GILBERT STRANG: OK, let me start one minute early. So this being MIT, I just came from a terrific faculty member, Andy Lowe in the Sloan School, and I have to tell you what he told us. And then I had to leave before he could explain why it's true, but this is like an amazing fact which I don't want to forget, so here you go. Everything will be on that board.

So it's an observation about us or other people, maybe not us. So suppose you have a biased coin. Maybe the people playing this game don't know, but it's 75% likely to produce heads, 25% likely to produce tails. And then the player has to guess for one flip after another heads or tails, and you get $1 if you're right, you pay $1 if you're wrong. So you just want to get as many right choices as possible from this coin flip that continues.

So what should you do? Well what I hope we would do is we would not know what the probabilities were, so we would guess maybe heads the first time, tails the second time, heads the third time, and so on. But the actual result would be mostly heads, so we would learn at some point that-- maybe not quite as soon as that. We would eventually learn that we should keep guessing heads, right? And that would be our optimal strategy, to guess heads all the time.

But what do people actually do? They start like this, the same way, and then they're beginning to learn that heads is more common. So maybe they do more heads than tails, but sometimes tails is right, and then after a little while, they maybe see that it's-- yeah. Well maybe they're not counting, they're just operating like ordinary people.

And what do ordinary people actually do in the long run? You would think guess heads every time, right? But they don't. In the long run, people and maybe animals and whatever guess heads three quarters of the time and tails one quarter of the time. Isn't that unbelievable? They're guessing tails a quarter of the time when the odds are never changing. Anyway, that's something that economists and other people have to explain, and if I had been able to stay another hour, I could tell you about the explanation.

Oh, I see I've written that on a board that I have no way to bury, so it's going to be there, and it's not the subject of 18.065 but it's kind of amazing. All right, so there's good math problems everywhere. OK. Can I just leave you with what I know, and if I learn more, I'll come back to that question. OK.

Please turn attention this way, right? Norms. A few words on norms, like that should be a word in your language. And so you should know what it means and you should know a few of the important norms. Again, a norm is a way to measure the size of a vector or the size of a matrix or the size of a tensor, whatever we have. Or a function. Very important.

A norm would be a function like sine x. From 0 to pi, what would be the size of that function? Well if it was 2 sine x, the size would be twice as much, so the norm should reflect that. So yesterday, or Wednesday, I told you that-- so p equals 2, 1, actually infinity, and then I'm going to put in the 0 norm with a question mark because you'll see that it has a problem. But let me just recall from last time. So p equal to 2 is the usual sum of squares square root. Usual length of a vector.

p equal 1 is this very important norm, so I would call that the l1 norm, and we'll see a lot of that. I mentioned that it plays a very significant part now in compressed sensing. It really was a bombshell in signal processing to discover-- and in other fields, too-- to discover that some things really work best in the l1 norm. The maximum norm has a natural part to play, and we'll see that, or its matrix analog.

So I didn't mention the l0 norm. All this lp business. So the lp norm, for any p, is you take the pth power-- to the pth power. Up here, p was 2. And you take the pth root. So maybe I should write it to the 1/p. Then that way, taking pth powers and pth roots, we do get the norm of 2v has a factor 2 compared to the norm of v.

So p equal to 2, you see it. We've got it right there. p equal 1, you see it here because it's just the sum of the absolute values. p equal to infinity, if I move p up and up and up, it will pick out-- as I increase p, whichever one is biggest is going to just take over, and that's why you get the max norm. Now the zero norm, where I'm using that word improperly, as you'll see.

So what is the zero norm? So let me write it [INAUDIBLE] It's the number of non-zero components. It's the thing that you'd like to know about in question of sparsity. Is there just one non-zero component? Are there 11? Are there 101? That you might want to minimize that

because sparse vectors and sparse matrices are much faster to compute with. You've got good stuff.

But now I claim that's not a norm, the number of non-zero components, because how does the norm of 2v compare with the norm of v, the zero norm? It would be the same. 2v has the same number of non-zeros as v. So it violates the rule for a norm. So I think with these norms and all the p's in between-- so actually, the math papers are full of, let p be between 1 and infinity, because that's the range where you do have a proper norm, as we will see.

I think the good thing to do with these norms is to have a picture in your mind. The geometry of a norm is good. So the picture I'm going to suggest is, plot all the vectors, let's say in 2D. So two-dimensional space, R2. So I want to plot the vectors that have v equal 1 in these different norms.

So let me ask you-- so here's 2D space, R2, and now I want to plot all the vectors that their ordinary l2 lengths equal 1. So what does that picture look like? I just think a picture is really worth something. It's a circle, thanks. It's a circle. It's a circle. This circle has the equation, of course, v1 squared plus v2 squared equal 1.

So I would call that the unit ball for the norm or whatever is a circle. OK, now here comes more interesting. What about in the l1, though? So again, tell me how to plot all the points that have v1 plus v2 equal 1. What's the boundary going to look like now? It's going to be, let's see. Well I can put down a certain number of points. There up at 1 and there at 1 and there at minus 1 and there at minus 1. That would reflect the vector 1, 0 and this would reflect the vector 0, minus 1. So yeah.

OK. So those are like four points easy to plot. Easy to see the l1 norm. But what's the rest of the boundary here? It's a diamond, good. It's a diamond. We have linear set equal to 1. Up here in the positive quadrant, it's just v1 plus v2 equal to 1, and the graph of that is a straight line. So all these guys-- this is all the points with v1 plus v2 equal 1. And over here and over here and over here. So the unit ball in the l1 norm is a diamond.

And that's a very important picture. It reflects in a very simple way something important about the l1 norm and the reason it's just exploded in importance. Let me continue, though. What about the max norm? v max or v infinity equal to 1. So again, let me plot these guys, and these guys are certainly going to be in it again because 0 [INAUDIBLE] plus or minus i and plus or minus j are good friends.

What's the rest of the boundary look like now? Now this means max of the v's equal to 1. So what are the rest of the points? You see, it does take a little thought, but then you get it and you don't forget it. OK, so what's up? I'm looking. So suppose the maximum is v1. I think it's going to look like that, out to 1, 0 and up to 0, 1. And up here, the vector would be 1.4 or something, so the maximum would be 1. Is that OK?

So somehow, what really sees, as you change this number p, you start with p equal to 1, or a diamond, and it kind of swells out to be a circle at p equal to 2, and then it kind of keeps swelling to be a square and p equal to infinity. That's an interesting thing. And yeah.

Now what's the problem with the zero norm? This is the number of non-zeros. OK, let me draw it. Where are the points with one non-zero? So I'm plotting the unit ball. Where are the vectors in this thing that have one non-zero? Not zero non-zero. So that's not included.

So what do I have? I'm not allowed the vector 1/3, 2/3 because that has two non-zeros, so where are the points with only one non-zero? Yeah, on the axes, yeah. That tells you. So it can be there and there. Oops, without that guy. And of course those just keep going out.

So it totally violates the-- so maybe the point that I should make about these figures-- so like, what's happening? When I go down to zero-- and really, that figure should be at the other end, right? Oh no, shoot. This guy's in the middle. This is a badly drawn figure. l2 is kind of the center guy. l1 is at one end, l infinity is at the other end, and this one has gone off the end at the left there.

The diamond has-- yeah, what's happened here, as that one goes down towards zero, none of these will be OK. These balls or these sets will lose weight. So they'll always have these points in, but they'll be like this and then like this and then finally in the unacceptable limit, but none of those-- this was not any good either. This was for people equal 1/2, let's say. That's a p equal to 1/2 and that's not a good norm. Yeah.

So maybe the property of the circle, the diamond, and the square, which is a nice math property of those three sets and is not possessed by this. As this thing loses weight, I lose the property. And then of course it's totally lost over there. Do you know what that property would be? It's what? Concave, convex.

Convex, I would say. Convex. This is a true norm as the convex unit. Well maybe for ball, I'm

taking all the v's less or equal to 1. Yeah, so I'm allowing the insides of these shapes. So this is not a convex set. That set, which I should maybe-- so not convex would be this one like so. And that reflects the fact that the rules for a norm are broken in the triangle. Inequality is probably broken in the-- other stuff, yeah. I think that's sort of worth remembering.

And then one more norm that's natural to think about is-- so S, as in the Piazza question, S does always represent a symmetric matrix in 18.065. And now my norm is going to be-- so I'm going to call it the S norm. So actually, it's a positive definite symmetric matrix. S is a positive definite symmetric matrix.

And what do I do? I'll take v transpose Sv. OK, what's our word for that? The energy. That's the energy in the vector v. And I'll take the square root so that I now have the length of two if I double v, from v to 2v. Then I got a 2 here and a 2 here, and when I take the square root, I get a overall 2 and that's what I want. I want the norm to grow linearly with the two or three or whatever I multiply by.

But what is the shape of this thing? So what is the shape of-- let me put it on this board. I'm going to get a picture like that. So what is the shape of v transpose Sv equal 1 or less or equal 1? This is a symmetric positive definite. People use those three letters to tell us. I'm claiming that we get a bunch of norms.

When do we get the l2 norm? What matrix S would this give us the l2 norm? The identity, certainly. Now what's going to happen if I use some different matrix S? This circle is going to change shape. I might have a different norm, depending on S. And a typical case would be S equal 2, 3, say. That's a positive definite symmetric matrix.

And now I would be drawing the graph of 2v1 squared plus 3v2 squared. That would be the energy, right? Equal 1. And I just want you to tell me what shape that is. So that's a perfectly good norm that you could check all its properties. They all come out easily. But I get a new picture-- a new norm that's kind of adjustable. You could say it's a weighted norm. Weights mean that you kind of have picked some numbers sort of appropriate to the particular problem.

Well, suppose those numbers are 2 and 3. What shape is the unit ball in this S norm? It's an ellipse, right. It's an ellipse. And I guess it will actually be-- so the larger number, 3, will mean you can't go as far as the smaller number, 2. I think it would probably be an ellipse like this, and the axes length of the ellipse would probably have something to do with the 2 and the 3.

OK, so now you know really all the vector norms that are sort of naturally used.

These come up in a natural way. As we said, the identity matrix brings us back to the two norm. So these are all sort of variations on the two norm. And these are variations as p runs from 1 up to 2 on to infinity and is not allowed to go below 1. OK, that's norms.

And then maybe you can actually see from this picture-- here is a, like, somewhat hokey idea of why it is that minimizing the area in this norm-- so what do I mean by that? Here would be a typical problem. Minimize, subject to Ax equal b, the l2-- sorry, I'm using x now-- the l1 norm of x. So that would be an important problem. Actually, it has a name. People have spent a lot of time thinking of a fast way to solve it.

It's almost like least squares. What would make it more like least squares would be, change that to 2. Yeah. Can I just sort of sketch, without making a big argument here, the difference between l equal 1 or 2 here. Yeah, I'll just draw a picture. Now I'll erase this ellipse, but you won't forget. OK.

So this is our problem. With l1, it has a famous name, basis pursuit. Well famous to people who work in optimization. For l2, it has an important name. Well it's sort of like least squares. Ridge regression. This is like a beautiful model problem. Among all solutions to Ax, suppose this is just one equation, like c1x1 plus c2x2 equals some right side, b. So the constraint says that the vectors x have to be on a line. Suppose that's a graph of that line.

So among all these x's, which one-- oh, I'm realizing what I'm going to say is going to be smart. I mean, it's going to be nice. Not going to be difficult. Let's do the one we know best, l2. So here's a picture of the line. Let me make it a little more tilted so you-- yeah, like 2, 3. OK.

This is the xy plane. Here's x1, here's x2. Here are the points that satisfy my condition. Which point on that line minimizes-- has the smallest l2 norm? Which point on the line has the smallest l2 norm? Yeah, you're drawing the right figure with your hands.

The smallest l2 norm-- l2, remember, is just how far out you go. It's circular here, so it doesn't matter what direction. They're all giving the same l2 norm, it's just how far. So we're looking for the closest point on the line because we don't want to go any further. We want to go a minimum distance with-- I'm doing l2 now.

So where is the point at minimum distance? Yeah, just show me again once more, with hands

or whatever. It'll be that. I didn't want 45 degree angles there. I'm going to erase it again and really-- this time, I'm going to get angles that are not 45 [INAUDIBLE] All right, brilliant. Got it. OK, that's my line.

OK, and what's the nearest point in the l2 norm? Here's the winner in l2, right? The nearest point. Everybody sees that picture? So that's a basic picture for minimizing something with a constraint, which is the fundamental problem of optimization, of neural nets, of everything, really. Of life. Well I'm getting philosophical.

But the question always is, and maybe it's true in life, too, which norm are you using? OK, now that was the minimum in l2. That's the shortest distance, where distance means what we usually think of it as meaning. But now, let's go for the l1 norm. Which point on the line has the smallest l1 norm?

So now I'm going to add the 2. So if this is some point a, 0 and this is some point 0, b right there. So those two points are obviously important. And that point, we could figure out the formula for because we know what the geometry is. But I've just put those two points in. So did I get a 0, b? Yeah, that's a zero.

So let me just ask you the question. What point on that line has the smallest l1 norm? Which has the smallest l1 norm? Somebody said it. Just say it a little louder so that you're on tape forever.

AUDIENCE: 0, b.

GILBERT STRANG: 0, b, this point. That's the winner. This is the l1 winner and this was the l2 winner. And notice what I said earlier, and I didn't see it coming, but now I realize this is a figure to put in the notes. The winner has the most zeros. It's the [? sparsest ?] vector. Well out of two components, it didn't have much freedom, but it has a zero component. It's on the axes. It's the things on the axes that have the smallest number of components.

So yeah, this is the picture in two dimensions. So I'm in 2D. And you can see that the winner has a zero component, yeah. And that's a fact that extends into higher dimensions too and that makes the l1 norm special, as I've said. Yeah. Is there more to say about that example?

For a simple 2D question, that really makes the point that the l1 winner is there. It's not further. You don't go further up the line, right? Because that's bad in all ways. When you go up further, you're adding some non-zero first component and you're increasing the non-zero second

component, so that's a bad idea. That's a bad idea. This is the winner.

And in a way, here's the picture. Oh yeah. I should prepare these lectures, but this one's coming out all right anyway. So the picture there is the nearest ball hits at that point. And what is it? Can you see that? So that star is outside the circle. This is the l1 winner and that's the blow up the l1 norm until it hits. That's the point where the l1 norm hits.

Do you see it? Just give it a little thought, that another geometric way to see the answer to this problem is, you start at the origin and you blow up the norm until you get a point on the line that satisfies your constraint. And because you were blowing up the norm, when it hits first, that's the smallest blow-up possible. That's the guy that minimizes. Yeah, so just think about that picture and I'll draw it better somewhere, too.

Well that's vector norms. And then I introduce some matrix norms, and let me just say a word about those. OK, a word about matrix norms. So the matrix norms were the-- so now I have a matrix A and I want to define those same three norms again for a matrix. And this was the 2 norm, and what was the 2 norm of a matrix?

Well it was sigma 1, it turned out to be. So that doesn't define it. Or we could define it. Just say, OK, the largest singular value is the 2 norm of the matrix. But actually, it comes from somewhere. So I want to speak about this one first, the 2 norm.

So it's the 2 norm of a matrix, and one way to see the 2 norm of a matrix is to connect it to the 2 norm of vectors. I'd like to connect the 2 norm of matrices to the 2 norm of vectors. And how shall I do that? I think I'm going to look at the 2 norm of Ax over the 2 norm of x.

So in a way, to me, that ratio is like the blow-up factor. If A was seven times the identity, to take an easy case-- if A is seven times the identity, what will that ratio be? Say it, yeah. Seven. If A is 7i, this will be 7x and this will be x, and norms, the factor seven comes out, so that ratio will be seven. OK.

For me, the norm is-- that's the blow-up factor. So here's the idea of a matrix norm. Now I'm doing matrix. Matrix norm from vector norm. And the answer will be the maximum blow-up. The maximum of this ratio. I call that ratio the blow-up factor. That's just a made-up name. The maximum over all x. All of x.

I look to see which vector gets blown up the most and that is the norm of the matrix. I've

settled on norms of vectors. That's done upstairs there. Now I'm looking at norms of matrices. And this is one way to get a good norm of a matrix that kind of comes from the 2 norm. So there would be other norms for matrices coming from other vector norms, and those, we haven't seen, but the 2 norm is a very important one.

So what is the maximum value of this? Of that ratio for a matrix A? The claim is that it's sigma 1. I'll just put a big equality there. Now, can we see, why is sigma 1 the answer to this problem? I can see a couple of ways to think about that but that's a very important fact. In fact, this is a way to discover what sigma 1 is without all the other sigmas. If I look for the x that has the biggest blow-up factor-- and by the way, which x will it be? Which x will win the max competition here and be sigma 1 times as large as-- the ratio will be sigma 1. That will be sigma 1. When is this thing sigma 1 times as large as that? For which x?

Not for an eigenvector. If x was an eigenvector, what would that ratio be? Lambda. But if A is not a symmetric matrix, maybe the eigenvectors don't tell you the exact way they go. So what vector would you now guess? It's not an eigenvector, it is a singular vector. And which singular vector is it probably going to be? v1. Yeah, v1 makes sense. Winner.

So the winner of this competition is x equal v1, the first right singular vector. And we better be able to check that. So again, this maximization problem, the answer is in terms of the singular vector. So that's a way to find this first singular vector without finding them all. And let's just plug in the first singular vector and see that the ratio is sigma 1.

So now let me plug it in. So what do I have? I want Av1 over length of v1. OK. And I'm hoping to get that answer. Well what's the denominator here? The length of v1 is one. So no big deal there. That's one.

What's the length of the top one? Now what is Av1? If v1 is the first right singular vector, than Av1 is sigma 1 times u1. Remember, the singular vector deals were Av equals sigma u. Avk equals sigma k uk. You remember that. So they're not eigenvectors. They're singular vectors.

So Av1 is the length of sigma 1 u1 and it's divided by 1. And of course, u1 is also a unit vector, so I just get sigma 1. OK. So that's another way to say that you can find sigma 1 by solving this maximum problem. And you get that sigma 1. OK. And I could get other matrix norms by maximizing that blow-up factor in that vector norm. I won't do that now, just to keep control of what we've got.

Now what was the next matrix norm that came in last time? Very, very important one for deep learning and neural nets. Somehow it's a little simpler than this guy. And what was that matrix norm? What letter whose name goes here? Frobenius. So capital F for Frobenius. And what was that?

That was the square root of the sum of all the-- add all the aij squares, for all over the matrix, and then take the square root. And then somebody asked a good question after class on Wednesday, what has that got to do with the sigmas? Because my point was that these norms are the guys that go with the sigmas, that have nice formulas for the sigmas, and here it is. It's the square root of the sum of the squares of all the sigmas. So let me write Frobenius again. But this notation with an F is now pretty standard, and we should be able to see why that number is the same as that number.

Yeah. I could give you a reason or I could put it on the problem set. Yeah, I think that's better on the problem set, because first of all, I get off the hook right away, and secondly, this connection between-- in Frobenius, that's a beautiful fact about Frobenius norm that you add up all the sigma squares-- it's just m times n of them because it's a filled matrix. So another way to say it is, we haven't written down the SVD today, A equal u sigma v transposed.

And the point is that, for the Frobenius norm-- actually, for all these norms-- I can change u. It doesn't change the norm, so I can make u the identity. u, as we all know, is an orthogonal matrix, and what I'm saying is, orthogonal matrix u doesn't change any of these particular norms. So suppose it was the identity. Same here. That could be the identity without changing the norm.

So we're down to the norm of Frobenius. So what's the Frobenius norm of that guy? What's the Frobenius norm of that diagonal matrix? Well you're supposed to add up the squares of all the numbers in the matrix and what do you get? You get that, right? So that's why this is the same as this because the orthogonal guy there and the orthogonal guy there make no difference in the norm. But that takes checking, right? Yeah. But that's another way to see why the Frobenius norm gives this.

And then finally, this was the nuclear norm. And actually, just before my lunch lecture on the subject of probability-- I've had a learning morning. The lunch lecture was about this crazy way that humans behave. Not us but other humans. Other actual-- well, no, I don't want to say that. Take that out of the tape.

Yeah, OK. Anyway, that was that lecture, but before that was a lecture for an hour plus about deep learning by somebody who really, really has begun to understand what is happening inside. How does that gradient descent optimization algorithm pick out, what does it pick out as the thing it learns. This is going to be our goal in this course. We're not there yet.

But his conjecture is that-- yeah, so it's a conjecture. He doesn't have a proof. He's got proofs of some nice cases where things commute but he hasn't got the whole thing yet, but it's pretty terrific work. So this was Professor Srebro who's in Chicago. So he just announced his conjecture, and his conjecture is that, in a modeled case, the deep learning that we'll learn about with the gradient descent that we'll learn about to find the best weights-- the point is that, in a typical deep learning problem these days, there are many more weights than samples and so there are a lot of possible minima. Many different weights give the same minimum loss because there are so many weights. The problem is, like, got too many variables, but it turns out to be a very, very good thing. That's part of the success.

And he believes that in a model situation, that optimization by gradient descent picks out the weights that minimize the nuclear norm. So this would be a norm of a lot of weights. And he thinks that's where the system goes. We'll see this. This comes up in compressed sensing, as I mentioned last time. But now I have to remember what was the definition.

Do you remember what the nuclear norm? He often used a little star instead of an N. I'll put that in the notes. Other people call it the trace norm. But I think this N kind of gives it a notation you can remember. So let's call it the nuclear norm. Do you remember what that one was? Yeah, somebody's saying it right.

Add the sigmas, yeah. Just the sum of the sigmas, like the l1 norm, in a way. So that's the idea, is that this is the natural sort of l1 type of norm for matrices. It's the l1 norm for that sigma vector. This would be the l2 norm of the sigma vector. That would be the l infinity norm. Notice that the vector numbers, infinity, 2, and 1, get changed around when you look at the matrix guy.

So that's an exciting idea and it remains to be proved. And expert people are experimenting to see, is it true? Yeah. So that's a big thing for their future. Yes.

OK, so today, we've talked about norms and this section of the notes will be all about norms. We've taken a big leap into a comment about deep learning and this is what I want to say the most. And I say it to every class I teach near the start of the semester. My feeling about what

my job is to teach you things, or to join with you in learning things, as happened today. It's not to grade you. I don't spend any time losing sleep-- you know, should that person take a one point or epsilon penalty for turning it in four minutes late? To Hell with that, right? We've got a lot to do here.

So anyway, we'll get on with the job. So homework three coming up, and you'll be using the notes that you already have posted in Stellar for those sections eight and nine and so on. And we'll keep going on Monday. OK, see you on Monday and have a great weekend.