GILBERT STRANG: So I'm going to talk about the gradient descent today to get to that central algorithm of neural net deep learning, machine learning, and optimization in general. So I'm trying to minimize a function. And that's the way you do it if there are many, many variables, too many to take second derivatives, then we settle for first derivatives of the function.

So I introduced, and you've already met the idea of gradient. But let me just be sure to make some comments about the gradient and the Hessian and the role of convexity before we see the big crucial example. So I've kind of prepared over here for this crucial example. The function is a pure quadratic, two unknowns, x and y, pure quadratic. So every pure quadratic I can write in terms of a symmetric matrix s.

And in this case, x1 squared was bx2 squared, the symmetric, the matrix is just 2 by 2. It's diagonal. It's got eigenvalues 1 and b sitting on the diagonal. I'm thinking of b as being the smaller one. So the condition number, which we'll see, is all important in the question of the speed of convergence is the ratio of the largest to the smallest.

In this case, the largest is 1 the smallest is b. So that's 1 over b. And when 1 over b is a big number, when b is a very small number, then that's when we're in trouble.

When the matrix is symmetric, that condition number is lambda max over lambda min. If I had an unsymmetric matrix, I would probably use sigma max over sigma min, of course. But here, matrices are symmetric.

We're going to see something neat is that we can actually take the steps of steepest descent, write down what each step gives us, and see how quickly they converge to the answer. And what is the answer? So I haven't put in any linear term here.

So I just have a bowl sitting on the origin. So of course, the minimum point is x equal 0, y equals 0. So the minimum point x star, is 0, 0, of course.

So the question will be how quickly do we get to that one. And you will say pretty small example, not typical. But the terrific thing is that we see everything for this example.

We can see the actual steps of steepest descent. We can see how quickly they converge to the x star, the answer, the place where this thing is a minimum. And we can begin to think what to do if it's too slow.

So I'll come to that example after some general thoughts about gradients, Hessians. So what does the gradient tell us? So let me just take an example of the gradient. Let me take a linear function, f of xy equals say, 2x plus 5y. I just think we ought to get totally familiar with these.

We're doing something. We're jumping into an important topic. When I ask you what's the gradient, that's a freshman question. But let's just be sure we know how to interpret the gradient, how to compute it, what it means, how to see it geometrically.

So what's the gradient of that function? It's a function of two variables. So the gradient is a vector with two components. And they are? The derivative of this factor x, which is 2 and the derivative of this factor y, which is 5.

So in this case, the gradient is constant. And the Hessian, which I often call H after Hessian, or del squared F would tell us we're taking the second derivatives, that will be the second derivatives obviously 0 in this case.

So what shape is H here? It's 2 by 2. Everybody recognizes 2 by 2 is H would have the-- I'll take a second derivative of that-- sorry, the first derivative of that with respect to x, obviously 0, the first derivative with respect to y, the first derivative of that with respect to x y. Anyway, Hessian 0 for sure.

So let me draw the surface. So x, y, and the surface, if I graph F in this direction, then obviously, I have a plane. And I'm at a typical point on the plane let's say. Yeah, yeah. So I'm at a point x, y, I should say. I'm at a point x, y. And let me put the plane through it. So how do I interpret the gradient at that particular point x, y? What does 2x plus 5y tell me? Or rather what does grad F tell me about movement from that point x, y? Of course, the gradient is constant. So it really didn't matter what point I'm moving from. But taking a point here. So what's the deal if I move? What's the fastest way to go up the surface? If I took the plane that went through that point x, y, what's the fastest way to climb the plane? What direction goes up fastest? The gradient direction, right? The gradient direction is the way up. How am I going to

put it in this picture? I guess I'm thinking of this plane as-- so what plane? You could well ask what plane have I drawn? Suppose I've drawn the plane 2x plus 5y equals 0 even? So I'll make it go through the arc. And I've taken a typical point on that plane. Now if I want to increase that function, I go perpendicular to the plane. If I want to stay level with the function, if I wanted to stay at 0, I stay in the plane. So there are two key directions. Everybody knows this. I'm just repeating. This is the direction of the gradient of F out of the plane, steepest upwards. This is the downwards direction minus gradient of F, perpendicular to the plane downwards. And that line is in the plane. That's part of the level set. 2x plus 5y equals 0 would be a level set. That's my pretty amateur picture. Just all I want to remember is these words level and steepest, up or down. Down with a minus sign that we see in steepest descent. So where in steepest descent. And what's the Hessian telling me about the surface if I take the matrix of second derivatives? So I have this surface. So I have a surface F equal constant. That's the sort of level surface. So if I stay in that surface, the gradient of F is 0. Gradient of F is 0 in-- on-- on is a better word-- on the surface. The gradient of F points perpendicular. But what about the Hessian, the second derivative? What is that telling me about that surface in particular when the Hessian is 0 or other surfaces? What does the Hessian tell me about-- I'm thinking of the Hessian at a particular point. So I'm getting 0 for the Hessian because the surface is flat. If the surface was convex upwards from-- if it was a convex or a graph of F, the Hessian would be-- so I just want to make that connection now. What's the connection between the Hessian and convexity of the-- the Hessian of the function and convexity of the function? So the point is that convexity-- the Hessian tells me whether or not the surface is convex. And what is the test?

AUDIENCE: [INAUDIBLE].

GILBERT STRANG: Positive definite or semi definite. I'm just looking for an excuse to write down convexity and strong. Do I say strict or strong convexity? I've forgotten. Strict, I think.

Strictly convex. So convexity, the Hessian is positive semi-definite, or which includes-- I better say that right here-- includes positive definite. If I'm looking for a strict convexity, then I must require positive definite. H is positive definite. Semi-definite won't do.

So semi-definite for convex. So that in fact, the linear function is convex, but not strictly convex. Strictly means it really bends upwards. The Hessian is positive definite. The curvatures are positive.

So this would include linear functions, and that would not include linear function. They're not strictly convex. Good, good, good.

Some examples-- OK, the number one example, of course, is the one we're talking about over here. So examples f of x equal 1/2 x transpose Sx. And of course, I could have linear terms minus a transpose x, a linear term. And I could have a constant. OK.

So this function is strictly convex when S is positive definite, because H is now S for that function, for that function H. Usually H, the Hessian is varying from point to point. The nice thing about a pure quadratic is its constant. It's the same S at all points.

Let me just ask you-- so that's a convex function. And what's its minimum? What's the gradient, first of all? What's the gradient of that? I'm asking really for differentiating thinking in vector, doing all n derivatives at once here. I'm asking for the whole vector of first derivatives.

Because here I'm giving you the whole function with x for vector x. Of course, we could take n to be 1. And then we would see that if n was 1, this would just be Sx squared, half Sx squared. And the derivative of a half Sx squared-- let me just put that over here so we're sure to get it right-- half of Sx squared.

This is in the n equal 1 case. And the derivative is obviously Sx. And that's what it is here, Sx.

It's obviously simple, but if you haven't thought about that line, it's asking for all the first derivatives of that quadratic function. Oh! It's not-- What do I have to include now here? That's not right as it stands for the function that's written above it. What's the right gradient?

AUDIENCE: [INAUDIBLE].

GILBERT STRANG: Minus a, thanks. Because the linear function, its partial derivatives are obviously just the components of a. And the Hessian H is S, derivatives of that guy. OK. Good. Good, good, good.

And the minimum value-- we might as well-- oh yeah! What's the right words for a minimum value? No, I'm sorry. The right word is minimum value like f min. So I want to compute f min.

Well, first I have to figure out where is that minimum reached? And what's the answer to that? We're putting everything on the board for this simple case. The minimum of f of f of f of x-- remember, it's x is-- we're in n dimensions-- is at x equal what? Well, the minimum is where

the gradient is 0.

So what's the minimizing x? S inverse a, thanks. Sorry. That's not right. It's here that I meant to write it.

Really, my whole point for this little moment is to be sure that we keep straight what I mean by the place where the minimum is reached and the minimum value. Those are two different things. So the minimum is reached at S inverse a, because that's obviously where the gradient is 0. It's the solution to Sx equal a.

And what I was going to ask you is what's the right word-- well, sort of word, made up word-- for this point x star where the minimum is reached? So it's not the minimum value. It's the point where it's reached. And that's called-- the notation for that point is

AUDIENCE: Arg min.

GILBERT STRANG: Arg min, thanks. Arg min of my function. And that means the place-- the point where f equals f min. I haven't said yet what the minimum value is.

This tells us the point. And that's usually what we're interested in. We're, to tell the truth, not that interested in a typical example and what the minimum value is as much as where is it? Where do we reach that thing? And of course, so this is x min.

This is then arg min of my function f. That's the point. And it happens to be in this case, the minimum value is actually 0. Because there's no linear term a transpose x.

Why am I talking about arg min when you've all seen it? I guess I think that somebody could just be reading this stuff, for example, learning about neural net, and run into this expression arg min and think what's that? So it's maybe a right time to say what it is. It's the point where the minimum is reached.

Why those words, by the way? Well, arg isn't much of a word. It sounds like you're getting strangled. But it's sort of short. I assume it's short.

Nobody ever told me this. I assume it's short for argument. The word argument is a kind of long word for the value of x. If I have a function f of x, f, I call it function and x is the argument of that function. You might more often see the word variable.

But argument-- and I'm assuming that's what that refers to, it's the argument that minimizes

the function. OK, good. And here it is, S inverse a. Now but just by the way, what is f min? Do you know the minimum of a quadratic?

I mean, this is the fundamental minimization question, to minimize a quadratic. Electrical engineering, a quadratic regulator problem is the simplest problem there. There could be constraints.

And we'll see it with constraints included. But right now, no constraints at all. We're just looking at the function f of x.

Let me to remove the b, because that just shifts the function by b. If I erase that, just to say it didn't matter. It's really that function. So that function actually goes through 0. As it is, when x is 0, we obviously get 0.

But it's still on its way down, so to speak. It's on its way down to this point, S inverse a. That's where it bottoms out. And when it bottoms out, what do you get for f? One thing I know, it's going to be negative because it passed through 0, and it was on its way below 0.

So let's just figure out what that f min is. So I have a half. I'm just going to plug in S inverse a, the bottom point into the function, and see where the surface bottoms out and at what level it bottoms out. So I have a half. So that's S inverse a is a transpose S inverse.

S symmetric, so I'll just write this inverse transpose. S, S inverse a from the quadratic term, minus a transpose. And x is S inverse a. Have you done this calculation? It just doesn't hurt to repeat it.

So I've plugged in S inverse a there, there, and there. OK, what have I got? Well, S inverse cancels S. So I have a half of a transpose S inverse a minus 1 of a transpose inverse a. So I get finally negative a half.

Half of it minus one of it of a transpose S inverse a. Sorry, that's not brilliant use of the blackboard to squeeze that in there. But that's easily repeatable. OK, good. So that's what a quadratic bowl, a perfect quadratic problem minimizes to that's its lowest level.

Ooh, I wanted to mention one other function, because I'm going to speak mostly about quadratics, but obviously, the whole point is that it's the convexity that's really making things work. So here, let me just put here, a remarkable convex function. And the notes tell what's the gradient of this function. They don't actually go as far as the Hessian.

Proving that this function I'm going to write down is convex, it takes a little thinking. But it's a fantastic function. You would never sort of imagine it if you didn't see it sometime. So it's going to be a function of a matrix, a function of-- those are n squared variables, x, i, j. So it's a function of many variables.

And here is this function. It's you take the determinant of the matrix. That's clearly a function of all the n squared variables. Then you take the log of the determinant and put in a minus sign because we want convex. That turns out to be a convex function.

And even to just check that for 2 by 2 well, for 2 by 2 you have four variables, because it's a 2 by 2 matrix. We could maybe check it for a symmetric matrix. I move it down to three variables. But I'd be glad anybody who's ambitious to see why that log determinant is a remarkable function. And let me see.

So the gradient of that thing is also amazing. The gradient of that function-- I'm going to peek so I don't write the wrong fact here. So the partial derivative of that function are the entries of-- these are the entries of a, a inverse. That's the-- of x inverse.

That's like, wow. Where did that come from? It might be minus the entries, of course. Yeah, yeah, yeah. So we've got n squared function-- what is a typical entry in x inverse?

What does a typical x inverse i, j? Just to remember that bit of pretty old fashioned linear algebra, the entry is of the inverse matrix, I'm sure to divide by what? The determinant, that's the one thing we know. And that's the reason we take the log, because when you take derivatives of a log, that will put determinant of x in the denominator.

And then the numerator will be the derivatives of the determinant of x. Oh! Can we get any idea what are the derivatives of the determinant? Oh my god. How did I never get into this? So are you with me so far?

This is going to be derivatives of determinant, the strength of all these variables divided by the determinant, because that's what the log achieved. So when I take the derivative of the log of something, that chain rule says take the derivative of that something divide by the function determinant of x.

So what's the derivative of the determinant of a matrix with respect to its 1, 1 entry? Yeah, sure. This is crazy. But it's crazy to be doing this. But it's healthy. OK.

So I have a matrix x, da, da, da, x, x, 1, 1, x, 1n, et cetera, xn, 1, x, n, n. OK. And what am I looking for? I'm looking for that for the derivatives of the-- do I want the derivatives of the determinant?

Yes. So what's the derivative of x of the determinant with respect to the first equals what? How can I figure out? So what's this asking me to do? It's asking me to change x, 1, 1 by delta x and see what's the change in the determinant.

That's what derivatives are. Change x, 1, 1 a little bit. How much did the determinant change? What has the determinant of the whole matrix got to do with x, 1, 1? You remember that there is a formula for determinants.

So I need that fact. The determinant of x is x, 1, 1 times something. Is that something that I really want to know? Plus x, 1, 2 times other something plus say, along the first row times another something. What are these factors that multiply the x's to give the determinant?

What [INAUDIBLE] a linear combination of the first row time certain factors gives the determinant? And how do I know that there will be such factors, because the fundamental property of the determinant is that it's linear in row 1 if I don't mess with other rows. It's a linear function of row 1. So it has a form x, 1, 1 times something. And what is something?

AUDIENCE: [INAUDIBLE].

GILBERT STRANG: The determinant of this. So what does x, 1, 1 multiply when you compute determinants? X, 1, 1 will not multiply any other guys in its row, because you're never multiplying two x's in the same row or the same column. What x, 1, 1 is multiplying all these guys. And in fact, it turns out to be is the determinant.

And what is this called? That one smaller determinant that I get by throwing away the first row and first column? It's called a-- Minor is good. Yes, minor is good.

I was saying there are two words that can be used, minor and co-factor. Yeah. And what is it? I mean, how do I compute it? What is the number?

This is a number. It's just a number. Maybe I think of the minor as this determinant-- Ah! Let me cancel that. Maybe I think of the minor as this smaller matrix, and the co-factor, which is the determinant of the minor.

And there is a plus or minus. Everything about determinants, there's a there's a plus or minus choice to be made. And we're not going to worry about that. But so anyway, so it's the co-factor. Let me call it C, 1, 1.

And so that's the formula for a determinant. That's the co-factor expansion of a determinant. OK. And that will connect back to this amazing fact that the gradient is the entries of x inverse, because the inverse is the ratio of co-factor to determinant.

So x inverse 1, 1 is that co-factor over the determinant. Yeah. So that's where this all comes from.

Anyway, I'm just mentioning that as a very interesting example of a convex function. OK. I'll leave that. That's just for like, education. OK.

Now I'm ready to go to work on gradient descent. So actually, the rest of this class and Friday's class about gradient descent are very fundamental parts of 18.065. And that will be one of our examples. And then the general case here.

So I'm using this. It would be interesting to minimize that thing, but we're not going there. Let's hide it, so we don't see it again. And I'll work with that example. So here's gradient descent.

Is xk plus 1 is xk minus Sk the step size times the gradient of f at xk. So the only thing left that requires us to input some decision making is a step size, the learning rate. We can take it as constant. If we take too big a learning rate, the thing will oscillate all over the place and it's a disaster. If we take too small a learning rate, too small steps, what's the matter with that?

Takes too long. Takes too long. So the problem is to get it just right. And one way that you could say get it right would be to think of optimize. Choose the optimal Sk.

Of course, that takes longer than just deciding an Sk in advance, which is what people do. So I'll tell you what people do is on really big problems is take an Sk-- estimate a suitable Sk, and then go with it for a while. And then look back to see if it was too big, they'll see oscillations. It'll be bouncing all over the place.

Or of course, an exact line search-- so you see that this expression often. The exact line search choose Sk to make my function f at xk plus 1 a minimum on the line, on the search line, a minimum in the search direction. The search direction is given by the gradient. That's the direction we're moving.

This is the distance we're moving, or measure of the distance we're moving. And an exact search would be to go along there. If I have a convex function, then as I move along this line, as I increase Sk, I'll see the function start down, because the gradient, negative gradient means down. But at some point it'll turn up again. And an exact line search would find that point and stop there.

That doesn't mean we would-- we will see in this example where we will do exact line searches that for a small value of b, it's extremely slow, that the condition number controls the speed. That's really what my message will be just in these last minutes and next time the sort of key lecture on gradient descent.

So an exact line search would be that. So what a backtracking line search-- backtracking would be take a fixed S like one. And then be prepared to come backwards. Cut back by half. See what you get at that point. Cut back by half of that to a quarter of the original step. See what that is. So the full step might have taken you back to the upward sweep. Halfway forward it might still be on the upward sweep. Might be too much, but so backtracking cuts the step size in pieces and checks until it-- So S0, half of S0, quarter of S0, or obviously, a different parameter, aS0, a squared S0, and so on until you're satisfied with that step. And there are of course, many, many refinements. We're talking about the big algorithm here that everybody has, depending on their function, has different experiences with.

So here's my fundamental question. Let's think of an exact line search. How much does that reduce the function? How much does that reduce the function? So that's really what the bounds that I want are.

How much does that reduce the function? And we'll see that the reduction involves the condition number, m over M. So why don't I turn to the example first? And then where we know exact answers.

That gives us a basis for comparison. And then our math goal is prove-- get S dead bounds on the size of f that match what we see exactly in that example where we know everything. We know the gradient. We know the Hessian. It's that matrix.

We know the condition number. So what happens if I start at a point x0 y0 that's on my surface? Sorry. What do I want to do here? Yeah.

I take a point, x0 y0 and I iterate. So the new xy k plus 1 is xyk minus the S, which I can

compute times the gradient of f. So I'm going to put in gradient f. What is the gradient here? The derivative is we expect to x.

So I have a 2xk and 2by. And this is the step size. And for this small problem where we're going to get such a revealing answer, I'm going to choose exact line search. I'm going to choose the best xk. And what's the answer?

So I just want to tell you what the iterations are for that particular function starting at x0 y0. So let me put start x0 y0. And I haven't done this calculation myself.

It's taken from the book by Steven Boyd and Vandenberghe called *Convex Optimization.* Of course, they weren't the first to do this either. But I'm happy to mention that book *Convex Optimization.* And Steven Boyd will be on campus this spring actually, in April for three lectures. This is April, maybe.

Yeah, OK. So it's this month in two or three weeks. And I'll tell you about that. So here are the xk's and the yk's and the f and the function values.

So where am I going to start? Yeah. So I'm starting from the point x0 y0 equal b1. Turns out that will make our formulas very convenient, x0 y0 equals b1. Good.

So OK. So xk is b times the key ratio b minus 1 over b plus 1 to the kth power. And yk happens to be-- it has this same ratio. And my function f has the same ratio too.

This is fk. It has that same ratio 1 minus b over 1 plus b to the kth times f0. That's the beautiful formula that we're going to take as the best example possible.

Let's just see. If k equals 0, I have xk equal b yk equal 1 b starting at b1. And that tells me the rate of decrease of the function. It's this same ratio. So what am I learning from this example?

What's jumping out is that this ratio 1 minus b over 1 plus b is crucial. If b is near 1, that ratio is small. If b is near 1, that's near 0 over 2. And I converge quickly, no problem at all. But if b is near 0, if my condition number is bad-- so the bad case, the hard case is small b.

Of course, when b is small, that ratio is very near 1. It's below 1. The ratio is below 1, so I'm getting convergence. I do get convergence. I do go downhill. But what happens is I don't go downhill very far until I'm headed back uphill again.

So the picture to draw for this-- let me change that picture to a picture in the xy plane of the

level sets. So the picture really to see is in the xy plane. The level sets f equal constant. That's what a level set is. It's a set of points, x and y where f has the same value.

And what do those look like? Oh, let's see. I think-- what do you think? What do the level sets look like for this particular function?

If I look at the curve x squared plus b y squared equal a constant, that's what the level set is. This is x squared plus by squared equal a constant. What kind of a curve is that?

AUDIENCE: [INAUDIBLE].

GILBERT STRANG: That's an ellipse. And what's up with that ellipse? What's the shape of it? Because there is no xy term, that ellipse is like, well lined up with the axes. The major axes of the ellipse are in the x and y directions, because there is no cross term here.

We could always have diagonalized our matrix if it wasn't diagonal. And that wouldn't have changed anything. So it's just rotating this space. And we've done that.

What do the levels set look like? They're ellipses. And suppose b is a small number, then what's with the ellipses? If b is small, I have to go pretty-- I have to take a pretty large y to match a-- change an x. I think maybe they're ellipses of that sort.

Are they? They're lined up for the axes. And I hope I'm drawing in the right direction. They're long and thin. Is that right?

Because I would have to take a pretty big y to make up for a small b. OK. So what happens when I'm descending? This is a narrow valley then. Think of it as a valley which comes down steeply in the y direction, but in the x direction I'm crossing the valley slow--

Oh, is that right? So what happens if I take a point there? Oh yeah, I remember what to do. So let's start at that point on that ellipse. And those were the levels sets f equal constant.

So what's the first search direction? What direction do I move from x0 y0? Do I move along the ellipse? Absolutely not, because along the ellipse f is constant. The gradient direction is perpendicular to the ellipse.

So I move perpendicular to the ellipse. And when do I stop? Pretty soon, because very soon I'm going back up again.

I haven't practiced with this curve. But I know-- and time is up, thank God. So what do I know is going to happen? And by Friday we'll make it happen? So what do we see for the curve, the track of the-- it's say it?

AUDIENCE: Zigzag.

GILBERT STRANG: It's a zigzag, yeah. We would like to get here, but we're not aimed here at all. So we zig, zig, zig zag, and very slowly approach that point. And how slowly? With that multiplier, 1 minus b over 1 plus b.

That's what I'm learning from this example, that that's a key number. And then you could ask, well, what about general examples? This was one specially chose an example with exact solution. Well, we'll see at the beginning of next time that for a convex function this is typical. This is 1 minus b is the critical quantity, or 1 over b, or the how small is b compared to 1?

So that will be the critical quantity. And we see it in this ratio 1 minus b over 1 plus b. So if b is 100, this is 0.99 over 1.01. It's virtually 1. OK.

So next time is a sort of a key lecture to see what I've just said, that this controls the convergence of steepest descent, and then to see an idea that speeds it up. That idea is called momentum or heavy ball. So the physical idea is if you had a heavy ball right there and wanted to get it down the valley toward the bottom, you wouldn't go perpendicular to the level sets. Not at all. You'd let the momentum of the ball take over and let it roll down.

So the idea of momentum is to model the possibility of letting that heavy ball roll instead of directing it by the steepest descent at every point. So there's an extra term in steepest descent, the momentum term that accelerates. OK. So Friday is the day. Good. See you then.