[AUDIO LOGO]

[MOUSE CLICK]

STEVEN JOHNSON: OK, so let's get started. So as I said, I'm going to do second derivatives, which is just going to be the derivative of the derivative. When you have functions from matrices to matrices or something like that, you have to think a little bit carefully just to make sure we understand what kind of thing the second derivative is.

So remember when we take basically the first derivative, what we have is this linear operator, f primed of x, that takes in a little change and gives us a df, which is the f of x plus dx minus f of x to first order, dropping higher-order terms. And so it's really natural to define the second derivative as the derivative of that.

So what should f double prime be? f double primed of x should take in-- let me call it dx prime. I mean, that's not a derivative. It just means a different-- we put it in here in red. This is going to be a different small change.

So prime here is not derivative, it's just different. We call it tilde or something like that. But we overload the prime a lot in mathematics.

So what should it be? Well, it should be the derivative. I mean, it should be exactly the same thing. It should be df primed. So it should be f primed of x plus dx primed minus f prime of x. So it should look exactly the same.

But now let's think about what this means. So this f prime here is not a number anymore. It could be a number, but in general, it's a linear operator. So this here is the difference of two linear operators.

And what does it mean to add and subtract and take differences of linear operators? And we did a little bit of this in problem set 1. You can think of linear operators themselves as a vector space. Just like we can take functions, sine x plus cosine x is another function. So we can think of linear operators as a vector space as well.

So if we have linear operators-- operators L1 and L2, then what we mean by their addition, L1 plus L2, or plus or minus, is the linear operator that takes a vector v and gives you L1 of V plus L2 of V.

And the same thing if you multiply a linear operator by a scalar, that's the linear operator that sends a vector V to alpha times L1 of V. And this should be familiar. So if I take two matrices and I add them, you'd actually know what that means. You say, oh, that means you add the elements up element-wise.

But where does that rule come from? That adding up two matrices element-wise is exactly the new matrix that, acting on a vector, gives you the first matrix times the vector plus the second matrix times the vector.

If I take a matrix and multiply it by 3, you can say, oh, you multiply all the entries by 3. But why? It's because if I take-- that's the linear operator that takes a vector, it's equivalent to multiplying it by that original matrix and then multiplying it by 3.

So that's where those rules come from, just like the rule for multiplying two matrices-- oh, rows times columns. Why? Because that's the new linear operator that's equivalent to multiplying the first matrix by the right matrix and then by the left matrix. That's where these rules come from.

So this thing here is now-- so what we really mean by this is this is a linear operator, so that this is the difference of two linear operators. These are linear operators that take in a dx and give you something else, give you a df.

So the difference of these two things is a linear operator that takes in a dx-- not a dx prime, a dx-- and gives you something else. So this is a linear operator that takes in-- we can call it f double prime of x. It takes in a dx primed that takes in another input. We can take it, write it as it takes in an input dx. And what it gives you is f primed of x plus dx primed, acting on dx, minus f prime of x acting on dx.

Does everyone see? So this f double prime is going to be-- it's just copying down the formula for the derivative, except applying it to f prime. But then-- so this is the linear operator you get when you take the original linear operator at a slightly shifted x minus f prime of x. So this is a linear operator that acts on a dx. And what it is, it's just the linear operator that takes this on dx minus this on dx.

But then if you think about this object, now, this object now takes two inputs. It takes the change x where dx prime is the change in where you're taking the derivative. And then the second one, dx, is the change in x that you're acting the derivative, f primed, on. So this is called a bilinear form.

So we have-- we're going to have f double primed that's taking of x that takes two inputs, dx prime, and dx. But writing those two pairs of brackets is a little bit annoying. So let me just write one pair of brackets and just give it two arguments.

And so it acts on two vectors. And it's linear in both. So in general, a bilinear form-- and I said let me just remind you of linearity. You have a bilinear form, call it B, takes in a vector u, and it takes in a vector v. And linearity means if you take B of u1 plus u2, comma v, that had better equal B of u1, v plus B of u2, v.

But it also has to be linear in the second argument. B of u, v1 plus v2 has to equal B of u, v1 plus B of u v2, et cetera. You can also scale the-- multiply one of the arguments by 2, and so forth. If I multiply both of the arguments by 2, if I do B of, like, 2u times 3v, that had better be 2 times 3 B of u, v. So if I multiply both arguments by 2, it doesn't multiply the output by 2. It multiplies it by 4.

So that's what this second derivative is. It takes in a change-- two changes, a change in where you're taking the derivative and a change in the thing that you're taking the derivative on. And you might wonder if it matters, like the order-- oops, and this should not have been a prime. If it matters, does the order of these two things matter?

And in general for bilinear forms, the order matters. In general, for bilinear forms, B of u, v is not equal to B of v, u. And, in fact, it may not even make sense to swap the arguments. You can have a bilinear form where u lives in one vector space, and v lives in a completely different vector space. So u is a scalar, and v is a column vector, and it doesn't even-- you can't even swap them. So I said this may not even make sense.

But here, dx and dx prime, they clearly live in the same vector space. They both changes to x. And so you could swap them and ask, what's the change? And, in fact-- but here, it turns out they are the same. So here, it turns out that f double prime of dx primed-- I'm making my color scheme right. I was using black, right?

f double prime of x-- there's the point at which you're evaluating the second derivative-- of dx primed comma dx always, in fact, equals f double primed of x dx comma dx primed. And so this is what's called a symmetric.

And we can show it pretty easily just from the definition. So in fact, as I'll show in a minute, this is actually-- you've seen this kind of thing before, and you didn't realize it. So you learned in 1802, points of variable calculus, that when I take a partial derivative, like partial f, partial x, partial y, like two n's, I can swap them, and it doesn't matter. It turns out that is going to be a special case of this.

So let me show it in general just from the definition. So why is this? So why symmetric? And so we just need to write out the definition a little bit. So just write out-- so f double primed of x of dx primed dx-- I'm not going to use colors here because it gets too annoying to swap pens back and forth.

OK, so there's two terms here. So there's a term that comes from f prime of x plus dx primed acting on dx minus f of x f primed of x acting on dx. That's just the definition I did before. f double primed is the difference of these two linear operators. So acting on dx is the difference of what they do on the dx.

But now let's take it one step further. I want you to expand out. What's the definition of f prime? Let's go back to that. f primed of-- f primed, and let's do the second one. That's easier. So the second term-- what's f primed of x dx? We said in the very first lecture, that's the same thing as f of x plus dx minus f of x. That's what that is. So what's f primed of x plus dx primed dx? It's f of x plus dx primed plus dx minus f of x plus dx primed.

Now, let me just regroup these terms a little bit. So I have one term that looks like f of x plus dx primed plus dx. And I have another term that looks like just a plus f of x. That's not very interesting. And then I have another term that looks like f of x plus dx with a minus sign. And I have another term that looks like f of x plus dx primed.

But now, if you stare at this clearly, a vector addition is commutative. I can swap those. I can swap these two terms. I can swap the addition and the input. I can also swap the addition of the output. I can just rearrange these two terms. And so this whole thing is exactly what-- the same as what you would get as if I took f double primed of x and evaluated it, and I swapped the outputs.

So it's just writing out the definition. This does not look symmetric because here, the dx prime is in the argument, and dx is what f is acting on. But when you write out the definition of the derivative, then you see that both dx and dx prime appear in the same way.

So let's do-- so the 1801 example is very easy. The first derivative is a number, and the second derivative is also a number. That's kind of boring. So it's the familiar-- this is a strict generalization of what you learned before. It's nothing new.

But let's do an 1802, multivariable calculus. So suppose we have a scalar function f of x. All right, so the output is a scalar. But the input is going to be-- this is going to be in R n. So this is an n component vector. Now, you can think of it as a column vector if you want.

So then what do we know? So we know that f primed of x is a row vector. It's the transpose of the gradient. So that way, f prime of x times dx is a scalar. That's the only way to get-- that's the only linear operator that takes a vector in and scalar out.

So now let's think about what is our second derivative, our f primed of x. It has to take in two vectors, dx prime and dx. Now, it doesn't really matter in which order I put the prime. But it has to give a scalar.

In the end, the output has to match. If you plug in both these things, it matches f. You can't just go back to the definition that f primed takes in a vector and gives you the same output as f because it gives you the df. So f double primed, when you plug in both vectors, or dx prime and dx, it has to give the same output as f plug in of dx. So it has to give something the same shape as f.

So this takes in two vectors and gives you a scalar. And it turns out there's only one kind of way to write down an operation that takes in two column vectors and outputs a scalar and is linear in both of the vectors. And that is to put the matrix here and put a dx here and a dx plug in transposed over there.

And it has to be symmetric. We know it has to be the same thing if you swap the dx and dx prime. But that means, actually, it has to be a symmetric matrix because this is also a number. So it equals the transpose of itself.

This is true for any number. You can always equal the transpose itself. So that equals dx transpose h transpose dx primed. So these two have to be equal. So you have a symmetric n by n matrix H, which is called the Hessian matrix.

How many of you have heard of Hessian matrices before? Yeah, it's a fair number of you. So yeah, so that's what a bilinear form looks like acting on column vectors. It's a matrix.

But it's nice to write it in 1802 terms more explicitly, like component-wise, just like the gradient you need to know when you first learn it. So it's partial f partial x1, partial f partial x2, and so forth. H is going to be the same thing, so same kind of thing.

So now if we do it explicitly, let's just write out slowly. So we know that the gradient of f is the vector that has partial f partial x1 to partial f partial xn. So that means what we want to take is the change, d, of gradient f, say transposed, I guess. Yeah.

So what's d of this, for example? So well, it's the same thing as d of partial f partial x1 to d of partial f partial xn. But partial f partial x1 is a scalar function of x, of all the-- it depends on all the components in x. And it spits out a number, which is partial f, partial x1.

So we know what the d of that looks like. The d of that is a gradient of partial f partial x1 transpose dx. Say all the way to gradient of partial f partial xn transpose dx.

So we can write that out as-- we can pull out the dx column vector. And that's gradient of partial f partial x1 transpose all the way to gradient of partial f partial xn transpose. But that matrix-- and with that, what is the gradient? That's partial f-- that's partial squared of f partial x1 partial x1 to partial f partial x1 partial xn. That's what the gradient transpose looks like. I take partial partial f x1 and take its derivative with respect to x1 to xn. So that's a mixed second derivative.

And then in the last row is the same thing, partial squared partial f partial xn. And then I take its derivative with respect to x1 all the way to its derivative with respect to xn. Derivative's xn dx. And this is a matrix of second derivatives.

This is exactly going to be-- I guess this was-- this is, I guess, H, or it's equal to-- I guess it's H transpose, because grad f was the transpose of the derivative. So this gives you the change. This is the d of f prime transpose.

So it's H transpose. But we know that that equals H. So it doesn't really matter. And so then, we get that the H, the IJ element is just partial squared f partial xi partial xj. But that equals xji because we know in general that the Hessian is a symmetric matrix, is a symmetric bilinear form, which means this is a symmetric matrix.

So that gives you this relationship you learned in multivariable calculus that when I take partial derivatives, I can swap the order. I mean, it just comes from the definition of the derivative. But the point is that that extends to more-- when you extend it to more general objects, it turns into this symmetric bilinear form business. Any questions? Yeah?

AUDIENCE: Just a quick clarification-- this should be an n1? Like, this should be reversed, right, just in the matrix itself?

STEVEN JOHNSON: Which? The partial derivatives, these should be reversed?

AUDIENCE: Yeah, technically.

STEVEN JOHNSON: I guess it depends on-- yes. It depends on what you mean by-- because we're so used to the fact that this-- you can take the derivative in either order. So actually, I don't know. Does it mean you take this derivative, then this derivative, or the other way around? If I put parentheses here, I think it's right.

AUDIENCE: OK.

STEVEN JOHNSON: Yeah, because it's-- but the notation is kind of ambiguous because it doesn't need to be explicit because you can swap the order. But yeah. So I think yeah, if I put derivative parentheses here, that's the explicit thing. I take partial f partial x1. I take its derivative with respect to x1 all the way to xn.

But if you put parentheses there, then it's the opposite. But at the end of the day, it's not going to matter because this is symmetric.

AUDIENCE: Yeah.

STEVEN JOHNSON: OK, so that's-- how many learned the Hessian matrix this way? Basically, it's the matrix of all the mixed second derivatives. Yeah. So that's usually how it's presented, right? And that's-- yes?

AUDIENCE: And just conclude that it's the Jacobian of the gradient, in a way?

STEVEN JOHNSON: Yeah, it's exactly the Jacobian of the gradient. Yes. Good, good, good. And we write that down. So this is the-- Yeah, exactly.

Yeah. But now we have it in a much more general setting. So I think it's nice to do-- let's do a more general example, an example that it's not so easy to do with 1801, or 1802 even.

Let's take our favorite function, our new favorite function, that takes in a matrix and gives you a scalar. So from the previous lecture, we learned what the derivative of this is. So if you take f primed of A, that's the linear operator acting on dA, what it gives you is determinant A times the trace of A inverse dA.

Or equivalent, we showed that the gradient of f was the determinant of A times the transpose of this, A inverse transpose, which is called the adjugate matrix. Yeah, the adjugate or the transpose of the adjugate-- sorry, the adjugate transpose. of A. I always forget which is which.

It's the cofactor matrix of A. So yes, we saw this, and there's various ways you can show this. Professor Edelman looked at a couple of different ways.

So now let's go one step further. And now let's take the second derivative. And whenever we're faced with something new and confusing, it's always good to fall back on the definition. So all we're doing is just going to take-- we're going to take d of this, d of this whole formula, determinant A trace A inverse dA.

I'm going to put a prime here, just to make it clear that what I'm changing is A, not dA. I'm going to use a dA prime. So d primed-- so what this is going to be, is this is going to be our f primed of A plus dA primed acting on dA.

Let me give myself a little more space. Minus f prime of A on dA. So the prime here is just-- it's not another derivative. I'm overloading my primes a little bit. It just means I'm using-- I'm changing A by dA primed. OK, and dA is going to be fixed. Yes?

AUDIENCE: In the previous example as well, that was technically dx prime, then, just to clarify that?

STEVEN JOHNSON: Yeah. Yeah, well, I didn't have any dx's here. So I didn't need a d prime.

AUDIENCE: Right.

STEVEN JOHNSON: Yeah, yeah. But I could have used a d prime there if I wanted. But it's really the same kind of thing. It's just I already have a dA here. And I want to be careful that I'm not changing this. This is now going to be-- this is-- actually, let me put this in blue here. This is dA, dA, dA. This is going to be fixed.

So dA is not changing. So we can think of it as a constant. So when I change things, I'm changing A by dA primed. And so now I can just use our derivative rules, so our product, and our chain rules, dot, dot, dot. And what do I get?

So now, I'm just treating dA as just a constant matrix I'm sticking in there. And then I'm taking the derivative the same way as before. Well, I have the derivative of this term times this plus this times the derivative of that term. And the derivative of this term, we just-- or the differential, sorry, of that term, we just saw what it was. This is our-- it's this. It's this, right?

So the first term is determinant of A times-- so the d of the determinant is the determinant of A times the trace of A prime-- not A prime, A inverse dA primed. All right, so this came from that term times the other term, trace of A inverse dA plus I still have my determinant of A times the derivative of the other term.

And trace is a linear operator. So I can just take the derivative inside. And so now-- or the differential inside-- and now I need dA inverse. And we know what dA inverse, that was a minus sign.

So let me change this to a minus sign of A inverse dA primed, A inverse, and then there's still a dA. Can everyone see that? So this term here is exactly d primed of A inverse-- well, with the minus sign.

And now the question is, is this-- this is it. This is-- it's not going to get much simpler than this. This is our bilinear form. This is bilinear.

And in dA primed and dA, so I stick in any dA primed, any dA. Clearly, this is linear in each one of them individually. I never-- I can multiply dA by dA prime. It's quadratic. But I can't multiply a dA by itself.

And is it symmetric? Well, let's look. This term is the same as this. So if I swap-- and these are just-- trace is just a number. So I can swap these two terms. And that's fine. And this is also a number.

What about this? If I swap dA and d primed, it looks a little bit different. But remember, the trace has the cyclic property. I can move the A inverse dA over to the beginning, and then it looks like trace of A inverse dA, A inverse dA primed. So this is symmetric using this cyclic property of the trace.

That's it. This doesn't simplify. We can't write it as a Hessian matrix unless I vectorize things. So if I do vec of dA, this could be some big matrix-- n squared by-- yeah, it's n squared by n squared matrix that has a vec of dA on one side and vec of dA primed on the other side. But it's not very natural to do that. And it's in some ways, it's easier to do this.

So I want to talk a little bit about why second derivatives? And of course, they come up in lots of cases. But let me just mention a few salient things. So first of all, they give you-- the first derivative gives you a linear approximation of a function linearization. The second derivatives give you quadratic approximations.

So if you have f of x plus a little change. Let's call it delta x. It's not infinitesimal anymore. So this is going to be an approximation. This is approximately f of x plus f primed of x delta x. That's our linear approximation from the first derivative, our finite difference approximation, if you think of it.

And now there'll be a new term that'll look like f double primed of x with a delta x, a delta x, and I'm missing something. What am I missing? Just think of 1801, Taylor series.

AUDIENCE: [INAUDIBLE]

STEVEN JOHNSON: Is 1/2, yeah. And then there's higher order terms, a little low of delta x squared. We're dropping terms. So we're dropping terms that are smaller than quadratic.

So if it's three times differentiable, we're dropping cubic terms. But maybe the function doesn't have it, their derivative. But definitely what we're dropping are our terms that are higher than quadratic.

So the 1/2. You can derive this pretty easily by just-- if you take two derivatives of this, you'd better get back to f primed-- f double primed. It's at a better match, the second derivative.

But because the x appears twice in this, then if you take two derivatives back to delta x, you should get back to f double primed. But because delta x appears twice in this, you need to have a half there in order to get back to f double primed.

So we just write that in. So if we take-- OK, the second derivative with a respect to delta x, you had better get back to f double primed of x. And that 1/2 factor is necessary because it appears twice.

So otherwise, you'd see it get twice, just like if-- for the same reason you have it in the Taylor series. If these are just scalars, when I take the-- this is a delta x squared. When I take the second derivative, I'm going to get a 2. And I really want it to match the second derivative of my function at that point.

OK, so linear-- approximating things by other things is useful. Approximating things by quadratic functions is useful. So for example, if you're doing optimization, we approximate f by approximately a quadratic and then optimize the quadratic to get a step. This is sometimes called-- this is a variety of names. It's sometimes called Sequential Quadratic Programming, or SQP.

Technically, you-- also, if you have constraints, you make linear approximations of the constraints or, I guess, affine approximations. But colloquially, it's called linear approximations of products of constraints.

So if you have a-- but you're minimizing an arbitrary nonlinear function, it's back to arbitrary nonlinear constraints. That's hard. But if you approximate the function by a quadratic function using a second derivative, approximate the constraints by linear, that's called a QP, or Quadratic Program. And there are good methods to solve those kinds of things.

Another equivalent-- equivalently, optimizing a quadratic function is the same thing as-- so optimizing a function is equivalent to finding-- locally to finding a root of the gradient. So we're equivalently finding a root of the gradient of f. And you're going to approximate it by a linear, or by an affine, or let's say, linear-- colloquially linear, because equivalently, you're approximating it by f primed of x plus your f double primed of x delta x.

And so if you're finding a root of a linear function, you have a nonlinear-- you're trying to find the root of a nonlinear function grad f, and you approximate it by a linear thing, and you find a root of that, what's the name for that method?

AUDIENCE: Newton--

STEVEN JOHNSON: Newton's method. So this is just Newton's method.

And so in practice, finding the Hessian, or finding the f double primed or the Hessian matrix is often expensive. If f of x again takes a Rn to a scalar, then H is an n-by-n matrix. And this is huge if n is large.

So if you have a neural network where n is a billion, the Hessian is a billion by a billion matrix. You can't even store this matrix, much less compute it. So it's hard to get exactly for-- I guess it's in high dimensions.

So often, what you would try and do is you try and approximate. If you don't want to give up on it entirely, you approximate the Hessian in various ways. And so these give you rise to a variety of methods called quasi-Newton methods, the most famous of which is called the BFGS method, which is Broyden, Fletcher, Goldfarb, and Shannon, I think. It's named after four people who amusingly discovered the same thing in the same year independently, like there's three or four separate papers [LAUGHS] on the same thing. There's also a closely related method called Newton-Krylov methods, and so forth.

So I don't have time to explain all these things. But these are some key words if you ever need to do this. So Hessians are useful for a small n. You can compute them explicitly. By even automatic differentiation, you can get Hessians for you.

But for big n's, you can't even store it, much less compute it. So then there are ways to kind of-- and it's a really intricate problem to approximate Hessians. Or you can compute the Hessian times in a particular direction quickly, like a Hessian times an operated on a dx. That you can get quickly. But yeah.