Matrix Calculus; Or, Throw Away The Matrix Cookbook!

2024-11-04 10:16:00

Over winter break, I spent some time watching Prof. Alan Edelman and Prof. Steven G. Johnson's course on Matrix Calculus. It was an IAP course at MIT, which I think means that it was taught informally to a broad audience over a shorter time-span. I've been looking for a course like this for a while——calculus had been a pain point for me for a while, and I found that analysis and algebra didn't do much to help me out. To give some illustrative examples, here are some questions that troubled me throughout my high school, college, and graduate studies:

1. I studied stats, and we'd often encounter problems whose solution was the total derivative of a function from vectors to scalars. This is fine, but often applying the 'chain rule' would require calculating derivatives of matrix terms as an intermediate step, which I never understood. Basically every derivative would end up being some stack of partials so it wasn't mechanically hard to find, but I could never remember whether it should be transposed, how the dimensions worked out, or why we were even allowed to do this.
2. On that note, it has never even been clear to me what df/dA even **is** when A is a matrix! Is it just a huge dictionary of partial derivatives? Does it have a 'shape'? Can we 'multiply' it with other things?
3. I really hate working with indices! I always get them mixed up, and even if I don't, excessive indexing makes calculations tedious and error-prone. In general, I'm a firm believer that mathematical 'broccoli'—stuff that you have to do to get where you want to go, but which causes no joy in the doing—should be avoided whenever possible. Yet how many times have I had to unpack three or four sums to differentiate an expression that can just be written as some compact matrix product?
4. To step back all the way to high school, what in the world is implicit differentiation? Why does the chain rule require you to 'multiply by the derivative of the inside'? Why can you sometimes treat df/dx as a fraction, and sometimes not?
5. What is a Jacobian? What is a derivative? What is a gradient? What is a partial derivative? Why do all of these things have different names, and

> why do they have different shapes? Why do I constantly have to transpose the gradient for it to do anything interesting?

I think that the intuition of matrix calculus helps answer all of these questions in a pretty satisfying way. It's helped me reduce my reliance on The Matrix Cookbook and other tools, and it's really tightened my understanding of both linear algebra and calculus. The course itself covers a fair amount of content that isn't directly connected to this theme, and because its audience had a variety of mathematical backgrounds, it sometimes goes through a lot of concepts that more experienced students may be comfortable skipping. Thus, I've decided to write a brief explainer covering my major takeaways from the course, as well as a few interesting examples.

Here's a list of topics covered for reference:

Table of Contents

Part I : Differentials and Derivatives

To start the process of developing matrix calculus, we need to be clear about what we mean by a 'derivative'. To appreciate how hard this could be, recognize that, in different contexts, the derivative could be thought of as any of the following:

- The slope of the line tangent to x.
- The instantaneous rate at which $f(x)$ is increasing at x.
- The direction of steepest ascent of f (for functions $f : \mathbb{R}^n \to \mathbb{R}$——the 'gradient')
- The matrix whose (i, j)-th entry is the change of the i-th entry of f with respect to the j-th entry of x, in the sense of the first two definitions (for functions $f : \mathbb{R}^n \to \mathbb{R}$).

- etc. etc. etc.

If you've taken a multivariable calculus class, you probably know some form of the following statement, which is in some sense the most satisfying summary of all these interpretations:

The derivative of a function at x is its *best linear approximation* at x.

More strongly, we think that the derivative provides a basically *perfect* linear approximation, so long as we're really close to x. This intuition remains correct throughout the blog post, but we are going to add some additional structure to make it as explicit and organized of a notion as possible.

Differences of Functions

To start, let's say that we have some x that we care about, and we want to see how $f(x)$ changes when you perturb it——say, by some amount Δx. Then a natural quantity to consider would be the difference between the perturbed and the original input. Let's call this Δf:

$$\Delta f \triangleq f(x+\Delta x) - f(x).$$

Let's play this game with a concrete function as an example: $f(x) = 4x^2 - 3x + 1$. We find

$$\Delta f = f(x+\Delta x) - f(x) = 4(x+\Delta x)^2 - 3(x+\Delta x) + 1 - 4x^2 + 3x - 1$$

$$= 4x^2 + 8x\Delta x + 4\Delta x^2 - 3x - 3\Delta x + 1 - 4x^2 + 3x - 1$$

$$= 8x\Delta x - 3\Delta x + 4\Delta x^2$$

$$= (8x-3)\Delta x + 4\Delta x^2.$$

A few things immediately jump out. First, assuming that we hold x constant, notice that we've constructed a *function* from input space to output space. That is,

$$\Delta f : \mathbb{R} \to \mathbb{R} \quad \text{is given by} \quad \Delta f(\Delta x) = (8x-3)\Delta x + 4\Delta x^2.$$

So Δf is just the name of a function (like f), and Δx is the name of its input. This is true no matter what we plug in for Δx; it doesn't need to be *infinitesimal*, or even particularly small! Let's say for example that $x = 0$. If I plug in $\Delta x = 5$, I get

$$\Delta f(5) = -3(5) + 4 \cdot 5^2 = 100 - 15 = 85.$$

If I compare this to $f(5) - f(0)$ explicitly, I also get that

$$f(5) - f(0) = (100 - 15 + 1) - 1 = 85.$$

So we see that Δf gives us exactly the change in f from $f(x)$ when we change the input from x to $x + \Delta x$! Moreover, just as a sanity check, notice that $\Delta f + f(x)$ gives us

$$\Delta f + f(x) = (8x - 3)\Delta x + 4\Delta x^2 + 4x^2 - 3x + 1$$

$$= 1 - 3(x + \Delta x) + 4(\Delta x^2 + 2x\Delta x + x^2) = 1 - 3(x + \Delta x) + 4(x + \Delta x)^2.$$

So we see that $\Delta f + f(x)$ is just f, reparametrized so that Δx is our input instead, and so we're now centering our function at $\Delta x = 0$ instead of $x = 0$!

We can do this for any function, and we'll get a corresponding function that exactly captures the change in f. It might not be so easy to write in a simple way (e.g. $f(x) = \sin(x) + \cos(x)$), but there's nothing stopping us from defining it in principle.

Differentials and Differentiability

However, you might notice that, at least in the example above, if Δx *is* small, then some of these terms don't matter. In particular, Δx^2 goes to 0 way faster than Δx does (e.g., if $\Delta x = 10^{-5}$, then $\Delta x = 10^{-10}$—only .001% of Δx!). Then if we choose some really tiny Δx, which we'll call dx, then the corresponding tiny Δf (which we'll likewise call df), is approximately given by

$$df \approx (8x - 3)dx.$$

I tend to think of the terms df and dx, which we'll call *differentials*, intuitively rather than formally, as some small vectors with nearly negligable norm (and negligable squared norm). But you can also think about dx as just another name for Δx indicating that it's small, and df as the 'leading term' in Δf—that is, a quantity satisfying

$$\Delta f = df + o(||dx||),$$

where $o(||dx||)$ is any function such that

$$\lim_{||dx|| \to 0} \frac{o(||dx||)}{||dx||} = 0.$$

Then intuitively, we'd like for this df to look basically like Δf when dx is small. We can formalize this to get at the notion of *differentiability* explicitly:

Definition (Differentiability). Let $f : V \to W$ be some function between two normed vector spaces, each with a corresponding metric $|| \cdot ||_V, || \cdot ||_W$. We say that f is *differentiable* at $x \in V$ if there exists some linear function $D_{f \leftarrow x} : V \to W$ such that

$$\lim_{||\Delta x||_V \to \mathbf{0}} f(x + \Delta x) - f(x) \triangleq df = D_{f \leftarrow x}[dx].$$

We call the linear operator $D_{f \leftarrow x}$ the *derivative* of f with respect to x. You might also know it as $\frac{df}{dx}$, J, $\nabla_x f$, and so on. Note also that, when this linear transformation is a matrix, vector, or scalar, we might overload notation by also referring to the associated matrix/vector/scalar of transformation as the derivative.

- *Note : I like this notation $D_{f \leftarrow x}$, even though I don't really see it used often. It emphasizes that D is just some linear transformation, which takes x points to f points. The reversed arrow direction is intended to make the chain rule more obvious, as we'll be composing linear transformations soon (see the chain rule section).*

Recall that, if $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ for some n, m, then it must be possible to write $D_{x \to f}$ as a matrix product with Δx. This is going to turn out to be the familiar *Jacobian* from multivariable calculus.

Let's get more familiar with this idea through a few simple examples.

Example 1: $v^T x$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = v^T x$ for some constant vector v. Then

$$f(x + dx) - f(x) = v^T(x + dx) - v^T x = v^T dx.$$

So we can say that $D_{f \leftarrow x} = v^T$. Note that, oddly enough, we find that

$$df = f(dx).$$

We also see that, if we had computed ∇f explicitly, we would've found $\nabla f = v$. Then we also have $df = \nabla f^T dx$.

Example 2: $x^T x$. Let's look at a slightly more complicated function: $f(x) = x^T x$. We see that

$$df = f(x+dx) - f(x) = (x+dx)^T (x+dx) - x^T x = x^T x + 2x^T dx + dx^T dx - x^T x$$

$$= 2x^T dx + 0,$$

since we established that $dx^T dx = ||dx||^2$ is negligable. Notice also that $D_{f \leftarrow x} = \nabla f^T = 2x^T$ in this case, so we again get that $df = \nabla f^T dx$!

Example 3: Ax Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be given by $f(x) = Ax$. Then

$$df = A(x + dx) - A(x) = A(dx).$$

Once again, you'll notice that $df = f(dx)$, and now $D_{f \leftarrow x} = A$, which is the jacobian of f! We can actually extend this result to an arbitrary linear function with ease:

Derivatives of Linear Functions

Theorem (Derivative of Linear Functions) For any linear function $T : V \to V$,

$$dT = T(dx).$$

Proof.

By linearity,

$$dT = T(x + dx) - T(x) = T(x) - T(x) + T(dx) = T(dx). \blacksquare$$

Though this may seem like a simple result, it actually comes in really handy for a ton of problems. For example, here are some linear functions that come up often in practice:

- diag(x), the matrix whose (i, i)-th entry is x_i, and all other entries 0.

- $\mathrm{tr}(A)$, the sum of the diagonal entries $\sum_{i=1}^{n} A_{ii}$.
- The transpose operator X^T, or more generally permutations of indices of matrices or vectors.
- The 'index-into' function $[A]_{ij}$, which picks out an entry from a matrix.
- Most generally, any matrix-to-matrix function whose elements are linear functions of the input matrix's elements.

Let's show this last one explicitly, just because it's easy and covers a ton of cases (including every one of the above).

Say $[M(X)]_{ij} = T_{ij}(X)$ for some linear functions $\{T_{ij}\}$. Then

$$[M(cX+Y)]_{ij} = T_{ij}(cX+Y) = cT_{ij}(X) + T_{ij}(Y)$$

$$= c[M(X)]_{ij} + [M(Y)]_{ij},$$

so $M(cX+Y) = cM(X) + M(Y)$.

Thus, we see that it's linear, and so the above theorem applies. ■

Part II : The Chain and Product Rules

All of this manual differentiation is pretty tedius, and it will become harder to do as we begin to tackle more complicated functions. As such, we should develop some general rules for performing differentiation, as in single-variable calculus. The easiest to show is the chain rule:

Theorem (Chain Rule) If $g : U \to V$, $f : V \to W$, and $h : U \to W$ is given by $h(x) = f(g(x))$, then

$$dh = D_{f\leftarrow g}[D_{g\leftarrow x}[dx]].$$

In particular, if U, V, and W are all spaces of real vectors, then their derivatives $D_{x\leftarrow g}$ and $D_{g\leftarrow f}$ are matrices, and so

$$dh = D_{f\leftarrow g} D_{g\leftarrow x} dx.$$

Proof.

Consider first $g(x+dx)$. Note that

$$g(x+dx) = g(x+dx) - g(x) + g(x) = dg + g.$$

Then

$$dh = f(g(x + dx)) - f(g(x))$$

$$= f(dg + g) - f(g) = D_{f \leftarrow g}[dg] = D_{f \leftarrow g}[D_{f \leftarrow x}[dx]]$$

by the definition of a differential. ■

This has a really intuitive interpretation—the linear transformation that takes dx to dh is just the transformation that takes dx to dg, followed by that taking dg to dh (df). This also helps us recognize that differentials behave really nicely: if you ever have some differential df in an expression, you can always just substitute in any defined $D_{f \leftarrow g}[dg]$ to change variables.

Next, let's prove the product rule:

Theorem (Product Rule) Let $f, g : V \to W$, and $h : V \to W$ be given by $h = f(x)g(x)$. Then

$$dh = D_{f \leftarrow x}[dx]g(x) + f(x)D_{g \leftarrow x}[dx].$$

You'll notice that including this dependence on x is kind of ugly; and, by the chain rule, we know that we can transform any expression in terms of some dg into an expression in terms of dx by just plugging in the differential definition. Then for the rest of the post, let's just repress the dependence on x, with the understanding that we can plug in dx's whenever we want.

This lets us rewrite the rule in a more simple form:

$$dh = dfg + fdg.$$

- Proof. *

$$dh = f(x + dx)g(x + dx) - f(x)g(x)$$

$$= (df + f)(dg + g) - fg = dfdg + fdg + dfg + fg - fg$$

$$= fdg + dfg + dfdg.$$

Finally, note that, since both df and dg have negligable square norm, so does $dfdg$, as it also has norm of square order. Then

$$= fdg + dfg + 0.\blacksquare$$

Application : Matrix Inverse

We now show that these properties allow us to derive some more interesting derivatives. In particular, let

$$F = A^{-1}.$$

Then $AF = I$, so by the chain rule (and the fact that $dI = 0$),

$$d(AF) = dAF + AdF = 0$$

$$\iff AdF = -dAF \iff dF = -A^{-1}dAA^{-1}.$$

Application : Elementwise Functions

Now, let's draw our attention to a popular set of functions. Namely, say that $F : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$[F(x)]_i = f(x_i)\ \forall i \in [n],$$

for some scalar function f. Then let's compute its derivative. We have that

$$[F]_i = f,$$

so

$$[dF]_i = d[F_i] = df = \frac{df}{dx}(x_i)dx_i.$$

Thus, letting $\frac{df}{dx} = \left[\frac{df}{dx}(x_1), \ldots, \frac{df}{dx}(x_n)\right]^T$, we find

$$dF = \frac{df}{dx} \odot dx = \text{diag}\left(\frac{df}{dx}\right) dx,$$

where $\odot$ is the Hadamard product (elementwise multiplication).

Part III : Tricky Derivatives

Though so far things have gone off mostly without a hitch, there's still a fair amount of art to matrix calculus. Just as in normal calculus, we need to construct a few tough 'rules' for differentiating common functions, and then the chain and product rules will help us cover lots of others built up out of these pieces. The determinant and matrix powers are the next such 'atomic rules' for us to parse.

The Determinant

For this, we'll use the definition of the determinant given below:

$$\det(A) = \sum_{\sigma \in S_n} \text{sign}(\sigma) a_{1\sigma(1)} \dots a_{n\sigma(n)}.$$

Consider in particular terms of the form $det(dA + \lambda I)$. We're broadly interested in figuring out what terms in this determinant have linear $O(dA)$ terms, since higher-order terms will go to zero in the limit. To get us there, let's first look at which terms in this big sum have λ^n? We need all n diagonal elements, so the only one that works is $\sigma = \mathbf{1}$, so we find

$$\prod_i (dA_{ii} - \lambda) = \lambda^n + \lambda^{n-1} \sum_i dA_{ii} + \lambda^{n-2} \sum_i \sum_j dA_{ii} dA_{jj} + \dots$$

For any other permutation, at least two terms are purely a_{ij} (since it's impossible for only 1 index to be misplaced), meaning that it's impossible to choose another term with only dA^1 terms.

Then consider

$$\det(I + dA).$$

As we showed above, this looks like

$$1^n + 1^{n-1}\text{tr}(dA) + [\text{terms involving } o(dA^2)].$$

Then

$$\det(A + dA) = \det(AI + AA^{-1}dA) = \det(A)\det(I + A^{-1}dA)$$

by properties of determinants,

$$= \det(A)(1 + \text{tr}(A^{-1}dA)).$$

then

$$d[\det(A)] = \mathrm{tr}(\det(\mathrm{A})A^{-1}dA).$$

Matrix Powers

Now, say that $F_n(A) = A^n$. We'll solve this by induction.

The chain rule gives us

$$dA^2 = dAA + AdA = \sum_{i=0}^{1} A^i dAA^{1-i}$$

.

Assume for some n that

$$dA^n = \sum_{i=0}^{n-1} A^i dAA^{n-1-i}.$$

Then by the chain rule,

$$dA^{n+1} = dAA^n + Ad(A^n) = dAA^n + \sum_{i=0}^{n-1} A^{i+1} dAA^{(n+1)-1-i}$$

$$= \sum_{i=0}^{(n+1)-1} A^i dAA^{(n+1)-1-i},$$

as desired. ■

There are other tricky derivative rules left to discover, but hopefully these are enough to get you started!

Part IV : Insights About the World

I want to spend the rest of this post unpacking some really nice inuition that this new formulation of the derivative gives us. Some of these ideas are really deep, while some are simple——the only common thread between all of them is that I think they're cool, and help me understand something about math.

Gradients via Inner Products

Let's think about functions of the form $f(x) : \mathbb{R}^n \to \mathbb{R}$; so-called *functionals.* To look at these functions, recall this foundational theorem from linear algebra:

Theorem (Riesz Representation Theorem) For any linear functional $F : V \to \mathbb{R}$ from real vector space V equipped with inner product $\langle \cdot, \cdot \rangle$ to $\mathbb{R}$ there exists a unique $\phi \in V$ such that

$$F(x) = \langle \phi, x \rangle \, \forall x \in V.$$

This is a general fact (which extends way beyond real vector spaces), but it has a really nice implication for us. It turns out, we can actually define the gradient relative to this representation:

Definition (Gradient) For a function $F : V \to \mathbb{R}$, the *gradient* at x $\nabla F(x)$ is the unique vector in V satisfying

$$dF = \langle \nabla F(x), dx \rangle.$$

—

This makes clear a lot of our examples from the first section, where derivatives seemed to be transposed gradients when we looked at functions from vectors to scalars.

But it doesn't stop there! Consider the function $F = \det(A)$, which maps from matrices to scalars. Recall the Frobenius inner product $\langle A, B \rangle = \mathrm{tr}(A^T B)$ We found in a previous section that

$$dF = \mathrm{tr}(\det(\mathrm{A}) A^{-1} dA).$$

This lets us immediately find that

$$\nabla_A F = \det(\mathrm{A}) A^{-T}.$$

Moreover, this formulation allows us to perform gradient ascent/descent just like in the vector-valued case.

Matrix Gradient Ascent

Say that we're at some point $x \in V$, and we seek to move in the direction where some function $F : V \to \mathbb{R}$ is increasing the quickest. If V was some column vector, it's already clear that this direction is given by the gradient. We now show that this remains true for any inner product space. Assume that $||dx|| = \epsilon$ is held constant, so we can just choose a maximizing direction to increase. Well, by the definition above, we find

$$dF = \langle \nabla F, dx \rangle,$$

so we seek

$$\operatorname*{argmax}_x \langle \nabla F, \epsilon \frac{x}{||x||} \rangle.$$

Well cauchy-schwarz says that this is maximized when x points in the same direction as ∇F!

In fact, we can also recover the first-order necessary optimality conditions for constrained optimization in vector spaces of matrices and tensors.

Kronecker Products

Recall the vector space isomorphism theorem——any vector space over the real numbers is isomorphic to some vector space $\mathbb{R}^n$. Well we can use this to make more exotic-looking derivatives into a more pallatable form.

For instance, the MIT course emphasizes the following identity for transforming matrices into vectors:

$$(A \otimes B)\text{vec}(C) = \text{vec}(BCA^T),$$

where $\otimes$ is the Kronecker Product. We can use this to turn ugly-looking matrix derivatives into nicer vector derivatives. For example, we saw that

$$d(A^TA) = dA^TA + A^TdA.$$

This could be annoying to work with, since the dA shows up in two places. But by applying the identity, we find

$$dA^TA = IdA^TA = (A^T \otimes I)\text{vec}(dA^T) = (A^T \otimes I)P\text{vec}(dA)$$

for suitable permutation matrix P. Doing the same for the second term and combining yields

$$\text{vec}(d(A^TA)) = (A^T \otimes I)(P + I)\text{vec}(dA),$$

which makes our derivative a matrix, as desired.

Implicit Differentiation

I was always really confused about why we could take expressions like

$$x^2 + y^2 = 1$$

and calculate 'derivatives' like this:

$$2xdx + 2ydy = 1 \iff \frac{dy}{dx} = -\frac{x}{y},$$

so-called 'implicit' differentiation. I think I've seen a proof before, but nothing that stuck with me. This linear transformation framing makes it really obvious what's going on. We ALWAYS have

$$dx^2 = 2xdx$$

and

$$dy^2 = 2ydy,$$

(and of course $d1 = 0$), so it's always true that

$$2xdx + 2ydy = 0$$

solving for dy yields

$$dy = -\frac{x}{y}dx,$$

and so $-x/y$ is our derivative by definition. No funny business!

Normal Vectors to Level Curves

Something that we often see in optimization is a situation where we have a function $F : \mathbb{R}^n \to \mathbb{R}$ and define a level curve by

$$F(x) = C$$

for some $C \in \mathbb{R}$. A fact about these level surface is that, at any point x, the vector $\nabla F(x)$ is normal to this surface at x. This is really useful for many geometric applications——for example, computing projections.

I've definitely seen some convincing proofs of this, but it's worth noting that our new machinery makes this fact immediate. Differentiating the constraint and using the Riesz representation theorem yields

$$dF = \langle \nabla F, dx \rangle = 0.$$

Since dx is constrained to keep x on the level surface, this tells us that, locally, the surface is orthogonal to ∇F, as desired.

Conclusion

Those are most of my thoughts for the time being. I'd be excited to hear if anyone has thoughts on how to

In any case, thanks for reading.

—Thomas