimum in the L1 norm?

AUDIENCE: Sparse.

GILBERT STRANG: Sparse, right. Sparse. So this is moving us up to matrices. And that's where compressed sensing comes in. Matrix completion comes in.

So matrix completion would just be-- I mentioned-- so this is completion. And I'll remember the words Netflix, which made the problem famous. So I have the matrix A, 3, 2, question mark, question mark, question mark, 1, 4, 6, question mark-- missing data.

And so I have to put it in something there, because if I don't put in anything, then the numbers I do know are useless, because no row or no column is complete. So it just would give up. Somebody that sent me the data, 3 and 2 and didn't tell me a ranking for the third movie, I'd have to say, well, I can't use it. That's not possible. So we need to think about there.

And the idea is that the numbers that minimized the nuclear norm are a good choice, a good choice. So that's just a connection here that we will say more about, but not-- we could have a whole course in compressed sensing and nuclear norm. Professor Parrilo in course 6 is an expert on this.

But you see the point that-- so you remember v1 came from the 0 norm. And what is the 0 norm of the vector? Well, it's not a norm. So you could say, forget it, no answer.

But what do we symbolically mean when I write the 0 norm of a vector? I mean the number of....? Non-zeros. The number of non-zeros. This was the number of non-zeros in the vector, in v.

But it's not a norm, because if I take 2 times the vector, I have the same number of non-zeros, same norm. I can't have the norm of 2v equal the norm of v. That would blow away all the properties of norms. So v0 is not a norm. And then we move it to that sort of appropriate nearest norm. And we get v1. We get the L1 norm, which is the sum of-- everybody remembers that this is the sum of the vi.

And you remember my pictures of diamonds touching planes at sharp points. Well, that's what

is going on here. That problem was called basis pursuit. And it comes back again in this section.

So I minimize this norm subject to the conditions. Now, I'm just going to take a jump to the matrix case. What's my idea here?

My idea is that for a matrix, the nuclear norm comes from what? What's the norm that we sort of start with, but it's not a norm? And when I sort of take the-- because the requirements for a norm don't fail-- they fail for what I'm about to write there. I could put A 0, but I don't want the number of non-zero entries. That would be a good guess. And probably in some sense it makes sense. But it's not the answer I'm looking for.

What do you think is the 0 norm of a matrix that is not a norm, but when I pump it up to the best, to the nearest good norm, I get the nuclear norm? So this is the question, it's what is A0? And it's what?

AUDIENCE: The rank.

GILBERT STRANG: The rank. The rank of matrix is the equivalent. So I don't know about the zero. Nobody else calls it A0. So I better not. It's the rank.

So again, the rank is not a norm, because if I double the matrix, I don't double the rank. So I have to move to a norm. And it turns out to be the nuclear norm.

And now, I'll just, with one minute, say it's the guess of some people who are working hard to prove it, that the deep learning algorithm of gradient descent finds the solution to the minimum problem in the nuclear norm. And we don't know if that's true or not yet.

For related examples, like this thing, it's proved. For the exact problem of deep learning, it's a conjecture.

So that's what in section 4.4. But that word lasso, you want to know what that is. Compressed sensing, I'll say a word about. So that will be Monday after Alex Townsend's lecture Friday. So he's coming to speak to computational science students all over MIT tomorrow afternoon. I'll certainly go to that, but then he said he would come in and take this class Friday. So I'll see you Friday. And he'll be here too.