[SQUEAKING]

[RUSTLING]

[CLICKING]

STEVEN G. JOHNSON: So I want to revisit the things that Alan talked about, but just a little bit more slowly and a bit more-- just try and lay out the rules for you as clearly as I can. And what we're going to try and do is, again, just revisit the notion of a derivative to try and write it in a way that we can generalize to other kinds of objects. And so I'm going to start with 18.01 and then go to 18.02 and so forth.

So as Alan said, the key notion of a derivative, just, I think, it's easy to get so good at taking derivatives, like knowing the rule for the derivative of sine or cosine or x squared. You're so good at doing them that you forget what they are, right? And so the very first thing you learned about a derivative is that it's the slope of the tangent.

But what that really is is linearization. So you have some arbitrary maybe nonlinear function f of x. And you're at a point x. And near that point, you're going to approximate the function with a straight line. That's the tangent. So it's really the linear approximation of f.

And then if you move a little bit away from x-- so let me call that delta x, so not d. Delta is going to be a finite change. d is going to be infinitesimal pretty soon. But if you move it just a finite amount, a little finite amount delta x away, of course, the function value changes.

But in a linear approximation, the new function value is the red dot here. So that linear approximation is, if you're taking the function f of x at x plus delta x, the new value is f of x. And then this linear thing is just the slope, which we call f prime of x times delta x. That's just the definition of the slope. It's the little change in y for a little change in x.

And, of course, these two terms are not exact. This red dot doesn't exactly match where you are in the real function. So there are also corrections. But the corrections are higher order. They're terms look like delta x squared, delta x cubed maybe. Maybe if the function is not higher-- it doesn't have higher derivatives, it might have square root of delta x, so delta x to the 1.1.

But these are all terms that are going to be higher powers of delta x terms that, if delta x is sufficiently small, these terms will become more and more negligible compared to this linear term. And a nice notation for this that's used a lot in computer science, less so outside of that, but it is used in calculus as well, maybe not in 18.01, is this asymptotic notation.

In computer science, probably many of you will have seen big O notation. Here, I'm going to use a variant called little o notation. So this is not a capital O. This is a lowercase o. So these terms we call little o of delta x.

It's a little o of delta x. It denotes any function that goes to 0 faster than linear, faster than delta x. So delta x squared goes to 0 faster than delta x as delta x goes to 0. I should say as delta x goes to 0.

And so you look at this thing and this probably looks a lot like a Taylor series. You do a Taylor expansion of f of x plus delta x around f of x. This is the first term. This is the second term. The third term, remember, is 1/2 f double prime delta x squared or something like that.

But that's the wrong way to look at this. A Taylor series is a much more advanced concept. It's something you can do much later in calculus and for good reason. Because not every function even has a Taylor series that converges. This is more basic. This is really the definition of a derivative.

The derivative is the linearization. If you make a small change in delta x, the change in f is a linear term plus smaller stuff. And that smaller stuff only gives you a Taylor series if it's basically a polynomial.

It might be smaller stuff like delta x to the 1.1 in which case it doesn't have a Taylor series. But this is always true. This is just this is what it means to be the slope of a tangent.

And so the nice thing about this is Alan says this notion, if we keep the delta x on the right, this is going to be much easier to generalize to other kinds of x's that are vectors or matrices or even other functions, other kinds of things. So this is the linearization. Whoops.

So we're going to have a delta f, which I'm going to have to define as-- whoops, let me put it in black-- f of x plus delta x. And again, the delta, the Greek letter delta, is not infinitesimal. It's just a small number. It's just a number. This is that thing.

But I'm going to drop terms, higher order terms. So there'll be an error there. Well, yeah, so actually let me put this another way. So this is going to be approximately f prime of x delta x plus higher order terms. This is the higher order.

So this is the small change in the input of the function. And this is the resulting small change in the output. And this is going to be the definition of the derivative. The derivative is whatever you do to delta x to first order to give you the linear change in the output for a small change in the input.

It's a little annoying, though, to keep these o's around. So we keep always having to-- whenever we have a finite change in the input and a finite change in the output, this is never exact. This is an approximate relationship. And we have to keep saying plus higher order terms, plus higher order terms, plus little o delta x. And it's annoying.

So it's easier to use to switch to differential notation. So I'm just going to change my delta to d. So df is going to be f of x plus dx minus f of x. And this is going to be f prime of x dx where this is-- I'm just going to right equal.

So we can think of this differential in dx as being arbitrarily small. So it's really a limit, some kind of limit, of course. And so you can also think of it as really it's just this. But I'm just implicitly going to drop any higher order terms.

ALAN EDELMAN: That's how I like to think of it.

STEVEN G. JOHNSON: Right? So we don't have to get too fancy with defining differentials. I mean, this is a definition, right? This is just shorthand for this where I don't have to write plus a little o over delta x all the time. Let's see.

And so it's important to keep in-- so this f prime to delta x, this is the derivative. Here df is the differential. So if I ask you for the derivative, I'm asking for f prime. I'm not asking for df. Of course, they're related.

And let's see if we can move this out of the way. Good. So what was I saying? Yes. So now, what I want to do is basically use this as the definition of a derivative, a more general definition of a derivative.

So the linear algebra notion is that we have what's called a linear operator. So basically, the change in the output df is going to be a linear operator. Let me write that as times, which I'm going to call f prime of x times the change in the input.

This is our-- whoops. Actually, but my input is in red, right? That's my color code. My color scheme is input is red and output is blue. So this is our dx.

So I'm going to interpret this more generally. If x is going to be some kind of vector or matrix or whatever, this is just going to be a linear operation on this. And of course, for numbers, a linear operation, this is just a number. If dx is a number, the only linear operation you can do is multiply by a number.

So let me just remind you if you haven't taken linear algebra for a while, a review. Let's talk about what a linear operator is. So suppose we have some given vectors, v, in some vector space. That would be capital V. And remember, a vector space is anything where you can basically add, subtract, and multiply by scalars.

We have a plus or minus and times scalar operations that stay in our vector space. That's what the informal definition of a vector space is. You can write out axioms and so forth, but that's basically what it means.

A linear operator, this is what we're going to mean by linearization in the derivative. It has to be linear. What does it mean to be linear in general? So we're going to call this-- I'm going to denote this by, let's say, L of-- let me denote it by square brackets or just by Lv.

When it's clear enough, I'll just write it as if it were a multiplication. Often, that'll be clear enough. This is really acting on-- when I write Lv, it's not necessarily an ordinary multiplication. This is just going to be acting on v.

So a linear operator is a rule that basically takes a vector and gives you a vector out maybe in a different vector space. So this is L takes a vector in, v in, and it gives you a vector-- that's a terrible L-- Lv out maybe in a different vector space.

And linearity means what you think it means. It means if you take, for example, L of v1 plus v2, if you take the sum of the inputs, that's the same thing as L of v1 plus L of v2. So if you add inputs, that's the same thing as adding the outputs or if you multiply by a scalar. And so as usual in linear algebra, Greek letters are going to denote scalars as you would. So that's equal to-- you can pull out the scalar.

And so the nice thing about linear operators is we can define them on lots of kinds of vector spaces. So let's just do a couple of examples just to make sure we're on the same page here. And so, for example, you could just have L is multiplication by a scalar.

So you can have just L of v is just alpha v. That's a perfectly good linear operation. And if your vector space, if your vs are scalars, this is the only option. If vs are, say, real numbers, that's a perfectly good vector space. Another one that you're very familiar with is if L is a multiplication by a matrix.

ALAN EDELMAN: Steven, maybe I'll just point out sometimes people like to ask me. Wait, I thought that the linear operators on scalars are, I think in high school notation, y equals mx plus b, right? It's scalar times--

STEVEN G. JOHNSON: Yeah, yeah.

ALAN EDELMAN: --plus an offset that may not be 0. So what's going on here? Is that linear or not linear?

STEVEN G. JOHNSON: Yeah. So what about-- yeah, so let's do that. Let me call it a different thing. What letter should I use? O, let's use O. Ov equals 2v plus 1 for v is real numbers.

ALAN EDELMAN: Is that linear or not linear? That is the question. The graph is a line. So we all think of it as linear, but go ahead.

STEVEN G. JOHNSON: Yeah. So does it satisfy the rules? That's the question. So if I multiply the input by 2, does it multiply the output by 2? No.

So if we do O of 2-- no, let's do 3. 3v, that's, what, 6v plus 1. And that's very much not equal to 3Ov, which that would be 6v plus 3.

So this one, it does have a name. It's related. These are sometimes called affine.

ALAN EDELMAN: Affine, but not linear even if the graph is demonstrably aligned?

STEVEN G. JOHNSON: Yeah.

ALAN EDELMAN: They're not linear in the sense of linear algebra.

STEVEN G. JOHNSON: Right? So another one is clearly multiplication by a matrix. That's why we do matrices in linear algebra. Because they're a nice way of writing down a linear operation if your vs are column vectors. They're not the only way of writing down linear operation.

So for example, if you take a column vector and multiply it by 3, you could write that down as a matrix with all 3s along the diagonal. But it's a lot easier to write that down as 3, I'd say, as a scalar, than to write it down as a matrix.

So another example, just to be more-- so another vector space. If you took 1806, you learned that we can have a-- whoops, I have to get my color scheme. Yeah, so my vectors are red.

Suppose the vector space V is the set of functions f of x that take real numbers in and give you real numbers out. Those are a perfectly good vector space. I can take two functions. I can add them or subtract them, get another function. I can take a function and multiply by 2, get another function.

ALAN EDELMAN: Wait, how do we get sine plus cosine?

STEVEN G. JOHNSON: I get sine x plus cosine x. It's just got some other function. So if you take sine x plus cosine x, that's the function f of x equals sine x plus cosine x. It's another rule that gives you-- takes real numbers to real numbers.

And so what would be your linear operators on this? Well, multiplication by a scalar, that, of course, works. So let's think of L on a function f of x is just 2 f of x. That takes a function in, function out. That's linear.

What about a linear operator on a function of f of x? Again, that gives you the derivative, just the ordinary 18.01 derivative. This is the 18.01 derivative. Obviously, that only works if the function is differentiable. So maybe we can look at the subspace of differentiable functions.

That's also a vector space. Because if I take two differentiable functions and add or subtract or multiply by constants, they're still differentiable. I could also do integration. So if f of x that takes a function f of x in and gives you the integral from, I don't know, 0 to x of f prime-- no, so f of x prime and dx prime if they're integrable. Again, we need to restrict what functions are allowed if we're taking derivatives or integrals, so things where these exist.

But this is perfectly linear. Why? Because if I take the function and I double it, if I double the integrand, it doubles the integrals. If I add two integrands, you add the integrals. Integration is a linear operation. Derivative is a linear operation.

Another fun one is suppose we take L of f of x. And the output is the function f of x squared. So this doesn't look linear. I have a square there.

But why is this linear? Why? Because, let's see, if I take L of two functions, if I have f of x plus g of x, that should be f of x squared plus g of x squared. I'm squaring the input, not the output. So that's equal to L of f plus L of g.

ALAN EDELMAN: So I'll just comment. Leave it to mathematicians to take what most people would think of as just a column of numbers and abstract it out and say that this finite dimensional column of numbers is somehow the same as continuous functions or differentiable functions, satisfies the same axioms. So we'll call it a vector space as well.

STEVEN G. JOHNSON: Yeah. But it's incredibly useful, though. Because very often, especially in physical sciences, you have something where conceptually you're solving for a functions. So you're solving for the fluid flow or something around an airplane wing.

And what you want is that then take that fluid flow and compute the drag on the airplane wing. And then in order to optimize it, you want the derivative of the drag with respect to that flow field, with respect to the function, or with respect to the shape of the airplane, which is a function. So it's very, very nice to be able to take derivatives connected to functions and work with vector functions as vector spaces.

And very soon we're going to be able to do that with this notion of a derivative. Because we're going to be able to define linear operators, functions that act on functions. And linear operators are functions. But that's getting a bit too far ahead of ourselves.

OK. So the point is that the 18.01-- so far, we haven't done any derivatives more than 18.01, at least in my half. Alan went a bit further. But already we can start to see, hopefully, how this is going to generalize. So if you have a function f of x and you make a small change in the input, delta x, and you ask for the small change in the output to first order, which we can denote with this d notation, the derivative is the linear operator that gives us that, the linearization of that function for a small change in the input.

And that is exactly equivalent to what you learned in 18.01. But it's going to be easier now to generalize this to other kinds of inputs and other kinds of outputs, where, in 18.01, we move this to the side where we take df, dx. We divide them. For numbers, that's fine. For other kinds of things, that becomes a little bit weirder to talk about.

Of course, you could define it as notation. But I think it's a lot clearer if you think of it in this sense once you start generalizing to other kinds of objects. So with that said, let's do that. Now, let's revisit 18.02. And let me do it in two parts.

So part one is going to be functions-- the first thing you usually do in 18.02, which is functions that take a vector in or multiple variables in. But we'll think of it as a vector in and a scalar out. So we're going to have a scalar. My output is blue, right? Yes. We'll keep the same color scheme, good.

So we're going to have a scalar function f of a vector input x. And I'll put a little vector sign above it. I won't always do that, but it's nice to be clear sometimes, which is a vector, which is a scalar. So x is going to live in our m. So this is going to be an m component column vector.

OK. And what we want to do is imagine what happens to the output when you change the input by a little bit. So we're going to take f of x plus dx. Think of this as a really small change. It's infinitesimal. We're going to drop anything that goes like dx squared or anything like that, any higher terms.

Let me just move it to [INAUDIBLE]. It's black-- minus f of x. And we're going to define this as f prime of x dx. So we wanted to note we have an arbitrary change in the inputs. dx is an arbitrary, very, very small vector.

And we want to ask, what's the change in the output differential, df? And the answer is going to be that this is going to be, for a very small dx, we can approximate this by a linear operator on dx. So this is going to be a linear operator, always going to be a linear operator.

And what is that linear operator do? This one takes a vector in and gives you a scalar. So this has to equal a-- df is a scalar, but dx is a vector. So what this has to be is it has to be kind of a row vector.

You can think of it more as a one-row matrix. Or there's fancier names for this, like covector or dual vector. We won't really use that. I just want to throw them out there.

So if you want to take a vector in and take a vector out, you need to multiply by a row vector. Another way of thinking about it is you need to take the dot product--

ALAN EDELMAN: The vector and the scalar out.

STEVEN G. JOHNSON: The vector and scalar out-- sorry. You need to multiply it by a one row thing. Another way of thinking about it is that, if you have a linear operation that takes a vector in and gives you a scalar out, the only type of thing that does that is a dot product.

If you take a dot product of the vector, you get a scalar. And that's the only linear operation that gives you a scalar from a vector in some sense. And so this is a dot product with some vector. That vector must be pretty special.

And so we'll give it a name. And we'll call that the gradient. So I think this is going to be the gradient of f. And this is the thing we take the dot product with to get our scalar df.

So you can think of this in linear algebra terms. So this dot product is the same thing as multiplying by a transpose. So this is the same thing as-- this is saying that f prime is really grad f transpose. Or equivalently, f prime dx is the operation of a dot product with a gradient.

This is going to be really powerful pretty soon because it's going to also allow us to generalize gradients to other kinds of vector spaces. As long as we have a dot product and a scalar function, you'll be able to define a gradient. So if you have a scalar function of-- if this is something that takes a matrix in and a scalar out, like a determinant, pretty soon we're going to be able to take the gradient of a determinant. We'll be able to define what that means.

ALAN EDELMAN: Just to be clear, what is on the other side of that equals sign that you just wrote?

STEVEN G. JOHNSON: So, yes, the f-- yeah, I should write that. That's the f prime. Yeah. So this is-- yeah. I need to-- my equals-- let's see. The gradient of f is the thing we take a dot product. And here it's f prime of x is this.

ALAN EDELMAN: Exactly. It's f prime of x. And it's not df.

STEVEN G. JOHNSON: Yeah. It's not-- no dx. Just the f prime by itself is a row vector. That's the transpose of the gradient.

So now, that's the definition of the gradient. It's only something we're usually going to define for scalar functions of vectors. And pretty soon, we'll be able to generalize that to other kinds of vectors, but will still be a scalar. And it's the thing you take the dot product with of dx, of the differential, and the change in the input with to get the change in the output, OK?

Yeah. So we did-- Alan did the example of x transpose x. Let's do another example just for fun. So suppose f of x. So x is going to be a vector. Suppose that x transpose Ax where-- so this is x. x here is going to have m components. And so we're going to let A be an m by m matrix.

ALAN EDELMAN: And are you assuming asymmetric or--

STEVEN G. JOHNSON: No, I'm not. I won't make it symmetric just for generality, OK? And so this is going to be-- A is not going to be an input. This is just going to be a constant matrix.

So when I do my d's, when I change x, A does not change. OK. So let me do it the long way first, and then we'll try and derive some rules. So let's do the long way-- long way, but still faster than doing it component by component, faster than the 18.02 component by component.

Because I could take the derivative of this with respect to x1, with respect to x2, with respect to all the components, and then build up the gradient. It starts to become really awkward really quickly.

I know it can be tempting when you're faced with new problems to fall back on what you know, right? And that's not a bad strategy. But we really want to encourage you to-- even though you know calculus really, really well, you know how to take derivatives really, really well, 18.02 and 18.01 style, we're going to try and learn something new here, a new way that can be really a lot more powerful besides [INAUDIBLE].

ALAN EDELMAN: I call it the way for big boys and big girls.

STEVEN G. JOHNSON: Yes, for big kids.

ALAN EDELMAN: For grown-ups.

STEVEN G. JOHNSON: The big-kid way.

ALAN EDELMAN: Kid way.

STEVEN G. JOHNSON: OK. So df, let's just do it slowly. So what we want to do is we want to take f. We're going to take-- I'm going to draw my vector symbols here. I guess I'll put them here. But I get tired of writing them all the time.

All my x's, and therefore my dx's, are vectors. So think of it as an arbitrary small change in an arbitrary direction. We want it to be able to handle anything like that.

ALAN EDELMAN: And in case it wasn't already obvious to everybody, what is the output of f? Is it a scalar, a vector, a matrix? It's a scalar, exactly. Just wanted to make sure everybody realized--

STEVEN G. JOHNSON: Sorry, yes.

ALAN EDELMAN: --that this is a scalar function of a vector.

STEVEN G. JOHNSON: Yeah. You could also write this as x dot product with ax. That's the same thing. OK. So I'm just going to do this out. I think it's still a little bit laboriously, but we'll have a better rule for-- we'll do the product rule in a minute. But let's do it without the benefit of that.

ALAN EDELMAN: Because you're effectively deriving the product rule in what's about to come.

STEVEN G. JOHNSON: Exactly, yes. So what do we do? So I'm going to plug-- I'm going to take f of x plus dx. I'm going to subtract fx. And I'm going to drop-- because it's d's, I'm going to drop anything that looks like a d squared, a dx squared, something that goes to 0 faster than dx.

So what's f of x plus dx? Well, I take x. I add dx, and I plug it into f. So there's a transpose there. There's an A. There's an x plus dx there. And then I subtract x transpose Ax.

And then I can just multiply everything out. Just think of dx as a really small vector. And just use your ordinary rules from linear algebra. I can just use the distributed whatever it's called, the distributive rule. I just have to make sure I don't change the orders of anything since these are vectors and matrices.

So I can multiply. There's this term times this term times this term. That's the first term. So there's an x transpose Ax. There's also this term times this term times this term. So that's a dx transpose Ax.

So dx is just a little vector. It's perfectly fine to transpose it. And then I also have this term times this term times this term. So that's x transpose Adx.

And then I have this term times this term times this term. But that has a dx squared. And so that I'm going to just-- let's see. Well, I actually am going to [INAUDIBLE] here. So that term is your dx transpose Adx. This term is gone. This is high order.

So in the limit as dx gets smaller and smaller and smaller, this term is negligible compared to these terms. And then I still have this term over here, can't forget that. Otherwise, these terms are negligible compared to this. But it's OK because I'm going to subtract that.

And that term cancels. And what's left is these two terms. And this is a perfectly good linear operation on dx, but it's not written in a very nice form. So the trick is to-- since these are numbers, I can transpose them. And this is something I feel like in 18.06 people get very confused by, that normally you're not allowed to change the order of anything.

Normally, a matrix is not equal to its transpose. But a number is always equal to its transpose. So this is since-- let me do a note over here.

Since dx transpose Ax is a scalar, a scalar is always equal to its transpose. So then we can take this and set it equal to its own transpose, which is the same thing as x transpose A transpose dx.

And again, make sure you understand that. This somehow is a source of endless confusion. Alan already talked about it when he said, for example, dx transpose x equals x transpose dx. That's a little easier to understand because it's just a dot product. You can swap things, a scalar product.

So this is also an instance of the same rule. But you know, it's a little bit more complicated looking. It's the same idea. Because this is a number, I can transpose it. And that swaps everything around. And the transpose of a product-- hopefully, you remember from linear algebra that the transpose of a product is the product of the transpose in reverse order.

So this term here, this whole term, equals x transpose A transpose dx. And what that allows me to do is it allows me to combine the two terms. Now, this term and this term, they both have a dx on the right. And so I can put that over in the right.

And I can put parentheses. And I have two terms, and both of them have an x transpose on the left. And the first term, one of the terms, has an A. The other term has an A transpose.

So that means this thing here is our derivative, f prime. Again, don't get that confused with the differential. f prime times dx is df. That's the change in the output. That's the little change in the output.

f prime is the rate of change. It's the thing you operate on dx. This is a row vector. Notice that this is a row vector.

ALAN EDELMAN: Could I ask the class? Then quickly tell me what is the gradient of x transpose dx. Anybody want to shout it out? What is the gradient?

AUDIENCE: You [INAUDIBLE]?

ALAN EDELMAN: Right. Do you want to say it in its full glory?

AUDIENCE: OK. A plus-- A plus A transpose times x.

ALAN EDELMAN: Good, yup. A plus A transpose times x is the gradient, exactly.

STEVEN G. JOHNSON: Right. It's just the transpose of this thing. So we transpose this thing. dx goes over on the right. This thing gets transposed. But this is symmetric, so it equals itself.

ALAN EDELMAN: So, Steven, on the clock here, we're already at 12:56. So you might want to kind of--

STEVEN G. JOHNSON: OK.

ALAN EDELMAN: --come to a conclusion.

STEVEN G. JOHNSON: Yeah. So this is just revisiting the notion of a gradient. And so next time, we're going to continue so that next time, basically, we're going to do 18.06 revisited part two. And we're going to have f is now going to be a vector function that takes a vector of outputs and also takes a vector of inputs.

And so we have outputs in, say, Rm inputs in, say, Rn. I probably should have used n just to be consistent before. And then what we're going to find is that then-- you can almost do it right now.

df has to be a linear operator that takes a small change in the input and gives you a small change in the output. And so this has to have m outputs, m components here. It has to have n components there. The only way you can get a linear operator that takes n inputs and m outputs is that this has to be an m by n matrix.

ALAN EDELMAN: It has to be expressible as an m by n matrix.

STEVEN G. JOHNSON: Exactly.

ALAN EDELMAN: But you don't have to write down the matrix.

STEVEN G. JOHNSON: Yes, that's right. So this is our f prime of x linear operator. And this, if we write it down as a matrix, we call it the Jacobian.

ALAN EDELMAN: I think what I'll do next time is show my favorite nonlinear operator on two-dimensional space, which is hyperbolic operators on corgis. So you'll see hyperbolic corgis on Friday if you come.

STEVEN G. JOHNSON: So probably on Friday maybe we'll switch. And I'll start with the first half and then finish this up. And then Alan can do the second half.

ALAN EDELMAN: OK, we could do it that way. OK.

STEVEN G. JOHNSON: Thanks, all. Any questions at this point? But let's--

ALAN EDELMAN: I'll also stick around. You can ask Steven or, you know--

STEVEN G. JOHNSON: Let me stop the recording.

ALAN EDELMAN: OK. Otherwise, see you on Friday, same room, 11:00 AM.