GILBERT STRANG: So I've got a list of things I'm hoping to do today. I'll begin with a few final words about saddle points. The reason I'm interested in saddle points is when we get to this deep learning direction, you know that the big step there is finding a minimum of the total cost function and gradient descent, which we'll certainly discuss as the usual method or stochastic gradient descent. And all kinds of issues arise, what happens if you have a saddle point or a degenerate minimum? All these possibilities and the understanding of deep learning is focusing more and more on what does that the gradient descent algorithm produce.

So I just thought minima and maxima we know about. Saddle points are kind of a little hazier. So this is a perfect example. And I'll just say a few more words about it.

Then, I want to talk about the Lab 3 that I boldly posted on the Stellar and also about projects, just to get us thinking about those. And then my real math topic for today and this week is basic ideas of statistics, particularly the covariance matrix. I'm sure you've met mean and variance. Those are the most used words. And we'll use them again. But then I want to go on to covariance. So that's what's coming today-- a few words on saddle points, a lot of words about the lab and anything you want to ask about projects, and then some basic statistics.

OK, saddle point so the example I'm taking is this Rayleigh quotient. And I'm taking a simple matrix S. I might as well take a diagonal matrix. It's symmetric, of course. And any symmetric matrix I could just change variables by a cube matrix, an orthogonal matrix to get to something like that.

And then the x, we're in 3D. So we got a sort of manageable size here. And the x vector is uvw. So this is the quotient. x transpose x, you see is just exactly 5 u squareds, 3 v squareds, and 1 w squared. And I divide by the length to normalize things.

So what are the main facts that we know that I'm not going to prove, but what are the main facts? What's the maximum value of R? What's the minimum value of R, of that function? And is there a saddle point? So saddle of R.

OK, what's the maximum value? How large could you make that ratio, capital R? I just think, you know, this isn't a standard topic in 18.06. But with an example like this, you'll see the whole point.

OK, so how large could I make R? Yeah, go ahead and say it.

AUDIENCE: Sigma.

GILBERT STRANG: Sigma 1. And what is it here? Let's just do with these numbers. How big can I make that ratio R? And what choice of uvw makes it big? So how big I can get it is?

AUDIENCE: 5.

GILBERT STRANG: 5. That ratio can't be more than 5. You see it would be 5.

Well, how do I get to 5? Maximum of R is 5. And what is the uvw that-- so I'll say at-- what choice of uvw would give us 5 here? You see it immediately, 1, 0, 0.

And what about the minimum of R? The minimum of this ratio, how do I make that ratio small? Well, I put I load stuff on the w instead of loading it up on to u. It's just clear. So what is the minimum value of R?

AUDIENCE: 1.

GILBERT STRANG: 1, because I'll load everything into w. So the minimum value will be 1. And that will be at the vector 0, 0, 1. I've loaded everything there.

And then the point of this short discussion is, is there another place where the derivatives, first derivatives, of R are all zero? That's where, of course, the first derivatives are 0 at the max, at the min. But we also have three variables here. And we're going to find a third point.

And what is that point? You probably guess. And what will be the saddle value? So you have to see some kind of a surface that-- I guess, what are we in? 4D. So we have base coordinates, uvw. And R goes vertically. And we plot that surface. And we don't really understand it unless we think a lot about it, which we haven't. But we can pretty well guess what's what.

And so what do you think is the saddle value? And where is it going to be reached? Everybody is going to tell me correctly. Saddle value would be? Three at this middle point.

And what are these three with respect to the matrix? They're its eigenvectors. What are these three numbers with respect to the matrix? There its eigenvalues. That's why that Rayleigh quotient is such an important function.

It's kind of a messy function. If you take its derivative, you've got to use the quotient rule or use Lagrange multiplier. That's the way to make it more manageable. But it's kind of messy.

But the results could not be better. The values there are the eigenvalues. And the places where you reach them are the eigenvectors. And so the max is the most important. So that's sigma 1. Here's sigma 3-- or lambda and sigma, because the matrix is symmetric positive definite. And here in the middle is sigma 2.

And if we want to compute eigenvectors, which I'm not planning to do today, just to make this remark-- computing eigenvectors, getting the largest one or the smallest one is a lot quicker in general than getting these ones in the middle. You have to use good codes and pay attention to computing those saddle point values.

So is there anything nice I can do with saddle points? How does one think about saddle points? So again, saddle point is defined by first derivatives equals 0. That's-- and the second derivatives-- OK, so that's a matrix. Here's a vector, the gradient vector. The derivative with respect to u, the derivative with respect to v, NDR DW, just a vector. And all those components are zero. The gradient vector is zero.

But what about second derivatives? Well, that's getting more. There are nine of those now, because I've got Ruu, second derivative with respect to u. But I've got mixed derivative. Second derivative of R with respect to Du Dv. So I have a 3 by 3 matrix.

Fortunately, that matrix is symmetric, because we're blessed by that wonderful fact that the derivative with respect to u and then v is the same as v and then u. So we get a symmetric matrix. Well, I won't write it down, but it's got the maximum, minimum, and saddle information built in.

Here's this one additional thought that I want to communicate about saddle points, because it's really nice to somehow get back to maxima and minima. So this idea for a saddle point is to be able to write it as the maximum of a minimum. So let me do that, and then I'm all done.

So I'm going to say that lambda 2, that value, is the maximum of over something of the minimum over something of x trans of our function. Now, of course, I have to tell you what you're maximizing over and what you're minimizing over. But that's the idea is that one way to get into the middle place there where the saddles are sitting is to have a maximum of a minimum.

And that leads-- that's what I'm about to complete here-- would lead you, for example, very quickly to the interlacing theorem that I spoke about for eigenvalues and for singular values of when you perturb S or when you throw away a row and column of S. The eigenvalues go in between. That is the kind of conclusion that this max min stuff is set up to produce.

So here, let me just tell you what it would be. I'm aiming to get lambda 2. So I'm going to take a maximum over two dimensional spaces-- subspaces of R3. We're in 3D. So you can see that's sort of like 2-dimensional spaces. Let me give that subspace a name like V. That'll do. Cap V, everybody can see that that's a cap V. And then this will be the minimum over V.

So it's is kind of tricky. So I take any subspace that's two dimensional. And I'll take one in a moment. And I'll figure out the minimum.

Well, suppose I take this subspace V, which is spanned by the first two-- it's supposed to be a 2D subspace-- spanned by the first. Suppose I try example. The span of 1, 0, 0, and 0, 1, 0. In other words, all vectors u, v, 0. That's a 2D space.

What is the minimum of that Rayleigh quotient over that two-dimensional space? So now I'm taking a minimum. I don't have to think about saddle points. So I'm looking at the thing, but w is zero now. Everybody sees that I've squeezed it down to 2D. So w is zero. So what is the minimum now?

So this thing would become-- for this space-- would become the 5u squared and the 3v squared over the u squared plus the v squared, but the W is 0. So what's the minimum of that?

AUDIENCE: 3.

GILBERT STRANG: 3. 3. OK, the minimum is 3 for this particular space. Let me call it V special. For that particular space, the minimum is 3, correct? Everybody sees that. Because I just have u and v to play with, the 5 and the 3. So if I put everything into v, I get to 3.

And now, I take the maximum. So the maximum is at least 3, because this particular choice of V gave me the answer 3. And now, I'm taking the maximum over all possible 2D spaces. And I got 3 for one of the possible spaces of V. And I might get higher than 3 for some other one. But actually I don't. The truth is that this turns out to be 3 which is, of course, exactly what we wanted.

So I'm saying that this particular two-dimensional space, the minimum of over that, minimum, the minimum there is 3. And now, I maximize over all others. And so the idea is that for any other one, the minimum value will be below 3. And, therefore, when I go for the max of the mins, I get 3. So I just repeat that and then be quiet about this whole subject.

So it's a maximum over subspaces of a minimum of the Rayleigh quotient. If that subspace is exactly the perfect choice, this one, I get the value 3. And I'm claiming that's the biggest value I can get, because if I pick any other subspace-- what if I picked a subspace that-- suppose another v would be all vectors 0, v, w. What would I get for the minimum of this thing? But now w is in the picture and u is not in the picture. What I get for the minimum there?

AUDIENCE: 1.

GILBERT STRANG: 1. I'd get 1. The minimum would be when I put everything into w and I got one. And then when I take the max, it's not a winner. It's thrown out. The winner will be the that space and the 3. So I guess I'm hoping that you sort of see in this small example that you can express this middle saddle value as-- it's reasonable to think of it as a maximum in some directions and a minimum in another.

Think of the-- well, try to think of some surface, which is going up in some direction. So it's a minimum in those directions. And it's going down in other directions. So it's a max in those directions. And saddle points is perched in there right at that place, at the saddle point. You know, if you're like hiking from here to California or something, you're going to pass a saddle point.

Actually, you see it on the Mass-- the Mass Pike has an amazing little sign. I don't know if you've noticed it. If you drive west on the Mass Pike, pretty far west of Boston, there's a little sign telling you the altitude or elevation, whatever. And it says there this is the highest point until you reach the Rockies basically. I'd say like, OK, Midwest is pretty flat, right? Because that's a long way away. You don't think of Massachusetts as like really in the big league with

high spots. But there it is. It's the highest one until you get-- and I think it tells you where the next one will be in Colorado.

Anyway, those highest points tend to be saddles. The very, very highest point-- where's that in Alaska or somewhere-- that's a max, of course, by definition. But there are a lot of saddle points in other places. And those would be maxima of minima or minima of maxima.

Good. I'm stopping there. We might see this again when we start gradient descent. But at least, because saddle points don't come up much in teaching calculus, I thought that was good.

OK, second point is models, Lab 3, and projects, anything you'd like to ask about projects. So, please, this is your chance to ask. You could also ask by email. If you have suggested or idea for our project, let me encourage you or a team to work on it or just yourself.

And if you want to think, OK, shall I get some feedback of does this sound sensible? Any suggestions? Send me an email. I'd be happy to-- of course, I'm a total beginner here, too. When I created this Lab 3, I was like desperate, not for model 1. For model 1, have you looked at-- it's reached Stellar. And it's only one printed page. Have people had a look at this?

So I'll just repeat quickly. Model 1 is an example where of overfitting. And what's going on with model 1? So model 1 says take-- 5 would be enough, but I probably said 10 or something. So I'll make it six points. And put a curve through them. So if you put a curve-- and the curve is going to be a polynomial.

So we're going to set a fit by polynomial. Everybody knows polynomial is C0 plus C1x plus whatever CK x to the K, let's say. For K equals 0-- well, I don't know if I even ask 0. That would be the best straight line. That would run along the average.

K equal to 1, that would be a straight line fit. And you would compute that by least squares, because of course no straight line is going to go through all the points. You're going to have some error by least squares.

2 would be fitting by a parabola. Again, you'll have some error, but smaller since parabolas include straight lines. So you can only reduce the total sum of squares error by going to a degree two.

Degree 3 and up to-- how many points shall we take? 1, 2-- let me just use the same letter I've

used here. Well, m is the number of points, but m varies between 1 point I guess and probably n point. Maybe K here. Up to n-- up to 6, let's say. And I want to make a comment about 6.

No, 5 would do it. Degree 5 will fit the 6 points that we've got 6 points here. But if I stop at degree 5, I was better there, because degree 5 polynomial also has a constant term. So it really has six coefficients. So there is a 1 degree 5 polynomial with six numbers, six coefficients, that goes through those six points. And so it's a perfect fit. That would be an exact fit of the data.

So here's the data. Create a polynomial of degree 5 that goes through those points exactly. And look at the result. And what would you see if you look at the result? Would it be smooth? Of course, it's a polynomial.

Would it be nice? No, it will be horrible. To get through those points-- did I get six points? Yeah. To get through those points, I'm guessing that that fifth degree polynomial, the perfect fit, is an example that occurs to practically everybody of overfitting, because making that decision, perfect fit, learn the data, training data exactly will send a polynomial-- I don't know what it looks like. I don't want to-- well, I do want to know, but not right now. Anyway, craziness.

And, of course, I'm going to ask-- it doesn't look like that probably-- I'm going to ask you to plot the results. Well, what's the least squares error when you fit by a straight line? When you fit by horizontal line, a constant, fit by a straight line, move up to parabolas, move up to a cubics?

But when you hit this, you're not making any error at all. You're not really needing to use least squares. You can solve Ax equals B. Ac equal b. So this is the b thing. And c is the vector of coefficients. And the matrix A is bad news when it's a 6 by 6, when you get up to a complete fit.

And I guess what I wanted just to see is-- a lot of things I don't know, like suppose I change six to 20 or something. Then I'm pretty sure that out there at 18, 19, 20, this thing is really off the map. And you could compute its max and you'd see a very big number. But, of course, for a straight line, that would be pretty safe. The slope would be pretty moderate.

And I don't know where you-- so that's probably underfitting to try to fit this by a straight line. It's not as close as you would want. But fitting by a full perfect filling, a high degree polynomial, is certainly overfitting. Where is the boundary? I'm sure people know about this, but I think it is

something we could learn from. So that's what that model 1 is about.

And just to make one final comment, that matrix A has a name in the case where it's a square matrix, where you're fitting exactly-- interpolating would be the word. So that exact fit, that corresponds to square matrix A. And the word for it is interpolation. And I guess it's Lagrange again. Seeing that guy too often here. So it would be Lagrange interpretation.

But the matrix has a different name. And whose name is associated with that matrix?

AUDIENCE: Vandermonde.

GILBERT STRANG: Vandermonde. Vandermonde. So this is the square matrix, which was-- so let me write it. It's called a Vandermonde matrix. And it's a matrix that has a crazy large inverse, because just as I'm saying, the C that comes out from the perfect fit, from the interpolation, from the square matrix, the C is going to be giant. And so you will construct a matrix, of course, to do this, and it will be identical to the-- so we've heard this word Vandermonde matrix in this class within the last week. Anybody remember where the word Vandermonde came up in class?

It was in professor Townsend's lecture. So you could go back to that video if you wanted as an example of a matrix which had a horrible inverse, a giant matrix. The Hilbert matrix was another example. I think he did two examples, Vandermonde and Hilbert.

So this Vandermonde matrix-- I could write it down, but I'll leave that to you-- has a big inverse. And its eigenvalues-- well, no singular values, because it's not symmetric-- it's singular values are way scattered. It has tiny little singular values plus an ordinary sized singular values.

So that's the example that I just think you could go with. And as far as I can see, sending it to auto grader as a Julia file, it would be even worse than usual, sending it to auto grader. I think it wouldn't know what to do as far as I can see. So I'm thinking of submissions coming to Gradescope. And I'm thinking of some plots to show what happens as K increases and some tables of data maybe and then maybe a paragraph of conclusion, like what degree is safe and when does it become risky and when does it become disaster. So stuff like that. Really, these are sort of open-ended labs and you use any language.

Questions about that example? Which is really that's what I'm expecting to be ready and quite a good example for Wednesday after the break. Question? Anyway, you can email me. You probably see that see what the model looks like.

Then the second one I've taken that first jump into, networks. Made a very simple network without any hidden layers at all actually. And just wrote down what I think might work. But you may find that you want to modify model 2. Go for it. I don't have any patent or personal stake in the way model 2 is written.

But the idea is fit data-- well, start with data, but don't make it too perfect, because we want some learning to happen here. So it's the classification problem. So it won't be least squares with variables like u and v and w. It's just plus 1 or minus 1, or 1, 0, or cat and dog, whatever that classification is. So that's the basic problem to start with in deep learning. For quite a long time, that's the natural problem.

So it's a classification problem. And the description here suggests one way to set up that training data and execute a neural net like experiment, but without getting very far away from ordinary linear algebra. So as I say, if you want to change this, develop it further, get some ideas about it, that's what the whole point is here.

Actually, the faculty meeting this week, maybe today-- what's today?

AUDIENCE: Wednesday.

GILBERT STRANG: Wednesday? Yeah, so it's this afternoon. And the faculty doesn't come to much. Of course, it's late in the afternoon. But faculty meeting this afternoon is about MIT's plans for requirements or courses in computational thinking. And in a way, this course within the math department is among the ones that are in that direction. Of course, in other departments, those are further along.

Anyway, when Raj Rao taught the course last spring, he had the Julia system better developed. And it was a chance to bring computers and bring laptops and do things in class. And you'll have that chance again when he visits in a month. OK, enough.

And I'm open to questions about the project. Should I maybe ask you to email me a rough idea of a project? And tell me if you are in a group or if you would like to find a group, maybe two or three people. I'm not thinking of groups of 50. Two or three would be sensible.

Questions about projects? I mean, I just introduced this idea of our project, and I apologize for not bringing it up the first week. But I just couldn't see-- I don't want to do exams on linear algebra. We passed that point. So this seemed the right way to go. But I'm not looking for a

PhD thesis here. Questions? Thoughts? I guess I hope you know you can ask. Yeah, oh good.

AUDIENCE: So could you maybe describe the scope of the project?

GILBERT STRANG: Right. How will I-- yeah, so the scope is connected to the time that you would devote to it. And what should I say about scope? Maybe the equivalent of three homeworks or something. Because I'll tamp down homeworks as project date gets closer. Does that give an idea? So it's not infinite, but it's not something tiny and trivial. Yeah, good.

AUDIENCE: Do you have an example projects of what were done in past years?

GILBERT STRANG: Well, that's the thing. There aren't really past years. We are the ones. So I will have next year if contribute some good ideas. Maybe I should I ask Professor Rao to maybe send us the projects he uses in Michigan? That would be some ideas. But remember that he hasn't, up to now anyway, moved the course toward deep learning. He did other topics, all of which would be fine. But quite a few people have had some 6.036 or know something about conventional neural nets. And I'm certainly excited to get to that topic.

So the project could get there or it could not. Both totally fine. OK, that's a good idea. I'll ask Raj for just the projects-- and you'll recognize a couple, because you've done a couple. But there are a bunch more. Then there was another question or thought?

And I'm remembering that I think maybe everybody got an email or a Stellar announcement that some members of the class took an initiative, which was wonderful, to open the possibility of people just showing up one evening a week in the Media Lab was it? Or was there a location? And has it happened or is it a future event?

AUDIENCE: It happened.

GILBERT STRANG: It happened. But I hadn't mentioned it in class, so probably you didn't have-- and we're not really into projects yet, so it was probably a quiet evening?

AUDIENCE: Yep.

GILBERT STRANG: Yeah. Yeah, and that's--

AUDIENCE: Productive but quiet.

GILBERT STRANG: Productive but quiet. OK. So will it happen again?

AUDIENCE: Sure, I think maybe now we'll be looking after spring break.

GILBERT STRANG: After spring break. OK, so post again on Stellar the plan for the next meeting that people could come to. So this is David Anderton, so you'll recognize his name. And did you have the meeting in the Media Lab?

AUDIENCE: Yeah, we had it on the Thursday and Friday.

GILBERT STRANG: OK. So with the break coming and spring hopefully coming after today's potential storm, when we come back, good. OK. Is that good? I hope some of that is helpful. You'll get an idea. You're seeing about as much as I know, which is model 1 is definitely doable and very significant. And Vandermonde matrices and so on are truly important. And their instability is a big issue. But then moving toward weights and training data and test data is where we want to go. Good. OK.

So do I have some time? I do just to speak about mean and variance, the two golden words of statistics-- and covariance, the matrix, the intersection of linear algebra with statistics, and then some famous inequality. So I'll continue with this on Friday and post some other material. So that it's coming from a later section of notes.

OK, can I just so I either have probabilities P1 up to Pn adding to 1, or I have our continuous distribution of probabilities, maybe from all x's from minus infinity to infinity, again, giving 1. Let me work with the discrete example. That's where people naturally start.

So what is the mean? So I've n possible outcomes with those probabilities. And I can ask you about the sample mean or I can ask you about the expected mean.

So the sample means, we've done an experiment. We've got some output. The expected mean means we know probabilities, but we haven't used them yet. So this uses actual output. And the sample mean is simply-- shall I just say, m for mean-- well, these two are importantly different. One is something where you've done the experiment. And this is before you do the experiment. And the letters get-- maybe mu, I'll change it to mu. I don't want to use S because S gets used with variance.

So it's just the average, average output from the sample. Like I've flipped a coin a million

times. And the output was 0 or 1. So I got a million 1s and 0s. And I take the average, and I'm expecting a number like half a million. Of course, I'm thinking of a fair coin. And the law of large numbers would say that this sample mean does approach 1/2 with probability 1 as the number of samples gets larger. So sample mean is straightforward.

The expected mean means-- these are actual sample outputs. They happened, whereas the expected mean is just the-- and I'll use m for that-- it's the probability of the first output times that output plus the probability of the second output times that output plus, plus Pn xn. So that will approach that with probability 1 as this number capital N-- notice the difference.

Capital N here is the number of samples, the number of trials. And it gets big. We keep doing things more and more. This little n is the number of possible different outputs with their probabilities. And there you see it.

And, of course, in the continuous case, we would take the interval of x p of x dx. So let me just, by analogy, that you should know what the continuous version is and what the discrete version is. OK, that's the mean.

Now, for variance. Sample variance-- and, shall I say, expected variance? I don't know. Just variance is what people would usually say. I don't know if I've remembered the right word there, sample variance. Is that-- I included this topic in linear algebra book. Anyway, yeah. OK.

So what's the sample variance? So I guess I'm-- hm-- yeah, so what is the sample variance? What's the variance about anyway? What's the key point of variance?

It's the distance from the mean. So this will be a distance from the sample mean, and this will be a distance from the expected mean. So not distance from zero, but distance from mu and m, from the center of the thing.

So the sample variance-- so, again, we have n samples. But for some wonderful reason in statistics you divide by n minus 1 this time. And the reason has to do with the fact that you used one-- this will involve the mean. So this would be the first output minus mu squared up to the n-th output minus mu squared.

So it's the average distance from mu-- average squared distance from mu-- but with this little twist that, of course, when n is large, it's not a very significant difference between n and n minus 1. I Think that's about right. All this is that I'm just doing one experiment over and over.

Covariance, which is the deeper idea, is where linear algebra comes in. I have a matrix because why? Because I'm doing multiple experiments at the same time. I'm flipping two coins. I'm flipping 15 coins. I'm doing other things. So that will be covariances when I'm doing several experiments at once. That will involve matrices of that size.

So what's the variance? I should have given you the usual notation. The expected value of x, that's what's the mean. And here, I'm looking at the expected value of what? So when I'm computing a variance, using probabilities-- so I'm using expectations, not trial runs-- expectations means use the probabilities. And it's the expectation of the distance from x to the mean squared.

And that is when I'm doing is an expectation for a discrete set, I think of the probability, the first probability, that goes with an output x1 and a second probability that goes with an output x2. And each time I subtract from the mean and square. So that's the variance that everybody calls sigma squared.

Now, two minutes left is enough to say a few more words about covariance. Oh, to get to covariance, I really have to speak about joint probabilities. That's a key idea-- joint probabilities. So I'm doing two experiments at once. So each one has its own probabilities.

But together, I have to ask-- so here are two easy cases. Suppose I'm flipping two coins. So that I might get heads heads. I may get heads tails, tails heads, or tails tails, four possibility, four possible outputs there, four possible pairs. And if you're flipping one coin and I'm flipping another one, those are independent results. Those are independent results. There won't be a covariance where by knowing what my flip was I would know more about your flip.

But now, the other possibility would be to glue the coins together. Now, if I do a flip, they always come up heads and heads or tails and tails. So the heads tails combination is not possible. In fact, one output is totally dependent on the other outputs. So that's the other extreme. We have independent outputs with covariance 0, and we have totally dependent outputs when the things are just glued together, when one result tells us what the other result is, then that's a situation where the covariance is a maximum. It couldn't be bigger than that.

And say in polling, if you were polling a family, say political polling, well, there would be some covariance expected there. The two or three or five people that are living in the same house wouldn't be independent, entirely independent, but nor would all five give the same answer. So their covariance matrix would have some off diagonal, but it would still be invertible.

And actually, what I wanted to tell you about next time at the start is that covariance matrix, which I have to define for you, will be symmetric positive definite, or semidefinite. What's the semidefinite case? Of course, that's the case where the coins are glued together.

OK, thanks. So you know what's coming Friday. I know that holiday is also coming Friday. And just you'll make a good plan, and I'll move on after the break. Good.