Review of Advanced Calculus
version date: 1/29/2022

Contents

1 Differentiability

Throughout sections 1–8 of these notes V and W are finite-dimensional vector spaces, of dimension at least 1 (unless stated otherwise), with zero-elements 0_V and 0_W respectively, and with norms $\| \, \|_V$ and $\| \, \|_W$ respectively. The subscripts on the zero-elements and norms will often be dropped when context makes clear which zero-element (that of V, W, or $\mathbf{R}$) or norm is intended.

The only reasons we exclude dimension zero by default are that (1) the limit in equation (1.1) below would not be defined if $\dim(V) = 0$, and (2) we don't want to have to say "assume $\dim(V) > 0$" or "assume $\dim(W) > 0$ whenever we want to introduce bases for these vector spaces. However, in the interests of thoroughness, we will sometimes make remarks for the zero-dimensional cases.

Definition 1.1 Let $U \subset V$ be an open set, let $F : U \to W$, and let $p \in U$. We say F is *differentiable at* p if F has a *good linear approximation* near p, i.e. if there exists a linear transformation $T : V \to W$ such that

$$\lim_{v \to 0_V} \frac{\|F(p+v) - F(p) - T(v)\|_W}{\|v\|_V} = 0; \tag{1.1}$$

equivalently, if for all $\epsilon > 0$ there exists $\delta > 0$ such that if $v \in V$ and $\|v\| < \delta$, then

$$\|F(p+v) - F(p) - T(v)\|_W \le \epsilon \|v\|_V. \tag{1.2}$$

We say that F *is differentiable* if F is differentiable at every point of U. ▲

More generally, if $U' \subset W$ is an open set, we define differentiability of $F : U \to U'$ (at a point or globally) the same way.

Note that (1.1) is equivalent to

$$\lim_{v \to 0_V} \frac{F(p+v) - F(p) - T(v)}{\|v\|_V} = 0_W \ . \tag{1.3}$$

Remark 1.2 The *good linear approximation* referred to in Definition 1.1 is the function $\tilde{F} : q \mapsto F(p) + T(q-p)$. This function is not linear in the sense of "linear transformation", but in the sense of "polynomial of degree at most 1": if we choose bases for V and W, then the component functions of $\tilde{F}$ (relative to the chosen basis of W) are polynomials of degree at most 1 in the coordinate functions of V (relative to the chosen basis of V). This is the only instance in these notes in which "linear function" will mean anything other than what "linear map" or "linear transformation" means in linear algebra. ▲

Claim 1.3 *Let U, F, p be as in Definition 1.1. Then there exists* at most one *linear transformation $T : V \to W$ such that* (1.1) *holds. In particular, if F is differentiable at p then the linear transformation T in equation* (1.1) *is unique.*

Proof: If linear transformations T_1 and T_2 are such that (1.1) holds, then for all $v \in V$

$$(T_2 - T_1)(v) = [F(p+v) - F(p) - T_1(v)] - [F(p+v) - F(p) - T_2(v)]\,.$$

Therefore, for all nonzero $v \in V$,

$$\frac{F(p+v) - F(p) - T_1(v)}{\|v\|} - \frac{F(p+v) - F(p) - T_2(v)}{\|v\|} = \frac{(T_2 - T_1)(v)}{\|v\|} = (T_2 - T_1)\left(\frac{v}{\|v\|}\right),$$

since the linearity of T_1 and T_2 implies that $T_2 - T_1$ is linear as well. Letting $v \to 0$, using the hypothesis that (1.3) is satisfied both with $T = T_1$ and with $T = T_2$, we deduce that

$$0_W = \lim_{v \to 0} (T_2 - T_1)\left(\frac{v}{\|v\|}\right).$$

Hence, by the "Substitution Lemma for limits", for every unit vector $e \in V$ we have

$$0 = \lim_{t \to 0^+} (T_2 - T_1)\left(\frac{te}{\|te\|}\right) = (T_2 - T_1)(e).$$

Since every $v \in V$ is a multiple of some unit vector, linearity implies that $(T_2 - T_1)(v) = 0$ for all $v \in V$, hence that $T_2 = T_1$. ■

Remark 1.4 The limit-statement (1.1) is *equivalent* to (1.2) because we have assumed $\dim(V) > 0$. If $\dim(V) = 0$ then (1.1) does not make sense, but (1.2) still does. Thus we could incorporate the case $\dim(V) = 0$ into Definition 1.1 by simply getting rid of equation (1.1), and using only the formulation involving (1.2) instead. The upshot is that if $\dim(V) = 0$ or $\dim(W) = 0$, then all maps from $f : V \to W$ are differentiable, with $Df|_p : V \to W$ being the zero linear transformation for all $p \in V$. Whenever $\dim(V) = 0$ or $\dim(W) = 0$, the zero linear transformation $V \to W$ is the *unique* linear transformation from V to W, so the uniqeness asserted in Claim 1.3 holds in these cases as well.

Definition 1.5 Suppose V, W, U, F and p are as in Definition 1.1. Assume F is differentiable at p. Then unique linear map $T : V \to W$ satisfying (1.1) is called the *derivative of F at p*, denoted $(DF)_p$ or $DF|p$ in these notes .

If $\dim(V) = 0$ or $\dim(W) = 0$, then for all $p \in V$ we define $(DF)_p$ to be the zero linear transformation from V to W. The condition involving (1.2) in Definition 1.1 holds with $T = DF|_p = 0$ in these cases.

▲

Claim 1.6 *Let $V = W = \mathbf{R}$, and let all other notation be as in Definition 1.1. Then F is differentiable at p (as defined in Definition 1.1) if and only if $F'(p)$ (as defined in Calculus 1) exists, and that in the differentiable case, the derivative $DF|_p$ is the linear map "multiplication by $F'(p)$" from $\mathbf{R}$ to $\mathbf{R}$.*

Proof: Exercise. ■

Remark 1.7 (Different meaning of "derivative") As Claim 1.6 shows, the derivative of F at p, as defined above, does *not* reduce to the "Calc 1 derivative" in the case $V = W = \mathbf{R}$. The latter is a *number* $F'(p)$, not a linear transformation. However, there is a natural one-to-one correspondence between real numbers and linear transformations $\mathbf{R} \to \mathbf{R}$:

$$\begin{aligned} \mathbf{R} &\longleftrightarrow \mathrm{Hom}(\mathbf{R}, \mathbf{R}), \\ c &\longleftrightarrow \text{the linear map } x \mapsto cx, \end{aligned}$$

i.e. the map "multiplication by c". Thus, when $V = W = \mathbf{R}$, either of $DF|_p$ and $F'(p)$ can be recovered from the other. Because Definition 1.5 does not reduce to the familiar meaning of "derivative" in the case $V = W = \mathbf{R}$, some authors prefer to call the linear transformation T in (1.1) the *differential* of F at p. ▲

Convention and notation for natural numbers. In these notes we use the convention that "the natural numbers start at 1;" i.e. that "natural number" means *(strictly) positive* integer. We let $\mathbf{N}$ denote the set of natural numbers.

Remark 1.8 Recall that for any fixed $n \in \mathbf{N}$, all norms on $\mathbf{R}^n$ are equivalent. It follows from this that the property of "differentiability at p" and (in the differentiable case) the linear transformation $DF|_p$ are independent of which norms are used on V and W. ▲

Example 1.9 If $F : V \to W$ is a linear map then at each point $p \in U$, $DF\big|_p = F$ ("a linear map is its own derivative [at each point]"). This follows from the uniqueness statement in Claim 1.3 and the fact if F is linear, then for all nonzero $v \in V$ we have

$$\frac{F(p+v) - F(p) - F(v)}{\|v\|} = 0.$$

▲

Exercise 1.10 Let $U \subset V$ be open, let $p \in U$, and let $F_1, F_2, \ldots F_k : U \to W$ be functions that are differentiable at p. Let $F = F_1 + F_2 + \cdots + F_k$. Show that F is differentiable at p, and that $DF|_p = DF_1|_p + DF_2|_p + \cdots + DF_k|_p$.

Exercise 1.11 Suppose $U_1 \subset U \subset V$, where both U_1 and U are open, let $p \in U_1$ and suppose $F : U \to V$ is differentiable at p. Then $F|_{U_1}$ (the restriction of F to the domain U_1) is differentiable at p, and $D(F|_{U_1})|_p = DF|_p$.

Definition 1.12 (a) Let $\mathrm{Hom}(V, W)$ denote the space of linear maps $V \to W$ (a vector space of dimension $(\mathrm{dim}V)(\mathrm{dim}W)$). For $T \in \mathrm{Hom}(V, W)$, the *operator norm* of T, denoted $\|T\|_{\mathrm{op}}$, is defined by

$$\|T\|_{\mathrm{op}} = \sup_{\{v \in V:\ \|v\|_V = 1\}} \|T(v)\|_W. \tag{1.4}$$

Note that the value of $\|T\|_{\mathrm{op}}$ may depend on the norms chosen on V and W. Notation such as "$\|T\|_{\mathrm{op}}^{(\|\ \|_V, \|\ \|_W)}$" would be more precise, but for reasons that should be all too apparent, we avoid using it.

Recall that every (a) linear transformation from one finite-dimensional normed vector space to another is continuous, (b) in any finite-dimensional normed vector space V, the unit sphere $S(V) := \{v \in V : \|v\| = 1\}$ is compact, and (c) any restriction of a continuous function is continuous. Thus, in the setting of (1.4), the function $S(V) \to \mathbf{R}$ defined by $v \mapsto \|T(v)\|$ is continuous (a composition of continuous functions), so the compactness of $S(V)$ implies that this function achieves a maximum value. Thus the supremum in (1.4) is finite (and is actually achieved; "sup" could be replaced by "max").

Claim 1.13 *(a) As the name and notation suggest, the operator norm is indeed a norm on the vector space* $\mathrm{Hom}(V, W)$.

(b) Let $T \in \mathrm{Hom}(V, W)$. Then $\|T\|_{\mathrm{op}} = \sup_{\{v \in V:\ v \neq 0\}} \left\| T\left(\frac{v}{\|v\|}\right) \right\|$, and for all $v \in V$ we have

$$\|T(v)\| \leq \|T\|_{\mathrm{op}} \|v\|. \tag{1.5}$$

(c) *Let* $(Z, \| \ \|_Z)$ *be a third finite-dimensional normed vector space. Then the corresponding operator norms on* $\mathrm{Hom}(V,W), \mathrm{Hom}(W,Z)$ *and* $\mathrm{Hom}(V,Z)$ *are related in the following "sub-multiplicative" way:*

$$\|S \circ T\|_{\mathrm{op}} \leq \|S\|_{\mathrm{op}} \, \|T\|_{\mathrm{op}} \quad \textit{for all } T \in \mathrm{Hom}(V,W), S \in \mathrm{Hom}(W,Z). \tag{1.6}$$

Proof: Exercise. ■

Proposition 1.14 (Differentiability implies continuity) *Notation as in Definition 1.1. If* F *is differentiable at* p*, then* F *is continuous at* p*.*

Proof: Assume F is differentiable at p, and let $T = DF|_p$. Let $\delta > 0$ be such that for all $v \in B_\delta(0_V)$ we have $\|F(p+v) - F(p) - T(v)\| \leq \|v\|$; such δ exists by Definitions 1.1 and 1.5. Then for all $q \in B_\delta(p)$, writing $v = q - p$ we have $\|v\| < \delta$, so

$$\begin{aligned} \|F(q) - F(p)\| = \|F(p+v) - F(p)\| &\leq \|F(p+v) - F(p) - T(v)\| + \|T(v)\| \\ &\leq \|v\| + \|T\|_{\mathrm{op}}\|v\| \\ &= (1 + \|T\|_{\mathrm{op}})\|q - p\|. \end{aligned}$$

Thus F is Lipschitz at p, hence continuous at p. ■

Remark 1.15 In Definition 1.1, the condition that V be finite-dimensional can be relaxed, at the cost of requiring that the linear transformation T be *bounded* (in the sense of linear transformations; see the handout "Some facts about normed vector spaces"). If $\dim(V)$ is finite then every linear transformation from V to W is bounded, regardless of whether $\dim(W)$ is finite. Conversely, if $\dim(V) = \infty$ then there always exist unbounded linear transformations from V to W (unless $\dim(W) = 0$, which we're continuing to assume is not the case).

With "linear transformation" replaced by "bounded linear transformation" in Definition 1.1, the proof of Proposition 1.14 carries through verbatim to the case in which $\dim(V) = \infty$. ▲

Exercise 1.16 Let W_1, W_2 be finite-dimensional vector spaces of positive dimension. Let $U \subset V$ be open, $p \in U$, and $g_i : U \to W_i$ differentiable at p for $i = 1, 2$. Define $f : U \to W_1 \oplus W_2$ by $f(q) = (g_1(q), g_2(q))$. Show that f is differentiable at p and compute $Df|_p(v)$ for arbitrary $v \in V$.

Note: one general approach to a problem of the form "show that a function F is differentiable at point q, and compute the derivative $DF|_q$" is to compute all the directional derivatives $(D_vF)(q)$. If this expression is not linear in v, then F is not differentiable at

q (and you were instructed to show something that was false). If "$(D_vF)(q)$" *is* linear in v, then the linear transformation T defined by $T(v) = (D_vF)(q)$ is the *only candidate* for $DF|_q$. You can then try to show that F is differentiable at q either by plugging this T into the definition of "differentiable at q" and showing that the relevant limit is zero, or by showing that, for all fixed v, the map $\tilde{q} \mapsto (D_vF)(q)$ is continuous in $\tilde{q}$ (in which case, automatically, F is not merely differentiable at q, but continuously differentiable at q). The former approach is the one to use in this problem, since nothing in the hypotheses implies continuity of the directional derivatives. ▲

Remark 1.17 The difference between the direct sum $W_1 \oplus W_2$ and the Cartesian product $W_1 \times W_2$ is that only $W_1 \oplus W_2$ is a vector space (whose underlyng point-set is $W_1 \times W_2$, and whose operations and zero element are defined componentwise). We can speak of *linear* maps with domain $W_1 \oplus W_2$, and *bi*linear maps with domain $W_1 \times W_2$, but not vice-versa. In the exercise above, we needed to use $W_1 \oplus W_2$ because we have not defined "differentiable map" for a function whose codomain is not a vector space (or subset of a vector space).

Exercise 1.18 Let $n, m, k \in \mathbf{N}$. Define $\mu : M_{m\times n}(\mathbf{R}) \oplus M_{n\times k}(\mathbf{R})$ by $\mu(A, B) = AB$ (matrix multiplication). Show that μ is differentiable, and compute its derivative.

2 Chain Rule Theorem

We now add a third finite-dimensional normed vector space $(Z, \| \ \|_Z)$ to the picture so that we can talk about compositions of differentiable functions, and state and prove the Chain Rule Theorem for functions between (subsets of) finite-dimensional vector spaces.

There are several ways of stating the Chain Rule. One way is better than all the others:

The derivative of a composition is the composition of the derivatives.

Some precision is sacrificed in this statement in order to emphasize the elegance and simplicity of the principle. The precise statement is equation (2.1) in the Chain Rule Theorem below.

Theorem 2.1 (Chain Rule Theorem) *Let V, W, Z be finite-dimensional vector spaces. Let $U_1 \subset V$, $U_2 \subset W$ be open sets, let $F : U_1 \to U_2, G : U_2 \to Z$ be functions, and let $p \in U_1$. Assume that F is differentiable at p and that G is differentiable at $F(p)$. Then the composition $G \circ F$ is differentiable at p, and*

$$D(G \circ F)\big|_p = Dg\big|_{F(p)} \circ DF\big|_p. \tag{2.1}$$

Remark 2.2 The *idea* behind the proof below is the portion from (2.8) through (2.9): starting with an arbitrary v, and writing "$F(p+v) - F(p)$" in place of w, we start with the left-hand side of (2.8) and use the triangle inequality and the linearity of S to attain (2.9). Then we work backwards to see what δ's are needed in order to get from (2.9) to get "(constant $\times\ \epsilon)\|v\|$" in (2.10). No cleverness or cooked-up functions are involved (as in most textbook-proofs of Theorem 2.1, for $V = W = Z = \mathbf{R}$ or for $V = \mathbf{R}^n, W = \mathbf{R}^m, Z = \mathbf{R}^k$); the strategy is completely natural. It is the *more advanced concept of differentiability* that makes this strategy so easy to find; armed with Definitions 1.1 and 1.5, and Definition 1.12 (of which the inequality (1.5) is a simple corollary), this strategy is the first thing we think of. This same natural proof works perfectly well when restricted to the case $V = W = Z = \mathbf{R}$.

Furthermore, if we remove the finite-dimensionality assumption on V and W, and modify Definition 1.1 by inserting the word "bounded" accordingly (see Remark 1.15), then our proof of the Chain Rule still works, *word for word*.

Proof of Theorem 2.1. Let $T = DF\big|_p, q = F(p)$, and $S = DG\big|_q$. Let $\epsilon > 0$. Let $\delta_1, \delta_2 > 0$ be such that for all $w \in W$ and $v \in V$,

$$\text{if } \|w\| < \delta_1 \text{ then } q + w \in U_2 \text{ and } \|G(q+w) - G(q) - S(w)\| \leq \epsilon\|w\| \tag{2.2}$$

and

$$\text{if } \|v\| < \delta_2 \text{ then } p + v \in U_1 \text{ and } \|F(p+v) - F(p) - T(v)\| \leq \min\{\epsilon, 1\}\|v\|; \tag{2.3}$$

such δ_1 and δ_2 exist by Definitions 1.1 and 1.5.

Let $\delta_3 = \min\{\delta_2, \frac{\delta_1}{1+\|T\|_{\text{op}}}\}$, and let $v \in V$ be any element with $\|v\| < \delta_3$. Then $\|v\| < \delta_2$, so by (2.3),

$$\|F(p+v) - F(p) - T(v)\| \leq \epsilon\|v\| \tag{2.4}$$

and, by the same argument as in the proof of Proposition 1.14,

$$\begin{aligned} \|F(p+v) - F(p)\| &\leq (1 + \|T\|_{\text{op}})\|v\| && (2.5) \\ &< (1 + \|T\|_{\text{op}})\delta_3 \leq \delta_1. && (2.6) \end{aligned}$$

Let $w = F(p+v) - F(p)$; thus $F(p+v) = F(p) + w = q + w$. From (2.6) we have $\|w\| < \delta_1$, so, by (2.2) and (2.5),

$$\|G(q+w) - G(q) - S(w)\| \leq \epsilon\|w\| \leq \epsilon(1 + \|T\|_{\text{op}})\|v\|. \tag{2.7}$$

Using (2.7), (2.4), and the linearity of S, we therefore have

$$\begin{aligned} \|(G \circ F)(p+v) - (G \circ F)(p) - (S \circ T)(v)\| &= \|G(q+w) - G(q) - S(T(v))\| && (2.8) \\ &\leq \|G(q+w) - G(q) - S(w)\| + \|S(w) - S(T(v))\| \\ &= \|G(q+w) - G(q) - S(w)\| + \|S(w - T(v))\| \\ &= \|G(q+w) - G(q) - S(w)\| + \|S\|_{\text{op}} \|F(p+v) - F(p) - T(v)\| && (2.9) \\ &\leq \epsilon(1 + \|T\|_{\text{op}})\|v\| + \|S\|_{\text{op}}\, \epsilon\|v\| \\ &= [(1 + \|T\|_{\text{op}} + \|S\|_{\text{op}})\, \epsilon]\|v\|. && (2.10) \end{aligned}$$

Since ϵ was arbitrary, it follows that $G \circ F$ is differentiable at p, with derivative $S \circ T$.
■

Exercise 2.3 Let $n, m, k \in \mathbf{N}$. Let $U \subset V$ be open, let $g : U \to M_{m\times n}(\mathbf{R})$ and $h : U \to M_{n\times k}(\mathbf{R})$ differentiable and define $f : U \to M_{m\times k}(\mathbf{R})$ by $f(p) = g(p)h(p)$. Note that $f = \mu \circ j$, where μ is the map in Exercise 1.18 and $\phi : U \to M_{m\times n}(\mathbf{R}) \oplus \; M_{n\times k}(\mathbf{R})$ is defined by $\phi(p) = (g(p), h(p))$, a map of the form in Exercise 1.16. Use the Chain Rule Theorem to prove that f is differentiable, and (using directional derivatives) express the derivative of f in terms of the derivatives of g and h.

If your answer is correct, then in the case $n = m = k = 1$, you should find with the aid of Claim 1.6 that you've recovered the "product rule" from Calculus 1. Thus, *the Calculus-1 product rule is a corollary of the (multivariable) chain rule.* Even though we learn the product rule in Calc 1 before we learn the chain rule, the "fully understood" chain rule (2.1) is more fundamental—but that's not something that can be explained or appreciated at the level of Calc 1.

3 Directional derivatives

Definition 3.1 Let U be an open subset of V, let $p \in U$ and let $F : U \to W$ (with no differentiability of F assumed). For $v \in V$, the (generalized) *directional derivative of F at p in direction v* is

$$(D_vF)(p) := (D_vF)_p := \frac{d}{dt}F(p+tv)\bigg|_{t=0} := \lim_{t\to 0}\frac{F(p+tv)-F(p)}{t}$$

if this limit exists. Note that $(D_vF)(p)$, when it exists, is an element of W.▲

We have inserted "(generalized)" since, unlike in Calculus 3, there is no requirement that v be a unit vector; v can even be the zero vector.

Example 3.2 In Definition 3.1, consider the case $V = \mathbf{R}^n, W = \mathbf{R}$; thus F is a real-valued function on an open set in $\mathbf{R}^n$. Let $\{e_j\}_{j=1}^n$ be the standard basis of $\mathbf{R}^n$, and let $\{x^j\}_{j=1}^n$ be the standard coordinate functions on $\mathbf{R}^n$. Then $(D_{e_j}F)(p) = \frac{\partial F}{\partial x^j}(p)$ (i.e. if one side of this equation exists, then so does the other, and the two sides are equal). Thus, partial derivatives are special cases of directional derivatives.

More generally, even if W is a general finite-dimensional vector space, it is common to make the definition

$$\frac{\partial F}{\partial x^j}(p) := (D_{e_j}F)(p) \tag{3.1}$$

(a vector-valued partial derivative). As is easily checked, if $W = \mathbf{R}^m$ and $\{f^i\}_{i=1}^m$ are the component functions of of F with respect to the standard basis on $\mathbf{R}^m$, then

$$\frac{\partial F}{\partial x^j} = \begin{pmatrix} \frac{\partial f^1}{\partial x^j} \\ \vdots \\ \frac{\partial f^m}{\partial x^j} \end{pmatrix}. \tag{3.2}$$

▲

Remark 3.3 Let F and p be given. Trivially, $(D_{0_V}F)(p) = 0_W$, but the limit in Definition 3.1 may or may not exist for a given nonzero v, and may exist for some nonzero v's but not others. However, if v is a vector for which $(D_vF)(p)$ exists, then it is easily shown that $(D_wF)pP)$ exists for all multiples w of v, and that we have the following homogeneity property:

$$(D_{\lambda v}F)(p) = \lambda(D_vF)(p) \quad \text{for all } \lambda \in \mathbf{R}. \tag{3.3}$$

▲

Proposition 3.4 *If F is differentiable at p then all directional derivatives of F exist at p and*

$$\underbrace{(D_vF)(p)}_{\substack{\textit{directional derivative} \text{ of} \\ F\text{, at } p\text{, in direction } v}} = \underbrace{DF\big|_p(v)}_{\substack{\textit{derivative} \text{ of } F \text{ at } p, \\ \text{applied to } v}}.$$

Proof: Suppose F is differentiable at p and let $T = DF\big|_p$. Then if $v \neq 0$,

$$\lim_{t\to 0} \frac{F(p+tv) - F(p)}{t} = \lim_{t\to 0}\left(\left[\|v\| \underbrace{\frac{F(p+tv) - F(p) - T(tv)}{t\|v\|}}_{\to 0 \text{ since } F \text{ is differentiable at } p}\right] + T(v)\right) = T(v)$$

■

Corollary 3.5 *Let $U \subset \mathbf{R}^n$ be open, $p \in U$, let $f : U \to \mathbf{R}$ be a function that is differentiable at p, let $\{x^j\}_{j=1}^n$ be the standard coordinates on $\mathbf{R}^n$, and let* $v = \begin{pmatrix} v^1 \\ \vdots \\ v^n \end{pmatrix} \in$

$\mathbf{R}^n$. *Then*

$$
\begin{aligned}
Df|_p(v) &= \sum_{j=1}^{n} \frac{\partial f}{\partial x^j}(p)\, v^j &(3.4)\\
&= \underbrace{\left(\frac{\partial f}{\partial x^1}(p),\ \ldots,\ \frac{\partial f}{\partial x^n}(p)\right)}_{1\times n \text{ matrix}} \underbrace{\begin{pmatrix} v^1 \\ \vdots \\ v^n \end{pmatrix}}_{n\times 1 \text{ matrix}}, &(3.5)
\end{aligned}
$$

where matrix multiplication is used on the right-hand side of (3.5)*, and* 1×1 *matrices are identified with real numbers.*

Proof: Let $\{e_j\}_{j=1}^n$ be the standard basis of $\mathbf{R}^n$. Then $v = \sum_{j=1}^n v^j e_j$. Using the linearity of $Df|_p$, Proposition 3.4, and Example 3.2, we therefore have

$$
Df|_p(v) = \sum_{j=1}^{n} v^j\ Df|_p(e_j) = \sum_{j=1}^{n} v^j\ (D_{e_j} f)(p) = \sum_{j=1}^{n} \frac{\partial f}{\partial x^j}(p)\, v^j,
$$

yielding (3.4). ■

Convention for the remainder of these notes. Whenever we introduce a function with a phrase like "Let $F : (U \subset V) \to W\ \ldots$", we assume that U is an open subset of V unless we specify otherwise.

4 The case $V = \mathbf{R}^n, W = \mathbf{R}^m$

We now specialize to the concrete case $V = \mathbf{R}^n, W = \mathbf{R}^m$. For purposes of matrix operations that will arise later, we treat elements of $\mathbf{R}^n$ and $\mathbf{R}^m$ as column vectors. When this is inconvenient typographically, so we will write column vectors as transposes of row vectors. If $(a^1, \ldots, a^n)^t$ is in the domain of a function f defined on a subset of $\mathbf{R}^n$, we write simply $f(a^1, \ldots, a^n)$ rather than $f((a^1, \ldots, a^n)^t)$ or $f\left(\begin{bmatrix} a^1 \\ \vdots \\ a^n \end{bmatrix}\right)$.

Throughout this section, unless stated otherwise, $\{e_i\}_{i=1}^n$ and $\{e_i'\}_{i=1}^m$ denote the standard bases of $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and $\{x^i\}_{i=1}^n$ and $\{y^i\}_{i=1}^m$ denote the standard coordinate functions on $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. For $1 \le i \le m$, define $\iota_i : \mathbf{R} \to \mathbf{R}^m$ by

$$
\iota_i(s) = s e_i' = (0, \ldots, 0, s, 0, \ldots, 0)^t,
$$

where the s is in the i^{th} slot.

Observe that each of the functions x^i, y^i, ι_i defined above is a linear map.

Definition 4.1 Let $U \subset \mathbf{R}^n$ be open, and let $F : U \to \mathbf{R}^m$ be a function. For each $i \in \{1, \ldots n\}$ let $f^i = y^i \circ F : U \to \mathbf{R}$ (the i^{th} *component function of* F [with respect to the standard basis of $\mathbf{R}^m$]). Then, as is easily checked,

$$F = \sum_{i=1}^{n} \iota_i \circ f^i \ . \tag{4.1}$$

We may write (4.1) in the more familiar form

$$F = \sum_{j=1}^{m} f^i \, e'_i = \begin{pmatrix} f^1 \\ \vdots \\ f^m \end{pmatrix},$$

with the understanding that this means $F(p) = \sum_{i=1}^{m} f^i(p)\, e'_i$ for all $p \in U$. At any point p for which all the partial derivatives $\frac{\partial f^i}{\partial x^j}(p)$ exist ($1 \le i \le n, 1 \le j \le m$), we define the *Jacobian matrix of* F *at* p to be the matrix whose $(ij)^{\text{th}}$ entry is $\frac{\partial f^i}{\partial x^j}(p)$:

$$J_F(p) = \begin{pmatrix} \frac{\partial f^1}{\partial x^1}(p) & \frac{\partial f^1}{\partial x^2}(p) & \cdots & \frac{\partial f^1}{\partial x^n}(p) \\ \frac{\partial f^2}{\partial x^1}(p) & \frac{\partial f^2}{\partial x^2}(p) & \cdots & \frac{\partial f^2}{\partial x^n}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f^m}{\partial x^1}(p) & \frac{\partial f^m}{\partial x^2}(p) & \cdots & \frac{\partial f^m}{\partial x^n}(p) \end{pmatrix}.$$

Example 4.2 Consider the case $n = 1$ in Definition 4.1. If the component functions $f^1, \ldots, f^m$ of F are differentiable at $t \in U$, then

$$J_F(t) = \begin{pmatrix} (f^1)'(t) \\ \vdots \\ (f^m)'(t) \end{pmatrix} = \lim_{h \to 0} \frac{F(t+h) - F(t)}{h} =: F'(t).$$

▲

Proposition 4.3 *Let* $U, F, \{f^i\}_{i=1}^m$*, and* p *be as in Definition 4.1. Then* F *is differentiable at* p *if and only if each component function* f^i *is differentiable at* p*,* $1 \le i \le m$*. In the differentiable case,*

$$DF|_p = \sum_{i=1}^{m} \iota_i \circ Df^i|_p \ ; \tag{4.2}$$

equivalently,

$$DF|_p(v) = J_F(p)v \quad \text{for all } v \in V, \tag{4.3}$$

where matrix-multiplication is implicit on the right-hand side of this equation.

Thus, Proposition 4.3 yields the following important relation between derivatives and Jacobians:

> *If $F : (U \subset \mathbf{R}^n) \to \mathbf{R}^m$ is differentiable at $p \in U$, then the derivative of F at p is the linear map $\mathbf{R}^n \to \mathbf{R}^m$ given by multiplication by the Jacobian matrix $J_F(p)$.*

Said another way,

> *If $F : (U \subset \mathbf{R}^n) \to \mathbf{R}^m$ is differentiable at $p \in U$, then the Jacobian matrix $J_F(p)$ is the matrix of the linear transformation $DF|_p : \mathbf{R}^n \to \mathbf{R}^m$ with respect to the standard bases of $\mathbf{R}^n$ and $\mathbf{R}^m$.*

For the case $n = m = 1$, we have already seen this fact in Remark 1.7. In this case, the Jacobian matrix $J_F(p)$ is a 1×1 matrix whose sole entry is $F'(p)$. The linear map $x \mapsto F'(p)x$ is exactly multiplication by this 1×1 matrix. Thus, the "Calc 1" derivative of a function $F : (U \subset \mathbf{R}) \to \mathbf{R}$ at p is the 1×1 Jacobian $J_F(p)$.

Proof of Proposition 4.3: First suppose that f^i is differentiable at p, $1 \leq i \leq m$. Let $i \in \{1, \ldots, m\}$. Since ι_i is linear, Example 1.9 implies that ι_i is differentiable and that $D\iota_i|_{f^i(p)} = \iota_i$. Hence, by the Chain Rule Theorem, $\iota_i \circ f^i$ is differentiable at p, and

$$D(\iota_i \circ f^i)|_p = D\iota_i|_{f^i(p)} \circ Df^i|_p = \iota_i \circ Df^i|_p. \tag{4.4}$$

Since (4.4) holds for each i, and $F = \sum_i \iota_i \circ f^i$, Exercise 1.10 implies the equality (4.2).

Conversely, suppose that F is differentiable at p, and let $i \in \{1, \ldots, m\}$. Then $f^i = y^i \circ F$. Since y^i is linear, the same Chain Rule argument as above shows that f^i is differentiable at p.

This establishes the "if and only if" statement in the Proposition, as well as the equality (4.2) in the differentiable case. For the equivalence between (4.2) and (4.3) (when $DF|_p$ exists), let $v = (v_1, \ldots, v_n)^t \in \mathbf{R}^n$. Then, by Corollary 3.5 (applied to the component functions f^i) and the definition of the maps ι_i,

$$\begin{aligned} DF|_p(v) &= \sum_{i=1}^m \iota_i \circ Df^i|_p(v) \\ &= \begin{pmatrix} \sum_{j=1}^n \frac{\partial f^1}{\partial x^j}(p)\, v_j \\ \vdots \\ \sum_{j=1}^n \frac{\partial f^m}{\partial x^j}(p)\, v_j \end{pmatrix} \\ &= J_F(p)v. \end{aligned}$$

Since v was arbitrary, (4.2) and (4.3) are equivalent. ■

With notation and hypotheses as in Theorem 2.1, let us now revisit the Chain Rule for the special case $V = \mathbf{R}^n$, $W = \mathbf{R}^m$, and $Z = \mathbf{R}^k$. From Proposition 4.3, for all $w \in \mathbf{R}^m$ and $v \in \mathbf{R}^n$ we have

$$\begin{aligned} Dg\big|_{f(p)}(w) &= \underbrace{J_g(f(p))}_{k\times m}\underbrace{w}_{\in\mathbf{R}^m} \in \mathbf{R}^k, \\ Df\big|_p(v) &= \underbrace{J_f(p)}_{m\times n}\underbrace{v}_{\in\mathbf{R}^n} \in \mathbf{R}^m, \\ \text{and}\quad D(g\circ f)\big|_p(v) &= \underbrace{J_{g\circ f}(p)}_{k\times n}\underbrace{v}_{\in\mathbf{R}^n} \in \mathbf{R}^k. \end{aligned}$$

Thus Theorem 2.1 implies

$$\underbrace{J_{g\circ f}(p)}_{k\times n} v = D(g\circ f)\big|_p(v) = Dg\big|_{f(p)}\left(Df\big|_p(v)\right) = \underbrace{J_g(f(p))}_{k\times m}\underbrace{J_f(p)}_{m\times n}\underbrace{v}_{\in\mathbf{R}^n} \in \mathbf{R}^k \quad \text{for all } v \in \mathbf{R}^n.$$

Therefore

$$J_{g\circ f}(p) = J_g(f(p))J_f(p), \tag{4.5}$$

i.e. "the Jacobian of a composition is the product of the Jacobians." This is the *second*-best statement of the Chain Rule.

Exercise 4.4 Check that (4.5) is exactly the chain rule you learned in Calculus 3, simply written in matrix notation.

In case your memory needs refreshing, the chain rule you learned in Calculus 3 should say the following, modulo notation (often abused in Calc 3): if $\{x^i\}_{i=1}^n$, $\{y^i\}_{i=1}^m$, and $\{z^i\}_{i=1}^k$ denote the standard coordinates on $\mathbf{R}^n \supseteq \text{domain}(f)$, $\mathbf{R}^m \supseteq \text{codomain}(f) \subseteq \text{domain}(g)$, and $\mathbf{R}^k \supseteq \text{codomain}(g)$ respectively, and we define $h = g\circ f$, $f^i = y^i \circ f$ for $1 \le i \le m$, $g^i = z^i \circ g$ for $1 \le i \le k$, and $h^i = z^i \circ h$ for $1 \le i \le k$, then

$$\frac{\partial h^i}{\partial x^j}(x) = \sum_{l=1}^m \frac{\partial g^i}{\partial y^l}(f(x))\,\frac{\partial f^l}{\partial x^j}(x), \quad 1 \le i \le k, \quad 1 \le j \le n. \tag{4.6}$$

(In equation (4.6) (i) some other Calc 3 notations for $\frac{\partial h^i}{\partial x^j}$ are "$\frac{\partial z^i}{\partial x^j}$" and the [literally *wrong*] "$\frac{\partial g^i}{\partial x^j}$", (ii) some other Calc 3 notations for $\frac{\partial g^i}{\partial y^l}(f(x))$ are "$\frac{\partial g^i}{\partial y^l}(y(x))$" and "$\frac{\partial z^i}{\partial y^l}(y(x))$", and (iii) another Calc 3 notation for "$\frac{\partial f^l}{\partial x^j}$" is $\frac{\partial y^l}{\partial x^j}$.) Equivalently, in terms of vector-valued partial derivatives,

$$\frac{\partial h}{\partial x^j}(x) = \sum_{l=1}^m \frac{\partial g}{\partial y^l}(f(x))\,\frac{\partial f^l}{\partial x^j}(x).$$

Exercise 4.5 Let $I \subset \mathbf{R}$, $U \subset \mathbf{R}^m$ be open, let $t_0 \in I$, $q \in U$, and suppose that $F : U \to \mathbf{R}^k$ is differentiable. Let $v \in \mathbf{R}^m$, and suppose that $\gamma : I \to \mathbf{R}^m$ is a differentiable function for which $\gamma(t_0) = q$ and $\gamma'(t_0) = J_\gamma(t_0) = v$ (see Example 4.2). Show that

$$\frac{d}{dt} F(\gamma(t))\Big|_{t=t_0} = J_F(q)v = DF|_q(v) = (D_v F)(q).$$

Exercise 4.6 Revisit the second paragraph of Exercise 2.3, where you saw how to derive the Calc-1 product rule from the general chain rule 2.1. Redo this derivation using using equation (4.5) instead of Claim 1.6; the Jacobian matrices J_μ and J_ϕ are, at each point, matrices of size 1×2 and 2×1, respectively.

Remark 4.7 (Operator norm on matrices) For $A \in M_{m\times n}(\mathbf{R})$, define $L_A : \mathbf{R}^n \to \mathbf{R}^m$ by $L_A(v) = Av$ (where we view v as a column vector); i.e. L_A is the linear transformation $\mathbf{R}^n \to \mathbf{R}^m$ whose matrix with respect to the standard bases is A. The map $M_{m\times n}(\mathbf{R}) \to \mathrm{Hom}(\mathbf{R}^n, \mathbf{R}^m)$ defined by $A \mapsto L_A$ is an isomorphism $M_{m\times n}(\mathbf{R}) \to \mathrm{Hom}(\mathbf{R}^n, \mathbf{R}^m)$. We use this isomorphism to define operator norms of *matrices*: given norms $\| \ \|, \| \ \|'$ on $\mathbf{R}^n, \mathbf{R}^m$, respectively, we define the corresponding operator-norm function on $M_{m\times n}(\mathbf{R})$ by

$$\|A\|_{\mathrm{op}} := \|L_A\|_{\mathrm{op}} \ .$$

(When $m = n$, it is implicit that we take $\| \ \|' = \| \ \|$ unless otherwise specified.) Thus, for all $A \in M_{m\times n}(\mathbf{R})$ and $v \in \mathbf{R}^n$,

$$\begin{aligned} \|A\|_{\mathrm{op}} &= \sup\{\|Av\|' : \|v\| = 1\} \quad \text{(from (1.4)), and} \\ \|Av\|' &\leq \|A\|_{\mathrm{op}}\, \|v\| \quad \text{(from (1.5)).} \end{aligned}$$

Furthermore, given also a norm $\|\cdot\|''$ on $\mathbf{R}^k$, the induced operator norms on $M_{k\times m}(\mathbf{R})$, $M_{m\times n}(\mathbf{R})$, and $M_{k\times n}(\mathbf{R})$ satisfy

$$\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}\, \|B\|_{\mathrm{op}} \quad \text{for all } A \in M_{k\times m}(\mathbf{R}), B \in M_{m\times n}(\mathbf{R}) \text{ (from (1.6)).}$$

Exercise 4.8 For $n \in \mathbf{N}$ let $GL(n, \mathbf{R}) \subset M_{n\times n}(\mathbf{R})$ be the set of invertible $n \times n$ matrices. For $n > 1$ there is no general formula expressing $\|A^{-1}\|_{\mathrm{op}}$ in terms of $\|A\|_{\mathrm{op}}$; e.g. it is *not* generally true that $\|A^{-1}\|_{\mathrm{op}} = 1/\|A\|_{\mathrm{op}}$. For example, if A is the 2×2 diagonal matrix with diagonal entries 2 and $1/2$, then for the operator norm determined by the Euclidean norm on $\mathbf{R}^2$, we have $\|A\|_{\mathrm{op}} = \|A^{-1}\|_{\mathrm{op}} = 2$.

But for any n, and any underlying norm on $\mathbf{R}^n$, there stilll some useful general inequalities involving $\|A^{-1}\|_{\mathrm{op}}$:

(a) Prove that $\|A^{-1}\|_{\mathrm{op}} \geq 1/\|A\|_{\mathrm{op}}$.

(b) Let $A \in GL(n, \mathbf{R})$. Since the unit sphere in $\mathbf{R}^n$ is compact, and $\|\cdot\| : \mathbf{R}^n \to \mathbf{R}$ is continuous, for any $A \in M_{n\times n}(\mathbf{R})$ the function $v \mapsto \|Av\|$ achieves a minimum value c, which must be strictly positive since A is invertible. Prove that $\|A^{-1}\|_{\mathrm{op}} \leq 1/c$.

(c) Fix $n \in \mathbf{N}$ and let I be the $n \times n$ identity matrix. Let $B \in M_{n\times n}(\mathbf{R})$ be such that $\|B\|_{\mathrm{op}} < 1$. Prove that $I + B$ is invertible and that $\|(I + B)^{-1}\|_{\mathrm{op}} \leq 1/(1 - \|B\|_{\mathrm{op}})$.

5 Conditions for differentiability

If F is differentiable at p then, as we have seen,

1. The directional derivatives $(D_v F)(p)$ exist for all directions v.
2. For every v, the equality $DF\big|_p(v) = (D_v F)(p)$ holds. Since $DF\big|_p$ is linear, the map $v \mapsto (D_v F)(p)$ must also be linear.

Thus, these are *necessary* conditions for F to be differentiable at p. As the next two examples show, these conditions are *not* sufficient. In these examples, for notational simplicity we write elements of $\mathbf{R}^2$ as row vectors.

Example 5.1 Let $f : \mathbf{R}^2 \to \mathbf{R}$ be the following function (any nonlinear function that is homogeneous of degree 1 would do):

$$f(x, y) = \begin{cases} \frac{x^3}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

Since $f(tx, ty) = tf(x, y)$ we have

$$\left(D_{(0,0)} f\right)((a, b)) = \lim_{t\to 0} \frac{f(t(a, b)) - f((0, 0))}{t} = f((a, b)) = \frac{a^3}{a^2 + b^2}.$$

Thus, for every (a, b), the directional derivative of f at $(0, 0)$ in the direction (a, b) exists. However, the map $(a, b) \to \left(D_{(0,0)} f\right)((a, b))$ is is not linear, so f is not differentiable at $(0, 0)$. ▲

The next example shows that even if $(D_v F)(p)$ exists for all v *and* the map $v \mapsto (D_v F)(p)$ is linear, F need not be differentiable at p.

Example 5.2 Consider the function $f : \mathbf{R}^2 \to \mathbf{R}$ given by

$$f(x,y) = \begin{cases} \frac{xy^3}{x^2+y^4} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

For this f we have both

$$f\big|_{x\text{-axis}} \equiv 0 \qquad \text{and} \qquad f\big|_{y\text{-axis}} \equiv 0,$$

so $\frac{\partial f}{\partial x}(0,0) = 0 = \frac{\partial f}{\partial y}(0,0)$. Moreover, for any $(a,b) \neq (0,0)$,

$$\frac{f((0,0)+t(a,b)) - f(0,0)}{t} = \frac{f(ta,tb)}{t} = \frac{1}{t} \cdot \frac{t^4ab^3}{t^2a^2+t^4b^4} = \frac{tab^3}{a^2+t^2b^4} \to 0 \qquad \text{as } t \to 0.$$

Thus, all directional derivatives of f exist at $(0,0)$ and are zero (so, in particular, the map $v \mapsto (D_{(0,0)}f)(v)$ is linear). Therefore, if f were differentiable at $(0,0)$ the derivative of f at $(0,0)$ would be the zero-map $\mathbf{R}^2 \to \mathbf{R}$. By the "Substitution Lemma for limits", it would then follow that if $\gamma : \mathbf{R} \setminus \{0\} \to \mathbf{R}^2 \setminus \{0\}$ is any function for which $\lim_{t\to 0} \gamma(t) = (0,0)$ (a curve approaching the origin as $t \to 0$), we must have $\lim_{t\to 0} \frac{f(\gamma(t))-f(0,0)}{\|\gamma(t)\|} = 0$. In particular this would hold for $\gamma(t) = (t^2, t) = (x(t), y(t))$ (approaching the origin along the parabola $x = y^2$). But for this curve γ, we have

$$\lim_{t\to 0} \frac{f(x(t),y(t)) - f(0,0)}{\|(x(t),y(t))\|} = \lim_{t\to 0} \frac{t^5/(2t^4)}{\sqrt{t^4+t^2}} = \lim_{t\to 0} \frac{1}{2\sqrt{1+t^2}} \frac{t}{|t|},$$

which does not exist. Hence f is not differentiable at $(0,0)$. ▲

In view of the previous two examples, one may ask whether there are *any* simple conditions on the directional derivatives of F that guarantee the existence of the derivative of F at a given point? The answer is yes; one such result, stated only for the case $V = \mathbf{R}^n, W = \mathbf{R}^m$, is Proposition 5.3 below. However, bear in mind that the result gives just a *sufficient* conditions for differentiability at a point, not *necessary* condition. Definition 1.1 *cannot* be simplified.

Proposition 5.3 *Let $U \subset \mathbf{R}^n$ be open, $F = (f^1, \ldots, f^m)^t : U \to \mathbf{R}^m$ a function, and $p \in U$. Let $\{x^i\}_{i=1}^n$ be the standard coordinates on $\mathbf{R}^n$. If each of the partial derivatives $\frac{\partial f^i}{\partial x^j}$ exists on some open neighborhood of p, and is continuous at p, then F is differentiable at p.*

Remark 5.4 The condition in Proposition 5.3 is the first condition (sufficient or necessary) we've seen for "F is differentiability at p" that involves knowing differentiability of something at points *other than* p. The fact that it involves any sort of differentiability at points other than p should serve as a reminder that this condition is unlikely to be *necessary* for differentiability at p. The function $f : \mathbf{R} \to \mathbf{R}$ defined by

$$f(x) = \begin{cases} x^2 \sin(1/x), & x \neq 0, \\ 0, & x = 0, \end{cases}$$

is an example of a function $\mathbf{R} \to \mathbf{R}$ that is differentiable everywhere, but for which the condition in Proposition 5.3 is *not* met at $x = 0$. ▲

Proof of Proposition 5.3: In view of Proposition 4.3, it suffices to prove Proposition 5.3 for the case $m = 1$. Thus, let $f : U \to \mathbf{R}$ be a function such that for $1 \leq j \leq n$, each of the partial derivatives $\frac{\partial f}{\partial x^j}$ exists on an open neighborhood of p, and is continuous at p.

Let $\{e_j\}_{j=1}^n$ be the standard basis of $\mathbf{R}^n$. Define a linear transformation $T : \mathbf{R}^n \to \mathbf{R}$ by $T(\sum_{j=1}^n v_j e_j) = \sum_{j=1}^n \frac{\partial f}{\partial x^j}(p) v_j$. Taking the norm on $V = \mathbf{R}^n$ to be the ℓ^1-norm $\| \ \|_1$, and the norm on $W = \mathbf{R}$ to be the standard norm on $\mathbf{R}$, we will show that for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $v \in \mathbf{R}^n$ with $\|v\|_1 < \delta$, (1.3) is satisfied, and therefore that f is differentiable at p. (As noted in Remark 1.8, the choices of norms on V and W do not affect whether f is differentiable at p, so we are free to choose any norms we find convenient.)

Let $\epsilon > 0$. For each $j \in \{1, \ldots, n\}$ let U_j be an open neighborhood of p such that for all $q \in U_j$, $|\frac{\partial f}{\partial x^j}(q) - \frac{\partial f}{\partial x^j}(p)| < \epsilon$. Let $U' = \bigcap_{1 \leq j \leq n} U_j$. Then U' is a finite intersection of open neighborhoods of p, hence an open neighborhood of p. For $r > 0$ and $q \in \mathbf{R}^n$, let $B_r^\infty(q)$ denote the open ball of radius δ and center q in $(\mathbf{R}^n, d_\infty)$, where d_∞ is the ℓ^∞-metric on $\mathbf{R}^n$. Since all norms on $\mathbf{R}^n$ are equivalent, U' contains $B_r^\infty(p)$ for some $r > 0$, hence contains the closed ball $\overline{B}_r^{\infty}(p)$ for any $r \in (0, r_1)$. Let $\delta > 0$ be such that $\overline{B}_\delta^{\infty}(p) \subset U'$. By definition of the metric associated with a norm, we have $\overline{B}_\delta^{\infty}(p) = \{p + v : v \in \overline{B}_\delta^{\infty}(0)\}$.

Let $\{x^j\}_{j=1}^n$ be the standard coordinates on $\mathbf{R}^n$, and let $v \in V$. For $1 \leq j \leq n$ let $p_j = x^j(p), v^j = x^j(v)$. (Thus $v = (v^1, \ldots, v^n)^t$ and $|v^j| < \delta$, $1 \leq j \leq n$.) For $0 \leq k \leq n$ define $q^{(k)} = p + \sum_{j=1}^k v^j e_j$, the point whose first k coordinates are those of $p + v$ and whose last $n - k$ coordinates are those of p. Then

$$\begin{aligned} f(p+v) - f(p) &= f(q^{(n)}) - f(q^{(0)}) \\ &= [f(q^{(n)}) - f(q^{(n-1)})] + [f(q^{(n-1)}) - f(q^{(n-2)})] + \cdots + [f(q^{(1)}) - f(q^{(0)})]. \end{aligned} \tag{5.1}$$

(Essentially, (5.1), read from the last bracketed expression to the first, says, "Walk from the 'corner' p of a 'cube' to the opposite 'corner' $p + v$ by walking first along an edge parallel to the 1st coordinate axis, then along an edge parallel to the 2nd coordinate axis, etc.") Let $k \in \{1, \ldots, n\}$. Observe that $q^{(k)} = q^{(k-1)} + v^k e_k$. For $t \in [-\delta, \delta]$, the point $z^{(k)}(t) := q^{(k-1)} + t e_k$ lies in $\overline{B}_\delta^{\infty}(p)$, on which $\frac{\partial f}{\partial x^k}$ exists and is continuous. But

$$\frac{d}{dt} f(z^{(k)}(t)) = \frac{\partial f}{\partial x^k}(z^{(k)}(t)),$$

so the function $t \mapsto f(z^{(k)}(t))$ is differentiable on an open interval that contains the closed interval with endpoints 0 and v^k. Hence we may apply the Mean Value Theorem and select c_k between 0 and v^k such that

$$f(q^{(k)}) - f(q^{(k-1)}) = f(z^{(k)}(v^k)) - f(z^{(k)}(0)) = \frac{\partial f}{\partial x^k}(z^{(k)}(c_k))v^k. \tag{5.2}$$

Define $\tilde{q}^{(k)} = z^{(k)}(c_k)$. Note that $\tilde{q}^{(k)}$ lies in the ball $B_\delta^\infty(p) \subset U'$, so $|\frac{\partial f}{\partial x^j}(\tilde{q}^{(k)}) - \frac{\partial f}{\partial x^j}(p)| < \epsilon$.

Writing $\tilde{q}^{(k)} = z^{(k)}(c_k)$ for each k, plugging (5.2) into (5.1), and using the definition of T, we have

$$\begin{aligned}
|f(p+v) - f(v) - T(v)| &= \left| \sum_{k=1}^n \frac{\partial f}{\partial x^k}(\tilde{q}^{(k)})\, v^k - \sum_{j=1}^n \frac{\partial f}{\partial x^j}(p)\, v^j \right| \\
&= \left| \sum_{j=1}^n \left(\frac{\partial f}{\partial x^j}(\tilde{q}^{(j)}) - \frac{\partial f}{\partial x^j}(p) \right) v^j \right| \\
&\leq \sum_{j=1}^n \left| \frac{\partial f}{\partial x^j}(\tilde{q}^{(j)}) - \frac{\partial f}{\partial x^j}(p) \right| |v^j| \\
&\leq \sum_{j=1}^n \epsilon |v^j| \\
&= \epsilon \|v\|_1.
\end{aligned}$$

Thus for all $v \in \mathbf{R}^n$ with $\|v\|_1 < \delta$, we have $|f(p+v) - f(v) - T(v)| \leq \epsilon \|v\|_1$. Since ϵ was arbitrary, it follows that f is differentiable at p. ■

Corollary 5.5 *Let $U \subset \mathbf{R}^n$ be open, $F : U \to \mathbf{R}^m$ a function, and $p \in U$. Suppose there is an open neighborhood of U' of p such that for all $v \in \mathbf{R}^n$, the directional derivative $(D_v F)(q)$ exists for every $q \in U'$, and the map $q \mapsto (D_v F)(q)$ is continuous. Then F is differentiable at p.*

Exercise 5.6 (a) Prove Corollary 5.5.

(b) Strengthen Corollary 5.5 by showing that $\mathbf{R}^n, \mathbf{R}^m$ can be replaced by arbitrary finite-dimensional vector spaces. I.e., prove the following corollary:

Corollary 5.7 *Let V, W be finite-dimensional vector spaces, $U \subset \mathbf{R}^n$ open, $F : U \to \mathbf{R}^m$ a function, and $p \in U$. Suppose there is an open neighborhood of U' of p such that for all $v \in V$, the directional derivative $(D_v F)(q)$ exists for every $q \in U'$, and the map $q \mapsto (D_v F)(q)$ is continuous. Then F is differentiable at p.*

Remark 5.8 Proposition 5.3 is stronger than Corollary 5.5; the Proposition shows that we can deduce differentiability at p from knowing the continuity at p of just all the first partials, of which there are only finitely many, whereas there are infinitely many directional derivatives. However, when V and W are not explicitly $\mathbf{R}^n$ and $\mathbf{R}^m$, there are no "standard coordinates", so the partials used in the Proposition do not make sense. We can always introduce bases for V and W (equivalently, introduce isomorphisms $V \to \mathbf{R}^{\dim(V)}$ and $W \to \mathbf{R}^{\dim(W)}$). A basis of V determines coordinate-functions on V, while a basis of W determines component-functions $\{f^i\}$ of the map F, so choices of bases allow us to define partial derivatives of component-functions with respect to coordinates on V. However, there are instances in which it is very easy to compute all directional derivatives, and show that they are continuous; introducing bases and computing partial derivatives of component functions simply becomes extra work. In these instances, Corollary 5.7 can be much more useful than Proposition 5.3. The exercise below illustrates one such instance. ▲

Exercise 5.9 Let $V = W = M_{n\times n}(\mathbf{R})$, the space of $n \times n$ matrice with real extrieds. Define $F : V \to V$ by $F(A) = A^2 := AA$. (For any square matrix A and positive integer k, we define $A^k = AA\ldots A$, the product of k copies of A.) (a) Compute $(D_B F)(A)$ for all $A, B \in M_{n\times n}$. (b) Show that for each $B \in V$, the map $A \mapsto (D_B F)(A)$ is continuous. (Hence F is differentiable.)

Note: This exercise illustrates one *method* for showing differentiability, but for this particular function it's not necessarily the fastest or easiest method. Differentiability of F can also be proven quickly by (i) using an obvious basis for $M_{n\times n}(\mathbf{R})$ (see Exercise 6.4) to yield an isomorphism $M_{n\times n}(\mathbf{R}) \to \mathbf{R}^{n^2}$, (ii) using the fact the entries of A^2 are polynomials in the entries of A, and (iii) proving that for any $m \in \mathbf{R}^m$, every polynomial in the standard coordinate functions is differentiable. But this approach wouldn't give you nice simple formulas for the derivative or directional derivatives of F.

Exercise 5.10 (derivative of the matrix-inversion map) For $n \in \mathbf{N}$, we define $GL(n, \mathbf{R}) \subset M_{n\times n}(\mathbf{R})$ to be the set of *invertible* $n\times n$ (real) matrices. Let $\iota : (GL(n, \mathbf{R}) \subset M_{n\times n}(\mathbf{R})) \to M_{n\times n}(\mathbf{R})$ be the matrix-inversion map: $\iota(A) = A^{-1}$. (Obviously the image of ι is $GL(n, \mathbf{R})$ itself; we've written the codomain as the larger set $M_{n\times n}(\mathbf{R})$ so that we have a function of the form "$F : (U \subset V) \to W$," where V and W are finite-dimensional vector spaces.)

(a) Show, the following two ways, that $GL(n, \mathbf{R})$ is an open subset of $M_{n\times n}(\mathbf{R})$, and that $\iota : GL(n, \mathbf{R}) \to M_{n\times n}(\mathbf{R})$ is continuous.

(i) Use the fact that, from the standard formula for the inverse of $A \in GL(n, \mathbf{R})$, (essentially "Cramer's Rule"), the entries of A^{-1} are rational functions of the entries of A, with nonzero denominator.

(ii) Use the identity

$$B^{-1} - A^{-1} = A^{-1}(A - B)B^{-1} \quad (\text{for } A, B \in GL(n, \mathbf{R})) \tag{5.3}$$

to show continuity of ι at a fixed, arbitrary $A \in GL(n, \mathbf{R})$, as follows:

– First use (5.3) to bound $\|B^{-1} - A^{-1}\|_{\mathrm{op}}$ in terms of $\|A^{-1}\|_{\mathrm{op}}, \|B^{-1}\|_{\mathrm{op}}$, and $\|B - A\|_{\mathrm{op}}$. (*Note*: For $n > 1$, there is no formula relating $\|A^{-1}\|_{\mathrm{op}}$ to $\|A\|_{\mathrm{op}}$; e.g. it is *not* generally true that $\|A^{-1}\|_{\mathrm{op}} = 1/\|A\|_{\mathrm{op}}$; see Exercise 4.8).

– Feed this upper bound on $\|B^{-1} - A^{-1}\|_{\mathrm{op}}$ into the triangle inequality $\|B^{-1}\|_{\mathrm{op}} \leq \|B^{-1} - A^{-1}\|_{\mathrm{op}} + \|A^{-1}\|_{\mathrm{op}}$ to derive an upper bound on $\|B^{-1}\|_{\mathrm{op}}$ in terms of $\|A^{-1}\|_{\mathrm{op}}$ and $\|B - A\|_{\mathrm{op}}$ (for $\|B - A\|_{\mathrm{op}}$ sufficiently small, as determined by A).

– Then feed that bound on $\|B^{-1}\|_{\mathrm{op}}$ back into your original bound on $\|B^{-1} - A^{-1}\|_{\mathrm{op}}$ to obtain an upper bound on $\|B^{-1} - A^{-1}\|_{\mathrm{op}}$ in terms of $\|A^{-1}\|_{\mathrm{op}}$ and $\|B - A\|_{\mathrm{op}}$ *alone* (again for $\|B - A\|_{\mathrm{op}}$ sufficiently small, as determined by A).

This final bound should show that $\|B^{-1} - A^{-1}\|_{\mathrm{op}} \to 0$ as $\|B - A\|_{\mathrm{op}} \to 0$.

(b) Suppose $B \in M_{n\times n}(\mathbf{R})$ with $\|B\|_{\mathrm{op}} < 1$. Let I be the $n \times n$ identity matrix.

(i) Show that the series $\sum_{n=0}^{\infty} B^n$ converges in norm, and hence converges. (Σ-notation-convention for power series in an $n \times n$ matrix B is that "B^0" means I.)

(ii) Show that $(I-B)\sum_{n=0}^{\infty} B^n = I$. Hence $I-B$ is invertible and $(I-B)^{-1} = \sum_{n=0}^{\infty} B^n$.

(iii) Use part (ii) to obtain a second proof of the inequality established in Exercise 4.8(c).

(iv) Show that the (matrix-valued) power seres $\sum_{n=0}^{\infty}(-1)^n t^n B^n$ in the real variable t has radius convergence at least $1/\|B\|_{\mathrm{op}}$, where we interpret "$1/\|B\|_{\mathrm{op}}$" as ∞ if $B = 0$. Hence on the open interval $(1/\|B\|_{\mathrm{op}}, 1/\|B\|_{\mathrm{op}})$, the term-by-term derivative of $\sum_{n=0}^{\infty}(-1)^n t^n B^n$ converges to $\frac{d}{dt}(I + tB)$. In particular, this holds at $t = 0$, so the directional derivatives of ι at I exist and satisfy $(D_B\iota)_I = -B$.

(c) Let $A \in GL(n, \mathbf{R})$. For any $C \in M_{n\times n}(\mathbf{R})$, the *left-translation* map $L_C : M_{n\times n}(\mathbf{R}) \to M_{n\times n}(\mathbf{R})$ defined by $L_C(X) = CX$, is linear, hence continuous. Use this fact, together with part (b) and the identity $(A + B)^{-1} = A^{-1}(I + BA^{-1})^{-1} = L_{A^{-1}}((I + BA^{-1})^{-1})$, to show that the directional derivatives of ι at I exist and are given by

$$(D_B\iota)_A = -A^{-1}BA^{-1} \; . \tag{5.4}$$

(d) Show that, for each $B \in M_{n\times n}(\mathbf{R})$, the map $GL(n, \mathbf{R}) \to M_{n\times n}(\mathbf{R})$ defined by $A \mapsto -A^{-1}BA^{-1}$, is continuous. Hence, by Corollary 5.5, the inversion-map ι is differentiable, and

$$(D\iota)_A(B) = (D_B\iota)_A = -A^{-1}BA^{-1}. \tag{5.5}$$

(In particular, $(D\iota)|_I = -\mathrm{id}_{M_{n\times n}(\mathbf{R})}$).

Remark 5.11 Power series are, by no means, the only way of showing that $(D\iota)|_I = -\mathrm{id}_{M_{n\times n}(\mathbf{R})}$. For example, we can use the identity

$$(I+B)\left((I+B)^{-1} - I + B\right) = B^2$$

(valid whenever $I+B$ is invertible; in particular, whenever $\|B\|_{\mathrm{op}} < 1$) to show that

$$\begin{aligned} \|(I+B)^{-1} - I + B\|_{\mathrm{op}} &\leq \|(I+B)\|_{\mathrm{op}}^{-1}(\|B\|_{\mathrm{op}})^2 \\ &\leq \frac{(\|B\|_{\mathrm{op}})^2}{1-\|B\|_{\mathrm{op}}} \text{ if } \|B\|_{\mathrm{op}} < 1 \text{ (by (b)(iii) above, or Exercise 4.8(c)).} \end{aligned}$$

Letting $T : M_{n\times n}(\mathbf{R}) \to M_{n\times n}(\mathbf{R})$ be the linear map $B \mapsto -B$, it follows that for any $\epsilon > 0$, the inequality (1.2) is satisfied whenever $\|B\|_{\mathrm{op}}/(1-\|B\|_{\mathrm{op}}) < \epsilon$, hence whenever $\|B\|_{\mathrm{op}}$ is sufficiently small. Thus ι is differentiable at I and $(D\iota)|_I = -\mathrm{id}_{M_{n\times n}(\mathbf{R})}$.

6 Continuous differentiability

Definition 6.1 *If $F : (U \subset V) \to W$ is differentiable we say F is* continuously differentiable (on U)*, or C^1* (on U)*, if the induced map $DF : U \to \mathrm{Hom}(V, W)$ given by $p \mapsto DF\big|_p$ is continuous.*

An immediate corollary of Proposition 5.3 is the following:

Corollary 6.2 *Let $F = (f^1, \ldots, f^m)^t : (U \subset \mathbf{R}^n) \to \mathbf{R}^m$ a function, and let $\{x^i\}_{i=1}^n$ be the standard coordinates on $\mathbf{R}^n$. Then F is continuously differentiable if and only if each of the partial derivatives $\frac{\partial f^i}{\partial x^j}$ exists throughout U and is continuous on U.*

Proof: First assume that F is continuously differentiable. Then $DF|_p$ exists for every $p \in U$, and the map $U \to \mathrm{Hom}(\mathbf{R}^n, \mathbf{R}^m)$ given by $p \mapsto DF|_p$ is continuous. By an earlier exercise, this implies that each of the partial derivatives $\frac{\partial f^i}{\partial x^j}$ exists throughout U and is continuous.

Conversely, assume that each of the partial derivatives $\frac{\partial f^i}{\partial x^j}$ exists throughout U and is continuous on U. By Proposition 5.3, F is differentiable at every point of U. By the same exercise mentioned above, the assumed continuity of the partials implies that the map $U \to \mathrm{Hom}(\mathbf{R}^n, \mathbf{R}^m)$ given by $p \mapsto DF|_p$ is continuous. Hence f is continuously differentiable. ■

Remark 6.3 Because the conditions in Corollary 6.2 are necessary and sufficient for continuous differentiability (not just-plain differentiability!) of $F : (U \subset \mathbf{R}^n) \to \mathbf{R}^m$, the condition "if each of the partial derivatives $\frac{\partial f^i}{\partial x^j}$ exists throughout U and is continuous on U" is often taken as the *definition* of "F is continuously differentiable on U", in place of Definition 6.1. (Definition 6.1 is *conceptually* the best definition of "continuous differentiability", but not the easiest definition to apply in practice.) Note, however, that as stated, this alternate definition applies only for functions from (an open subset of) $\mathbf{R}^n$ to $\mathbf{R}^m$. For more general finite-dimensional vector spaces V and W, we must introduce bases, and the associated coordinate functions, in order to make a similar definition. It is not hard to show that, in this more general situation, the continuous-partial-derivatives condition is independent of the choice of bases. ▲

Exercise 6.4 In the context of Definition 6.1, let $V = \mathbf{R}^n$ and $W = \mathbf{R}^m$, write F as $(f^1, \ldots, f^m)^t$. Show that the map $p \mapsto DF\big|_p$ is continuous if and only if the map

$$p \mapsto J_F(p)$$

is continuous as a map from U to the space $M_{m\times n}$ of $m \times n$ matrices, which in turn is equivalent to all of the real-valued functions $\frac{\partial f^i}{\partial x^j}$ being continuous. (*Suggestion*: Use the fact that $\{E_{ij} : 1 \le i \le m, 1 \le j \le n\}$ is a basis of $M_{m\times n}$, where E_{ij} is the $m \times n$ matrix whose $(i,j)^{\text{th}}$ entry is 1 and all of whose other entries are 0.)

7 "Mean Value Theorem" for vector-valued functions?

The Mean Value Theorem (MVT), a beautiful and important theorem for real-valued functions of a real variable, asserts that for any continuous function $f : [a,b] \to \mathbf{R}$ that is differentiable on (a,b), there exists $c \in (a,b)$ such that $f(b) - f(a) = f'(c)(b-a)$.

It is easy to see that the MVT cannot generalize to functions whose codomains have dimension greater than 1. For example, if we define $f : [0, 2\pi] \to \mathbf{R}^2$ by $f(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, then the hypotheses of the MVT are met except for the codomain not being $\mathbf{R}$, but

$$f(2\pi) - f(0) \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\neq\; 2\pi f'(c) \quad \text{for any } c,$$

since $f'(t)$ is a nonzero vector for every t.

However, an important corollary of the MVT is that if $I \subset \mathbf{R}$ is an open interval, and $f : I \to \mathbf{R}$ is continuously differentiable, then for all $a < b \in I$ we have $|f(b) - f(a)| \le M|b-a|$, where $M = \max_{t\in[a,b]} |f'(t)|$. Of course, if f' is bounded on the entire interval I, then the above inequality also holds with M replaced by $\sup_{t\in I} |f'(t)|$. This *corollary* of the MVT does generalize to the setting of Definition 1.1 (allowing the dimensions of both the domain and codomain to be arbitrary natural numbers), provided the domain is *convex*:

Definition 7.1 A subset U of a vector space V is called *convex* if for all $p, q \in U$ the line segment joining p and q lies entirely in U (i.e. if $tp + (1-t)q \in U$ for all $t \in [0,1]$).

The lemma below should really not be called a "Mean Value Theorem", but that's what my own Advanced Calculus 1 professor called it, and I've never found a better name that's not inconveniently long! I think of this result as a lemma, rather than an end result, since its importance lies in its applications (none of which are in these notes).

Lemma 7.2 ("Mean Value Theorem" for vector-valued functions) *Let $U \subset V$ be an open set and let $F : U \to W$ be a C^1 map.*

(a) Let $K \subset U$ be a compact, convex set, Then for all $p, q \in K$,

$$\|F(q) - F(p)\| \le M\|q - p\|, \tag{7.1}$$

where $M = \sup_{x\in K} \|(DF)|_x\|_{\text{op}} < \infty$.

(b) Suppose DF (viewed as the map $x \mapsto DF|_x$) is bounded on U, and that U is convex. Then inequality (7.1) *holds with $M = \sup_{x\in U} \|(DF)|_x\|_{\text{op}}$.*

Proof: If hypotheses (a) hold, define $U' = K$; if hypotheses (b) hold define $U' = U$. Under hypotheses (a), the function $K \to \mathbf{R}$ defined by $x \mapsto \|(DF)|_x\|_{\text{op}}$ is continuous, hence achieves a maximum value. Thus $M = \sup_{x\in K} \|(DF)|_x\|_{\text{op}} = \max_{x\in K} \|(DF)|_x\|_{\text{op}} < \infty$. Hence, under hypotheses (a) or (b), the set U' is convex and the function DF is bounded on U'.

Let $p, q \in U'$, let $v = q - p$ (so $q = p + v$), and define $\gamma : \mathbf{R} \to V$ by $\gamma(t) = p + tv$. Then, since the image of γ lies in U', we have

$$F(q) - F(p) = \int_0^1 \frac{d}{dt} F(\gamma(t))\, dt = \int_0^1 DF|_{p+tv}(\gamma'(t))\, dt = \int_0^1 DF|_{p+tv}(v)\, dt.$$

(In the first equality, we have used the Fundamental Theorem of Calculus for Vector-Valued Functions, which can be proven "easily" from the FTC for real-valued functions.[1]

[1]The proof is *truly* easy only if we mistakenly forget that an m-dimensional vector space is only *isomorphic* to $\mathbf{R}^m$, not *equal* to $\mathbf{R}^m$, and then define vector-valued integrals componentwise. If we don't assume $W = \mathbf{R}^m$, but choose a basis of W and *then* define the integral componentwise, we must check that values of integrals are independent of our choice of basis. See "Notes on Riemann Integration" (dgarchive.com/classes/6257_s22), Section 1.9, for a treatment of vector-valued integration that does not start with choosing a basis. Exercises 1.20, 1.21, 1.22(c), and 1.23(b) (pp. 58–59) address the FTC for VVF's. The form of the "triangle inequality for integrals" (valid for *any* norm on W) used below is Corollary 1.96 (p. 55). To prove a special case of this, Rudin's *Principles of Mathematical Analysis* gives a "clever trick" that works, directly, only for the case $W = \mathbf{R}^m$ and $\|\ \|$ = Euclidean norm. In the finite-dimensional case, Rudin's trick can be generalized by using the dual pairing $W^* \times W \to \mathbf{R}$ instead of the standard inner product on $\mathbf{R}^m$. This trick can be further generalized to any Banach space W (including infinite-dimensional spaces) using the Hahn-Banach Theorem. But using such a cannon to kill a flea—and even using Rudin's trick in the simplest case—obscures the fundamental reason *why* the "triangle inequality for integrals" holds for functions taking values in any Banach space, namely the *ordinary* triangle inequality.

Continuity of DF is being used to ensure that the function $t \mapsto DF|_{p+tv}(v)$ is continuous, hence integrable. This is the only reason we needed to assume that F is C^1—i.e. that the map $q \mapsto DF|_q$ is continuous—rather than just a differentiable function whose derivative is bounded.) Hence

$$\begin{aligned}
\|F(q) - F(p)\| &= \left\| \int_0^1 (DF|_{p+tv})(v)\, dt \right\| \\
&\leq \int_0^1 \|(DF|_{p+tv})(v)\|\, dt \quad \text{(by the "triangle inequality for integrals")} \\
&\leq \int_0^1 \|DF|_{p+tv}\|_{\text{op}}\, \|v\|\, dt \\
&\leq \big(\sup_{x \in \text{image}(\gamma)} \|(DF)\big|_x\|_{\text{op}} \big) \|v\| \\
&\leq \big(\sup_{x \in U'} \|(DF)\big|_x\|_{\text{op}} \big) \|v\| \\
&= M\|q - p\|.
\end{aligned}$$

■

8 Higher-order Derivatives

Suppose that $F : (U \subset V) \to W$ is differentiable. Let $Df : U \to \text{Hom}(V, W)$ be the map defined by $p \mapsto Df|_p$. Since $\text{Hom}(V, W)$ is another finite-dimensional vectore space, we have defined what "differentiable function $U \to \text{Hom}(V, W)$" means. We say that f is *twice differentiable* (either at a specified point, or—if no specific evaluation-point is mentioned—throughout U) if Df is differentiable. Recursively, we define f to be k times differentiable if Df is $(k-1)$-times differentiable, and call $D^k f|_p := D(D^{k-1} f)|_p$ the k^{th} *derivative of f at p*. If $D^k f$ is continuous, we say that f is C^k (k-times continuously differentiable). Note that, when these derivatives exist, they live in progressively larger and more formidable-looking vector spaces:

$$\begin{aligned}
&Df|_p \in \text{Hom}(V, W) \\
&D^2 f|_p \in \text{Hom}(V, \text{Hom}(V, W)) \\
&D^3 f|_p \in \text{Hom}(V, \text{Hom}(V, \text{Hom}(V, W))) \\
&\vdots
\end{aligned}$$

Fortunately, these vector spaces are canonically isomorphic to simpler-looking ones. We will see this shortly, but first, we examine how $D^2 f$ is related to directional derivatives (for a twice differentiable function). For this we need a lemma:

Lemma 8.1 *Let $g : (U \subset V) \to \mathrm{Hom}(V, W)$ be a differentiable map, and let $p \in U$. Then for all $u, v \in V$*

$$(Dg|_p(u))\,(v) = (D_u g)|_p(v) = D_u\big(q \mapsto \overbrace{\underbrace{(g(q))}_{\substack{\in \\ \mathrm{Hom}(V,W)}}\ (v)}^{\in\ W}\ \big)|_p$$

Proof: First, given any interval (or any topological space) I, any $t_0 \in I$, and any map $t \mapsto A(t)$ from I to $\mathrm{Hom}(V, W)$ for which $\lim_{t\to t_0} A(t)$ exists, then for all $v \in V$, letting $L = \lim_{t\to t_0} A(t)$,

$$\|A(t)u - L(u)\| \le \|A(t) - L\|_{\mathrm{op}}\,\|v\| \qquad \to 0 \ \text{ as } t \to t_0\ .$$

Hence $L(v) = \lim_{t\to t_0}(A(t)v)$; i.e.

$$\left(\lim_{t\to t_0} A(t)\right) v = \lim_{t\to t_0}(A(t)v). \tag{8.2}$$

Let $u, v \in V$. Applying (8.2) with $A(t) = \frac{g(p+tu)-g(p)}{t}$ and I an open interval containing 0,

$$\begin{aligned}
(Dg|_p)\,(u) &= (D_u g)_p(v) \qquad\qquad \text{(by Proposition 3.4)}\\
&= \left(\lim_{t\to 0}\frac{g(p+tu)-g(p)}{t}\right)(v)\\
&= \lim_{t\to 0}\left(\frac{g(p+tu)-g(p)}{t}(v)\right)\\
&= \lim_{t\to 0}\left(\frac{(g(p+tu))(v)-(g(p))(v)}{t}\right)\\
&= D_u(q \mapsto (g(q))(v)).
\end{aligned}$$

■

Suppose now that $f : (U \subset V) \to W$ is twice differentiable at p and let $u, v \in V$. Then by definition of $D^2 f$ and Proposition 3.4,

$$(D^2 f)_p(u) = D(Df)|_p(u) = D_u(Df)|_p$$

(an element of $\mathrm{Hom}(V, W)$). Hence

$$\begin{aligned}
\big((D^2 f)_p(u)\big)\,(v) &= (D_u(Df)|_p)\,(v)\\
&= (D_u\,(q \mapsto (Df)|_q(v)))\,|_p\\
&= (D_u\,(q \mapsto (D_v f)|_q))\,|_p.
\end{aligned} \tag{8.3}$$

Equation (8.3) is the sought relation between the second derivative $D^2 f$ and (iterated) directional derivatives.

Definition 8.2 Let $V_1, V_2, \ldots V_k$ be finite-dimensional vector spaces. A map $B : V_1 \times V_2 \to W$ is *bilinear* if it is linear in each variable separately; i.e. if for all $v_1 \in V_1$, the map $V_2 \to W$ defined by $u \mapsto B(v_1, u)$ is linear and for all $v_2 \in V_2$, the map $V_1 \to W$ defined by $u \mapsto B(u, v_2)$ is linear.

Similarly we define *trilinear* maps $V_1 \times V_2 \times V_3 \to W$, and, more generally *$k$-linear* or *multilinear* maps $V_1 \times V_2 \times \cdots \times V_k \to W$.

In these notes we use the terminology $\mathrm{Bihom}(V_1 \times V_2, W)$ for the set of all bilinear maps $V_1 \times V_2 \to W$, and analogously define the terminology $\mathrm{Trihom}(V_1 \times V_2 \times V_3, W)$ and k-$\mathrm{Hom}(V_1 \times V_2 \times \cdots \times V_k, W)$. ▲

For any nonempty set A, the set $\mathrm{Func}(A, W)$ of all maps from A to W has a canonical vector-space structure, induced by pointwise operations $\big((f+g)(a) = f(a) + g(a)$; $(cf)(a) = c(f(a)$; $0_{\mathrm{Func}(A,W)} =$ the constant function $a \mapsto 0_W\big)$. By convention, when we say that $S \subset \mathrm{Func}(A, W)$ "is" a vector space, we mean that S is a vector subspace of $\mathrm{Func}(A, W)$.

Exercise 8.3 *Let* $V_1, V_2, \ldots V_k$ *be finite-dimensional vector spaces. Then* k-$\mathrm{Hom}(V_1 \times \cdots \times V_k, W)$ *is a vector space.*

Lemma 8.4 *Let* $V_1, V_2, \ldots V_k$ *be finite-dimensional vector spaces. Then* $\mathrm{Hom}(V_1, \mathrm{Hom}(V_2, W))$ *is canonically isomorphic to* $\mathrm{Bihom}(V_1 \times V_2, W)$ *(via the isomorphism given in the proof below).*

More generally, $\mathrm{Hom}(V_1, \mathrm{Hom}(V_2, \mathrm{Hom}(V_3, \ldots, \mathrm{Hom}(V_k, W))))$ *is canonically isomorphic to* k-$\mathrm{Hom}(V_1 \times V_2 \cdots \times V_k, W)$. *In particular, this holds when* $V_1 = V_2 = \cdots = V_k = V$.

Proof: [Sketch] For $L \in \mathrm{Hom}(V_1, \mathrm{Hom}(V_2, W))$, define a map $\hat{L} : V_1 \times V_2 \to W$ by

$$\hat{L}(v_1, v_2) = \big(L(v_1)\big)(v_2)$$

As is easily checked, for every $L \in \mathrm{Hom}(V_1, \mathrm{Hom}(V_2, W))$, the map $\hat{L} : V_1 \times V_2 \to W$. Hence the assignment $L \mapsto \hat{L}$ defines a map φ from the vector space $\mathrm{Hom}(V_1, \mathrm{Hom}(V_2, W))$ to the vector space $\mathrm{Bihom}(V_1 \times V_2, W)$. As is easily checked, φ is linear and invertible. (Its inverse is the map $\psi : \mathrm{Bihom}(V_1 \times V_2, W) \to \mathrm{Hom}(V_1, \mathrm{Hom}(V_2, W))$ defined by $(\psi(B))(v_1) = B(v_1, \cdot)$.)

This yields the $k = 2$ case of the lemma. Induction then yields the general case. ■

For a twice differentiable map $f : (U \subset V) \to W$, we may post-compose $D^2 f$ with the canonical isomorphism in Lemma 8.4 to obtain a map $U \to \mathrm{Bihom}(V \times V, W)$.

(Standard) convention: For f as above, we avoid introducing extra notation, and simply regard $D^2 f$ as a map from U to $\mathrm{Bihom}(V \times V, W)$. More generally, if f is k-times differentiable, we regard $D^k f$ as a map from U to k-$\mathrm{Hom}(V \times V \times \cdots \times V, W)$.

For example, we view $(D^2 f)|_p$ as the bilinear map $V \times V \to W$ given by

$$\begin{aligned}(D^2 f)|_p(u,v) &= \text{“old”} \ \big((D^2 f)_p(u)\big)(v) \\ &= D_u\left(q \mapsto (D_v f)|_q\right) \qquad \text{(by equation (8.3)).}\end{aligned}$$

Similar principles apply to derivatives of higher order.

Example 8.5 (the case $V = \mathbf{R}^n, W = \mathbf{R}$) Let $\{e_i\}_{i=1}^n$ be the standard basis of $\mathbf{R}^n$ and $\{x^i\}_{i=1}^n$ the standard coordinates on $\mathbf{R}^n$. Let $U \subset \mathbf{R}^n$ be an open set and suppose that $f : U \to \mathbf{R}$ is k-times differentiable, where $k \geq 2$. Then

$$(D^2 f)(e_i, e_j) = D_{e_i}(D_{e_j} f) = \frac{\partial}{\partial x^i}\left(\frac{\partial f}{\partial x^j}\right) = \frac{\partial^2 f}{\partial x^i\, \partial x^j}\ ,$$

and more generally,

$$(D^k f)(e_{i_1}, e_{i_2}, \ldots, e_{i_k}) = \frac{\partial^k f}{\partial x^{i_1} \ldots \partial x^{i_k}}\ .$$

▲

Theorem 8.6 (Equality of cross-partials) *Let $U \subset \mathbf{R}^n$ be open and let $f : U \to \mathbf{R}$ be a function all of whose first and second partial derivatives exist. Let $i, j \in \{1, \ldots, n\}$. If both $\frac{\partial^2 f}{\partial x^i\, \partial x^j}$ and $\frac{\partial^2 f}{\partial x^j\, \partial x^i}$ are continuous, then they are equal to each other.*

Proof: Omitted.

Corollary 8.7 *Let $U \subset V$ be open, $f : U \to W$ a C^2 function, and let $p \in U$. Then the bilinear map $(D^2 f)|_p : V \times V \to W$ is symmetric.*

Exercise 8.8 Prove Corollary 8.7. (First use Theorem 8.6 to handle the case $V = \mathbf{R}^n, W = \mathbf{R}$. Use this to generalize to the case $V = \mathbf{R}^n, W = \mathbf{R}^m$, and finally to general vector spaces of these dimensions.)

Exercise 8.9 Let $n = \dim(V)$, let $m = \dim(W)$, and suppose that $L : V \to \mathbf{R}^n$ and $L' : W \to \mathbf{R}^m$ are isomorphisms. (Note: since linear maps between finite-dimensional vector spaces are continuous, isomorphisms carry open sets to open sets.) Let $k \geq 1$. Then a map $f : (U \subset V) \to W$ is k-times differentiable (respectively, C^k) if and only if the map $\tilde{f} := L' \circ f \circ L^{-1}|_{L(U)} : (L(U) \subset \mathbf{R}^n) \to \mathbf{R}^m$ is k-times differentiable (resp. C^k). Furthermore, if f and $\tilde{f}$ are k-times differentiable, then the following relation holds for all $p \in U$:

$$(D^k f)|_p(v_1, \ldots, v_k) = (L')^{-1}\left((D^k \tilde{f})|_{L(p)}(L(v_1), \ldots, L(v_k))\right)$$

Exercise 8.10 Let $k \in \mathbf{N}$, let $U \subset \mathbf{R}^n$ be open, and let $f = \begin{pmatrix} f^1 \\ \vdots \\ f^m \end{pmatrix} : U \to \mathbf{R}^m$ be a k-times differentiable function. Let $J_f : U \to M_{m\times n}(\mathbf{R})$ be the Jacobian matrix of f (a matrix-valued function of the evaluation-point).

(a) Prove that the following are equivalent:

(i) For all $j_1, \dots, j_k \in \{1, \dots, n\}$ and all $i \in \{1, \dots, m\}$, the function $\frac{\partial^k f^i}{\partial x^{j_1} \dots \partial x^{j_k}} : U \to \mathbf{R}$ is continuous.

(ii) For all $j \in \{1, \dots, m\}$ and all $i \in \{1, \dots, n\}$, the function $\partial f^i / \partial x^j$ is C^{k-1}.

(iii) The function $J_f : U \to M_{m\times n}(\mathbf{R})$ is C^{k-1}.

(iv) The function $f : U \to \mathbf{R}^m$ is C^k.

(b) Show that if f is C^k, then for all $j_1, \dots, j_k \in \{1, \dots, n\}$, all $i \in \{1, \dots, m\}$, and every permutation σ of $\{1, \dots, k\}$,

$$\frac{\partial^k f^i}{\partial x^{j_1} \dots \partial x^{j_k}} = \frac{\partial^k f^i}{\partial x^{j_{\sigma(1)}} \dots \partial x^{j_{\sigma(k)}}} \,. \tag{8.4}$$

In particular, (8.4) holds for a permutation that re-orders the indices $j_1, \dots, j_k$ in increasing order. Hence, every k^{th} order partial derivative of f^i is equal to one of the form $\frac{\partial^k f^i}{\partial x^{j_1} \dots \partial x^{j_k}}$ with $j_1 \le j_2 \le \cdots \le j_k$.

Note: in parts (a)(i) and (ii), if we use vector-valued partial derivatives as defined in equations (3.1)–(3.1), we can omit the index i and the "for all $i \in \{1, \dots, m\}$." ▲

Exercise 8.11 Let $k \in \mathbf{N}$ and let $f : (U \subset V) \to W$. Prove that the following are equivalent.

(i) f is C^k.

(ii) For all $v_1, v_2, \dots, v_k \in V$, the map $U \to W$ defined by $p \mapsto D^k f|_p(v_1, v_2, \dots, v_k)$ is continuous.

(You should find this a straightforward consequence of Exercise 8.10(a)'s equivalence "(i) $\Longleftrightarrow$ (iv).")

Exercise 8.12 Consider again the inversion map $\iota : GL(n, \mathbf{R}) \to M_{n\times n}(\mathbf{R})$. In Exercise 5.10, you showed that ι is differentiable, and that for $A \in GL(n, \mathbf{R})$, the directional derivatives of ι at A are given by

$$(D_B\iota)_A = -A^{-1}BA^{-1} \quad \text{(for all } B \in M_{n\times n}(\mathbf{R})) \ . \tag{8.5}$$

(a) Use Exercise 2.3, equation (8.5), and the fact that (for any $B \in M_{n\times n}(\mathbf{R})$) the left-translation map $L_B : M_{n\times n}(\mathbf{R}) \to M_{n\times n}(\mathbf{R}),\ X \mapsto BX$ is linear (hence its own derivative at any point in $M_{n\times n}(\mathbf{R})$), to show that for all $B, C \in M_{n\times n}(\mathbf{R})$,

$$(D^2\iota)|_A(B, C) = A^{-1}\, B\, A^{-1}\, C\, A^{-1} + A^{-1}\, C\, A^{-1}\, B\, A^{-1} \ . \tag{8.6}$$

(*Remark*: When $n = 1$, equation (8.6) yields $(D^2\iota)|_x(1, 1) = 2x^{-3}$. But from Example 8.5, $(D^2\iota)|_x(1, 1) = d^2\iota/dx^2$. Hence we recover the Calculus 1 result that $\frac{d^2}{dx^2}\,(x^{-1}) = 2x^{-3}$. Equation (8.6) is the matrix version of this Calc 1 result.)

(b) Generalize part (a) to show that, for any $k \in\in \mathbf{N}$, $A \in GL(n, \mathbf{R})$, and $B_1, \ldots, B_k \in M_{n\times n}(\mathbf{R})$,

$$(D^k\iota)|_A(B_1, \ldots B_k) = (-1)^k \sum_{\sigma\in S_k} A^{-1}\, B_{\sigma(1)}\, A^{-1}\, B_{\sigma(2)}\, A^{-1} \ldots A^{-1}\, B_{\sigma(k)} A^{-1} \ ,$$

where S_k is the symmetric group of degree k, the group of permutations of $\{1, 2, \ldots, k\}$. Compare this with the Calc 1 formula for $\frac{d^k}{dx^k}\, x^{-1}$.

Lemma 8.13 *For all* $k \in \mathbf{N} \cup \{\infty\}$*, the space* $C^k(U) := \{C^k \text{ maps from } U \text{ to } \mathbf{R}\}$ *is a vector space.*

Proof: Exercise. ■

Lemma 8.14 (the product of C^k functions is C^k) *For any* $k \in \mathbf{N} \cup \{\infty\}$*, if* $f, g : (U \subset V) \to \mathbf{R}$ *are* C^k *functions, then so is* fg*.*

Proof: By definition of "C^∞", it suffices to prove the result just for $k \in \mathbf{N}$. For a given $k \in \mathbf{N}$, Exercise 8.9 shows that if suffices to prove the result for $V = \mathbf{R}^n$ (with $n \in \mathbf{N}$ arbitrary).

Let $n \in \mathbf{N}$, let $\{x^i\}_{i=1}^n$ be the standard coordinates on $\mathbf{R}^n$, and let $f, g : (U \subset \mathbf{R}^n) \to \mathbf{R}$ be C^1 functions.

Let $i \in \{1, \ldots, n\}$. Then

$$\frac{\partial}{\partial x^i}(fg) = \frac{\partial f}{\partial x^i}\, g + f\, \frac{\partial g}{\partial x^i} \ . \tag{8.7}$$

Since f is C^1, the function $\frac{\partial f}{\partial x^i}$ is continuous. Since g is differentiable, g is continuous. Hence the product $\frac{\partial f}{\partial x^i}g$ is continuous. Similarly, $f\frac{\partial g}{\partial x^i}$ is continuous. Hence the RHS of equation (8.7) is continuous, and therefore so is the LHS.

Thus $\frac{\partial}{\partial x^i}(fg)$ is continuous for each $i \in \{1, \ldots, n\}$. By Corollary 6.2, fg is C^1.

Now let $k \in \mathbf{N}$, and assume that the product of any two C^k functions $U \to \mathbf{R}$ is C^k. Let $f, g : U \to \mathbf{R}$ be C^{k+1} functions. Then by Exercise 8.10, for each $i \in \{1, \ldots, n\}$ the partial derivatives $\partial f/\partial x^i$ and $\partial g/\partial x^i$ are C^k. Since f and g are C^{k+1}, they are also C^k. Hence each product on the RHS of equation (8.7) is a product of C^k functions. By our inductive hypothesis, each of these products is C^k, so by Lemma 8.13, the RHS of equation (8.7) is C^k. Hence all first partials of fg are C^k. By Exercise 8.10 again, fg is C^{k+1}.

By induction, it follows that for all $k \in \mathbf{N}$, if $f, g : U \to \mathbf{R}$ are both C^k, then so is fg. ■

Corollary 8.15 (the composition of C^k functions is C^k) *Let Z be a finite-dimensional vector space. For any $k \in \mathbf{N} \cup \{\infty\}$, if $f : (U_1 \subset V) \to (U_2 \subset W)$ and $g : U_2 \to Z$ are C^k functions, then so is $g \circ f : U_1 \to W$.*

Proof: By definition of "C^∞", it suffices to prove the result just for $k \in \mathbf{N}$. For a given $k \in \mathbf{N}$, Exercise 8.9 shows that it suffices to prove the result for $V = \mathbf{R}^n$, $W = \mathbf{R}^m$ (with $n, m \in \mathbf{N}$ arbitrary).

Let $n, m \in \mathbf{N}$, and let $U_1 \subset \mathbf{R}^n$ and $U_2 \subset \mathbf{R}^m$ be open sets. Let $\{x^j\}_{j=1}^n$ and $\{y^i\}_{i=1}^m$ denote the standard coordinates on $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. Let $f : U_1 \to U_2 \subset \mathbf{R}^m$ and $g : U_2 \to \mathbf{R}$ be C^1 functions, and let $h = g \circ f$. Let $f^i = y^i \circ f$ (the i^{th} component of f) for $i \in \{1 \ldots, m\}$. Since $k \geq 1$, the Chain Rule Theorem implies that h is differentiable and that its (first-order) partial derivatives are given by equation (4.6). Hence, for all $j \in \{1, \ldots, n\}$,

$$\frac{\partial h}{\partial x^j} = \sum_{i=1}^m \left(\frac{\partial g}{\partial y^i} \circ f\right) \frac{\partial f^i}{\partial x^j} . \tag{8.8}$$

This formula implies that h is not merely differentiable, but *continously* differentiable (C^1), as follows. Let $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$. Since f is differentiable, f is continuous. Since g is C^1, the function $\partial g/\partial y^i$ is continuous, and hence so is the composition $(\partial g/\partial y^i) \circ f$. Since f is C^1, the function $\partial f^i/\partial x^j$ is continuous. Thus the i^{th} summand on the RHS of equation (8.8) is continuous. Thus the RHS of equation (8.8) is a finite sum of continuous functions, hence is continuous.

Hence $\partial h/\partial x^j$ is continuous, $1 \leq j \leq n$, and therefore h is C^1 (by Corollary 6.2).

This establishes that given any C^1 functions $f : U_1 \to U_2$ and $g : U_2 \to \mathbf{R}$, the composition $g \circ f : U \to \mathbf{R}$ is C^1.

Now let $k \in \mathbf{N}$, and assume that for any two C^k functions $\tilde{f} : U_1 \to U_2$ and $\tilde{g} : U_2 \to \mathbf{R}$, the composition $\tilde{g} \circ \tilde{f}$ is C^k. Let $f : U_1 \to U_2$ and $g : U_2 \to \mathbf{R}$ be C^{k+1} functions, and let $h = g \circ f$. Let $j \in \{1, \ldots, m\}$ and let $i \in \{1, \ldots, n\}$. Then $\partial g/\partial y^i$ is C^k (again using Exercise 8.10), and f is C^k since f is C^{k+1}. Thus, by the inductive hypothesis, $\frac{\partial g}{\partial y^i} \circ f$ is C^k. The function $\frac{\partial f^i}{\partial x^j}$ is also C^k (again using Exercise 8.10). Hence the i^{th} summand on the RHS of equation (8.7) is a product of two C^k functions, hence is C^k. It then follows from Lemma (8.13) that the RHS of (8.7) is C^k, and therefore so is the LHS.

Hence every first partial derivative of h is C^k. Using Exercise 8.10 one last time, we conclude that h is C^{k+1}.

If follows by induction that, for all $k \in \mathbf{N}$, if $f : U_1 \to U_2$ and $g : U_2 \to \mathbf{R}$ are both C^k, then so is $g \circ f$. ■

Remark 8.16 The key to proving Corollary 8.15 *efficiently* in the case $V = \mathbf{R}^n$, $W = \mathbf{R}^m$ was to avoid wasting time trying to produce a *formula* for higher-order partial derivatives of compositions. The student with lots of free time to spare is invited to try to produce an exact formula just for the case $m = n = 1$. There *is* an approach, of intermediate efficiency, that involves stating (precisely) and showing that, when f and g are k-times diferentiable, every k^{th}-order partial derivative of $g \circ f$ function is a polynomial in $\{$ (partial derivatives of g of order $\leq k) \circ f\} \cup \{$ (partial derivatives of f of order $\leq k)\}$. It is not hard to do this, or to see that the terms and coefficients of the relevant polynomials must satisfy some simple constraints. However, it takes time and space to write all this down carefully, and the work is not needed if Corollary 8.15 is all we're after.[2]

[2] There *is* one application, which will not arise in these notes, in which the terms of these polynomials that involve the exactly-k^{th}-order partials of g are important.