1 Notes

The goal for today is to review multivariable differentiation. Let's start with the Fréchet derivative.

Definition 1.1. Let V and W be normed vector spaces (you can assume for our purposes that the vector spaces are isomorphic to $\mathbb{R}^n$ for some n), $U \subseteq V$ an open subset of V. A functor $f : U \to W$ is called *Fréchet differentiable* at $x \in U$ if there exists a (continuous) linear operator $A : V \to W$ such that

$$\lim_{\|h\| \to 0} \frac{\|f(x+h) - f(x) - Ah\|_W}{\|h\|_V} = 0.$$

Note that such a linear operator is unique, since if A and B both satisfy the condition, then we have

$$\lim_{\|h\| \to 0} \frac{\|Ah - Bh\|_W}{\|h\|_V} = 0,$$

so for any $x \in V$, as we take $t \in \mathbb{R}$ to 0, we have

$$\lim_{t \to 0} \frac{\|(A-B)(tx)\|_W}{\|tx\|_V} = \frac{\|(A-B)x\|_W}{\|x\|} = 0,$$

and thus $Ax = Bx$.

We can thus introduce the following notation, one of $D_x f$, Df_x, $Df(x)$ is used to say that f is differentiable at x and denote the derivative at that point.

We have the following basic properties of the derivative

Proposition 1.1. *(1) If f is a (continuous) affine function, in other words $f(x) = Ax + t$ for some (continuous) linear operator $A : V \to W$ and constant $t \in W$, then for all $x \in V$,*

$$D_x f = A.$$

In particular, if f is constant, then $D_x f = 0$ everywhere.

(2) (Chain rule) If V, V', V'' are normed vector spaces, $U \subseteq V$, $U' \subseteq V'$, $f : U \to V'$, $g : U' \to V''$, $x \in f^{-1}(U')$, and $D_x f$ and $D_{f(x)} g$ exist, then $D_x(g \circ f) = (D_{f(x)} g) \circ D_x f$.

(3) If $m : V \times V' \to W$ is (continuous and) bilinear, then

$$D_{(v,v')} m(h, h') = m(h, v') + m(v, h').$$

1.1 Definitions specific to $\mathbb{R}^n$

If $f : \mathbb{R}^n \to \mathbb{R}^m$, then Df can be written as a matrix, which we call the *Jacobian*. In the particular case that $f : \mathbb{R}^n \to \mathbb{R}$, then we also call Df the *gradient* of f, which is also written ∇f.

Note that we can write $f : \mathbb{R}^n \to \mathbb{R}^m$ as a tuple of functions $(f_1, \ldots, f_m)$, and $D_x f$ is the matrix (when it exists) with rows $(D_x f_1, \ldots, D_x f_m)$. To prove this, note that $f_i = \pi_i \circ f$, and π_i is linear, so by (1) and (2) above, $D_x f_i = \pi_i \circ D_x f$.

We can also define the jth partial derivative of f as

$$D_{j,x} f := D_{t=0} f(x + te_j),$$

where e_j is the jth standard basis vector. By chain rule, again, we can see that if f is differentiable at x, then $D_{j,x}f$ is the jth column of D_xf, which is $(D_xf)e_j$.

The index notation for partial derivatives rather than the conventional notation is chosen to match Spivak's notation.

More generally, if $u \in \mathbb{R}^n$ is a vector, we can define the directional derivative of f in the direction of u to be

$$(\nabla_u f)(x) := D_{t=0}(f(x+ut)),$$

and observe that by chain rule, when f is differentiable at x, we have

$$(\nabla_u f)(x) = (D_xf)u.$$

Finally, there is a partial converse to the observation that when the derivative of f exists then all of f's partial derivatives exist and are given by the entries in the Jacobian of f.

Theorem 1.1 (Spivak 2-8)**.** *If $f : \mathbb{R}^n \to \mathbb{R}^m$, and D_jf_i exist and are continuous in a neighborhood of x for all $1 \le i \le m$ and $1 \le j \le n$, then f is differentiable at x and $(D_xf)_{ij} = D_jf_i$.*

We call a function f satisfying the hypotheses of the theorem *continuously differentiable.* Functions all of whose higher order partials are differentiable are called C^∞ functions.

2 Examples and Problems

(1) Verify the properties listed in Proposition 1.1.

(2) Generalize the property (3) of Proposition 1.1 to arbitrary (continuous) k-multilinear maps.

(3) (Weak chain rule for partial derivatives, Spivak 2-9) Let $g_1, \dots, g_m : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable at a, and let $f : \mathbb{R}^m \to \mathbb{R}$ be differentiable at $(g_1(a), \dots, g_m(a))$. Define $F : \mathbb{R}^n \to \mathbb{R}$ by $F(x) = f(g_1(x), \dots, g_m(x))$. Verify:

$$D_{i,a}F = \sum_{j=1}^{m} D_{j,(g_1(a),\dots,g_m(a))}f \cdot D_{i,a}g_j$$

Why do we need to assume that the g_i are continuously differentiable?

(4) Show that if $U \subseteq V$, the directional derivative $\nabla : V \times C^\infty(U,W) \to C^\infty(U,W)$ is linear in its first variable and satisfies *the Leibniz rule* in its second, meaning that for $t, s \in \mathbb{R}$, $u, v \in V$, $f : U \to W$ a C^∞ function, we have

$$\nabla_{tu+sv,x}f = t\nabla_{u,x}f + s\nabla_{v,x}f,$$

and for $a : U \to \mathbb{R}$, $f, g : U \to W$,

$$\nabla_{u,x}(f+g) = \nabla_{u,x}f + \nabla_{u,x}g \text{ and } \nabla_{u,x}af = (\nabla_{u,x}a)f(x) + a(x)\nabla_{u,x}f.$$

(5) Suppose $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ admits a local inverse at $x \in U$, i.e., a continuous function $g : W \to \mathbb{R}^n$ such that $f(x) \in W$, $g(f(u)) = u$ for $u \in f^{-1}(W)$, and $f(g(w)) = w$ for $w \in W$. Then if f is differentiable at x, show that D_xf is invertible if and only if g is differentiable at $f(x)$.

(Originally I forgot to say that we need to assume the derivative is invertible, but if the derivative is not invertible, the inverse can fail to be differentiable, as with $f(x) = x^3$ at 0, and it turns out this is an if and only if. I also forgot to assume that g is continuous, which I suspect isn't actually necessary, but that wasn't supposed to be part of the exercise.)

(6) Compute the following derivatives, where $M_{k\times\ell}$ is the set of $k \times \ell$ real matrices.

  (a) Let $f : M_{k\times\ell} \times M_{\ell\times n} \to M_{k\times n}$ be given by $f(A, B) = AB$. What is $D_{(X,Y)}f$?

  (b) Let $f : M_{n\times n} \to \mathrm{Sym}(n)$ be given by $f(A) = A^T A$. What is $D_X f$?

  (c) Let $f : M_{n\times n} \to \mathbb{R}$ be given by the determinant: $f(A) = \det A$. What is $D_{I_n} f$, where I_n is the $n \times n$ identity matrix? Can you work out $D_B f$, where B is an invertible $n \times n$ matrix?

3 A Proof of the Chain Rule

When you set out to prove the chain rule, you quickly realize that you need to control the variation of a differentiable function in a neighborhood of the point at which it's differentiable. Here is the appropriate lemma.

Lemma 3.1. *If $f : U \subseteq V \to W$ is differentiable at $x \in U$, then for all $\epsilon > 0$ there exists a $\delta > 0$ such that for all $h, h' \in V$ with $\|h\|, \|h'\| < \delta$ we have*

$$\|f(x + h) - f(x + h')\|_W \leq \|D_x f\|\|h - h'\| + \epsilon(\|h\| + \|h'\|).$$

Note the corollary that f is Lipschitz at points where it is differentiable, in the sense that by taking $h' = 0$ in the previous lemma, we have that for all $\epsilon > 0$, we can find a $\delta > 0$ such that for $\|h\| < \delta$,

$$\|f(x + h) - f(x)\| \leq (\|D_x f\| + \epsilon)\|h\|.$$

Proof of lemma. For a given ϵ, pick δ such that for $0 < \|h\| < \delta$,

$$\frac{\|f(x + h) - f(x) - D_x f h\|}{\|h\|} < \epsilon.$$

Then for all h with $\|h\| < \delta$, we have that

$$\|f(x + h) - f(x) - D_x f h\| \leq \epsilon\|h\|.$$

Now we have for $\|h\|, \|h'\| < \delta$,

$$\begin{aligned}\|f(x + h) - f(x + h')\| &\leq \|f(x + h) - f(x + h') - D_x f(h - h')\| + \|D_x f(h - h')\| \\ &= \|f(x + h) - f(x) - D_x f h - (f(x + h') - f(x) - D_x f h')\| + \|D_x f(h - h')\| \\ &\leq \epsilon\|h\| + \epsilon\|h'\| + \|D_x f\|\|h - h'\|,\end{aligned}$$

which is what we want. ■

Now we can prove the chain rule.

Proof of chain rule. We want to show that if $f : U \subseteq V \to V'$, $g : U' \subseteq V' \to V''$, $x \in f^{-1}(U)$, and $D_x f$ and $D_{f(x)} g$ exist, then $D_x(g \circ f) = D_{f(x)} g \circ D_x f$.

In other words, we need to show that

$$\lim_{\|h\|\to 0} \frac{\|g(f(x+h)) - g(f(x)) - D_{f(x)}g(D_x f h)\|}{\|h\|} = 0.$$

Now we can add and subtract $g(f(x) + D_x f(h))$ in the norm on the top, so by triangle inequality, it suffices to show that both

$$\lim_{\|h\|\to 0} \frac{\|g(f(x) + D_x f h) - g(f(x)) - (D_{f(x)})g(D_x f h)\|}{\|h\|} = 0,$$

and

$$\lim_{\|h\|\to 0} \frac{\|g(f(x+h)) - g(f(x) + D_x f(h))\|}{\|h\|} = 0.$$

For the first limit, since g is differentiable, for arbitrary $\epsilon > 0$, for $\|w\|$ small enough, we have

$$\|g(f(x) + w) - g(f(x)) + D_{f(x)} g w\| \le \epsilon \|w\|.$$

Taking $w = D_x f h$, since $D_x f$ is continuous, we have that for $\|h\|$ small enough,

$$\|g(f(x) + D_x f h) - g(f(x)) + D_{f(x)} g D_x f h\| \le \epsilon \|D_x f\| \|h\|.$$

Dividing by $\|h\|$ gives us that the desired limit is zero.

For the second limit, we apply the lemma. We have that for any ϵ, when $\|h\|$ is small enough,

$$\|g(f(x+h)) - g(f(x) + D_x f h)\| \le \|D_{f(x)} g\| \|f(x+h) - f(x) - D_x f h\| + \epsilon \|f(x+h) - f(x)\| + \epsilon \|D_x f h\|.$$

Reducing the bound on $\|h\|$ as necessary, for $\|h\|$ small enough, the right hand side is bounded by

$$\|D_{f(x)} g\|(\epsilon \|h\|) + \epsilon(\|D_x f\| + \epsilon)\|h\| + \epsilon \|D_x f\| \|h\|.$$

Then when we divide by $\|h\|$, we get that the second limit is also zero.

■

4 Solutions

(1) Proposition 1.1 (1) follows immediately from the definition, since if $f = Ax + t$, then $f(x + h) - f(x) - Ah = 0$ on the nose.

Proposition 1.1 (2) is proved in the section above.

Proposition 1.1 (3) can be proven in the following manner. Since $m(x + h, x' + h') = m(x, x') + m(h, x') + m(x, h') + m(h, h')$, the proof reduces to showing that

$$\lim_{\|(h,h')\|\to 0} \frac{\|m(h, h')\| \|(h, h')\|}{=} 0.$$

But

$$\|m(h, h')\| = \|h\| \|h'\| \|m(h/\|h\|, h'/\|h'\|)\| \le M \|(h, h')\| \|(h, h')\|,$$

for some constant M, since (in the finite dimensional case) m is bounded on pairs of norm one vectors because m is continuous and pairs of norm one vectors form a compact set isomorphic to $S^{n-1} \times S^{m-1}$ if $V \cong \mathbb{R}^n$ and $V' \cong \mathbb{R}^m$. (In the infinite dimensional case, a bilinear function m should be continuous if and only if we have such a bound, though I might be mistaken).

Dividing by $\|(h, h')\|$ and taking the limit we get the result.

(2) The same proof as in part (3) of the previous example gives that if $m(x_1, \ldots, x_k)$ is a multilinear map $V_1 \times \cdots \times V_k \to W$, then

$$D_(x_1, \ldots, x_k) m(h_1, \ldots, h_k) = \sum_{i=1}^{k} m(x_1, \ldots, x_{i-1}, h_i, x_{i+1}, \ldots, x_k).$$

(3) The reason that we assume that the g_i are continuously differentiable at a is so that we can combine them to form a function $g : \mathbb{R}^n \to \mathbb{R}^m$ whose components are $(g_1, \ldots, g_m)$, and apply Theorem 1.1 to conclude that g is differentiable at a.

Then our function $F = f \circ g$, so we can apply the chain rule to conclude that $D_a F = D_{g(a)} f \circ D_a g$.

The ith partial is given by multiplying with the ith standard basis vector e_i, so multiplying e_i on the right, we have $D_{i,a} F = D_{g(a)} f \circ D_{i,a} g$, which is the application of the gradient of f to the column vector whose jth entry is the ith partial of g_j. Expanding this out into a sum gives the desired result.

(4) Linearity in the first variable follows from the fact that $\nabla_{u,x} f = (D_x f)u$, and $D_x f$ is linear.

The second variable is additive because

$$\nabla_{u,x}(f+g) = D_x(f+g)u = (D_x f + D_x g)u = D_x f u + D_x g u = \nabla_{u,x} f + \nabla_{u,x} g.$$

The Leibniz rule is satisfied because multiplication $\cdot : \mathbb{R} \times V \to V$ is bilinear, so we have

$$\begin{aligned}
\nabla_{u,x} af &= (D_x(af))u \\
&= (D_x \cdot \circ (a, f))u \\
&= (D_{(a(x), f(x))} \cdot)(D_x(a, f))u \\
&= (D_{(a(x), f(x))} \cdot)(D_x a, D_x f)u \\
&= (D_{(a(x), f(x))} \cdot)(D_x a u, D_x f u) \\
&= (D_{(a(x), f(x))} \cdot)(\nabla_{u,x} a, \nabla_{u,x} f) \\
&= (\nabla_{u,x} a) f(x) + a(x) \nabla_{u,x} f.
\end{aligned}$$

(5) Without loss of generality, we can assume $U = f^{-1}(W)$, so $f : U \to W$, $g : W \to U$ are inverses. We are given that f is differentiable at $x \in U$. Let $y = f(x)$, and let $A = D_x f$. We are also given that A is invertible. We want to show that g is differentiable at y.

If g is differentiable at y, then $f \circ g = 1_W$ and $g \circ f = 1_U$, so we have $D_x f \circ D_y g = I_n$ and $D_y g \circ D_x f = I_n$, so it's a necessary condition that A be invertible for g to be differentiable. Moreover, when g is differentiable at y, its derivative must be A^{-1}.

Therefore, we need to prove

$$\lim_{\|h\| \to 0} \frac{\|g(y+h) - g(y) - A^{-1}h\|}{\|h\|} = 0.$$

Let $h' = g(y+h) - g(y)$, so that $f(x+h') = f(g(y+h)) = y+h$.

Then if we fix $\epsilon > 0$, we have that for $\|h\|$ small enough, $\|h'\|$ is small enough that we have $\|f(x+h') - f(x) - Ah'\| \leq \epsilon \|h'\|$, since f is differentiable.

But $f(x+h') = y + h = f(x) + h$, so this says that $\|h - Ah'\| \leq \epsilon\|h'\|$ for $\|h\|$ small enough.

Hence $\|Ah'\| \leq \|h\| + \epsilon\|h'\|$. Since we are in finite dimensions, and A is injective, we have that $x \mapsto \|Ax\|$ attains a minimum (nonzero) value on the sphere of unit vectors, call that C. Therefore for any h', we have $C\|h'\| \leq \|Ah'\|$, and we have $\|h\| \geq (C-\epsilon)\|h'\|$. In fact, this should still work for infinite dimensional Banach spaces by the Open Mapping Theorem, but that isn't really on topic.

We can also transform our inequality $\|h - Ah'\| \leq \epsilon\|h'\|$ by applying A^{-1}. Therefore for $\|h\|$ small enough, we have

$$\|h' - A^{-1}h\| \leq \|A^{-1}(h - Ah')\| \leq \|A^{-1}\|\|h - Ah'\| \leq \epsilon\|A^{-1}\|\|h'\| \leq \epsilon\|A^{-1}\|(C-\epsilon)\|h\|.$$

Which, since C and $\|A^{-1}\|$ are constants, implies that by reducing ϵ, we have for $\|h\|$ small enough, that $\|g(y+h) - g(y) - A^{-1}h\| = \|h' - A^{-1}h\| \leq \epsilon\|h\|$. Dividing by $\|h\|$, we have that g is differentiable at y with derivative $(D_x f)^{-1}$.

(6) (a) Multiplication of matrices is bilinear, so $D_{(X,Y)}f(A,B) = XB + AY$.

(b) The map $A \mapsto (A^T, A)$ is linear, so its derivative is itself, and the map we care about is the composite of this map with matrix multiplication, so our derivative is $D_X f(A) = X^T A + A^T X$.

(c) The determinant is multilinear in the columns of the input matrix, so if $f(A) = \det A$, then

$$D_{I_n} f(A) = \sum_{i=1}^{n} \det(e_1, \ldots, e_{i-1}, A_i, e_{i+1}, \ldots, e_n) = \sum_{i=1}^{n} A_{ii} = \operatorname{tr} A.$$

When B is invertible, we have that $\det(A) = \det(B)\det(B^{-1}A)$, and $\det(B^{-1}A)$ is the composite of det and left multiplication by B^{-1}, which I will denote $\lambda_{B^{-1}}$, which is linear

$$D_B f(H) = \det(B) D_{B^{-1}B} \det \circ D_B \lambda_{B^{-1}} H = \det(B)\operatorname{tr}(B^{-1}H).$$

In fact, since $\det(B)B^{-1}$ is the adjugate matrix of B, invertible matrices are dense in all matrices, and the determinant is a polynomial function, and thus continuously differentiable, we have that this formula implies that for an arbitrary B, $D_B f(H) = \operatorname{tr}((\operatorname{adj} B)H)$. (This observation comes from the wiki page)

This is called Jacobi's formula.