

The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare in general is available at ocw.mit.edu.

PROFESSOR: OK. We're ready to go. So this will be the last lecture that I give maybe, almost, and the last one that will videotape, and then Monday we start on the presentations. And we'd better get rolling on those because time will run out on us, and I want everybody who would like to to have a chance to talk about work.

OK. Can I begin? So I promised to speak about inverse problems, ill-posed problems. It's a giant subject, but there's one thing I didn't do, one beautiful thing that I didn't do well at the end of the last lecture. And if you don't mind I'd like to go back to that because the number one example for optimization is b squared, minimizing b minus Ax squared. Here's b ; here are all the Ax 's. Think of it as a line, but of course it could be a n -dimensional sub space. And then this is the winner. And the reason it's the winner is that it's a right angle, it's a true projection.

And there's a second problem, which is minimize this thing, the distance to w , where w lies on this perpendicular direction. And, of course, the closest point w to the b in this direction is there. So those are the actual winners that get a little star. And when we're there, the length of this squared plus the length of this squared equals the length of b squared. We have right angles, and Pythagoras a squared plus b squared equals c squared is right. But then so what I added just quickly at the end was the key point about suppose I take any allowed Ax and any allowed w . Then I would expect -- and I look at that distance which is larger than this, and I look at this distance which is larger than this, so of course, now I'm not looking at the optimal ones but instead looking at these solid line ones, and then the b minus Ax squared is bigger than it has to be. The b minus w squared is bigger, so the sum is b squared plus another term. So that's weak duality that for any choice some inequality holds, since this is never negative -- another way to say weak duality would be to say that this is always greater or equal to. But what I didn't throw in last time was there's this beautiful expression for what the difference is. If you don't

have the optimum, this is what you miss by. And one part of the beauty is that that tells us right away how to recognize the optimum. Duality holds, equality holds, without this term when this term is 0 when b is equal to Ax plus w , and of course that's the good dash line case when that vector plus that vector is exactly b . Ax and the w make b . Those are the dash lines that's optimal.

So somehow the optimality condition is this one, and this is a condition that connects the two problems. See here we had a minimization of this. w didn't appear. In the dual problem we had a minimization of this. x didn't appear. But when we get both problems right then they connect. And this is the step that in my three step framework for applied math that dominated 18085. There was the x , there was the b minus Ax , there was the shift over to w . So matrix A entered there, A matrix A transpose entered here to give A transpose w equals 0. And here is the bridge. That's the bridge between them which in this simple model problem is -- I usually write that bridge with a physical constant c , and in this simple model problem c was 1, or c was the identity. So then this equals this, which is our optimality condition. So when we have that bridge between one problem and the dual problem we've got them both right. So I'm happy to find that.

But I have a problem for you. You may say where did that identity come from? So that's simply an identity. And it came from just multiplying it out. If I multiply out take b minus Ax transposed times b minus Ax . Take this transposed times itself. b transposed b This transpose this. It works. The terms cancel, and this is an identity. Now that's a simple identity in geometry, so my puzzle for you is take this -- let me draw the identity here. So here's my picture. I'm going up to any w . And here is b in this picture. So b minus w is there. This says that some vector squared plus that squared equals that squared plus that squared. So now I just have to find all those vectors and ask you prove it.

So let me name all these vectors. So here is b minus Ax . Right? That's b minus Ax . And here's w . OK. Let me get them all right. I have to get b in there and here is b minus w , and here's b minus Ax . And what's the fourth one? The fourth one is this mysterious one. So I go up b minus Ax w . I'll have to put that here. I'm going up this

same w , straight up. There is the mysterious fourth guy: b minus Ax minus w is there, and this was the b minus Ax .

And this identity holds. It's gotta be like Pythagoras could have figured it out but how? Let me draw this very same picture again. So it's this line, this line, this line, and this line. And let me just call those a , b , c , and d . And they connect. Oh sorry, there's a rather important fact. So this statement here then says that a squared -- no what does it say -- a squared plus c squared maybe is the first two, and this one is b squared plus d squared and why? Please submit a proof to get an A. OK. Let me get the picture right. So this is just dotted lines here. So that's a rectangle, and there is d . So I have four lines starting at the same point. Two of the lines go to the corners of a rectangle.

So it's a geometry problem, which struck me early this morning -- too early this morning. But it must be, of course, it can't be news to the world. But I think I've got it right: a squared plus c squared equals b squared plus d squared. We take any point and connect to the four corners of a rectangle, and that holds. And the question is why? It holds from algebra, but somehow we also ought to be able to get it out of Pythagoras. So as far as I know none of these angles are special. That might look equal to that or something, but it's not, I don't think. I don't think. You know, that point is off this d point, it's anywhere. It's off center, and here's the rectangle that it's connecting to the corners of.

So there you go. Open problem. Why is that true? And I'm sure it is, so this isn't a wild goose chase, but I'm just wondering what proofs can we fight for that. OK. Can I leave you with that? I'll leave that on the board and hope you'll listen to my presentation about ill-posed problems, but I hope you copied that little picture and will either email me or hard copy, or whatever. Anyway let me know a good way to prove it.

OK. Now the lecture begins. Today's lecture begins. Oh, I had one other comment about the interior point method with the barrier problem. I got down at the end to an equation -- you remember I was taking the derivative, solving the barrier problem,

which was minimizing $c x$ minus some multiple of the barrier $\log x$ sub i . So I took the derivative of this quantity and set it to 0. And the derivative in respect to x gave me the equation c equals α over x . Anyway I kind of a lost my nerve, but all I want to say is this is right and it leads to Newton's method that we spoke about. I won't go back to that.

All right. Inverse problems. I think maybe the best thing I can do in one lecture about inverse problems is first of all to get a general picture of what are they. Secondly to mention areas that we will all know about, where these inverse problems enter. And then thirdly to look a little bit at the integral equations that often describe inverse problems. Inverse problems come from many sources, not only integral equations, but integral equations are responsible for quite a few. But let's think about others.

OK, so really I plan now to list various examples. And number one I've already spoken about find velocities from positions. Take the derivative. So taking the derivative is a process that makes things bigger, so when we go the other way we're -- so the difficulty with the problem is that a small change in the position data may be a very large change in the velocity. For example, suppose the position is the correct position, say x of t , plus some noise term that's going to be small but small in size, small in amplitude, but not small in derivative. Maybe like $\sin t$ over ϵ . So that would be a case in which a small noise term in the position -- so this is the position -- has a big effect on the derivative. And that's why the problem is ill posed. The problem is ill posed -- and it goes with our intuition that high frequency. This 1 over ϵ down below is producing a high frequency oscillation here. And of course everybody realizes the velocity is the correct, the x dt, plus the derivative of this, which brings out a 1 over ϵ , maybe it's a cosine. Maybe it's a cosine. So the 1 over ϵ cancels the ϵ ; that's a cosine of t over ϵ .

So this was small. A small change in position produced an order of 1 change in the velocity. So if we only know position within ϵ , we're in trouble. And this is exactly the point of ill-posed problems. Then our velocity could be that. I could make this example worse if I increase the frequency further -- put an ϵ squared

there. Then the amplitude would still be a small, but when I take the derivative, there'd be a 1 over epsilon, the amplitude would actually be very large. So I was just modest to keep epsilon and epsilon there, so that they canceled each other and produced an effect on the derivative.

So that's the problem. If you have noisy data about position, how can you get velocity? OK, so that's like example number one. It's very important, and I actually I have no magic recipe for it. But let me mention other problems that you'll know about like well seismology. A typical inverse problem in seismology would be find the density find earth density, say from travel times of waves, from wave travel time, which is what we can measure. So that's seismology and also, of course, everybody understands that this is what oil exploration depends on. You set off an explosion on the surface of the earth. The wave travels into the earth and some part of it bounces back, and in fact maybe several pieces bounce back at different times. And from those results you have to sort of back project to find the density. So back projection is a word that comes into several of these applications. Of course, another one would be the medical ones: CT scans, MRI, PET -- all these ways to take measurements. And from those measurements you have to find the density, so you're looking for density of tissue because you hope that would allow you to distinguish a tumor from normal tissue. So that's a giant area of application. Oh, another one would be find the density of the earth -- let's say another way to find the density of the earth, another bit of information we have from the gravitational field. You see, that's what we can measure: the effect of gravity, the effect of the earth's density. We measure the effect, and we want to know the cause. That's the problem. And this reminds me that there is a special lecture coming by Professor Wunsch, Carl Wunsch at MIT -- he's outstanding -- and that's Wednesday, May 10 at 4 o'clock. And in fact his abstract which you might see somewhere -- I can post it on the course website -- his abstract tells, he's solving a very, very large-scale, ill-posed optimization problem. Least squares problem. Perfect for this course.

So those are familiar. The books I've been looking at just list whole lots of examples. Let me see. Oh, scattering. Let me just keep going here. This is number five. Scattering. From scattering data find the shape of the obstacle that's responsible for

the scattering. So that's a giant example of many air force applications and many other applications. But we recognize if you want to identify some object by scattering, by radar data and other scattering data. Well there are just lots of others. Oh, and then we have a Laplace or a Poisson equation, which is the divergence of some inhomogeneous material property equals the sum f of xy . OK, so what we're usually doing in this course and most courses is the direct problem of find u , so the direct problem is find u . And what's the inverse problem? The inverse problem is we know u and f and we have to find c . So find the coefficient. Well, it may not be instantly clear whether it's possible. In fact, it probably is not possible, and that's what makes the problem ill posed. Yet if you have enough measurements of inputs and outputs, you could reconstruct the matrix. Of course, a person like me is going to think about the matrix question. Suppose I'm looking for the matrix A Can I call this number seven? And since it's in matrix notation, it doesn't take much space. Usually I know the matrix, and I know b and I want x . In the inverse problem I know b and x and I want to know the matrix. What was the matrix that produced it. Well obviously one pair bx is not going to be enough to produce the matrix, but enough will. But then if there's noise -- this is the point, of course, that there's always noise. So that's what I now have to deal with. The main thing is how to deal with noise in the data, noise in the measurements, because if the problem is so posed, we saw even in that simple cooked up example that a small amount of noise could produce a big difference in the solution.

OK. Right. So those are examples. Now I wanted because math courses and this one never mention integral equations, I thought I would write one on the board. And these examples fit in this -- if I describe them mathematically or another whole list of problems that I'm seeing in the books on ill-posed problems -- very often they are integral equations of the first kind. Again, the direct problem is -- I'd better write it down -- the direct problem -- this is known . So the direct problem is given the k , which is a bit like c over here, find the u , solve for u . Solve for the unknown u . It's a linear problem. It's an Ax equals b problem, only it's in function space. And of course one way to solve it will be, probably the way to solve it numerically will be somehow make it discrete, bring it down to a matrix problem. That's what we would

eventually do. But in function space, integral equations played a very very important historical role in the development of function spaces. And now then the inverse problem would be given u find the x I guess. Something like that. That would be possible. That would be one possible: inverse number one.

But I wanted to make some comments on integral equations, just so you would have seen them. The integral could go up to x or it could go up to what. And Volterra and Fredholm are the names associated with those two possibilities, but these are both - whether Volterra or Fredholm -- they're both ill posed, whereas if I want to make them well posed. Well we've seen how to make a problem well posed. I have this operator A , which has a terrible inverse or no inverse at all, and the way I improve it is add a little multiple of the identity. I'm supposing that I know about A , that it's not negative, that it's eigenvalues can be very very small or 0 but not negative. So when I add a little bit, it pushes the eigenvalues away from 0 up at least as far as α . And over here if I add in αu of x , that's what it does of course. I've added α times the identity operator here and that's given me an equation of the second kind. So those three minutes were just to say something about language and to look at an integral equation -- something we don't do enough of. Integral equations, well some problems on nice domains, you can turn differential equations into integral equations, and it pays off big time to do that. Professor White in EE, if he taught a course like this it would end up with half a dozen lectures on integral equations because he's an expert in converting the differential equation to an integral equation. He would convert Laplace's equation to an integral equation, and the Green's function would enter and he would solve it there.

OK, so how does velocity fit? Well, everybody can see that the integral of velocity -- so example, the velocity example is that the integral from 0 to x of the velocity, that's of course integral from 0 to x or, or 0 to t maybe would be a better -- So suppose velocity is the position dx , ds . So it's x of t minus x of 0. So this is position This is velocity. And in the inverse problem, position is known, say by GPS and velocity is unknown, to find by GPS. So GPS will give you a measurement of position. Maybe you know something about GPS. You know that there are satellites up there whose position is known very exactly, and they have a very very accurate atomic clock on

them, so that times are accurately known. So they send signals down to your little hundred dollar receiver, which of course has a ten dollar clock in it. So there'll be some errors, partly due to the clock, largely due to the cheap time keeper. But actually the way you get real accuracy out of GPS is to have two receivers, and then you can cancel the clock errors and get less than a meter accuracy. If you take account of all the sources of error, you can get it down to centimeters. So GPS is giving you -- just your single receiver is still good enough. It's measuring the travel time and since the signals are coming with the speed of light, that tells us the distance from each satellite. Right? Here is your receiver R. Here is satellite number one two three, four up in the sky, and you know these distance. But you don't know the time very well. So with four satellites I'm able to find the position of the receiver and somewhere in space and some moment of time that this clock is not good enough to tell us. So we have to solve for that.

So four receivers sending signals to -- I'm sorry, four satellites sending signals to the receiver, we can solve that problem and find the position and time. That's the fundamental idea of GPS. Now of course you get better results if there's a fifth receiver, a fifth satellite, and a sixth. The more the better. And of course then it's going to be least squared. Because you're still looking for four unknowns, but now you have six distances pseudo ranges, so we would use least squared, so by least squared. OK.

But what about velocity? Suppose your receiver is moving, as of course it is if you rent a car that has GPS installed to tell you where to turn. And of course it has to have a map system installed so that it can look up for the map position. So for many purposes you need velocity, and I'm not an expert on that subject at all. I just comment that one way to get velocity is to take differences, so that velocity is approximately x of t plus Δt minus x or x of t . Two, let's say: x of t sub 2 minus x of t sub 1 over t sub 2 minus t sub 1. But if we want to get velocity near a certain time then these t 's better be near that time because the velocity's changing so we better be measuring it at the time we're wanting it. And then if they're very close, then we're dividing by a small number and the difference -- the noise, the error in

measurements, x , is multiplied by that 1 over Δt .

And one way to avoid it is to go into the frequency domain. So this is like an interesting option in a lot of these problems. Can you operate better in the frequency domain? Of course the ill-posedness is not going to go away. It comes as we saw from high frequency. But if we go into the frequency domain, and if these GPS satellites are sending at a certain frequency and as we move, of course, the Doppler effect of course is the fact that as the receiver moves, the frequency is observed to change a little. Right? Just says like the noise of a train going by is the familiar example. Nobody ever sees a train going by anymore. But it's the same idea. But you hear traffic go by. Actually, I guess that that's how we cross the street, come to think of it, by listening to the noise of cars, and our global internal Doppler says the cars are going away from us, in which case we don't worry, or it says the car's coming fast, in which case we're careful, or it says the car's coming slowly, and we get across first. So we use Doppler. I don't know exactly how, how our human audio system builds in Doppler. Anyway, Doppler would be a change to the frequency domain and a restatement and perhaps an improvement in the in the problem.

OK, so those are examples without math. Now here's a small bit of math as the lecture ends. So the only math was this simple example. So I guess I got one more board over here. I'm going to put this geometry problem that you've been thinking about out of sight.

OK, so what's the key? It's this Tikhonov. Tikhonov Regularization. Tikhonov Regularization. And it's adding α . Add α to the least squares problem. And I thought it was amusing to notice, Tikhonov was born in 1906, so this is a hundred years exactly since he proposed this method. It's one of the about five methods that the books describe. And it's the one I'll speak about here at the end. I'll just mention that one which might come up in a project possibly. Other methods are used in iterative methods, like conjugate gradients, and stop when you're ahead. See if you push conjugate gradients on and on and on, then eventually it's going to produce your exact ill-posed matrix with the big invert and unrealistic solution. So you stop.

The same way for an integral equation. We discretize that so we get a matrix process, which we can solve. But if we refine the mesh so that the matrix gets bigger, it gets more ill posed. So the closer we get, the closer the discrete problem gets to the true integral equation, the more [? sick ?] it is. So there's some point at which you are OK, and then if you go too far, you're worse off. So that happens with conjugate gradients.

Now what's the Tikhonov idea? So the Tikhonov idea is look at them as, you'll know, it's a minimum. I'll just call it A again. Ax minus b squared plus αx squared. That would be the simplest penalty term, regularization term. So this leads to the normal equation: $A^T A + \alpha I, \hat{u}$, which depends on α , equals b . OK. so \hat{u}_α is the solution to that.

OK, now let's let noise come in. So noise in b . Noise yields a b_δ , say a δ amount of noise. b_δ is the measured observation. See we don't know the exact b . This is what we measured. And we can suppose that the size of the noise is -- let's say -- δ . So δ measures the noise. OK. I mean what's my question here? Always the question is what do you take for α ? What should that parameter be? if you take α very small or 0, then your problem is ill posed, and the answer you get is destroyed by the noise. If you take α very large, then you're overriding the real problem -- you're over regularizing it. You're over smoothing it. So we don't want to take α too large, but we can't take α too small either. And the theory will say that if we use an α -- So the theory will say this, essentially this. It'll say that the \hat{u}_α , the difference between \hat{u}_α and \hat{u}_α with the noise coming from the -- Do you see what I mean by it? This comes from the true b , which we don't know but it's the answer we would like to know. This comes from the measured b with noise in it, and it turns out that this is of size δ over square root of α . That's a rather neat result. It gives us a guide because we want that to be small.

So we're assuming that we have some idea about δ , and it tells us again that if I take α very small, then I'm not learning anything. So you see actually that we want, we need, δ over square root of α to go to 0. Maybe we're reducing the

noise. Maybe we have a sequence of measurements that get better and better. So δ goes to 0. We would like to get to the right answer then, but as δ goes to 0, we better not let α go to 0 faster. The message here is -- so I haven't derived this estimate but just written a conclusion, which is that as the noise goes to 0, that means our measurements are getting better and better, we do want to let α go to 0 but we can't overdo it. And a good choice, a good choice is let α be δ to the $2/3$. That turns out from these little estimates to be the best choice, because then square root of α is δ to the $1/3$. This then is δ divided by δ to the $1/3$ third. This is δ to the $2/3$, and we can't do better. So that's the balancing. That's the balance when you take that value of α . Then as you reduce the noise, the error gets reduced in the same proportion. It's not the full reduction in δ , but fraction the $2/3$ power.

OK, so that board summarizes, you could say, the theory of regularization. And there's a lot more to say about that theory, but I think this is perhaps the right point to stop. OK so that's applications. One other application, if I can add one last one because it's quite important and it comes up in life sciences. So in life sciences you might have in genomics you might have a lot of genes acting. So the action of n genes, n being large, produces some expression, the expression of the gene, and it depends on how much of those genes are present. But what everybody wants to know is which genes are important, which genes control the blue or brown eyes, or male or female. Not clear, right? We have an idea from biological experiments where the important genes lie on the whole genome, where to find them in the chromosome, but this is what we measure. We observe male or female, and we can change the genes, but there are more unknowns than, more dimensions than sample points. We're really up against it here. And we're really up against it because -- so what's going to measure the importance of a gene. How important is gene number two? Well, the importance of gene number two is identified by the size of the derivative. That quantity, if it's big, tells me that the expression depends strongly on gene number two. If it's small, it says gene number two can be ignored. That's exactly what the Whitehead and Broad Institutes want to know. Well, for cancer of course as well. And my only point here in this lecture is to say that again

we're trying to estimate a derivative. We're estimating a derivative from few samples in high dimension, and it's certainly ill posed. And it certainly has to be regularized, and it certainly has to be studied and solved.

So that's maybe a seventh example, and with that I'll stop the final lecture and just bring down one last time this little bit of geometry to give you something interesting to do for the weekend.

OK, so I'll see you Monday then with a talk about different methods for solving linear equations and others coming. And time is, you know, this semester will run out on us. So please send me emails to volunteer. And don't think you have to be perfect. If you've got transparencies, got the topic in, mind got the question in mind, got some numerical results, you're ready.

OK. Good. Have a good weekend. Bye.