The General Derivative

Robert L. Bocchino Jr.

Revised May 2021

This document shows how the derivative studied in first-year calculus (that is, the derivative of a function from real numbers to real numbers) is a special case of a much more general theory: namely, the derivative of a map between normed vector spaces over the real numbers. Traditionally one learns this more general theory, if at all, after two years of calculus (including a year of "multivariable calculus," which covers only parts of the theory) and at least one semester of undergraduate real analysis. The goal here is to present the general theory in a way that is accessible with a minimum of technical background.

There are two prerequisites for reading this document. The first is a good understanding of calculus in one real variable. For that, a year of calculus at the advanced high school or beginning undergraduate level should suffice. The second is a certain amount of "mathematical maturity." Basically this means a willingness to (a) think abstractly and (b) puzzle over something that doesn't make sense until it does, possibly by trying examples on your own.

If you have not studied second-year calculus, don't worry. No knowledge of that subject is assumed. If you have, then this document may help you mentally organize the ad-hoc computational rules involving gradients, partial derivatives, dot products, matrices, etc., that you learned in that course. The ad-hoc rules are of course useful in applications such as physics and engineering. However, they obscure the simplicity and elegance of the underlying theory.

1. Real Numbers

We begin by examining the real numbers. These are ordinary numbers like 1, 1/8, 1.25, $\sqrt{2}$, and π. They are called "real numbers" because they are the numbers that one encounters in day-to-day life for counting and measuring, and because they exclude the square root of -1, which for historical reasons is called an "imaginary number." We denote the set of real numbers $\mathbf{R}$. We recall some basic properties of $\mathbf{R}$.

Addition: R has an operation called **addition**. According to this operation, we can **add** any two real numbers r_1 and r_2 to form a real number called the **sum** of r_1 and r_2 and denoted $r_1 + r_2$. For example, given real numbers 2 and 3, we can form the sum $2 + 3$, which is the real number 5. We say that addition of real numbers is **associative** because for any three real numbers, say 1, 2, and 3, we have $(1 + 2) + 3 = 1 + (2 + 3)$. We say that addition of real numbers is **commutative** because for any two real numbers, say 2 and 3, we have $2 + 3 = 3 + 2$. There is a **zero element** for addition, denoted zero or 0, with the property that for any real number, say 5, we have $5 + 0 = 0 + 5 = 5$.

Subtraction: Every real number r has an **additive inverse** denoted $-r$, such that $r + (-r) = 0$. For example, the additive inverse of 5 is -5, and $5 + (-5) = 0$. The additive inverse is unique. For example, if r is an additive inverse for 5, then $5 + r = 0$ and $5 + (-5) = 0$, so $5 + r = 5 + (-5)$. Adding -5 to both sides shows that $r = -5$.

When we add r_1 and $-r_2$, we also say that we **subtract** r_2 from r_1. For example, we may form the additive inverse of 2 and add it to 5, like this: $5 + (-2)$. Equivalently, we say that we are "subtracting 2 from 5" and write the operation like this: $5 - 2$. The notion $5 - 2$ is a shorthand for $5 + (-2)$. Note that subtraction consists of two operations: forming the additive inverse and adding. Also note that subtraction is not commutative: $5 - 2 \neq 2 - 5$. Nor is it associative: $(3 - 2) - 1 \neq 3 - (2 - 1)$.

Multiplication: R has an operation called **multiplication**. According to this operation, we can **multiply** any two real numbers r_1 and r_2 to form a real number called the **product** of r_1 and r_2 and denoted $r_1 r_2$ or $r_1 \cdot r_2$. (We usually use the dot when writing the product of concrete numbers, and omit it when writing products involving abstract numbers denoted by letters.) Like addition, multiplication is associative and commutative. There is a **unit element** for multiplication, denoted one or 1, such that for any real number, say 10, we have $10 \cdot 1 = 1 \cdot 10 = 10$. Multiplication **distributes over** addition. For example, $2 \cdot (3 + 5) = 2 \cdot 3 + 2 \cdot 5$, and this relation holds for any real numbers in place of 2, 3, and 5. Similarly, $(3 + 5) \cdot 2 = 3 \cdot 2 + 5 \cdot 2$. (This fact follows from the previous one and commutativity,

because $(3+5)\cdot 2 = 2\cdot(3+5) = 2\cdot 3 + 2\cdot 5 = 3\cdot 2 + 5\cdot 2$.) For any real number r, we have $0r = r0 = 0$. (This fact follows from the unit element, distributivity, and the properties of addition and subtraction. For example,

$$5 = 5\cdot 1 = 5\cdot(1+0) = 5\cdot 1 + 5\cdot 0 = 5 + 5\cdot 0.$$

Adding -5 to both sides shows that $0 = 5\cdot 0$.)

Division: Every real number r except zero has a **multiplicative inverse** denoted r^{-1} or $1/r$, such that $r\cdot r^{-1} = 1$. The multiplicative inverse is unique, by a similar argument to the one that we made for the additive inverse.

When we multiply r_1 and r_2^{-1}, we also say that we **divide** r_1 by r_2. For example, we may form the multiplicative inverse of 2 and multiply 5 by it, like this: $5\cdot(1/2)$. Equivalently, we say that we are "dividing 5 by 2" and write the operation like this: $5/2$. The notation $5/2$ is a shorthand for $5\cdot(1/2)$ or $5\cdot 2^{-1}$. Note that division consists of two operations: forming the multiplicative inverse and multiplying. Also note that division is not commutative: $5/2 \neq 2/5$. Nor is it associative: $(8/4)/2 \neq 8/(4/2)$. Finally, remember that the multiplicative inverse of zero (equivalently, division by zero) is not defined. Expressions such as 0^{-1} and $10/0$ have no meaning in the real number system.

Norm or absolute value: Each real number r has a **norm** or **absolute value** written $|r|$ and equal to r if $r \geq 0$, otherwise $-r$. The norm of a real number r measures the magnitude of r and disregards its sign. The norm has the following properties:

1. For any real number r, $|r| \geq 0$, and $|r| = 0$ if and only if $r = 0$.
2. For any two real numbers r_1 and r_2, $|r_1 r_2| = |r_1||r_2|$.
3. For any two real numbers r_1 and r_2, $|r_1 + r_2| \leq |r_1| + |r_2|$.

The last inequality is called the **triangle inequality**.

It is straightforward to verify each of these properties. The only one that is not entirely obvious is the triangle inequality. To verify that one, you can consider four cases: when neither r_1 nor r_2 is less than zero; when both are; when only r_1 is; and when only r_2 is. By commutativity, the last two cases are identical, so you only have to prove one of them.

2. Normed Vector Spaces

Now we turn our attention to normed vector spaces. These are the spaces in which we define the general derivative. In this paper we will focus on normed vector spaces over the real numbers **R**. Other normed vector spaces (for example, normed vector spaces over the complex numbers) are possible.

Vector spaces: A **vector space** V over **R** is a set of **vectors** satisfying the following rules.

1. **Vector addition:** We can add any two vectors v_1 and v_2 to form their sum, which is a vector denoted $v_1 + v_2$. The addition behaves like addition in **R**: it is associative and commutative; there is a vector 0 such that $v + 0 = 0 + v = v$ for all vectors v, and every vector v has a unique additive inverse $-v$ such that $v + (-v) = 0$.
2. **Scalar multiplication:** The real numbers are called the **scalars** of V. Given any scalar r and any vector v, we can form the product rv. We also write vr, and this has the same meaning as rv. This product satisfies the following rules:
    a. For all vectors v, $1v = v$.
    b. Scalar multiplication is associative in the sense that if r_1 and r_2 are scalars and v is a vector, then $r_1(r_2 v) = (r_1 r_2)v$.
    c. Scalar multiplication distributes over addition of vectors in the sense that if r_1 and r_2 are scalars and v_1 and v_2 are vectors, then $(r_1 + r_2)v = r_1 v + r_2 v$ and $r(v_1 + v_2) = rv_1 + rv_2$.

By the same argument we gave in the previous section, we can derive the fact that $0v = 0$ for all vectors v.

Normed vector spaces: We say that a vector space V is a **normed vector space** if each vector v in V has a norm $|v|$, where $|v|$ is a real number, and the norm satisfies the following properties:

1. For any vector v, $|v| \geq 0$, and $|v| = 0$ if and only if $v = 0$.
2. For any scalar r and vector v, $|rv| = |r||v|$.

3. For any two vectors v_1 and v_2, $|v_1 + v_2| \leq |v_1| + |v_2|$.

Notice the similarity to the properties stated in the previous section for the norm on $\mathbf{R}$.

Here are some examples of normed vector spaces:

Example 1: $\mathbf{R}$ is a normed vector space over itself. Vector addition is addition in $\mathbf{R}$. Scalar multiplication is multiplication in $\mathbf{R}$. The vector norm is the norm on $\mathbf{R}$ (i.e., the absolute value).

Example 2: Let $\mathbf{R}^2$ denote the set of pairs of real numbers (x, y). It is a normed vector space over $\mathbf{R}$. Vector addition is componentwise: $(x, y) + (x', y') = (x + x', y + y')$. Scalar multiplication is componentwise: $r(x, y) = (rx, ry)$. The norm is the standard Euclidean norm: $|(x, y)| = \sqrt{x^2 + y^2}$.

Example 3: Let $\mathbf{R}^n$ denote the set of n-tuples of real numbers $x = (x_1, \ldots, x_n)$. It is a normed vector space over $\mathbf{R}$. Vector addition and scalar multiplication are componentwise. The norm is the Euclidean norm $|x| = \sqrt{x_1^2 + \ldots + x_n^2}$.

Example 4: Let $\mathbf{R}^\infty$ denote the set of infinite sequences $x = x_1, x_2, \ldots$ of real numbers such that all but finitely many of the elements x_i are zero. It is a normed vector space over $\mathbf{R}$. Vector addition and scalar multiplication are componentwise. The norm $|x|$ is the square root of the sum of all the elements x_i^2. This sum is finite because all but finitely many of the x_i are zero.

Example 5: Consider the set of **bounded** functions f from $\mathbf{R}$ to $\mathbf{R}$, where bounded means that there is a fixed real number r called a **bound** for f such that $|f(x)| \leq r$ as x ranges over $\mathbf{R}$. This set is a normed vector space over $\mathbf{R}$:

- Vector addition is given by $f + g = x \mapsto f(x) + g(x)$. This notation says that the sum of the functions f and g is the function that takes x to $f(x) + g(x)$. For example, if $f(x) = x^2$ and $g(x) + x$, then $(f + g)(x) = x^2 + x$.
- The zero element is the **zero map** $x \mapsto 0$.
- The additive inverse of f is $x \mapsto -f(x)$. For example, the additive inverse of $f = x \mapsto x^2$ is $-f = x \mapsto -x^2$. We also write $f(x) = x^2$ and $(-f)(x) = -x^2$.
- Scalar multiplication is given by $rf = x \mapsto rf(x)$. For example, if $f(x) = x^2$, then $(3f)(x) = 3x^2$.
- The norm $|f|$ is given by the supremum of $|f(x)|$, written sup $|f(x)|$. This is the smallest number that is greater than or equal to every $|f(x)|$ for x in $\mathbf{R}$. This value exists if f is bounded. For example, $|\sin| = 1$.

Example 6: Let V be a normed vector space over $\mathbf{R}$, and let V^n denote the set of n-tuples of elements of V given by $v = (v_1, \ldots, v_n)$. It is a normed vector space over $\mathbf{R}$. Vector addition and multiplication are componentwise. The norm is the sup norm $|v| = \sup(|v_1|, \ldots, |v_n|)$. For example, let $V = \mathbf{R}^2$, and let $n = 2$. Then $V^n = (\mathbf{R}^2)^2$. One element of V is $v = ((1, 2), (3, 4))$. Its norm is $|v| = \sup(|(1, 2)|, |(3, 4)|) = \sup(\sqrt{1^2 + 2^2}, \sqrt{3^2 + 4^2}) = 5$.

Finite and infinite dimensions: Examples 1 through 3 are called **finite dimensional** vector spaces, because in each case, we can choose a finite number of vectors in the space and express every other vector as a sum of numbers times those vectors. For example, we may write every vector (x, y) in $\mathbf{R}^2$ as $x(1, 0) + y(0, 1)$. We say that $(1, 0)$ and $(0, 1)$ are **basis vectors** for $\mathbf{R}^2$. On the other hand, examples 4 and 5 are **infinite dimensional** vector spaces, because in each case there is no such finite set of basis vectors. Example 6 is finite dimensional if and only if V is.

It is possible to take derivatives in infinite dimensional vector spaces. However, the theory is a bit more complex than in the finite dimensional case. We will discuss this issue a bit more in § 10. Otherwise, for the rest of this document, we will assume that our vector spaces are finite dimensional. In fact, the theory of vector spaces tells us that a finite dimensional vector space V over $\mathbf{R}$ is **isomorphic** to $\mathbf{R}^n$ for some n. This means that there is a one-to-one structure-preserving map between V and $\mathbf{R}^n$. (A map is like a function, except that it takes vectors to vectors instead of numbers to numbers. In this case the structure-preserving map is a linear map; we will define this term in § 4.) For example, $(\mathbf{R}^2)^2$ is isomorphic to $\mathbf{R}^4$, via the mapping $((a, b), (c, d)) \mapsto (a, b, c, d)$. So in fact, "up to isomorphism," as mathematicians like to say, for the rest of this document we will always work over $\mathbf{R}^n$ for some n. However, we will still use letters such as V to denote vector spaces. That way we don't have to give an an explicit isomorphism to $\mathbf{R}^n$, even though we know there is one.

Multiple norms: In general, a vector space V may have more than one norm. For example, in examples 2, 3, and 4, we could have used the sup norm instead of the Euclidean norm. Therefore, when working with a normed vector space, we will need to specify which norm we mean.

3. Map, Limits, and Continuity

Maps: In first-year calculus, we take derivatives of functions $f(x)$, where x and $f(x)$ are numbers. Here we wish to extend this idea to **maps** (or mappings) $f(x)$ where x and $f(x)$ are vectors. A map is an association between points x of a vector space X and points $f(x)$ of a vector space Y. We write such a map $f: X \to Y$. In this document we will not require that f be defined on all points of X. Sometimes a map $f: X \to Y$ that is not defined on all points of X is called a **partial map**. The subset of X on which f is defined is called the **domain** of f.

We do not require that the vectors x in X be explicitly given as tuples of the form $(x_1, \ldots, x_n)$ for numbers or vectors $x_1, \ldots, x_n$. Further, where we do have $x = (x_1, \ldots, x_n)$, the notation $f(x)$ gives us the flexibility to treat x as one thing (a vector) or as a composition of several things (the elements x_i, also called the coordinates of the vector x). We treat the coordinates of Y similarly. We will explore this idea further in § 6 below.

To study maps of vector spaces, we need to define limits and continuity for these maps. This is not difficult. We shall see that these concepts carry through almost identically from numbers to vectors.

Limits: To define limits, we just need a concept of the distance between points. In $\mathbf{R}$ that distance is given by the absolute value $|x|$ of real numbers x. Two numbers x_1 and x_2 are close together if the norm or absolute value of their difference $|x_1 - x_2|$ is small. In a normed vector space, just as in $\mathbf{R}$, we have "subtraction" (really the addition of an additive inverse) and a norm, so we can do the same thing. We let the distance be given by $|x_1 - x_2|$, the norm of the difference between the vectors x_1 and x_2. In § 2, we defined the norm in such a way that arguments based on the absolute value (for example, the triangle inequality) carry through as arguments based on the norm.

Let X and Y be normed vector spaces. The **limit** of a map $f(x): X \to Y$ as x approaches x_0, written $\lim_{x \to x_0} f(x)$, is defined in the analogous way as it is when $X = Y = \mathbf{R}$. We say that the limit exists and is equal to y if the distance from $f(x)$ to y in Y can be made as small as desired by choosing x sufficiently close to x_0. For vector spaces, the distance is given by the norm. In symbols, we say $\lim_{x \to x_0} f(x) = y$ if for all $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ whenever $|x - x_0| < \delta$. This is the same definition as in the case of a single real variable, after replacing the absolute value with the more general concept of the norm. Just keep in mind that when we write expressions such as $|x - x_0|$, in general x and x_0 refer to vectors, not numbers.

Continuity: As for a single real variable, we say that a mapping $f(x): X \to Y$ is **continuous** at x_0 if $\lim_{x \to x_0} f(x) = f(x_0)$. We say that f is continuous on a set of points S of X if it is continuous at every point x_0 in S. We say that f is **uniformly continuous** on S if (a) it is continuous on S and (b) given ε in the definition of the continuity limit, we can choose a single δ that satisfies the definition at all points in S. In general this is not true. For example, the function $f(x) = 1/x$ is not uniformly continuous in the open interval $(0, 1)$ because as x approaches zero, the curve gets steeper and steeper, so for any fixed value of value of ε. we need to choose smaller and smaller values of δ.

Note on terminology: We avoid the terms "single variable" and "multivariable" that are sometimes used to distinguish first-year from more advanced calculus. In fact, even in the most advanced and general form of calculus, *a map $f(x)$ has one variable x*. It is very important to keep this fact in mind, because it greatly simplifies the theory. When generalizing calculus, we do not "add more variables." Instead, (1) we let x be a vector instead of just a number; and (2) we may (but are not required to) represent a vector as a collection of numbers. We use the term "single real variable" to refer to the first-year calculus case, in which the variable x refers to a single real number.

4. Linear Maps

In the general setting, the derivative is a special kind of map between vector spaces called a **linear map**. So next we discuss linear maps.

Let V and W be vector spaces. A linear map $\lambda: V \to W$ is a map from V to W that satisfies the following rules:

1. For all real numbers r and all vectors v in V, we have $\lambda(rv) = r\lambda(v)$.
2. For all vectors v_1 and v_2 in V, we have $\lambda(v_1 + v_2) = \lambda(v_1) + \lambda(v_2)$.

The set of linear maps from V to W forms a vector space, by the same argument that we used for Example 5 in § 2. We denote this vector space $L(V, W)$. It is a finite dimensional vector space over $\mathbf{R}$. In § 4.2, we will define the norm that makes $L(V, W)$ into a normed vector space.

4.1. Linear Products

In the special case of $V = W = \mathbf{R}$, each linear map corresponds to multiplication by a real number. That is, every element of $L(\mathbf{R}, \mathbf{R})$ is a function $\lambda(x) = rx$ for some real number r. Indeed, for any real number r', we have $\lambda(r') = \lambda(r'1) = r'\lambda(1)$, so $r = \lambda(1)$. Further, multiplication by a real number r is linear because it is associative and it distributes over addition. Thus we see that the product of real numbers gives a one-to-one linear map from $\mathbf{R}$ to $L(\mathbf{R}, \mathbf{R})$.

In general, when V and W are normed vector spaces, and $f: V \to W$ is a one-to-one linear map, we call f an **isomorphism**. The word "isomorphism" comes from the Greek words meaning "equal shape." If an isomorphism exists between V and W, then V and W have the same shape, in the sense that we can transform one into the other by a structure-preserving map.

Thus the product of numbers gives an isomorphism from $\mathbf{R}$ to $L(\mathbf{R}, \mathbf{R})$, the space of linear maps from $\mathbf{R}$ to $\mathbf{R}$. To characterize linear maps in vector spaces, we develop similar isomorphisms for suitably defined products of vectors. Let U, V, and W be vector spaces. We define a **linear product** to be an isomorphism P from U to $L(V, W)$. By definition, for each vector u in U, P associates linear map $P(u): V \to W$. Therefore P provides a way to interpret the vectors of U as linear maps from V to W. Equivalently, we may think of P as multiplying or combining a vector u in U and a vector v in V, yielding a vector $P(u)(v)$ in W. When the product P is clear from the context, we write $u \cdot v$ or uv instead of $P(u)(v)$.

For any vector space V, scalar multiplication is a linear product $P: \mathbf{R} \to L(V, V)$. In particular, multiplication in $\mathbf{R}$ is a linear product $P: \mathbf{R} \to L(\mathbf{R}, \mathbf{R})$. We use M to denote multiplication of real numbers. That is, $M(r_1)(r_2) = r_1 r_2$ means "multiply r_1 by r_2," and $M(r)$ means "multiplication by r."

Given a linear product $P: U \to L(V, W)$, we can use P to construct new linear products. There are three basic constructions. Each one is a kind of extended multiplication, in which one or both of the factors may have several elements, and we use P to combine the individual elements. We continue to use the letter M for this extended multiplication. We use subscripts to indicate the number of elements in each factor.

Multiplying one element by several elements: Let V^n be the vector space consisting of tuples $(v_1, \ldots, v_n)$ of vectors of V, as defined in Example 6 of § 2. For any $n > 0$, we define $M_{1n}(P): U \to (V^n, W^n)$ to be the linear product given by

$$u \cdot (v_1, \ldots, v_n) = (u \cdot v_1, \ldots, u \cdot v_n),$$

where the dot on the left represents $M_{1n}(P)$, and the dots on the right represent P. In other words, we use P to multiply u by each element v_i. Notice that when $V = W = \mathbf{R}$, $M_{1n}(M)$ is scalar multiplication in $\mathbf{R}^n$. More generally, when P is scalar multiplication on the left in V, $M_{1n}(P)$ is scalar multiplication on the left in V^n.

Multiplying several elements by one element: Let U^n be the vector space consisting of tuples $(u_1, \ldots, u_n)$ of vectors of U. For any $n > 0$, we define $M_{n1}(P): U^n \to L(V, W^n)$ to be the linear product given by

$$(u_1, \ldots, u_n) \cdot v = (u_1 \cdot v, \ldots u_n \cdot v),$$

where the dot on the left represents $M_{n1}(P)$, and the dots on the right represent P. In other words, we use P to multiply each element u_i by v. Notice that when P is scalar multiplication on the right in U, $M_{1n}(P)$ is scalar multiplication on the right in U^n.

Multiplying several elements by several elements: For any $n > 0$, we define $M_{nn}: U^n \to L(V^n, W)$ to be the linear product given by

$$(u_1, \ldots, u_n) \cdot (v_1, \ldots, v_n) = u_1 \cdot v_1 + \cdots + u_n \cdot v_n,$$

where the dot on the left represents $M_{nn}(P)$, the dots on the right represent P, and the plus signs represent addition in W. In other words, we add n terms, each of which is $P(u_i)(v_i)$.

Notice that the definitions overlap for $M_{11}(P)$, and they all agree in this case. Also, for any P, $M_{11}(P) = P$. In particular, $M_{11}(M) = M$.

It is a straightforward exercise to verify that each of the products $M_{ij}(P)$ is an isomorphism.

By starting with real multiplication M and iteratively applying the product constructors M_{ij}, we can multiply different kinds of vectors, and we can represent linear maps as vectors. To keep the notation tidy, we write $M_{ab\,cd}(P)$ instead of $M_{ab}(M_{cd}(P))$. Also, we omit the argument P when there is only one way to multiply the elements in

question. For example, in $\mathbf{R}^n$ we write M_{nn} instead of $M_{nn}(M)$.

Example 1 (The dot product): In $\mathbf{R}^n$, the linear product M_{nn} is called the **scalar product** or **dot product**. For example, $(1,2)\cdot(3,4) = 1\cdot 3 + 2\cdot 4 = 11$, where the dot between the vectors means M_{22}, and the dot between the numbers means M. M_{nn} interprets any vector in $\mathbf{R}^n$ as a linear map from $\mathbf{R}^n$ to $\mathbf{R}$. Equivalently, it multiplies two vectors in $\mathbf{R}^n$, yielding a number in $\mathbf{R}$.

Example 2 (Multiplying matrices by vectors): $M_{n1\ mm}$ interprets vectors in $(\mathbf{R}^m)^n$ as linear maps from $\mathbf{R}^m$ to $\mathbf{R}^n$. Equivalently, it multiplies vectors u in $(\mathbf{R}^m)^n$ by vectors v in $\mathbf{R}^m$, yielding vectors w in $\mathbf{R}^n$. We can read these facts from the notation $M_{n1\ mm}$:

1. From the left elements in the subscripts, reading right to left, u is an element of $(\mathbf{R}^m)^n$.
2. From the right elements in the subscripts, again right to left, v is an element of $(\mathbf{R}^m)^1 = \mathbf{R}^m$.
3. Combining m elements with m elements yields one element, and combining n elements with one element yields n elements. So w is an element of $(\mathbf{R}^1)^n = \mathbf{R}^n$.

To illustrate, we use $M_{21\ 33}$ to multiply the vector $((0,1,2),(1,2,3))$ in $(\mathbf{R}^3)^2$ by the vector $(3,4,5)$ in $\mathbf{R}^3$, yielding a vector in $\mathbf{R}^2$. We start with the product

$$((0,1,2),(1,2,3))\cdot(3,4,5),$$

where the dot means $M_{21\ 33}$. In English, that says, "Treat the left factor as two elements, treat the right factor as one element, and when multiplying elements, use M_{33}." By definition this is

$$((0,1,2)\cdot(3,4,5),(1,2,3)\cdot(3,4,5)),$$

where the dot means M_{33}. By definition this is

$$(0\cdot 3 + 1\cdot 4 + 2\cdot 5, 1\cdot 3 + 2\cdot 4 + 3\cdot 5),$$

where the dot means M. So the answer is $(14,26)$.

It quickly becomes unwieldy to write out vectors of vectors using commas and parentheses. Therefore vectors in $(\mathbf{R}^m)^n$ are traditionally written as **matrices** or grids of numbers, with n rows of m columns each. For example, the vector $((0,1,2),(1,2,3))$ appearing as the left factor of a multiplication is usually written as a matrix

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

When multiplying matrices by vectors according to the formula $uv = w$, we interpret u as a column of row vectors and each of v and w as a single column vector. So the product

$$((0,1,2),(1,2,3))\cdot(3,4,5) = (14,26)$$

is usually written

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 14 \\ 26 \end{bmatrix}.$$

Example 3 (Multiplying matrices by matrices): $M_{1o\ n1\ mm}$ interprets vectors in $(\mathbf{R}^m)^n$ as linear maps from $(\mathbf{R}^m)^o$ to $(\mathbf{R}^n)^o$, for $o \geq 1$. As an example, let us use $M_{12\ 21\ 33}$ to evaluate

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 3 & 5 \\ 4 & 6 \\ 5 & 7 \end{bmatrix}.$$

$M_{12\ 21\ 33}$ says that we treat the right factor as two elements and multiply the left factor by each one. According to the conventions of matrix notation, the elements of the right factor are the columns. So this gives

$$\left[\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix}\right].$$

At the outer level there are now two elements. Applying $M_{21\,33}$ at each element as we did in Example 2 yields

$$\begin{bmatrix} 14 & 20 \\ 26 & 38 \end{bmatrix},$$

This process is called **matrix multiplication**. In general, when carrying out a matrix multiplication $uv = w$,

1. We may represent u as a column of rows u_i, where i is the row number.
2. We may represent v a row of columns v_j, where j is the column number.
3. We may represent w as a collection of numbers w_{ij}, where i is the row number and j is the column number.

A handy rule for carrying out the multiplication is that each number w_{ij} is the dot product of the row vector u_i and the column vector v_j. For example:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}[v_1 \; v_2] = \begin{bmatrix} u_1 \cdot v_1 & u_1 \cdot v_2 \\ u_2 \cdot v_1 & u_2 \cdot v_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}.$$

Example 4 (Matrices of vectors): The product $M_{21\,22\,22}$ multiplies vectors in $((\mathbf{R}^2)^2)^2$ by vectors in $(\mathbf{R}^2)^2$, yielding vectors in $\mathbf{R}^2$. For example, let u be the matrix of row vectors

$$u = \begin{bmatrix} (1,2) & (3,4) \\ (5,6) & (7,8) \end{bmatrix}.$$

To compute $M_{21\,22\,22}(u)(v)$, we use matrix multiplication as for $M_{21\,22}$, except that when multiplying each row of the matrix u by the column v, we compute a sum of dot products. For example,

$$\begin{bmatrix} (1,2) & (3,4) \\ (5,6) & (7,8) \end{bmatrix} \begin{bmatrix} (9,10) \\ (11,12) \end{bmatrix} = \begin{bmatrix} (1,2)\cdot(9,10) + (3,4)\cdot(11,12) \\ (5,6)\cdot(9,10) + (7,8)\cdot(11,12) \end{bmatrix}$$

Similarly it is possible to construct matrices of matrices, etc., and interpret them as linear maps.

4.2. The Norm of $L(V, W)$

As observed above, $L(V, W)$ is a vector space. We make it into a normed vector space as follows.

First, consider $L(\mathbf{R}, \mathbf{R})$. As noted above, a map λ in this space is a product $M(r)$ by some real number r. In this space, it is natural to use the norm $|\lambda| = |M(r)| = |r|$. For example, the norm of the linear map "multiplication by –5" is 5.

In the more general context, we let $|\lambda|$ be the supremum of all values $|\lambda(x)|$ such that $|x| \le 1$. For example, let $\lambda\colon \mathbf{R}^2 \to \mathbf{R}$ be the dot product with $(1, 1)$. Then for any $x = (x_1, x_2)$, we have $\lambda(x) = x_1 + x_2$. If we use the sup norm on $\mathbf{R}^2$, then $|\lambda|$ is the maximum value of $|x_1 + x_2|$ subject to the constraint that $|x_1| \le 1$ and $|x_2| \le 1$. Therefore $|\lambda| = 2$. (Notice that it is convenient to use the sup norm on $\mathbf{R}^2$ here. If we use the Euclidean norm, then computing $|\lambda|$ is a nontrivial calculus problem.)

Observe that this norm agrees with the norm discussed above for $L(\mathbf{R}, \mathbf{R})$, because $|M(r)(x)| = |rx| = |r||x|$ attains its maximum value for $|x| \le 1$ when $|x| = 1$, so $|M(r)| = |r|$.

Observe also that for any linear map λ in $L(V, W)$ and any x in V, we have $|\lambda(x)| = |\ \lambda(x/|x|)\ |x|\ | = |\lambda(x/|x|)|\ |x|$. Since $|x/|x|| = |x|/|x| = 1$, the first factor is bounded by $|\lambda|$. Therefore for any linear map λ we have

$$|\lambda(x)| \le |\lambda||x|.$$

5. The Definition of the Derivative

Now we can define the general derivative.

Definition for one real variable: We recall the definition of the derivative from first-year calculus. Given a function $f\colon \mathbf{R} \to \mathbf{R}$, the **derivative** of f, denoted $f'(x)$ or $\dfrac{df}{dx}$, is a function that assigns to a point x the real number

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

if f is defined for all points within some fixed distance of x (so we exclude isolated points in the domain of f), and the limit exists. At points x where these conditions are true, we say that f is **differentiable**.

General definition: It turns out that more or less the same definition works for an arbitrary normed vector space. However, to make it work we need to revise it slightly. First, we rewrite equation (1) as follows:

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - f'(x)h}{|h|} = 0. \tag{2}$$

It is clear that (2) holds if and only if (1) does. All we have done is rearranged terms and replaced h with $|h|$; in general this alters the sign but does not affect the magnitude of the expression in the limit as h approaches zero. Next, we define a **differential operator** D. Given a function $f: \mathbf{R} \to \mathbf{R}$, we let $Df: \mathbf{R} \to L(\mathbf{R}, \mathbf{R})$ be the map $x \mapsto h \mapsto f'(x)h$, where $f'(x)$ is the ordinary one-variable derivative at x. In other words, for each f, x, and h, $Df(x)(h) = f'(x)h$. Notice that $Df(x) = M(f'(x))$, where M is the linear product "multiply" defined in § 4.1.

With this definition, we can rewrite (2) as follows:

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - Df(x)(h)}{|h|} = 0. \tag{3}$$

Equivalently, we can write

$$f(x+h) = f(x) + Df(x)(h) + \phi(h), \tag{4}$$

where ϕ is a function that goes to zero faster than $|h|$, in the sense that $\lim_{h \to 0} \frac{\phi(h)}{|h|} = 0$. When this is true, we say that ϕ is $o(h)$, pronounced "little oh of h." To convert (4) into (3), just move every term on the right but $\phi(h)$ to the left, divide by $|h|$, and take the limit. Equation (3) emphasizes the connection with the derivative for one real variable, while equation (4) emphasizes that the derivative $Df(x)$ is a linear map that approximates $f(x)$ near x. Equation (4) is the one we usually use in proofs.

Now we can generalize to maps $f: X \to Y$, where X and Y are normed vector spaces. We generalize the one-variable differential operator D in an obvious way: we let $Df(x)$ be an element of $L(X, Y)$. So in general, for a map $f: X \to Y$, if the derivative Df exists at x, then $Df(x)$ is a linear map from X to Y, and Df is a map from X to $L(X, Y)$.

In general we must represent $Df(x)$ as a vector v instead of a number r. We can think of $v = \lambda$ as an element of the space $L(X, Y)$; or, if we can represent the map λ as a vector or matrix of numbers (or of vectors or matrices) as discussed in § 4.1, then we can think of v as an element of one of these spaces. Thus we generalize the derivative from first-year calculus, in which we consider only $Df(x) = M(f'(x))$ represented by a single number $f'(x)$ for each x.

Now in either form (3) or form (4), the definition of the derivative is valid for normed vector spaces X and Y. To apply the definition, we let x and h be vectors in X instead of numbers in $\mathbf{R}$. By § 2, we know how to take the norm $|h|$ of vectors h. By § 3, we know how to compute limits of expressions that take their values as vectors y in Y.

Uniqueness: If the derivative Df of a map f exists at a point x (i.e., a vector of X), then it is unique there. In other words, there is at most one linear map that satisfies equations (3) and (4) at any point x.

To see this, suppose that linear maps λ_1 and λ_2 both satisfy (4) at a point x, and consider the difference map $\lambda = \lambda_1 - \lambda_2 = h \mapsto \lambda_1(h) - \lambda_2(h)$. We want to show that $\lambda(h) = 0$ for all h. Since λ is linear, we have $\lambda(0) = \lambda(0 \cdot 0) = 0 \cdot \lambda(0) = 0$, so $\lambda(0) = 0$. Further, for any vector $h \neq 0$, we have $\lambda(h) = \lambda(|h|(h/|h|)) = |h|\lambda(h/|h|)$, where $h/|h|$ has norm 1. So it suffices to show that $\lambda(h_0) = 0$ whenever h_0 has norm 1.

Choose any vector h_0 with norm 1, and let $h = rh_0$ for a real variable r that goes to zero. Then h goes to zero. Each of λ_1 and λ_2 satisfies (4), and subtracting the two equations shows that $\lambda(h)$ is $o(h)$. Now consider $\phi(h) = |\lambda(h)| = |\lambda(rh_0)| = |r\lambda(h_0)| = |r||\lambda(h_0)|$. Since $\phi(h)$ is $o(h)$ and $|r| = |h|$, we have that $\phi(h)/|r| = |\lambda(h_0)|$ goes to zero as r goes to zero. But h_0 is a constant. Therefore we must have $\lambda(h_0) = 0$.

This argument shows that we are justified in speaking of "the" derivative of f at x. Also, if we can find a linear map λ that satisfies the required properties at a point x, then we know that $\lambda = Df(x)$.

Continuity: Taking the limit as h goes to zero in both sides of (4) yields

$$\lim_{h \to 0} f(x+h) = f(x) + \lim_{h \to 0} Df(x)(h) + \lim_{h \to 0} \phi(h).$$

By definition, the second limit on the right-hand side is zero. As to the first limit, we have $|Df(x)(h)| \le |Df(x)||h|$, so it too is zero. Therefore we have

$$\lim_{h \to 0} f(x+h) = f(x)$$

This equation shows that if f is differentiable at x, then it is continuous at x.

Notation: Some authors write f' interchangeably with Df. Here we reserve f' for the special case of a single real variable, so it is clear when we are treating that case. Keep in mind that in the case of a single real variable, in our notation we always have $Df(x) = M(f'(x))$.

6. Coordinate Systems

If $X_1, \ldots, X_n$ are normed vector spaces, then the vector space $X = X_1 \times \cdots \times X_n$ consisting of tuples $x = (x_1, \ldots, x_n)$ with each x_i in X_i is also a normed vector space. This construction is a straightforward generalization of the vector space $V^n = V \times \cdots \times V$ (n times) that we discussed in Example 6 of § 2. The only difference is that X may be composed of several different vector spaces, instead of several copies of the same vector space.

If $X = X_1, \ldots, X_n$ with $n > 1$, then we say that X is a **coordinate system**, and we call the vectors x_i the **coordinates** of a point x in X. In particular, $\mathbf{R}^n = \mathbf{R} \times \cdots \times \mathbf{R}$ (n times) for $n > 1$ is a coordinate system in which each coordinate x_i of a point x is a number. However, we are not limited to considering $\mathbf{R}^n$. For example, we can let $X = (\mathbf{R}^2)^2$, consisting of all vectors $((x_{11}, x_{12}), (x_{21}, x_{22}))$, where the elements x_{ij} are real numbers.

In this section, we show how to compute derivatives in coordinate systems. We will assume that our coordinate systems use either the Euclidean norm (on $\mathbf{R}^n$) or the sup norm. These norms give us the following important fact. If h is a vector and $h = (h_1, \ldots, h_n)$, and for some h_i $\phi(h_i)$ is $o(h_i)$, then $\phi(h_i)$ is $o(h)$. The reason is that $|h| \ge |h_i|$, so if $\phi(h_i)/|h_i|$ tends to zero, then so does $\phi(h_i)/|h|$, because the values of the second expression are at least as small as the values of the first expression.

In the rest of this section, we let X and Y be normed vector spaces, and we let f be a map from X to Y. We consider three cases: when Y is a coordinate system, when X is a coordinate system, and when both X and Y are coordinate systems.

6.1. When Y Is a Coordinate System

We let Y be a two-dimensional coordinate system, i.e., we consider maps $f\colon X \to Y_1 \times Y_2$ for normed vector spaces X, Y_1, and Y_2. Once we have worked out the two-dimensional case, it will be obvious how to proceed in any number of dimensions. Working in two dimensions simplifies the notation.

The first step is to break f into two maps, one for each of the coordinates in Y. By definition, f assigns to each vector x in X a value $f(x) = y = (y_1, y_2)$. If we look at just the first y coordinate, we get a map $f_1(x) = y_1$; and if we look at just the second y coordinate, we get a map $f_2(x) = y_2$. We call f_1 and f_2 the **coordinate maps** of f. Then by definition, for all x we may write

$$f(x) = (f_1(x), f_2(x)).$$

If f_1 and f_2 are differentiable at x, then by the definition of the derivative, we have

$$f(x+h) = (\ f_1(x+h), f_2(x+h)\ ) = (\ f_1(x) + Df_1(x)(h) + \phi_1(h), f_2(x) + Df_2(x)(h) + \phi_2(h)\ )$$

$$= (\ f_1(x), f_2(x)\ ) + (\ Df_1(x)(h), Df_2(x)(h)\ ) + (\ \phi_1(h), \phi_2(h)\ ),$$

where each ϕ_i is $o(h)$. Setting $\phi(h) = (\phi_1(h), \phi_2(h))$, we can write

$$f(x+h) = f(x) + (\ Df_1(x)(h), Df_2(x)(h)\ ) + \phi(h),$$

where ϕ is $o(h)$ because the ϕ_i are. Applying the definition of the derivative again, we see that $Df(x)$ is the linear map from X to $Y_1 \times Y_2$ whose ith coordinate map is $Df_i(x)$. In other words,

$$Df(x)(h) = (Df_1(x)(h), Df_2(x)(h)).$$

Conversely, a map is linear if and only if its coordinate maps are linear. Therefore if f is differentiable at x, then we have

$$Df(x)(h) = (\lambda_1(h), \lambda_2(h))$$

for linear coordinate maps λ_1 and λ_2, and we can reverse the argument to show that $\lambda_1 = Df_1(x)$ and $\lambda_2 = Df_2(x)$.

By the discussion above, when Y is an n-dimensional coordinate system, we can represent $Df(x)$ as the linear product

$$Df(x) = M_{n1}(Df_1(x), \ldots, Df_n(x)), \tag{1}$$

where $M_{n1}(\lambda_1, \ldots, \lambda_n)(h) = (\lambda_1(h), \ldots, \lambda_n(h))$. In § 4.1, we required that the elements in the left factor of M_{n1} all be in the same vector space. However, that restriction was not necessary, and now we drop it.

Example: Let $X = \mathbf{R}$, $Y = \mathbf{R}^3$, and $f(x) = (x, x^2, x^3)$. Then $Df_1(x) = M(1)$, $Df_2(x) = M(2x)$, and $Df_3(x) = M(3x^2)$. So

$$Df(x)(h) = (M(1)(h), M(2x)(h), M(3x^2)(h)) = (h, 2xh, 3x^2h).$$

In second-year calculus, we are taught that for a function from numbers to vectors, we may compute the derivative coordinate by coordinate, so the derivative is $(1, 2x, 3x^2)$. Thus our derivative $Df(x)$ is the linear map that takes the number h to h times the vector representing the derivative from second-year calculus.

6.2. When X Is a Coordinate System

We let $X = X_1 \times X_2$ and consider maps $f\colon X_1 \times X_2 \to Y$, for normed vector spaces X_1, X_2, and Y. Again, once we cover the two-dimensional case, the generalization to n dimensions will be obvious. Each point x in X may be expressed in coordinates as $x = (x_1, x_2)$. We write $f(x_1, x_2)$ to denote $f(x) = f((x_1, x_2))$.

First, assume that f is differentiable at x. Then there exists a linear map $Df(x)$ such that for all $h = (h_1, h_2)$, we have

$$Df(x)(h) = Df(x)(h_1, h_2) = Df(x)(h_1, 0) + Df(x)(0, h_2) \tag{2}$$

because $Df(x)$ is linear. We define $D_1 f(x)(h_1) = Df(x)(h_1, 0)$ and $D_2 f(x)(h_2) = Df(x)(0, h_2)$. For each i, the map $D_i f$ is called the ith **partial derivative** of f. An alternate notation for the ith partial derivative of f is $\dfrac{\partial f}{\partial x_i}$.

Equation (2) shows that

$$Df(x)(h) = D_1 f(x)(h_1) + D_2 f(x)(h_2).$$

Therefore, to compute $Df(x)$, it suffices to compute the partial derivatives $D_i f(x)$. Further, if we let $f_{(-,x_2)}\colon X_1 \to Y$ be the mapping that fixes the second coordinate at x_2 and takes x_1 to $f(x_1, x_2)$ and we let $h = (h_1, 0)$, then we can rewrite

$$f(x + h) = f(x) + Df(x)(h) + \phi(h)$$

where ϕ is $o(h)$ as

$$f_{(-,x_2)}(x_1 + h_1) = f_{(-,x_2)}(x_1) + D_1 f(x)(h_1) + \phi_1(h_1), \tag{3}$$

where ϕ_1 is $o(h_1)$. Equation (3) shows that $D_1 f(x)$ is the derivative in one real variable with respect to x_1 of $f_{(-,x_2)}$. Similarly, $D_2 f(x)$ is the derivative in one real variable with respect to x_2 of $f_{(x_1,-)}$. Therefore, when $X_i = Y = \mathbf{R}$, we can use first-year calculus techniques to compute partial derivatives. We give an example below.

Conversely, we show that if the partial derivatives $D_i f(x)$ exist, and all but possibly one are continuous for all points $x + h$ for small enough $|h|$, then f is differentiable at x. To see this, assume that both partial derivatives exist and $D_1 f$ is continuous near x, and look at the difference

$$d(h) = f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2) = f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2 + h_2) + f(x_1, x_2 + h_2) - f(x_1, x_2).$$

By the definition of the derivative, we want $d(h)$ to be equal to $Df(x)$ plus a term that is $o(h)$. For small enough h_2 that the partial derivatives exist, we have

$$d(h) = D_1 f(x_1, x_2 + h_2)(h_1) + \phi_1(h_1) + D_2(x)(h_2) + \phi_2(h_2),$$

where each ϕ_i is $o(h_i)$. To get what we want, we just need to show that the difference between the first term on the right and $D_1 f(x)(h_1)$ is $o(h)$. Let $\lambda(h_2) = D_1 f(x_1, x_2 + h_2) - D_1(x)$. Then it will suffice to show that $|\lambda(h_2)(h_1)|$ is $o(h)$. By the argument given in § 4.2, we have $|\lambda(h_2)(h_1)| \leq |\lambda(h_2)\|h_1|$. Because $D_1 f$ is continuous, $\lim_{h_2 \to 0} |\lambda(h_2)| = |\lambda(0)| = 0$. Therefore $\lambda(h_2)(h_1)$ is $o(h)$, as was to be shown.

By the discussion above, when X is an n-dimensional coordinate system, we can represent $Df(x)$ as the linear product

$$Df(x) = M_{nn}(D_1 f(x), \ldots, D_n f(x)), \tag{4}$$

where $M_{nn}(\lambda_1, \ldots, \lambda_n)(h_1, \ldots, h_n) = \lambda_1(h_1) + \ldots + \lambda_n(h_n)$.

Example: Let $f(x_1, x_2) = x_1^2 x_2$. Then $D_1 f(x_1, x_2) = M(2x_1 x_2)$ and $D_2 f(x_1, x_2) = M(x_1^2)$. In this case,

$$Df(x_1, x_2)(h_1, h_2) = M(2x_1 x_2)(h_1) + M(x_1^2)(h_2) = 2x_1 x_2 h_1 + x_1^2 h_2.$$

In second-year calculus, we are taught that for a function from vectors to numbers, the **gradient** of f, written grad f or ∇f, is equal to $\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right) = (2x_1 x_2, x_1^2)$. Thus our derivative $Df(x)$ is the map that takes the vector h to the number $\nabla f \cdot h$.

6.3. When X and Y Are Coordinate Systems

We consider maps $f: X_1 \times X_2 \to Y_1 \times Y_2$. To handle this case, we just compose the two cases we have already considered. There are two orders in which one can do this.

Order 1: We can treat f as a map $X \to Y_1 \times Y_2$, ignoring the fact that $X = X_1 \times X_2$ is a coordinate system. By § 6.1, we have

$$Df(x) = M_{21}(Df_1(x), Df_2(x)), \tag{5}$$

Next we apply § 6.2 to each of the coordinate map $f_1: X_1 \times X_2 \to Y_1$ and $f_2: X_1 \times X_2 \to Y_2$ to obtain

$$Df_1(x) = M_{22}(D_1 f_1(x), D_2 f_1(x)) \tag{6}$$

$$Df_2(x) = M_{22}(D_1 f_2(x), D_2 f_2(x)),$$

Putting together equations (5) and (6) yields

$$Df(x) = M_{21}(M_{22}(D_1 f_1(x), D_2 f_1(x)), M_{22}(D_1 f_2(x), D_2 f_2(x))). \tag{7}$$

Writing (7) in matrix form yields

$$\begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) \end{bmatrix}. \tag{8}$$

Note that using matrix multiplication to apply (8) agrees with the way that we evaluate the linear products in (7).

In the case where $X = \mathbf{R}^m$ and $Y = \mathbf{R}^n$, so the linear maps $D_i f_j(x)$ are real numbers, (8) is called the **Jacobian matrix** corresponding to the derivative of f at x. Using the alternate notion for partial derivatives, it is written this way:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}$$

Extending this computation to higher dimensions is straightforward. In general we get an $m \times n$ matrix of partial derivatives. The only real difficulty is to remember which are the rows and which are the columns. You just have to remember that the partial derivatives vary within a row, and the coordinate maps vary within a column.

Order 2: Alternatively, we can also go in the reverse order, treating f first as a map $X_1 \times X_2 \to Y$. In this case we get

$$Df(x) = M_{22}(D_1 f(x), D_2 f(x)).$$

Computing each partial derivative yields

$$Df(x) = M_{22}(M_{21}(D_1 f_1(x), D_1 f_2(x)), M_{21}(D_2 f_1(x), D_2 f_2(x))).$$

Writing this in matrix form as a column of rows yields the **transpose** of the matrix (8). In general, the transpose of a matrix A is A with the rows and columns interchanged. The order of M_{22} and M_{21} are also swapped compared to (8), indicating that if we want to apply this vector by standard matrix multiplication, we need to take its transpose first.

In practice, we use order 1 because it works better given the conventions of matrix multiplication. Either way is fine in principle, though, as long as one is careful about translating vectors into linear maps.

Example 1: Let $f: \mathbf{R}^2 \to \mathbf{R}^2$ be the map $f(x_1, x_2) = (x_1^2 + x_2, x_1 x_2)$. Computing the Jacobian matrix and applying it to $h = (h_1, h_2)$ yields

$$\begin{bmatrix} 2x_1 & 1 \\ x_2 & x_1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} 2x_1 h_1 + h_2 \\ x_2 h_1 + x_1 h_2 \end{bmatrix}.$$

Thus $Df(x)$ is the linear map $h \mapsto (2x_1 h_1 + h_2, x_2 h_1 + x_1 h_2)$. We can also write the product as

$$\begin{bmatrix} \nabla f_1(x) \\ \nabla f_2(x) \end{bmatrix} [\ h\ ] = \begin{bmatrix} \nabla f_1(x) \cdot h \\ \nabla f_2(x) \cdot h \end{bmatrix}.$$

Example 2: Let $f: (\mathbf{R}^2)^2 \to (\mathbf{R}^2)^2$ be the map $((a, b), (c, d)) \mapsto ((ab, bc), (cd, da))$. Then $Df(x)$ at a point $x = (a, b, c, d)$ is a 2×2 matrix of 2×2 matrices. At the outer level, the leftmost, topmost matrix element is the derivative of $f_{(-,(b,c))} = (a, b) \mapsto (ab, bc)$, treating b and c as constants. This derivative is

$$\begin{bmatrix} b & a \\ 0 & c \end{bmatrix}$$

As an exercise, you can work out the rest.

Example 3: $(\mathbf{R}^2)^2$ is isomorphic to $\mathbf{R}^4$ via the isomorphism $((a, b), (c, d) \mapsto (a, b, c, d)$. Work out $Df(x)$ at $x = (a, b, c, d)$ for the map $f: \mathbf{R}^4 \to \mathbf{R}^4$ given by $(a, b, c, d) \mapsto (ab, bc, cd, da)$. It is a 4×4 matrix of numbers. Compare this matrix to the $(2 \times 2) \times (2 \times 2)$ matrix that you got for Example 3.

7. Properties of the Derivative

We now establish some basic properties of the derivative in the general setting. Each one generalizes a property from first-year calculus in a straightforward way. In this section, X, Y, and Z are normed vector spaces.

7.1. The Derivative of a Constant Map

Let $f: X \to Y$ be the constant map $f(x) = y$. Then for all values of x and h we have

$$f(x + h) = f(x)$$

Therefore $Df(x)(h) = 0$, so $Df(x)$ is the zero map $h \mapsto 0$:

$$Df = 0$$

7.2. The Derivative of a Linear Map

Let $\lambda: X \to Y$ be a linear map. We assert

$$D\lambda = \lambda.$$

In other words, for all x and h in X, we assert $D\lambda(x)(h) = \lambda(h)$; in particular the map $D\lambda(x)$ does not depend on x.

This rule generalizes the rule from first-year calculus that $(rx)' = r$.

Proof: The assertion follows immediately from the definition of the derivative, because

$$\lambda(x + h) - \lambda(x) = \lambda(x + h - x) = \lambda(h).$$

Therefore the $Df(x)(h)$ term in the definition is $\lambda(h)$, and the $\phi(h)$ term is zero.

7.3. The Derivative of a Sum

For any maps $f: X \to Y$ and $g: X \to Y$, we assert

$$D(f+g) = Df + Dg$$

at all points x where $Df(x)$ and $Dg(x)$ exist. As usual,

$$f + g = x \mapsto f(x) + g(x)$$

and

$$Df + Dg = x \mapsto h \mapsto (Df(x) + Dg(x))(h) = x \mapsto h \mapsto Df(x)(h) + Dg(x)(h).$$

This rule generalizes the rule from first-year calculus that $(f+g)'(x) = f'(x) + g'(x)$.

Proof: To justify the assertion, we start with the difference $(f+g)(x+h) - (f+g)(x)$. By the definition of $f+g$ and by rearranging terms, we can write this as

$$f(x+h) - f(x) + g(x+h) - g(x).$$

If the derivatives of f and g exist at x, then we can rewrite the difference as

$$Df(x)(h) + Dg(x)(h) + \phi(h),$$

where $\phi(h)$ is $o(h)$. By the definition of $Df(x) + Dg(x)$, this gives

$$(f+g)(x+h) - (f+g)(x) = (Df(x) + Dg(x))(h) + \phi(h).$$

After rearranging terms, this gives us what we want.

7.4. The Derivative of a Product

Let X, Y_1, Y_2, and Y be normed vector spaces. We define a **bilinear map** $m: Y_1 \times Y_2 \to Y$ to be a map such that

1. For all y_2 in Y_2, the map $y_1 \mapsto m(y_1, y_2)$ is linear in y_1; and
2. For all y_1 in Y_1, the map $y_2 \mapsto m(y_1, y_2)$ is linear in y_2.

Note that a linear product $P: Y_1 \to L(Y_2, Y)$ induces a bilinear map $m: Y_1 \times Y_2 \to Y$ given by $m(y_1, y_2) = P(y_1)(y_2)$.

Fix a bilinear map m and maps $f: X \to Y_1$ and $g: X \to Y_2$. Let $m(f, g)$ denote the map $x \mapsto m(f(x), g(x))$. Then we assert the following rule for all x where $Df(x)$ and $Dg(x)$ exist:

$$D(m(f,g))(x)(h) = m(Df(x)(h), g(x)) + m(f(x), Dg(x)(h)). \tag{1}$$

If we interpret $g(x)$ as the constant map $h \mapsto g(x)$ and $f(x)$ as the constant map $h \mapsto f(x)$, then we can write the rule as follows:

$$D(m(f,g))(x)(h) = m(Df(x), g(x))(h) + m(f(x), Dg(x))(h)$$

Using the rule for a sum of maps, we can then write

$$D(m(f,g))(x)(h) = (m(Df(x), g(x)) + m(f(x), Dg(x)))(h)$$

or

$$D(m(f,g))(x) = m(Df(x), g(x)) + m(f(x), Dg(x)).$$

If we write $m(f,g)$ as the product fg and $m(y_1, y_2)$ as the product $y_1 y_2$, then the rule becomes

$$D(fg)(x) = Df(x)g(x) + f(x)Dg(x).$$

According to our notation for sums and products of functions, this is

$$D(fg)(x) = (Dfg + fDg)(x)$$

or

$$D(fg) = (Df)g + f(Dg), \tag{2}$$

where as discussed above we interpret (2) to mean (1). In form (2), the rule generalizes the rule from first-year calculus that

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

Proof: To justify the rule, consider the expression

$$(fg)(x+h) = f(x+h)g(x+h).$$

If the derivatives exist at x, then this expands into

$$(f(x) + Df(x)(h) + \phi_1(h))(g(x) + Dg(x)(h) + \phi_2(h)).$$

where ϕ_1 and ϕ_2 are $o(h)$. Because the product is bilinear, we can multiply the factors term by term. Doing this generates the following terms:

1. $(fg)(x) + (Df(x)(h))g(x) + f(x)(Dg(x)(h))$
2. $f(x)\phi_2(h) + Df(x)(h)\phi_2(h) + \phi_1(h)\phi_2(h) + \phi_1(h)g(x) + \phi_1(h)Dg(x)(h) + \phi_1(h)\phi_2(h)$
3. $(Df(x)(h))(Dg(x)(h))$

The terms in (1) are the initial terms in the definition of the derivative. We just need to show that each of the terms in (2) and (3) is $o(h)$.

As to (2), we observe that each term is a product $\psi_1(h)\psi_2(h)$, where each factor ψ_i is a mapping $X \to Y_i$, each $|\psi_i(h)|$ is bounded by a constant C_i for small $|h|$, and at least one of the ψ_i is $o(h)$. This statement is clear for the factors involving f, g, and ϕ_i. For the factor $Df(x)(h)$ we have $|Df(x)(h)| \leq |Df(x)\|h|$, so the factor is bounded by the constant $|Df(x)|$ for $|h| \leq 1$; and similarly for the factor $Dg(x)(h)$.

Assume that $\psi_2(h)$ is $o(h)$. Then, writing m for the linear map $Y_1 \to L(Y_2, Y)$ given by $y_1 \mapsto (y_2 \mapsto m(y_1, y_2))$, we have

$$\begin{aligned}
\psi_1(h)\psi_2(h) &= m(\psi_1(h))(\psi_2(h)) \\
&\leq |m(\psi_1(h))\|\psi_2(h)| \\
&\leq |m\|\psi_1(h)\|\psi_2(h)| \\
&\leq |m|C_1|\psi_2(h)|
\end{aligned}$$

This expression is $o(h)$ because $\psi_2(h)$ is. The symmetric argument goes through when ψ_1 is $o(h)$.

As to (3), we have

$$\begin{aligned}
(Df(x)(h))(Dg(x)(h)) &= m(Df(x)(h))(Dg(x)(h)) \\
&\leq |m\|Df(x)(h)\|Dg(x)(h)| \\
&\leq |m\|Df(x)\|h\|Dg(x)\|h|
\end{aligned}$$

When we divide by $|h|$, one of the $|h|$ factors cancels, but the other one remains. So after dividing by $|h|$ the expression tends to zero.

7.5. The Chain Rule

Composition of functions: Let $f: X \to Y$ and $g: Y \to Z$ be maps. We write $g \circ f$ to denote the map $x \mapsto g(f(x))$. This map is called the **composition** of g and f. A handy way to read $(g \circ f)(x)$ is "apply f to x and then apply g to the result."

The derivative of a composition: Suppose that $Df(x)$ exists and $Dg(f(x))$ exists. Then we assert the following rule:

$$D(g \circ f)(x) = Dg(f(x)) \circ Df(x)$$

In other words, $D(g \circ f)(x)$ is the map $h \mapsto (Dg(f(x)) \circ Df(x))(h) = Dg(f(x))(Df(x)(h))$. This rule generalizes the rule from first-year calculus that $(g \circ f)'(x) = g'(f(x))f'(x)$.

Proof: To justify the assertion, we start with the equation

$$g(y+k) - g(y) = Dg(y)(k) + |k|\psi_1(k) \tag{3}$$

We obtain (3) from the definition of the derivative by relabeling and by letting $\psi_1(k) = \phi(k)/|k|$, so that $\lim_{k \to 0} \psi_1(k) = 0$. Now let $y = f(x)$ and $k = k(h) = f(x+h) - f(x)$. Since f is continuous at x, $k(h)$ goes to zero as h goes to zero. Plugging these formulas into (3) yields

$$g(f(x+h)) - g(f(x)) = Dg(f(x))(k(h)) + |k(h)|\psi_1(k(h)). \tag{4}$$

Since f is differentiable at x, by the definition of k and the definition of the derivative we have

$$k(h) = Df(x)(h) + |h|\psi_2(h), \tag{5}$$

where $\lim_{h \to 0} \psi_2(h) = 0$. Plugging (5) into (4) yields

$$g(f(x+h)) - g(f(x)) = Dg(f(x))(Df(x)(h)) + Dg(f(x))(|k(h)|\psi_2(k(h))) + |k(h)|\psi_1(k(h)). \tag{6}$$

We just need to show that each of the last two terms in (6) is $o(h)$. The first term is bounded by $|k(h)||Dg(f(x))||\psi_2(k(h))|$, so each term is equal to $|k(h)|$ times a factor that goes to zero as h goes to zero. Thus it suffices to show that $|k(h)|/|h|$ is bounded. To do this we use the triangle inequality and (5) to obtain

$$\begin{aligned} |k(h)| &= |Df(x)(h) + |h||\psi_2(h)| \\ &\leq |Df(x)(h)| + |h||\psi_2(h)| \\ &\leq |Df(x)||h| + |h||\psi_2(h)| \end{aligned}$$

Dividing the right-hand side by $|h|$ yields $|Df(x)| + |\psi_2(h)|$, which is a constant plus a term going to zero, and therefore bounded.

Example: Let $f: \mathbf{R} \to \mathbf{R}^2$ be the map $f(x) = (x, x^2)$. Let $g: \mathbf{R}^2 \to \mathbf{R}$ be the map $g(y_1, y_2) = y_1^2 + y_2^2$. Then

- $Df(x)(h) = (1, 2x)h$.
- $Dg(y_1, y_2)(h) = (2y_1, 2y_2) \cdot h$, where the dot represents the dot product.
- $Dg(f(x))(h) = Dg(x, x^2)(h) = (2x, 2x^2) \cdot h$.
- $Dg(f(x))(Df(x)(h)) = (2x, 2x^2) \cdot (1, 2x)h = (2x + 4x^2)h$.

Therefore, by the chain rule, $D(g \circ f)(x)(h) = (2x + 4x^2)h$.

7.6. Composition with a Linear Map

Let $f: X \to Y$ be a map, and let $\lambda: Y \to Z$ be a linear map. By the chain rule, at all x where the derivatives on the right-hand side exist, we have

$$D(\lambda \circ f)(x) = D\lambda(f(x)) \circ Df(x).$$

By the rule for the derivative of a linear map, we have $D\lambda(f(x)) = \lambda$. Therefore

$$D(\lambda \circ f) = \lambda \circ Df$$

wherever $Df(x)$ exists. In particular, where $M(r)$ means "multiply by r," we have

$$D(M(r) \circ f) = M(r) \circ Df.$$

In other words,

$$D(y \mapsto rf(y))(x) = h \mapsto rDf(x)(h).$$

By the rule for multiplying functions by numbers, we can also write

$$D(rf)(x) = rDf(x).$$

This rule generalizes the rule from first-year calculus that $(rf(x))' = rf'(x)$.

7.7. The Fundamental Theorem of Calculus

Single real variable case: In first-year calculus, we learn that if $f: \mathbf{R} \to \mathbf{R}$ is a function defined on an interval $[a, b]$, and if f is integrable on $[a, b]$, and if

$$F(x) = \int_a^x f(t)dt, \tag{7}$$

Then $F'(x) = f(x)$ at all points x in $[a, b]$ where $f(x)$ is continuous. In the notation of the general derivative, which is a linear map and not a number, we have that $DF(x)(h) = f(x)h$, so

$$DF(x)(1) = f(x) \tag{8}$$

at all points where $f(x)$ is continuous. This theorem is called the **fundamental theorem of calculus**. It establishes the basic relationship between the integral and the derivative, the two main areas of study in calculus.

Proof: Here is a proof of the fundamental theorem in one real variable that uses the definition of the general derivative. Let

$$\phi(h) = \int_x^{x+h} (f(t) - f(x))dt. \tag{9}$$

From the rules for integration, we have

$$\phi(h) = \int_x^{x+h} f(t)dt - \int_x^{x+h} f(x)dt = F(x+h) - F(x) - f(x)h.$$

Therefore

$$F(x+h) = F(x) + f(x)h + \phi(h). \tag{10}$$

On the other hand, by estimating the integral in (9), we get

$$|\phi(h)| \le |h| \sup |f(t) - f(x)|$$

for $x \le t \le x + h$. Because f is continuous at x, the supremum goes to zero as h goes to zero. Therefore $\phi(h)$ is $o(h)$. Putting this result together with (10) establishes the theorem.

The general case: In the general case, we have maps f and F from $\mathbf{R}$ to Y, where Y is a normed vector space. The general theorem again says that if (7) holds, then (8) holds at all points x in $[a, b]$ where $f(x)$ is continuous. The only difference is that $DF(x)(1) = f(x)$ is a vector in Y instead of a number.

The same proof goes through in this case. $f(x)h$ represents scalar multiplication of the vector $f(x)$ by the number h. We just need to extend the theory of integration to cover maps from numbers to vectors. For an elementary treatment of this theory, see [Lang 1997]. For a more advanced treatment based on measure theory, see [Lang 1993].

Going in the other direction, if we are given a map $F: \mathbf{R} \to Y$ with a continuous derivative, then by the definition of the derivative we know that (8) holds for some map $f: \mathbf{R} \to Y$; and by the theorem we know that there is a constant C such that $\int_a^x DF(t)(1)\ dt = F(x) + C$. Therefore we have

$$\int_a^b DF(t)(1)\ dt = F(b) - F(a).$$

In the case of a single real variable, $DF(t)(1) = M(F'(t))(1) = F'(t)$, so this becomes the usual formula

$$\int_a^b F'(t)\ dt = F(b) - F(a).$$

7.8. The Mean Value Theorem

Single real variable case: In first-year calculus, we learn that if $f(x)\colon \mathbf{R} \to \mathbf{R}$ is differentiable on the interval $[a, b]$, then for some c between a and b we have $f'(c) = (f(b) - f(a))/(b - a)$. This statement is called the **mean value theorem**. It says that for some c in $[a, b]$ the instantaneous rate of change $f'(c)$ agrees with the average or mean rate of change between a and b.

When $f'(x)$ is continuous on $[a, b]$, the mean value theorem follows from the fundamental theorem of calculus and from the intermediate value theorem for continuous functions. From the fundamental theorem, we have

$$f(b) - f(a) = \int_a^b f'(x)dx.$$

Let $m = (f(b) - f(a))/(b - a)$. If $f'(x) < m$ on all of $[a, b]$, then the integral is less than $m(b - a)$ so less than $f(b) - f(a)$. So for some value c_1 in $[a, b]$, we must have $f'(c_1) \geq m$. Similarly, if $f'(x) > m$ on all of $[a, b]$, then the integral is greater than $f(b) - f(a)$. So for some value c_2 in $[a, b]$, we must have $f'(c_2) \leq m$. Then by the intermediate value theorem, we must have $f'(c) = m$ for some c between c_1 and c_2.

The general case: In the general setting, we cannot integrate a map from vectors to vectors with respect to one real variable dx. We can do one of the following:

1. Integrate a map $f(x)\colon \mathbf{R} \to Y$ from numbers to vectors with respect to dx, as discussed in the previous section.
2. Integrate a map $f(x)\colon X \to Y$ from vectors to vectors by resolving X into coordinates $x_1, \ldots, x_n$ and using either a **multiple integral** or an or an area or volume element called a **differential form** with respect to $dx_1, \ldots, dx_n$.

Here we want to stick with case (1). So in the general setting we revise the theorem slightly.

By the fundamental theorem, still in one real variable, we can write

$$f(x + h) - f(x) = \int_x^{x+h} f'(y)dy.$$

Let $y(t) = x + th$ and $dy = h\ dt$. Then when $t = 0$, we have $y(t) = x$, and when $t = 1$, we have $y(t) = x + h$. Therefore by a standard change of variables, using the chain rule from first-year calculus, we can write

$$f(x + h) - f(x) = \int_0^1 f'(x + th)h\ dt$$

Then passing to the notation of the general derivative Df, still in one real variable, we can write

$$f(x + h) - f(x) = \int_0^1 Df(x + th)(h)\ dt \tag{11}$$

In this form, we can let x and h be vectors. The variable t is a real number that scales the magnitude of the vector th, and we integrate with respect to t. When Df is continuous, the statement is valid, because by the generalized fundamental theorem (§ 7.7), we have

$$f(x + h) - f(x) = (f \circ y)(1) - (f \circ y)(0) = \int_0^1 D(f \circ y)(t)(1)\ dt,$$

and by the generalized chain rule

$$D(f \circ y)(t)(k) = (Df(y(t)) \circ Dy(t))(k) = Df(x + th)(hk),$$

so $D(f \circ y)(t)(1) = Df(x + th)(h)$. We call (11) the **generalized mean value theorem**.

8. Second and Higher Derivatives

The second derivative: Let X and Y be normed vector spaces, and let $f\colon X \to Y$ be a map. At points x where f is differentiable, the derivative $Df(x)$ is a linear map from X to Y. Therefore Df is a map from X to $L(X,Y)$. As discussed in § 4.2, $L(X,Y)$ is a normed vector space. Therefore we can take the derivative $D(Df)$ of the map Df at points x where the definition of $D(Df)$ is satisfied. The derivative $D(Df)$ is called the **second derivative** of f. We usually write $D^2 f$ instead of $D(Df)$. Note that if $D^2 f(x)$ exists, then by the definition of the derivative, $Df(x)$ exists as well.

Comparing $D^2 f$ to Df at each point where $D^2 f(x)$ exists, we see the following:

1. $Df(x)$ is a linear map $h \mapsto Df(x)(h)$. For small h, $Df(x)(h)$ approximates $f(x+h) - f(x)$.
2. $D^2 f(x)$ is a linear map $k \mapsto D(Df(x))(k)$. For small k, $D^2 f(x)(k)$ approximates $Df(x+k) - Df(x)$. But $Df(x+k)$ itself is a linear map $h \mapsto Df(x+k)(h)$. Similarly, $Df(x)$ is a linear map $h \mapsto Df(x)(h)$. Therefore for small k and all h, $D^2 f(x)(k)(h)$ approximates $(Df(x+k) - Df(x))(h) = Df(x+k)(h) - Df(x)(h)$.

When working with second derivatives, it is useful to keep the following facts in mind:

1. $D^2 f = x \mapsto k \mapsto h \mapsto D^2 f(x)(k)(h)$ is a map from X to $L(X, L(X,Y))$.
2. If each of x, k, and h is an element of X, then
   a. $D^2 f(x)$ is a linear map from X to $L(X,Y)$.
   b. $D^2 f(x)(k)$ is a linear map from X to Y. It approximates the map $Df(x+k) - Df(x)$, which in general is not linear.
   c. $D^2 f(x)(k)(h)$ is an element of Y.

Example 1 (Functions from R to R): Let $f\colon \mathbf{R} \to \mathbf{R}$ be a function. Assume that $f'(x)$ and $f''(x)$ exist at x, where f' and f'' are the first and second derivatives from first-year calculus. By § 5, we know that $Df = M \circ f'$. By the rule for composition with a linear map (§ 7.6), we have

$$D^2 f = D(M \circ f') = M \circ Df' = M \circ (M \circ f'').$$

Therefore

$$D^2 f(x)(k)(h) = f''(x)kh.$$

For example, if $f(x) = x^3$, then $f'(x) = 3x^2$, $f''(x) = 6x$, and $D^2 f(x)(k)(h) = 6xkh$.

Example 2 (Maps from $\mathbf{R}^2$ to R): Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be a map, and assume that $D^2 f$ exists at x. By § 6.2, we have $Df = M_{22} \circ v$, where $v(x) = (D_1 f(x), D_2 f(x))$. By the rule for composition with a linear map, we have $D^2 f = M_{22} \circ Dv$. By § 6.3, $Dv(x)(k) = A(x)k$, where

$$A(x) = \begin{bmatrix} D_1 D_1 f(x) & D_2 D_1 f(x) \\ D_1 D_2 f(x) & D_2 D_2 f(x) \end{bmatrix}. \tag{1}$$

The notation $D_i D_j f(x)$ says to take the partial derivative D_j of $f(x)$ and then to take the partial derivative D_i of $D_j f(x)$. Therefore we have

$$D^2 f(x)(k)(h) = (A(x)k) \cdot h,$$

where the dot denotes M_{22}. The matrix of double partial derivatives in (1) is called the **Hessian matrix**.

Example 3 (Maps from $\mathbf{R}^2$ to $\mathbf{R}^2$): Let $f\colon \mathbf{R}^2 \to \mathbf{R}^2$ be a map, and assume that $D^2 f$ exists at x. By § 6.3, we have $Df(x)(h) = A(x)h$, where

$$A(x) = \begin{bmatrix} a_1(x) \\ a_2(x) \end{bmatrix} = \begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) \end{bmatrix}.$$

By the rule for composition with a linear map, we have $D^2 f(k)(h) = (DA(x)k)h$. By § 6.3 again, we have that $DA(x)(k) = B(x)k$, where

$$B(x) = \begin{bmatrix} b_{11}(x) & b_{12}(x) \\ b_{21}(x) & b_{22}(x) \end{bmatrix} = \begin{bmatrix} D_1 a_1(x) & D_2 a_1(x) \\ D_1 a_2(x) & D_2 a_2(x) \end{bmatrix}$$

$$= \begin{bmatrix} (D_1D_1f_1(x), D_1D_2f_1(x)) & (D_2D_1f_1(x), D_2D_2f_2(x)) \\ (D_1D_1f_2(x), D_1D_2f_2(x)) & (D_2D_1f_2(x), D_2D_2f_2(x)) \end{bmatrix}$$

and $B(x)k = M_{21\,22\,21}(B(x))(k)$. This product operates as follows, where $k = (k_1, k_2)$:

1. The answer has one column and two rows. Row i is $b_{i1}(x)k_1 + b_{i2}(x)k_2$.
2. Each element $b_{ij}(x)k_j$ is a vector in $\mathbf{R}^2$ times a number, yielding a vector in $\mathbf{R}^2$. So the answer is a single column of two vectors in $\mathbf{R}^2$, i.e., a 2×2 matrix.

Therefore we have

$$D^2f(x)(k)(h) = (B(x)k)h.$$

The second derivative as a bilinear map: Given a second derivative D^2f evaluated at x, we may think of it in either of two ways:

1. As a map $D^2f(x)$: $L(X \to L(X, Y))$ given by $k \mapsto h \mapsto D^2f(x)(k)(h)$.
2. As a map $D^2f(x)$: $X \times X \to Y$ given by $(k, h) \mapsto D^2f(x)(k)(h)$.

Both ways are valid. The map (2) is bilinear in the variables k and h. We write it $D^2f(x)(k, h)$.

The symmetry of the second derivative: Suppose that $D^2f(x)$ exists and is continuous at x. Then $D^2f(x)$ is **symmetric**, i.e., $D^2f(x)(k, h) = D^2f(x)(h, k)$ for all k and h in X.

To prove this assertion, we can use the generalized mean value theorem to estimate the difference

$$d(k, h) = D^2f(x)(k, h) - D^2f(x)(h, k).$$

Then we can make an argument similar to the one for the uniqueness of the derivative, given in § 5, to show that $d(k, h) = 0$ for all k and h. See [Lang 1993] for the details.

Together with Example 2 above, this result shows that when the second derivative is continuous, the Hessian matrix (1) is symmetric, and therefore we have $D_{ij}f(x) = D_{ji}f(x)$ for all i and j.

Higher derivatives: We can iterate the process of taking derivatives indefinitely. For each $n > 1$, we get an nth derivative $D^nf = D(D^{n-1}f)$. As with the second derivative, we can think of $D^nf(x)$ in either of two ways:

1. As a map $h_n \mapsto \cdots \mapsto h_1 \mapsto D^nf(x)(h_n)\cdots(h_1)$.
2. As a map $(h_n, \ldots, h_1) \mapsto D^nf(x)(h_n)\cdots(h_1)$.

In the second form, $D^nf(x)$ is symmetric and linear in each of the variables h_i. See [Lang 1993] for the proofs.

9. Taylor's Formula

In this section, we derive Taylor's formula in the general setting. As in the case of one real variable, Taylor's formula lets us approximate a function as a sum of derivatives.

9.1. Integration by Parts

We begin by generalizing the technique of **integration by parts** learned in first-year calculus. To simplify the discussion, we will focus on the specific case that we need to derive Taylor's formula. It is not hard to generalize the result.

Fix the following:

1. A normed vector space Y.
2. A map $u(t)$: $\mathbf{R} \to Y$ defined on the interval $[0, 1]$.
3. A function $v(t)$: $\mathbf{R} \to \mathbf{R}$ defined on the interval $[0, 1]$.

Then $(uv)(t) = u(t)v(t)$ represents scalar multiplication on the right of the vector $u(t)$ in Y by the real number $v(t)$. Assume that $D(uv)$ exists and is continuous everywhere on $[0, 1]$. By the fundamental theorem of calculus (§ 7.7), we have

$$\int_0^1 D(uv)(t)(1)\ dt = (uv)(1) - (uv)(0). \tag{1}$$

By the product rule for differentiation (§ 7.4) and the rules of integration, the left-hand side of (1) is

$$\int_0^1 u(t)v'(t)\ dt + \int_0^1 Du(t)(1)v(t)\ dt. \tag{2}$$

Putting (1) together with (2) and rearranging terms yields

$$\int_0^1 u(t)v'(t)\ dt = u(1)v(1) - u(0)v(0) - \int_0^1 Du(t)(1)v(t)\ dt. \tag{3}$$

9.2. The Error Term in the Derivative

Now we use the generalized mean value theorem together with integration by parts to compute the error term $\phi(h)$ in the definition of the derivative. Let X and Y be normed vector spaces, and let $f\colon X \to Y$ be a map. Assume that f has a continuous derivative Df at x. Let h be a vector in X, and assume that $f(x+th)$ and $D^2 f(x+th)$ are defined for all t in the interval $[0,1]$. Then we assert that

$$f(x+h) = f(x) + Df(x)(h) + \int_0^1 (1-t)D^2 f(x+th)(h)(h)\ dt. \tag{4}$$

When x is a single real variable, this is

$$f(x+h) = f(x) + f'(x)h + \int_0^1 (1-t)f''(x+th)h^2\ dt.$$

Proof: By the generalized mean value theorem (§ 7.8), we have

$$f(x+h) = f(x) + \int_0^1 Df(x+th)(h)\ dt \tag{5}$$

Set $u(t) = Df(x+th)(h)$ and $v(t) = -(1-t)$. Then the integral in (5) is $\int_0^1 u(t)v'(t)\ dt$, and we can apply integration by parts. We have $u(1)v(1) = 0$ and $u(0)v(0) = -Df(x)(h)$. Write $u = (\lambda \circ g)$, where λ is the linear map $y \mapsto y(h)$, and $g(t) = Df(x+th)$. By the rule for composition with a linear map, we have $Du = \lambda \circ Dg$. By the chain rule, $Dg(t)(k) = D^2 f(x+th)(hk)$. Therefore

$$Du(t)(1)v(t) = -(1-t)D^2 f(x+th)(h)(h).$$

(4) then follows from (5) by substituting these terms into the right-hand side of (3).

Equation (4) gives us an explicit formula for the error term $\phi(h)$ in the definition of the derivative. We have

$$\phi(h) = \int_0^1 (1-t)D^2 f(x+th)(h)(h)\ dt.$$

Indeed $\phi(h)$ is $o(h)$, because

$$|\phi(h)| \le \int_0^1 |(1-t)D^2 f(x+th)(h)(h)|\ dt$$

$$\leq \int_0^1 |D^2 f(x+th)\|h\|h|\ dt$$

$$\leq C|h\|h|$$

where C is the supremum of $|D^2 f(x+th)|$ for t in $[0,1]$. Then $|\phi(h)|/|h| = C|h|$ which goes to zero as h goes to zero.

9.3. Higher Terms

We can continue integrating by parts. Doing this gives us Taylor's formula.

Starting with (4), let $u(t) = (1-t)D^2 f(x+th)(h)(h)$ and $v(t) = -(1-t)^2/2$. Then integration by parts yields

$$f(x+h) = f(x) + Df(x)(h) + \frac{D^2 f(x)(h)(h)}{2} + \int_0^1 \frac{(1-t)^2}{2}\, D^3 f(x+th)(h)(h)(h)\ dt. \tag{6}$$

This is gives the degree two term in Taylor's formula.

Let us write $(h)\cdots(h)$ (n times) as $(h)^n$. Then by induction we get the following:

$$f(x+h) = f(x) + Df(x)(h) + \cdots + \frac{D^n f(x)(h)^n}{n!} + \int_0^1 \frac{(1-t)^n}{n!}\, D^{n+1}(x+th)(h)^{n+1}\ dt. \tag{7}$$

Formula (7) is the generalized Taylor formula of degree n. It is a kind of extended polynomial. When $f(x)$ is a map from **R** to **R**, it is the familiar polynomial formula

$$f(x+h) = f(x) + f'(x)h + \cdots + f^{(n)}(x)h^n + \int_0^1 \frac{(1-t)^n}{n!}\, f^{(n+1)}(x+th)h^{n+1}\ dt,$$

where $f^{(n)}$ means the nth derivative of f as a function of one real variable. By an argument similar to the one given in the previous section, we can see that the integral remainder is $o(|h|^n)$. Also, for $2 \leq i \leq n$, the degree i term is $o(|h|^{i-1})$.

10. Infinite Dimensions

As discussed in § 2, the theory developed above is for finite-dimensional normed vector spaces over **R**. It is not difficult to extend the theory infinite-dimensional vector spaces. When working in infinite dimensions, the main issues are as follows:

1. We must define the concept of a **complete** normed vector space, and we must assert that our vector spaces are complete. In finite dimensions over **R**, we get completeness for free.
2. We must assert that our linear maps λ are continuous; or, equivalently, that the norm $|\lambda|$ as we defined it in § 4.2 is finite. In finite dimensions over **R**, all linear maps satisfy this condition.

See [Lang 1993] for the details.

References

Lang, Serge. *Real and Functional Analysis.* Third Edition. Springer Verlag 1993.

Lang, Serge. *Undergraduate Analysis.* Second Edition. Springer Verlag 1997.