GILBERT STRANG: Well, OK, I am happy to be back, and I am really happy about the project proposals that are coming in. This is like, OK, this is really a good part of the course. And so keep them coming, and I'm happy to give whatever feedback I can on those proposals, and do make a start there. They're really good, and if some are completed before the end of the semester and we can to offer you a chance to report on them, that that's good too. So well done with those proposals.

So today, I'm jumping to part six. So part six and part seven are optimization which is the fundamental algorithm that goes into deep learning. So we've got to start with optimization. Everybody has to get that picture, and then part seven will be the structure of CNNs, Convolution Neural Nets, and all kinds of applications.

And so can we start with optimization? So first, can I like get the basic facts about three terms of a Taylor series? So that's the typical. It's seldom that we would go up to third derivatives in optimization.

So that's the most useful approximation to a function. Everybody recognizes it. Here, I'm thinking of F as just one function, and x as just one variable, but now I really want to go to more variables. So what do I have to change if F is a function of more variables? So now, I'm thinking of x as-- well, now let me see.

Yeah, I want n variables here. x is x1 up to xn. So just to get the words straight so we can begin on optimization, so what will be the similar step so the function F at x-- remember, x is n variables. OK?

Now, what do I have? Delta x, so what's the point about delta x now? It's a vector, delta x1 to delta xn, and what about the derivative of F? It's a vector too, the derivative of F with respect to x1, the derivative of F with respect to x2, and so on.

What do I have to change about that? I know those guys are vectors, so it's their dot product. So it's delta x transpose at vector times this dF/dx. So now I'm replacing this by all the

derivatives, and it's the gradient. So the gradient of F at x is the derivatives-- let's see.

It's essential to get the notation straight here. Yeah, so it'll be the partial derivatives of the function F. So grad F is the partial derivatives of F with respect to x1 down to partial derivative with respect to xn. OK, good.

That's the linear term, and now what's the quadratic term? 1/2, now delta x isn't a scalar anymore. It's a vector. So I'm going to have delta x transpose and a delta x, and what goes in between is the second derivatives, but I've got a function of n variables.

So now, I have a matrix of second derivatives, and I'll call it H. This is the matrix of second derivatives, Hjk is the second derivative of F with respect to xj and xk, and what's the name for this guy? The Hessian, Hessian matrix.

How the Hessians got into this picture I don't know. The only Hessians I know are the ones who fought in the Revolutionary War for somebody. Who? Which side were they on? I think maybe the wrong side. The French were on our side and--

Anyway, Hessian matrix, and what are the facts about that matrix? Well, the first fact is that it's [INAUDIBLE] and the key fact is it's symmetric. Yeah. OK, and again, it's an approximation. And everybody recognizes that if n is very large, and we have a function of many variables. Then, we had n derivatives to compute here, and about 1/2 n squared derivatives.

The 1/2 comes from the symmetry, but the key point is the n squared derivatives to compute there. So computing the gradient is feasible if n is small or moderately large. Actually, by using automatic differentiation, the key idea of back propagation, back prop, you can speed up the computation of derivatives quite amazingly. But still for the size of deep learning problems that's out of reach. OK.

So that's the picture, and then I will want to use this to solve equations. There is a parallel picture for a vector f. So now, this is a vector function. This is f1 of x up to fn of x, an x is x1 to xn. So I have n functions of n variables, n functions of n variables.

Well, that's exactly what I have in the gradient. Think of these two as parallel, the parallel being f corresponds to the gradient of F, n functions of n variables. OK. Now maybe, what I'm after here is to solve f equals 0. So I'm going to think about the f at x plus delta x, so it starts with f of x.

And then we have the correction times the matrix of first derivatives, and what's the name for that matrix of first derivatives? Well, if I'm just given n functions-- yeah, what am I after here? I'm looking for the Jacobian. So here we'll go the Jacobian, J. This is the Jacobian named after Jacoby, Jacobian matrix.

And what are its entries? J, the jk entry is the derivative of the J function with respect to the kth variable, and I'm stopping at first order there. OK, so these are sort of like facts of calculus, facts of 18.02 you could say. Multivariable calculus, that's the point.

Notice that we're doing just like the first half of 18.02, just do differential calculus, derivatives, Taylor series. We're not doing multiple integrals. That's not part of our world here. OK, so that's the background.

Now, I want to look at optimization. So over here, I want to optimize-- well, over here, let me try to minimize F of x, and I'll be in the vector case here. And over here, I want to solve f equals 0, and of course, that means f of 1 equals 0 all the way along to fn equals 0. Here, I have n equations, and n unknowns.

Let me start with that one, and I'll start with Newton's method, Newton's method to solve these n equations and n unknowns. OK, so Newton, Newton's method which is often not presented in 18.02. That's a crime, because that's the big application of gradients in Jacobians.

OK, so I'm trying to solve n equations and n unknowns, and so I want f at x plus delta x to be 0. Right? So I want f of x plus delta x to be 0. So f at x plus delta x is-- I'm putting in a 0. I'm just copying that equation-- is f at where I am. Let me use K for the case iteration.

So I'm at a point xK. I want to get to a point xK plus 1. And so I have 0 is f of x plus J, at that point, times delta x which is xK plus 1 minus xK. Good. That's Newton's method.

Of course, 0 isn't quite true. Well, 0 will be true if I'm constructing xK plus 1 here. I'm constructing xK plus 1. OK. So let me just rewrite that, and we've got Newton's method. So we're looking for this change, xK plus 1 minus xK. I'll put it on this side as plus xK, so that's this.

Now, I have to invert that and put it on the other side of the equation. So that will go with a minus. This guy will be inverted and f at xK. So that's Newton's methods. It's natural.

So let me just repeat that. You see where the xK plus 1 minus xK is sitting? Right? And I

moved f of xK to the other side with a minus sign, and then I multiplied through by J inverse, so I got that. So that's Newton's method for a system of equations, and over there, I'm going to write down Newton's method for minimizing a function. This is such basic stuff that we have to begin here.

Let me even begin with an extremely straightforward example of Newton's method here. Suppose my function-- suppose I've only got one function actually. Suppose I only had one function. So suppose my function is x squared minus 9, and I want to solve f of x equals 0. I want to find the square root of 9.

OK, so what is Newton's method for it? My point is just to see how Newton's method is written and then rewrite it a little bit so that we see the convergence. OK, so of course, the Jacobian is 2x. So Newton's method says that xK plus 1-- I'm just going to copy that Newton's method-- minus 1 over 2xK. Right? That's the derivative times f at xK which is xK squared minus 9.

OK. We followed the formula, this determines xK plus 1, and let's simplify it. So here I have xK minus that looks like 1/2 of xK, so I think I have 1/2xK, and then this times this is 9/2 of 1 over xK. Is that right? 1/2 of xk from this stuff and plus 9/2 of 1 over xK. OK.

Can I just like check that I know the answer is 3? Can I be sure that I get the right answer, 3? That if xK was exactly 3, then of course, I expect xK plus 1 to stay at 3. So does that happen? So 1/2 of 3 and 9/2 of 1/3, what's that, 1/2 of 3 and 9/2 of 1/3?

OK, that's 3/2 and 3/2. That's 6/2, and that's 3. OK. So we've checked that the method is consistent which just means we kept the algebra straight. But then the really important point about Newton's method is to discover how fast it converges. So now let me do xK plus 1 minus 3.

So now, I'm looking at the error which is, I hope, approaching 0. Is it approaching 0? How quickly is it approaching 0? These are the fundamental questions of optimization.

So I'm going to subtract 3 from both sides somehow. OK, from here, I guess, I'm going to subtract 3. So I was just checking that it was correct. OK. Now, so xK plus 1 minus 3, I'm going to subtract 3 from both sides. I'm going to subtract 3 there, and then I hope that-- that box is what goes down here. Right?

Subtracted 3 from both sides, so I'm hoping now things go to 0. OK, so what do I have there? Let me factor out the 1 over xK. So what do I have then left? 1 over xK, so there's a 9/2 from

there, 1 over xK.

So I really have 1/2 of xK squared, because I've divided by an xK. And this minus 3, I better put minus 3xK, because I'm dividing by xK. I claim that that's-- now I've got it. And let's see, let me take out the 2-- 2, forget these 2s, and make that a 6. So I have 1 over 2xK times 9 plus xK squared minus 6.

Anything good about that? We hope so. We hope that that is something attractive. So this is, again, the error at set K plus 1, and it's 1 over 2xK times this thing in brackets-- 9 plus xK squared minus 6xK. And we recognize that as xK minus 3 squared.

xK squared minus 6 of them plus 9, that's xK minus 3 squared. OK, that was the goal, of course. That's the goal that shows why Newton's method is fantastic. If you can execute it, if you can start near enough, notice that-- so how do I describe this great equation? It says that the error is squared at every step, squared at every step.

So if I'm converging to a limit, it will satisfy the-- it'll be 3, or I guess minus 3, is that possible? Yeah, minus 3 is another solution here. So we've got two solutions. Newton's method could converge to 3. Am I right, it could converge to minus 3?

So I'd have a similar equation sort of centered at minus 3, or does it always do one of those? It could blow up. So there are sort of regions of attraction. They're all the starting points that approach 3, and the whole point of that equation is with quadratic convergence the error being squared at every step. It zooms in on 3.

Then, there is all the starting points that would go to minus 3, and then there are the starting points that would blow up. And those, maybe for this very simple problem, the picture is not too difficult to sort out those three regions. And this is allowing for a vector, two equations or n equations, then we're in n variables, and really you get beautiful pictures.

You get some of the type of pictures that gave rise to these books on fractals, picture books on fractals for these basins of attraction. Does the starting point lead you to one of the solutions, or does it lead you to infinity? Here, that would be interesting to just draw it for this, but the essential point is the quadratic convergence, if it's close enough.

You see that it has to be close. If x0 is pretty near 3, then this is about 1/6 of that, and there would be a good region of attraction in this case. OK. So that's Newton's method for

equations.

And now I want to do Newton's method. I just want to convert all those words over to Newton's method for optimization. So remember, these boards were solving f equals 0. This board is minimizing capital F, and what's the connection between them? Well of course, this corresponds to solving the gradient equals 0.

At a minimum, if I'm minimizing, I'm finding a point where all the first derivatives are 0. So that will be the match between these. This grad F in this picture is the small f in that picture. OK.

Now, I guess here I have-- and this is sort of the heart of our applications to deep learning-- we have very complicated loss functions to minimize, functions of thousands or hundreds of thousands of variables. OK. So that means that we would like to use Newton's method, but often we can't. So I need him to put down here two methods-- one that doesn't involve those high second derivatives and Newton's that does.

So first, I'll write down a method that does not involve, so method one, and this will be steepest descent. And what is that? That says that xK plus 1-- the new x is the old x minus-- steepest descent means that I move in the steepest direction which is the direction of the gradient of F. I move some distance, and I better have freedom to decide what that distance should be. So this is a step size, s, or in the language of deep learning, it's often called the learning rate, so if you see learning rate. OK.

So and it's natural to choose sK. We're going along, do you see what this right-hand side looks like? I'm at a point in n dimensions. We're in n dimensions here. We have functions of n variables.

There is a vector. There is a direction to move down the steepest slope of the graph. And here is a distance to move, and we will stop. We'll have to get off this step, normally. If we stay on it, it will swing back, it'll take us off to infinity.

You would like to choose sK so that you minimize capital F. You take the point on this line, so this a line in R n, a direction in R n. And for all the points on that line, in that direction, F has some value, and what you expect is that initially, because you chose it sensibly, the value of F will drop. But then at a certain point, it will turn back on you and increase.

So that would be the natural stopping point. I would call that an exact line search. So I exact line search would be, exact line search is the best s. Of course, that would take time to

compute, and you probably, in deep learning, that's time you can't afford, so you fix the learning rate s. Maybe you choose 0.01 to be pretty safe.

OK, so that's method one, steepest descent. Now, method two will be Newton's method. So now, we have xK plus 1 equal to xK minus something times delta F, and now I'm going to do the right thing. I'm going to live right here, and the right thing is the Hessian, the second derivative.

This was cheap. We just took the direction and went along it. Now, we're getting really the right direction by using the second derivative, so that's H inverse. OK, and what I've done is to set that 0.

Do you see that's Newton's method? It's totally parallel to this guy. Actually, I'm really happy to have these two on the board parallel to each other, because you have to keep straight, are you solving equations, or are you minimizing functions? And you're using different letters in the two problems, but now you see how they match.

The Jacobian of-- so again the matches, think of f as the gradient of F. That's the way you should think of it. So the Jacobian of the gradient is the Hessian. The Jacobian of the gradient is the Hessian, and that makes sense, because the first derivative of the first derivative is the second derivative. Only we're doing matrix y, so the Jacobian of the gradient-- we're doing a vector matrix sentence instead of a scalar sentence-- the Jacobian of the gradient is a Hessian. Yeah, right.

OK, so that's what I wanted to start with, just to get those basic facts down. And so the basic facts were the three-term Taylor series. And then the basic algorithms followed naturally from it by setting f F at the new point to 0, if that's what you were solving or by assuming you had the minimum. Right, good, good, good, good. OK.

Now, what? Now, we have to think about solving these problems, studying. Do they converge? What rate do they converge? Well, the rate of convergence is like why I separated off this example.

So the convergence rate for Newton's method will be quadratic. The error gets squared, and of course, that means super-fast convergence, if you start and close enough. The rate of convergence for a steepest descent is, of course, not. You're not squaring errors here, because you're just taking some number instead of the inverse of the correct matrix, so you

can't expect super speed.

So a linear rate of convergence would be right. You would like to know that the error is multiplied at every step by some constant below 1. That would be a linear rate compared to being squared at every step. OK, and so this will be our basic formula that we build on for really large scale problems.

And there are methods, of course, people are going to come up with methods that they're sort of a cheap Newton's method. Levenberg-Marquardt, and it's in the notes at the end of this section, at the end of 6.4 that we'll get to. So Levenberg-Marquardt is a sort of cheap man's Newton's method. It does not compute to Hessian, but it says, OK, from the gradient, I can see one term in the Hessian. So it grabs that term, but it's not fully second order.

OK. So now, we have to think about problems, and I guess the message here is, at our starting point, has to be convexity. Convexity is the key word for these problems, for the function that we want to minimize. If that's a convex function, well first of all, the convex function is likely to have one minimum. And the picture that's in our mind of steepest descent, this picture of a bowl, a bowl is the graph of the convex function.

So I'm turning to convexity now. I'll leave that board there, because that's pretty crucial, and speak about the idea of convexity. Convex function, convex set, so let's call the function f of x, and a typical convex set would be I'll call it K. OK. So we just want to remember what does that word can convex mean, and how do you know if you have a convex function or a convex set?

OK, let me start with convex set. So because here is my general problem, my convex minimization, which you hope to have, and in many applications, you do have. So you minimize a convex function for points in a convex set. So that's like the ideal situation. That's the ideal situation, to get something on your side, something powerful, convexity.

The function is convex, and you say, well, let me draw a convex function, the graph. OK, so I'll draw a convex function, say a bowl. So that's a graph of f of x, and then here are the x's. Let me maybe put x1 and x2 in the base and the graph of f of 1x x2 up here. OK. Actually, I'm over there.

I should be calling this function F, I think. Is that right? Yeah, a little f would be the gradient of this guy. Yeah, I think so. OK.

Now, I'm minimizing it over certain x's, not all x's. I might be minimizing, for example, K might be the set where Ax equals B. K might be, in that case, a subspace or a shifted subspace. I said subspace, but then 18.06 is reminding me in my mind that I only have a subspace when B is 0.

You know the word for a subspace that's sort of moved over? Affine, so I'll just put that word down here. Bunch of words to learn for this topic, but they're worth learning. OK.

So it's like a plane but not necessarily through the origin. If B is 0, it doesn't go through it. If B it's not 0, it doesn't go through the origin. OK. Anyway, or I have some other convex set. Let me just put this convex set K in the base for you, and did I make it convex? I think pretty luckily I did.

So now what's the? Well, the convex sets the constraint, so this is the constraint set. Constraint is that x must be in the set K. OK, and I drew it as a convex blob. Here was an example where it's flat, not a blob but a flat plane.

But let me come back to what does convex mean. What's a convex set? Yeah, we have to do that, should have done that before. In the notes, I had the fun of figuring out, if I took a triangle, is that a convex set? Let's just be sure.

So what's a convex set? That is a convex set, because if I take any two points in the set and draw the line between them, it stays in the set. So that's convexity, any edge, line, from x1 to x2 stays in the set. OK, good.

So here's my little exercise to myself. What if I took the union of two triangles? All I want to get you to do is just visualize convex and not convex possibilities. Suppose I have one triangle, even if it was obtuse, that's still convex, right? No problem.

But now what if I put those two triangles together, take their union? Well, if I take them sitting with a big gap between, like I've lost. I mean, I never had a chance that way, because if it was the union of these two-- well, you know what I'm going to say. If I'm doing that point and that point, of course, it goes outside and stupid. All right.

What if what if that triangle, that lower triangle, overlaps the upper triangle? Is that a convex set? Everybody's right saying no. Why how do I see that the union of those two triangles is not a convex set? Guys, you tell me where to pick two points, where the line goes out. Well, I take one from that corner and one from that corner, and the line between them went outside. So

union is usually not convex.

Well, if I think of the union of two sets, my mind automatically goes to the other corresponding possibility which is the intersection of the two sets. So if I take the intersection of two sets. Now, what's the deal with that? When I had two triangles, two separated triangles, what can we say about the intersection of those two triangles?

AUDIENCE: [INAUDIBLE]

GILBERT STRANG: It's empty. So should we regard the empty set as a convex set? Yes. Isn't it?

AUDIENCE: Yeah, it's vacuous.

GILBERT STRANG: Vacuous, so it hasn't got any problems. Right? OK, but now the intersection is always convex. I'm assuming the two sets that we start with are. Now, that's an important fact, that the intersection of convex sets. Let's just draw a picture that shows an example.

So what's the intersection? Just this part and it's convex. OK, can you give me a little proof that the intersection is convex? So I take two points in the intersection-- let me start the proof.

To test if something's convex, how do you test it? You take two points in the set in the intersection, and you want to show that the line between them is in the intersection. OK, why is that?

So take two points, take x1 in the intersection. We've got two sets here, and that's the symbol for intersection, and we've got another point in the intersection. And now, we want to look at the line between them, the line from x1 to 2x. What's the deal with that one? Is that fully in K1?

AUDIENCE: Yes.

GILBERT STRANG: Why is it fully in K1? I took two points in the intersection, I'm looking at the line between them, and I'm asking, is it in the first set K1? And the answer is yes, because those points were in K1, and K1's convex. And is that line between them in K2? Yes, same reason, the two endpoints were in K2, so the line between them is in K2.

So the intersection of convex sets is always convex. The intersection of convex sets is convex. Good. So you'll see in the note these possibilities with two triangles. Sometimes, you can take the union but not very often. OK.

Now, what's the next thing I have to do? Convex functions, we got convex sets, what are convex functions, and we're good. Because this is our prototype of a problem, and I now want to know what it means for that F to be-- oh, I'm sorry. I now know what it means for the set K to be convex set, but now I have to look at the other often more important part of the problem. What's the function I'm minimizing, and I'm looking for functions with this kind of a picture. OK.

The coolest way is to connect the definition of a convex function to the definition of a convex set. This is really the nicest way. It's a little quick. It just swishes by you. But tell me, do you see a convex set in that picture? [INAUDIBLE]

You see a convex set in that picture. That's the picture of a graph of a convex function. It's a picture of a bowl. Are the points on that surface, is that a convex set? No, certainly not. No, but where is a convex set to be found here, in that picture? Yes.

AUDIENCE: The set of y, if y is greater than [INAUDIBLE]

GILBERT STRANG: Yes, the points on and above the bowl, inside the bowl, we could say, these points. So convex function, yes, a function's convex when the points on and above the graph are convex set. You could say, OK, mathematicians are just being lazy. Having got one definition straight for a convex set, now they're just using that to give an easy definition of a convex function. Actually, it's quite useful for functions that could maybe equal infinity, sort of generalized functions.

But it's not the quickest way to tell if the function is convex. It's not our usual test for convex functions. So now I want to give such a test. OK. So now, the definition of convex function, of a smooth convex, yeah. This fact, I shouldn't rush off away from it, from the definition of a convex function as having a convex set above its graph. The really official French name for the set above the graph is the epigraph, but I won't even write that word down. OK.

Why do I come back to that for a minute? Because I would like to think about two functions, F1 and F2. Out of two functions, I can always create the minimum or the maximum.

So suppose I have to convex functions, convex function F1 and F2. OK. Then, I could choose a minimum. I could choose my new function. Shall I call it little m for minimum? m of x is the minimum of F1 and F2.

And I could choose a maximum function which would be the maximum of F1 of x and F2 of x at the same point x. It's just a natural to think, OK, I have two functions. I've got a bowl and I've

got another bowl, and suppose they're both convex.

So I'm just stretching you to think here. If I've got the graphs of two convex functions, and I would like to consider the minimum of those two functions and also the maximum of those two functions. I believe life is good. One of these will be convex, and the other won't.

And can you identify which one is convex and which one is not convex? What about the minimum? Is that a convex function? So just look at the graph. What does the minimum look like? The minimum is this guy until they meet somehow on some surface and then this guy.

Is that convex? We have like one minute to answer that question. Absolutely no. It's got this bad kink in it. What about the maximum of the two functions? So the maximum is the one that is above, all the points or things that are above or on.

There is the maximum function. That was the minimum function. It had a kink. The maximum function is like that, and it is convex, so maximum yes, minimum no. OK, and we could have a maximum of 1,500 functions. If the 1,500 functions are all convex, the maximum will be, because it's the part way above everybody's graph, and that would be the graph of the maximum. OK, good.

And now finally, let me just say, how do you know whether a function is convex? How to test, how of test. OK, so let me take just a function of one variable. What's the test you learned in calculus, freshman calculus actually, just show that this is a convex function? What's the test for that?

AUDIENCE: Use second derivative.

GILBERT STRANG: Second derivative should be?

AUDIENCE: Positive.

GILBERT STRANG: Positive or possibly 0, so second derivative greater or equals 0 everywhere. That's convex. OK, final question, suppose F is a vector. So this is a vector, and so I have n functions of n variable. No, I don't. I have one, sorry, I've got one function, but I'm in n variables. So this was just one.

What's the test for convexity? So it would be passed, for example, by x1 squared plus x2

squared. Would it be passed by-- so here would be the question-- would it be passed by x transpose some symmetric matrix S? That would be a quadratic, a pure quadratic.

Would it be convex? What would be the test? I'm looking for an n dimensional equivalent of positive second derivative. The n dimensional equivalent of positive second derivative is convexity, and we have to recognize what's the test. So I could apply it to this function, or I could apply it to any function of n variables. It should be OK.

What's the test here? Here, I have a matrix instead of a number. So what's the requirement going to be? Times out, yeah? [INAUDIBLE] Positive definite or semidefinite, or semidefinite just as here. Yeah.

So the test is positive, semidefinite, Hessian. And here, the Hessian is actually that S, because the second derivatives will produce-- I'll put a 1/2 in there-- the second derivatives will produce S equal the Hessian H. So here, the S-- so positive semidefinite, Hessian in general, second derivative matrix for a quadratic.

OK. So its convex problems that we're going to get farther with. We run into no saddle points. We run into no local minimum. Once we found the minimum, it's the global minimum. These are the good problems. OK, again, happy to see you today, and I look forward to Wednesday.