PROFESSOR: OK. So I thought I'd begin today with, as we're coming to the end of the sort of focus on linear algebra and moving on to a little probability, a little more optimization, and a lot of deep learning. So this was like, by way of review, to write down the big factorizations of a matrix.

And so my idea, and I kind of enjoyed it, is checking that the number of free parameters, say an L and U or a Q and R or every-- each of those, that the number of free parameters agrees with the number of parameters in A itself, like n squared, usually.

So A usually has n squared. And then can we replace A if-- after we've computed L and U, can we throw away A? Yes, because all the information is in L and U. And it fills that same n by n matrix. Well, that's kind of obvious because L is lower triangular, and the diagonal, all ones, are not free parameters. And U is triangular, upper triangular. And it's diagonal to the pivots. Those are free parameters so that-- but can I just write down the count?

So I'll go through each of these just quickly after I've figured out how-- these are sort of the building blocks. So how many free parameters are there in these two triangular matrices? Well, I think the answer is 1/2 n, n minus 1, and 1/2 n, n plus 1. That's a familiar number. You recognize that as the sum of 1 plus 2, up to n. And you have one free para-- in the upper triangular U. You've got one free parameter up in the corner, two in the next one. And as you're coming down, you end up with n on the main diagonal. And they add up to that. And you see that those two are different by n, which is what we want.

OK. Diagonal. The answer is obviously n.

How about the eigenvector matrix? This whole exercise is like something I've never seen in a textbook. But for me it brings back all these key-- really the condensed course in linear algebra is on that top line. So how many free parameters in an eigenvector matrix?

OK. And of course, if you're sort of thinking, what's the rule for free parameters? My answer is going to be, for the number of free parameters, so this is an n by n matrix with the n

eigenvectors in it. But there's a certain freedom there. And what is that? What freedom do we have in choosing the eigenvector matrix?

Every eigenvector can be multiplied by a scalar. If x is an eigenvector, so is 2x. So is 3x. So we could make a convention that the first component was always 1. Maybe that wouldn't be the most intelligent convention in the world. But it would show that that top row of ones were not to be counted. So I get n squared minus n for that.

Oh, yeah. Well, having done those two, let me look at this one. Does that come out a total of n squared? Yes, because the eigenvector x has n squared minus n by this reasoning, little hokey reasoning that I just gave. And then there are n more for the eigenvalue matrix. And there's nothing left for the eigen-- the inverse because it's determined by x. So do you see the count adding up to n squared for those?

Now, I left open the orthogonal one. I think we kind of talked about that during the-- when we met it. And it's a little less obvious. But do you remember? So I'm talking about an n by n orthogonal matrix, Q. So how many free parameters in column one? That column is what we always call Q1. Does it have n free parameters? Or is there a condition that cuts that back?

There is a condition, right? And what's the condition on the first column that removes one parameter? It's normalized. Its length is 1. So I only get n minus 1 from the first column. And now if I move over to the second column, how many free parameters there? Again, it's a unit vector. But also, it is orthogonal to the first. So two parameters got you-- two rules got imposed. And two parameters got removed. So this is n minus 2. And then finally, whatever. So I think that that-- sum of these guys is exactly the same that we had up here. I think it's also 1/2 n, n minus 1, or 1/2n squared minus n. Yeah. Yeah, so not as many as you might think because the matrix is size n squared.

Now, can I use those? Because these are the-- like the building blocks. Can I just check these? Let's see. I'll just go along the list.

L times U. So L had this. And U had that. And when I add those, it adds up to n squared. Right? The minus cancels the plus. And the 1/2n squared twice gives me n squared. So good for that one.

What about QR? Well, R is upper triangular like so. And then Q, we just got it right there. So for Q times R, it's that plus that again, adding to n squared. Good for that one. n squared for

that one.

And this one we just did. n squared minus n in x. n on the diagonal. Total n squared.

What about this guy? What about the big, really fundamental one that I would normally write to matrix as S instead of A to remind us that it-- that the matrix here is symmetric? So I'm not expecting n squared for a symmetric ma-- oh, I should've put that on my list.

What's the count for a symmetric matrix? Because this is an S here. So I'm not expecting to get n squared. I'm only expecting to get the number of symmetric S.

What's the number of free parameters that I would-- that I start with that I hope will reappear in Q and lambda? What's the deal for a symmetric matrix? Let's see. I'm free to choose.

Is it the same count as this? Yeah, because I'm free to choose the upper triangular part and the diagonal, but I'm not free to choose the lower. So I'd say that's 1/2n times n minus 1. And plus 1. Sorry. The diagonal's in there.

OK. So do I get that total, 1/2 of n squared plus n, from these guys? Well, I probably do. The diagonal guy gives me n. This gives me n. And that's a Q, which is my other favorite number there. And when I add that to that, that becomes a plus sign. And I'm good. Yeah. You see how I enjoy doing this, right? But I'm near the end. But the last one is kind of not well known.

OK. Q times S. Do you remember that factorization? That's called the polar decomposition. It's an orthogonal times the symmetric. And it is often used in engineering as a way to decompose a displacement, strain matrix. Anyway, Q times S. And it-- actually, it's very, very close to the SVD. And I have friends who say, better to compute QS than the SVD and then just move along.

Anyway, Q times S. So Q is this guy. And S. What's S? Symmetric. That's this guy. So that's Q. Let me write that letter Q and S so I don't lose it. What do those add up to? N squared. Happy.

OK. So finally, the SVD. Finally, the SVD. What's the count? Now I've got rectangular stuff in there. I'm ready for this one. And I have to think a little bit.

And we may have done this. Let's suppose that m is less or equal n. Suppose that. Yeah. Otherwise, we would just transpose and look at SVD. So let's say m less or equal n. So let's say it's got full rank. And what's the largest rank that the matrix can have? m, clearly. Full rank

m.

So the SVD will be m by m. Let's remember the U, the sigma, and the V transpose. This will be m by n. And this will be n by n. For the full scale SVD. And if the rank is equal to m, then I really expect to get-- I expect it to add up to the total for A. For A, the original A has mn, right? It's an m by n matrix. The matrix A is m by n with the m less or equal n, giving me these things. So it has mn parameters.

So do we get m times n from this? I hope we do. I know how many we get from sigma. What? How many was the count for sigma? m. And what's the count for V? So that's an n by n. And what's the count for U? OK. Yeah.

They're orthogonal matrices. So I should be able to go up to that line. This was an m by n one. Is that a 1/2n, n minus 1? Am I copying that correctly out of this circle there? That's an m by m orthogonal matrix. Oh, but I have to write m. That was foolish. OK. m. m. Yeah, because that matrix is of size m. So that's an m.

And then I have that. And then I have whatever V transpose n by n. Oh, what's the deal in there? Hmm. Do I want all of the 1/2n, n minus 1? Oh, God. I thought I had got this straight. Let's see. I could subtract this from this and find out what I should say. Whoa. Students have been known to do this too. Let's see. Well, let's try to think anyway.

So I have this n by n symmet-- this n by n orthogonal matrix. First, it could be any orthogonal matrix. Yeah. But is it only the first m columns that I really need? The rest I could just throw away. Let me try to imagine that it's just the first. Well, then I won't have any n in here. So maybe I better take a 1/2n-- no. Help. Oh, oh, yes, of course. Ha. I've got only m columns that matter, the-- everybody sort of now understands that SVD. The rank is m. Don't forget that.

OK. Then the first R, the first m columns of V are important. Those are the singular vectors that go with nonzero singular values that really matter. And the rest really don't matter. So I'm going to just-- I have to count how many-- so, sorry. V, the important part of V has how many on the m columns. But it's an n by n matrix. And those columns are orthogonal. So the answer is not mn for this guy. I have to go through this foolish reasoning again.

I have n minus 1, plus n minus 2, plus so on, plus n minus m. There were n minus 1 parameters in the first orthogonal vector-- unit vector, n minus 2 in the second one, up to n minus m in the third. And then V has some more columns that are coming, really, from a null

space, that are not important.

I believe this is the right thing to do. I'm hoping you agree. And I mean, I'm hoping even more that those add up to m times n. OK. I have a 1/2n s-- oh, I really have to total this thing. OK. This had m terms. So there's m of these n's. And then I have to subtract off 1 plus 2 plus 3, up to m. And so what am I subtracting off? What's that sum? 1 plus 2 plus 3, stopping at m? It's one of these guys, 1/2-- is it 1/2m, m plus 1? Yeah, 1/2m, m plus 1. Sorry. 1/2m, m plus 1.

I'm supposed to enjoy this. And now it gets a little nervous. But OK. So I believe that that is that. OK. Well, we have the mn. That's a good sign that we're shooting for. So does the rest of it add to nothing? Well, I guess, yeah, I guess it does. When I put these two together, I have 1/2m, m plus 1. And then I'm subtracting it away again. So I get mn. Hooray.

Well, it had to happen, or we wouldn't-- anything-- before I erase that board and consign that to history, is there-- should I pause a little more? Minute? This will be, like, I'm hoping, a one-page appendix to the notes and the book. And you'll see it.

But I do have one more count to do. And then I'm good with this review and ready to move onward to the topic of saddle points and ready to move onward after that. Well, I'll say a little bit about the next lab homework that I'm creating. And then our next topic will be, like, covariance matrices, a little statistics this week. Then we get a week off we could-- to digest it. And then come back for gradient descent, and deep learning, and those things. OK. Everybody happy with that?

So what's my final question? My final question is the SVD for any matrix of rank R. So it's an m by n matrix. But the rank is only R. It's a natural question-- how many parameters are there in a rank R matrix? We may even have touched on this question.

And I have two ways to answer it. And one way is the SVD. And that will be similar to what I just pushed up there. So if the rank is R, the SVD of this typical rank R matrix will be U sigma V transpose. But U, now this is the-- like the condensed thing, where I've thrown away stuff that's automatically zero because if the rank is only R, like if the rank was 1, suppose the rank was 1, then I'd have 1 column times 1 sigma times 1 row, right? And I could do that count for R equal 1.

Now I have R columns. So this is m by R. Then sigma is diagonal, of course. So I'm going to get R numbers out of that. And this one is now R by n.

In other words, maybe I should, like, save this little bit here that was helpful. But now I've got m is reduced to R. So I believe that if I count these three, I'll get the right number of parameters for a rank R matrix. And that's not so obvious because the rank R matrices are not a-- we don't have a subspace. If I add a rank R matrix to another rank R matrix, well, the rank could be as big as 2R and probably will be.

You know, it's just a little interesting to get your hands on matrices of rank R because they're kind of a thin, like a, well, a mass-- person would call it a manifold, some kind of a surface within matrix space. Have you ever thought about matrix space? So that's vector space because we can add matrices. We can multiply them by constants. We can take linear combinations. We could call them vectors if we like. There would be a vector space of m by n matrices. What would be the dimension of that space? So the vector space of all 3 by 4 matrices. That has what dimension? 12. 12, because you've got 12 numbers to choose. And it is a space because you can add.

Now if I say 3 by 4 matrices of rank 2, I don't have a space anymore. That word, space, is seriously preserved for meaning vector space, meaning I can take combinations. But if I take a rank 2 matrix plus a rank 2 matrix, I'm not-- so it's sort of a surface within 12d, the 2-- the 3 by 4 matrices of rank 2. And we're about to find the dimension of that surface. Does your mind sort of visualize a surface in 12 dimensions? Yeah, well, give it a shot anyway. But that surface could have-- be 11 dimensional, so to speak, like, meaning locally, the-- it wouldn't have to be a pl-- an 11 dimensional plane going through the origin. In fact, it wouldn't go through the origin because the origin won't have rank R. So it's some kind of a surface. And maybe it's got some different pieces. Probably, some smart person knows what that surface looks like. But we're just going to find out something about its number of parameters, its local dimension. Well, I know that this answer is R because I've got R sigmas. And this one, I'm pretty good at. But now it's not-- it's R by n, so it's-- instead of-- here R was m. But now, down here, R is R. So I think it's rn minus 1/2. What's that? Is that an m? So now it's an r. r plus 1. I think. I think.

And what about the U? So U is going to be similar, except instead of the n here, we've got an m. So I think for you, we'll have m minus 1, plus m minus 2, plus-- so let me write it here. m minus 1. So U, I'm talking about U here, it's got R columns. The first one has m minus 1 because I throw away 1 because it's a unit vector, up to m minus r. That's r's column. OK. And now so how-- what does that add up to? Well, I put all the m's together. So that's rm, or let me say mr. And then I'm subtracting on 1 plus 2 plus 3, up to r. Now tell me again what that adds

up to. 1 plus 2 plus 3, stop at r. That's what we had here. And we've got it for V. And we've got it again here. Minus 1/2 r, r plus 1. Are you OK with that?

And now I just want to add them up. So I have mr. And I have nr. And then I have two of these. So let me get it here. mr and nr. And now I have to look at-- so mr, check. nr, check. Now I have two of these guys. So they combine into r squared plus r. And then I-- r squared, yeah, minus r squared plus r. Sorry. They combine into minus r squared plus r. And then here's r coming in with a plus. I think we have a minus r squared. And that is the right answer. Yeah.

OK. So I took a bit longer than I intended. But this is a number that's sort of interesting. I mentioned saddle points sort of, like, separately from maxima and minima just because they are definitely not as easy to work with. You understand what I mean by saddle points? The matrices involved have-- are not positive definite. Those would go with a maximum. They're not negati-- they're not-- well, those would go with maxima and minima. But we're looking in between. So saddle points. OK. Well, I'll get going on those.

OK. I sort of realized that there are two main sources of saddle points. One of them is when I have problems that-- when I-- let's say I minimize. So this will be the constraint. The saddle points have come from the constraint. So Lagrange is going to be responsible for these saddle points. So we might have some minimum problem like minimize, some positive definite thing. And of course, if we don't say anymore, the minimum is zero. Right? Because otherwise, it's positive.

But we're going to put on constraints, Ax equal b. So this is the classical constrained optimization problem, quadratic cost function, linear constraints. We could solve this exactly. But let's just see where saddle points is going to arise.

So this S is positive definite. But now how do we deal with that problem? Well, Lagrange said what to do. Lagrange said, look at the Lagrangium. Well, OK. He introduced lambda. This x is in n dimensions. That's an n by n matrix. But I have m constraints. So the matrix A is m by n. I've m constraints. And then I'm going to follow the rules and introduce m, Lagrange multipliers. That's an m.

And then the neat part of the Legra-- and what? What is this? Well, it's-- I take the function, and then I introduce-- remember, lambda's a vector now, not just a number. We had some application where it was just-- there was just one constraint. But now I have m constraints. So I take a lambda. So lambda transposed times Ax minus b. And the plus or the minus sign here

is not important. I mean, you can choose it because that will determine the sign of lambda. But either way, it's correct. OK.

So we've introduced a function that now depends on x and also on lambda. And there is a l-- and they multiply each other in there. And my point is that Lagrange says, take the derivatives with respect to x and lambda. So that's the cool thing that he's contributed. He says if you only create my function, now you can take x derivative and lambda derivative. That will give you n equation for the x-- from this one, from the x derivative and the m equation from the lambda derivative. It will be n plus m. It will determine the good x and the lambda. But I'm saying that's all true and all important. But I'm saying that the x and that pair x lambda will be a saddle point of this function. This function has saddle points, not a maximum.

OK. Let's just take the derivatives and see what we get. So the derivatives with respect to x, d by dx-- x is now a vector, so I really should say the gradient in the x direction. I get Sx. And here, the derivative with respect to x, what that's-- that is A transpose lambda because this is the dot product of A transpose lambda with x. You know, I've put parentheses around it and followed the transpose rule. So that's the dot product of A transpose lambda with x. It's linear in x. So it's derivative. It's just A transpose lambda. And that's zero.

And now I take the other one, the lambda derivative. The lambda derivative, this doesn't depend on lambda. The lambda derivative is just Ax minus b. It brings back the constraints. So that's pretty simple. It doesn't even require much thought because you just know the constraints are coming back. And of course, b should be put over on this side because it's a constant. So there we see two-- a block. We see an important, very important class of problems. And the matrix we're seeing, we could write this in block matrix form, S minus A transpose.

Oh, I'm going to change that into a plus because I'm more of a plus person. OK. A transpose and A, yeah, yeah. When I took the derivative with respect to lambda, I didn't put the minus sign in here. And I didn't want to. So let's make it a plus. A and then there's nothing there. And the x, and the lambda, and the zero, and the B.

That is the model of a constrained minimum, a minimum problem with constraint. It's the model because the function here is quadratic and the constraints are linear. In Course 6, it's everywhere, constantly appearing as the simplest model. OK. And my point today is just that the solution x lambda, that total solution, the x together with a lambda, that that is a saddle

point of the Lagrangian function L. It's a saddle point, not a minimum.

It's sort of a minimum in the x direction because this is positive definite. As a function of x, it's going up. But somehow the appearance of lambda makes this matrix indefinite. It starts positive definite, but it has this A transpose A and that 0.

It couldn't be p-- actually, if I look at that matrix, I see it's not positive definite. What do I see? Why do I say that immediately? When I look at that matrix, it's not a positive definite matrix because when I see that 0 on the diagonal, that shoots positive definite. Couldn't be.

Take, as an example, S equal 3, 1, and 1, 0. Take that matrix. Just random. I made it 2 by 2 instead of size m plus n. Do you see? Or how do I know that the eigenvalues of that matrix, one is plus and one is minus? The determinant is negative. So that tells me right away that one is plus and one is minus. Thanks. Yes. Yeah, yeah. The determinant is negative. And somehow here, the determinate, a similar calculation, would produce A transpose A or something with a minus because I'm going this way.

Well, I could do better than that. But you saw the point. That simple example of this has eigenvalues of both signs. Let me just quickly say, and I'll put it in the notes or in that chapter, I guess that all this is coming-- is still 3.2. That was originally 4.2. And you will see it.

So what do I want to say? I'd like to say that that example is pretty convincing to me that these KKT matrices, if you talk to people in optimization, that's Karush, Kuhn, and Tucker, three famous guys, and these are the KKT conditions that they derived following Lagrange. Right.

And my point is-- and this is a typical sort, so it's an indefinite matrix. I believe it has that if I do an elimination, yeah, tell me this. This is a good way to look at it. Suppose I do elimination on this one or on this one. Well, suppose I do elimination there. What is the first pivot? 3. Positive.

So now let me turn down to here. What if I do elimination on this block matrix? Then I start up here. And that first pivot is? Positive again, right? This S is a positive definite matrix. Don't forget. In fact, the first n pivots will all be positive because the first n pivots, you're working away in this corner. And if you're only thinking about the first n, this corner is size n by n, then you don't even see A. You're doing some subtractions. And I'll do those. But the pivots themselves are coming-- all coming from S. And S is positive definite. So we know that one of the tests for a positive definite matrix is all pivots are positive. So I think all n of the first pivots will be positive. And when we use them, let's just see what happens when we use them.

So here is the KKT matrix that I start with. And what do I end up with? Well, really, what I'm doing is I'm multiplying that block row by something to-- and subtracting to kill that A. So these rows-- well, near enough. Let me do block elimination. Block elimination is, like, easier. I don't have to write down all little tiny numbers. So I just want to multiply this row by something. Tell me what. And subtract from this second row.

Suppose they're numbers or letters. I guess they are letters. What do I multiply that first row by and subtract? Let's see. If these were just little tiny numbers, as like in 3, 1, 1, 0, what do I multiply that row by and subtract from this?

I multiply by A over S, right? I do multiply by A over S, which puts an A there. Then I subtract.

So here I'll multiply by A over S. But these are matrices, so I multiply by S-- by AS inverse, right? When I multiply by AS inverse times this S, I get A. And then I subtract. And I get the 0. And when I multiply by this guy and subtract, I get minus because I'm subtracting this thing, minus AS inverse, A transpose. That was block elimination, which just, in other words, it's just-- you've learned about 2 by 2 matrices, 3x plus 4y equals 7 and stuff. Now I'm just doing it with blocks instead of single numbers.

But you see, this produced those positive pivots. And what can you tell me about that matrix? What kind of-- what can you tell me about the signs or the eigenvalues or whatever of this matrix? Suppose S was the identity. What could you tell me about minus AA transpose? Minus AA transpose. And my voice should emphasize that minus.

It's that matrix there is negative definite. So all the next set of m pivots that come from here will all be negative. So I get m or rather n, n positive, and n negative pivots. And then I remember that the pivots actually have the same sign as the eigenvalues. That's just a beautiful fact. We know that for positive definite ones. The eigenvalues are all positive. The pivots are all positive.

But it's even better than that. If we have some mixture for the signs of the pivots, that tells us the signs of the eigenvalues. That's a really neat fact. So I'll just write that down. Plus and minus signs of pivots give us the plus and minus signs of the eigenvalues. So I've sneaked in a nice matrix there that-- for symmetric matrices. This is symmetric matrices.

OK. That's what I wanted to say about constraint and saddle points coming from there. And then I now want to say something about constraints and-- not constraints now. I'm going to

look at a second source of saddle points. So these will be saddles from this remarkable function that we know. So I now have a symmetric matrix S. Could be even positive definite. Usually, it is here.

Do you know what the name for R is? It's a ratio or a quotient. It's named after somebody starting with R. Who's that? Rayleigh. Right. It's Rayleigh quotient.

And what is the largest value, possible value of the Rayleigh quotient? We've seen this idea. It is the maximum value of that Rayleigh quotient, of that ratio, is lambda max. Right. Lambda 1, the biggest one. And the x that does it is the eigenvector. Right? So max is lambda 1 and at x equal q1 because q1 transpose Sq 1, over transpose q1. So I'm plugging in this winner. And Sq1 is lambda 1q1. Right? It's the first eigenvector. And so a lambda 1 comes out. So I get lambda 1.

I know everything about that. And what I know is if I put in any x, what do I know? If I put in any x whatever and look at this number, what do I know about that number? It's smaller than lambda 1. Or it might hit lambda 1. But it's not bigger.

That's why maxima are easy. You put in any vector, and you know what's happening. You know, it doesn't-- it's not above the max, obviously.

And what about the min? That's equally simple, of course. It's at the bottom. So what would be the minimum of that Rayleigh-- of that quotient if I was looking for what eigenvector and eigenvalue will I find when I look at the bottom of this? I will find lambda n, the last guy. Lambda min. At the winning x will be its eigenvector. And again, this stuff will equal lambda n.

So that's easy. I know that if I put in any vector whatever, just choose any vector in dimensions and compute r of x, what do I now also know about our-- that R of that vector? It's greater than lambda n. Below the max, above the min.

Now what about the other lambdas? Well, the point is that those are saddle points. The beautiful thing about this Rayleigh quotient is its derivative equals 0 right at the saddle point-- at the eigenvectors. And its value at the eigenvectors is the eigenvalue.

You see what I'm saying? I have lambda 1 here, a max, lambda n here, a min. And in between I have a bunch of other lambdas, which are saddle points. And if I put an x into r of x and look to see what happens, I have no idea whether I'm on this side, below it, or this side, above

lambda i. So the saddle points are more difficult and take a little more patience.

So that's the other source of saddle points. Let me just emphasize again what I'm saying. At lambda at x equal qk, I have some number the-- r of x has some number of positive eigenvalues and some number of negative ones for the things above and below qk.

OK. I've run out of time to follow up on the saddle point part of the-- on the details of this picture. That will be on the notes. And I might come back to it at the very start of next time.

Before that, you will have the lab number three. And then I think we should discuss it because I haven't done this lab. It's intended to give you some feeling for overfitting and also intended to give you a little introduction to deep learning. And so I'll get it to you, and we can talk about it Wednesday. And again, it won't be due until the Wednesday after break.

OK. Thanks. So I'll see you Wednesday.