[SQUEAKING]

[RUSTLING]

[CLICKING]

STEVEN JOHNSON: So I want to briefly start to talk about going beyond 18.02 derivatives. So we've already gone, I would say, beyond 18.02 in the sense that they never do the-- they have Jacobians maybe, but they never really do the chain rule. They never really write even this definition in-- where is it?

Go back-- saying that the change in the function is the Jacobian times dX, which I view as almost the definition of the Jacobian. I think that they never really do that. They never really do, like, multi-dimensional Taylor expansion or something like that.

But now what we want to do in this class is go really beyond 18.02 to have functions where the inputs and outputs are not just column vectors necessarily or scalars. But they could live in other vector spaces, in sort of more general vector spaces.

So let me give an example. And you probably learned in 18.06 hopefully, remember, that a vector space is not just a column vector. It's really more abstractly anything you can add and subtract or multiply by scalars. And so there are other things that you can put in vector spaces-- you can think of as vector spaces.

And so for example-- would be matrix inputs and/or outputs. OK? For example, suppose we have the function f that takes a matrix as input. And suppose it's an n by n matrix as input. So now there's n-squared inputs. It's an n-squared dimensional vector space, if I allow arbitrary matrices.

And I could say, for example, f of A equals a inverse, right, or A-cubed, or even more complicated things. For example, I could say f of A equals-- you take a matrix as input. And you get the u, the upper triangular matrix from Gaussian elimination. Right?

If you do Gaussian elimination, let's assume that then we have to restrict ourselves to maybe nonsingular matrices or something. You take a matrix in. And it gives you an upper triangular matrix out. OK?

Or you could also have things-- these are things with matrix outputs and matrix inputs. You could also have scalar outputs, but you have matrix inputs. Right? And what kind of functions would those be?

So for example, the determinant is a perfectly good function that takes a matrix in and gives you a scalar out-- or the trace. OK? Or you could have even more complicated things. We could have-- or the sigma 1 of A, the largest singular value. Right?

You take a matrix in. You compute the singular values. Take the largest one. That's a number. That's a function that takes a matrix in and gives you a scalar out, right-- or many other examples. And so we're going to see how to take derivatives of things like this. We'd like to be able to take derivatives of things like this.

And so let's do a couple of them. They're not that hard to actually analyze-- the simple ones. Determinant-- trace is actually pretty easy, but I won't do that. But determinant is a little bit more complicated. But Alan's going to-- we're going to show you some clever tricks for that pretty soon.

And matrix factorizations, like SVD or Gaussian elimination, are also a little bit tricky. But once you see the trick, it's not so bad. But things like A inverse and A-cubed are actually pretty straightforward.

So for example, let's do, f of A is A-cubed. Right? So if we do-- I think actually Alan already did f of A equals A-squared. But now we'll do a-cubed. Right?

So now we can just do the product rule. Our df, right-- A-cubed is just A times A times A. And again, I'm going to assume these are square matrices. Otherwise, you couldn't multiply them by themselves. [INAUDIBLE] right, are blue. Should I put-- my matrix input is red. Right?

And what's the product rule? Well, it's just three terms. I have a dA times A-squared plus A times dA times A plus A-squared times dA. Does everyone see that? So I can just take the d of this term, the d of that term, the d of this term. And A is no longer a constant matrix. A is not a fixed parameter. It's the input parameters.

So I can-- remember? So all this is just saying, suppose I take my function f, and I take A, and I add an arbitrary change-- this is an arbitrary n-by-n matrix with little entries, infinitesimally small entries. And I ask how much does the function change, does A-cubed change-- and again, dropping terms that's proportional to dA-squared or higher, right?

So it's nothing too mysterious. It's just asking how much does it change, keeping only linear terms. That's what the derivative means. OK? And if we keep only linear terms, we get these product rule. We get these three terms.

So our f prime of A is a linear operator. And I think this is one of the cases where it's really useful to have the square brackets here, right? Because it's not anything multiplying dA on the left. It's the linear operator that takes any little change dA-- any arbitrary change there-- and returns this matrix. It returns A-squared, right?

ALAN EDELMAN: So no square on--

STEVEN JOHNSON: OK?

ALAN EDELMAN: --the middle term.

STEVEN JOHNSON: Oh, yeah. The middle term is not squared. Thanks. Good that Alan is back. So I notice that this is very much not equal to what you would get for-- if this were a scalar, the derivative of A-cubed would be 3A-squared dA.

ALAN EDELMAN: [INAUDIBLE] it would be equal, of course.

STEVEN JOHNSON: Yeah. When would it be equal, right? Unless what? dA and-- anyone guess? For particular perturbations, dA, what kind of special--

ALAN EDELMAN: Well, you may not have picked it up on the microphone, but there was an answer that dA and A commute.

STEVEN JOHNSON: Yes. Yes. So unless dA and A commute, which is not true, right-- so if you're interested in perturbations that are proportional to A, right, or something like that, right-- or some power of A, something that commutes with A-- then you can use the 3A-squared dA. They could-- basically, commuting matrices almost act like scalars, right? But--

ALAN EDELMAN: Well, I think people find it kind of surprising that this is kind of the derivative of A-cubed in, really, simplest form. I mean, you can't-- there's not-- there's nothing you can do to simplify that any further.

STEVEN JOHNSON: Yeah. And then, particularly, it's not equal to any matrix times dA. Right? I can't just multiply dA-- multiplying a matrix dA-- on the left of my matrix-- it is a linear operation, but I cannot write this linear operation in this form. Right? And so writing-- think about it. The Jacobian of this is going to be-- we can't really-- this is a linear operation, but I can't really write down a Jacobian matrix that easily.

And the only caveat is, pretty soon, Alan will show you actually how you can do this unless we convert dA to a column vector. And pretty soon, we will-- you can do that. You can take the dA, and take all of its n-squared variables, and just stack them up, right? And then you can write this back as a matrix times this. But--

ALAN EDELMAN: And by the way, I hate doing that as much as I hate indices.

STEVEN JOHNSON: Yes.

ALAN EDELMAN: Because-- and I'm saying that out loud because students are very prone to do this mechanically. And I kind of want to put pressure in 180 degrees the opposite way-- not to do it. But of course, I will do it right away anyway. But I hate it. [CHUCKLES] I'll explain further in a little bit.

STEVEN JOHNSON: Yeah. And so let me do the other example. And then I'm going to stop, right? And of course, these are not the only vector spaces. But maybe later, I'll talk about if you have functions that take a function as input and give you, say, a number as output. Then you can also talk about derivatives. Because functions form a vector space. And that leads you to something called calculus of variations, which is another really cool application of the same kind of idea.

So the other one that's really, really practically important is taking the derivative of matrix inverses. Why is it important? Because matrix inverses come up in solving lots of real-world problems, like solving the stress on an airplane wing. And if you want to compute the derivative of the output with respect to the parameters in the matrix, right, you need to differentiate through a matrix inverse somehow.

ALAN EDELMAN: But let me quickly ask, if A were a scalar, just what's the derivative of 1 over A? Shout it out. You all know it. What's the derivative of the function that takes [INAUDIBLE] of 1 over A?

AUDIENCE: Negative 1 over A-squared.

ALAN EDELMAN: Right. So we're looking for something that's kind of like that. But it's more matrixy, right? So just to--

STEVEN JOHNSON: Yeah.

ALAN EDELMAN: [INAUDIBLE]

STEVEN JOHNSON: So here's a case where if I did it the 18.02 style, I would go mad, right? I could write down the inverse-- like, some explicit, horrible formula in terms of Cramer's rule and determinants of minors, and cofactors, and things, right; and then try to take the derivative of each component of that with respect to each entry of A. And it's just--

ALAN EDELMAN: You-- even in the 2-by-2 case, I would hate to see that.

STEVEN JOHNSON: Even in the 2-by-2 case, it would be-- oh, you would go crazy. But so there's a trick here that makes it really easy, right? So we know that A inverse times A is equal to what? Yeah, hopefully somebody said "I", right? So again, these are going to be n-by-n. Obviously, this is only going to work for n-by-n nonsingular matrices. And this is the n-by-n identity matrix, right?

So then what I can do is I can say, therefore, let's ask what happens-- what's d of this? And d of this, again, means I change A by a little bit. And I ask how much does this change by. Oops. And the answer is, of course, it doesn't matter how I change A.

If I change A by a little bit, the product is I. So the derivative is 0. But by the product rule, this must be equal to d of A inverse times A plus A inverse times d of dA. All right? That's just the product tool.

That must equal-- is this clear to everyone? So if I take this thing and I ask-- if you change A by a little bit, how does this change? I can do it two ways. I can apply the product rule, which means it's d of this, the change of A inverse times A plus A inverse times the change in A. Or I can say, well, A inverse is I. So I can ask, how much does I change if I change A? And the answer is, not at all.

So this-- 0 has to equal that. So if I just put those two together, right, and I solve for-- that means that dA inverse, right, times A has to equal minus A inverse dA.

But that means that df, which is, of course, dA inverse, OK-- how can I get dA inverse? I need to move this to the right-hand side. So I just multiply both sides of this by A inverse on both sides. Right?

And so if I multiply this by A inverse on the right, OK, then I just get dA inverse. And on the other side, I get a minus A inverse dA A inverse. And that's it. This is-- and so this is my f prime of A acting on dA.

ALAN EDELMAN: So you see when it's a scalar, it's minus 1 over A-squared times dA-- the result you remember. But in a way, I like to think of it as, when you have one-- like the scalar case, the minus 1 over A-squared.

Should that 1 over A-squared be on the left, or should it be on the right? Well, there's no reason to prefer one over the other. And the math, of course, makes it very fair by putting one A inverse on the left and one A inverse on the right. And that's kind of how I like to look at that answer.

STEVEN JOHNSON: Yeah. And we're going to--

[INTERPOSING VOICES]

STEVEN JOHNSON: And we'll see that this [INAUDIBLE] is incredibly useful. Yeah. Because differentiating through matrix inverses is something that actually happens all the time, especially if you're optimizing an engineering problem. Because engineering problems involve solving systems of equations. And then if you want to optimize the result, you need to differentiate through this. I'm going to-- should I stop here, Alan? Did you want to talk for a few minutes?

ALAN EDELMAN: [INAUDIBLE] have 10 minutes. But I think it'd be fun to show some of the examples in the Pluto notebook at this point.

STEVEN JOHNSON: Sure, sure.

ALAN EDELMAN: So let me go ahead. And I'll do that. I'll grab the mic and put it on my shirt. And I'll grab the screen as well.

OK. So I've got these-- this notebook here, which I'd like to just show you. And again, I only have about 9 or 10 minutes. So it'll just be a little bit of a taste.

But in effect, what I'm really doing is showing you, again, what Steven just showed you-- maybe slightly different examples or maybe the same, I don't remember-- but kind of also a little bit of a different view. Because I think when it comes to this sort of thing, it's really helpful to see the same thing from multiple viewpoints.

And so-- let's see. So first of all, just a little bit of Julia over here-- these 2-by-2 matrix Jacobians. So I'm using the Symbolics package in Julia at the moment, besides some linear algebra. I'm not even sure if I need the linear algebra.

So just to see some variables, here are some variables-- p, q, r, s, and theta. And let's see. What did I do here? Yeah. So here I'm going to take a rotation matrix and multiply-- let's see. What did we do here? Where was q? Well, it doesn't really matter.

The point is that I'm going to be working with 2-by-2-- let's go to this one-- 2-by-2 matrices. I like to go in column major order. This is what Julia does. And so p, q, r, s goes down the columns, right? So this is a 2-by-2 matrix over here. OK?

And as Steven started to say, it's not unusual to flatten a matrix with this vec operator, right? And so here's this matrix-- p, q, r, and s-- this matrix X. And the vec of the matrix converts the matrix into a vector. This is the thing that I hate to do, but it's not bad for kind of beginning students to show you the concept of flattening.

And so if we consider the matrix square function, here it is in all of its glory symbolically, right? So if I were to square that matrix X that had p, q, r, and x-- s-- p, q, r, s, right? This is what you would get if you multiplied X times X, right?

And so when we talk about taking the derivative, you could think of this as a function, right? You could actually look at the vec of X-squared. And if you want to-- in a way, this is forced. But in a way, this is also a good idea.

You could see that I have a vector from four input variables-- p, q, r, and s-- to these four outputs, right, these expressions-- p-squared plus qr and so forth. So if you were to work with the vec function-- if you worked the four inputs to the four outputs, what would the size of the Jacobian be? What would be the shape? If you have the four inputs and four outputs, the shape would be--

AUDIENCE: 4-by-4.

ALAN EDELMAN: 4-by-4. And let's think about the general matrix squared just for a moment. If I have an n-by-n matrix that I'm squaring, right, if I vec that matrix, it'll be a column that's n-squared long. And the output will be a column that's n-squared long. And so what would the Jacobian be for general n?

AUDIENCE: n-Squared matrix

ALAN EDELMAN: It would be an n-squared by n-squared matrix if you write it as a Jacobian, right? And of course, you could do that. And I want to urge you to-- I don't want to say never do that, but not to do that too instinctively. But nonetheless, we can do the symbolic Jacobian. And in fact, Julia will do it for you. And here, you have it.

And so let's see. So Steven made a point that if you move this way, you're changing the inputs. And if you move this way, you're changing the outputs. So let's focus on this one output over here, p-squared plus qr, right? If I take the derivative with respect to p, I'll get 2p, right?

The next derivative I'll take is with respect to q. And I'll get the r, which goes this way, right? Then if I take the derivative with respect to r, I'll get the q, which is over here. And finally, the derivative with respect to s-- well, this doesn't depend on s at all. So the derivative is 0, right?

And so if I focus on this one output and take all the four input derivatives, I go this way. If I did the same down here, I would get this row, right? The derivative of this with respect to p is 0, which is that 0 over there. OK?

And so somehow or another, this is capturing the derivative of A-squared explicitly as a matrix. You see, I think by now you're probably getting my point of view that if you take 18.06 or any elementary linear algebra class-- in the old days-- and maybe some of you took a course like this-- you never would see a matrix.

But if you take Gil Strang's course and a lot of other linear algebra classes, you think that you only see a matrix by element-by-element. You don't see the linear operators anymore. You just-- it's almost like missing the forest for the trees because-- or is it the trees?

Yeah, you're missing the forest for the trees. Because in most modern linear algebra classes, you're focusing element by element. And so that's kind of how you're-- you almost psychologically learn to think that way. And you don't think about linear operators anymore. I don't know how you all learned linear algebra. But in the old days, it was no matrices, pure linear operators. OK?

So here's an explicit matrix. And of course, we can do this numerically as well. So here, I'm replacing p, q, r and s with the numbers 1, 3, 2, 4. OK? And let's see. And of course, I could do this numerically by making a perturbation of-- so I have my matrix M, which is 1, 2, 3, 4, right? And I have this-- E is 0.001-- just a few small numbers. And of course, you can see M plus E-squared minus M-squared is, in fact, the change to the square, right?

So here-- let's see. So maybe I'll add this right now. So in the language that Steven had just done, we would probably write this as dX-squared is-- oops, oops, oops. Where did I go? dX squared is equal to X-- he just did the cube, but the same product rule gives you the answer for the square, right? So I'll just write this right here. I won't execute that. Right?

So this is how we're writing-- see, I want to make sure that you clearly understand that, one way or another, the information in this 4-by-4 matrix and, in some sense, the information in this linear operator somehow has to be the same thing mathematically, right? It's a matter of presentation, but it's not so much a matter of what's going on. I mean, some people would say that basis was chosen. But forgetting about that, from our point of view, this thing must express the same thing. OK?

And so I'm going to write this operator. I don't like to write the dX's all the time. So I'm going to write this linear operator, this linear transformation of dX. I'm going to call it E. All right? And I'm not going to use X. I'm going to use M. But you'll see it's just ME plus EM. It's really the same thing here with different letters.

And if I go linear transformation of E, I get the exact same answer to first order. I mean, you could see a little bit of the second order term showing up. But here, we're taking a difference. Here, we're computing a linear transformation. And you see that, to first order, we're getting the same numbers.

I don't know if you've ever stopped to think how magic calculus is, where you can numerically take a little perturbation-- or you have a formula for the answer that's going to turn out to be the same. I mean, we're all so used to it, you may have never stopped to see the magic in all of that.

But it's really very magical that you could numerically subtract two nearby things, and the mathematical gods above give you a formula for that difference. I think it's-- I bet you've never stopped to appreciate that, but it's really a wonderful thing. So here, the ability to compute ME plus EM gives you this exact infinitesimal answer that corresponds to this numerical difference. And I think that's just a wonderful, wonderful thing.

Now, there is a notation called the Kronecker product which kind of lets you express these sorts of things quite well, quite nicely, kind of going back and forth between the explicit matrix and the linear operator point of view. And I don't know-- how many of you have said you've seen the Kronecker product somewhere, somehow? So about three or four. It's a small-- oh, five-- a small minority of the students in the room.

So I like to tell the story that, in one of the courses years ago, I showed the Kronecker product to-- and I just brought it up just like I'm doing for you today. And about a month or two later, a student came back to me and said, you know, I'm so glad you showed me the Kronecker product. I actually used the Kronecker product in a paper. And I won a big prize. So thank you for showing me the Kronecker product. So I can't guarantee you'll get a prize for it, but that did happen to me.

All right. So let's see. So here, I've got-- I'll add the variables a, b, c, and d. And I'll just take two-- I'll just show you the Kronecker product quickly. And I'll kind of use it on Monday.

But here are two 2-by-2 matrices. And you could imagine trying to take all possible products of entries in the first matrix with entries in the second matrix. OK? And so let's-- oh, yeah, let's do that here.

So let me take the Kronecker product of these two matrices just to show you what it looks like. And I think you'll notice that on the-- we have a's, b's, c's, and d's in all possible combinations on the left and p's, q's, r's, and s's in all kinds of possible combinations on the right. This gets shown a little bit better with letters and these cute characters that we have available in Julia.

So here, I'm going to do it with the Kronecker product of-- let's see. Here's a, b, c, d, e, f, I guess. So that's a 2-by-3 matrix, with the pizza, the alien-- what do I have-- the panda, and the smiley cat. And again, I think with your eyeballs, you could kind of see what's happening. Here's the 2-by-2 matrix with the identity. Here's what happens when the identity is on the left. OK?

And so it's just all possible products with the convention-- and I think, with the identity, you might even see it most clearly-- the convention that you take the top-left entry, say, of the first matrix with all of the entries in the second matrix, right? And then you go-- the bottom left is down here, the top right of the left entry, and the-- you see how that works, right? And so that's the order.

And OK. I'm going to end now. The class is over. But I will mention that to get this-- this Jacobian matrix that you saw before from the square can be expressed using Kronecker products. This is how it's done.

So I'm going to pick up-- I'll probably practically start from the beginning on Monday, I imagine, since class is over. But I just wanted you to quickly get the sense of the Kronecker product as a way of expressing the-- it's a great way of expressing some of these matrix functions and the derivatives of some of these matrix functions. All right?

So I'll stop right here. And--

STEVEN JOHNSON: And we'll post pset 1 later today, by early [INAUDIBLE]

ALAN EDELMAN: We're hoping for 5:00 PM. But if it doesn't happen, please don't hate us. But I think we're going to target 5:00 PM. And we'll see you all on Monday. So have a good weekend, everybody.