Continuous Calculus

Peter J. Olver
School of Mathematics
University of Minnesota
Minneapolis, MN 55455
olver@umn.edu
http://www.math.umn.edu/∼olver

Contents

1. Introduction.

One evening in late June, 2020, for no apparent reason (or at least none I can now recall) I started thinking about the topological definition of continuity, [**12**]. In brief, a function is said to be *continuous* if and only if the inverse image of any open set is open. This sounds very simple — and certainly simpler than the limit-based definition used in calculus. And I started wondering why not try to develop basic calculus using this as the starting point, and, possibly, eliminating all references to limits, epsilons, and deltas while still retaining rigor. And, after some thought, I realized it could be done. Continuity is basic, and limits, including limits of sequences, and derivatives follow from it in a reasonably straightforward manner, while bypassing epsilons and deltas entirely! You will see the results of this line of reasoning below.

Not only can the development be made completely rigorous, I believe it is more elementary and eminently more understandable by the beginning mathematics student, who will be better able to appreciate the rigor behind the calculational tools. Moreover, this approach introduces them to the basics of point set topology at an early stage in their mathematical career, rather than having to start from scratch in a later course in the subject or in preparation to study real analysis. The casual reader of these notes can, of course, skip over the proofs to assimilate the basic ideas and computational techniques, returning to them at a later stage as necessary. On the other hand, the reader already familiar with basic calculus may find new insight and appreciation for the subject using this alternative foundation of the subject. Perhaps a better description of these notes would be *Topological Calculus*. However, I refrained from using this more sophisticated title so as not to potentially scare off prospective readers through the use of unfamiliar words at the outset.

In the second half of these notes, I have also adopted a somewhat non-traditional approach to integration — both single and double integrals — that is based entirely on a simplified definition of area, although it is not that far removed from the standard constructions of Riemann sums. The proof of the Fundamental Theorem of the Calculus, that the derivative of the integral of a continuous function is the function itself, is a straightforward consequence of our continuity-based approach to differentiation. While there is much more in these notes than could be expected of a basic course, a suitably toned down subset could, I believe, form the basis for an introductory course in calculus, both more rigorous and more understandable to the student than the usual limit-based approach.

The reader is referred to the standard textbooks [**1**, **6**, **7**, **27**] for traditional introductions to calculus, as well the less conventional texts [**26**, **28**]. All rely on limits, and use ε's and δ's throughout, although in a typical introductory calculus course, we tend skip over the rigorous foundations entirely because they are simply too challenging for most students to come to grips with! The one precedent for this approach that I have been able to locate so far is the online booklet of Daniel J. Bernstein, [**2**], which I became aware of after realizing that the continuity-based definition of the derivative goes back to Carathéodory, [**4**, **13**];

see Sections 7 and 12. For an alternative limit-free approach to the foundations of calculus, see [**17**].

Prerequisites: The reader should be familiar with elementary set theory, proof by induction, the real number line, and the concept of a function, including the elementary functions, by which we mean polynomials, rational functions, algebraic functions (involving roots), exponential functions, (natural) logarithms, and the trigonometric functions and their inverses. Elementary properties of these functions will be stated and used without proof. Otherwise these notes are self-contained, except for the proofs of a couple of the more difficult theorems that arise in multivariable calculus, where we refer to the literature. In particular, no familiarity with topology is assumed, and the concepts of open and closed sets, and continuity are developed from scratch. (On the other hand, we do not attempt to construct the real number line; see [**24**, **26**] for several versions.)

I will employ standard notation for sets and functions. In particular, $\{\, F \mid C \,\}$ will denote a set, where F gives the formula for the members of the set and C is a list of conditions. For example, $\{\, x \in \mathbb{R} \mid 0 \le x \le 1 \,\}$ means the closed unit interval from 0 to 1, also written $[0,1]$ and also sometimes abbreviated as $\{\, 0 \le x \le 1 \,\}$. The empty set is denoted by the symbol $\varnothing$. The usual notations $x \in A$ will mean x is an element of the set A, while $y \notin A$ means that y is not an element of A. Further, $A \subset B$ indicates that A is a subset of B; if we need to also say $A \ne B$, we write $A \subsetneq B$. We use $A \cup B$ for the union of the sets A, B, and $A \cap B$ for their intersection. The notation $A \setminus B$ for their set-theoretic difference, meaning the set of all elements of A which do not belong to B; note that this does not require $B \subset A$. For brevity, I will at times say something like "the subset $x \in S \subset A$" as a slightly shorter way of saying "the subset $S \subset A$ containing the element x".

The notation $f\colon X \to Y$ means that the function f maps the domain set X to the image set Y, with formula $y = f(x) \in Y$ for $x \in X$. The range of f, meaning $\{\, y \in Y \mid y = f(x) \text{ for some } x \in X \,\}$ is, in general, a subset of Y. Composition of functions is denoted by $f \circ g$, so that $f \circ g(x) = f\big[\, g(x) \,\big]$.

Finally *Q.E.D.* signifies the end of a proof, an acronym for "quod erat demonstrandum", which is Latin for "which was to be demonstrated".

2. Elementary Topology of the Real Line.

The setting of one variable calculus is the real line $\mathbb{R}$, and we assume the reader is familiar with its basic properties. Multivariable calculus takes place in higher dimensional space and we will discuss how these constructions can be adapted there starting in Section 9. We will use the notation $\mathbb{Z}$ for the integers, $\mathbb{Q}$ for the rational numbers, and $\mathbb{N} = \{\, n \in \mathbb{Z} \mid n > 0 \,\}$ for the natural numbers, excluding 0. All are subsets of $\mathbb{R}$.

The *basic open sets* of $\mathbb{R}$ are the open intervals:

$$(a,b) = \{\, x \in \mathbb{R} \mid a < x < b \,\}. \tag{2.1}$$

Here $a < b$ are any real numbers, the *endpoints* of the interval. We also allow $a = -\infty$ or $b = \infty$ or both, so that $\mathbb{R} = (-\infty, \infty)$ is itself a basic open set. Two other important cases are $\mathbb{R}^+ = (0, \infty) = \{x > 0\}$, the *positive real numbers*, and $\mathbb{R}^- = (-\infty, 0) = \{x < 0\}$, the *negative real numbers*. The interval (a, b) is *bounded* if both $a, b \in \mathbb{R}$; otherwise it is *unbounded*. Once we decide on the basic open sets, we can form general open sets.

Definition 2.1. An *open set* is a union of basic open sets.

Thus, the open sets in $\mathbb{R}$ are unions of open intervals. The notion of open set is said to define a *topology* on a space, in this case the real number line, [**12**]. The word "topology" was coined by mathematicians in the nineteenth century, and is based on the Greek word topos ($\tau o \pi o \varsigma$) that means "place", hence topology is the "study of place".

Here are a few simple examples of open sets:

- Any open interval is of course open, being the union of a single basic open set, namely itself.
- The empty set is open, being the union of zero open intervals.
- The union of two open intervals is open, but not necessarily an interval. For example $(0, 1) \cup (2, 3)$ is the set of points $x \in \mathbb{R}$ such that either $0 < x < 1$ or $2 < x < 3$. On the other hand, $(0, 2) \cup (1, 3) = (0, 3)$ is an open interval. Finally $(0, 1) \cup (1, 2)$ is the pair of intervals obtained by deleting the point 1 from the interval $(0, 2)$; in other words $(0, 1) \cup (1, 2) = (0, 2) \setminus \{1\}$. One can similarly take unions of any finite number of open intervals to obtain an open set that is, on occasion, an interval itself. The reader may enjoy writing down necessary and sufficient conditions for this to be the case.
- But we can also take unions of infinitely many intervals. For example,
$$S = \bigcup_{n=-\infty}^{\infty} (n, n+1)$$
is the open set obtained by deleting all the integers: $S = \mathbb{R} \setminus \mathbb{Z}$.

Example 2.2. A particularly interesting example of the last case is provided by the open set S obtained as the union of the intervals
$$\left(\frac{3k-2}{3^n}, \frac{3k-1}{3^n} \right), \qquad \text{where} \qquad n \in \mathbb{N}, \quad \text{and} \quad k = 1, \ldots, 3^{n-1}. \tag{2.2}$$
Thus $S \subset [0, 1]$ consists of an infinite collection of disjoint intervals of progressively smaller and smaller size. Its complement $C = [0, 1] \setminus S$ is known as the *Cantor set*, named after the nineteenth century Russian-born, German-based mathematician Georg Cantor, the founder of modern set theory and point set topology, although this set was first discovered in 1874 by the British mathematician Henry John Stephen Smith. (Mathematics is full of objects and results that were introduced or proved by X but named after Y.) The Cantor set plays a particularly important role in the study of fractals, [**15**], and in advanced real analysis, [**24**].

Let us collect together some basic facts concerning open sets.

Proposition 2.3. *The union of any collection of open sets is open.*

This is clear because each open set in the collection is itself a union of basic open sets, and hence the union of the collection is equal to the union of all the basic open sets associated with each set in the collection.

Proposition 2.4. *The intersection of any* finite *collection of open intervals is either empty or an open interval.*

Proof: Let us first prove the result for two open intervals: $I_1 = (a, b)$, $I_2 = (c, d)$, where $a < b$ and $c < d$. We also assume, without loss of generality, that $a \le c$ as otherwise we interchange the two intervals, $I_1 \longleftrightarrow I_2$, before proceeding. There are then three possible subcases: (a) If $a < b \le c < d$ then $I_1 \cap I_2 = \varnothing$; (b) if $a \le c < b \le d$, then $I_1 \cap I_2 = (c, b)$; (c) if $a \le c < d \le b$, then $I_1 \cap I_2 = (c, d)$. Thus in each case, the intersection is either empty or an open interval. By a straightforward induction, the same holds for any finite collection of open intervals. *Q.E.D.*

Proposition 2.5. *The intersection of any* finite *collection of open sets is open.*

Proof: It suffices to prove this for two open sets S, T since then one can use an evident induction to prove it for any finite intersection. Let $S = \bigcup I$ be a union of open intervals, and similarly for $T = \bigcup J$. But then

$$S \cap T = \bigcup I \cap \bigcup J = \bigcup \,(I \cap J),$$

where the right hand side is the union of the intersections of all pairs of intervals forming S and T respectively. Proposition 2.4 implies that each term in the final union is either empty or an open interval, and hence, discarding all the empty sets, one expresses $S \cap T$ as a union of open intervals. *Q.E.D.*

The intersection of an infinite collection of open sets need not be open. For example,

$$\bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n} \right) = \{ 0 \},$$

and the latter set, being a single point, is not open. (Why?)

Lemma 2.6. *A set $S \subset \mathbb{R}$ is open if and only if for every $x \in S$ there is a (bounded) open interval I such that $x \in I \subset S$.*

Proof: Since an open set S is, by definition, the union of open intervals, every $x \in S$ belongs to one or more of them. If one requires boundedness, and the interval in question is unbounded, we can just choose any bounded interval that contains x and is contained within it. Conversely, if each $x \in S$ belongs to a (bounded) open interval: $I_x \subset S$ so that $x \in I_x$, then $S = \bigcup_{x \in S} I_x$ can be expressed as their union. It may be an uncountable union, but that is still allowed in the original definition. *Q.E.D.*

Remark: In Lemma 2.6, one could even impose a bound on the length of the intervals $I = (a, b)$ allowed, i.e., require $b - a < r$ for some fixed $r > 0$.

Two sets A, B are said to be *disjoint* if their intersection is empty: $A \cap B = \varnothing$. A collection of subsets is (pairwise) *disjoint* if any two members are disjoint. The proof of the following result will appear below.

Proposition 2.7. *If $S \subset \mathbb{R}$ is open, then it is the union of disjoint open intervals.*

Theorem 2.8. *The real line $\mathbb{R}$ cannot be written as the disjoint union of two or more open sets.*

Proof: We can use Proposition 2.7 to write each of the open sets as the disjoint union of open intervals, and hence it suffices to show that $\mathbb{R}$ cannot be written as the disjoint union of two or more open intervals. Let (a, b) be one of the intervals. Then, for any other interval (c, d) in the collection, either $d < a$ or $b < c$ since the two intervals are assumed to be disjoint. But this implies neither a nor b lies in any of the intervals in the collection. At least one of a, b is finite, as otherwise $(a, b) = (-\infty, \infty) = \mathbb{R}$ and there would be no other intervals in the collection. Thus, a finite endpoint of any interval in the collection does not belong to any of the intervals, and hence does not lie in their union either. *Q.E.D.*

Remark: With a little more care, one can similarly prove that an open interval (a, b) cannot be written as the disjoint union of two or more open subsets.

The set-theoretic complement of an open set is known as a closed set.

Definition 2.9. A set $C \subset \mathbb{R}$ is *closed* if the set $S = \mathbb{R} \setminus C$ is open.

For example, $\mathbb{R}$ is closed because $\mathbb{R} \setminus \mathbb{R} = \varnothing$ is open. The same holds for the empty set $\varnothing$, both of which are also open. In fact, the only subsets of $\mathbb{R}$ which are both open and closed are $\mathbb{R}$ itself and the empty set $\varnothing$.

Proposition 2.10. *If $\varnothing \neq S \subset \mathbb{R}$ and S is both open and closed, then $S = \mathbb{R}$.*

Proof: Suppose $\varnothing \neq S \neq \mathbb{R}$. According to Proposition 2.7, S is the union of one or more disjoint open intervals. On the other hand, since S is also closed its complement $C = \mathbb{R} \setminus S$ is open, not empty, and hence also the union of one or more disjoint open intervals. But this implies $\mathbb{R} = S \cup C$ is the union of two or more disjoint open intervals, in contradiction to Theorem 2.8. *Q.E.D.*

A *closed interval*

$$[a, b] = \{ x \in \mathbb{R} \mid a \leq x \leq b \} \tag{2.3}$$

where $a \leq b$ are both finite, is, not surprisingly, closed. Indeed, we can write its complement

$$\mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$$

as the union of two unbounded open intervals. In particular, a single point $\{a\} = [a, a]$ forms a closed set.

Some sets are neither open nor closed, examples being the *half-open intervals*

$$(a, b] = \{a < x \leq b\}, \qquad [a, b) = \{a \leq x < b\}, \tag{2.4}$$

when $a < b$ are both finite. The definitions (2.4) make sense if $a = -\infty$ in the first case, or $b = \infty$ in the second, although the resulting unbounded intervals are closed sets. Indeed, the complement of $(-\infty, b]$ is the open unbounded interval $(b, \infty) = \mathbb{R} \setminus (-\infty, b]$, and similarly for $[a, \infty)$.

Proposition 2.11. *The intersection of any collection of closed sets is closed. The union of any* finite *collection of closed sets is closed.*

This result follows easily from Propositions 2.3 and 2.5. It allows us to construct further examples of closed sets, as follows. Since a single point $\{a\}$ forms a closed set, any finite set of points is closed. Discrete infinite sets of points need not be closed. For example $\mathbb{Z}$ is closed since, as we noted above, its complement $\mathbb{R} \setminus \mathbb{Z}$ is open. On the other hand, $\widetilde{A} = \{1/n \mid n \in \mathbb{N}\}$ is not closed since 0 is in its complement, but no interval containing 0 is because if $a < 0 < b$, then $\widetilde{A} \cap (a, b) = \{1/n \mid 1/b < n \in \mathbb{N}\}$. On the other hand, $A = \widetilde{A} \cup \{0\}$ is closed because its complement is the union of the open intervals $(-\infty, 0), (1, \infty)$ and $(1/(n+1), 1/n)$ for all $n \in \mathbb{N}$. See Section 6 for further details on this important example.

Lemma 2.12. *If S is open and C is closed then $S \setminus C$ is open. Similar, if C is closed and S is open, then $C \setminus S$ is closed.*

Indeed, given that S is open, we can write $S \setminus C = S \cap (\mathbb{R} \setminus C)$ as the intersection of two open sets, proving the first statement. The second statement is established by an analogous argument.

The *closure* of a set A, written $\overline{A}$, is defined as the intersection of all the closed sets containing it. Note that the closure $\overline{A}$ is the smallest closed set containing A; in other words, if $C \supset A$ is closed, then $\overline{A} \subset C$. In particular, if A itself is closed, $\overline{A} = A$. For example, the closure of an open interval (a, b) is the closed interval $[a, b]$. The integers $\mathbb{Z} \subset \mathbb{R}$ is already closed. Interestingly, the closure of the rational numbers $\mathbb{Q}$ is the entire set of real numbers: $\overline{\mathbb{Q}} = \mathbb{R}$. This is because every open interval $I \subset \mathbb{R}$ contains infinitely many rational numbers, and hence $\mathbb{R} \setminus \overline{\mathbb{Q}} = \varnothing$. For the same reason, the closure of the set $\mathbb{I} = \mathbb{R} \setminus \mathbb{Q}$ of irrational numbers is also $\mathbb{R}$. One says that the rational and the irrational numbers, which are neither open nor closed, are *dense* in the real number line.

Finally, we mention the fundamental *completeness* property of the real line, that is a consequence of its axiomatic construction, [**24, 26**]. A nonempty set $\varnothing \neq A \subset \mathbb{R}$ is *bounded from above* if there exists $a \in \mathbb{R}$ such that $x \leq a$ for all $x \in A$. Any number satisfying this is called an *upper bound* for A. The *least upper bound* property states that, for any set bounded from above, there exists a minimal upper bound, meaning an upper bound $M \in \mathbb{R}$ satisfying $M \leq a$ for any upper bound a. The least upper bound is written $M = \sup A$, where sup is short for *supremum*, which is a fancier name for it. If (b, M) is any open interval, then $A \cap (b, M) \neq \varnothing$ since otherwise $b < M$ would also be an upper bound of A, and M would not be the least upper bound. The least upper bound may or may not be an element of A. If A is open, then $M = \sup A \notin A$ since otherwise an interval containing M would be contained in A, and thus M could not be an upper bound. On the other hand, if A is closed then $M = \sup A \in A$. Indeed, if $M \in \mathbb{R} \setminus A$, because the latter set is open, there is an open interval I such that $M \in I \subset \mathbb{R} \setminus A$, but then $A \cap I = \varnothing$ which would imply that any point $b \in I$ is an upper bound for A and hence M cannot be the least upper bound.

Vice versa, a set $\varnothing \neq A \subset \mathbb{R}$ is *bounded from below* if there exists a *lower bound* $c \in \mathbb{R}$ such that $c \leq x$ for all $x \in A$. There also exists a *greatest lower bound*, written $m = \inf A$, where inf is short for *infimum*, such that $c \leq m$ for any lower bound c. The above statements for the least upper bound also hold for the greatest lower bound.

By convention, if A is not bounded from above, we write $\sup A = \infty$, while if it is not bounded from below, we write $\inf A = -\infty$. If $A = \{a^1, \ldots, a^n\} \subset \mathbb{R}$ is a finite subset, then its supremum is just the maximal element and its infimum is its minimal element. We will continue to use the notations max and min when dealing with finite subsets, in which case $\sup A = \max A$ and $\inf A = \min A$. Infinite sets, of course, may or may not have maximum and minimum elements, but they always have a supremum and an infimum.

Remark: The least upper bound property is not valid for the rational numbers $\mathbb{Q}$. Indeed, the least upper bound of $A = \{\, r \in \mathbb{Q} \mid r^2 < 2 \,\}$ is $\sup A = \sqrt{2}$, which is not rational. This means that there is no smallest rational upper bound. Indeed, any rational number $q \in \mathbb{Q}$ such that $q > \sqrt{2}$ is an upper bound of A, but there is no least such number. (Why?) This distinction between $\mathbb{R}$ and $\mathbb{Q}$ is one of the main reasons why one needs to extend the rational numbers to the real numbers in order to perform calculus and mathematical analysis.

Let us state a couple of elementary results about upper and lower bounds — suprema and infima — whose proofs are left as exercises for the reader.

Lemma 2.13. *Let $A, B \subset \mathbb{R}$. Then*

$$\begin{aligned} \sup(A \cup B) &= \max\{\sup A,\ \sup B\}, \qquad & \inf(A \cup B) &= \min\{\inf A,\ \inf B\}, \\ \sup(A \cap B) &= \min\{\sup A,\ \sup B\}, \qquad & \inf(A \cap B) &= \max\{\inf A,\ \inf B\}. \end{aligned} \tag{2.5}$$

Lemma 2.14. *If $A \subset B \subset \mathbb{R}$, then $\sup A \leq \sup B$ and $\inf A \geq \inf B$.*

Lemma 2.15. *Given $A \subset \mathbb{R}$, let $-A = \{\, -x \mid x \in A \,\}$. Then $\sup(-A) = -\inf A$ and $\inf(-A) = -\sup A$.*

Lemma 2.16. *Let $A, B \subset \mathbb{R}$. Define their* Minkowski sum *to be the set*

$$A + B = \{ x + y \mid x \in A,\ y \in B \}. \tag{2.6}$$

Then

$$\sup(A + B) = \sup A + \sup B, \qquad \inf(A + B) = \inf A + \inf B. \tag{2.7}$$

The Minkowski sum is named after the nineteenth century German/Lithuanian mathematician Hermann Minkowski, one of the founders of special relativity. In particular, if $B = \{b\}$ is a singleton, where $b \in \mathbb{R}$, then we write $A + b = A + \{b\} = \{ x + b \mid x \in A \}$, which is the set obtained by translating the set A by a distance b along the real line. Since $\sup B = \inf B = b$, we deduce

$$\sup(A + b) = \sup A + b, \qquad \inf(A + b) = \inf A + b. \tag{2.8}$$

Lemma 2.17. *Let $A \subset \mathbb{R}$ be a subset such that whenever $x < y < z$ and $x, z \in A$, then $y \in A$. Then A is an interval — open, closed, or half open.*

Proof: Let $m = \inf A$ and $M = \sup A$. Thus, if both are finite, $A \subset [m, M]$, with analogous statements if one or both are infinite. If $m = M$, then $A = \{m\} = [m, m]$ is a singleton, which is a closed interval. Otherwise, suppose $m < c < M$. Let $x \in (m, c) \cap A$ and $y \in (c, M) \cap A$, both of which are nonempty by the preceding remarks. Then $x < c < z$, and hence $c \in A$. Thus every point $c \in (m, M)$ belongs to A. If $m, M \notin A$, then $A = (m, M)$ is an open interval. If $m, M \in A$ then $A = [m, M]$ is a closed interval. If one but not the other belong to A, then A is the corresponding half open interval. *Q.E.D.*

Finally we supply an as yet unrealized proof.

Proof of Proposition 2.7: The case when $S = \varnothing$ is trivial, since it is the union of zero open intervals. Otherwise, given $x \in S$, let

$$a_x = \inf \{ a \mid x \in (a, b) \subset S \}, \qquad b_x = \sup \{ b \mid x \in (a, b) \subset S \}.$$

Then it is not hard to show that $I_x = (a_x, b_x) \subset S$ is the largest open interval contained in S that also contains x, i.e., if I is any interval with $x \in I$, then $I \subset I_x$. By maximality, $I_x = I_y$ if $y \in I_x$ or, equivalently, $x \in I_y$, and thus $I_x \cap I_y = \varnothing$ if $x \notin I_y$, or, equivalently, $y \notin I_x$. Thus the collection of all such intervals (without repeats) is pairwise disjoint and their union is the entire set S. *Q.E.D.*

3. Functions.

We will now look at real-valued functions defined on $\mathbb{R}$ and subsets thereof. Let $D \subset \mathbb{R}$ be the *domain* of a function $f\colon D \to \mathbb{R}$. Thus, to each point $x \in D$, the function f assigns a unique value $y = f(x) \in \mathbb{R}$. If $A \subset D$ is any subset, we define its *image* under f to be the subset

$$f(A) = \{ y = f(x) \mid x \in A \} \subset \mathbb{R}. \tag{3.1}$$

In particular, the *range* E of f is the image of the entire domain set: $E = f(D)$. For example, the following functions all have domain $D = \mathbb{R}$. The range of $f(x) = x$ is $\mathbb{R}$; the range of $f(x) = x^2$ is $[0, \infty)$; the range of $f(x) = 1/(1 + x^2)$ is $(0, 1]$; the range of $f(x) = e^x$ is $(0, \infty)$; the range of $f(x) = \sin x$ is $[-1, 1]$.

In general, if $A \subset D$, then one can restrict f to the subset A, defining $\widetilde{f}: A \to \mathbb{R}$ by $\widetilde{f}(x) = f(x)$ whenever $x \in A$, with range $\widetilde{f}(A) = f(A)$. We will write $\widetilde{f} = f \mid A$ and call $\widetilde{f}$ the *restriction* of f to the subset A. We will often drop the tilde and identify f with its restriction to any subdomain, writing $f: A \to \mathbb{R}$ instead. Conversely, given $\widetilde{f}: A \to \mathbb{R}$, we will call a function $f: D \to \mathbb{R}$ an *extension* of $\widetilde{f}$ if $\widetilde{f} = f \mid A$ whenever $x \in A$. One can always extend functions, since we can let $f(x)$ for $x \in D \setminus A$ be anything we like; in other words we can choose f such that its restriction to the complementary subset $\widehat{f} = f \mid (X \setminus A)$ is arbitrary. On the other hand, if we want the extension to have nice properties, such as continuity or differentiability, then it is not always clear that such an extension exists and, in fact, it may not. This will be discussed in detail at the appropriate points in the text.

A function $f: D \to \mathbb{R}$ is said to be *one-to-one* if assigns different values to different points in its domain; in other words, if $x \neq \widetilde{x}$ then $f(x) \neq f(\widetilde{x})$. A one-to-one function has a well-defined inverse, namely $f^{-1}(y) = x$ whenever $y = f(x)$. The domain E of f^{-1} is the range of f and vice versa, so $f(D) = E$ while $f^{-1}(E) = D$. Note that the inverse of the inverse function is the original function: $(f^{-1})^{-1} = f$.

Example 3.1.

- The constant function $f(x) \equiv 1$, whose range is the singleton $\{1\}$, is not one-to-one because every point has the same value.
- The function $f(x) = x$ is one-to-one and is its own inverse: $f^{-1}(y) = y$. More generally, any non-constant linear function† $f(x) = mx + l$ with $m \neq 0$ is one-to one with inverse $f^{-1}(y) = y/m - l/m$.
- The function $f(x) = x^2$ is not one-to-one as a function from $\mathbb{R}$ to $[0, \infty)$ since $f(-x) = f(x)$ for any x. If we restrict to f to the domain $D^+ = [0, \infty)$, then it is one-to-one, with range $E^+ = f(D^+) = [0, \infty)$. Its inverse is given by the (positive) square root function: $f^{-1}(y) = +\sqrt{y}$ for $y \geq 0$. On the other hand, f is also one-to-one when restricted to $D^- = (-\infty, 0]$, with the same range: $E^- = f(D^-) = [0, \infty)$; here its inverse is the negative square root function: $f^{-1}(y) = -\sqrt{y}$ for $y \geq 0$.
- The function

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

† We will follow the calculus convention of calling such functions *linear* since their graphs are straight lines. This is at odds with the conventions of linear algebra, [**22**], where only those with $l = 0$ are linear. The general case, when $l \neq 0$, is called an *affine function*.

is one-to-one and is its own inverse:

$$f^{-1}(y) = \begin{cases} 1/y, & y \neq 0, \\ 0, & y = 0. \end{cases}$$

- The exponential function $f(x) = e^x$ is one-to one with range $\mathbb{R}^+ = (0, \infty)$. The inverse function is the natural logarithm: $f^{-1}(y) = \log y$ for $y > 0$; see the discussion following (15.6) for details.

Even though functions that are not one-to-one do not have (single-valued) inverses, we can nevertheless define the inverse image of a set. Namely, given $f: D \to \mathbb{R}$ and a set $S \subset \mathbb{R}$, we write

$$f^{-1}(S) = \{\, x \in \mathbb{R} \mid f(x) \in S \,\} \subset D. \tag{3.2}$$

In particular, if $E = \operatorname{range} f$ and $S \cap E = \varnothing$, then $f^{-1}(S) = \varnothing$. Note that the set-valued inverse respects unions and intersections:

$$f^{-1}\left(\bigcup S\right) = \bigcup f^{-1}(S), \qquad f^{-1}\left(\bigcap S\right) = \bigcap f^{-1}(S). \tag{3.3}$$

4. Continuity.

We have now arrived at the heart of our topological approach to calculus: the definition of continuity. To begin with, for simplicity, we assume the domain of our function f is all of $\mathbb{R}$, although we do not impose any restrictions on its range.

Definition 4.1. A function $f: \mathbb{R} \to \mathbb{R}$ is called *continuous* if, whenever $S \subset \mathbb{R}$ is open, then $f^{-1}(S) \subset \mathbb{R}$ is open.

Why this requirement is called "continuous" will gradually become evident as we explore its many consequences; see also [**21**] for a physical motivation based on the continuity of motion. Thus, in view of Lemma 2.6, to verify that $f^{-1}(S)$ is open one needs to show that, every point $x \in f^{-1}(S)$ is contained in an open interval $x \in I \subset f^{-1}(S)$. Furthermore, since any open set is a union of (bounded) intervals, in view of property (3.3), to prove continuity one only needs to prove that if $J \subset \mathbb{R}$ is an open (bounded) interval, then $f^{-1}(J)$ is open, i.e., has the above property. Note that $f^{-1}(J)$ need not itself be an open interval.

We also note that f is continuous if and only if whenever $C \subset \mathbb{R}$ is closed, then $f^{-1}(C)$ is closed. Indeed, setting $S = \mathbb{R} \setminus C$, which is open, then $f^{-1}(C) = \mathbb{R} \setminus f^{-1}(S)$ is closed, and vice versa. However, their effect on open sets is the standard default for the analysis of the continuity of functions.

Let us look at some elementary examples of continuous and discontinuous functions, which are sketched in Figures 1 and 2. In each plot, a representative open interval on the y axis is colored red and its inverse image under the function on the x axis is colored green. Open circles indicate the endpoints of open intervals; closed circles indicate points or the endpoints of closed intervals.

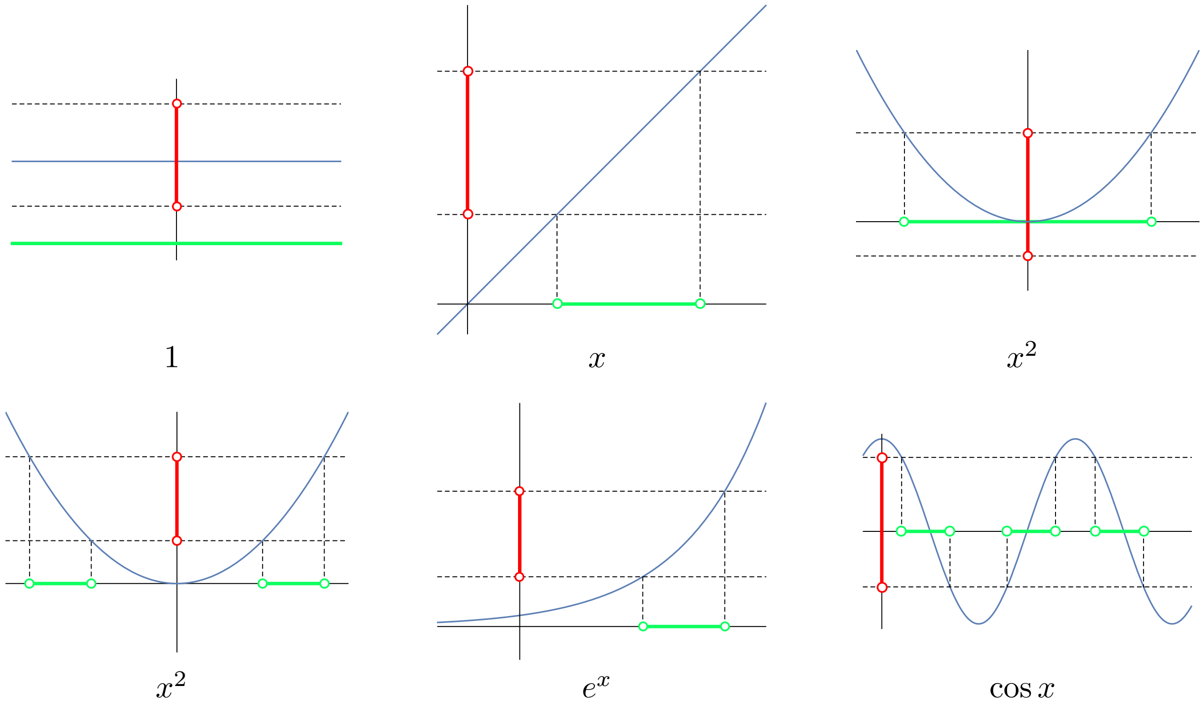


Figure 1. Continuous Functions.

Example 4.2.

- A constant function, e.g., $f(x) \equiv 1$ is continuous. Indeed, $f^{-1}(S) = \mathbb{R}$ if $1 \in S$ while $f^{-1}(S) = \varnothing$ if $1 \notin S$. In either case $f^{-1}(S)$ is open (even if S itself is not open).
- The function $f(x) = x$ is continuous. Indeed, if S is open so is $f^{-1}(S) = S$.
- The function $f(x) = x^2$ is continuous. Note that the range of f is the closed half line $[0, \infty)$. To prove continuity, we use the preceding remark and thus only need to look at the effect of f^{-1} on open intervals $J = (a, b)$ for $a < b$. There are three cases: (a) If $b \le 0$, then $f^{-1}(J) = \varnothing$. (b) If $a < 0 < b$, then $f^{-1}(J) = (-\sqrt{b}, \sqrt{b})$ is an open interval. (c) If $0 \le a < b$, then $f^{-1}(J) = (-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b})$ is the union of two open intervals. Since each one of these is open, the result follows. Later we will prove that all polynomial functions are continuous.
- The exponential function $f(x) = e^x$ is continuous. Indeed, if $J = (a, b)$ then $f^{-1}(J) = (\log a, \log b)$ is an open interval. *Remark*: Technically, this relies on the fact that the corresponding logarithms $x = \log y$ fill out the entire interval $\log a < x < \log b$. While it is possible to prove this from basic principles, a slicker strategy is to first prove continuity of the natural logarithm via integration, and then deduce continuity of its inverse, the exponential function; see Example 15.4 for details.
- We remark, without proof, that the trigonometric functions $\sin x$ and $\cos x$ are continuous on all of $\mathbb{R}$. The tangent function $\tan x$ is continuous where defined, i.e. on $D = \mathbb{R} \setminus \{\, (n + \frac{1}{2})\pi \mid n \in \mathbb{Z} \,\}$. (See below for continuity on subdomains.) As with the exponential and logarithm, the quickest proofs of these facts rely on integration.

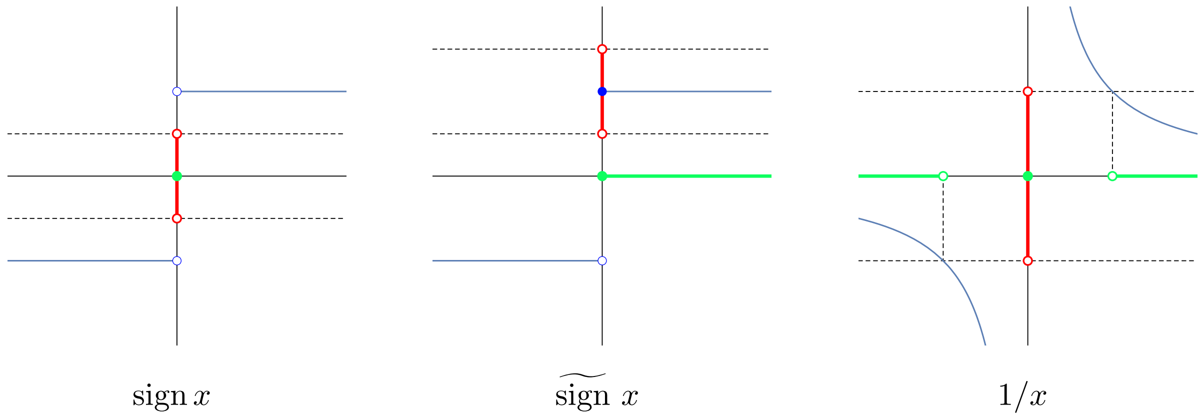


Figure 2. Discontinuous Functions.

- The *sign function*

$$\operatorname{sign} x = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0, \end{cases} \tag{4.1}$$

is *not* continuous. Indeed, setting $f(x) = \operatorname{sign} x$, if $S = \left(-\frac{1}{2}, \frac{1}{2}\right)$, then $f^{-1}(S) = \{0\}$, which is not open. Keep in mind that to demonstrate that a function is not continuous, it suffices to find one open set $S \subset \mathbb{R}$ such that $f^{-1}(S)$ is not open. Moreover, it is not possible to redefine the sign function at $x = 0$ and make it continuous. For example, if we set

$$\widetilde{f}(x) = \widetilde{\operatorname{sign}}\ x = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0, \end{cases}$$

then $\widetilde{f}^{-1}\left(\frac{1}{2}, \frac{3}{2}\right) = [0, \infty)$ is closed, not open.

- Similarly, the function

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

is not continuous since, for example, $f^{-1}(-1, 1) = \{0\} \cup (-\infty, -1) \cup (1, \infty)$. The second and third sets are open intervals, but the first, being a single point, is not open, and their disjoint union is also not open. Again, one cannot redefine $f(x)$ at $x = 0$ in order to make a continuous function. (On the other hand, as we will see below, $f(x) = 1/x$ is continuous on the open subset $\mathbb{R} \setminus \{0\}$ obtained by deleting the origin.)

The next result states that basic algebraic operations on functions maintain continuity.

Theorem 4.3. *If $f(x)$ and $g(x)$ are continuous, then*

- *Their sum $f(x) + g(x)$ and difference $f(x) - g(x)$ are both continuous.*
- *Their product $f(x)\, g(x)$ is continuous.*
- *Their quotient $f(x)/g(x)$ is continuous provided $g(x) \neq 0$.*

- *Their maximum* $\max\{ f(x),\ g(x) \}$ *and minimum* $\min\{ f(x),\ g(x) \}$ *are both continuous.*
- *Their composition* $f \circ g(x) = f[\, g(x) \,]$ *is continuous.*

In particular, the function $-f(x)$, being the product of the constant function -1 and the continuous function $f(x)$ is continuous. We defer the proof of the first four bullet points until Section 9, when it will be easy. As for the composition, this is an immediate consequence of the fact that

$$(f \circ g)^{-1}(S) = g^{-1}\big[\, f^{-1}(S) \,\big] = g^{-1}(T), \quad \text{where} \quad T = f^{-1}(S).$$

Thus, if S is open, continuity of f implies that $T = f^{-1}(S)$ is open. Furthermore, continuity of g implies that $g^{-1}(T) = (f \circ g)^{-1}(S)$ is open, which proves continuity of $f \circ g$.

We can use sums, products, and quotients to build up complicated continuous functions. For instance, knowing that x is continuous, and any constant function is continuous, we deduce $m x$ is continuous for any $m \in \mathbb{R}$, from which it follows $m x + l$ for $l \in \mathbb{R}$ is also continuous. Similarly, the product of x with itself, $x^2 = x\, x$, is continuous, which implies $x^3 = x\, x^2$ is also continuous, and, by induction, any power x^n for $n \in \mathbb{N}$ is continuous. We can in turn multiply by any constant $c \in \mathbb{R}$, so $c\, x^n$ is also continuous. Then summing up such functions we arrive at the conclusion that any polynomial function is continuous. Moreover, if $p(x)$ and $q(x)$ are polynomials with $q(x) \neq 0$ — an example is $q(x) = x^2 + 1$ — then the rational function $r(x) = p(x)/q(x)$ is continuous.

The following result provides a useful test for continuity.

Theorem 4.4. *Let* $\lambda > 0$ *be a positive constant. If* $f\colon \mathbb{R} \to \mathbb{R}$ *satisfies*

$$|\, f(x) - f(y) \,| \leq \lambda \,|\, x - y \,| \tag{4.2}$$

for all $x, y \in \mathbb{R}$*, then* f *is continuous.*

Proof: Let $(\, c, d \,)$ be an open interval. Suppose $a \in f^{-1}(c, d)$, so $c < f(a) < d$. Choose $r > 0$ so that $c < f(a) - r < f(a) + r < d$. Then if $|\, x - a \,| < r/\lambda$, condition (4.2) implies

$$|\, f(x) - f(a) \,| \leq \lambda \,|\, x - a \,| < r$$

which, by our choice of r, implies $f(x) \in (\, c, d \,)$. Thus, the open interval $(\, a - r/\lambda, a + r/\lambda \,) \subset f^{-1}(c, d)$. Since this applies to every $a \in f^{-1}(c, d)$, we conclude that $f^{-1}(\, c, d \,)$ is open, and hence that f is continuous. *Q.E.D.*

A function that satisfies the condition (4.2) for some $\lambda > 0$ is called *Lipschitz continuous*, after the nineteenth century German mathematician Rudolf Lipschitz. One can weaken the condition (4.2) by only requiring that it hold when $|\, x - y \,| < s$ for some $s > 0$. Lipschitz continuity is stronger than mere continuity, although functions that are continuous but not Lipschitz continuous are badly behaved. See also Theorem 7.16 below.

Remark: One can replace (4.2) by the condition

$$|\, f(x) - f(y) \,| \leq \lambda \,|\, x - y \,|^{\alpha} \tag{4.3}$$

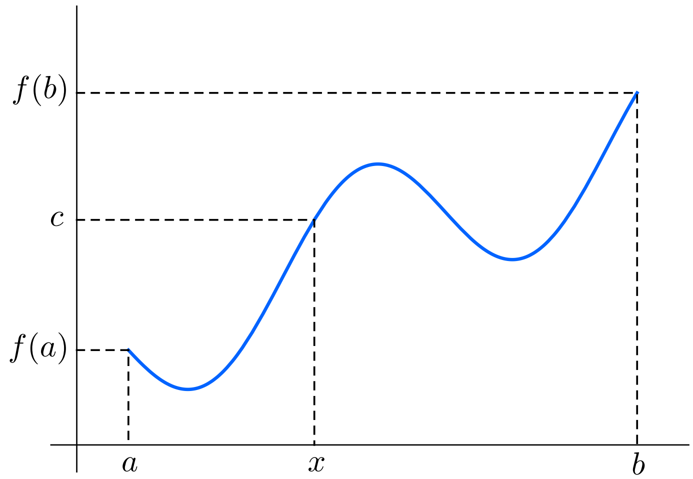


Figure 3. The Intermediate Value Theorem.

where $\alpha > 0$ is another constant; even more generally, one can replace $|\,x - y\,|^\alpha$ by $H\big(\,|\,x-y\,|\,\big)$ where $H\colon[\,0,\infty\,) \to [\,0,\infty\,)$ is any continuous strictly increasing function with $H(0) = 0$. Again, one can weaken either condition by only requiring it when $|\,x - y\,| < s$. Functions that satisfy (4.3) are said to be *Hölder continuous*, after Lipschitz's younger compatriot Otto Hölder. Since $t^\alpha > t$ when $0 < t < 1$, Hölder is weaker than Lipschitz, although not all continuous functions are Hölder continuous either.

Lemma 4.5. *Let $y \in \mathbb{R}$. Suppose $f\colon\mathbb{R} \to \mathbb{R}$ is continuous, and $f(x) \geq y$ for all $x \neq a$. Then $f(a) \geq y$. Similarly, if $f(x) \leq y$ for $x \neq a$, then $f(a) \leq y$.*

Proof: We just prove the first statement, leaving the second as an exercise for the reader. Suppose $f(a) < y$. Let $b < f(a) < c < y$. Then $f^{-1}(b,c) = \{\,a\,\}$ which is not open, contradicting our assumption that f is continuous. *Q.E.D.*

Remark: This result is not true for the strict inequality. Indeed, $f(x) = x^2$ satisfies $f(x) > 0$ for all $x \neq 0$ but $f(0) = 0$.

The next result is of fundamental importance, and is known as the Intermediate Value Theorem. Roughly speaking, it says that if a continuous function takes on two values, then it necessarily takes on all intermediate values; see Figure 3.

Theorem 4.6. *Let $f\colon\mathbb{R} \to \mathbb{R}$ be continuous. Suppose $a < b$. If $c \in \mathbb{R}$ satisfies either $f(a) < c < f(b)$, or $f(a) > c > f(b)$, then there exists $x \in [\,a,b\,]$ such that $f(x) = c$*

Proof: Suppose not, i.e., c is not in the range of f. Then $S_- = f^{-1}(-\infty,c)$ and $S_+ = f^{-1}(c,\infty\,)$ are nonempty open subsets since either $a \in S_-$ and $b \in S_+$ or the reverse. Moreover, since for every $x \in \mathbb{R}$, its image $f(x)$ lies in either $(\,-\infty,c\,)$ or $(\,c,\infty\,)$, we have $\mathbb{R} = S_- \cup S_+$ with $S_- \cap S_+ = \varnothing$. But this expresses $\mathbb{R}$ as the disjoint union of two nonempty open sets, which, by Theorem 2.8, is not possible. *Q.E.D.*

The Intermediate Value Theorem 4.6 is the reason we informally say that a function is continuous if one can draw its graph without lifting the pencil from the paper. Indeed, it implies that to draw the graph starting at one point $(a, f(a))$ and ending at another one $(b, f(b))$, one must pass through all the intervening points $(x, f(x))$ in a continuous manner without skipping any values.

5. Topology of and Continuity on Subsets of the Real Line.

In many situations, a function of interest is not defined on all of $\mathbb{R}$, but we are nevertheless interested in its continuity on its domain of definition. The goal of this section is to understand this in some detail.

If $A \subset \mathbb{R}$ is any subset, we can define a *topology* on A by declaring that any set of the form $U = S \cap A$, where $S \subset \mathbb{R}$ is open, is an *open subset* of A. In particular, the *basic open subsets* of A are the intersections $I \cap A$ of open intervals $I = (a, b)$ with A. As in the case of the real line, $U \subset A$ is open if and only if every $x \in U$ is contained in a basic open subset: $x \in I \cap A \subset U$. For example, if $A \subset \mathbb{R}$ itself is open, then, because the intersection of two open sets, $U = S \cap A$, is open, the open subsets of A are the open subsets $U \subset \mathbb{R}$ that are contained in A, whence $U \subset A \subset \mathbb{R}$.

A function $f: A \to \mathbb{R}$ is *continuous* (when restricted to A) if, whenever $S \subset \mathbb{R}$ is open, $f^{-1}(S) \subset A$ is open in A, i.e., a union of basic open subsets of A. As before, one only needs to check this when $S = J \subset \mathbb{R}$ is an open (bounded) interval. The following lemma states that the restriction of continuous functions to subsets remain continuous.

Lemma 5.1. *Suppose $B \subset A \subset \mathbb{R}$. Given $f: A \to \mathbb{R}$, let $\widetilde{f}: B \to \mathbb{R}$ denote its restriction to B, so that $\widetilde{f}(x) = f(x)$ whenever $x \in B \subset A$. If f is continuous on A, then its restriction $\widetilde{f}$ is continuous on B.*

Proof: Let $V \subset \mathbb{R}$ be open. Continuity of f on A implies $f^{-1}(V)$ is an open subset of A, and hence of the form $f^{-1}(V) = S \cap A$, where $S \subset \mathbb{R}$ is open. But then $\widetilde{f}^{-1}(V) = f^{-1}(V) \cap B = S \cap A \cap B = S \cap B$ since $B \subset A$. But the latter set is open in B, and hence $\widetilde{f}$ is continuous. *Q.E.D.*

For example, the function $f(x) = 1/x$ is continuous on its (open) domain $A = \mathbb{R} \setminus \{0\}$. Indeed, given a bounded open interval (a, b), we have

$$f^{-1}(a,b) = \begin{cases} (1/b, 1/a), & 0 < a < b, \\ (1/b, \infty), & 0 = a < b, \\ (-\infty, 1/a) \cup (1/b, \infty), & a < 0 < b, \\ (-\infty, 1/a), & a < b = 0, \\ (1/b, 1/a), & a < b < 0, \end{cases}$$

each of which is an open interval or the union of two open intervals, thus proving continuity. Any rational function $p(x)/q(x)$ is continuous on its domain $D = \{ x \in \mathbb{R} \mid q(x) \neq 0 \}$, i.e., outside the zeros of q.

Conversely, given a continuous function on a subset, it is not always possible to extend it to a continuous function on a larger subset. The function $f(x) = 1/x$ is continuous on $(0, \infty)$, but cannot be extended to a continuous function on any set containing the origin; this is because $f(x)$ is not bounded, and a continuous extension would violate the Boundedness Theorem 5.16 below. Even bounded continuous functions may not have continuous extensions; an example is the sign function $\operatorname{sign} x$, as noted above. A more interesting example is the trigonometric function $\sin(1/x)$, which is continuous on $(0, \infty)$, being the composition of continuous functions. It also cannot be continuously extended to any set containing the origin due to its increasingly rapid oscillations between -1 and 1 as x gets closer and closer to 0. On the other hand, as we will see, a function that is continuous on a closed interval can easily be continuously extended to all of $\mathbb{R}$.

Remark: The Intermediate Value Theorem 4.6 applies as stated when f is only defined on an open interval. It is *not* true on more general open sets; for example, the function $f(x) = 1/x$ is continuous on $\mathbb{R} \setminus \{0\}$, and $f(-1) = -1$, $f(1) = 1$, but there is no point $x \in \mathbb{R} \setminus \{0\}$ such that $f(x) = 0$.

Lemma 5.2. *Let $I \subset \mathbb{R}$ be an open interval. A continuous function $f\colon I \to \mathbb{R}$ is one-to-one if and only if it is* strictly monotone*, meaning either it is strictly increasing, so that $f(x) < f(y)$ whenever $x < y$, or strictly decreasing, so that $f(x) > f(y)$ whenever $x < y$.*

Proof: Suppose f is not strictly monotone. If $f(a) = f(b)$ for $a < b$, then f is not one-to-one. So not being strictly monotone would imply that one can find $a < b < c$ in I such that either $f(a) < f(b) > f(c)$ or $f(a) > f(b) < f(c)$. Let us deal with the first case, leaving the second to the reader. Let $z \in \mathbb{R}$ satisfy both $f(a) < z < f(b)$ and $f(c) < z < f(b)$. By the Intermediate Value Theorem 4.6, we can find $a < x < b$ such that $f(x) = z$, and, again, $b < y < c$ such that $f(y) = z$. But this implies $x \neq y$ whereas $f(x) = f(y)$, in contradiction to our assumption that f is one-to-one. *Q.E.D.*

Theorem 5.3. *Let $I \subset \mathbb{R}$ be an open interval. If $f\colon I \to \mathbb{R}$ is continuous and one-to-one, then its range $J = f(I)$ is an open interval and its inverse $f^{-1}\colon J \to I$ is continuous.*

Proof: According to Lemma 5.2, f must be strictly monotone. Let us, for definiteness, assume that f is strictly increasing, so $f(x) < f(y)$ whenever $x < y$. Then, given an open interval $(a, b) \subset I$, if $a < x < b$, then $f(a) < f(x) < f(b)$. Moreover, by the Intermediate Value Theorem 4.6, given any c with $f(a) < c < f(b)$ there exists $x \in (a, b)$ such that $f(x) = c$. Putting these together, we conclude that the image of the open interval (a, b) under f is the open interval $f(a, b) = (f(a), f(b))$. This implies f maps open sets to open sets. Since $f = (f^{-1})^{-1}$ is the inverse of f^{-1}, we conclude that f^{-1} satisfies the condition of Definition 4.1, and hence is continuous. *Q.E.D.*

Remark: Since Proposition 2.7 implies every open set is the union of a disjoint collection of open intervals, Theorem 5.3 holds as stated with "open interval" replaced by "open set".

Example 5.4. The function $f_3(x) = x^3$ is continuous and strictly monotone. Its inverse is the cube root function $f_3^{-1}(y) = y^{1/3}$. Similar remarks apply to $f_n(x) = x^n$ where n is any odd positive integer. On the other hand, the $f_2(x) = x^2$ is not monotone on $\mathbb{R}$, and hence not one-to-one, and its range $f_2(\mathbb{R}) = [0, \infty)$ is not open. Each y has two possible real square roots. If we restrict f_2 to the positive reals $\mathbb{R}^+ = (0, \infty)$ then the resulting function $f_2^+ : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly monotone increasing, and its inverse $(f_2^+)^{-1}(y) = +\sqrt{y}$, the positive square root function, is continuous on $\mathbb{R}^+$. Similarly, the restriction of f_2 to the negative reals, denoted $f_2^- : \mathbb{R}^- \to \mathbb{R}^+$, where $\mathbb{R}^- = \{x < 0\}$, is strictly monotone decreasing, and so its inverse $(f_2^-)^{-1} : \mathbb{R}^+ \to \mathbb{R}^-$ given by $(f_2^-)^{-1}(y) = -\sqrt{y}$ is continuous.

Let us next discuss the case of a half-open interval $A = (a, b]$, where $a < b$ and we allow $a = -\infty$. The basic open subsets are the open subintervals (c, d) with $a \le c < d \le b$, and the half-open subintervals $(c, b]$ with $a \le c < b$, including $(a, b]$ itself. If $f : (a, b] \to \mathbb{R}$ is continuous when restricted to $(a, b]$, then we say that f *is continuous from the right* at the endpoint b. On the other hand, according to Lemma 5.1 and the preceding paragraph, the restriction of f to the open subinterval $(a, b) \subset (a, b]$ coincides with ordinary continuity there. The analogous "mirror image" results hold for a half-open interval $B = [a, b)$, and one speaks of *continuity from the left* at the endpoint a of a continuous function $f : B \to \mathbb{R}$. Finally, assuming $a < b$, if $C = [a, b]$ is a closed interval, then its basic open sets are the open subintervals (c, d) where $a \le c < d \le b$, the half-open subintervals $[a, c)$ and $(c, b]$ for $a < c < b$, and the entire closed interval $[a, b]$. If $f : [a, b] \to \mathbb{R}$ is continuous, we say it is continuous from the left at the endpoint a and continuous from the right at the endpoint b. Its restriction $f : (a, b) \to \mathbb{R}$ to the open subinterval is continuous in the usual sense.

The following result allows one to stitch together continuous functions on subdomains to obtain a continuous function on a larger domain. First we state a basic lemma.

Lemma 5.5. *Let $a < b < c$, where we allow $a = -\infty$ and/or $c = \infty$. Suppose U is open in $(a, b]$ and V is open in $[b, c)$. Then $U \cup V \subset (a, c)$ is open (also as a subset of $\mathbb{R}$) if and only if either $b \notin U$ and $b \notin V$, or $b \in U$ and $b \in V$.*

Proof: The first case implies $U \subset (a, b)$ and $V \subset (b, c)$, which implies that both U and V are open subsets of $\mathbb{R}$ and hence their union is open. In the second case, since $b \in U \subset (a, b]$, there is a half open interval $(p, b] \subset U$ for some $a < p < b$. Similarly, there is a half open interval $[b, q) \subset V$ for some $b < q < c$. But then the open interval $(p, q) \subset U \cup V$ contains b. Thus, every point in $U \cup V$, including b, belongs to an open interval that is contained therein, which, by Lemma 2.6, proves $U \cup V$ is open. *Q.E.D.*

Theorem 5.6. *Let $a < b < c$, where we allow $a = -\infty$ and/or $c = \infty$. Suppose $f : (a, b] \to \mathbb{R}$ and $g : [b, c) \to \mathbb{R}$ are continuous and $f(b) = g(b) = z$. Then the combined*

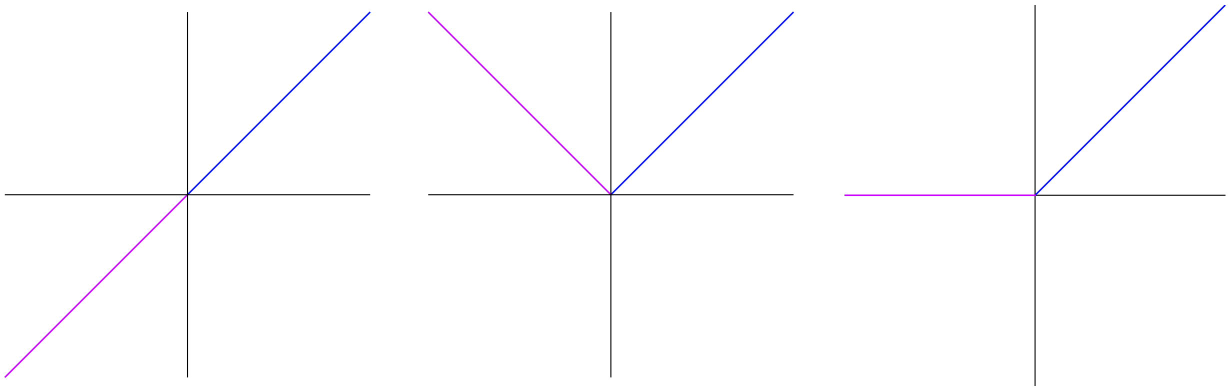


Figure 4. Extensions of $f(x) = x$.

function $h\colon (a,c) \to \mathbb{R}$ *given by*

$$h(x) = \begin{cases} f(x), & a < x < b, \\ z, & x = b, \\ g(x), & b < x < c, \end{cases} \tag{5.1}$$

is continuous. The same holds when one or both domains are closed intervals.

Proof: Let $S \subset \mathbb{R}$ be open, and let $U = f^{-1}(S)$, which is open in $(a,b]$, and $V = g^{-1}(S)$, which is open in $[b,c)$. If $z \notin S$, then $b \notin U$ and $b \notin V$; on the other hand, if $z \in S$, then $b \in U$ and $b \in V$. In either case, Lemma 5.5 implies $U \cup V = h^{-1}(S)$ is open in (a,c), and hence h is continuous. *Q.E.D.*

Remark: Under the set-up of the Stitch Theorem 5.6, if $f(b) \neq g(b)$, no matter how one defines $h(b)$ in (5.1), one cannot stitch the two functions together to form a continuous function on the combined interval (a,c).

Example 5.7. The function $g(x) = x$ is continuous on the positive half line $[0,\infty)$. The function $f(x) = -x$ is continuous on the negative half line $(-\infty,0]$. Moreover $f(0) = g(0) = 0$. Thus, Theorem 5.6 implies that the combined absolute value function $h(x) = |\,x\,|$ is continuous on all of $\mathbb{R}$. Alternatively, one can write $|\,x\,| = \max\{x, -x\}$ and use Theorem 4.3 to prove continuity.

There are, of course, many other ways to extend the function $g(x)$ to the negative half line while maintaining continuity, besides setting $h(x) = x$ or $|\,x\,|$. For example, setting $f(x) \equiv 0$ for all $x \leq 0$ leads to the continuous extension

$$h(x) = \max\{0,x\} = \begin{cases} x, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{5.2}$$

known as the *ramp function*, or in modern data analysis of neural networks, the *rectified linear unit* or *ReLU function*, [**14**]; see Figure 4.

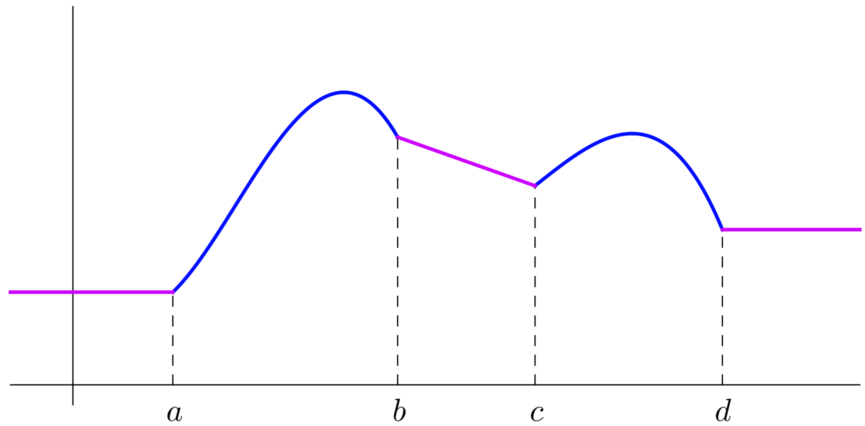


Figure 5. Extension of a Continuous Function.

The last example indicates one how one can easily extend a continuous function defined on a closed interval to a continuous function defined on the entire real line. Its proof follows by applying Theorem 5.6 to both endpoints.

Proposition 5.8. *Let $a < b$, and suppose $f\colon [a,b] \to \mathbb{R}$ is continuous. Then*

$$h(x) = \begin{cases} f(a), & x \leq a, \\ f(x), & a \leq x \leq b, \\ f(b), & x \geq b, \end{cases} \tag{5.3}$$

defines a continuous function $h\colon \mathbb{R} \to \mathbb{R}$.

Of course, as before, (5.3) is but one of many possible extensions. More generally, suppose $A = [a,b] \cup [c,d]$ is the disjoint union of two closed intervals, with $a < b < c < d$. Then

$$h(x) = \begin{cases} f(a), & x < a, \\ f(x), & a \leq x \leq b, \\ \dfrac{f(c)\,(x-b) - f(b)\,(x-c)}{c-b}, & b < x < c, \\ f(x), & c \leq x \leq d, \\ f(d), & x > d, \end{cases} \tag{5.4}$$

defines a continuous function $h\colon \mathbb{R} \to \mathbb{R}$. Here one has extended f by constant values on the two complementary unbounded intervals $(-\infty, a]$ and $[d, \infty)$, but on the intermediate interval $[b,c]$ one uses the linear *interpolating function* whose graph is the straight line passing through the points $(b, f(b))$ and $(c, f(c))$; see Figure 5. One can clearly adapt this construction to a finite collection of disjoint closed intervals, but it does not necessarily work on an infinite collection.

All results, both above and below, that were stated when the domain of the function is all of $\mathbb{R}$ also hold, suitably rephrased, when the domain of f is an open interval. They also hold, provided care is taken at the endpoints, on half-open and closed intervals. However, they do not necessarily apply to more general subsets, which must be handled on their own merits, or lack thereof.

There are, of course, many other subsets $A \subset \mathbb{R}$ besides intervals, and the continuity of functions $f\colon A \to \mathbb{R}$ thereon is also of interest. A particularly important case is when the subset $A \subset \mathbb{R}$ is a countably discrete set of points possessing a "limit point".

Example 5.9. Consider the countably infinite closed (as per the remarks after Proposition 2.11) subset

$$A = \{\,0\,\} \cup \widetilde{A}, \quad \text{where} \quad \widetilde{A} = \{\,1/n \mid \ n \in \mathbb{N}\,\}. \tag{5.5}$$

Let us determine the basic open subsets of A. Let $a < b$ be real numbers. If $0 < a < 1$, then $(a,b) \cap A$ consists of a finite number of adjacent points $1/n \in \widetilde{A}$, namely those for which $1/b < n < 1/a$. In particular, given $n \in \mathbb{N}$, if we choose any a,b such that $n-1 < a < n < b < n+1$, then $(a,b) \cap A$ is the single point $\{1/n\} \in A$. Thus, in this case, unlike $\mathbb{R}$ or subintervals of nonzero length in $\mathbb{R}$, singleton sets can be open! If $a = 0 < b$, then $(a,b) \cap A$ consists of an infinite collection of adjacent points, namely those for which $n > 1/b$. On the other hand, if $a < 0 < b$, then $(a,b) \cap A$ consists of an infinite collection of adjacent points, namely those for which $n > 1/b$, combined with the point 0. Finally if $a \geq 1$ or $b \leq 0$, then $(a,b) \cap A = \varnothing$. General open subsets are unions of these basic ones. In particular, *any* subset of $\widetilde{A} = \{\,1/n \,|\, n \in \mathbb{N}\,\}$, either finite or infinite, is open. On the other hand, the singleton set $\{\,0\,\}$ is not open. Indeed, if the open set $S \subset A$ contains 0, then it contains a basic open set $(a,b) \cap A \subset S$ for some $a < 0 < b$, which means that it contains the infinite subset $\widetilde{A}_b = \{\,1/n \,|\, 1/b \leq n \in \mathbb{N}\,\} \subset S$. Note that S can also contain any collection of points in $\widetilde{A} \setminus \widetilde{A}_b = \{\,1/n \,|\, 0 < n < 1/b,\ n \in \mathbb{N}\,\}$; since all subsets of $\widetilde{A}$ are open, this does not affect the openness of S.

Consider a function $f\colon A \to \mathbb{R}$. Continuity of f requires that $f^{-1}(c,d) \subset A$ be open for any (bounded) open interval (c,d). Let $z = f(0)$. If $z \notin (c,d)$, then $0 \notin f^{-1}(c,d)$ and, since any subset of $\widetilde{A}$ is open, there are no conditions imposed by the continuity of f. On the other hand, if $z \in (c,d)$, then continuity requires that $f^{-1}(c,d)$ be an open subset of A containing 0, and hence, by the preceding paragraph, must contain an infinite subset of the form $\widetilde{A}_b = \{\,1/n \,|\, 1/b \leq n \in \mathbb{N}\,\}$ for some $b > 0$. In other words, $f\colon A \to \mathbb{R}$ is continuous if, for every $c < f(0) < d$, there exists a natural number $m = m_{c,d} \in \mathbb{N}$ such that $c < f(1/n) < d$ for all $n \geq m$. Note that the number m depends on c,d, and we expect that the smaller $d - c$ is, the larger m needs to be.

Keep in mind that if $f\colon \mathbb{R} \to \mathbb{R}$ is continuous, then its restriction to A is automatically continuous as a function $f\colon A \to \mathbb{R}$. Conversely, given a continuous function $f\colon A \to \mathbb{R}$, one can construct a continuous extension $h\colon \mathbb{R} \to \mathbb{R}$ by adapting the construction in (5.4). Namely, on the interval $(\,1/(n+1), 1/n\,)$, one uses the linear interpolating function between the values of f at the endpoints, and then extends h to be constant for $x \geq 1$ and for $x \leq 0$.

In other words,

$$h(x) = \begin{cases} f(0), & x \le 0, \\ \dfrac{f(1/n)(x - 1/(n+1)) - f(1/(n+1))(x - 1/n)}{1/n - 1/(n+1)} & \\ \quad = n\, f\left(\dfrac{1}{n}\right) \big(\,(n+1)x - 1\,\big) - (n+1)\, f\left(\dfrac{1}{n+1}\right)(n\,x - 1), & \dfrac{1}{n+1} \le x < \dfrac{1}{n}, \\ f(1), & x \ge 1, \end{cases} \tag{5.6}$$

for all $n \in \mathbb{N}$. The proof that this function is continuous on the entire real line is left as an exercise for the reader. We will return to this interesting and important example when we discuss limits in Section 6.

Let us use this example to motivate the definition of a limit point — also known as an *accumulation point* — of a set.

Definition 5.10. Let $A \subset \mathbb{R}$. A point $\ell \in \mathbb{R}$ is called a *limit point* of A if, whenever S is an open subset containing ℓ, we have $A \cap (S \setminus \{\ell\}) \neq \varnothing$.

We delete ℓ from S in the limit point condition since otherwise every point in A would be a limit point, which is not what we want. It is, however, possible that some points in A are limit points. For example, if $A = (\,a, b\,)$ is an open interval with $a < b$ finite, then every point in the closed interval $[\,a, b\,]$ is a limit point of A. On the other hand, the singleton set $\{\,a\,\}$ has no limit points.

One need only check the limit point condition when $S = I$ is a (sufficiently small) interval containing l. Thus, in the case of Example 5.9, $\ell = 0$ is a limit point for the sets A and $\widetilde{A}$. Indeed, given any interval $0 \in (\,a, b\,)$, so $a < 0 < b$, then

$$A \cap \big[\,(\,a, b\,) \setminus \{\,0\,\}\,\big] = \{\,1/n \mid 1/b < n \in \mathbb{N}\,\} \neq \varnothing.$$

On the other hand, there are no other limit points, because if $\ell \neq 0$, then any sufficiently small interval containing ℓ either contains no points in A, or, when $\ell = 1/n \in A$, contains just one point in A, namely $\ell = 1/n$ itself.

Lemma 5.11. *If $A \subset B$ then every limit point of A is also a limit point of B.*

Of course, B may have more limit points than A. One situation where this is not the case is the following:

Lemma 5.12. *Let $F \subset A$ be a finite subset. Then every limit point of A is a limit point of $A \setminus F$ and vice versa.*

Another fundamental property of the real line is the Limit Point Theorem, also known as the Bolzano–Weierstrass Theorem after two of the nineteenth century developers of modern rigorous mathematical analysis: the Czech precursor Bernard Bolzano and the influential German founder Karl Weierstrass.

Theorem 5.13. *Let $A \subset \mathbb{R}$ be a bounded infinite subset. Then the set of limit points of A is not empty.*

Proof: Suppose $A \subset J_0 = [a, b]$ where $a < b$. We cut $[a, b]$ in half, producing the intervals $[a, m]$ and $[m, c]$ where $m = \frac{1}{2}(a + b)$ is the midpoint, and so each subinterval has length $\frac{1}{2}(b - a)$. Since A is infinite, at least one of the two subintervals contains infinitely many points in A. We choose it — if both contain infinitely many points, it does not matter which of the two we select. Let the chosen interval be labelled $J_1 = [a_1, b_1]$, where either $a_1 = a$, $b_1 = m$, or $a_1 = m$, $b_1 = b$. We then divide the interval J_1 into two half-size intervals, of lengths $\frac{1}{4}(b - a)$ and choose one of the two, labelled $J_2 = [a_2, b_2]$, that contains infinitely many points in A. We repeat the process indefinitely, so that, for each $k \in \mathbb{N}$, we have constructed an interval

$$J_k = [a_k, b_k] \qquad \text{of length} \qquad b_k - a_k = 2^{-k}(b - a) \tag{5.7}$$

that contains infinitely many points in A. Note further that, by the construction methodology, the endpoints of these intervals satisfy

$$a \le a_1 \le a_2 \le a_3 \le \ \cdots \ \le b_3 \le b_2 \le b_1 \le b.$$

Thus, taking the respective supremum and infinimum, we see

$$c = \sup\{ a_k \} \le d = \inf\{ b_k \}.$$

We claim that $c = d$ and, moreover, their common value $\ell = c = d$ is a limit point of A. First, to prove equality, suppose $c < d$. Then, for any $k \in \mathbb{N}$,

$$a_k \le c < d \le b_k,$$

but these inequalities are not possible when $b_k - a_k < d - c$, which, according to (5.7), will occur once k is sufficiently large. Finally, to prove $\ell = c = d$ is a limit point, given any open interval $I = (\ell - r, \ell + r)$ centered on ℓ of length $2r > 0$, since $a_k \le \ell \le b_k$, once k is sufficiently large so that $b_k - a_k < \frac{1}{2} r$, we must have $J_k = [a_k, b_k] \subset I$, and hence I also contains infinitely many points in A. *Q.E.D.*

Remark: The result in Theorem 5.13 is false if A is unbounded. For example $\mathbb{N}$ and $\mathbb{Z}$ have no limit points.

Proposition 5.14. *A subset $C \subset \mathbb{R}$ is closed if and only if all its limit points ℓ belong to it: $\ell \in C$.*

Proof: If C is closed, then $S = \mathbb{R} \setminus C$ is open. If $x \in S$, then, since $S \cap C = \varnothing$, the point x does not satisfy the limit point criterion using S itself as the open set. Conversely, if every limit point belongs to C, then given $x \in S = \mathbb{R} \setminus C$, there must exist an open interval I containing x with $I \cap C = \varnothing$, as otherwise x would be satisfy the limit point criterion. But this means $I \subset S$. Since this holds for all $x \in S$, Lemma 2.6 implies S is open and so C is closed. *Q.E.D.*

Warning: While every limit point of a closed subset C is contained in it, not every point of C is a limit point. For example, the singleton set $C = \{a\}$ has no limit points; nor does any finite subset $C = \{a_1, \ldots, a_n\} \subset \mathbb{R}$. Similarly, the only limit point of the closed set A in Example 5.9 is 0.

Proposition 5.15. *Let $A \subset \mathbb{R}$ and let L be the set of all its limit points. Then $C = A \cup L$ is closed and in fact $C = \overline{A}$ is the closure of A.*

Proof: If $x \in \mathbb{R} \setminus C$, then $x \notin A$ and $x \notin L$ is not a limit point of A. This implies that there is an open interval I containing x such that $I \cap C = \varnothing$. But this implies $\mathbb{R} \setminus C$ is open, and hence C is closed. Thus $\overline{A} \subset C$ because the closure is the smallest closed subset containing A. On the other hand, if $\widetilde{C} \supset A$ is any other closed subset, then Lemma 5.11 implies $L \subset \widetilde{C}$, and hence $C \subset \widetilde{C}$. Thus, $C = \overline{A}$. *Q.E.D.*

In particular, if C is a closed subset, Proposition 5.14 implies that it is its own closure $C = \overline{C}$. Combining Propositions 5.14 and 5.15, we deduce that a limit point of limit points is itself a limit point. In other words, the set of limit points of $C = A \cup L$ also equals L, the set of limit points of A.

The closure of an open interval (a, b) when $a < b$ are finite, as well as the corresponding half open intervals $[a, b)$ or $(a, b]$, is the closed interval $[a, b]$, which, by the above remarks, is its own closure. The only limit point of the set $\widetilde{A}$ in Example 5.9 is 0, and hence its closure is $A = \widetilde{A} \cup \{0\}$.

A particularly important class are the continuous functions on bounded closed intervals, also known as *compact intervals*. The next result, known as the Boundedness Theorem, says that, on such a domain, a continuous function cannot go off to infinity.

Theorem 5.16. *Let $a < b$ be finite, and suppose that $f: [a, b] \to \mathbb{R}$ is continuous. Then f is* bounded, *meaning that its range $f[a, b]$ is a bounded subset of $\mathbb{R}$, and so there exist $c \leq d$ such that $c \leq f(x) \leq d$ for all $x \in [a, b]$.*

Remark: It is essential that the interval be closed and hence bounded for this result to be valid. For example, the continuous function $f(x) = x$ is not bounded on any unbounded interval. Similarly, the function $f(x) = 1/x$ is continuous but not bounded on the open interval $(0, 1)$ nor on the half open interval $(0, 1]$.

Proof: Suppose that $f(x)$ is not bounded from above for $x \in [a, b]$. In this case, for every $n \in \mathbb{N}$, there is a point $x_n \in [a, b]$ such that $f(x_n) \geq n$. According to the Limit Point Theorem 5.13, the infinite set $\{x_n\}$ has a limit point $\ell \in [a, b]$. Let $p < f(\ell) < m \in \mathbb{N}$. Then, by continuity of f, the open subset $f^{-1}(p, m)$ contains ℓ and hence contains infinitely many points x_n. But this implies $f(x_n) < m$ for all $x_n \in f^{-1}(p, m)$, which contradicts our specification of the points x_n. The proof of the existence of a lower bound is similar. *Q.E.D.*

As a consequence of the preceding developments, we arrive at the Extreme Value Theorem, which states that a continuous function on a bounded closed interval is not only

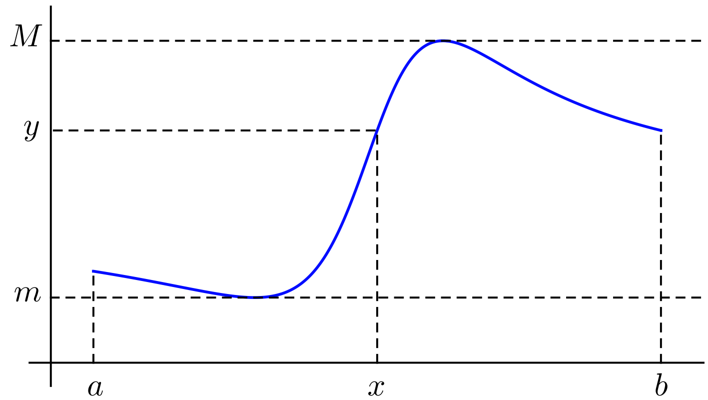


Figure 6. The Extreme Value Theorem.

bounded from above and below, but actually achieves its maximum and minimum values, as well as every value in between. Again, the fact that the interval is both closed and bounded is essential; on open or unbounded intervals there are many counterexamples, as noted above.

Theorem 5.17. *Let $a < b$ be finite. If $f\colon [a,b] \to \mathbb{R}$ is continuous, then its range is a closed interval: $f[a,b] = [m, M]$, where $M = \sup f[a,b]$ and $m = \inf f[a,b]$. This means that, given any $m \le y \le M$, there exists at least one point $x \in [a,b]$ such that $f(x) = y$.*

Proof: By definition of supremum, $f(x) \le M$ for all $x \in [a,b]$. Suppose that $f(x) < M$ for all $x \in [a,b]$. Then the function $g(x) = M - f(x) > 0$ for all $x \in [a,b]$. Thus, by Theorem 4.3, the function $h(x) = 1/g(x)$ is continuous. The Boundedness Theorem 5.16 implies $h(x) \le \beta$ for all $x \in [a,b]$ for some $0 < \beta \in \mathbb{R}$. But this implies $g(x) \ge 1/\beta$, and hence $f(x) = M - g(x) \le M - 1/\beta$ for all $x \in [a,b]$. Thus, $M - 1/\beta < M$ is an upper bound for $f[a,b]$, which contradicts the assumption that M is the least upper bound. A similar proof works for the greatest lower bound m; alternatively one can replace $f(x)$ by $\widetilde{f}(x) = -f(x)$ whose least upper bound is $-m$. In this manner, we have proved the existence of at least one point $x_M \in [a,b]$ with $f(x_M) = M$ and at least one point $x_m \in [a,b]$ with $f(x_m) = m$. The final statement of the theorem then follows from the Intermediate Value Theorem 4.6. *Q.E.D.*

On the other hand, suppose $f\colon (a,b) \to R$ is continuous on an open interval. The Intermediate Value Theorem 4.6 combined with Lemma 2.17 implies that its range $f(a,b)$ is an interval, but, unless f is one-to-one in which case Theorem 5.3 applies, then it is *not* necessarily true that the image $f(a,b)$ is an open interval. For example, if $f(x) = x^2$ and $a < 0 < b$ then the image of the open interval (a,b) is the half open interval $[0,c) = f(a,b)$ where $c = \max\{a^2, b^2\}$.

The one case we omitted above is when the closed interval consists of a single point $A = \{a\} = [a,a]$, but this is completely trivial. There are only two subsets, $\varnothing$ and

A itself, which are both relatively open (and closed). Thus *any* function $f: A \to \mathbb{R}$ is continuous. However, there is a more refined concept of continuity of a function $f: \mathbb{R} \to \mathbb{R}$ at a point a, that means more than just its restriction to $A = \{a\}$ be continuous, which is always true. The following definition captures the essence.

Definition 5.18. A function $f: \mathbb{R} \to \mathbb{R}$ is said to be *continuous at* $a \in \mathbb{R}$ if, for every open subset $f(a) \in V \subset \mathbb{R}$, there exists an open subset $a \in U \subset f^{-1}(V)$.

Clearly, if f is continuous, then it is continuous at each point a since we can set $U = f^{-1}(V)$ which is open by the definition of continuity. Similarly, if f is piecewise continuous, then it is continuous at every point a which is not a discontinuity.

Remark: By *piecewise continuous*, we mean that the function is continuous on a disjoint collection of open intervals such that its domain is contained in the union of the corresponding closed intervals. An example is the sign function (4.1) which is continuous on the subintervals $(-\infty, 0)$ and $(0, \infty)$. Another example is the piecewise constant function $f(x) = n$ when $n \le x < n+1$ for $n \in \mathbb{Z}$, which is continuous on the intervals $(n, n+1)$. The term *piecewise constant* similarly means that the function is constant on each interval belonging to such a collection.

Proposition 5.19. *A function $f: \mathbb{R} \to \mathbb{R}$ is continuous if and only if it is continuous at every point $a \in \mathbb{R}$.*

Proof: We just showed that the direct statement is valid. To prove the converse, suppose $V \subset \mathbb{R}$ is open. We need to show $f^{-1}(V)$ is open. Let $a \in f^{-1}(V)$. Then, according to Definition 5.18 we can find an open set U containing a and contained in $f^{-1}(V)$, but, in view of Lemma 2.6, this suffices to prove that $f^{-1}(V)$ is open. *Q.E.D.*

Example 5.20. Here is a nontrivial example of a function that is not continuous on any interval, but is continuous at a single point. Set

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q}, \\ x, & x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases} \tag{5.8}$$

Since it assumes differing values on the rationals and the irrationals, the function (5.8) is highly discontinuous; indeed, it is not continuous when restricted to any interval, closed or open, of nonzero length. However, it is continuous at the origin. To show this, consider an open interval $(-s, s)$ for $s > 0$. Then

$$f^{-1}(-s, s) = (-s, s) \cup \mathbb{Q},$$

which is not open, but does contain the open interval $(-s, s)$. Since this holds for any such interval, it also holds, by our usual arguments, for any open $0 \in V \subset \mathbb{R}$. Be that as it may, I would argue that such pathological discontinuous functions, while important drivers behind the development of modern real analysis, [**24**], play no appreciable role in any of the usual applications of calculus, especially those encountered by students from other disciplines, and are only of interest to the mathematical diehard.

Remark: Definition 5.18 is, in fact, equivalent to the usual limit-based definition of continuity used in almost all calculus texts, including [**1**, **6**, **26**, **28**]. The delta-epsilon mantra learned by all students of calculus and analysis over the past 150 years goes as follows: *The function $f(x)$ is continuous at $x = a$ if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for every x satisfying $|\, x - a \,| < \delta$, we have $|\, f(x) - f(a) \,| < \varepsilon$.* To make the connection explicit, given $\varepsilon > 0$, let $J_\varepsilon = (\, f(a) - \varepsilon, f(a) + \varepsilon \,)$ be the open interval of length 2ε centered at $f(a)$. Definition 5.18 tells us that there is an open set $a \in U \subset f^{-1}(J_\varepsilon)$, and hence we can find $\delta > 0$ such that the open interval $I_\delta = (\, a - \delta, a + \delta \,) \subset U$. Thus, $x \in I_\delta$ if and only if $|\, x - a \,| < \delta$. Since $I_\delta \subset f^{-1}(J_\varepsilon)$, this implies $f(x) \in J_\varepsilon$, which is equivalent to saying $|\, f(x) - f(a) \,| < \varepsilon$. More generally, given any open V containing $f(a)$, one can select $\varepsilon > 0$ such that $J_\varepsilon \subset V$, and then set $U = I_\delta \subset f^{-1}(J_\varepsilon) \subset f^{-1}(V)$ in order to satisfy the condition in Definition 5.18. Proposition 5.19 demonstrates that continuity at each point implies continuity of the function f, meaning that it satisfies our basic Definition 4.1. Thus, our continuity-based approach is entirely equivalent to the traditional delta-epsilon approach, and loses nothing in generality. I hope you will be convinced that the continuity approach is easier both conceptually and developmentally, and thereby makes a compelling alternative to the standard treatment of the foundations of calculus.

Closed bounded intervals are particular examples of an important class of sets.

Definition 5.21. A set $K \subset \mathbb{R}$ is called *compact* if it is both *closed* and *bounded*.

The simplest compact subsets are closed bounded intervals and finite unions thereof. A less trivial example is provided by the Cantor set C constructed in Example 2.2.

As an immediate consequence of Theorems 5.13 and 5.14, one deduces:

Theorem 5.22. *If K is compact and $A \subset K$ is an infinite subset, then A has a limit point $\ell \in K$.*

The Boundedness Theorem 5.16 holds as stated, with an identical proof, when the interval $[\, a, b \,]$ is replaced by a compact set K. The first part of the Extreme Value Theorem 5.17 remains valid, namely that a function on a compact set achieves its minimum and maximum values; the adapted proof is left to the reader.

Theorem 5.23. *Let $K \subset \mathbb{R}$ be compact. If $f\colon K \to \mathbb{R}$ is continuous, then $m = \inf f(K)$ and $M = \sup f(K)$ are both finite. Moreover, there exist points $x_m, x_M \in K$ such that $f(x_m) = m$, $f(x_M) = M$.*

On the other hand, unless the set is also *connected*, meaning it is a closed bounded interval, the function does not necessarily achieve all the intermediate values between its minimum and maximum. For example, the function $f(x) = x - 1$ restricted to $K = [0, 1] \cup [2, 3]$ has $-1 = f(0) = \inf f(K)$, $2 = f(3) = \sup f(K)$, but there is no $x \in K$ such that $f(x) = y$ when $0 < y < 1$.

The most important topological feature of compact sets, including compact intervals, is that they satisfy the finite covering property, as we now describe.

Definition 5.24. Let $S \subset \mathbb{R}$. A collection of open subsets $\mathcal{U} = \{U\}$ is called a *cover* of S if for every $x \in S$ there exists an open set $U \in \mathcal{U}$ such that $x \in U$. Equivalently, S is contained in the union of all the open sets in the cover: $S \subset \bigcup_{U \in \mathcal{U}} U$.

Note especially that we do not require that the open sets in the cover satisfy $U \subset S$.

Example 5.25. Let $K = [0,1]$. Then, given a natural number $n > 0$, the (finite) collection of open intervals $I_k = ((k-1)/n, (k+1)/n)$ for $k = 0, \ldots, n$, all of length $2/n$, forms a cover of K. Indeed, every point $0 \le x < 1$ lies in an interval $k/n \le x < (k+1)/n$ for some $k = 0, \ldots, n-1$, and hence lies in I_k (and, unless $x = k/n$, also in I_{k+1}), while $x = 1 \in I_n$.

Example 5.26. As another example, consider the collection $\mathcal{V}$ consisting of the infinitely many open intervals $J_n = (1/(n+1), 1/n)$ for $n \in \mathbb{N}$. Then $\mathcal{V}$ forms a cover of the open interval $S = (0,1)$ but not the closed interval $K = [0,1]$, since it does not include its endpoints. If we enlarge $\mathcal{V}$ by including a couple of open intervals that cover the endpoints, say $\mathcal{U} = \mathcal{V} \cup \{(-s,s),\ (1-s, 1+s)\}$ for some $s > 0$, then $\mathcal{U}$ covers all of $[0,1]$. But now only finitely many of the preceding intervals J_n are required to cover $[0,1]$, since as soon as $n > 1/s$ then $J_n \subset (-s,s)$, and hence is not needed in the cover. On the other hand, no finite subcollection of $\mathcal{V}$ will cover the open interval $(0,1)$.

These examples are indicative of a general property of compact sets. Only finitely many members of any cover are needed to cover the set. This important fact is known as the Heine–Borel Theorem, named after the nineteenth century German mathematician Eduard Heine and the early twentieth century French mathematician Émile Borel.

Theorem 5.27. *Let $K \subset \mathbb{R}$ be compact. Let $\mathcal{U}$ be a cover of K by open sets. Then $K \subset U_1 \cup \cdots \cup U_n$ can be covered by finitely many of the open sets $U_1, \ldots, U_n \in \mathcal{U}$.*

Remark: In point set topology, [**12**], the finite cover property in Theorem 5.27 is often taken as the definition of a compact set. In this event, the corresponding theorem becomes that any compact subset of $\mathbb{R}$ is closed and bounded. The proof of this latter result is left as an exercise for the reader.

Proof: Given $x \in K$, let

$$r(x) = \sup\{\, r > 0 \mid \text{ there exists } U \in \mathcal{U} \text{ such that } (x-r,\ x+r) \subset U \,\}.$$

In other words, $r(x)$ is the supremum of the sizes (half-lengths) of open intervals centered at x that are contained in an open set in the cover. Since, by the assumed covering property, every $x \in U$ for at least one such U, and, being open, U contains such an open interval, we deduce that $r(x) > 0$ for all $x \in K$. Furthermore, since $(x-s,\ x+s) \subset (x-r,\ x+r)$ whenever $s < r$, we have $(x-s,\ x+s) \subset U$ for some $U \in \mathcal{U}$ whenever $0 < s < r(x)$.

Now set

$$s^* = \inf \{ r(x) \mid x \in K \}.$$

We claim that $s^* > 0$. If true, then we proceed to cover K as follows: First note that for any $0 < s < s^*$, every point $x \in K$ satisfies $r(x) \geq s^* > s$, and hence is contained in an interval $(x - s, x + s) \subset U$ for some $U \in \mathcal{U}$. Let us first suppose that $K = [a, b]$ is a closed bounded interval with $a < b$. (The case $a = b$ is trivial.) Choose $n \in \mathbb{N}$ such that $s = (b - a)/(n - 1) < s^*$, and let $a = x_1 < x_2 < \cdots < x_n = b$ be equally spaced points, so $x_{k+1} - x_k = s$ for $k = 1, \ldots, n - 1$. By the same reasoning as in Example 5.25, the intervals $I_k = (x_k - s, x_k + s)$ cover K. But then each $I_k \subset U_k$ for some $U_k \in \mathcal{U}$, and hence $U_1, \ldots, U_n$ also cover $[a, b]$, forming the claimed finite cover.

More generally, given a compact set $K \subset \mathbb{R}$, since it is bounded, there exists a compact interval $[a, b] \supset K$. Let $V = \mathbb{R} \setminus K$, which is open because K is closed. Then the enlarged collection $\widetilde{\mathcal{U}} = \mathcal{U} \cup \{V\}$ covers $[a, b]$ because $\mathcal{U}$ covers K, while V covers the remainder $[a, b] \setminus K$. Let $U_0 = V$ and $U_1, \ldots, U_n \in \mathcal{U}$ be a finite cover of $[a, b]$ constructed as before. Then clearly $U_1, \ldots, U_n$ covers K.

Finally, to prove the preceding claim, suppose $s^* = 0$. This implies that there exist $\widetilde{x}_n \in K$ such that $r(\widetilde{x}_n) < 1/n$ for each $n \in \mathbb{N}$. According to Theorem 5.22, the infinite set $\{\widetilde{x}_n\}$ has a limit point $\ell \in K$. Let $r^* = r(\ell) > s > 0$, so $(\ell - s, \ell + s) \subset U$ for some open set $U \in \mathcal{U}$. Since ℓ is a limit point, there exist infinitely many $\widetilde{x}_n \in \left(\ell - \frac{1}{2}s, \ell + \frac{1}{2}s\right)$. For any such $\widetilde{x}_n$, the interval $\left(\widetilde{x}_n - \frac{1}{2}s, \widetilde{x}_n + \frac{1}{2}s\right) \subset (\ell - s, \ell + s) \subset U^*$, which implies $r(\widetilde{x}_n) > \frac{1}{2}s > 0$ for all such n, which is incompatible with our earlier specification that $r(\widetilde{x}_n) < 1/n$ when $n > 2/s$. Thus, we must have $s^* > 0$. *Q.E.D.*

6. Limits.

A striking feature of the present approach to calculus is that it reverses the traditional development the subject, by defining limits using continuity rather than the other way around. Thus, continuity is fundamental, and limits are a consequence thereof. While it would thus be possible to avoid limits entirely in the presentation, the underlying language and intuition remains useful when assimilating and developing intuition for our continuity-based constructions.

Definition 6.1. Let $I \subset \mathbb{R}$ be an open interval, and let $a \in I$. Suppose that $f: I_a \to \mathbb{R}$ is continuous on the *punctured interval* $I_a = I \setminus \{a\}$. We say that $f(x)$ has *limiting value* $z \in \mathbb{R}$ at $x = a$ if the *extended function*

$$\widehat{f}: I \longrightarrow \mathbb{R} \qquad \text{given by} \qquad \widehat{f}(x) = \begin{cases} f(x), & x \neq a, \\ z, & x = a, \end{cases} \tag{6.1}$$

is continuous, in which case we write

$$\lim_{x \to a} f(x) = z. \tag{6.2}$$

The Stitch Theorem 5.6 and the remark following it imply that the limiting value of the function f, if it exists, is unique. In other words, there is at most one possible value z that makes the the extended function (6.1) continuous.

In this case, one says that such a function $f(x)$ has a *removable discontinuity* at $x = a$ because one can remove the apparent discontinuity by appropriately defining its value at a, namely by setting $f(a) = z$. In such cases, we will often, to avoid notational clutter, use the same symbol f for the function and its extension. In particular, if $f: I \to \mathbb{R}$ is continuous, then

$$\lim_{x \to a} f(x) = f(a) \tag{6.3}$$

at all points $a \in I$. Many discontinuities are not removable, examples being the discontinuity at $x = 0$ of the functions $\operatorname{sign} x$, $1/x$, and $\sin(1/x)$.

Example 6.2. Consider the rational function

$$f(x) = \frac{x^2 - 1}{x - 1},$$

which, ostensibly, has a discontinuity at $x = 1$ because its denominator vanishes there. However, the discontinuity is removable, with

$$\lim_{x \to 0} f(x) = 1.$$

Indeed, factorizing the numerator, one immediately finds that, when $x \neq 0$,

$$f(x) = \frac{(x - 1)(x + 1)}{x - 1} = x + 1,$$

and hence the extended continuous function is merely $\widehat{f}(x) = x + 1$ with $\widehat{f}(0) = 1$.

On the other hand, the singularity of the rational function

$$f(x) = \frac{x^2 + 1}{x - 1}$$

at $x = 1$ is not removable. Indeed, we can write

$$f(x) = x + 1 + \frac{2}{x - 1}$$

and the last term is unbounded as x tends to 1.

One can also consider one-sided limits at a point. Let $a < b$. If $f: (a, b) \to \mathbb{R}$ is continuous, and a is finite, the extended function

$$\widehat{f}: [a, b) \longrightarrow \mathbb{R} \qquad \text{given by} \qquad \widehat{f}(x) = \begin{cases} f(x), & a < x < b, \\ z, & x = a, \end{cases}$$

is continuous on the indicated half-open interval, then we say that z is the *right-hand limit* of f at the point a, and write

$$\lim_{x \to a^+} f(x) = z. \tag{6.4}$$

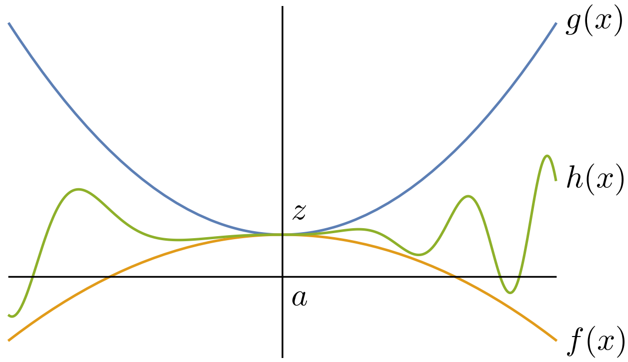


Figure 7. The Squeeze Theorem.

Similarly, if

$$\widetilde{f}\colon (a,b] \longrightarrow \mathbb{R} \qquad \text{given by} \qquad \widetilde{f}(x) = \begin{cases} f(x), & a < x < b, \\ z, & x = b, \end{cases}$$

is continuous, then we say that z is the *left-hand limit* of f at the point b, and write

$$\lim_{x \to b^-} f(x) = z. \tag{6.5}$$

As a consequence of the Stitch Theorem 5.6, we deduce that a function maintains continuity at a point if and only if its right- and left-hand limits are the same there:

Lemma 6.3. *Let $a < b < c$. If $f\colon (a,b] \to \mathbb{R}$ and $g\colon [b,c) \to \mathbb{R}$ are continuous on their respective domains, then the combined function*

$$h(x) = \begin{cases} f(x), & a < x < b, \\ z, & x = b, \\ g(x), & b < x < c, \end{cases}$$

is continuous if and only if $z = f(a) = g(b)$. In other words,

$$\lim_{x \to b} h(x) = z \qquad \textit{if and only if} \qquad z = \lim_{x \to b^-} f(x) = \lim_{x \to b^+} g(x).$$

The next result is known as the Squeeze Theorem, since it says that a function squeezed between two continuous functions at a point must itself be continuous there; see Figure 7. It is very useful for analyzing removable discontinuities and limits.

Theorem 6.4. *Suppose $f\colon I \to \mathbb{R}$ and $g\colon I \to \mathbb{R}$ are continuous on an open interval I, while $h\colon I_a \to \mathbb{R}$ is continuous on the punctured interval $I_a = \mathbb{R} \setminus \{a\}$ for some $a \in I$. Suppose that $f(x) \le h(x) \le g(x)$ for all $a \ne x \in I$, and, further, $f(a) = g(a) = z$. Then a is a removable discontinuity of h, with*

$$\lim_{x \to a} h(x) = z.$$

Proof: We set

$$\widehat{h}(x) = \begin{cases} h(x), & x \ne a, \\ z, & x = a, \end{cases}$$

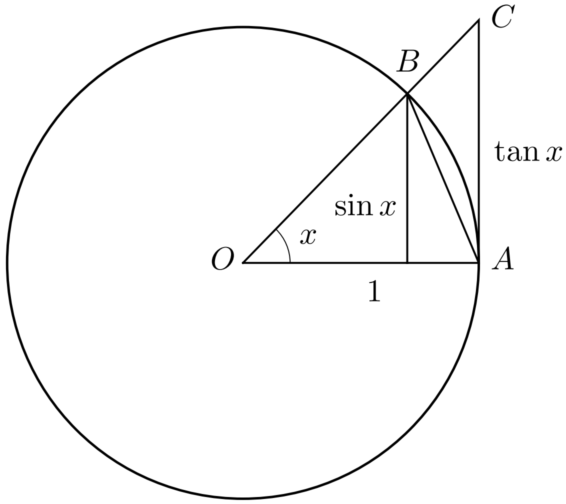


Figure 8. Sine and Tangent Inequalities.

to be the extension of $h(x)$. We need to prove that $\widehat{h}$ is continuous. Given an open subset $S \subset \mathbb{R}$, if $z \notin S$, then $\widehat{h}^{-1}(S) = h^{-1}(S)$ is open by continuity of h. On the other hand, suppose $z \in S$. Let $(c, d) \subset S$ be an open interval containing z, so that $c < z < d$. Consider the open set $U = f^{-1}(c, d) \cap g^{-1}(c, d)$. Note that $a \in U$ since $f(a) = g(a) = z \in (c, d)$. If $x \in U$, then $c < f(x) < d$ and $c < g(x) < d$, and therefore,

$$c < f(x) \leq \widehat{h}(x) \leq g(x) < d$$

for all $x \in U$, including $x = a$. Thus $U \subset \widehat{h}^{-1}(c, d)$, and hence $\widehat{h}^{-1}(S)$ is open, thereby proving continuity of $\widehat{h}$. *Q.E.D.*

One can also make use of a one-sided squeeze, in which one imposes the conditions of Theorem 6.4 on only one side of the point a, leading to the right or left hand limit of the squeezed function h. As before, if they agree, then h has their common value as its one-sided limit.

Example 6.5. Consider the following trigonometric ratio:

$$f(x) = \frac{\sin x}{x},$$

which is defined for $x \neq 0$. We claim that $x = 0$ is a removable discontinuity with limiting value

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{\sin x}{x} = 1. \tag{6.6}$$

In Figure 8, we plot the unit circle, centered at the origin $O = (0, 0)$ and a line at angle $0 < x < \frac{1}{2}\pi$ with respect to the horizontal, which intersects the unit circle at the point $B = (\cos x, \sin x)$. The point $C = (1, \tan x)$ also lies on this line, directly above $A = (1, 0)$.

The area[†] of the triangle OAB, which has base 1 and height $\sin x$ is $\frac{1}{2}\sin x$. This triangle is contained in the circular sector (pizza slice) bounded by the two line segments OA and OB and the circular arc from A to B. Thus, the area of the triangle OAB is strictly less than the area of the circular sector, which is $\frac{1}{2}x$, since it forms a fraction of $x/(2\pi)$ of the area of the unit disk, which is π. Finally, the circular sector is contained in the triangle OAC and hence its area is strictly less than the area of this right triangle, which is $\frac{1}{2}\tan x$. Combining the preceding deductions, and cancelling the common factor of $\frac{1}{2}$, we have established the following inequalities:

$$\sin x < x < \tan x = \frac{\sin x}{\cos x} \qquad \text{for} \qquad 0 < x < \tfrac{1}{2}\pi. \tag{6.7}$$

Rearranging these two inequalities produces

$$\cos x < \frac{\sin x}{x} < 1. \tag{6.8}$$

Initially, owing to (6.7), the inequalities (6.8) are only valid for $0 < x < \frac{1}{2}\pi$, but the first two functions are even and so replacing x by $-x$ does not change either inequality. Thus, (6.8) is valid for $-\frac{1}{2}\pi < x < \frac{1}{2}\pi$. Moreover, both $\cos x$ and the constant function 1 are continuous and have value 1 at $x = 0$. Thus, the Squeeze Theorem 6.4 implies (6.6).

Example 6.6. Consider the trigonometric function

$$f(x) = x \,\sin \frac{1}{x}\,,$$

which is defined and continuous for $x \neq 0$. Since $|\sin(1/x)| \leq 1$ for $x \neq 0$, we have

$$-|\,x\,| \leq f(x) \leq |\,x\,| \qquad \text{for all} \qquad x \neq 0.$$

The functions $\pm|\,x\,|$ are both continuous, as noted above, and have limiting value 0 at $x = 0$. Thus, applying the Squeeze Theorem 6.4, we conclude that

$$\lim_{x \to 0} f(x) = 0. \tag{6.9}$$

Since the composition of two continuous functions is continuous, we immediately deduce the following useful statement:

$$\text{If } g \text{ is continuous, and } \lim_{x \to a} f(x) = z, \text{ then } \lim_{x \to a} g\big[\,f(x)\,\big] = g\left[\,\lim_{x \to a} f(x)\,\right] = g(z). \tag{6.10}$$

Thus, for example, setting $g(x) = 1/x$, which is continuous for $x \neq 0$, and using (6.6), we deduce that

$$\lim_{x \to 0} \frac{x}{\sin x} = \frac{1}{\displaystyle\lim_{x \to 0} \frac{\sin x}{x}} = 1.$$

This simple result allows us to build up a large repertoire of known limits.

† See Section 13 for a full discussion of the concept of area.

The above definition of limit adapts immediately to the case when the function is defined on a general subset $A \subset \mathbb{R}$; just replace I by A in Definition 6.1. One can also use Definition 5.18 to define the limit of a function at a single point, independent of its continuity or lack thereof elsewhere. Namely, we set

$$\lim_{x \to a} f(x) = z \quad \text{provided} \quad \widehat{f}(x) = \begin{cases} f(x), & x \neq a, \\ z, & x = a, \end{cases} \quad \text{is continuous at } a. \tag{6.11}$$

Thus, for example, the function $f(x)$ which takes the value x at all rational numbers and x at all irrationals — see (5.8) — has $\lim_{x \to 0} f(x) = f(0)$. Using the remarks following Example 5.20, one can easily recast condition (6.11) in the traditional delta-epsilon definition of limits, whose precise statement is left to the reader.

Example 6.7. Finally, let us look at the discrete setting of Example 5.9:

$$A = \{0\} \cup \widetilde{A}, \quad \text{where} \quad \widetilde{A} = \{\, 1/n \mid n \in \mathbb{N} \,\}.$$

We can identify a function $f \colon \widetilde{A} \to \mathbb{R}$ with a *sequence*, setting $f_n = f(1/n) \in \mathbb{R}$ for $n \in \mathbb{N}$, so that $f_1, f_2, f_3, \ldots$ forms an infinite sequence† of real numbers. First of all, since any subset of $\widetilde{A}$ is open, any function $f \colon \widetilde{A} \to \mathbb{R}$ is continuous, and so, unlike the case of functions on $\mathbb{R}$ or subintervals thereof, its limit at any point $1/n \in \widetilde{A}$ is *not* uniquely defined. In other words, given $n \in \mathbb{N}$ and a function $\widehat{f} \colon \widetilde{A}_{1/n} \to \mathbb{R}$ on the punctured domain $\widetilde{A}_{1/n} = \widetilde{A} \setminus \{1/n\}$, *any* value $z = f(1/n) \in \mathbb{R}$ will make the extended function $f \colon \widetilde{A} \to \mathbb{R}$ continuous, and hence satisfy the condition $\displaystyle\lim_{x \to 1/n} f(x) = z$ on $\widetilde{A}$. This is strange, and will not be employed in the sequel.

More interesting is what happens at the limit point $0 \in A$. Here we say the limit of the sequence $\{\, f_n \mid n \in \mathbb{N} \,\}$ defined by $f \colon A \setminus \{0\} = \widetilde{A} \to \mathbb{R}$ is

$$\lim_{n \to \infty} f_n = \lim_{1/n \to 0} f(1/n) = z \tag{6.12}$$

if the extended function

$$\widehat{f} \colon A \longrightarrow \mathbb{R} \qquad \text{given by} \qquad \widehat{f}(x) = \begin{cases} f(x), & x \in \widetilde{A}, \\ z, & x = 0, \end{cases} \tag{6.13}$$

is continuous on A. See Example 5.9 for details on how to assess its continuity. In this case, the limiting value, if it exists, is unique. Indeed, suppose, to the contrary, that $z \neq w$ and both

$$\widehat{f}(x) = \begin{cases} f(x), & x \in \widetilde{A}, \\ z, & x = 0, \end{cases} \qquad \widetilde{f}(x) = \begin{cases} f(x), & x \in \widetilde{A}, \\ w, & x = 0, \end{cases}$$

are continuous. Choose disjoint open intervals, $I \cap J = \varnothing$, with $z \in I$ and $w \in J$. Then, by continuity, both $\widehat{f}^{-1}(I) = \{0\} \cup f^{-1}(I)$ and $\widetilde{f}^{-1}(J) = \{0\} \cup f^{-1}(J)$ are open in A,

† In general, the term *sequence* refers to an infinite collection of objects labelled by the natural numbers.

and hence so is their intersection. But since I, J are disjoint, $f^{-1}(I) \cap f^{-1}(J) = \varnothing$, and hence $\widehat{f}^{-1}(I) \cap \widetilde{f}^{-1}(J) = \{0\}$, which is not an open subset of A, in contradiction to our original assumption. *Q.E.D.*

Thus, the continuity characterization of limits includes the limiting behavior of a sequence of real numbers $f_1, f_2, f_3 \ \ldots \ \in \mathbb{R}$, for which we identify $f_n = f(1/n)$ with the values of a function on the points $1/n \in \widetilde{A}$. Alternatively, we can identify the sequence with a suitably continuous extension of the function $f\colon \widetilde{A} \to \mathbb{R}$, defining a continuous function $h\colon \mathbb{R}^+ \to \mathbb{R}$ such that $h(1/n) = f(1/n)$ for $n \in \mathbb{N}$. The sequence has a limit $z \in \mathbb{R}$, as in (6.12), if and only if one can extend h to a continuous function $\widehat{h}\colon \mathbb{R} \to \mathbb{R}$ such that $\widehat{h}(x) = h(x)$ for all $x > 0$ while $\widehat{h}(x) = z$ for all $x \leq 0$, cf. (5.6).

Since composition of functions restricted to subsets remains continuous, we can adapt (6.10) to establish a similarly useful result for discrete limits:

$$\text{If } g \text{ is continuous, and} \quad \lim_{n\to\infty} f_n = z, \ \text{ then } \ \lim_{n\to\infty} g(f_n) = g\left(\lim_{n\to\infty} f_n\right) = g(z). \tag{6.14}$$

It is also worth noting that this construction extends to other countably infinite subsets with one or more limit points. If a continuous function defined on such a subset can be extended to a continuous function on the closure — the union of the set and all its limit points — then the extension serves to uniquely define the limiting values of the original function at all the limit points. Details are left to the reader. We also leave it for the reader to speculate and/or investigate what the continuity and limiting behavior of functions $f\colon \mathbb{Q} \to \mathbb{R}$ on the rational numbers might look like.

Example 6.8. Consider the function

$$f(x) = \sin \frac{\pi}{x}, \qquad x \neq 0,$$

which is continuous on the punctured line $\mathbb{R} \setminus \{0\}$. The corresponding sequence is $f_n = f(1/n) = \sin \pi n = 0$ for all $n \in \mathbb{N}$, which is identically zero and hence has limit 0 at $x = 0$. However, $x = 0$ is not a removable discontinuity, and there is no continuous extension of $f(x)$. Indeed, suppose $f(0) = z > -1$. Let $c > \max\{1, z\}$. Then $A = f^{-1}(-1, c) = \mathbb{R} \setminus \left\{ 1/\left(2n + \frac{3}{2}\right) \;\middle|\; n \in \mathbb{Z} \right\}$, which is not open because $0 \in A$ but there is no open interval $I \subset A$ that contains 0. A similar construction applies when $z < 1$.

7. Derivatives.

We are now in a position to define the derivative directly using continuity instead of basing it on any limiting process. We begin with the simplest case of the derivative of a function at the origin.

Definition 7.1. Let $f:\mathbb{R}\to\mathbb{R}$ be continuous and satisfy $f(0)=0$. We say that $f(x)$ has a *derivative* at $x=0$ if there exists $z\in\mathbb{R}$ such that the function

$$q(x)=\begin{cases} f(x)/x, & x\neq 0,\\ z, & x=0,\end{cases} \tag{7.1}$$

is continuous. In this case we write $z=f'(0)$.

According to Theorem 4.3, the quotient $q(x)=f(x)/x$ is continuous except at $x=0$. The existence of a derivative implies that it has a removable discontinuity at $x=0$, and hence by extending it by setting $q(0)=f'(0)$ produces a continuous function on all of $\mathbb{R}$.

Example 7.2. The function $f(x)=3x$ satisfies $f(0)=0$. To find its derivative, we note that, for $x\neq 0$,

$$q(x)=\frac{f(x)}{x}=\frac{3x}{x}=3.$$

Thus, $x=0$ is a removable discontinuity, and we immediately conclude that $f'(0)=3$.

Next consider the function $f(x)=x^2$. In this case

$$q(x)=\frac{f(x)}{x}=\frac{x^2}{x}=x,$$

again initially for $x\neq 0$, but the limiting value that makes $q(x)$ continuous at $x=0$ is $q(0)=0=f'(0)$. The same clearly holds for any monomial function: $f(x)=x^n$ for $2\leq n\in\mathbb{N}$.

We can similarly define derivatives at other points for functions with nonzero values. Suppose $f(a)=c$. In Definition 7.1, we replace $f(x)$ by $f(x)-f(a)=f(x)-c$, which vanishes when $x=a$; similarly we replace x by $x-a$ which has the same property. In this manner, we arrive at the general definition of the derivative.

Definition 7.3. Let $f:\mathbb{R}\to\mathbb{R}$ be continuous. We say that $f(x)$ has a *derivative* at $a\in\mathbb{R}$ if there exists $z\in\mathbb{R}$ such that the *difference quotient* function

$$q(x)=\begin{cases} \dfrac{f(x)-f(a)}{x-a}, & x\neq a,\\ z, & x=a.\end{cases} \tag{7.2}$$

is continuous. In this case, we write $f'(a)=z=q(a)$, or, equivalently

$$f'(a)=\lim_{x\to a}\frac{f(x)-f(a)}{x-a}. \tag{7.3}$$

Remark: The preceding definition of derivative via continuity of the difference quotient function $q(x)$ dates back to the early twentieth century Greek mathematician Constantin Carathéodory, who states an equivalent version in his textbook on complex analysis, [**4**], namely, that $f:\mathbb{R}\to\mathbb{R}$ has a derivative at $a\in\mathbb{R}$ if there exists a continuous function $q:\mathbb{R}\to\mathbb{R}$ such that

$$f(x)=f(a)+q(x)\,(x-a). \tag{7.4}$$

Again, $f'(a) = q(a)$. Clearly, if (7.4) holds, $f(x)$, being a combination of continuous functions and constants, if itself necessarily continuous. Interestingly, Carathéodory does not use this definition in his real analysis text, [**5**]; see also [**13**]. Also, he initially defines continuity through limits, and not in the topological terms we use here.

If $f(x)$ has a derivative at all points in its domain, we say f is *differentiable* and write $f'(x)$ for the value of its derivative at the point x.

Example 7.4. The preceding function $f(x) = 3x$ has constant derivative: $f'(x) = 3$ at all $x \in \mathbb{R}$. Indeed, the difference quotient is

$$\frac{f(x) - f(a)}{x - a} = \frac{3x - 3a}{x - a} = 3 \qquad \text{when} \qquad x \neq a,$$

and hence has limiting value $f'(a) = 3$.

Example 7.5. The derivative of any constant function $f(x) = c$ is zero everywhere. Indeed,

$$\frac{f(x) - f(a)}{x - a} = \frac{c - c}{x - a} = 0 \qquad \text{when} \qquad x \neq a.$$

Later we will see that these are essentially the only functions with identically zero derivative; see Proposition 7.14. Here "essentially" means that the function should be defined on an open interval. A piecewise constant function that assumes different constant values on a disjoint union of subintervals also has zero derivative, but is not constant on its entire disconnected domain, although its restriction to any of the subintervals is constant.

Example 7.6. Let $f(x) = \sin x$. The difference quotient at $x = 0$ is $q_0(x) = \sin x / x$, whose limit $q_0(0) = 1 = f'(0)$ we computed in Example 6.5. Thus the derivative of $\sin x$ at $x = 0$ is 1. At a general point $x = a$, the difference quotient can be rewritten using a standard trigonometric identity:

$$\begin{aligned} q_a(x) = \frac{\sin x - \sin a}{x - a} &= \frac{2\,\cos\frac{1}{2}(x+a)\ \sin\frac{1}{2}(x-a)}{x - a} \\ &= \frac{\sin\frac{1}{2}(x-a)}{\frac{1}{2}(x-a)}\ \cos\tfrac{1}{2}(x+a) = q_0\left[\,\tfrac{1}{2}(x-a)\,\right]\cos\tfrac{1}{2}(x+a). \end{aligned}$$

Thus, at the point $x = a$,

$$f'(a) = q_a(a) = q_0(0)\cos a = \cos a.$$

Replacing a by x we conclude that the derivative of $\sin x$ is $\cos x$. Similarly, to compute the derivative of $\cos x$, we write the corresponding difference quotient

$$\frac{\cos x - \cos a}{x - a} = \frac{-2\ \sin\frac{1}{2}(x+a)\ \sin\frac{1}{2}(x-a)}{x - a} = -\,q_0\left[\,\tfrac{1}{2}(x-a)\,\right]\sin\tfrac{1}{2}(x+a),$$

which has the limiting value $-\sin a$ at $x = a$, and hence the derivative of $\cos x$ is $-\sin x$.

Example 7.7. As note above, the absolute value function $f(x) = |\,x\,|$ is continuous When restricted to $\mathbb{R}^+$ it has the derivative $f'(a) = +1$, for all $a > 0$, because the quotient

$$\frac{f(x) - f(a)}{x - a} = \frac{x - a}{x - a} = 1, \qquad \text{provided} \qquad a > 0, \quad x > 0, \quad x \neq a,$$

On the other hand,

$$\frac{f(x) - f(a)}{x - a} = \frac{-x + a}{x - a} = -1, \qquad a \neq x < 0,$$

and hence $f'(a) = -1$ when $a < 0$. Thus, on $\mathbb{R} \setminus \{0\}$, the derivative of the absolute value function is the sign function, $f'(x) = \operatorname{sign} x$ when $x \neq 0$. However, the absolute value function does not have a derivative at $a = 0$ because the difference quotient

$$q(x) = \frac{|\,x\,|}{x} = \operatorname{sign} x$$

does not have a removable discontinuity there; see the discussion following (4.1). Thus, the derivative of a continuous function, even when it exists, need not be continuous.

Using Definition 5.18 of the continuity of a function at a single point, one can define the derivative of a function at a point without relying on continuity of the difference quotient, or the function itself away from the point. Namely, we say that $f(x)$ has a *derivative at the point* $a \in \mathbb{R}$ if the associated difference quotient (7.2) is continuous at a, in which case $z = f'(a) = q(a)$, or, equivalently, (7.3) holds. Since one can write the function in terms of its difference quotient using (7.4), existence of the derivative at a implies continuity of f at a. An example of such a discontinuous function is

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q}, \\ x^2, & x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases} \tag{7.5}$$

The difference quotient at $a = 0$ is $q(x) = f(x)/x$ for $x \neq 0$, which coincides with the function (5.8) taking the values 0 or x according to whether x is rational or irrational, and hence is made continuous by setting $q(0) = f'(0) = 0$. On the other hand, (7.5) is not continuous at any $a \neq 0$ and so does not have a derivative at any other point.

After one learns the basic rules for differentiation, the derivative of any combination of elementary functions can be calculated directly in terms of the derivatives of their constituents without the need to revert to the difference quotient. Included in each of the following rules is the statement that if the functions f and g are differentiable at a point a then the prescribed combination is also differentiable at a and has the indicated formula. Continuity of the required combinations follows from Theorem 4.3 (which remains to be proved). Under these assumptions, as in (7.2), we set

$$F(x) = \begin{cases} \dfrac{f(x) - f(a)}{x - a}, & x \neq a, \\ f'(a), & x = a, \end{cases} \qquad G(x) = \begin{cases} \dfrac{g(x) - g(a)}{x - a}, & x \neq a, \\ g'(a), & x = a. \end{cases} \tag{7.6}$$

Sum Rule: The first rule is that the derivative of the sum of two functions equals the sum of their derivatives. Given

$$s(x) = f(x) + g(x), \qquad \text{then} \qquad s'(a) = f'(a) + g'(a). \tag{7.7}$$

Indeed, since

$$\frac{s(x) - s(a)}{x - a} = \frac{f(x) + g(x) - f(a) - g(a)}{x - a} = \frac{f(x) - f(a)}{x - a} + \frac{g(x) - g(a)}{x - a} = F(x) + G(x).$$

The right hand side is continuous, being the sum of two continuous functions, with limiting value $F(a) + G(a) = f'(a) + g'(a)$ at $x = a$. Thus, the left hand side is also continuous with limiting value $s'(a)$, which implies (7.7) holds. The same holds for their difference, with an identical proof. Setting

$$d(x) = f(x) - g(x), \qquad \text{implies} \qquad d'(a) = f'(a) - g'(a), \tag{7.8}$$

Product Rule: The derivative of the product of two functions is *not* the product of their derivatives! The correct formula says that the derivative of

$$p(x) = f(x)\, g(x), \qquad \text{is} \qquad p'(a) = f'(a)\, g(a) + f(a)\, g'(a). \tag{7.9}$$

The proof follows from the algebraic formula

$$\begin{aligned} \frac{p(x) - p(a)}{x - a} &= \frac{f(x)\, g(x) - f(a)\, g(a)}{x - a} \\ &= \frac{f(x) - f(a)}{x - a}\, g(x) + f(a)\, \frac{g(x) - g(a)}{x - a} = F(x)\, g(x) + f(a)\, G(x). \end{aligned}$$

The right hand side is continuous with limiting value

$$F(a)\, g(a) + f(a)\, G(a) = f'(a)\, g(a) + f(a)\, g'(a)$$

at $x = a$, thereby establishing (7.9).

Reciprocal Rule: The next formula is for the reciprocal of a nonzero continuous function. Setting

$$r(x) = \frac{1}{g(x)} \qquad \text{with} \qquad g(x) \neq 0, \qquad \text{then} \qquad r'(a) = -\,\frac{g'(a)}{g(a)^2}. \tag{7.10}$$

To prove this, use the identity

$$\frac{r(x) - r(a)}{x - a} = \frac{1/g(x) - 1/g(a)}{x - a} = -\,\frac{1}{g(x)\, g(a)}\, \frac{g(x) - g(a)}{x - a} = -\,\frac{G(x)}{g(x)\, g(a)}.$$

The limiting value of the final expression at $x = a$ is $-\,g'(a)/g(a)^2$, thus producing the final formula (7.10).

Quotient Rule: Let

$$q(x) = \frac{f(x)}{g(x)} \qquad \text{with} \qquad g(x) \neq 0. \qquad \text{Then} \qquad q'(a) = \frac{f'(a)\, g(a) - f(a)\, g'(a)}{g(a)^2}.$$

This is proved by combining the product rule (7.9) and the reciprocal rule (7.10). Indeed, $q(x) = f(x)\,r(x)$, and hence

$$q'(a) = f'(a)\,r(a) + f(a)\,r'(a) = \frac{f'(a)}{g(a)} - f(a)\,\frac{g'(a)}{g(a)^2} = \frac{f'(a)\,g(a) - f(a)\,g'(a)}{g(a)^2}.$$

Composition or Chain Rule: Finally consider the composition of continuous functions:

$$c(x) = f \circ g(x) = f(g(x)) \qquad \text{Then} \qquad c'(a) = f'(g(a))\,g'(a). \tag{7.11}$$

To prove this we use the formula

$$\frac{c(x) - c(a)}{x - a} = \frac{f(g(x)) - f(g(a))}{x - a} = \frac{f(g(x)) - f(g(a))}{g(x) - g(a)}\;\frac{g(x) - g(a)}{x - a}\,,$$

which is valid provided $g(x) \neq g(a)$. Since the function $F(x)$ given in (7.6) is continuous, its composition with $g(x)$, namely

$$H(x) = F \circ g(x) = \begin{cases} \dfrac{f(g(x)) - f(g(a))}{g(x) - g(a)}, & g(x) \neq g(a), \\ f'(g(a)), & g(x) = g(a), \end{cases}$$

is also continuous. But this implies continuity of

$$\frac{c(x) - c(a)}{x - a} = H(x)\,G(x)\,, \quad \text{with limit} \quad c'(a) = H(a)\,G(a) = f'(g(a))\,g'(a),$$

as claimed.

As a simple application, we find that, for any $a \in \mathbb{R}$,

$$\begin{aligned} g(x) &= f(x + a) \quad \text{has derivative} \quad g'(x) = f'(x + a), \\ h(x) &= f(a - x) \quad \text{has derivative} \quad h'(x) = -\,f'(a - x). \end{aligned} \tag{7.12}$$

To see this, note that $g = f \circ l$ where $l(x) = x + a$ has derivative $l'(x) = 1$ everywhere. Thus, the formula for $g'(x) = f'(l(x))\,l'(x) = f'(x+a)$ follows immediately from the Chain Rule. A similar calculation establishes the second formula in (7.12).

Example 7.8. Let us prove that the derivative of

$$f(x) = x^n \quad \text{is} \quad f'(x) = n\,x^{n-1} \quad \text{for} \quad n \in \mathbb{N}. \tag{7.13}$$

We will prove this by induction, with the cases $n = 0, 1$ already known. Given that (7.13) holds for a given n, consider the function

$$g(x) = x^{n+1} = x\,x^n = x\,f(x).$$

Applying the Product Rule (7.9) and the fact that the derivative of x is 1,

$$g'(x) = f(x) + x\,f'(x) = x^n + x\,(n\,x^{n-1}) = (n+1)\,x^n,$$

thus completing the induction step. More generally, let

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \;\cdots\; + a_2 x^2 + a_1 x + a_0$$

where $a_0, \ldots, a_n$ are constants, be a polynomial of degree n. Then we can repeatedly employ the Sum and Product Rules to compute its derivative:

$$p'(x) = n\, a_n x^{n-1} + (n-1)\, a_{n-1} x^{n-1} + \;\cdots\; + 2\, a_2 x + a_1,$$

which is a polynomial of degree $n-1$.

Theorem 7.9. *Suppose $f: \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotone function, and let $g = f^{-1}: \mathbb{R} \to \mathbb{R}$ be its continuous inverse. If $f'(x) \neq 0$, then g is differentiable at $y = f(x)$ and*

$$g'(y) = \frac{1}{f'(g(y))}. \tag{7.14}$$

Proof: By definition of inverse,

$$g \circ f(x) = x.$$

Taking the derivative of both sides, and using the Chain Rule (7.11), yields

$$g'\,[\, f(x)\,]\; f'(x) = 1.$$

Under the assumption $f'(x) \neq 0$ we can divide both sides of the latter equation by it. Replacing $x = g(y)$, so $y = f(x)$, produces (7.14). *Q.E.D.*

For example, consider the function $f(x) = x^3$. Its inverse is the function $g(y) = y^{1/3}$. Since $f'(x) = 3\,x^2$, we find

$$g(y) = y^{1/3}, \qquad g'(y) = \frac{1}{3\,x^2} = \tfrac{1}{3}\, y^{-2/3},$$

provided $y \neq 0$. In fact, $g(y) = y^{1/3}$ is not differentiable at $y = 0$ since its difference quotient $q(y) = g(y)/y = y^{-2/3}$ is unbounded there. Note that this formula is in accordance with our earlier rule (7.13), where in this case we take $n = \frac{1}{3}$. Indeed, it can be shown that the rule holds for all $n \in \mathbb{R}$, where we exclude $x = 0$ when $n < 1$. Of course, when $n = 0$, we have $x^0 = 1$, with derivative 0 for all x. More generally, by applying the Chain Rule (7.11), we deduce that if $f(x)$ is differentiable, then its $n^{\rm th}$ power

$$F(x) = \bigl[\, f(x)\,\bigr]^n$$

is differentiable provided that $f(x) \neq 0$ when $0 \neq n < 1$, with

$$F'(x) = n\,\bigl[\, f(x)\,\bigr]^{n-1}\, f'(x). \tag{7.15}$$

Given $b \neq a$, the value of the difference quotient $q(b) = \dfrac{f(b) - f(a)}{b - a}$ equals the slope of the *secant line* passing through the two points $(a, f(a))$ and $(b, f(b))$ on the graph of f, which can be identified as the graph of the linear function

$$s_b(x) = \frac{f(b) - f(a)}{b - a}\,(x - a) + f(a). \tag{7.16}$$

As $b \to a$, the difference quotient has the limiting value $q(a) = f'(a)$, and so the secant line functions (7.16) converge, in a continuous manner, to the linear function

$$t(x) = f'(a)(x - a) + f(a), \tag{7.17}$$

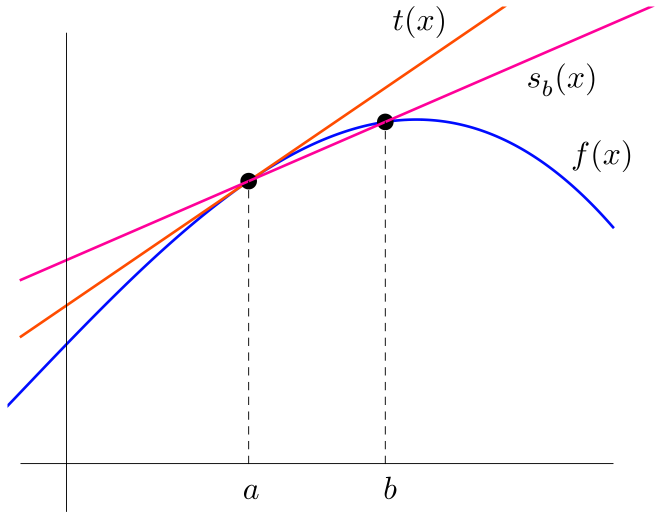


Figure 9. Secant and Tangent Lines.

whose graph can be identified with the *tangent line* to the graph of f at the point $(a, f(a))$. This is illustrated in Figure 9. In other words, as b gets closer and closer to a, the secant lines approach the tangent line, whose slope equals the derivative $f'(a)$.

If $f\colon [a,b] \to \mathbb{R}$ is defined on a closed interval, then one can introduce notions of right- and left-hand derivatives at the endpoints. Specifically, we define the *right-hand derivative* $f'(a^+)$ at the right hand endpoint a to be the value (if such exists) that makes the extended function

$$F_a(x) = \begin{cases} \dfrac{f(x) - f(a)}{x - a}, & a < x \le b, \\ f'(a^+), & x = a, \end{cases}$$

continuous on $[a,b]$. Similarly, we define the *left-hand derivative* $f'(b^-)$ at the left hand endpoint b as the value that makes

$$F_b(x) = \begin{cases} \dfrac{f(x) - f(b)}{x - b}, & a \le x < b, \\ f'(b^-), & x = b, \end{cases}$$

continuous on $[a,b]$. This extends in the obvious manner to half open intervals. Applying Lemma 6.3, if f is defined on an open interval containing a, then $f'(a)$ exists if and only if $f'(a^+) = f'(a^-)$, in which case all three are equal.

The next two results state that the derivative of a function at a (local) minimum or maximum must be zero, provided it exists. We start with a simple version.

Lemma 7.10. *Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is continuous, $f(0) = 0$, and $f(x) \ge 0$ for all $x \ne 0$. If f is differentiable at 0 then $f'(0) = 0$.*

Proof: As in the definition, set $q(x) = f(x)/x$. If $f(x) > 0$ for all $x \ne 0$, then $q(x) > 0$ for $x > 0$ while $q(x) < 0$ for $x < 0$. By the Intermediate Value Theorem 4.6, there is a

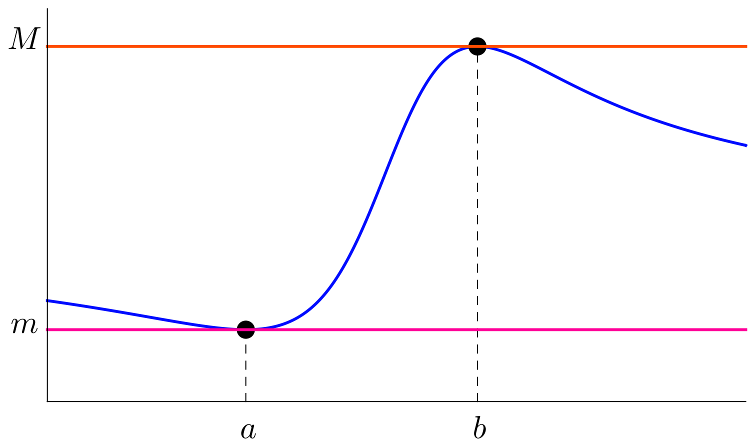


Figure 10. Tangent Lines at Minimum and Maximum.

point c where $q(c) = 0$, but the only possibility is $c = 0$ since $q(x) \neq 0$ when $x \neq 0$.

If we only know $f(x) \geq 0$, then we must work a little harder to establish the result, because $q(x)$ will vanish whenever $f(x) = 0$ and so we cannot immediately conclude that $c = 0$. However, we do know $q(x) \geq 0$ for all $x > 0$. According to Lemma 4.5, adapted to one-sided limits, this implies $\lim_{x \to 0^+} q(x) = f'(0^+) \geq 0$. On the other hand, $q(x) \leq 0$ for all $x < 0$, which implies $\lim_{x \to 0^-} q(x) = f'(0^-) \leq 0$. Since we are assuming $q(x)$ is continuous, its right and left limits must coincide, and this implies

$$\lim_{x \to 0} q(x) = f'(0) = f'(0^-) = f'(0^+) = 0. \qquad Q.E.D.$$

Warning: This argument does not apply to right- and left-hand derivatives of a function at a minimum or maximum when they do not agree. For example, $f(x) = |\,x\,|$ has a minimum at $x = 0$, but $f'(0^-) = -1,\ \ f'(0^+) = +1$.

Note that the assumption of Lemma 7.10 implies that $f(0) = 0$ is a (global) minimum of the function $f(x)$ since everywhere else it is non-negative. Adapting the preceding proof to general local maxima and minima, we are able to establish the following result.

Theorem 7.11. *Suppose that $m = f(a)$ is a local minimum for $f(x)$, meaning that $f(x) \geq m$ for all $x \in I$ where I is an open interval containing the point a. Then, provided the derivative exists, $f'(a) = 0$. Similarly, if $M = f(b)$ is a local maximum, so $f(x) \leq M$ for all $x \in J$, where $b \in J$ is an open interval, then, when it exists, $f'(b) = 0$.*

One can interpret this result as saying that the tangent line at a local maximum or a local minimum must be horizontal; that is, it is the graph of a constant function; see Figure 10.

Warning: If $f\colon [\,a, b\,] \to \mathbb{R}$ is defined on a closed interval, and its maximum and/or minimum occurs at one of the endpoints, say a, then it is *not* true that $f'(a) = 0$. For

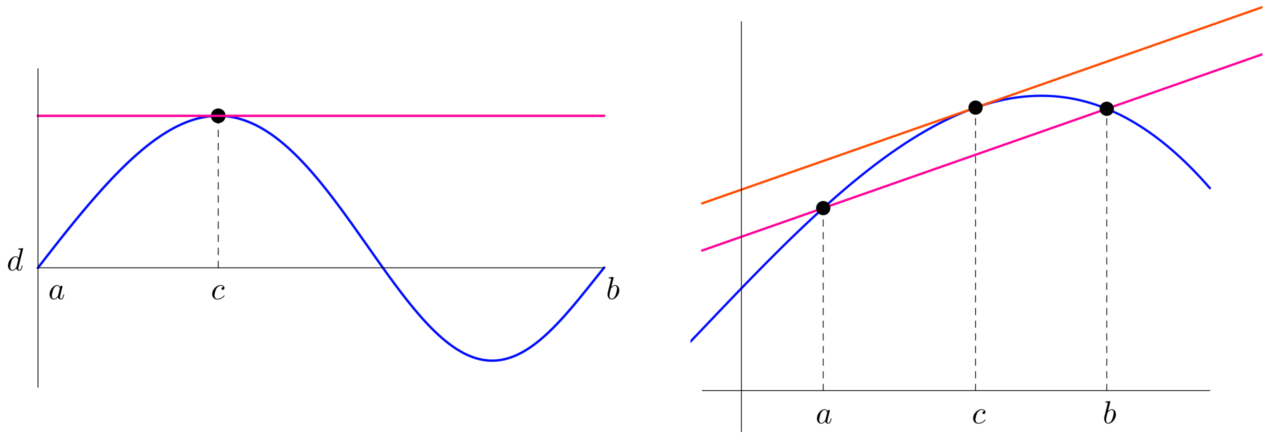


Figure 11. Rolle's Theorem and the Mean Value Theorem.

example, the function $f(x) = 3x$ on the closed interval $[1,2]$ has minimum $3 = f(1)$ and maximum $6 = f(2)$, but $f'(x) \equiv 3$ everywhere, including the endpoints.

The next result is sometimes known as Rolle's Theorem, in honor of the seventeenth century French mathematician Michel Rolle, and is illustrated in Figure 11.

Theorem 7.12. *Let $a < b$. Suppose that $f\colon [a,b] \to \mathbb{R}$ is differentiable and satisfies $f(a) = f(b) = d$. Then there exists $a < c < b$ such that $f'(c) = 0$.*

Proof: Let $M = \sup f[a,b] \geq d$ and $m = \inf f[a,b] \leq d$. According to the Extreme Value Theorem 5.17, there exists at least one point $x_M \in [a,b]$ such that $f(x_M) = M$ and at least one point $x_m \in [a,b]$ such that $f(x_m) = m$. If $M > d$, then $a < x_M < b$. Theorem 7.11 implies $f'(x_M) = 0$ and we can set $c = x_M$. Similarly, if $m < d$, then $a < x_m < b$ and $f'(x_m) = 0$, so that $c = x_m$ works. Finally, the only way we can have both $M = m = d$ is if $f(x) \equiv d$ is constant, but then $f'(x) = 0$ for all $a < x < b$. *Q.E.D.*

Rolle's Theorem is a special case of a very useful technical result known as the Mean Value Theorem. It states that if f is differentiable on $[a,b]$, then there is at least one point on its graph such that the associated tangent line has the same slope as the secant line passing through the points $(a, f(a))$ and $(b, f(b))$; see Figure 11.

Theorem 7.13. *Let $a < b$ and suppose $f\colon [a,b] \to \mathbb{R}$ is differentiable. Then there exists $a < c < b$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}\,. \tag{7.18}$$

Proof: Define

$$g(x) = f(x) - s(x), \qquad \text{where} \qquad s(x) = \frac{f(b) - f(a)}{b - a}\,(x - a) + f(a)$$

is, according to (7.16), the linear function whose graph is the secant line passing through the points $(a, f(a))$ and $(b, f(b))$. Thus, $g(x)$ is continuous on $[a,b]$, and satisfies $g(a) =$

$g(b) = 0$, which are the conditions of Rolle's Theorem 7.12. This implies there is a point $a < c < b$ such that

$$0 = g'(c) = f'(c) - s'(c) = f'(c) - \frac{f(b) - f(a)}{b - a},$$

which establishes (7.18). *Q.E.D.*

Proposition 7.14. *Let $f: \mathbb{R} \to \mathbb{R}$ be differentiable. Then $f'(x) \equiv 0$ for all x if and only if $f(x) \equiv c$ is a constant function.*

Proof: We already know that the derivative of a constant function is everywhere 0. To prove the converse, let $a < b$. Then according to (7.18)

$$\frac{f(b) - f(a)}{b - a} = f'(c) = 0$$

for some point $a < c < b$. This implies $f(b) = f(a)$. Since this holds for all a, b, it implies that f has the same constant value everywhere. *Q.E.D.*

Remark: The result holds when f is defined on an interval — open or closed. It does not hold on more general subsets. For example, the function $f(x) = \operatorname{sign} x$ has $f'(x) = 0$ for all $x \neq 0$, but is not constant. It is, however, constant on each subinterval $(-\infty, 0)$ and $(0, \infty)$ of its domain $\mathbb{R} \setminus \{0\}$.

Corollary 7.15. *Two differentiable functions $f(x)$ and $g(x)$ have the same derivative, $f'(x) = g'(x)$ if and only if they differ by a constant $c \in \mathbb{R}$, so that $f(x) = g(x) + c$.*

To prove this, merely apply Proposition 7.14 to the difference $f(x) - g(x)$, using (7.8).

Let us next show that continuously differentiable functions are Lipschitz continuous, as characterized in (4.2).

Theorem 7.16. *If $f: [a, b] \to \mathbb{R}$ and its derivative*[†] *$f': [a, b] \to \mathbb{R}$ are continuous, then f satisfies the Lipschitz continuity condition* (4.2) *for $x, y \in [a, b]$.*

Proof: Continuity of $f'(x)$ and the Boundedness Theorem 5.16 imply that $|f'(x)| \leq \lambda$ has an upper bound $\lambda \geq 0$ for all $x \in [a, b]$. By the Mean Value Theorem 7.13, if $a \leq x < y \leq b$, then

$$f(y) - f(x) = f'(z)\,(y - x)$$

for some $x < z < y$. Thus,

$$|f(y) - f(x)| = |f'(z)|\,|y - x| \leq \lambda\,|y - x|,$$

which is exactly (4.2). *Q.E.D.*

[†] At the endpoints a, b of the interval, we employ the right- and left-hand derivatives, respectively.

There is a useful extension of the Mean Value Theorem due to Cauchy, that will be needed in the ensuing result.

Theorem 7.17. *Let $a < b$ and suppose $f, g : [a, b] \to \mathbb{R}$ are differentiable. Then there exists $a < c < b$ such that*

$$\big[\, g(b) - g(a) \,\big]\, f'(c) = \big[\, f(b) - f(a) \,\big]\, g'(c). \tag{7.19}$$

Note that when $g(x) = x$, so $g'(x) \equiv 1$, the Cauchy Mean Value Formula (7.19) reduces to $(b - a) f'(c) = f(b) - f(a)$, which is equivalent to the ordinary Mean Value Formula (7.18). Interestingly its similar proof is, if anything, slightly simpler.

Proof: Consider the function

$$h(x) = \big[\, g(b) - g(a) \,\big]\, f(x) - \big[\, f(b) - f(a) \,\big]\, g(x).$$

Note that

$$h(a) = g(b)\, f(a) - f(b)\, g(a) = h(b),$$

and hence, by Rolle's Theorem 7.12, there exists $a < c < b$ such that

$$0 = h'(c) = \big[\, g(b) - g(a) \,\big]\, f'(c) - \big[\, f(b) - f(a) \,\big]\, g'(c). \qquad \textit{Q.E.D.}$$

The following result, named after Guillaume de l'Hôpital, another seventeenth century French mathematician, is useful for computing limits of ratios when both numerator and denominator vanish.

Theorem 7.18. *Suppose $f(x), g(x)$ satisfy $f(a) = g(a) = 0$. If the ratio $f'(x)/g'(x)$ has a continuous extension at $x = a$, then so does $f(x)/g(x)$ with the* same *value there*:

$$\ell = \lim_{x \to a} \frac{f'(x)}{g'(x)} = \lim_{x \to a} \frac{f(x)}{g(x)}\,. \tag{7.20}$$

In particular, if $g'(a) \neq 0$, then the limiting value is $\ell = f'(a)/g'(a)$.

Proof: Define $Q(x) = f(x)/g(x)$, $R(x) = f'(x)/g'(x)$, for $x \neq a$, while $Q(a) = R(a) = \ell$. Our task is to prove that if R is continuous at $x = a$, then so is Q. For this purpose, we invoke the Cauchy Mean Value Formula (7.19) on the interval with endpoints a and x, keeping in mind that both f and g vanish at $x = a$. Thus, there exists y between a and x such that

$$Q(x) = \frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(y)}{g'(y)} = R(y). \tag{7.21}$$

Note that the hypothesis requires that $g'(x) \neq 0$ for $x \neq a$ sufficiently near a, so (7.21) implies $g(x) \neq 0$ for such $x \neq a$ also, and hence both ratios are defined nearby a.

Now suppose $U \subset \mathbb{R}$ is any open subset containing the limiting value: $\ell \in U$. Then, by continuity of R, there is an open interval $I = (\, a - r, a + r\,) \subset R^{-1}(U)$. If $x \in I$ lies in this interval, then the point y in (7.21), because it is closer to a than x, also lies in the interval: $y \in I$. Thus, $Q(x) = R(y) \in U$, and hence $I \subset Q^{-1}(U)$, which establishes the continuity of Q at $x = a$. *Q.E.D.*

Example 7.19. The ratio

$$\frac{e^x - 1}{x}$$

is of the form (7.20) at the point $a = 0$, with $f(x) = e^x - 1$ and $g(x) = x$ satisfying $f(0) = g(0) = 0$. Since† $f'(x) = e^x$, $g'(x) = 1$, l'Hôpital's Rule implies that the limiting value at $x = 0$ is

$$\lim_{x \to 0} \frac{e^x - 1}{x} = \frac{f'(0)}{g'(0)} = \frac{e^0}{1} = 1.$$

Next, consider the ratio

$$\frac{e^x - 1 - x}{x^2}$$

where now $f(x) = e^x - 1 - x$ and $g(x) = x^2$. In this case, $f'(x) = e^x - 1$, $g'(x) = 2x$, and so we cannot apply (7.20) as stated because $g'(0) = 0$. However, the previous limiting formula continues to hold, and we find that

$$\lim_{x \to 0} \frac{e^x - 1 - x}{x^2} = \lim_{x \to 0} \frac{e^x - 1}{2x} = \frac{1}{2} \lim_{x \to 0} \frac{e^x - 1}{x} = \frac{1}{2}.$$

8. Higher Order Derivatives and Taylor's Theorem.

Suppose $f(x)$ is a continuous function and possesses a derivative $g(x) = f'(x)$ at each point. If $g(x)$ is also continuous and has a derivative, then we call its derivative the *second derivative* of f and write $f''(x) = g'(x)$. And so on for third and higher order derivatives. The n^{th} derivative is often denoted by $f^{(n)}(x)$. Just as not every continuous function has a derivative everywhere, not every differentiable function has a second derivative, and so on.

As we saw, if $f(0) = 0$, the derivative $f'(0)$ exists if the function $q(x) = f(x)/x$ has a removable discontinuity with $q(0) = f'(0)$. Similarly, if $f'(0) = 0$, the second derivative $f''(0)$ exists if the function $r(x) = f'(x)/x$ has a removable discontinuity with $r(0) = f''(0)$. Somewhat surprisingly the second derivative $f''(0)$ is not obtained from the value of $f(x)/x^2$ at its removable discontinuity. To see what is going on, let us first write

$$f(x) = x\,q(x) \qquad \text{where} \qquad q(0) = f'(0) = 0. \tag{8.1}$$

Similarly, to find the derivative $q'(0)$, we set

$$q(x) = x\,r(x) \qquad \text{where} \qquad r(0) = q'(0). \tag{8.2}$$

On the other hand, if we differentiate (8.1) using the product rule followed by (8.2), we find

$$f'(x) = x\,q'(x) + q(x) = x\,\big[\,q'(x) + r(x)\,\big] = x\,s(x), \qquad \text{where} \qquad s(x) = q'(x) + r(x).$$

† See (15.7) below for a justification that the derivative of the exponential function $e(x) = e^x$ is itself: $e'(x) = e^x$.

Since we are assuming $f'(x)$ has a derivative at $x = 0$, the function $s(x)$ is continuous, with value

$$f''(0) = s(0) = q'(0) + r(0) = 2\,r(0).$$

Recapitulating, if $f(0) = f'(0) = 0$, then the removable discontinuity of the function

$$r(x) = \frac{f(x)}{x^2} \qquad \text{is given by} \qquad r(0) = \tfrac{1}{2}\, f''(0).$$

An alternative means of verifying this result is to apply l'Hôpital's Rule (7.20) twice to the ratio $r(x)$.

More generally, suppose f is differentiable, and $f'(x)$ has derivative $f''(a)$ at $x = a$. Recall the Carathéodory formulation (7.4) of the derivative:

$$f(x) = f(a) + (x-a)\,q(x), \qquad \text{where} \qquad q(a) = f'(a).$$

Similarly, we can write

$$q(x) = q(a) + (x-a)\,r(x), \qquad \text{where} \qquad r(a) = q'(a).$$

On the other hand, differentiating the first expression and substituting the second yields

$$f'(x) = q(x) + (x-a)\,q'(x) = g(a) + (x-a)\left[\,q'(x) + r(x)\,\right] = f'(a) + (x-a)\,s(x),$$

where

$$s(x) = q'(x) + r(x).$$

Thus,

$$f''(a) = s(a) = q'(a) + r(a) = 2\,r(a).$$

Putting the pieces together, we deduce that if $f(x)$ and $f'(x)$ are continuous, then $f''(a)$ exists if and only if the function $r(x)$ in

$$f(x) = f(a) + f'(a)(x-a) + (x-a)^2\, r(x)$$

is continuous, in which case $r(a) = \frac{1}{2}\, f''(a)$.

The general version of this result is known as Taylor's Theorem, named after the early eighteenth century British mathematician Brook Taylor. Suppose $f(x)$ is a function whose first n derivatives $f(a), f'(a), f''(a),\ \ldots\ , f^{(n)}(a)$ exist. Consider the polynomial function

$$P_n(x) = f(a) + f'(a)\,(x-a) + f''(a)\,\frac{(x-a)^2}{2} + f'''(a)\,\frac{(x-a)^3}{6} + \ \cdots\ + f^{(n)}(a)\,\frac{(x-a)^n}{n!} \tag{8.3}$$

known as the *Taylor polynomial of degree* n for the function $f(x)$ at the point $x = a$. Here

$$n! = n\,(n-1)(n-2)\ \cdots\ 3\cdot 2\cdot 1 \tag{8.4}$$

denotes the *factorial* of $n \in \mathbb{N}$. Note that, using (7.12, 13), the basic monomial function

$$p_n(x) = \frac{(x-a)^n}{n!} \qquad \text{has derivative} \qquad p_n'(x) = \frac{n\,(x-a)^{n-1}}{n!} = \frac{(x-a)^{n-1}}{(n-1)!} = p_{n-1}(x). \tag{8.5}$$

This implies that the derivatives of the Taylor polynomial (8.3), up to order n, coincide with the derivatives of f at the point $x = a$:

$$P_n(a) = f(a), \qquad P_n'(a) = f'(a), \qquad P_n''(a) = f''(a), \qquad \ldots \qquad P_n^{(n)}(a) = f^{(n)}(a). \tag{8.6}$$

On the other hand, since it has degree n, its $(n+1)^{\text{st}}$ derivative $P_n^{(n+1)}(x) \equiv 0$ for all x.

Theorem 8.1. *Suppose the first n derivatives $f(x), f'(x), f''(x), \ldots, f^{(n)}(x)$ are continuous. Let $P_n(x)$ be its Taylor polynomial of degree n at the point $x = a$, (8.3). Then the $(n+1)^{\text{st}}$ derivative $f^{(n+1)}(a)$ exists at $a \in \mathbb{R}$ if and only if the* remainder function *$R_{n+1}(x)$ defined by the* Taylor formula

$$f(x) = P_n(x) + R_{n+1}(x)\,\frac{(x-a)^{n+1}}{(n+1)!} \tag{8.7}$$

is continuous, in which case its value is $R_{n+1}(a) = f^{(n+1)}(a)$.

To prove the Theorem, we use (8.5) to write

$$R_{n+1}(x) = \frac{f(x) - P_n(x)}{p_{n+1}(x)},$$

and then repeatedly apply l'Hôpital's Rule (7.20) $n+1$ times using (8.5, 6) to arrive at its limiting value $R_{n+1}(a) = f^{(n+1)}(a)$.

Furthermore, there is a formula for the Taylor remainder function that generalizes the Mean Value Theorem 7.13.

Theorem 8.2. *Given the Taylor expansion* (8.7), *with continuous $R_{n+1}(x)$, then for $b > a$ there exists $a < y < b$, while for $b < a$, there exists $b < y < a$, such that*

$$R_{n+1}(b) = f^{(n+1)}(y). \tag{8.8}$$

The case $n = 0$ of Theorem 8.2 coincides with the Mean Value Theorem 7.13, because $P_0(x) = f(a)$ and so the Taylor expansion (8.7) reduces to

$$f(x) = f(a) + (x-a)\,R_1(x), \qquad \text{whereby} \qquad R_1(x) = \frac{f(x) - f(a)}{x - a}$$

is the difference quotient.

Proof: The following proof is modeled on that in [**7**]. Let us assume $a < b$; the proof for $b < a$ is essentially identical. Consider the function

$$Q_n(x) = f(x) + f'(x)\,(b-x) + f''(x)\,\frac{(b-x)^2}{2} + f'''(x)\,\frac{(b-x)^3}{6} + \cdots + f^{(n)}(x)\,\frac{(b-x)^n}{n!}. \tag{8.9}$$

Comparing with (8.3) we see that $Q_n(x)$ is the Taylor polynomial based at x evaluated at the point b. It is a polynomial in b, but a complicated function of the initial point x. In particular, at $x = a$ it coincides with the Taylor polynomial (8.3) based at a evaluated at the point $x = b$:

$$Q_n(a) = P_n(b), \qquad \text{while} \qquad Q_n(b) = f(b). \tag{8.10}$$

Let us compute its derivative, using the product and chain rules. By (7.12, 13), the derivative of

$$q_n(x) = \frac{(b-x)^n}{n!} \qquad \text{is} \qquad q_n'(x) = -\frac{(b-x)^{n-1}}{(n-1)!} = -q_{n-1}(x).$$

Thus,

$$Q_n'(x) = f'(x) + \big[-f'(x) + f''(x)\,(b-x)\big] + \left[-f''(x)\,(b-x) + f'''(x)\,\frac{(b-x)^2}{2}\right] + \cdots$$
$$\cdots + \left[-f^{(n)}(x)\,\frac{(b-x)^{n-1}}{(n-1)!} + f^{(n+1)}(x)\,\frac{(b-x)^n}{n!}\right].$$

Observe that the first and second terms cancel; the third and fourth terms also cancel; and so on; indeed, all the terms cancel out except the very last one, and so

$$Q_n'(x) = f^{(n+1)}(x)\,\frac{(b-x)^n}{n!}. \tag{8.11}$$

Now consider the function

$$h(x) = Q_n(x) - f(b) + C\,\frac{(b-x)^{n+1}}{(n+1)!}, \quad \text{where} \quad C = \frac{(n+1)!\,\big[f(b) - Q_n(a)\big]}{(b-a)^{n+1}} \tag{8.12}$$

is a constant, chosen so that

$$0 = h(a) = Q_n(a) - f(b) + C\,\frac{(b-a)^{n+1}}{(n+1)!}, \tag{8.13}$$

In addition, in view of the second equation in (8.10),

$$h(b) = Q_n(b) - f(b) = 0.$$

Thus, on the interval $[a,b]$, the function $h(x)$ satisfies the conditions $h(a) = h(b) = 0$ of Rolle's Theorem 7.12, and we conclude that there is a point $a < y < b$ such that

$$0 = h'(y) = Q_n'(y) - C\,\frac{(b-y)^n}{n!} = f^{(n+1)}(y)\,\frac{(b-y)^n}{n!} - C\,\frac{(b-y)^n}{n!},$$

using (8.11). Thus, $C = f^{(n+1)}(y)$. Substituting back into (8.13) and using the first equation in (8.10) yields

$$f(b) = Q_n(a) + f^{(n+1)}(y)\,\frac{(b-a)^{n+1}}{(n+1)!} = P_n(b) + f^{(n+1)}(y)\,\frac{(b-a)^{n+1}}{(n+1)!}.$$

Replacing b by x completes the proof of the remainder formula (8.8). *Q.E.D.*

9. Topology of the Real Plane.

Let us now generalize the preceding constructions and results to functions of two variables. Some aspects are more challenging than the one variable case considered so far, but once the two variable version is properly understood, further extensions to functions of more than two variables is reasonably straightforward. Some of the applications will include new results for functions of a single variable, as well as supplying the missing proofs in the preceding sections.

Recall that the two-dimensional plane $\mathbb{R}^2$ is the set of all ordered pairs of real numbers $\mathbf{v} = (x, y)$ where $x, y \in \mathbb{R}$ are its *Cartesian coordinates*, named after the seventeenth century French philosopher René Descartes who founded modern analytic geometry, and thereby revolutionized mathematics by linking geometry and algebra. The points $\mathbf{v} \in \mathbb{R}^2$ are also known as *vectors*, although we will not employ any of the underlying vector space properties, [**22**]. The *basic open sets* in $\mathbb{R}^2$ are the *open rectangles*

$$R_{a,c}^{b,d} = \{\, (x,y) \mid\ a < x < b, \quad c < y < d \,\} \tag{9.1}$$

where $a < b$ and $c < d$ are real numbers, in which case the rectangle is *bounded*, or $\pm\infty$ when unbounded. When the side lengths are equal and finite, so $d - c = b - a$, the rectangle is a *square*. We note that the intersection of any finite collection of open rectangles is either empty or an open rectangle; this is proved by induction starting with the case of two rectangles, and the details are left to the reader.

The definition of an *open set* $S \subset \mathbb{R}^2$ as a union of basic open sets is exactly as in Definition 2.1. We adapt Lemma 2.6, with an identical proof, to the current situation.

Lemma 9.1. *A set $S \subset \mathbb{R}^2$ is open if and only if for every $\mathbf{v} \in S$ there is an open rectangle R such that $\mathbf{v} \in R \subset S$.*

In fact, when verifying this criterion, one can restrict one's attention to squares, or even to squares or rectangles whose side lengths are sufficiently small. Let us look at some simple examples.

Example 9.2. The right half plane $S = \{\, (x, y) \mid x > 0 \,\}$ is open because it is a rectangle with $a = 0$, $b = \infty$, $c = -\infty$, $d = \infty$, so that $S = R_{0,-\infty}^{\infty,\infty}$. On the other hand, $T = \{\, (x, y) \mid x \geq 0 \,\}$ is not open because any open rectangle R containing a point $(0, y) \in T$ on the y axis must also contains points with negative x coordinate, and so $R \not\subset T$.

Example 9.3. Here we show that the *unit disk*, meaning the interior of the unit circle

$$D = \left\{\, (x,y) \;\middle|\; x^2 + y^2 < 1 \,\right\} \tag{9.2}$$

is open. (A simpler but non-direct proof can be found below.) Let $\mathbf{v} = (x, y) \in D$ and let $s = \sqrt{1 - x^2 - y^2} > 0$ denote the distance from $\mathbf{v}$ to the closest point on the unit circle $C = \{\, x^2 + y^2 = 1 \,\}$. Then we assert that the open square Q centered at $\mathbf{v}$ with side length $\sqrt{2}\, s$, namely $Q = R_{a,c}^{b,d}$ with $a = x - s/\sqrt{2}\,$, $b = x + s/\sqrt{2}\,$, $c = y - s/\sqrt{2}\,$, $d = y + s/\sqrt{2}\,$, is contained in D. Indeed, its corner points lie a distance s from the center (and so one of them is the closest point on the unit circle) and every point in Q lies closer to its center, and hence must be contained in the disk; see Figure 12. Since every point in D belongs to such an open square, we conclude that D is open. A similar argument proves that the exterior of the unit circle $E = \{\, (x, y) \mid x^2 + y^2 > 1 \,\}$ is also open.

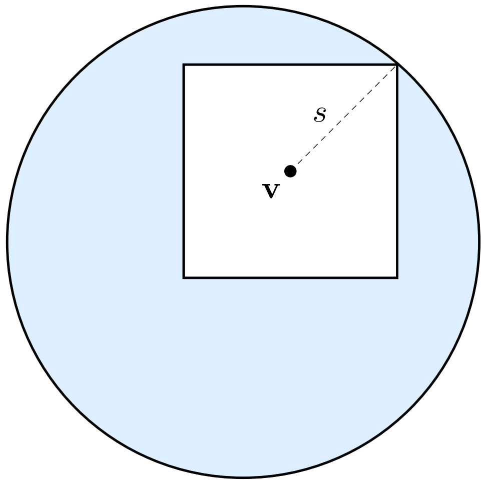


Figure 12. The Unit Disk and an Interior Square.

Remark: The analog of Proposition 2.7, which states that any open subset of $\mathbb{R}$ is a disjoint union of open intervals, is *not* valid in $\mathbb{R}^2$. For example, the interior of the unit circle cannot be written as the disjoint union of open rectangles. The reader may enjoy formally establishing this fact.

Lemma 9.4. *If $S \subset \mathbb{R}$ is open, then $U = \{\,(x,y) \mid x \in S\,\}$ and $V = \{\,(x,y) \mid y \in S\,\}$ are open subsets of $\mathbb{R}^2$.*

Proof: Given $(x,y) \in U$, let $(a,b) \subset S$ be an open interval containing x. Then the unbounded rectangle $R^{b,\infty}_{a,-\infty} = \{\,(x,y) \mid a < x < b\,\} \subset U$. Thus, every point $(x,y) \in U$ is contained in an open rectangle $R \subset U$, proving that U is open. The proof for V is identical. *Q.E.D.*

Lemma 9.5. *Let $S \subset \mathbb{R}^2$ be open. Let $p, q \in \mathbb{R}$. Then $U_q = \{\, x \in \mathbb{R} \mid (x,q) \in S\,\}$ and $V_p = \{\, y \in \mathbb{R} \mid (p,y) \in S\,\}$ are open subsets of $\mathbb{R}$.*

Proof: Let $x \in U_q$. Let $R^{b,d}_{a,c} \subset S$ be an open rectangle containing the point $(x,q) \in S$. Then $I = (a,b) \subset U_q$ is an open interval containing x. Since this occurs for every point in U_q we conclude that U_q is open. The proof for V_p is identical. *Q.E.D.*

Define the *Cartesian product* of subsets $A, B \subset \mathbb{R}$ to be the subset

$$A \times B = \{\,(x,y) \mid \ x \in A, \ y \in B\,\} \subset \mathbb{R}^2. \tag{9.3}$$

For example, the Cartesian product of a pair of open intervals $(a,b), (c,d) \subset \mathbb{R}$ is the open rectangle (9.1):

$$R^{b,d}_{a,c} = (a,b) \times (c,d).$$

Lemma 9.6. *If $S, T \subset \mathbb{R}$ are open, then their Cartesian product $S \times T \subset \mathbb{R}^2$ is open.*

Proof: It suffices to note that every point $x \in S$ is contained in an open interval, $x \in I \subset S$, and every point $y \in T$ is contained in an open interval $y \in J \subset T$. Thus, every point $(x, y) \in S \times T$ is contained in the corresponding Cartesian product rectangle: $(x, y) \in R = I \times J \subset S \times T$. *Q.E.D.*

Proposition 9.7. *If $f\colon \mathbb{R} \to \mathbb{R}$ is a continuous functions, then the sets*

$$S^+ = \{\, (x, y) \mid\ y > f(x) \,\}, \qquad S^- = \{\, (x, y) \mid\ y < f(x) \,\}, \tag{9.4}$$

lying above and below its graph are open.

Proof: We prove S^+ is open. Suppose $(x, y) \in S^+$ so $y > f(x)$. Let $c < f(x) < d < y < e$. Then $x \in f^{-1}(c, d)$, which is open. Let $a < x < b$ be such that $(a, b) \subset f^{-1}(c, d)$. We claim that the rectangle $R^{b,e}_{a,d} \subset S^+$. Indeed, given $(z, w) \in R^{b,e}_{a,d}$, so $z \in (a, b) \subset f^{-1}(c, d)$, we have $f(z) < d < w$, and hence $(z, w) \in S^+$. Thus, each point $(x, y) \in S^+$ is contained in a rectangle $R^{b,e}_{a,d} \subset S^+$, which implies that S^+ is open. The analogous proof for S^- is left to the reader. *Q.E.D.*

If $f(x)$ is discontinuous, this result may or may not be true. For the sign function (4.1),

$$S^+ = \{\, (x, y) \mid\ y > \operatorname{sign} x \,\}, \qquad \text{and} \qquad S^- = \{\, (x, y) \mid\ y < \operatorname{sign} x \,\},$$

are not open, whereas

$$S^+ = \left\{\, (x, y) \;\middle|\; y > 1/x^2,\ x \neq 0 \,\right\}, \qquad S^- = \left\{\, (x, y) \;\middle|\; y < 1/x^2,\ x \neq 0 \,\right\},$$

are open. We can even add in the y axis $Y = \{x = 0\}$ to the latter set, whereby

$$\widehat{S}^- = S^- \cup Y = \left\{\, (x, y) \;\middle|\; x = 0 \quad \text{or} \quad y < 1/x^2 \,\right\}$$

is also open. This can be proved using the method in Example 9.10 below.

Corollary 9.8. *If $f, g\colon \mathbb{R} \to \mathbb{R}$ are continuous functions, then the* strip

$$S = \{\, (x, y) \mid\ f(x) < y < g(x) \,\} = \{y > f(x)\} \cap \{y < g(x)\} \tag{9.5}$$

between their graphs is open.

Remark: Corollary 9.8 does not require that $f(x) < g(x)$ everywhere, and so their graphs are allowed to cross. Of course if $f(x) \geq g(x)$ everywhere, then $S = \varnothing$.

The theorem and corollary are also valid if the functions f, g are only defined on an open set. In particular, given an open interval $I \subset \mathbb{R}$, either bounded or unbounded,

$$S = \{\, (x, y) \mid\ x \in I, \quad f(x) < y < g(x) \,\} \tag{9.6}$$

is open.

Example 9.9. The open unit disk (9.2) can be characterized as the "strip"

$$D = \left\{\, (x, y) \;\middle|\; x^2 + y^2 < 1 \,\right\} = \left\{\, (x, y) \;\middle|\; -1 < x < 1, \quad -\sqrt{1 - x^2} < y < \sqrt{1 - x^2} \,\right\}$$

lying between the graphs of the two indicated continuous square root functions. Thus, the fact that it is open follows immediately from (9.6).

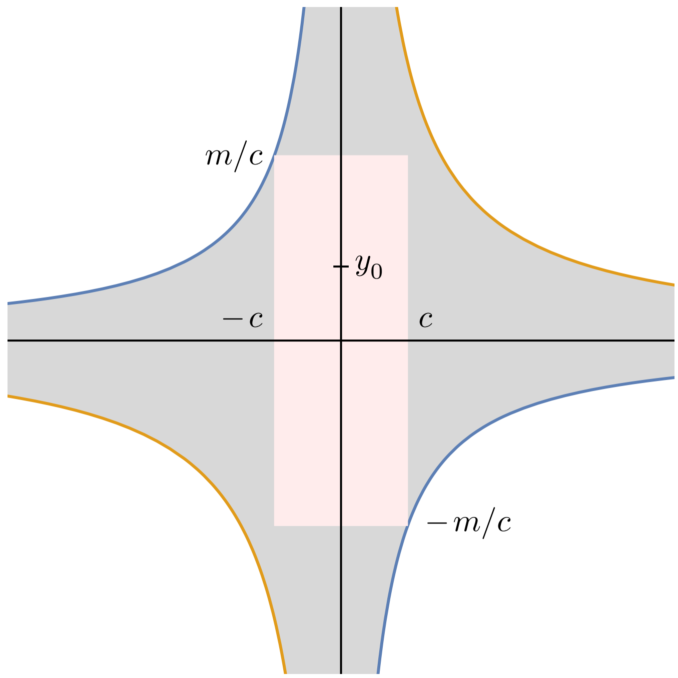


Figure 13. Hyperbolic Strip.

Example 9.10. Let $a < b$. Let us prove that the *hyperbolic strip*

$$S = \{ (x,y) \mid a < x\,y < b \} \tag{9.7}$$

lying between the hyperbolas $y = a/x$ and $y = b/x$ is open. Openness of a strip (9.6) implies that

$$S^+ = \{ (x,y) \mid x > 0, \quad a/x < y < b/x \} \quad \text{and} \quad S^- = \{ (x,y) \mid x < 0, \quad b/x < y < a/x \}$$

are both open. If $0 \le a < b$ or $a < b \le 0$ then $S = S^+ \cup S^-$ and hence is open. Otherwise, when $a < 0 < b$, the y axis $Y = \{ x = 0 \}$ also belongs to S, so $S = S^+ \cup S^- \cup Y$. Now since S^+ is open, each point $(x,y) \in S^+$ is contained in an open rectangle $R \subset S^+$. The same holds for S^-. Let $m = \min\{ -a, b \} > 0$. Given $(0, y_0) \in Y$, we can check that $(0, y_0) \in R^{c, m/c}_{-c, -m/c} = \{ (x,y) \mid |\,x\,| < c, \ |\,y\,| < m/c \} \subset S$ provided $0 < c < m/|\,y_0\,|$; see Figure 13. (If $y_0 = 0$ then any value of $c \in \mathbb{R}^+$ works.) Thus every point in S belongs to an open rectangle contained in S, proving that S is open.

As before, the complement of an open set is a closed set; in other words, $C \subset \mathbb{R}^2$ is *closed* if and only if $S = \mathbb{R}^2 \setminus C$ is open. An example is the closed rectangle

$$\overline{R}^{b,d}_{a,c} = \{ (x,y) \mid a \le x \le b, \quad c \le y \le d \}. \tag{9.8}$$

The closure $\overline{A}$ of a set $A \subset \mathbb{R}^2$ is the intersection of all closed subsets containing A, which is also the smallest closed subset containing it.

If S is open with closure $\overline{S}$, we define the common *boundary* of S and $\overline{S}$ to be the subset

$$\partial S = \partial \overline{S} = \overline{S} \setminus S. \tag{9.9}$$

We note that ∂S is closed because its complement is the union of two open sets:

$$\mathbb{R}^2 \setminus \partial S = S \cup (\mathbb{R}^2 \setminus \overline{S}).$$

For example, the closure of the open unit disk (9.2) is the closed unit disk

$$\overline{D} = \left\{ (x,y) \;\middle|\; x^2 + y^2 \leq 1 \right\},$$

and their boundary is the unit circle

$$C = \partial D = \partial \overline{D} = \left\{ (x,y) \;\middle|\; x^2 + y^2 = 1 \right\}, \tag{9.10}$$

which is itself a closed set. Similarly, the closed rectangle (9.8) is the closure of the open rectangle (9.1). Their boundary is the rectangular curve consisting of the four line segments joined at their endpoints:

$$\begin{aligned} \partial R^{b,d}_{a,c} = \partial \overline{R}^{b,d}_{a,c} = \overline{R}^{b,d}_{a,c} \setminus R^{b,d}_{a,c} = \{ (x,c) \mid a \leq x \leq b \} \cup \{ (x,d) \mid a \leq x \leq b \} \\ \cup \{ (a,y) \mid c \leq y \leq d \} \cup \{ (b,y) \mid c \leq y \leq d \}. \end{aligned} \tag{9.11}$$

By a *half open rectangle* we mean a subset of the form

$$\widetilde{R} = R^{b,d}_{a,c} \cup B, \qquad \text{where} \qquad \varnothing \neq B \subsetneq \partial R^{b,d}_{a,c}.$$

The more common cases are when B consists of 1, 2, or 3 of the line segments (9.11).

The definition of limit point for subsets $A \subset \mathbb{R}^2$ is modeled directly on Definition 5.10. There is a direct analog of the Limit Point Theorem 5.13, which is proved in an analogous fashion. First, a subset $A \subset \mathbb{R}^2$ is *bounded* if it is contained in a bounded closed rectangle:

$$D \subset \overline{R}^{b,d}_{a,c} \qquad \text{for some} \qquad a, b, c, d \in \mathbb{R}, \quad a \leq b, \quad c \leq d. \tag{9.12}$$

A set that is both closed and bounded is called *compact*. Thus, closed rectangles, closed disks, and their boundaries are all examples of compact sets. The planar version of the Limit Point Theorem 5.13 is proved in an analogous fashion.

Theorem 9.11. *Let $A \subset \mathbb{R}^2$ be a bounded infinite subset. Then the set of limit points of A is not empty.*

Proof: Suppose $A \subset R_0 = \overline{R}^{b,d}_{a,c}$ where $a < b$ and $c < d$. We cut R_0 in four half-size subrectangles, namely $\overline{R}^{p,d}_{a,c}, \overline{R}^{b,d}_{p,c}, \overline{R}^{b,q}_{a,c}, \overline{R}^{b,d}_{a,q}$ where $p = \frac{1}{2}(a+b)$, $q = \frac{1}{2}(c+d)$. Since A is infinite, at least one of these contains infinitely many points in A, and denote it as $R_1 = \overline{R}^{b_1,d_1}_{a_1,c_1}$. Its side lengths are half those of the original rectangle: $b_1 - a_1 = \frac{1}{2}(b-a)$, $d_1 - c_1 = \frac{1}{2}(d-c)$. Next, we divide R_1 into four half-size subrectangles, and choose one of them, denoted $R_2 = \overline{R}^{b_2,b_2}_{a_2,a_2}$, with side lengths $b_2 - a_2 = \frac{1}{4}(b-a)$, $d_2 - c_2 = \frac{1}{4}(d-c)$, that still contains infinitely many points in A. We repeat the process indefinitely, so that, for each $k \in \mathbb{N}$, we have constructed a rectangle

$$R_k = \overline{R}^{b_k,d_k}_{a_k,c_k} \quad \text{with side lengths} \quad b_k - a_k = 2^{-k}(b-a), \quad d_k - c_k = 2^{-k}(d-c), \tag{9.13}$$

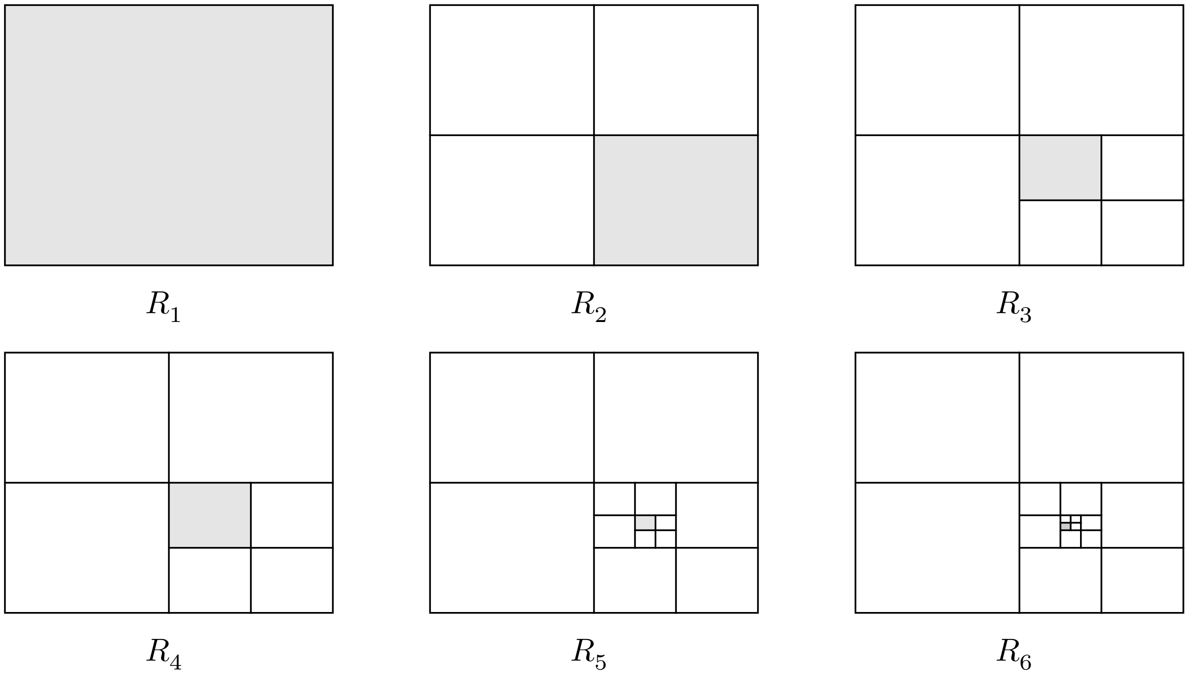


Figure 14. Subdividing a Rectangle.

that contains infinitely many points in A; see Figure 14, where the selected rectangles are shaded in gray. Observe that,

$$a \leq a_1 \leq a_2 \leq a_3 \leq \cdots \leq b_3 \leq b_2 \leq b_1 \leq b,$$
$$c \leq c_1 \leq c_2 \leq c_3 \leq \cdots \leq d_3 \leq d_2 \leq d_1 \leq d.$$

We set

$$x = \sup\{\, a_k \,\} = \inf\{\, b_k \,\}, \qquad y = \sup\{\, c_k \,\} = \inf\{\, d_k \,\},$$

where the equalities between the respective suprema and infinima are proved in exactly the same manner as in the proof of the one-dimensional Theorem 5.13, using the fact that both side lengths of the rectangles R_k, as given in (9.13), tend to 0 as $k \to \infty$. Finally, again in a similar manner, we verify that $\ell = (x, y)$ is a limit point of A because any open rectangle containing ℓ necesarily contains the small rectangles R_k for k sufficiently large, and hence contains infinitely many points in A. Details are left to the reader. *Q.E.D.*

Corollary 9.12. *If $K \subset \mathbb{R}^2$ is compact and $A \subset K$ is an infinite subset, then A has at least one limit point $\ell \in K$.*

The definition of an open cover $\mathcal{U}$ of a subset $S \subset \mathbb{R}^2$ is identical to the one-dimensional version Definition 5.24. The two-dimensional counterpart of the Heine–Borel Theorem 5.27 is then proved in a very similar fashion.

Theorem 9.13. *Let $K \subset \mathbb{R}^2$ be compact. Let $\mathcal{U}$ be a cover of K by open sets. Then there exist finitely many open sets $U_1, \ldots, U_n \in \mathcal{U}$ that cover $K \subset U_1 \cup \cdots \cup U_n$.*

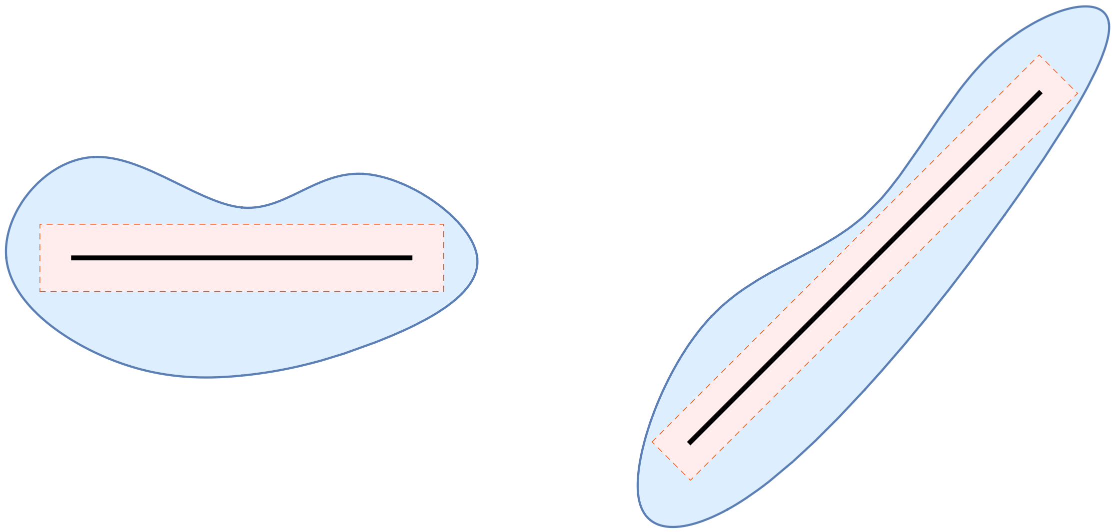


Figure 15. Horizontal and Diagonal Strips.

The only differences with the one-dimensional proof are as follows. In the first part, one replaces the open intervals $(x-r, x+r)$ by open squares $R^{x+r,y+r}_{x-r,y-r}$ centered on the points $(x,y) \in K$. In the second part of the argument, to prove the result for a compact rectangle $K = \overline{R}$, one takes a finite grid consisting of the points $(x_i, y_j) = (x_0 + i\,s, y_0 + j\,s) \in \overline{R}$ for $i, j \in \mathbb{Z}$, and noting that the corresponding open squares $Q_{ij} = R^{x_i+s,y_j+s}_{x_i-s,y_j-s}$ cover $\overline{R}$ and are each contained in an open set belonging to the cover: $Q_{ij} \subset U_{ij} \in \mathcal{U}$. The remainder of the proof proceeds exactly as before.

Two particular consequences of this result will be of use when we discuss uniform continuity and convergence.

Lemma 9.14. *Let $I = \{\,(x,0) \mid a \le x \le b\,\}$, where $a \le b$, be a compact line segment contained in the x axis. Suppose $I \subset U$ where $U \subset \mathbb{R}^2$ is open. Then there exists an open rectangular strip $R_r = R^{b+r,r}_{a-r,-r}$ of half width $r > 0$ such that $I \subset R_r \subset U$.*

The result is illustrated in the first plot in Figure 15, where the line segment I is black, the open set U is colored blue and the rectangle R_r is colored pink.

Proof: Since U is open, each point $(x,0) \in I$ is contained in an open square: $(x,0) \in Q_{x,s} = R^{x+s,s}_{x-s,-s} \subset U$ for some $s = s(x) > 0$. The collection of all such squares covers I. Because I is compact, the Heine–Borel Theorem 9.13 implies that there exists a finite subcover $Q_i = Q_{x_i,s_i} \subset U$ for $i = 1, \ldots, n$. Let $r = \min\{s_1, \ldots, s_n\} > 0$ be the minimum half width of the squares in the finite subcover. Then every point $(x,y) \in R_r$ in the rectangular strip of half width r belongs to one these squares, and so

$$I \subset R_r \subset \bigcup_{i=1}^{n} Q_i \subset U. \qquad Q.E.D.$$

If we subject the preceding interval to a 45° rotation, we establish a useful alternative;

details of the latter proof are left to the reader. (One can either work directly by say using rotated squares to cover the segment, or simply use the rotation to reduce it to the previous version, after noting that rotations take open sets to open sets.) The resulting configuration is illustrated in the second plot in Figure 15 with the same color scheme.

Lemma 9.15. *Let $I = \{ (x,x) \,|\, a \le x \le b \}$, where $a \le b$, be a compact diagonal line segment. Suppose $I \subset U$ where $U \subset \mathbb{R}^2$ is open. Then, there exists an open diagonal strip*

$$S_r = \{ (x,y) \mid \ |x-y| < r,\ 2(a-r) < x+y < 2(b+r) \},$$

of half-width $r > 0$ such that $I \subset S_r \subset U$.

Remark: Further results of this nature can be similarly established, but these two will suffice for our later purposes.

Remark: Both results require compactness of the line segment. For example, the entire x axis $X = \{ (x,0) \,|\, x \in \mathbb{R} \}$ is contained in the open set

$$X \subset U = \left\{ (x,y) \;\middle|\; |y| < \frac{1}{1+x^2} \right\}$$

is open, but there is no horizontal strip $R_r = \{ (x,y) \,|\, |y| < r \}$ for any $r > 0$ such that $X \subset R_r \subset U$. Similarly, the open unit disk $U = \{ x^2 + y^2 < 1 \}$ contains the non-closed line segment $I = \{ (x,0) \,|\, -1 < x < 1 \}$ forming its diameter, but there is no open rectangle R satisfying $I \subset R \subset U$.

Now let us consider more general subsets $A \subset \mathbb{R}^2$. As before, we construct a topology on A by defining a *relatively open subset* to be the intersection, $S \cap A$, with an open subset $S \subset \mathbb{R}^2$. In particular, the *basic open subsets* are the intersections of A with open rectangles: $R^{b,d}_{a,c} \cap A$. Any relatively open subset is a union of basic open subsets. If A itself is open, every relatively open subset $S \cap A$ is also an open subset of $\mathbb{R}^2$, although its basic open sets may no longer be rectangular.

For example, suppose $a < b$ and $c < d$, and let

$$A = \overline{R}^{b,d}_{a,c} = \{ (x,y) \mid \ a \le x \le b,\ c \le y \le d \} \subset \mathbb{R}^2$$

be a closed rectangle. Then the basic open subsets of A are the intersections $R \cap A$, where R is an open rectangle, which are either open rectangles, or half open rectangles, such as $\{ a \le x < e,\ f < y < g \}$ for $a < e \le b$ and $c \le f < g \le d$, or $\{ a \le x \le b,\ f < y \le d \}$ for $c \le f < d$, and so on, as well as the closed rectangle A itself.

Definition 9.16. A set S is called *disconnected* if it can be written as the disjoint union of two nonempty open subsets. In other words, there exist relatively open sets $S_1, S_2 \subsetneq S$ such that

$$S_1 \cup S_2 = S, \qquad S_1 \cap S_2 = \varnothing, \qquad S_1 \ne \varnothing, \qquad S_2 \ne \varnothing. \tag{9.14}$$

Vice versa, a set is *connected* if it is not disconnected.

Proposition 9.17. *If S is connected, the only subsets $A \subset S$ that are simultaneously relatively open and relatively closed are the set S itself and the empty set.*

Proof: Indeed, if $\varnothing \neq S_1 \subsetneq S$ is both open and closed, its complement $S_2 = S \setminus S_1 \neq \varnothing$ is also open and closed, and hence $S = S_1 \cup S_2$ would be the union of disjoint nonempty open subsets and thus disconnected. *Q.E.D.*

In $\mathbb{R}$ the only nonempty connected sets are the intervals; see Theorem 2.8. In $\mathbb{R}^2$ where there are many connected sets, including a rectangle (open or closed or half-open), a rectangular boundary, a circle, as well as its interior and exterior, etc. The most basic examples are the empty set, which is trivial, and the entire plane, as we now demonstrate.

Theorem 9.18. *The space $\mathbb{R}^2$ is connected.*

Proof: Suppose not, so that there exist disjoint open subsets $S_1, S_2 \subsetneq \mathbb{R}^2$ such that $S_1 \cup S_2 = \mathbb{R}^2$, $S_1 \cap S_2 = \varnothing$. Suppose $(0,0) \in S_1$. Given $(a,b) \neq (0,0)$, consider the line $L_{a,b} = \{ (t a, t b) \mid t \in \mathbb{R} \}$ passing through both points. We claim that the disjoint subsets $T_i = \{ t \in \mathbb{R} \mid (t a, t b) \in S_i \} \subset \mathbb{R}$, for $i = 1, 2$, are both open, and $T_1 \cup T_2 = \mathbb{R}$. Indeed, since S_i is open, any point $(t a, t b) \in S_i$ is contained in an open rectangle $R \subset S_i$, but then $\{ s \in \mathbb{R} \mid (s a, s b) \in R \} \subset T_i$ forms an open interval, proving that T_i is open. Since $0 \in T_1 \neq \varnothing$, Theorem 2.8 implies $T_2 = \varnothing$ and hence $(a,b) \in L_{a,b} \subset S_1$. Since this holds for every point (a,b), we conclude that $\mathbb{R}^2 = S_1$ and $S_2 = \varnothing$, thus proving connectivity. *Q.E.D.*

A similar proof can be used to prove the connectivity of rectangles, triangles, disks, etc. Indeed, the method of proof will show connectivity of any *star shaped region*, by which we mean a region D such that if $(a,b) \in D$, then so is every point on the line segment connecting it to the origin, i.e., $(t a, t b) \in D$ for all $0 \leq t \leq 1$. One can also replace the origin by any other point. Moreover, unions of connected regions with nonempty overlap are also connected; the proof of this result is left as an exercise.

Proposition 9.19. *If S and T are connected, and $S \cap T \neq \varnothing$, then $S \cup T$ is connected.*

10. Continuous Functions on the Plane.

Let us now investigate the continuity properties of functions $F\colon \mathbb{R}^2 \to \mathbb{R}$. Here, the value of the function at a point $(x,y) \in \mathbb{R}^2$ is the real number $F(x,y) \in \mathbb{R}$. The requirement that F be continuous is word for word the same as above. Specifically, F is *continuous* if and only if $F^{-1}(S) \subset \mathbb{R}^2$ is open whenever $S \subset \mathbb{R}$ is open. As before, we only need to check the inverse images of basic open sets, i.e., the open intervals. Thus, F is continuous if and only if, whenever $I \subset \mathbb{R}$ is an open interval, then $F^{-1}(I)$ is an open subset of $\mathbb{R}^2$. This is equivalent to the statement that, for every $(x,y) \in F^{-1}(I)$, there is an open rectangle R such that $(x,y) \in R \subset F^{-1}(I)$, or, alternatively, $F(R) \subset I$.

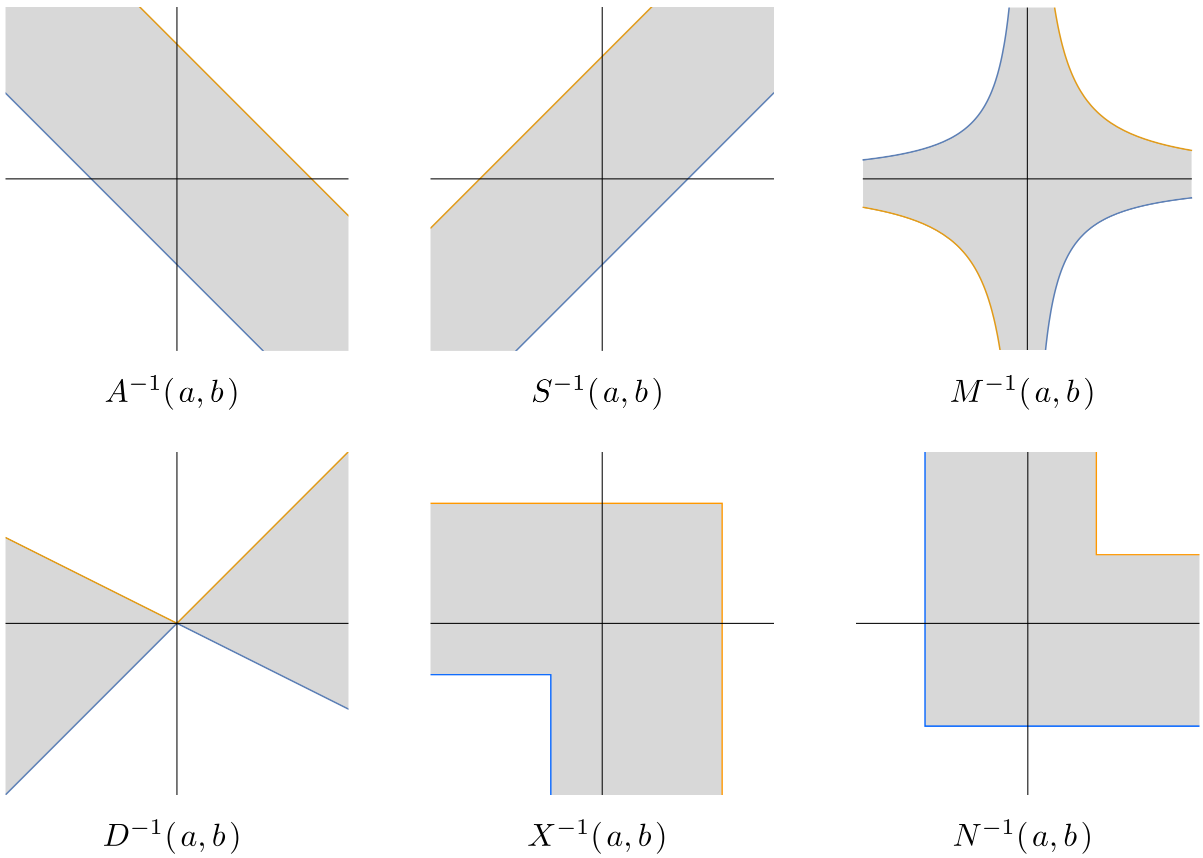


Figure 16. Inverse Images of Intervals.

Example 10.1.

(a) The addition function $A(x,y) = x+y$ from $\mathbb{R}^2$ to $\mathbb{R}$ is continuous. Indeed, if $(a,b) \subset \mathbb{R}$ is any open interval, then

$$A^{-1}(a,b) = \{ (x,y) \mid a < x+y < b \}$$

is the strip lying between the straight lines $y = a - x$ and $y = b - x$, which, as a consequence of Corollary 9.8, is open; see Figure 16.

(b) The subtraction function $S(x,y) = x - y$ from $\mathbb{R}^2$ to $\mathbb{R}$ is continuous. Indeed, if $(a,b) \subset \mathbb{R}$ is any open interval, then

$$S^{-1}(a,b) = \{ (x,y) \mid a < x-y < b \}$$

is the open strip lying between the straight lines $y = x - b$ and $y = x - a$.

(c) The multiplication function $M(x,y) = x\,y$ from $\mathbb{R}^2$ to $\mathbb{R}$ is continuous. Indeed, if $(a,b) \subset \mathbb{R}$, then

$$M^{-1}(a,b) = \{ (x,y) \mid a < x\,y < b \}$$

is a hyperbolic strip whose openness was established in Example 9.10.

(*d*) The division function $D(x,y) = x/y$ from $\mathbb{R}^2 \setminus \{ y = 0 \}$ to $\mathbb{R}$ is continuous. Indeed, if $(a,b) \subset \mathbb{R}$, then

$$D^{-1}(a,b) = \{ (x,y) \mid \; y > 0, \;\; a\,y < x < b\,y \} \cup \{ (x,y) \mid \; y < 0, \;\; b\,y < x < a\,y \}$$

is the union of two open wedges lying between the straight lines $x = a\,y$ and $x = b\,y$.

(*e*) The maximum function $X(x,y) = \max \{ x, y \}$ from $\mathbb{R}^2$ to $\mathbb{R}$ is continuous. Indeed, if $(a,b) \subset \mathbb{R}$, then

$$X^{-1}(a,b) = \{ (x,y) \mid \; a < \max \{ x, y \} < b \}$$

is the union of the two overlapping unbounded rectangles $R^{b,b}_{-\infty,a}$ and $R^{b,b}_{a,-\infty}$. Similarly, the minimum function $N(x,y) = \min \{ x, y \}$ is continuous since

$$N^{-1}(a,b) = \{ (x,y) \mid \; a < \min \{ x, y \} < b \}$$

is the union of $R^{\infty,b}_{a,a}$ and $R^{b,\infty}_{a,a}$.

The next result provides an easy way to build continuous functions of two variables from a continuous function of one variable. More interesting combinations will appear later.

Lemma 10.2. *If $f\colon \mathbb{R} \to \mathbb{R}$ is continuous, then the functions $F, G\colon \mathbb{R}^2 \to \mathbb{R}$ given by $F(x,y) = f(x)$ and $G(x,y) = f(y)$ are both continuous.*

Proof: Given an open set $S \subset \mathbb{R}$, we have

$$\begin{aligned} F^{-1}(S) &= \{ (x,y) \mid \; f(x) \in S \} = \left\{ (x,y) \;\middle|\; x \in f^{-1}(S) \right\}, \\ G^{-1}(S) &= \{ (x,y) \mid \; f(y) \in S \} = \left\{ (x,y) \;\middle|\; y \in f^{-1}(S) \right\}. \end{aligned}$$

Since $f^{-1}(S) \subset \mathbb{R}$ is open as a consequence of the continuity of f, Lemma 9.4 implies both $F^{-1}(S)$ and $G^{-1}(S)$ are open subsets of $\mathbb{R}^2$. *Q.E.D.*

Conversely, given a continuous function of two variables, we can fix the value of one of its variables to produce a continuous function of one variable, sometimes called "partial functions" since they each form part of the full function.

Proposition 10.3. *Let $F\colon \mathbb{R}^2 \to \mathbb{R}$ be continuous. Fix $a, b \in \mathbb{R}$ and set $f(x) = F(x,b)$ and $g(y) = F(a,y)$. Then the scalar functions $f, g\colon \mathbb{R} \to \mathbb{R}$ are both continuous.*

Proof: Let $S \subset \mathbb{R}$ be open. Then $f^{-1}(S) = \{ x \mid (x,b) \in F^{-1}(S) \}$. But continuity of F implies $F^{-1}(S) \subset \mathbb{R}^2$ is open. Lemma 9.5 implies $f^{-1}(S) \subset \mathbb{R}$ is open. A similar argument can be applied to g. *Q.E.D.*

Example 10.4. Consider the function

$$F(x,y) = \begin{cases} \dfrac{x\,y}{x^2 + y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0). \end{cases} \tag{10.1}$$

Then, given $a, b \neq 0$, the partial functions

$$f(x) = F(x,b) = \frac{b\,x}{x^2+b^2}, \qquad g(y) = F(a,y) = \frac{a\,y}{y^2+a^2},$$

are both continuous. Also, when $a = b = 0$,

$$f(x) = F(x,0) \equiv 0, \qquad g(y) = F(0,y) \equiv 0,$$

are also trivially continuous. Surprisingly, the function $F(x,y)$ is *not* continuous! Indeed, consider the open interval $\left(-\frac{1}{2}, \frac{1}{2}\right)$. Its inverse image is

$$\begin{aligned} A = F^{-1}\left(-\tfrac{1}{2}, \tfrac{1}{2}\right) &= \{(0,0)\} \cup \left\{ (x,y) \;\middle|\; |\,x\,y\,| < \tfrac{1}{2}(x^2+y^2) \right\} \\ &= \{(0,0)\} \cup \{ (x,y) \mid y \neq \pm x \}. \end{aligned}$$

The second equality follows from the fact that

$$\tfrac{1}{2}(x^2+y^2) > \pm x\,y \quad \text{if and only if} \quad (x \pm y)^2 = x^2 \pm 2\,x\,y + y^2 \neq 0.$$

The latter set, consisting of all points in $\mathbb{R}^2$ not on the two diagonals $y = x$ and $y = -x$, is open, but including the origin makes A into a non-open subset. Indeed, since $F(x, \pm x) = \pm\frac{1}{2}$ for all $0 \neq x \in \mathbb{R}$, there is no open rectangle R containing $(0,0) \in A$ such that $R \subset A$ because all points in R of the form $(x, \pm x)$ with $x \neq 0$ do not belong to A.

We conclude from the above example that the converse of Proposition 10.3 is false. Continuity of the two partial functions $f(x) = F(x,b)$ and $g(y) = F(a,y)$ is *not* sufficient to ensure the continuity of the full function $F(x,y)$!

Proposition 10.5. *Let $F\colon \mathbb{R}^2 \to \mathbb{R}$ be continuous. If $S \subset \mathbb{R}^2$ is connected, then $F(S) \subset \mathbb{R}$ is connected, and hence an interval.*

Proof: Suppose, to the contrary, that $F(S) \subset U_1 \cup U_2$ where $U_1, U_2 \subset \mathbb{R}$ are disjoint open subsets with $U_i \cap F(S) \neq \varnothing$ for $i = 1, 2$. Then $S \subset F^{-1}(U_1) \cup F^{-1}(U_2)$ where the $F^{-1}(U_i)$ are also nonempty and disjoint, and, because F is continuous, open. But this contradicts the assumed connectivity of S. *Q.E.D.*

Warning: It is *not* true that if $S \subset \mathbb{R}$ is connected, then so is $F^{-1}(S)$. For example, if

$$F(x,y) = x^2, \quad \text{then} \quad F^{-1}(1,4) = \{ (x,y) \mid 1 < x < 2 \} \cup \{ (x,y) \mid -2 < x < -1 \}.$$

Proposition 10.5 provides an immediate proof of the two variable version of the Intermediate Value Theorem 4.6.

Theorem 10.6. *Let $F\colon \mathbb{R}^2 \to \mathbb{R}$ be continuous. Suppose $S \subset \mathbb{R}^2$ is connected, and $F(a,b) = \alpha < F(c,d) = \beta$ for (a,b), $(c,d) \in S$. Then, given any $\alpha < z < \beta$, there exists $(x,y) \in S$ such that $F(x,y) = z$.*

Proof: Suppose not. Then we can write $S \subset S_1 \cup S_2$, where $S_1 = F^{-1}(-\infty, z)$, $S_2 = F^{-1}(z, \infty)$, which are disjoint, open subsets, and also nonempty since $(a,b) \in S_1$ and $(c,d) \in S_2$, which contradicts the connectivity of S. *Q.E.D.*

As before, we can restrict a function $F\colon \mathbb{R}^2 \to \mathbb{R}$ to a subdomain $A \subset \mathbb{R}^2$. If A is open, the continuity condition does not change because a relatively open subset $S \subset A$ is just an open subset of $\mathbb{R}^2$ that is contained in A. On the other hand, for more general subsets A one must analyze the structure of relatively open subsets of A in order to understand continuity of functions defined on A. Any continuous $F\colon \mathbb{R}^2 \to \mathbb{R}$ automatically restricts to a continuous function $F\colon A \to \mathbb{R}$ for any A contained in its domain. But, as in the one variable situation, it may not be possible to reverse the process.

The Extreme Value Theorem for functions of two variables states that a continuous function on a compact subset is bounded from above and below and achieves its maximum and minimum values. Both compactness requirements — bounded and closed — are essential for its validity.

Theorem 10.7. *Let $K \subset \mathbb{R}^2$ be compact. Let $F\colon K \to \mathbb{R}$ be a continuous function. Then its minimum $m = \inf F(K)$ and maximum $M = \sup F(K)$ are both finite. Moreover, there exist points $(x_m, y_m), (x_M, y_M) \in K$ such that $F(x_m, y_m) = m,\ \ F(x_M, y_M) = M$.*

Proof: The proof is an evident adaptation of the proofs of the one-dimensional version. One first establishes boundedness of F as in Theorem 5.16. Then the existence of the points $(x_m, y_m), (x_M, y_M) \in K$ follows the method used in the first part of the proof of the one variable Extreme Value Theorem 5.17; see also Theorem 5.23. *Q.E.D.*

If, in addition, K is connected, we can use the planar Intermediate Value Theorem 10.6 to establish a refinement similar to Theorem 5.17.

Theorem 10.8. *Let $K \subset \mathbb{R}^2$ be a connected compact set. If $F\colon K \to \mathbb{R}$ is continuous, then its range $F(K) = \{\, F(x,y) \mid (x,y) \in K \,\} \subset \mathbb{R}$ is a bounded closed interval: $F(K) = [\,m, M\,]$ where $M = \sup F(K)$ and $m = \inf F(K)$, which means that, given any $m \le z \le M$, there exists at least one point $(x,y) \in K$ such that $F(x,y) = z$.*

Let us next study functions that map the real line, or subsets thereof, to the two-dimensional plane. A vector-valued function $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^2$ is called *continuous* if $\mathbf{f}^{-1}(S) \subset \mathbb{R}$ is open whenever $S \subset \mathbb{R}^2$ is open. As always, we need only check when S is a basic open subset, i.e., an open rectangle. We can write $\mathbf{f}(t) = (f(t), g(t))$, where $f, g\colon \mathbb{R} \to \mathbb{R}$. We will mostly (but not always) use $t \in \mathbb{R}$ as the argument of $\mathbf{f}$ so as to distinguish it from points $(x,y) \in \mathbb{R}^2$.

Lemma 10.9. *If $f, g\colon \mathbb{R} \to \mathbb{R}$ are continuous, then $\mathbf{f}(t) = (f(t), g(t))$ defines a continuous function $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^2$.*

Proof: Given an open rectangle $R^{b,d}_{a,c}$, we see that $\mathbf{f}(t) \in R^{b,d}_{a,c}$ if and only if $a < f(t) < b$ and $c < g(t) < d$. Thus

$$\mathbf{f}^{-1}(R^{b,d}_{a,c}) = \{\, t \in \mathbb{R} \mid \ a < f(t) < b \,\} \cap \{\, t \in \mathbb{R} \mid \ c < g(t) < d \,\} = f^{-1}(a,b) \cap g^{-1}(c,d),$$

which, being the intersection of two open sets, is open. Since this holds for every open rectangle, we conclude that $\mathbf{f}$ is continuous. *Q.E.D.*

The proof that the composition of continuous functions is, in general, continuous proceeds exactly as before; see the earlier proof of the last statement of Theorem 4.3. With this in hand, we are finally in a position to give a short proof of the remaining statements in Theorem 4.3, on the continuity of algebraic combinations of continuous functions. Given continuous functions $f, g: \mathbb{R} \to \mathbb{R}$, let $\mathbf{f}(x) = (f(x), g(x))$ be the vector-valued function, $\mathbf{f}: \mathbb{R} \to \mathbb{R}^2$, which is itself continuous by Lemma 10.9. Let $A, S, M, D, X, N: \mathbb{R}^2 \to \mathbb{R}$ be the continuous functions in Example 10.1. Then, we can express

$$\begin{aligned} f(x) + g(x) &= A \circ \mathbf{f}(x), & f(x)\, g(x) &= M \circ \mathbf{f}(x), & \max\{ f(x), g(x) \} &= X \circ \mathbf{f}(x), \\ f(x) - g(x) &= S \circ \mathbf{f}(x), & f(x)/g(x) &= D \circ \mathbf{f}(x), & \min\{ f(x), g(x) \} &= N \circ \mathbf{f}(x). \end{aligned}$$

Each of these combinations can thus be expressed as the composition of two continuous functions, and hence is also a continuous function, thus completing the proofs of the heretofore unproved statements in Theorem 4.3.

Proposition 10.10. *If $\mathbf{f}: [a, b] \to \mathbb{R}^2$ is continuous, then its image $C = \mathbf{f}[a, b] \subset \mathbb{R}^2$ is a closed subset.*

Proof: Let $S = \mathbb{R}^2 \setminus C$. Given $(c, d) \in S$, consider the function

$$h(t) = \max \left\{ \, |\, f(t) - c \,|, \; |\, g(t) - d \,| \, \right\},$$

Since $h: [a, b] \to \mathbb{R}$ is built up from combinations of continuous functions, we can use Theorem 4.3 to verify that it is continuous. Thus, Theorem 5.16 implies that h is bounded on $[a, b]$. Let $\alpha \in \mathbb{R}$ be a lower bound for $h[a, b]$, so $h(t) \geq \alpha$ for all $a \leq t \leq b$. We claim that the open square $Q = R^{c+\alpha, d+\alpha}_{c-\alpha, d-\alpha} \subset S$. Indeed, if $(x, y) \in Q$ then both $|\, x - c \,|$ and $|\, y - d \,|$ are $< \alpha$, but this implies $(x, y) \in S$ since if $x = f(t)$, $y = g(t)$, then $h(t) < \alpha$ and hence $t \notin [a, b]$. This means that every point $(c, d) \in S$ belongs to an open square contained in S, which proves that S is open, and hence its complement $C = \mathbb{R}^2 \setminus S$ is closed. *Q.E.D.*

Given a continuous vector-valued function $\mathbf{f}: I \to \mathbb{R}^2$, where $I \subset \mathbb{R}$ is an interval (open, closed, half open, bounded or unbounded) we can identify its range with a parametrized plane curve:

$$C = \mathbf{f}(I) = \{ \, (f(t), g(t)) \mid \; t \in I \, \} \subset \mathbb{R}^2. \tag{10.2}$$

In the case $I = [a, b]$ is a closed interval, and hence the curve $C = \mathbf{f}[a, b]$ is a closed subset, we call $\mathbf{a} = \mathbf{f}(a)$ and $\mathbf{b} = \mathbf{f}(b)$ the *endpoints* of C.

The curve C is called *simple* if $\mathbf{f}$ is one-to-one, meaning that the curve does not cross itself; formally, if $s \neq t$ then $\mathbf{f}(s) \neq \mathbf{f}(t)$.

Theorem 10.11. *If $C = \mathbf{f}[a, b]$ is a simple curve, then its complement $\mathbb{R}^2 \setminus C$ is a connected open set.*

This result is, perhaps surprisingly, not easy to prove, and requires the sophisticated mathematical techniques developed to prove Theorem 10.12 below.

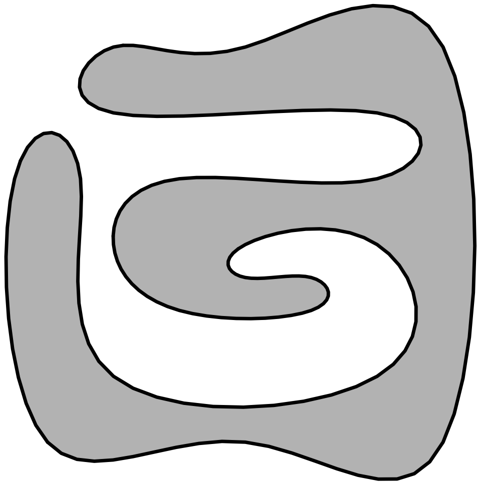


Figure 17. A Simple Closed or Jordan Curve.

A *closed curve* is the image of a continuous function $\mathbf{f}: [a,b] \to \mathbb{R}^2$ on a closed interval that satisfies $\mathbf{f}(a) = \mathbf{f}(b)$. Examples include the unit circle (9.10) and the rectangular boundary of an open rectangle (9.11). The closed curve is *simple* if $\mathbf{f}: [a,b) \to \mathbb{R}^2$ is one-to-one; equivalently, the only values $a \le s < t \le b$ for which $\mathbf{f}(s) = \mathbf{f}(t)$ are $s = a$ and $t = b$. A simple closed curve is also known as a *Jordan curve*, named after the nineteenth French mathematician Camille Jordan, who first rigorously proved the following theorem.

Warning: The term "closed" could be misleading here. Proposition 10.10 says that the image of $\mathbf{f}$ is a closed subset of $\mathbb{R}^2$ independent of the extra condition, but that is not what is usually implied by the term "closed curve".

Theorem 10.12. *Let $C \subset \mathbb{R}^2$ be a simple closed curve. The complement $\mathbb{R} \setminus C$ is the union of two disjoint nonempty open connected sets, one bounded — the curve's interior — and one unbounded — the curve's exterior.*

Figure 17 illustrates the theorem with a moderately complicated Jordan curve, whose interior is shaded and whose exterior is white. The Jordan Curve Theorem 10.12 seems intuitively obvious, but is surprisingly difficult to prove rigorously. One of the reasons for this difficulty is that the curve in question could be a complicated fractal. Jordan's famous proof, cf. [**8**], was a milestone in modern mathematics and, in particular, inspired the initial development of the field of algebraic topology, [**29**].

Finally, we consider vector-valued functions of two variables, that is $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2$, which is given by $\mathbf{F}(x,y) = \big(F(x,y), G(x,y)\big)$, where its components $F, G: \mathbb{R}^2 \to \mathbb{R}$ are scalar-valued. We say that $\mathbf{F}$ is *continuous* if $\mathbf{F}^{-1}(S) \subset \mathbb{R}^2$ is open whenever $S \subset \mathbb{R}^2$ is open. As always, we need only check the condition when S is a basic open subset, i.e., an open rectangle. The following analog of Lemma 10.9 is proved similarly, and thus left to the reader.

Lemma 10.13. *If* $F, G: \mathbb{R}^2 \to \mathbb{R}$ *are continuous, then* $\mathbf{F}(x,y) = \big(F(x,y), G(x,y)\big)$ *defines a continuous function* $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2$.

As always, the composition of continuous functions is easily shown to be continuous. Thus, by combining Lemma 10.13 with Example 10.1, we deduce that algebraic combinations of continuous functions of two variables, e.g.,

$$F(x,y) + G(x,y), \qquad F(x,y) - G(x,y), \qquad F(x,y)\,G(x,y),$$
$$\max\{\,F(x,y),\ G(x,y)\,\}, \qquad \min\{\,F(x,y),\ G(x,y)\,\},$$

are all continuous. The quotient function $F(x,y)/G(x,y)$ is continuous on the subset $D = \{\,(x,y) \mid G(x,y) \neq 0\,\}$. Note that since $D = G^{-1}(\mathbb{R} \setminus \{0\})$, continuity of G implies $D \subset \mathbb{R}^2$ is open, and nonempty unless $G(x,y) \equiv 0$ is the zero function.

We can specialize these combinations using Lemma 10.2. A particularly interesting combination is the function

$$G(x,y) = f(x) - f(y) \tag{10.3}$$

which is continuous whenever $f: \mathbb{R} \to \mathbb{R}$ is continuous. Moreover, G vanishes on the *diagonal*

$$\Delta = \{\, x = y \,\} = \{\, (x,x) \mid\ x \in \mathbb{R}\,\} \subset \mathbb{R}^2, \tag{10.4}$$

meaning $G(x,x) = 0$ for all $x \in \mathbb{R}$.

The next result establishes the *uniform continuity* of continuous functions defined on closed intervals.

Theorem 10.14. *Let* $f: [a,b] \to \mathbb{R}$ *be continuous. Given* $s > 0$*, there exists* $r > 0$ *such that* $|\,f(x) - f(y)\,| < s$ *whenever* $x, y \in [a,b]$ *and* $|\,x - y\,| < r$.

Proof: Let $G(x,y)$ be given by (10.3), which is defined on the closed square $Q = \overline{R}^{b,b}_{a,a} = [a,b] \times [a,b]$. Since G is continuous, $U = G^{-1}(-s,s) \subset Q$ is relatively open, and hence $U = \widehat{U} \cap Q$ where $\widehat{U} \subset \mathbb{R}^2$ is open and contains the diagonal: $\{\,(x,x) \mid x \in [a,b]\,\} \subset U \subset \widehat{U}$. Let $\widehat{S}_r \subset \widehat{U}$ be the diagonal strip produced by Lemma 9.15, and set $S_r = \widehat{S}_r \cap Q$. Then, given $a \le x, y \le b$ with $|\,x - y\,| < r$, this implies $(x,y) \in S_r \subset G^{-1}(-s,s)$ which in turn implies $|\,f(x) - f(y)\,| < s$. *Q.E.D.*

There is also a two variable analog of the Uniform Continuity Theorem 10.14, which can be stated as follows.

Theorem 10.15. *Let* $K \subset \mathbb{R}^2$ *be a compact subset, and* $F: K \to \mathbb{R}$ *a continuous function. Given* $s > 0$*, there exists* $r > 0$ *such that* $|\,F(x,y) - F(z,w)\,| < s$ *whenever* $(x,y),\ (z,w) \in K$ *with* $|\,x - z\,|,\ |\,y - w\,| < r$.

The proof relies on a similar analysis of the function $G(x,y,z,w) = F(x,y) - F(z,w)$ depending on four variables, so that† $G: \mathbb{R}^4 \to \mathbb{R}$, and we skip the details. For this

† Given $n \in \mathbb{N}$, we use $\mathbb{R}^n$ to denote the n-dimensional space consisting of all ordered n-tuples of real numbers (vectors) $\mathbf{v} = (x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

and many other reasons, e.g., the fact that we live in a three-dimensional universe, or, relativistically speaking, four-dimensional space-time, it would be worthwhile developing the calculus of functions of n variables all at once, but we have chosen to retain our focus on the two variable case to avoid excessive notational clutter. Moreover, adapting the results and proofs from functions of two variables to functions of more than two variables is almost always straightforward.

According to the remark following Theorem 5.3, a continuous one-to-one function on $\mathbb{R}$ maps open subsets to open subsets, and its inverse is also continuous. The two-dimensional analog of this result is true, but much harder to prove. It was finally established, in all dimensions, by the Dutch mathematician L.E.J. Brouwer in 1912, cf. [**9**]. This celebrated result is known as the Theorem on Invariance of Domain.

Theorem 10.16. *Let $U \subset \mathbb{R}^2$ be open. If $\mathbf{F}: U \to \mathbb{R}^2$ is continuous and one-to-one, then its range $V = \mathbf{F}(U)$ is open and its inverse $\mathbf{F}^{-1}: V \to U$ is continuous.*

Functions that are continuous and one-to one, hence with continuous inverse, are known as *homeomorphisms*. Theorem 10.16 and its generalizations imply that the dimension of a set cannot change under a homeomorphism. This intuitively "obvious" result stands in contrast to Cantor's surprising construction, in 1878, of a highly discontinuous function $f: I \to Q$ that maps the one-dimensional unit interval $I = [0,1]$ in a one-to-one manner to the two-dimensional unit square $Q = \overline{R}^{1,1}_{0,0} = \{ (x,y) \mid 0 \le x, y \le 1 \}$. Later, in 1890, the Italian mathematician Giuseppe Peano constructed a *space filling curve*, meaning a continuous map $f: I \to Q$ whose range is all of the unit square: $Q = f(I)$. However, due to the invariance of dimension, there is no one-to-one continuous function with this property. See also [**10**, **11**] for a detailed history and discussion of Cantor, Peano, and Brouwer's results and their subsequent reception and impact on what became known as dimension theory and on mathematics in general.

11. Uniform Convergence and the Cauchy Criterion.

In this optional section, we further investigate the convergence of sequences of real numbers, and also consider sequences of functions. In Examples 5.9 and 6.7, we saw how the convergence of a sequence of real numbers is equivalent to the continuity of an associated function on a certain subset of $\mathbb{R}$. Here we similarly investigate the convergence of sequences of scalar functions. In other words, we are given functions $f_n(x)$ for $n \in \mathbb{N}$ defined on a common interval, say $a \le x \le b$, and are interested in whether or not they converge, as $n \to \infty$, to a function $f_0(x)$ defined on the same interval, and if so, what the properties of the limiting function are.

To this end, let us return to the countably infinite subset

$$A = \{ 0 \} \cup \widetilde{A}, \quad \text{where} \quad \widetilde{A} = \{ 1/n \mid n \in \mathbb{N} \} \tag{11.1}$$

used in Example 5.9 to analyze convergence of sequences of real numbers. Here, given a compact interval $[a,b] \subset \mathbb{R}$, we introduce the following Cartesian product subsets of $\mathbb{R}^2$:

$$\begin{aligned} \widetilde{B} &= [a,b] \times \widetilde{A} = \{\, (x, 1/n) \mid \ a \le x \le b, \quad n \in \mathbb{N} \,\}, \\ B &= [a,b] \times A = I \cup \widetilde{B}, \qquad \text{where} \qquad I = [a,b] \times \{0\} = \{\, (x,0) \mid \ a \le x \le b \,\}. \end{aligned} \tag{11.2}$$

As always, the basic open subsets of both $\widetilde{B}$ and B are obtained by intersecting them with open rectangles $R \subset \mathbb{R}^2$. In the case of $\widetilde{B}$, by the same reasoning as in Example 5.9, these are the subsets of the form

$$\{\, (x, 1/n) \mid \ x \in J, \quad k \le n \le l \,\}, \quad \text{and} \quad \{\, (x, 1/n) \mid \ x \in J, \quad k \le n \,\}, \tag{11.3}$$

where $J \subset [a,b]$ is a relatively open interval, while $0 < k \le l$ are natural numbers. Any union of basic open subsets is also an open set, and we conclude that each open subset of $\widetilde{B}$ is a union of subsets of the form $U_n \times \{1/n\}$ for $n \in \mathbb{N}$, where $U_n \subset [a,b]$ is any relatively open subset, including the empty set. As for B, the same sets (11.3) remain basic open sets, along with the subsets

$$\{\, (x,y) \mid \ x \in J, \quad y = 0 \ \text{ or } \ y = 1/n \ \text{ for } \ n \ge k \,\}, \tag{11.4}$$

for some $k \in \mathbb{N}$. Again, the most general open subset is a union of the basic open subsets.

Now let us study continuity of functions on these sets. Consider a function $\widetilde{F} \colon \widetilde{B} \to \mathbb{R}$, which we can identify with a sequence of functions $f_n \colon [a,b] \to \mathbb{R}$ for $n \in \mathbb{N}$, where $f_n(x) = \widetilde{F}(x, 1/n)$. Since $\widetilde{F}$ is continuous, Proposition 10.3 tells us that its partial functions are continuous, and hence each individual f_n is continuous. On the other hand, continuity of $\widetilde{F}(x, 1/n)$ as a function of $1/n$ for fixed x imposes no constraints, in accordance with the discussion at the beginning in Example 6.7.

What about a function $F \colon B \to \mathbb{R}$? As above, its partial functions $f_n(x) = F(x, 1/n)$ for fixed $n \in \mathbb{N}$, as well as $f_0(x) = F(x, 0)$ are all continuous. On the other hand, again according to Example 6.7, continuity of $F(x,y)$ for fixed x as a function of $y \in A$, meaning that $y = 1/n$ or $y = 0$, is equivalent to convergence of the corresponding sequence $f_n(x)$, i.e.,

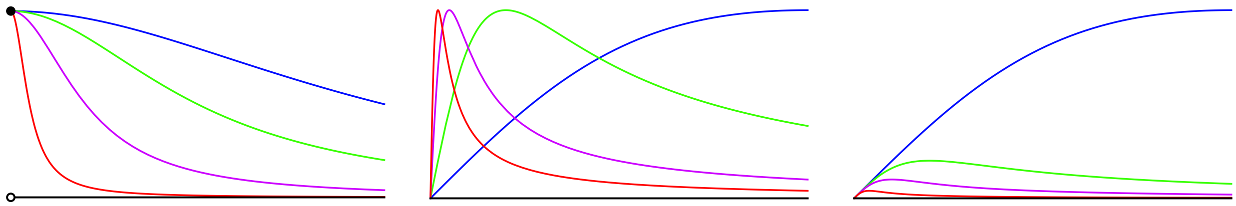
$$\lim_{n \to \infty} f_n(x) = \lim_{n \to \infty} F(x, 1/n) = F(x, 0) = f_0(x) \quad \text{for each} \quad x \in [a,b]. \tag{11.5}$$

The condition written in (11.5) is known as *pointwise convergence* of the sequence of functions $f_n(x)$ to the limiting function $f_0(x)$, since the equations hold at each individual point $x \in [a,b]$. If $F(x,y)$ is continuous, then the functions $f_n(x)$ and their pointwise limit $f_0(x)$ are all continuous. However, one can easily construct continuous functions whose pointwise limit is not continuous. An example is provided by

$$f_n(x) = \frac{1}{1 + n^2 x^2}, \quad \text{for} \quad 0 \le x \le 1, \quad n \in \mathbb{N},$$

with discontinuous pointwise limit

$$\lim_{n \to \infty} f_n(x) = f_0(x) = \begin{cases} 1, & x = 0, \\ 0, & x \ne 0. \end{cases}$$



Pointwise Convergence to Discontinuous Function — Pointwise Convergence to Continuous Function — Uniform Convergence to Continuous Function

Figure 18. Convergence of Functions.

It follows that the corresponding function $F: B \to \mathbb{R}$ is not continuous. This example is illustrated in the first plot in Figure 18; the blue, green, purple, red curves are the graphs of $f_1(x), f_2(x), f_5(x), f_{20}(x)$, respectively, while the discontinuous limit function f_0 is in black.

On the other hand, as we saw in Example 10.4, continuity of its partial functions does not guarantee continuity of a function. For, example restricting the function (10.1) to $B = [0,1] \times A$, suppose

$$F(x,y) = \begin{cases} f_n(x) = \dfrac{nx}{1+n^2x^2}, & y = 1/n, \quad n \in \mathbb{N}, \\ f_0(x) = 0, & y = 0, \end{cases} \qquad \text{for} \qquad 0 \le x \le 1. \tag{11.6}$$

Its partial functions $f_n(x), f_0(x)$ are individually continuous; moreover, they satisfy the pointwise convergence criterion (11.5):

$$\lim_{n \to \infty} f_n(x) = f_0(x).$$

Nevertheless, the full function (11.6) is not continuous on B by the same reasoning used in Example 10.4; indeed, its values $F(1/n, 1/n) = f_n(1/n) = 1/2$ do not converge to $0 = F(0,0)$ as $n \to \infty$. The convergence mechanism is illustrated in the second plot in Figure 18; the blue, green, purple, red graphs represent $f_1(x), f_5(x), f_{20}(x), f_{50}(x)$, respectively. Even though the functions eventually converge to zero at every point $x \in [0,1]$, their graphs all ascend to the same maximum value of $1/2$ while the overall width of this excursion gets thinner and thinner as $n \to \infty$.

Continuity of a function $F: B \to \mathbb{R}$ imposes an additional requirement beyond the individual functions $f_n(x)$ being continuous and converging pointwise to the continuous limit function $f_0(x)$. This amounts to a "uniformity" of the convergence mechanism, expressing, in a sense, that the individual pointwise limits (11.5) converge "at the same rate". To find what this entails, and motivated by our proof of the Uniform Continuity Theorem 10.14, we look at the difference between the function and its limit, setting

$$G(x,y) = F(x,y) - F(x,0) \quad \text{for} \quad x \in [a,b], \quad y \in A.$$

Since both summands are continuous, G is also continuous. Moreover, $G(x,0) = 0$ for all

$a \le x \le b$. Let $J = (-s, s)$ be an open interval centered around 0. Then $U = G^{-1}(J) \subset B$ is relatively open and contains the interval $I = \{ (x,0) \,|\, a \le x \le b \} \subset G^{-1}(0) \subset U$. According to Lemma 9.14 (restricted to B), this implies that U contains a set given by intersecting a rectangular strip with B, and thus of the form

$$S_{1/k} = \{ (x,y) \;|\; a \le x \le b,\; y = 0 \;\text{ or }\; y = 1/n \;\text{ for all }\; n \ge k \}, \quad \text{for some} \quad k \in \mathbb{N}.$$

The fact that $S_{1/k} \subset U = G^{-1}(J)$ means that

$$|\, G(x, 1/n) \,| = |\, F(x, 1/n) - F(x, 0) \,| = |\, f_n(x) - f_0(x) \,| < s \qquad \text{whenever} \quad a \le x \le b, \quad n \ge k. \tag{11.7}$$

Condition (11.7) can be interpreted as saying that the functions $f_n(x)$ are *uniformly close* to their limit $f_0(x)$, because the measure of closeness, as determined by $s > 0$, does not depend on the point x, which it may well do for pointwise convergence. Condition (11.7) is known as *uniform convergence* of the functions $f_n(x)$ to $f(x)$. Uniform convergence implies pointwise convergence, but the converse is not necessarily valid; the functions (11.6) converge pointwise to the zero function $f_0(x) \equiv 0$, but, in view of Theorem 11.1 below, cannot converge uniformly since the corresponding F is not continuous. As an illustration of uniform convergence, the third plot in Figure 18 graphs the functions $f_n(x) = x/(1 + n^2 x^2)$ on the interval $[0, 1]$, which converge uniformly to $f_0(x) \equiv 0$; here, the blue, green, purple, red graphs represent $f_1(x), f_5(x), f_{10}(x), f_{20}(x)$. Because of uniformity, and in contrast to the first and second plots, one can draw an arbitrarily thin strip around the limiting function that contains the entire graphs of each f_n when n is sufficiently large.

Remark: The preceding analysis remains valid when the closed bounded interval $[a, b]$ is replaced by any compact subset $K \subset \mathbb{R}$. It does not hold if one replaces it by a more general subset S, e.g., an open interval or an unbounded interval, either closed or open, including all of $\mathbb{R}$. However even in this case, continuity of the function $F: S \times A \to \mathbb{R}$ is still stronger than continuity of its partial functions, i.e., pointwise convergence of its continuous constituents $f_n(x)$. Continuity of F still implies that $U = G^{-1}(J) \subset S \times A$ is relatively open, but the final statement about the existence of a strip of uniform width contained in U is not necessarily valid in this more general, non-compact situation, as we already noted after Lemma 9.14. On the other hand, if we restrict the functions to any compact subset $K \subset S$, the convergence does satisfy the uniformity condition (11.7) even though it might fail on all of S. This type of convergence of functions is known as *uniform convergence on compact subsets*, which is sometimes abbreviated as *compact convergence*.

We have thus shown that a continuous function $F: B \to \mathbb{R}$ can be identified with a uniformly converging (or, more generally, compactly converging) sequence of continuous functions with a continuous limiting function. But this result is unsatisfying, since it assumes that the limiting function $f_0(x) = F(x, 0)$ is continuous from the outset. In fact, it turns out that the uniform convergence result is valid without this assumption, and can be stated as follows. For simplicity, we will just work on an interval.

Theorem 11.1. *Let $L \subset \mathbb{R}$ be an interval, and let $B = L \times A$, $\widetilde{B} = L \times \widetilde{A}$, where $A, \widetilde{A}$ are as in* (11.1). *Let $F: B \to \mathbb{R}$, and suppose that its restriction $\widetilde{F}: \widetilde{B} \to \mathbb{R}$ is continuous. Set $G(x,y) = F(x,y) - F(x,0)$. Then F is continuous if and only if it satisfies the uniformity condition: If $0 \in V \subset \mathbb{R}$ is open, then $G^{-1}(V)$ contains an open subset U satisfying $L \times \{0\} \subset U \subset G^{-1}(V)$.*

Remark: The uniformity condition on G is similar to the Definition 5.18 of continuity of a function at a point, in that, in accordance with the remarks preceding the statement of the Theorem, we do not assume continuity of G on its entire domain, but only "at the subset" $L \times \{0\}$. Continuity of F then implies both uniform convergence, or, more generally, compact convergence of the sequence of continuous functions $f_n(x) = F(x, 1/n)$, and continuity of the limit function $f_0(x) = F(x,0)$, which in turn implies that G is also continuous.

Proof: If F is continuous on all of B, then $f_0(x) = F(x,0)$ is continuous, and hence G is continuous, proving that it satisfies the uniformity condition.

To prove the converse, let $V \subset \mathbb{R}$ be any open subset. If $0 \notin V$, then $F^{-1}(V) = \widetilde{F}^{-1}(V)$ is open since $\widetilde{F}$ is assumed continuous. Thus, we only need analyze the case when $0 \in V$. Suppose $(x_0, y_0) \in F^{-1}(V)$. If $y_0 \neq 0$, then $(x_0, y_0) \in \widetilde{F}^{-1}(V) \subset F^{-1}(V)$, and the first set is again open by the continuity of $\widetilde{F}$. Thus, the only nontrivial case is when $y_0 = 0$, and we will construct a relatively open rectangle $R \subset B$ such that $(x_0, 0) \in R \subset F^{-1}(V)$, thus establishing continuity of F.

Let $z_0 = F(x_0, 0)$. Choose $t > 0$ so that $I = (z_0 - t, z_0 + t) \subset V$. For this value of t, let $J = \left(-\frac{1}{4}t, \frac{1}{4}t\right)$. By our assumption of uniformity, there exists an open set U satisfying $L \times \{0\} \subset U \subset G^{-1}(J)$, which is equivalent to the inequality

$$|\,G(x,y)\,| < \tfrac{1}{4}t \qquad \text{whenever} \qquad (x,y) \in U. \tag{11.8}$$

Since U is open and contains $(x_0, 0)$, it contains a relativity open square centered there: $Q = R^{x_0+s,s}_{x_0-s,-s} \cap B \subset U$ for some $s > 0$. Choose $n_* \in \mathbb{N}$ such that $0 < y_* = 1/n_* < s$. Since $f_*(x) = F(x, y_*)$ is continuous, we can choose $0 < r < s$ such that

$$|\,f_*(x) - f_*(x_0)\,| < \tfrac{1}{4}t \qquad \text{whenever} \qquad |\,x - x_0\,| < r. \tag{11.9}$$

Now suppose $(x,y) \in R = R^{x_0+r,s}_{x_0-r,-s} \cap B \subset Q \subset U$. Then, adding and subtracting several terms,

$$\begin{aligned} F(x,y) - F(x_0,0) &= \big[\,F(x,y) - F(x,0)\,\big] - \big[\,F(x,y_*) - F(x,0)\,\big] \\ &\qquad + \big[\,F(x,y_*) - F(x_0,y_*)\,\big] + \big[\,F(x_0,y_*) - F(x_0,0)\,\big] \\ &= G(x,y) - G(x,y_*) + \big[\,f_*(x) - f_*(x_0)\,\big] + G(x_0,y_*), \end{aligned}$$

which, using (11.8), (11.9), and the fact that $(x,y), (x,y_*), (x_0,y_*) \in R \subset U$, implies

$$\begin{aligned} |\,F(x,y) - z_0\,| &\le |\,G(x,y)\,| + |\,G(x,y_*)\,| + |\,f_*(x) - f_*(x_0)\,| + |\,G(x_0,y_*)\,| \\ &< \tfrac{1}{4}t + \tfrac{1}{4}t + \tfrac{1}{4}t + \tfrac{1}{4}t = t. \end{aligned}$$

This means that $F(x,y) \in I = (\,z_0 - t, z_0 + t\,)$, and hence the relatively open rectangle $R \subset F^{-1}(I) \subset F^{-1}(V)$. *Q.E.D.*

Corollary 11.2. *Let $L \subset \mathbb{R}$ be an interval. Let $f_n\colon L \to \mathbb{R}$ for $n \in \mathbb{N}$ be continuous and converge compactly to $f_0\colon L \to \mathbb{R}$ as $n \to \infty$. Then f_0 is continuous.*

We next conduct a detailed analysis of another particularly interesting example of the restriction of a continuous function to a subset.

Example 11.3. Let us return to the countably infinite subset $A = \{\,0\,\} \cup \{\,1/n \mid n \in \mathbb{N}\,\}$ appearing in (11.1). As we saw, a function $f\colon A \to \mathbb{R}$ is continuous if and only if the corresponding sequence $f_n = f(1/n)$ has a limit, namely $\lim\limits_{n\to\infty} f_n = f(0)$.

Consider the following Cartesian product subsets of $\mathbb{R}^2$:

$$\begin{aligned} C &= A \times A = \{\,(x,y) \mid \; x = 0 \;\text{ or }\; 1/x \in \mathbb{N} \;\text{ and }\; y = 0 \;\text{ or }\; 1/y \in \mathbb{N}\,\}, \\ \widetilde{C} &= \widetilde{A} \times \widetilde{A} = \{\,(1/m, 1/n) \mid \; m,n \in \mathbb{N}\,\}. \end{aligned} \tag{11.10}$$

Note first that

$$C = \{\,(0,0)\,\} \cup \big(\,\widetilde{A} \times \{\,0\,\}\,\big) \cup \big(\,\{\,0\,\} \times \widetilde{A}\,\big) \cup \widetilde{C}.$$

The relatively open subsets are obtained by intersecting C or $\widetilde{C}$ with open rectangles. An analysis similar to the one-dimensional version in Example 5.9 shows that any subset $\widetilde{S} \subset \widetilde{C}$ is open. Now suppose $S \subset C$ is open. If the point $(0, 1/n) \in S$, then S must also contain a countably infinite subset of the form $\{\,(1/m, 1/n) \mid M \le m \in \mathbb{N}\,\} \subset \widetilde{C}$ for some $M \in \mathbb{N}$; similarly if $(1/m, 0) \in S \subset C$, then S must contain a subset of the form $\{\,(1/m, 1/n) \mid N \le n \in \mathbb{N}\,\} \subset \widetilde{C}$ for some $N \in \mathbb{N}$. Finally, if $(0,0) \in S \subset C$, then S must also contain a subset of the form $\{\,(1/m, 1/n) \mid M \le m \in \mathbb{N},\; N \le n \in \mathbb{N}\,\} \subset \widetilde{C}$ for some $M, N \in \mathbb{N}$. Replacing M, N by their maximum, we simplify the latter criterion to just be that S contain a subset of the form

$$U_M = \{\,(1/m, 1/n) \mid \; M \le m, n \in \mathbb{N}\,\} \subset \widetilde{C}. \tag{11.11}$$

Any subset satisfying these three requirements is open. The limit points of C are its points on the x and y axis, namely $(0,0), (1/m, 0), (0, 1/n)$ for $m, n \in \mathbb{N}$.

A function $\widetilde{F}\colon \widetilde{C} \to \mathbb{R}$ can be identified with a doubly infinite sequence:

$$f_{m,n} = \widetilde{F}\left(\frac{1}{m}, \frac{1}{n}\right) \in \mathbb{R} \qquad \text{for all} \qquad m,n \in \mathbb{N}.$$

Any such function is continuous since all subsets of $\widetilde{C}$ are open. What about a function $F\colon C \to \mathbb{R}$? In this case, the values of F on the points in C with one or both coordinates equal to 0 serve to define limits of the doubly infinite sequence. Let us set

$$a = F(0,0), \qquad b_m = F\left(\frac{1}{m}, 0\right), \qquad c_n = F\left(0, \frac{1}{n}\right), \qquad f_{m,n} = F\left(\frac{1}{m}, \frac{1}{n}\right).$$

Then continuity of F implies that all the limits of the doubly infinite sequence $f_{m,n}$ exist:

$$\begin{aligned}
c_n &= \lim_{m\to\infty} f_{m,n} = \lim_{1/m\to 0} F\left(\frac{1}{m},\frac{1}{n}\right) = F\left(0,\frac{1}{n}\right),\\
b_m &= \lim_{n\to\infty} f_{m,n} = \lim_{1/n\to 0} F\left(\frac{1}{m},\frac{1}{n}\right) = F\left(\frac{1}{m},0\right),\\
a &= \lim_{m,n\to\infty} f_{m,n} = \lim_{(1/m,1/n)\to(0,0)} F\left(\frac{1}{m},\frac{1}{n}\right)\\
&= \lim_{m\to\infty} b_m = \lim_{1/m\to 0} F\left(\frac{1}{m},0\right) = \lim_{n\to\infty} c_n = \lim_{1/n\to 0} F\left(0,\frac{1}{n}\right).
\end{aligned}$$

As usual, if $F\colon\mathbb{R}^2\to\mathbb{R}$ is continuous, its restriction $F\colon C\to\mathbb{R}$ is also continuous, and hence obeys the above limiting conditions. Vice versa, given a continuous $F\colon C\to\mathbb{R}$, we can reconstruct a continuous $F\colon\mathbb{R}^2\to\mathbb{R}$ that agrees with it when restricted to C by using a suitable interpolation method, e.g., by using piecewise linear functions on triangles. We leave it to the motivated reader to fill in the details.

Now, let us look at the following subset:

$$\widehat{C} = \{\,(0,0)\,\} \cup \widetilde{C}, \tag{11.12}$$

i.e., we delete all the limit points of $C \supset \widetilde{C}$ except for the origin. As before, any subset of $\widetilde{C}$ is open in $\widehat{C}$. On the other hand, if $(0,0)\in S\subset\widehat{C}$ is open, then, as above, S must contain a subset of the form (11.11).

Now, let us return to our one-dimensional discrete set $\widetilde{A} = \{\,1/n \mid n\in\mathbb{N}\,\}$. Let

$$\widetilde{f}\colon\widetilde{A}\longrightarrow\mathbb{R}, \qquad \text{with} \qquad f_n = \widetilde{f}(1/n)$$

be any function, or, equivalently, any sequence $\{\,f_1, f_2, f_3, \,\ldots\,\}$. We are interested in whether or not the sequence $\{\,f_n\,\}$ converges or, equivalently, whether there exists a continuous extension

$$f\colon A\longrightarrow\mathbb{R} \quad \text{such that} \quad f(1/n) = \widetilde{f}(1/n) \quad \text{for all} \quad n\in\mathbb{N},$$

in which case

$$\lim_{n\to\infty} f_n = \lim_{1/n\to 0} f(1/n) = f(0).$$

The following convergence criterion is named after the nineteenth century French mathematician Augustin-Louis Cauchy, a key founder of modern analysis.

Definition 11.4. The function $\widetilde{f}\colon\widetilde{A}\to\mathbb{R}$, or corresponding sequence $f_n = \widetilde{f}(1/n)$, satisfies the *Cauchy criterion* if

$$\widehat{G}\left(\frac{1}{m},\frac{1}{n}\right) = \widetilde{f}\left(\frac{1}{m}\right) - \widetilde{f}\left(\frac{1}{n}\right) = f_m - f_n, \qquad \widehat{G}(0,0) = 0, \tag{11.13}$$

defines a continuous function $\widehat{G}\colon\widehat{C}\to\mathbb{R}$.

The function $\widehat{G}$ in (11.13) is constructed in analogy with $G(x,y) = f(x) - f(y)$ in (10.3), whose continuity was noted above. Note that it vanishes on the diagonal

$$\widehat{C} \cap \Delta = \{ (0,0) \} \cup \{ (1/m, 1/m) \mid m \in \mathbb{N} \},$$

cf. (10.4), so that

$$\widehat{G}(0,0) = \widehat{G}(1/m, 1/m) = 0 \quad \text{for all} \quad m \in \mathbb{N}.$$

An important analytical result states that a sequence that satisfies the Cauchy criterion has a unique limit point.

Theorem 11.5. *If the function $\widetilde{f}: \widetilde{A} \to \mathbb{R}$ or, equivalently, the infinite sequence $f_n = \widetilde{f}(1/n)$ satisfies the Cauchy criterion, meaning the function $\widehat{G}: \widehat{C} \to \mathbb{R}$ defined in* (11.13) *is continuous, then there exists a continuous extension $f: A \to \mathbb{R}$ such that $f(1/n) = \widetilde{f}(1/n)$ for all $n \in \mathbb{N}$, and hence the corresponding sequence converges with limiting value $f(0)$. Moreover, one can extend $\widehat{G}$ to a continuous function $G: C \to \mathbb{R}$ by setting $G(x,y) = f(x) - f(y)$ for all $x, y \in A$.*

Proof: Given any $r > 0$, consider the open interval $(-r, r)$, which contains the origin. Thus, by continuity,

$$\begin{aligned} \widehat{G}^{-1}(-r,r) &= \{ (0,0) \} \cup \left\{ \left(\frac{1}{m}, \frac{1}{n} \right) \;\middle|\; \left| \widehat{G}\left(\frac{1}{m}, \frac{1}{n} \right) \right| < r \right\} \\ &= \{ (0,0) \} \cup \left\{ \left(\frac{1}{m}, \frac{1}{n} \right) \;\middle|\; | f_m - f_n | < r \right\} \end{aligned}$$

is open in $\widehat{C}$, meaning that it contains a subset U_M of the form (11.11) for some $M \in \mathbb{N}$. In other words, the sequence $\{ f_n \}$ satisfies the Cauchy criterion if and only if for every $r > 0$ there exists $M \in \mathbb{N}$ such that[†]

$$| f_m - f_n | < r \qquad \text{whenever} \qquad m, n \geq M. \tag{11.14}$$

The first thing to note is that (11.14) implies that the subsequence

$$S_M = \{ f_n \mid n \geq M \} \tag{11.15}$$

is bounded[‡]. To see this, setting $m = M$ in (11.14) implies

$$f_M - r \leq f_n \leq f_M + r \qquad \text{whenever} \qquad n \geq M.$$

There are two cases to consider. Suppose first that the subsequence (11.15) forms a finite set, say $S_M = \{x_1, \ldots, x_k\}$, so that for each $n \geq M$, we have $f_n = x_i$ for some $1 \leq i \leq k$ depending on n. Choose

$$0 < \widetilde{r} < \min \left\{ | x_i - x_j | \;\middle|\; 1 \leq i < j \leq k \right\}.$$

† Experts will now recognize the standard Cauchy criterion for convergence of sequences.

‡ Adding in a finite number of points does not alter boundedness (although it can alter the overall bounds) and so this also implies that the entire sequence $\{ f_n \mid n \in \mathbb{N} \}$ is bounded.

Then (11.14), with $\widetilde{r}, \widetilde{M}$ replacing r, M, implies that $f_n = f_{\widetilde{M}}$ for all $n \geq \widetilde{M}$. Thus, setting $f(0) = f_{\widetilde{M}}$ defines a function that is constant on the subset $\{0\} \cup \{\, 1/n \,|\, n \geq \widetilde{M}\,\}$, and hence continuous.

Otherwise, the bounded infinite set defined by the subsequence (11.15) contains infinitely many points. Thus, according to the Limit Point Theorem 5.13, it has at least one limit point, say ℓ. We claim that the extended function $f\colon A \to \mathbb{R}$ given by

$$f(x) = \begin{cases} f_n = \widetilde{f}(1/n), & x = 1/n \in \widetilde{A}, \\ \ell, & x = 0, \end{cases} \tag{11.16}$$

is continuous. This will establish the result, demonstrating that

$$\lim_{n \to \infty} f_n = \ell.$$

Thus, given $r > 0$, the limit point criterion implies that there exists $m \geq M$ such that $f_m \in (\,\ell - r, \ell + r\,) \setminus \{\,\ell\,\}$. But then

$$|\, f_n - \ell \,| \leq |\, f_n - f_m \,| + |\, f_m - \ell \,| < 2\,r \quad \text{for all} \quad n \geq M.$$

We conclude that

$$f^{-1}(\ell - 2\,r, \ell + 2\,r\,) \supset \{\,0\,\} \cup \{\, 1/n \,|\; M \leq n \in \mathbb{N}\,\}\,. \tag{11.17}$$

The right hand side is an open subset of A; the left hand side differs from it by a subset of $\widetilde{A}$, and hence is also open in A. Since this holds for all $r > 0$, we deduce the continuity of the extended function (11.16).

In either case, once the continuity of f is proved, the final statement of the theorem is then easily established. *Q.E.D.*

One can extend the Cauchy criterion to treat removable discontinuities of continuous functions. Given $a \in \mathbb{R}$, consider the subset

$$S_a = \{\,(a,a)\,\} \cup W_a, \qquad \text{where} \qquad \begin{aligned} W_a &= \{\,(x,y) \,|\; (x-a)\,(y-a) \neq 0\,\} \\ &= \mathbb{R}^2 \setminus \big(\{\,x = a\,\} \cup \{\,y = a\,\}\big), \end{aligned} \tag{11.18}$$

is the open subset obtained by deleting the horizontal and vertical lines passing through (a,a). The basic open subsets of S_a consist of all open rectangles that do not contain (a,a) along with sets of the form $\{\,(a,a)\,\} \cup (R \cap W_a)$ where $R \subset \mathbb{R}^2$ is an open rectangle containing (a,a). A continuous function $f\colon \mathbb{R} \to \mathbb{R}$ defines a continuous function $G\colon \mathbb{R}^2 \to \mathbb{R}$ given by $G(x,y) = f(x) - f(y)$, which remains continuous when restricted to S_a. The following continuous *Cauchy criterion* allows us to test for removable discontinuities in a function without knowing what the limiting value at the discontinuity is.

Theorem 11.6. *Suppose* $\widetilde{f}\colon \mathbb{R} \setminus \{\,a\,\} \to \mathbb{R}$ *is continuous. Define* $\widetilde{G}\colon S_a \to \mathbb{R}$*, where* S_a *is given in* (11.18)*, by*

$$\widetilde{G}(x,y) = \begin{cases} \widetilde{f}(x) - \widetilde{f}(y), & x \neq a \quad \text{and} \quad y \neq a, \\ 0 & x = y = a. \end{cases}$$

Then $\widetilde{f}$ has a removable discontinuity at $x = a$ if and only if $\widetilde{G}$ is continuous.

Proof: The statement that $\widetilde{f}$ has a removable discontinuity at $x = a$ means that there exists a continuous function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x) = \widetilde{f}(x)$ for all $x \neq a$. In this case, the function $G\colon \mathbb{R}^2 \to \mathbb{R}$ given by $G(x,y) = f(x) - f(y)$ is continuous. Moreover, $G(x,y) = \widetilde{G}(x,y)$ when $(x,y) \in S_a$, and hence $\widetilde{G}\colon S_a \to \mathbb{R}$ is continuous.

To prove the converse, let us assume that $a = 0$ without loss of generality, and thus the set (11.12) considered in the preceding example satisfies $\widehat{C} \subset S_0$. Let $\widehat{G}\colon \widehat{C} \to \mathbb{R}$ be the restriction of $\widetilde{G}$, so $\widehat{G}(x,y) = \widetilde{G}(x,y)$ whenever $(x,y) \in \widehat{C}$. Continuity of $\widetilde{G}$ implies continuity of $\widehat{G}$, which, by Theorem 11.5, implies that there is a continuous extension $f\colon A \to \mathbb{R}$ of the restriction of $\widetilde{f}$ to $\widetilde{A}$ with limiting value $\ell = \lim_{n\to\infty} \widetilde{f}(1/n) = f(0)$. We claim that the further extension

$$f(x) = \begin{cases} \widetilde{f}(x), & x \neq 0, \\ \ell, & x = 0, \end{cases}$$

is also continuous, which serves to prove the result.

Let $s > 0$. Continuity of $\widetilde{G}$ implies $\widetilde{G}^{-1}(-s,s) \subset S_0$ is open and contains the origin, and hence contains a subset of the form

$$W_0 \cap Q_r = \{ (0,0) \} \cup \{ (x,y) \mid 0 < |x|, |y| < r \} \subset \widetilde{G}^{-1}(-s,s),$$

where $Q_r = R^{r,r}_{-r,-r}$ is the square of half width $r > 0$ centered at the origin. In other words, if $0 < |x|, |y| < r$ then $|G(x,y)| = |f(x) - f(y)| < s$. Moreover, since $\ell = \lim_{n\to\infty} f(1/n)$, there exists $n \in \mathbb{N}$ such that $1/n < r$ and $|f(1/n) - \ell| < s$. But then, whenever $|x| < r$,

$$|f(x) - \ell| \leq |f(x) - f(1/n)| + |f(1/n) - \ell| < 2s$$

Thus, $(-r,r) \subset f^{-1}(\ell - 2s, \ell + 2s)$. The fact that this holds for all $0 < s \in \mathbb{R}$ proves that f is continuous. *Q.E.D.*

Let us end this section by combining the preceding two constructions, and formulate a Cauchy criterion for the uniform convergence of continuous functions. Recalling (11.2), (11.10), (11.12), consider the following Cartesian product subsets of $\mathbb{R}^3$:

$$\begin{aligned} D &= [a,b] \times C = [a,b] \times A \times A \\ &= \{ (x,y,z) \mid a \leq x \leq b, \ \ y = 0 \ \text{ or } \ 1/y \in \mathbb{N}, \ \ z = 0 \ \text{ or } \ 1/z \in \mathbb{N} \}, \\ \widetilde{D} &= [a,b] \times \widetilde{C} = [a,b] \times \widetilde{A} \times \widetilde{A} = \{ (x, 1/m, 1/n) \mid m,n \in \mathbb{N} \}, \\ \widehat{D} &= [a,b] \times \widehat{C} = [a,b] \times \big(\{ (0,0) \} \cup \widetilde{D} \big). \end{aligned} \tag{11.19}$$

We will leave the reader to investigate the nature and continuity of functions $F\colon D \to \mathbb{R}$ and fix our attention on $\widehat{D}$.

Let $\widetilde{F}\colon \widetilde{B} \to \mathbb{R}$ be a continuous function on $\widetilde{B} = \{ (x, 1/n) \mid a \leq x \leq b, \ n \in \mathbb{N} \}$ corresponding to a sequence of continuous functions $f_n(x) = \widetilde{F}(x, 1/n)$ for $n \in \mathbb{N}$ as before. Consider the function $\widehat{G}: \widehat{D} \longrightarrow \mathbb{R}$ defined by

$$\begin{aligned} \widehat{G}\left(x, \frac{1}{m}, \frac{1}{n}\right) &= \widetilde{F}\left(x, \frac{1}{m}\right) - \widetilde{F}\left(x, \frac{1}{n}\right) = f_m(x) - f_n(x), \qquad && m, n \in \mathbb{N}, \\ \widehat{G}(x, 0, 0) &= 0, && a \leq x \leq b. \end{aligned} \tag{11.20}$$

As in Theorem 11.5, the *Cauchy criterion* for the uniform convergence of the sequence of functions $f_n(x)$ is the continuity of the associated function $\widehat{G}$.

Theorem 11.7. *Given continuous functions $f_n: [a, b] \to \mathbb{R}$ for $n \in \mathbb{N}$, the function $\widehat{G}: \widehat{D} \to \mathbb{R}$ defined in* (11.20) *is continuous if and only if the sequence $f_n(x)$ converges uniformly to a continuous limit function $f_0(x) = \lim_{n \to \infty} f_n(x)$.*

Proof: We will establish the direct statement, leaving the converse to the reader. Therefore, we assume the functions f_n and $\widehat{G}$ are continuous, and prove that the sequence of functions f_n converge uniformly to a continuous limit f_0.

The first remark is that if we fix $x \in [a, b]$, then the partial function $G_x(y, z) = \widehat{G}(x, y, z)$ is a continuous function $G_x: \widehat{C} \to \mathbb{R}$, thus satisfying the Cauchy criterion contained in Definition 11.4. Since $G_x(1/m, 1/n) = f_m(x) - f_n(x)$, Theorem 11.5 implies that the sequence $f_n(x)$ converges to a limit $f_0(x)$ as $n \to \infty$, again for each fixed $x \in [a, b]$. In this way, we easily deduce pointwise convergence of the functions.

Continuity of $\widehat{G}$ requires that if $r > 0$, the inverse image of the open interval $(-r, r)$, namely

$$\widehat{G}^{-1}(-r, r) = \{ (x, 0, 0) \mid a \leq x \leq b \} \cup \left\{ \left(x, \frac{1}{m}, \frac{1}{n}\right) \;\middle|\; a \leq x \leq b, \ | f_m(x) - f_n(x) | < r \right\},$$

is open in $\widehat{C}$. Adapting Lemma 9.14 to the current three-dimensional situation, we see that, because the open set contains the compact line segment $\{ (x, 0, 0) \mid a \leq x \leq b \}$, it necessarily contains a solid rectangular strip of width $2t > 0$ surrounding it, of the form

$$S_t = \{ (x, 1/m, 1/n) \mid \ a \leq x \leq b, \ \ 1/m, \ 1/n < t \} \subset \widehat{G}^{-1}(-r, r).$$

Choosing $1/t \leq M \in \mathbb{N}$, this is equivalent to the Cauchy criterion

$$| f_m(x) - f_n(x) | < r \quad \text{whenever} \quad m, n \geq M \quad \text{for all} \quad a \leq x \leq b. \tag{11.21}$$

Now, fix m and x, and let $n \to \infty$ in (11.21). In view of (6.14) and the pointwise convergence of $f_n(x)$ to $f_0(x)$, this implies

$$| f_m(x) - f_0(x) | \leq r \quad \text{whenever} \quad m \geq M \quad \text{for all} \quad a \leq x \leq b,$$

which is equivalent to the uniform convegence criterion (11.7), thus completing the proof of the direct statement. The proof of the converse is left to the reader.

In fact, we have shown that $\widetilde{F}: \widetilde{B} \to \mathbb{R}$ admits a continuous extension $F: B \to \mathbb{R}$, so that $F(x, 1/n) = \widetilde{F}(x, 1/n)$ for $n \in \mathbb{N}$ and $F(x, 0) = f_0(x)$. This in turn produces a continuous extension $G: C \to \mathbb{R}$ of $\widehat{G}$ with $G(x, y, z) = F(x, y) - F(x, z)$ for all $(x, y, z) \in C$. *Q.E.D.*

12. Differentiating Functions of Two Variables.

The next step in the program is to define the derivative of a function of two variables and investigate its properties. In this case, we can no longer form a single difference quotient to define its derivative. Instead, we introduce an analog of the Carathéodory formulation (7.4).

Definition 12.1. Let $F: \mathbb{R}^2 \to \mathbb{R}$ be continuous. We say that F is *differentiable* at the point $(a,b) \in \mathbb{R}^2$ if there exist continuous functions $Q, R: \mathbb{R}^2 \to \mathbb{R}$ such that

$$F(x,y) = F(a,b) + Q(x,y)\,(x-a) + R(x,y)\,(y-b). \tag{12.1}$$

The functions $Q(x,y), R(x,y)$ are *not* uniquely determined by the equation (12.1). Indeed, if $S(x,y)$ is any continuous function then

$$\widetilde{Q}(x,y) = Q(x,y) + S(x,y)\,(y-b), \qquad \widetilde{R}(x,y) = R(x,y) - S(x,y)\,(x-a),$$

will also satisfy (12.1). Nevertheless, as we will soon see, their values $Q(a,b)$ and $R(a,b)$ are uniquely determined, and serve to specify the two components of the *derivative* $F'(a,b)$. More specifically, these values are said to define the *partial derivatives* of the differentiable function F at the point (a,b), and we will use the following subscript notation:

$$F_x(a,b) = Q(a,b), \qquad F_y(a,b) = R(a,b).$$

The reason for the term "partial derivative" is partly because they each form "part" of the derivative, which is a pair of real numbers or, equivalently, a vector commonly known as the *gradient* of F, and often denoted by the symbol

$$\nabla F(a,b) = F'(a,b) = \big(F_x(a,b), F_y(a,b) \big). \tag{12.2}$$

Another reason for the name is that the partial derivatives are obtained by differentiating the partial functions associated with $F(x,y)$. Namely if we, say, fix $y = b$ then the result is a scalar function $f: \mathbb{R} \to \mathbb{R}$ given by $f(x) = F(x,b)$, which, according to Proposition 10.3, is continuous. Setting $y = b$ in (12.1) produces

$$F(x,b) = F(a,b) + Q(x,b)\,(x-a)$$

or, equivalently,

$$f(x) = f(a) + (x-a)\,q(x) \qquad \text{where} \qquad q(x) = Q(x,b).$$

This is the Carathéodory formula (7.4) for the function $f(x)$, and hence

$$q(a) = Q(a,b) = F_x(a,b) = f'(a),$$

which in turn implies $F_x(a,b) = f'(a)$ is uniquely determined. Similarly, fixing $x = a$ leads to the partial function $g(y) = F(a,y)$, and now (12.1) reduces to

$$F(a,y) = F(a,b) + R(a,y)\,(y-b)$$

or, equivalently,

$$g(y) = g(b) + (y-b)\,r(y) \qquad \text{where} \qquad r(y) = R(a,y),$$

and hence

$$r(b) = R(a,b) = F_y(a,b) = g'(b).$$

Example 12.2. Consider the function

$$F(x,y) = x^2 y^3 + 5\,y^2 + 2\,x - 1.$$

We easily compute its partial derivatives by just regarding one variable as a fixed constant and differentiating with respect to the other using all the usual rules of differentiation for a function of the single variable. Thus, the x and y partial derivatives are, respectively,

$$F_x(x,y) = 2\,x\,y^3 + 2, \qquad F_y(x,y) = 3\,x^2 y^2 + 10\,y.$$

Differentiability of F at the origin $(x,y) = (0,0)$ follows from the identity

$$F(x,y) = -1 + x\,Q(x,y) + y\,R(x,y),$$

where, for instance,

$$Q(x,y) = x\,y^3 + 3\,y^2 + 2, \qquad R(x,y) = 3\,y.$$

As above, Q and R are not uniquely determined. However, their values at the origin are, with $F_x(0,0) = Q(0,0) = 2$, $F_y(0,0) = R(0,0) = 0$, in accordance with the preceding calculation. Differentiability at a general point (a,b) follows similarly using a slightly more complicated algebraic manipulation. However, one can use the general rules for differentiation to verify the differentiability of the polynomial function $F(x,y)$ without having to perform the explicit manipulations.

Example 12.3. *Warning*: The existence of partial derivatives does not necessarily imply that the function is differentiable. For example, consider the function

$$F(x,y) = \begin{cases} \dfrac{x\,y}{x^2 + y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0), \end{cases}$$

from Example 10.4. Since

$$f(x) = F(x,0) \equiv 0, \qquad g(y) = F(0,y) \equiv 0,$$

we clearly have

$$f'(0) = 0, \qquad g'(0) = 0,$$

and hence both partial derivatives of F at $(0,0)$ exist and are equal to $0 = F_x(0,0) = F_y(0,0)$. But the function $F(x,y)$ is not continuous, and hence there are no continuous functions Q, R such that (12.1) holds. Thus, despite possessing both partial derivatives, F is *not* differentiable at the origin.

Algebraic combinations, including sums and differences, products, quotients (when the denominator doesn't vanish), maxima, and minima of differentiable functions are differentiable. We leave their verification to the interested reader. Furthermore, since partial derivatives can be computed by taking an ordinary derivative while fixing one of the two variables, they obey all the algebraic rules we established in Section 7. In particular, the partial derivative of the sum and the difference of functions is the sum and difference of their partial derivatives. The product rule for the partial derivatives of

$$H(x,y) = F(x,y)\,G(x,y)$$

takes the form

$$\begin{aligned} H_x(a,b) &= F_x(a,b)\,G(a,b) + F(a,b)\,G_x(a,b),\\ H_y(a,b) &= F_y(a,b)\,G(a,b) + F(a,b)\,G_y(a,b). \end{aligned}$$

As for the reciprocal $R(x,y) = 1/F(x,y)$ provided $F(x,y) \neq 0$, we can differentiate the product

$$1 = F(x,y)\,R(x,y),$$

producing

$$\begin{aligned} 0 &= F_x(a,b)\,R(a,b) + F(a,b)\,R_x(a,b) = \frac{F_x(a,b)}{F(a,b)} + F(a,b)\,R_x(a,b),\\ 0 &= F_y(a,b)\,R(a,b) + F(a,b)\,R_y(a,b) = \frac{F_y(a,b)}{F(a,b)} + F(a,b)\,R_y(a,b), \end{aligned}$$

and hence

$$R_x(a,b) = -\,\frac{F_x(a,b)}{F(a,b)^2}, \qquad R_y(a,b) = -\,\frac{F_y(a,b)}{F(a,b)^2}.$$

We leave it to the reader to write out the quotient rule, as well as fill in the omitted proofs.

As for the Chain Rule, there are several versions depending upon which type of functions are being composed. Consider first the case $h(t) = F(f(t), g(t))$ where $f, g\colon \mathbb{R} \to \mathbb{R}$ are differentiable at $t = a$, while $F\colon \mathbb{R}^2 \to \mathbb{R}$ is differentiable at $(x,y) = (f(a), g(a))$. Then the composite function $h(t)$ is differentiable at $t = a$, and

$$h'(a) = F_x(f(a), g(a))\,f'(a) + F_y(f(a), g(a))\,g'(a). \tag{12.3}$$

To prove this, we use (12.1) and the equations

$$f(t) = f(a) + (t-a)\,q(t), \quad g(t) = g(a) + (t-a)\,r(t), \quad \text{where} \quad f'(a) = q(a), \quad g'(a) = r(a),$$

to write

$$\begin{aligned} h(t) &= F(f(t), g(t))\\ &= F(f(a), g(a)) + Q(f(t), g(t))\big[\,f(t) - f(a)\,\big] + R(f(t), g(t))\big[\,g(t) - g(a)\,\big]\\ &= h(a) + (t-a)\,\big[\,Q(f(t), g(t))\,q(t) + R(f(t), g(t))\,r(t)\,\big]. \end{aligned}$$

The function multiplying $t - a$ is continuous, and hence

$$h'(a) = Q(a,b)\,q(a) + R(a,b)\,r(a) = F_x(f(a), g(a))\,f'(a) + F_y(f(a), g(a))\,g'(a).$$

The other two versions of the chain rule are proved similarly. The composition $G(x,y) = f(F(x,y))$ at the point (a,b) has derivatives

$$G_x(a,b) = f'(F(a,b))\,F_x(a,b), \qquad G_y(a,b) = f'(F(a,b))\,F_y(a,b). \tag{12.4}$$

Finally, if $\mathbf{G}: \mathbb{R}^2 \to \mathbb{R}^2$ is given by $\mathbf{G}(x,y) = \big(I(x,y), J(x,y) \big)$, then the partial derivatives of their composition

$$K(x,y) = F \circ \mathbf{G}(x,y) = F\big(I(x,y), J(x,y) \big)$$

at the point (a,b) are

$$\begin{aligned} K_x(a,b) &= F_x(a,b)\, I_x(a,b) + F_y(a,b)\, J_x(a,b), \\ K_y(a,b) &= F_x(a,b)\, I_y(a,b) + F_y(a,b)\, J_y(a,b). \end{aligned} \tag{12.5}$$

The formulas (12.3–5) can all be subsumed under a general Chain Rule formula involving Jacobian matrices for the derivative of all possible compositions, [**1**, **16**].

As we discovered, the existence of partial derivatives is not sufficient to ensure differentiability of the function. However, if they are continuous, then this does suffice.

Theorem 12.4. *Let $F: \mathbb{R}^2 \to \mathbb{R}$ be continuous and assume both partial derivatives $F_x, F_y: \mathbb{R}^2 \to \mathbb{R}$ are also continuous. Then F is differentiable.*

Proof: Given $(a,b) \in \mathbb{R}^2$, let us first write

$$F(x,y) = F(a,y) + (x-a)\, P(x,y). \tag{12.6}$$

Since, for any fixed y, we are assuming $f_y(x) = F(x,y)$ has a continuous derivative, namely the x partial derivative $F_x(x,y) = f_y'(x)$, in view of the Definition 7.3 of the one-dimensional derivative, we can write

$$P(x,y) = \begin{cases} \dfrac{F(x,y) - F(a,y)}{x-a}, & x \neq a, \\ F_x(a,y), & x = a, \end{cases} \tag{12.7}$$

where, for each fixed y, the partial function $p_y(x) = P(x,y)$ is continuous as a function of x. We claim that $P: \mathbb{R}^2 \to \mathbb{R}$ is continuous as a function of both variables. First, in view of the continuity of sums and quotients, its restriction $\widetilde{P}: \mathbb{R}^2 \setminus \{x = a\} \to \mathbb{R}$ is continuous, and so we only need worry about points where $x = a$.

Let $S \subset \mathbb{R}$ be open. To prove that $P^{-1}(S) \subset \mathbb{R}^2$ is open, given any $(x_0, y_0) \in P^{-1}(S)$ we need to find an open set U such that $(x_0, y_0) \in U \subset P^{-1}(S)$. If $x_0 \neq a$, then we have $(x_0, y_0) \in \widetilde{P}^{-1}(S) \subset P^{-1}(S)$ and, by continuity of the restriction $\widetilde{P}$, the subset $U = \widetilde{P}^{-1}(S)$ is open. Now suppose $x_0 = a$. Then $P(x_0, y_0) = P(a, y_0) = F_x(a, y_0) \in S$, and hence $(x_0, y_0) \in F_x^{-1}(S)$, which is open by the assumed continuity of the partial derivative, although it is not necessarily contained in $P^{-1}(S)$. However, let R be an open rectangle with $(x_0, y_0) \in R \subset F_x^{-1}(S)$. Continuity of $F_x(x,y)$ implies that we can apply the Mean Value Theorem 7.13, and write

$$F(x,y) = F(a,y) + (x-a)\, F_x(z,y)$$

for some point z, depending on a, x, y, lying in the interval between a and x. Thus, we can re-express the difference quotient function (12.7) as simply

$$P(x,y) = F_x(z,y) \tag{12.8}$$

where, when $x = a$, we set $z = a$ to match the formula (12.7) for P there. Moreover, because z lies between a and x, if $(x,y) \in R$, so is $(z,y) \in R \subset F_x^{-1}(S)$, and hence $F_x(z,y) \in S$ which, by (12.8), implies $P(x,y) \in S$. Since this holds for all $(x,y) \in R$, we conclude that $R \subset P^{-1}(S)$, thereby completing the proof that the latter set is open.

Finally, we use the differentiability of the partial function $g(y) = F(a,y)$ at $y = b$ to write

$$F(a,y) = F(a,b) + (y-b)\,q(y), \tag{12.9}$$

where $q(y)$ is continuous with $q(b) = F_y(a,b)$. Substituting (12.9) into (12.6) produces the required formula:

$$F(x,y) = F(a,y) + (x-a)\,P(x,y) + (y-b)\,Q(x,y),$$

where $P(x,y)$, constructed earlier, and $Q(x,y) = q(y)$ are both continuous, thus establishing the differentiability of F at (a,b). *Q.E.D.*

Note that according to the proof, one only needs to assume continuity of the partial derivatives nearby the point (a,b) in order to conclude differentiability of the function there. In other words, differentiability is a local property and does not depend on the function's overall global behavior. Interestingly, we also only required continuity of one of the partial derivatives and existence of the other at the point to establish the conclusion. (We used continuity of the x partial derivative, but could easily reproduces the proof using continuity of the y partial derivative instead.)

We can proceed on to higher order partial derivatives by playing the same game as in the univariate case. Starting with (12.1), suppose both $Q(x,y)$ and $R(x,y)$ are themselves differentiable at (a,b), so that

$$\begin{aligned} Q(x,y) &= Q(a,b) + S(x,y)\,(x-a) + T(x,y)\,(y-b), \\ R(x,y) &= R(a,b) + V(x,y)\,(x-a) + W(x,y)\,(y-b), \end{aligned} \tag{12.10}$$

for S, T, V, W continuous, with

$$Q_x(a,b) = S(a,b), \qquad Q_y(a,b) = T(a,b), \qquad R_x(a,b) = V(a,b), \qquad R_y(a,b) = W(a,b).$$

On the other hand, differentiating (12.1) and substituting using (12.10), we find

$$\begin{aligned} F_x(x,y) &= Q(x,y) + Q_x(x,y)\,(x-a) + R_x(x,y)\,(y-b) \\ &= Q(a,b) + \big[\,S(x,y) + Q_x(x,y)\,\big]\,(x-a) + \big[\,T(x,y) + R_x(x,y)\,\big]\,(y-b), \\ F_y(x,y) &= R(x,y) + Q_y(x,y)\,(x-a) + R_y(x,y)\,(y-b) \\ &= R(a,b) + \big[\,V(x,y) + Q_y(x,y)\,\big]\,(x-a) + \big[\,W(x,y) + R_y(x,y)\,\big]\,(y-b), \end{aligned}$$

and hence

$$\begin{aligned} F_{xx}(a,b) &= S(a,b) + Q_x(a,b) = 2\,S(a,b), \\ F_{xy}(a,b) &= T(a,b) + R_x(a,b) = T(a,b) + V(a,b), \\ F_{yx}(a,b) &= V(a,b) + Q_y(a,b) = T(a,b) + V(a,b), \\ F_{yy}(a,b) &= W(a,b) + R_y(a,b) = 2\,W(a,b). \end{aligned}$$

The second expression is the result of first applying the x partial derivative to F and then applying the y partial derivative to F_x, whereas the third expression is in the opposite order: first apply the y partial derivative to F and then apply the x partial derivative to F_y. Interestingly, their right hand sides are exactly the same. Combining (12.1) and (12.10), we deduce

$$\begin{aligned} F(x,y) = F(a,b) &+ F_x(a,b)\,(x-a) + F_y(a,b)\,(y-b) \\ &+ S(x,y)\,(x-a)^2 + U(x,y)\,(x-a)\,(y-b) + W(x,y)\,(y-b)^2, \end{aligned} \tag{12.11}$$

where $U(x,y) = T(x,y) + V(x,y)$. If (12.11) holds where S, U, W are continuous, we say that F is *twice differentiable* at (a,b). As before, these functions are not uniquely determined; however, their values at the point (a,b) are, and equate to the second order partial derivatives of F there:

$$F_{xx}(a,b) = 2\,S(a,b), \qquad F_{xy}(a,b) = F_{yx}(a,b) = U(a,b), \qquad F_{yy}(a,b) = 2\,W(a,b). \tag{12.12}$$

The fact the that the two "mixed derivatives" have the same formula is an important result, known as the Equality of Mixed Partials.

Theorem 12.5. *Suppose $F(x,y)$ is twice differentiable at the point (a,b). Then*

$$F_{xy}(a,b) = F_{yx}(a,b). \tag{12.13}$$

It is not hard to show that algebraic combinations of twice differentiable functions remain twice differentiable. Thus, all polynomial functions are twice differentiable as are all rational functions as long as the denominator doesn't vanish. In analogy with Theorem 12.4 we also have.

Theorem 12.6. *If all second order partial derivatives of F are continuous, then F is twice differentiable.*

The proof strategy is similar, using the order $n = 1$ version of the Taylor expansion (8.7) coupled with the remainder formula (8.8). Details are left to the reader.

Example 12.7. Returning to the function

$$F(x,y) = x^2y^3 + 5\,y^2 + 2\,x - 1$$

considered in Example 12.2, with

$$F_x(x,y) = 2\,x\,y^3 + 2, \qquad F_y(x,y) = 3\,x^2y^2 + 10\,y,$$

we readily compute its second order partial derivatives by differentiating with respect to one of the variables whilst holding the other variable fixed:

$$F_{xx}(x,y) = 2\,y^3, \qquad F_{xy}(x,y) = F_{yx}(x,y) = 6\,x\,y^2, \qquad F_{yy}(x,y) = 6\,x^2y + 10,$$

where the two mixed partial derivatives are equal because F is twice differentiable.

Example 12.8. Consider the function

$$F(x,y) = \begin{cases} x\,y\,\dfrac{x^2-y^2}{x^2+y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0). \end{cases}$$

A direct computation shows that

$$F_x(x,y) = \frac{y\,(x^4+4\,x^2y^2-y^4)}{(x^2+y^2)^2}, \qquad F_y(x,y) = \frac{x\,(x^4-4\,x^2y^2-y^4)}{(x^2+y^2)^2}, \qquad (x,y)\neq(0,0).$$

On the other hand,

$$F(x,y) = x\,Q(x,y) \qquad \text{where} \qquad Q(x,y) = \begin{cases} y\,\dfrac{x^2-y^2}{x^2+y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0). \end{cases}$$

It can be shown that Q is continuous — the only problematic point is the origin — and hence $F_x(0,0) = 0$. Now, set

$$g(y) = F_x(0,y) = -y, \qquad \text{and hence} \qquad g'(0) = F_{xy}(0,0) = -1, \qquad \text{while}$$
$$h(x) = F_y(x,0) = x, \qquad \text{and hence} \qquad h'(0) = F_{yx}(0,0) = +1.$$

Thus, both mixed partial derivatives of F exist at the origin, but are *not* equal. This is because there are no continuous functions S, U, W such that (12.11) holds, which in this case would be

$$F(x,y) = x^2\,S(x,y) + x\,y\,U(x,y) + y^2\,W(x,y).$$

For instance, if we choose $S = W = 0$, the function

$$U(x,y) = \begin{cases} \dfrac{x^2-y^2}{x^2+y^2}, & (x,y) \neq (0,0), \\ 0, & (x,y) = (0,0), \end{cases}$$

satisfies the preceding equation, but is not continuous. Indeed, its restrictions to the x and y axes

$$U(x,0) = \begin{cases} 1, & x \neq 0, \\ 0, & x = 0, \end{cases} \qquad U(0,y) = \begin{cases} -1, & y \neq 0, \\ 0, & y = 0, \end{cases}$$

are not continuous, in violation of Proposition 10.3.

By analogy with the one-dimensional version, we will call equation (12.11) the first order Taylor expansion of the function $F(x,y)$. The general version of Taylor's Theorem for a function of two variables can be formulated as follows. We use the notation F_{ij} for non-negative integers $i, j \geq 0$ to denote the partial derivative of order $k = i + j$ obtained by differentiating i times with respect to x and j times with respect to y. For example, $F_{10} = F_x$, $F_{20} = F_{xx}$, $F_{11} = F_{xy}$, $F_{23} = F_{xxyyy}$, and so on. By generalizing Theorem 12.5, a consequence of differentiability will be the fact that it does not matter in which order the x and y derivatives are performed.

Given $n \geq 0$, we assume that the partial derivatives of order $i+j \leq n$ exist at the point (a,b), and define the *Taylor polynomial* of degree n for F there as

$$P_n(x,y) = \sum_{0 \leq i+j \leq n} F_{ij}(a,b) \, \frac{(x-a)^i \, (y-b)^j}{i! \, j!}. \tag{12.14}$$

We say that $F(x,y)$ is order $(n+1)$-*differentiable* at (a,b) if there exist continuous functions $R_{ij}(x,y)$ for all non-negative integers $0 \leq i,j \leq n+1$ satisfying $i+j = n+1$, such that

$$F(x,y) = P_n(x,y) + \sum_{i+j=n+1} R_{ij}(x,y) \, \frac{(x-a)^i \, (y-b)^j}{i! \, j!}, \tag{12.15}$$

in which case their values at the point equal the corresponding $(n+1)^{\text{st}}$ order partial derivative of F there: $R_{ij}(a,b) = F_{ij}(a,b)$. Moreover, there exists a point (z,w) in the open rectangle with opposite corners (a,b) and (x,y) such that

$$R_{ij}(x,y) = F_{ij}(z,w), \qquad i+j = n+1. \tag{12.16}$$

Finally, by a similar argument as in the order 0 and 1 cases, we can prove that if all the $(n+1)^{\text{st}}$ partial derivatives of F are continuous, then we deduce that $F(x,y)$ is order $(n+1)$-differentiable.

13. Area.

We now turn to the problem that was the first immediately successful application of Newton's calculus — computing the area of regions in the plane. The areas of simple shapes are well known from elementary geometry, but more complicated regions with curved boundaries require the full power of calculus – in particular the Fundamental Theorem 15.3.

Let us begin with the simplest possible region: a rectangle.

Definition 13.1. Let $a \leq b$ and $c \leq d$ be real numbers. The *area* of the bounded closed rectangle

$$\overline{R}{}^{b,d}_{a,c} = \{\, (x,y) \mid \ a \leq x \leq b, \ \ c \leq y \leq d \,\}, \qquad \text{is} \qquad \text{area}\, \overline{R}{}^{b,d}_{a,c} = (d-c)\,(b-a). \tag{13.1}$$

In other words, the area of a rectangle is the product of its length $b-a$ and width $d-c$. (We will adopt the convention that distances in the x direction are called *length* and those in the y direction *width*.) In particular, if $a = b$ and/or $c = d$, the rectangle degenerates into a line segment or a single point, both of which have 0 area.

We will take rectangles as our basic building blocks, and use them to define the area of general subsets $D \subset \mathbb{R}^2$. First, we postulate that the area of a disjoint union of rectangles is the sum of their individual areas:

$$\text{area} \left(\bigcup_{i=1}^n R_i \right) = \sum_{i=1}^n \text{area}\, R_i \qquad \text{provided} \qquad R_i \cap R_j = \varnothing \quad \text{for all} \quad i \neq j. \tag{13.2}$$

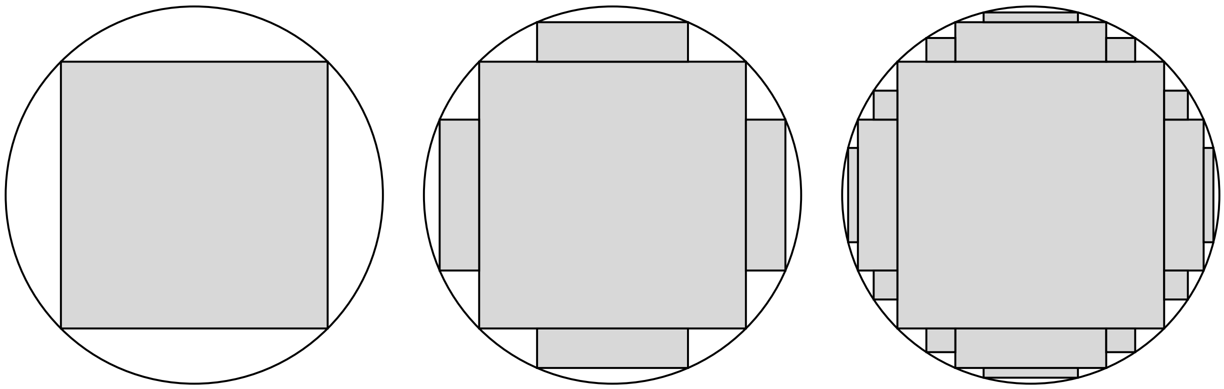


Figure 19. The Method of Exhaustion for a Disk.

We can relax this requirement by allowing the rectangles to intersect along their boundaries, which have zero area. We thus call a collection of closed rectangles $R_1, \ldots, R_n$ *essentially disjoint* if the intersection of any pair $R_i \cap R_j$ has zero area, keeping in mind that the empty set also has zero area. The formula (13.2) continues to hold when the collection is essentially disjoint.

A fundamental property is that a subset necessarily has less area than any enclosing set:

$$\text{if} \quad D \subset E \quad \text{then} \quad \text{area}\, D \leq \text{area}\, E. \tag{13.3}$$

Thus, given a set $D \subset \mathbb{R}^2$, let $R_1, \ldots, R_n$ be an essentially disjoint collection of closed rectangles such that each $R_i \subset D$. Their union, denoted by R^*, is a subset of D, and hence

$$\text{area}\, D \geq \text{area}\, R^* = \text{area}\left(\bigcup_{i=1}^{n} R_i\right) = \sum_{i=1}^{n} \text{area}\, R_i, \qquad \text{where} \qquad R^* = \bigcup_{i=1}^{n} R_i. \tag{13.4}$$

Note that if $D \neq \varnothing$, then there exist such collections since, at the very least, we could take $R_i = \overline{R}_{a_i,a_i}^{b_i,b_i} = \{(a_i, b_i)\}$, where $\{(a_1, b_1), \ldots, (a_n, b_n)\}$ is any finite set of points in D for some $n \in \mathbb{N}$.

The *Method of Exhaustion*, dating back to Archimedes, relies on the intuition that, as we increase the number of rectangles so that they fill out (exhaust) more and more of their enclosing region, their total area forms a better and better approximation to its area; an example is sketched in Figure 19. Archimedes' idea inspires our general definition of the area of a planar region.

Definition 13.2. Let $D \subset \mathbb{R}^2$. Given any collection of essentially disjoint closed rectangles $R_1, \ldots, R_n \subset D$, let area R^* denote the area of their union, as in (13.4). Let

$A_D = \{ \text{area } R^* \} \subset [0, \infty)$ be the set of all such areas. We define the *area* of D to be

$$\text{area } D = \begin{cases} 0, & D = \varnothing, \\ \sup A_D, & A_D \text{ is bounded}, \\ \infty, & A_D \text{ is unbounded}. \end{cases} \tag{13.5}$$

In other words, the area of a nonempty subset $D \neq \varnothing$ is the least upper bound of the total areas of any collection of subrectangles contained therein,

$$\text{area } D = \sup \left\{ \sum_{i=1}^n \text{area } R_i \;\middle|\; \begin{array}{l} n \in \mathbb{N}, \quad R_1, \ldots, R_n \subset D \\ \qquad \text{are essentially disjoint closed rectangles} \end{array} \right\}, \tag{13.6}$$

thereby quantifying Archimedes' Method of Exhaustion. There are other ways of "exhausting the area" of a domain, e.g., by using triangles or (regular) polygons. But to avoid unnecessary extra complications in our constructions, we will restrict our attention to rectangles, in that they are, at least from the standpoint of area, the "simplest" planar regions.

Remark: There are more sophisticated definitions of area, such as Lebesgue measure, [**24**], that are better suited to advanced mathematical analysis and may produce different values for the area (measure) of more pathological subsets such as fractals, [**15**]. However, for our purposes this naïve but intuitive definition suffices.

Before proceeding, our first task is to check that the preceding definition of area is consistent, meaning that it gives the same answer when D is itself a closed rectangle. In other words, we need to show that if $R = \overline{R}_{a,c}^{b,d}$ has area given by the original formula (13.1) and $R_1, \ldots, R_n$ are any collection of essentially disjoint closed rectangles with each $R_i \subset R$, then

$$\text{area } R^* = \sum_{i=1}^n \text{area } R_i \leq \text{area } R. \tag{13.7}$$

Since R itself forms a possible collection, this implies consistency: area $R = \sup A_R$.

First suppose we chop up (or *subdivide*) R into disjoint open rectangles by slicing it with a finite collection of horizontal and vertical lines, as illustrated in the first plot in Figure 20. In other words, we select numbers

$$a = x_0 < x_1 < x_2 < \cdots < x_{k-1} < x_k = b, \qquad c = y_0 < y_1 < y_2 < \cdots < y_{l-1} < y_l = d,$$

which we will refer to as *partitions* of the intervals $[a, b]$ and $[c, d]$, and consider the $n = k\,l$ closed rectangles

$$R_{i,j} = \overline{R}_{x_{i-1}, y_{j-1}}^{x_i, y_j} \subset R \qquad \text{for} \qquad \begin{array}{l} i = 1, \ldots, k, \\ j = 1, \ldots, l, \end{array} \qquad \text{such that} \qquad R = \bigcup_{i=1}^k \bigcup_{j=1}^l R_{i,j},$$

which we call a *partition* of R. Their boundaries are all contained in the lines $\{ x = x_i \}$ and $\{ y = y_j \}$, which have zero area, and hence the $R_{i,j}$ are essentially disjoint. Their

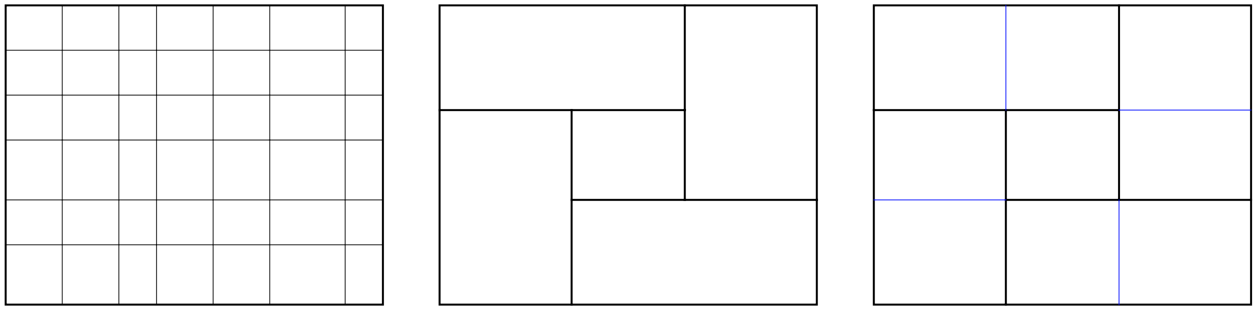


Figure 20. Subdividing a Rectangle.

combined areas are

$$\begin{aligned}\text{area}\left(\bigcup_{i=1}^{k}\bigcup_{j=1}^{l} R_{i,j}\right) &= \sum_{i=1}^{k}\sum_{j=1}^{l} \text{area } R_{i,j} = \sum_{i=1}^{k}\sum_{j=1}^{l} (x_i - x_{i-1})\,(y_j - y_{j-1}) \\ &= \left[\sum_{i=1}^{k} (x_i - x_{i-1})\right]\left[\sum_{j=1}^{l} (y_j - y_{j-1})\right] \\ &= (x_k - x_0)\,(y_l - y_0) = (d-c)\,(b-a) = \text{area } R.\end{aligned}$$

Thus, not surprisingly, the computation of area is consistent when we subdivide a rectangle in this manner.

More generally, suppose $S_m = \overline{R}^{b_m,d_m}_{a_m,c_m}$ for $m = 1, \ldots, n$ are an essentially disjoint set of closed rectangles with each $S_m \subset R$, so that $a \le a_m < b_m \le b$ and $c \le c_m < d_m \le d$. An example appears in the middle plot in Figure 20. We can further subdivide the S_m into a system of rectangles of the form in the previous paragraph; the process is illustrated in the last plot in Figure 20, in which each of the larger rectangles in the second plot has been cut in half. For this purpose, we let

$$\begin{aligned} X &= \{\,a, a_1, \ldots, a_n, b_1, \ldots, b_n, b\,\} = \{x_0, \ldots, x_k\}, \\ Y &= \{\,c, c_1, \ldots, c_n, d_1, \ldots, d_n, d\,\} = \{y_0, \ldots, y_l\}, \end{aligned}$$

where the x_i's and y_j's are of the form in the preceding paragraph. Here, the set notation for X, Y implies that we order the a_m's and b_m's and eliminate any repeats to obtain the x_i's, and similarly for the c_m's and d_m's to obtain the y_j's. The corresponding horizontal and vertical lines subdivide not only R into the collection of $R_{i,j}$'s but also each S_m is itself subdivided by a subset of the $R_{i,j}$'s, and hence, by the preceding argument

$$\text{area } S_m = \sum_{R_{i,j} \subset S_m} \text{area } R_{i,j}. \tag{13.8}$$

Since they are mutually essentially disjoint, each $R_{i,j}$ is contained in at most one S_m, although some may not appear at all if we have omitted any part of R in our original collection $S_1, \ldots, S_n$. In this way,

$$\text{area } R = \sum_{i=1}^{k} \sum_{j=1}^{l} \text{area } R_{i,j} \geq \sum_{m=1}^{n} \text{area } S_m,$$

with equality if and only if every $R_{i,j}$ is contained in one of the S_m's. In other words, the area of a rectangle is the same no matter how we subdivide it into subrectangles. This thus establishes consistency of our definition of area.

Given the construction in the preceding paragraph, let $Z = \{Z_1, \ldots, Z_q\}$ be the set of all rectangles $R_{i,j}$ that are *not* contained in one of the S_m's. Observe that

$$Z^* = \bigcup_{p=1}^{q} Z_p = \overline{R \setminus S^*} \quad \text{is the closure of the complement to} \quad S^* = \bigcup_{m=1}^{n} S_m.$$

Moreover, since every $R_{i,j}$ appears in either S^* or in Z^*,

$$S^* \cup Z^* = R, \qquad \text{and} \qquad \text{area } S^* + \text{area } Z^* = \text{area } R. \tag{13.9}$$

We also note that an open rectangle

$$R_{a,c}^{b,d} = \{\, (x,y) \mid \ a < x < b, \ \ c < y < d \,\}$$

has the same area as the corresponding closed rectangle

$$\text{area } R_{a,c}^{b,d} = \text{area } \overline{R}_{a,c}^{b,d} = (b-a)\,(d-c). \tag{13.10}$$

Indeed, if $0 < r \leq \min\left\{\, \frac{1}{2}(b-a), \frac{1}{2}(d-c) \,\right\}$, the closed rectangle $\overline{R}_{a-r,c-r}^{b-r,d-r} \subset R_{a,c}^{b,d}$ and has area

$$\text{area } \overline{R}_{a-r,c-r}^{b-r,d-r} = (b-a-2r)(c-d-2r).$$

The supremum of these areas over all such $r > 0$ equals (13.10). The same clearly holds for any half open rectangle that has the same closure.

More generally, given a subset $D \subset R$, we can use the proceeding collections of rectangles to approximate its area, and thus

$$\text{area } D = \sup \left\{ \sum_{R_{i,j} \subset D} \text{area } R_{i,j} \right\}. \tag{13.11}$$

Indeed, given any other essentially disjoint collection of rectangles $S_1, \ldots, S_n \subset D$, we subdivide them as above and use (13.8) to compute their total area in terms of the areas of the $R_{i,j}$'s.

There are many examples of subsets of $\mathbb{R}^2$ whose area is of interest. Some that come readily to mind include triangles, parallelograms, polygons, circles, ellipses, and so on. With some work, one establishes the well-known geometrical formulas for their areas. However, calculus provides a much easier and more powerful route to these formulas, as well as many others that are not covered in elementary geometry, e.g., the area between two intersecting parabolas.

In this exposition, we will concentrate on bounded regions, meaning that $D \subset R$ is contained in a bounded rectangle R. Since area $D \leq$ area $R < \infty$, any bounded region has finite area: $0 \leq$ area $D < \infty$. On the other hand, many of our constructions can be extended straightforwardly to the unbounded case, and there are unbounded regions with finite area; for example, a straight line has zero area. Regions with unbounded but asymptotically thin cusps or whiskers can also have finite area.

Let us next present some general properties of area. First, let us prove the subset property (13.3): $D \subset E$ implies area $D \leq$ area E. If D is empty, the result is trivial. Otherwise, if $R_1, \ldots, R_n$ are essentially disjoint open rectangles contained in D, then they are also contained in E. Thus, recalling Definition 13.2, $A_D \subset A_E$, and hence area $D = \sup A_D \leq \sup A_E =$ area E.

A second important property is how area behaves under unions of sets. We would intuitively expect that area $(D \cup E) \leq$ area $D +$ area E with equality if and only if D, E are disjoint, or more generally, when area $(D \cap E) = 0$. Surprisingly, *this is not correct in general*! For general sets, the best we can do is the opposite inequality:

$$\text{area}\, D + \text{area}\, E \leq \text{area}\, (D \cup E) \quad \text{provided} \quad \text{area}\, (D \cap E) = 0. \tag{13.12}$$

To be clear, the fact that we may have an inequality is not intuitive. But Example 13.3 below shows that the strict inequality can hold. In other words, when we take the union of two sets, we may amass more area than expected. Indeed, if $R_1, \ldots, R_m$ are an essentially disjoint collection of closed rectangles contained in D and $S_1, \ldots, S_n$ are an essentially disjoint collection of closed rectangles contained in E, then the combined collection $R_1, \ldots, R_m, S_1, \ldots, S_n$ is also essentially disjoint, since each $R_i \cap S_j \subset D \cap E$, and thus has area 0. Taking the suprema over all such collections, as in (13.6),

$$\begin{aligned} \text{area}\, D + \text{area}\, E &= \sup \left(\sum_{i=1}^{m} \text{area}\, R_i \right) + \sup \left(\sum_{j=1}^{m} \text{area}\, S_j \right) \\ &= \sup \left(\sum_{i=1}^{m} \text{area}\, R_i + \sum_{j=1}^{m} \text{area}\, S_j \right) \leq \text{area}\, (D \cup E). \end{aligned}$$

However, we cannot replace the inequality in (13.12) by an equality since there may well be other collections of rectangles contained in $D \cup E$ that cannot be split up into subrectangles of D and E. Here is an example.

Example 13.3. Consider following two subsets of the unit square $S = \overline{R}^{1,1}_{0,0}$:

$$\begin{aligned} D &= \{\, (x,y) \mid\ 0 \leq x, y \leq 1, \quad y = 0 \ \text{ if } \ x \in \mathbb{Q}, \quad y < 1 \ \text{ if } \ x \notin \mathbb{Q} \,\}, \\ E &= \{\, (x,y) \mid\ 0 \leq x, y \leq 1, \quad y > 0 \ \text{ if } \ x \in \mathbb{Q}, \quad y = 1 \ \text{ if } \ x \notin \mathbb{Q} \,\}. \end{aligned}$$

First note that they are disjoint: $D \cap E = \varnothing$. Neither D nor E contains any open rectangle (why?), and hence area $D =$ area $E = 0$. However, $D \cup E = S$, with

$$1 = \text{area}\, S = \text{area}\, (D \cup E) > \text{area}\, D + \text{area}\, E = 0.$$

We can go further: setting $\widehat{D} = D \cup \overline{R}_{0,1}^{1,4/3}$ and $\widehat{E} = E \cup \overline{R}_{0,1}^{1,4/3}$, then

$$\text{area}\, \widehat{D} + \text{area}\, \widehat{E} = 2\, \text{area}\, \overline{R}_{0,1}^{1,4/3} = \tfrac{2}{3} < \text{area}\, (\widehat{D} \cup \widehat{E}) = \text{area}\, \overline{R}_{0,0}^{1,4/3} = \tfrac{4}{3},$$

even though $\widehat{D} \cap \widehat{E} = \overline{R}_{0,1}^{1,4/3}$ only has area $\frac{1}{3}$.

Such pathological examples are counterintuitive, and it will be good to avoid them, which we do as follows. In general, we will call bounded subsets $D, E \subset \mathbb{R}^2$ *area compatible* if their intersection is empty or more generally has zero area, and the area of their union is the sum of their individual areas, i.e., they satisfy

$$\text{area}\, D + \text{area}\, E = \text{area}\, (D \cup E), \qquad \text{area}\, (D \cap E) = 0. \tag{13.13}$$

All standard geometrical subsets — rectangles, triangles, polygons, disks, ellipses, etc. — and unions thereof are mutually area compatible. Only pathological subsets like those in Example 13.3, certain fractals, etc. are incompatible. Area compatibility will play an essential role in our approach to integration. There is in general no simple test to check if two subsets are or are not area compatible. However, one special case is contained in the following useful result.

Lemma 13.4. *Let $a \in \mathbb{R}$. Suppose $D \subset \{\, (x,y) \mid x \le a \,\}$ and $E \subset \{\, (x,y) \mid x \ge a \,\}$. Then D, E are area compatible. Area compatibility also holds when $D \subset \{\, (x,y) \mid y \le a \,\}$ and $E \subset \{\, (x,y) \mid y \ge a \,\}$.*

Proof: Let $R_1, \ldots, R_n$ be an essentially disjoint collection of closed rectangles, with each $R_i \subset D \cup E$. We split the collection into two subcollections $\mathcal{R}_D, \mathcal{R}_E$ that are, respectively, contained in D and E. If $R_i \subset D$, then it is in $\mathcal{R}_D$; if $R_i \subset E$, then it is in $\mathcal{R}_E$. The remaining R_i can be divided into two subrectangles $R_i^- = R_i \cap \{\, x \le a \,\} \subset D$ and $R_i^+ = R_i \cap \{\, x \ge a \,\} \subset E$, and the two halves are included in their respective subcollections, so $R_i^- \in \mathcal{R}_D,\ R_i^+ \in \mathcal{R}_E$. Note that

$$\text{area}\, R_i^+ + \text{area}\, R_i^- = \text{area}\, R_i.$$

Thus,

$$\sum_{i=1}^{m} \text{area}\, R_i = \sum_{R \,\subset\, \mathcal{R}_D \cup \mathcal{R}_E} \text{area}\, R \ \le\ \text{area}\, D + \text{area}\, E,$$

and hence

$$\text{area}\, (D \cup E) = \sup \left\{ \sum_{i=1}^{m} \text{area}\, R_i \right\} \le \text{area}\, D + \text{area}\, E,$$

which, when compared with (13.12), proves their area compatibility (13.13). The second statement is proved in an identical manner. *Q.E.D.*

More generally, one expects sets with “nice” boundaries to be area compatible, while those with fractal boundaries or worse are less likely to be. For our purposes, though, the above lemma suffices.

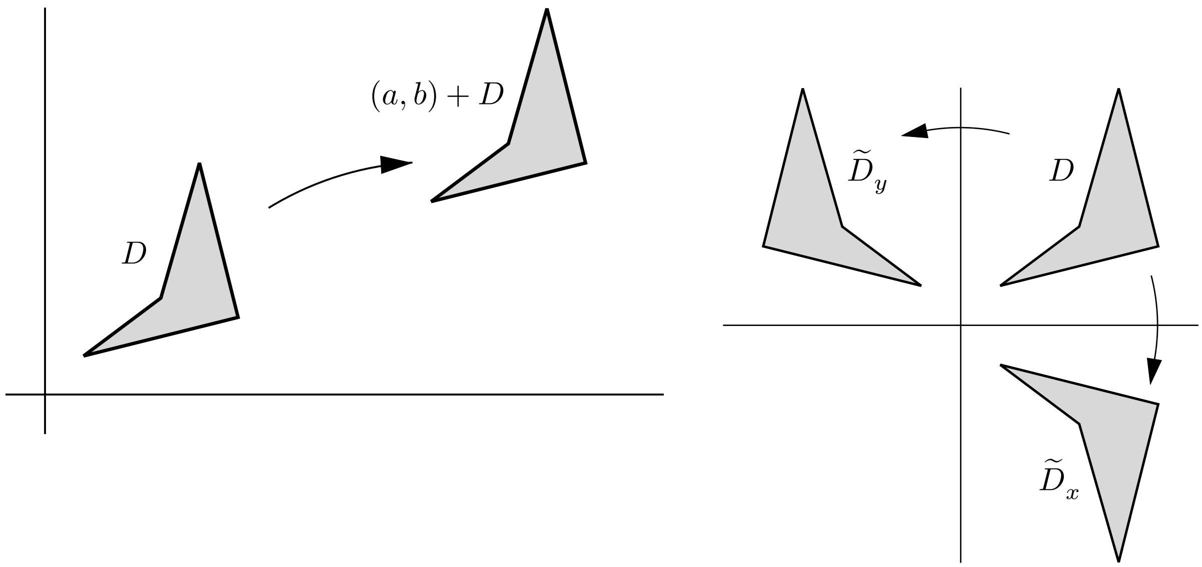


Figure 21. Translated and Reflected Regions.

Before proceeding further, we note that areas are unchanged under several elementary transformations of the plane. In each case, we need only prove the result for closed rectangles; the general case will follow from our Definition 13.2 of area.

Lemma 13.5. *If* $D \subset \mathbb{R}^2$ *and* $(a, b) \in \mathbb{R}^2$*, define the* translated region

$$\widetilde{D} = (a, b) + D = \{\, (x + a, y + b) \mid (x, y) \in D \,\}.$$

Then area $\widetilde{D}$ = area D.

See Figure 21. Indeed, translation preserves the area of rectangles:

$$\text{if} \quad R = \overline{R}^{f,d}_{e,c}, \quad \text{then} \quad \widetilde{R} = (a, b) + R = \overline{R}^{f+a,d+b}_{e+a,c+b},$$

and hence

$$\text{area } \widetilde{R} = \big[\,(f + a) - (e + a)\,\big]\big[\,(d + b) - (c + b)\,\big] = (f - e)\,(d - c) = \text{area } R.$$

As noted above, this suffices to prove the result in general.

Lemma 13.6. *If* $D \subset \mathbb{R}^2$*, then the* reflected regions

$$\widetilde{D}_x = \{\, (x, -y) \mid (x, y) \in D \,\}, \qquad \widetilde{D}_y = \{\, (-x, y) \mid (x, y) \in D \,\}, \tag{13.14}$$

both have the same area as D.

Here $\widetilde{D}_x$ is the reflection of D with the x axis serving as the mirror, while $\widetilde{D}_y$ is the reflection of D through the y axis; see Figure 21. Again, one checks that neither reflection affects the areas of rectangles.

Other examples of area-preserving transformations are the *rotations*

$$(x, y) \longmapsto (x \cos\theta - y \sin\theta,\ x \sin\theta + y \cos\theta), \qquad \text{where} \qquad 0 \le \theta < 2\pi, \tag{13.15}$$

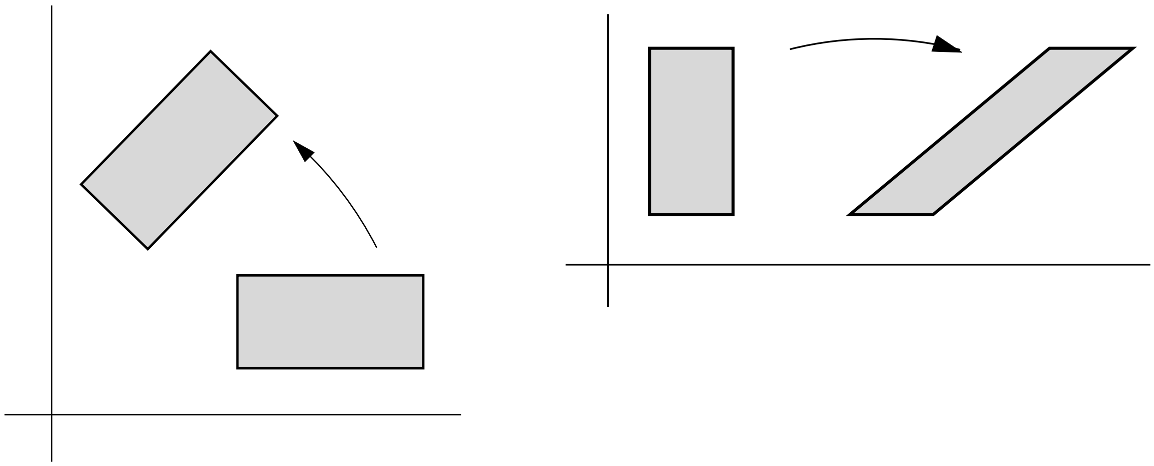


Figure 22. Rotation and Shear of Rectangles.

and the *shearing transformations*

$$\begin{aligned} (x,y) &\longmapsto (x+\beta\, y,\ y), \\ (x,y) &\longmapsto (x,\ \gamma\, x+y), \end{aligned} \qquad \text{where} \qquad \beta,\gamma\in\mathbb{R}, \tag{13.16}$$

along, respectively, the x and y axes. The former rotate rectangles into "rectangular diamonds" of the same area; the latter shear rectangles into parallelograms again of the same area; see Figure 22. Proof that they both preserve areas of rectangles is a bit more challenging, and will be left to the reader. One can, of course, compose such transformations to produce other area-preserving transformations. In general, an *affine transformation*

$$(x,y) \longmapsto (\alpha\, x+\beta\, y+a,\ \gamma\, x+\delta\, y+b) \tag{13.17}$$

where $\alpha,\beta,\gamma,\delta,a,b\in\mathbb{R}$, preserves area if and only if the absolute value of its *determinant* equals

$$|\,\alpha\,\delta-\beta\,\gamma\,| = 1. \tag{13.18}$$

All such affine transformations map parallelograms, including rectangles, into parallelograms of the same area, [**22**]. There also exist a wide variety of nonlinear transformations that preserve area, [**20**].

On the other hand, stretching the plane in one or both directions multiplies the area by the corresponding factor(s).

Lemma 13.7. *Let $\lambda,\mu\in\mathbb{R}$. If $D\subset\mathbb{R}^2$, then the* stretched regions

$$\widehat{D}_\lambda = \{\,(x,\lambda y)\mid\ (x,y)\in D\,\}, \qquad \widetilde{D}_\mu = \{\,(\mu x, y)\mid\ (x,y)\in D\,\}, \tag{13.19}$$

satisfy

$$\text{area}\ \widehat{D}_\lambda = |\,\lambda\,|\ \text{area}\ D, \qquad \text{area}\ \widetilde{D}_\mu = |\,\mu\,|\ \text{area}\ D.$$

Again the result is proved by showing that it has the given effect when $D = R$ is a rectangle. If $\lambda = 0$ and/or $\mu = 0$ in (13.19) the region is squashed down to a subset of

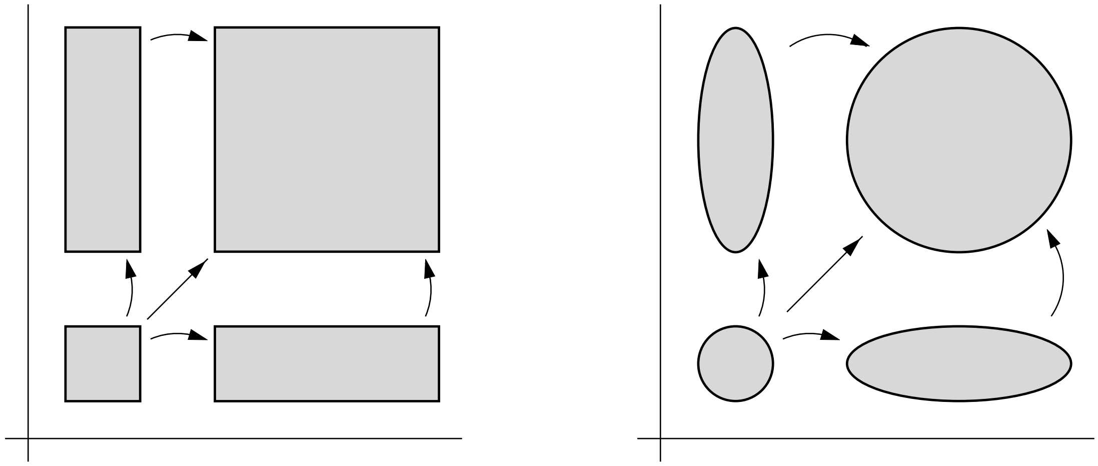


Figure 23. Stretching a Square and a Disk.

the x or y axis, respectively, with zero area. If $\lambda < 0$ or $\mu < 0$ the region is reflected and stretched, with the cases $\lambda = \mu = -1$ being the pure reflections in Lemma 13.6. We can combine individual scalings; for example if we scale both x and y by $\lambda = \mu$ (the effect of zooming on a camera or a computer screen), then the area of the scaled region is multiplied by λ^2, which is the reason area is measured in units of length^2. If one subjects a circle to any combination of stretches, the result is, in general, an ellipse, and hence one can use Lemma 13.7 to deduce the formula for the area of an ellipse from that of a circle. Figure 23 illustrates the effect of stretching transformations on a square and a circular disk. The horizontal arrows indicate stretches in the x direction; vertical arrows indicate stretches in the y direction; while the diagonal arrow is the scaling transformation obtained by combining the two individual stretches.

More generally, an affine transformation (13.17) scales area by the overall factor $\delta = |\,\alpha\,\delta - \beta\,\gamma\,|$. The calculus formula for the behavior of area under a nonlinear transformation of the plane can be found in [**1**, **6**].

14. Integration.

In calculus, we are particularly interested in computing the areas of regions that lie under the graphs of functions, as illustrated in Figure 24. Indeed, the areas of the geometric examples mentioned above can all be determined once we know how to compute the areas of this particular type of region since, assuming area compatibility, we can glue such regions together to make more complicated regions.

To begin with, suppose $f\colon [a,b] \to [0,\infty)$ is a positive function, not necessarily continuous, so $f(x) \geq 0$ for all $a \leq x \leq b$. We define its *lower region*

$$L = \{\, (x,y) \mid \ a \leq x \leq b,\ 0 \leq y \leq f(x) \,\}, \tag{14.1}$$

meaning that part of the plane lying on or above the x axis and on or below its graph

$$G = \{\, (x, f(x)) \mid \ a \leq x \leq b \,\}. \tag{14.2}$$

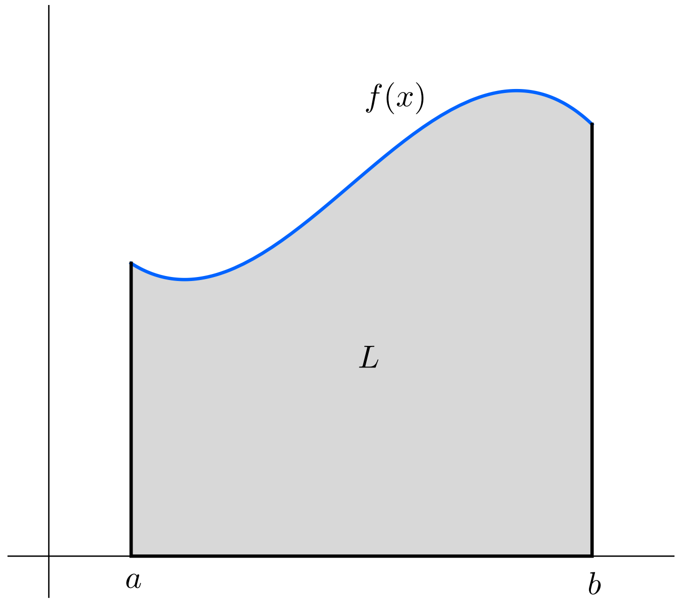


Figure 24. The Lower Region under the Graph of a Function.

As per Definition 13.2, the area of L equals the supremum of the total areas of essentially disjoint collections of closed subrectangles $R_1, \ldots, R_n \subset L$.

Let us further assume that f is bounded from above on the interval $[a,b]$, so there exists $\beta \geq 0$ such that $0 \leq f(x) \leq \beta$ for all $a \leq x \leq b$; equivalently, $f[a,b] \subset [0,\beta]$. Let

$$U_\beta = \{ (x,y) \mid a \leq x \leq b, \ f(x) \leq y \leq \beta \} \tag{14.3}$$

be the complementary *upper region*, consisting of points lying on or above the graph of f and on or below the line $\{y = \beta\}$. Observe, as sketched in Figure 25, that the union of the lower region L and the upper region U_β is the closed rectangle of width β sitting on the interval $[a,b]$:

$$R_\beta = L \cup U_\beta = \overline{R}_{a,0}^{b,\beta} = \{ (x,y) \mid a \leq x \leq b, \ \ 0 \leq y \leq \beta \} .$$

On the other hand, the intersection of the upper and lower regions is the graph: $L \cap U_\beta = G$, which has area 0 because it contains no rectangles of nonzero area. The basic premise of integration is that the upper and lower regions be area compatible, in the sense of (13.13), which serves to inspire the following key definition.

Definition 14.1. A function $f\colon [a,b] \to [0,\beta]$ is *integrable* if its upper and lower regions (14.1, 3) are area compatible, meaning

$$\text{area } L + \text{area } U_\beta = \text{area } R_\beta = \beta\,(b-a). \tag{14.4}$$

In this case, we define the *integral* of f on the interval $[a,b]$ to be the area of its lower region:

$$\int_a^b f(x)\,dx = \text{area } L = \text{area } \{ (x,y) \mid a \leq x \leq b, \ 0 \leq y \leq f(x) \} . \tag{14.5}$$

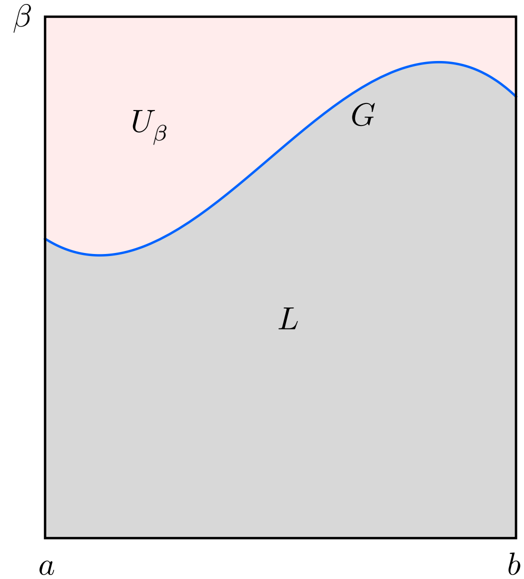


Figure 25. Upper and Lower Regions and Graph.

The left hand side of (14.5) is sometimes referred to as a *definite integral*, which indicates that the endpoints a, b of the interval of integration are to be viewed as "fixed". See Section 15 for a full explanation of this terminology.

Remark: The integral symbol $\displaystyle\int$ is an elongated "S" which stands for "sum", indicating that we are calculating areas by summing up (and then taking a supremum) of the areas of rectangles. The notation dates back to the German mathematician and philosopher Gottfried Leibniz, who was Newton's contemporaneous rival inventor of the calculus. While Newton can legitimately claim priority, Leibniz's notation proved to be the better, which is one of the reasons much of the initial post-Newton development of calculus took place on the European continent rather than in England. The integration variable x, which is indicated by the symbol dx, is a "dummy variable" and any other letter, except the variables a and b that appear in the *integration limits* indicating the interval to be integrated over, can be used instead:

$$\int_a^b f(x)\,dx = \int_a^b f(y)\,dy = \int_a^b f(t)\,dt,$$

and so on. These are just different ways of denoting the same area. In Leibniz's thinking, dx represents the width of an "infinitesimally" narrow rectangle lying under the graph of the function, and so the product $f(x)\,dx$ is to viewed as its "infinitesimal" area. The total area or integral is then obtained by "summing" these infinitely many infinitesimal quantities. Mathematicians have endeavored to make this idea rigorous, [**23**], but we will not pursue this line of reasoning here.

For example, any constant function $f(x) \equiv c$, where $c \geq 0$, is integrable on any closed bounded interval $[a, b]$. (Later we will see that this is also true when $c < 0$.) Indeed, the

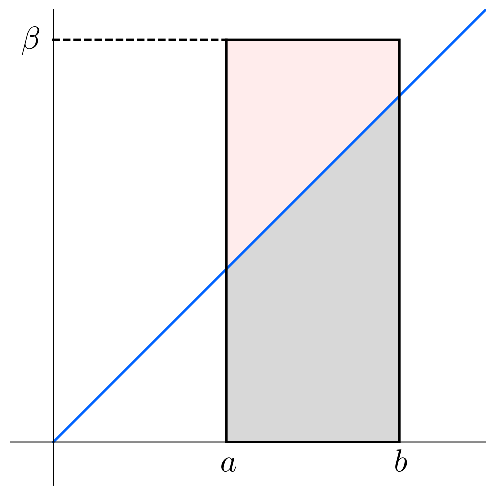


Figure 26. Integral of x.

associated lower region $L = \{ (x,y) \mid a \le x \le b,\ 0 \le y \le c \}$ is just a closed rectangle, namely $\overline{R}^{b,c}_{a,0}$, and hence its area is $c\,(b-a)$. On the other hand, given $\beta \ge c$, the upper region is also a closed rectangle, $U_\beta = \{ (x,y) \mid a \le x \le b,\ c \le y \le \beta \} = \overline{R}^{b,\beta}_{a,c}$, with area $(\beta - c)\,(b-a)$. The sum of their areas is thus the area of the entire rectangle of width β:

$$\text{area}\, L + \text{area}\, U_\beta = c\,(b-a) + (\beta - c)\,(b-a) = \beta\,(b-a) = \text{area}\, R_\beta,$$

verifying the integrability criterion (14.4). Thus, the integral of a constant function is the area of its lower region:

$$\int_a^b c\, dx = c\,(b-a). \tag{14.6}$$

A second relatively easy case is the function $f(x) = x$ on the interval $[a,b]$ where $0 \le a \le b$ to ensure positivity. The lower region $L = \{ (x,y) \mid a \le x \le b,\ 0 \le y \le x \}$ is a trapezoid with vertical side lengths a, b — see Figure 26 — and hence its area is known to be $\frac{1}{2}(a+b)(b-a) = \frac{1}{2}b^2 - \frac{1}{2}a^2$. Further, given $\beta > b$, the upper region $U_\beta = \{ (x,y) \mid a \le x \le b,\ x \le y \le \beta \}$ is also a trapezoid, now with vertical side lengths $\beta - a$ and $\beta - b$, and hence its area is $\frac{1}{2}\left[(\beta - a) + (\beta - b) \right](b-a) = \beta\,(b-a) - \frac{1}{2}b^2 + \frac{1}{2}a^2$. The sum of their areas is

$$\text{area}\, L + \text{area}\, U_\beta = \left[\tfrac{1}{2}b^2 - \tfrac{1}{2}a^2 \right] + \left[\beta\,(b-a) - \tfrac{1}{2}b^2 + \tfrac{1}{2}a^2 \right] = \beta\,(b-a) = \text{area}\, R_\beta,$$

thus establishing the integrability criterion (14.4). We conclude that

$$\int_a^b x\, dx = \tfrac{1}{2}b^2 - \tfrac{1}{2}a^2. \tag{14.7}$$

We will defer trying to compute any more complicated examples until we have established the remarkable connection between integration and differentiation known as the Fundamental Theorem of Calculus. With this powerful tool in hand, both of the preceding integrals, and many more, will reduce to easy one-line calculations. It is important to note at the outset that all "reasonable" functions, including, as we will shortly learn, those that are not everywhere positive, are integrable. Later, in (14.15), we will exhibit an example of a pathological non-integrable function.

The integrability requirement (14.4) is, reassuringly, independent of the upper bound β. Indeed, if $\gamma > \beta$, then the corresponding upper region

$$\begin{aligned} U_\gamma &= \{\,(x,y)\mid a\le x\le b,\ f(x)\le y\le \gamma\,\} \\ &= \{\,(x,y)\mid a\le x\le b,\ f(x)\le y\le \beta\,\} \cup \{\,(x,y)\mid a\le x\le b,\ \beta\le y\le \gamma\,\} = U_\beta \cup \overline{R}^{b,\gamma}_{a,\beta} \end{aligned}$$

is obtained by gluing a closed rectangle of width $\gamma-\beta$ lying directly above U_β, meeting at the horizontal line $y=\beta$. In view of Lemma 13.4,

$$\text{area}\, U_\gamma = \text{area}\, U_\beta + \text{area}\, \overline{R}^{b,\gamma}_{a,\beta} = \text{area}\, U_\beta + (\gamma-\beta)\,(b-a).$$

Since we are assuming (14.4) holds, this implies

$$\begin{aligned} \text{area}\, L + \text{area}\, U_\gamma &= \text{area}\, L + \text{area}\, U_\beta + \text{area}\, \overline{R}^{b,\gamma}_{a,\beta} \\ &= \beta\,(b-a) + (\gamma-\beta)\,(b-a) = \gamma\,(b-a) = \text{area}\, R_\gamma, \end{aligned}$$

and thus (14.4) also holds with γ replacing β.

Some basic properties of integration follow. First, if $a=b$, every function is integrable on the singleton $[a,a]=\{\,a\,\}$, with

$$\int_a^a f(x)\,dx = 0. \tag{14.8}$$

Indeed, both the lower and upper regions have zero area, so the integrability requirement (14.4) is trivially satisfied.

Proposition 14.2. *If $c\ge 0$ is constant and f is integrable on $[a,b]$, then the function $c\,f(x)$ is also integrable, and*

$$\int_a^b \big[\,c\,f(x)\,\big]\,dx = c\int_a^b f(x)\,dx. \tag{14.9}$$

Proof: The lower region of the function $c\,f(x)$ is obtained by stretching the lower region of $f(x)$ by the factor c:

$$\{\,(x,y)\mid\ a\le x\le b,\ 0\le y\le c\,f(x)\,\} = \{\,(x,c\,y)\mid\ a\le x\le b,\ 0\le y\le f(x)\,\}.$$

Since $c\ge 0$, Lemma 13.7 implies that the area of the stretched region is c times the area of the original, which implies (14.9), subject to verification that $c\,f(x)$ is integrable. For this, using $c\,\beta$ for the upper bound for $c\,f(x)$, the upper region is stretched by a similar amount, as is the corresponding rectangle: $R_\beta \longmapsto R_{c\,\beta}$, and one thus verifies a scaled version of the integrability condition (14.4). *Q.E.D.*

It will be useful to extend formula (14.9) to all real $c\in\mathbb{R}$, which will allow us to integrate negative functions. If $0\le f(x)\le\beta$ on $[a,b]$, then its negative satisfies $-\beta\le -f(x)\le 0$,

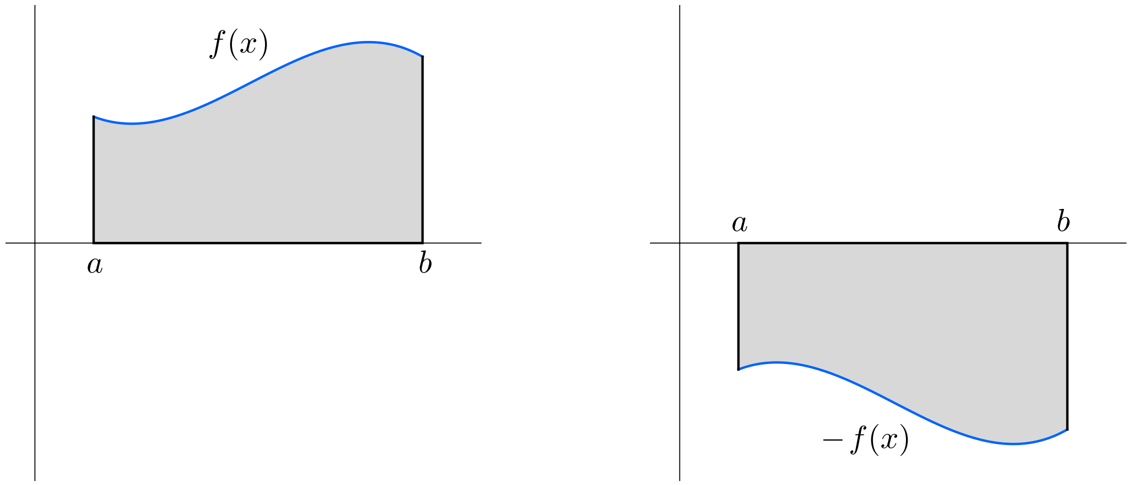


Figure 27. Integration Regions for a Function and its Negative.

and is deemed to have a negative integral:

$$\int_a^b \big[-f(x) \big] \, dx = -\int_a^b f(x)\, dx, \tag{14.10}$$

which is the case $c = -1$ of (14.9). The right hand side is the negative of the area of the lower region L of f, namely the region (14.1) lying below the graph of $f(x)$ and above the x axis. The corresponding region for $-f(x)$ is

$$\widetilde{L} = \{ \, (x,y) \mid \; a \leq x \leq b, \;\; -f(x) \leq y \leq 0 \, \} = \{ \, (x,-y) \mid \; a \leq x \leq b, \; 0 \leq y \leq f(x) \, \},$$

which is the reflection of the lower region L through the x axis, which, according to Lemma 13.6, has the same area $\widetilde{L}$ = area L. Equivalently, $\widetilde{L}$ is the region lying above the graph of $-f(x)$ and below the x axis; see Figure 27. The right hand side of formula (14.10) is the *negative* of the area of this region, and hence the integral of a negative function is *minus* the area of the region $\widetilde{L}$ lying between its graph and the x axis. In other words, from the point of view of integration, regions below the x axis make negative contributions to the integral. For example, if $f(x) = -x$ for $0 \leq a \leq x \leq b$, then

$$\int_a^b (-x)\, dx = -\int_a^b x\, dx = \tfrac{1}{2}\, a^2 - \tfrac{1}{2}\, b^2$$

is minus the area of the trapezoid lying between its graph and the x axis. More generally, if the graph of $f(x)$ lies both above and below the x axis, its integral is the sum of the areas of the subregions lying above the axis minus the sum of the areas of the regions lying below; in Figure 28, the blue regions make positive contributions to the integral while the purple regions make negative contributions, and the integral is the total of all such contributions. A justification will appear shortly.

One can always convert a bounded but non-positive function into an everywhere positive function by adding a large positive constant to it, effectively translating its graph upwards.

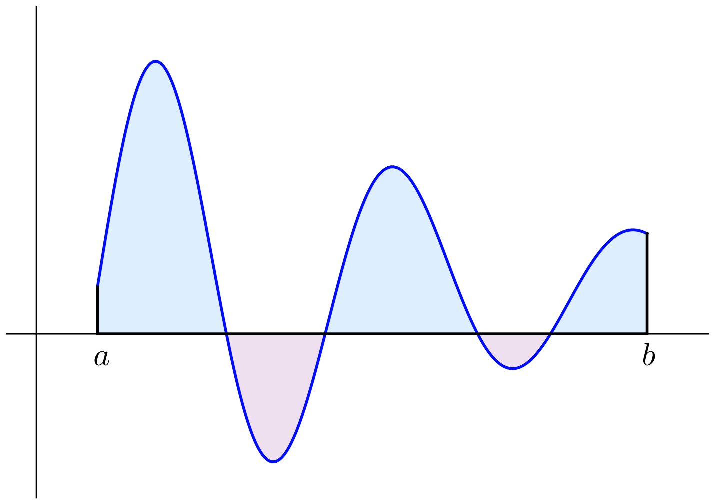


Figure 28. Integral of a General Function.

The basic formula is

$$\int_a^b \big[\, f(x) + c \,\big]\, dx = \int_a^b f(x)\, dx + \int_a^b c\, dx = \int_a^b f(x)\, dx + c\,(b-a), \tag{14.11}$$

which is a special case of a later formula, namely (14.23), but is easy to prove directly when f and c are both positive; we then extend its validity to all integrable functions and all $c \in \mathbb{R}$. In the former case, the lower region for $f(x) + c$ is

$$\begin{aligned} &\{\, (x,y) \mid\ a \le x \le b,\ 0 \le y \le f(x) + c \,\} \\ &\qquad = \{\, (x,y) \mid\ a \le x \le b,\ 0 \le y \le c \,\} \cup \{\, (x,y) \mid\ a \le x \le b,\ c \le y \le f(x) + c \,\}, \end{aligned}$$

which is the union of the rectangle $\overline{R}{}^{b,c}_{a,0}$ and a vertically translated version of the lower region of f:

$$\{\, (x,y) \mid\ a \le x \le b,\ c \le y \le f(x) + c \,\} = \{\, (x,y+c) \mid\ a \le x \le b,\ 0 \le y \le f(x) \,\};$$

see Figure 29. Lemma 13.5 says that the translated region has the same area as the original, hence (14.11) reduces to the result that the area of this disjoint union is just the sum of the individual areas, which follows from Lemma 13.4. One can prove integrability of $f(x) + c$ by replacing the upper bound β for $f(x)$ by $\beta + c$, and noting that its upper region is the same translate of the upper region of $f(x)$:

$$\{\, (x,y) \mid\ a \le x \le b,\ f(x) + c \le y \le \beta + c \,\} = \{\, (x,y+c) \mid\ a \le x \le b,\ \ f(x) \le y \le \beta \,\},$$

and hence also has the same area.

For example, if $f(x) = x - 1$ for $0 \le x \le 1$, then we can chose $c = 1$ so that $f(x) + 1 = x$ and hence, according to (14.7),

$$\tfrac{1}{2} = \int_0^1 x\, dx = \int_0^1 (x-1)\, dx + (1-0) = 1 + \int_0^1 (x-1)\, dx, \qquad \text{so} \qquad \int_0^1 (x-1)\, dx = -\tfrac{1}{2}.$$

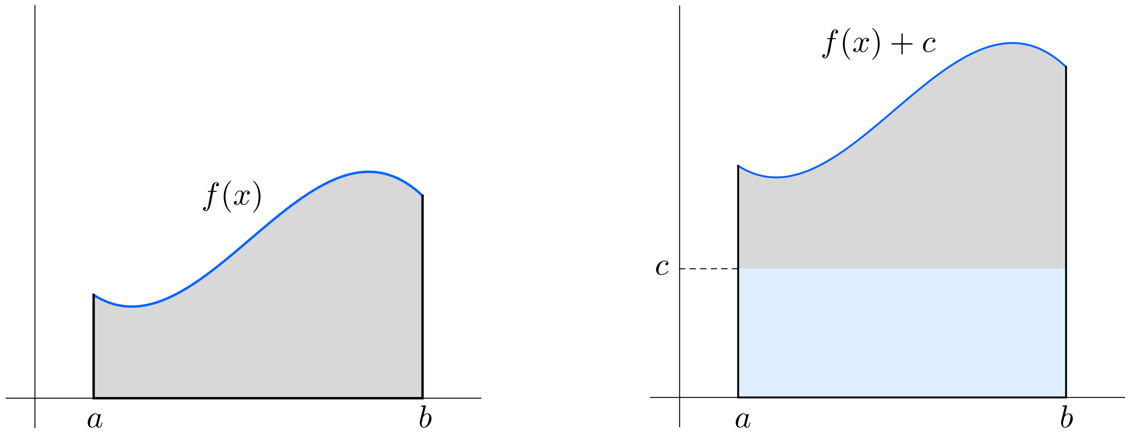


Figure 29. Vertically Translated Function.

The function $x-1$ lies entirely below the x axis when $0 \le x \le 1$, and we are effectively computing the negative of the area of the triangle that lies between its graph and the x axis.

In general we only need to establish an integration property for positive functions, and then extend it as stated to general functions by the same technique of adding in a suitably large constant.

Example 14.3. Given $a < b$, a function that is constant on a finite number of disjoint open subintervals is known as a *step function*. In other words, given $I_j = (a_j, b_j)$ for $j = 1, \ldots, m$, with

$$a \le a_1 < b_1 \le a_2 < b_2 \le \ \cdots \ \le a_m < b_m \le b \tag{14.12}$$

and constants $c_1, \ldots, c_m \in \mathbb{R}$, the corresponding step function is

$$\sigma(x) = \begin{cases} c_j, & a_j < x < b_j, \\ 0, & \text{otherwise.} \end{cases} \tag{14.13}$$

For simplicity, we have assigned the value 0 outside the given rectangles; indeed, the values at the endpoints a_j, b_j of the subintervals play no role in its integration. An example is the sign function (4.1), restricted to a bounded interval $[a, b]$.

If the step function is non-negative, so all $c_j \ge 0$, then the corresponding lower region consists of the disjoint closed rectangles $\overline{R}^{b_j, c_j}_{a_j, 0}$ that lie below its graph over each interval. The integral of the step function (14.13) is the sum of their areas:

$$\int_a^b \sigma(x)\,dx = \sum_{i=1}^{m} c_j\,(b_j - a_j). \tag{14.14}$$

We leave the verification of the integrability criterion (14.4) for a suitable upper bound $\beta \ge \max\{c_1, \ldots, c_m\}$ to the reader. The same integration formula (14.14) holds in general, with the right hand side (14.14) being the sum of the areas of the rectangles lying above the x axis minus the sum of the areas of the rectangles lying below.

For example, consider the simple positive step function

$$\sigma(x) = \begin{cases} 5, & -1 < x < 0, \\ 3, & 0 < x < 2, \\ 0, & \text{otherwise.} \end{cases}$$

On the interval $[-1, 2]$, its lower region consists of two open rectangles, $R^{0,5}_{-1,0}$ and $R^{2,3}_{0,0}$, of respective areas 5 and 6. Thus, its integral equals the total area:

$$\int_{-1}^{2} \sigma(x)\, dx = 5 \cdot 1 + 3 \cdot 2 = 11.$$

On the other hand, the step function

$$\sigma(x) = \begin{cases} -2, & -3 < x < 0, \\ 4, & 0 < x < 1 \\ -5, & 1 < x < 3, \\ 0, & \text{otherwise} \end{cases} ,$$

is negative on some subintervals. Its integral

$$\int_{-3}^{3} \sigma(x)\, dx = -2 \cdot 3 + 4 \cdot 1 - 5 \cdot 2 = -12$$

combines the area of the rectangle $R^{1,4}_{0,0}$ above the x axis with the negatives of the areas of the rectangles $R^{0,0}_{-3,-2}$ and $R^{3,0}_{1,-5}$ that lie below.

As noted above, most "reasonably behaved" functions are integrable. In particular, this includes all bounded monotone (nondecreasing or nonincreasing) functions.

Theorem 14.4. *$f\colon [a,b] \to \mathbb{R}$ is monotone and bounded, then f is integrable.*

Another fundamental result states that all continuous functions are integrable. Proofs of both of these results will appear below.

Theorem 14.5. *If $f\colon [a,b] \to \mathbb{R}$ is continuous, then f is integrable.*

Remark: There are many bounded discontinuous functions that are nevertheless integrable. Any *piecewise integrable function*, meaning that f is integrable when restricted to each of a finite union of subintervals, is integrable on the entire interval. Thus piecewise monotone and piecewise continuous functions, including piecewise constant step functions, as in Example 14.3, are all integrable. This result follows by repeated application of Proposition 14.7 below.

On the other hand, the pathological function

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q}, \\ 1, & x \in \mathbb{R} \setminus \mathbb{Q}, \end{cases} \tag{14.15}$$

is not integrable on any closed interval $a \le x \le b$. Indeed, using the upper bound $\beta = 1$, the lower and upper regions are

$$\begin{aligned} L &= \{\, (x,y) \mid \ a \le x \le b, \ \ 0 \le y \le f(x) \,\} \\ &= \{\, (x,y) \mid \ a \le x \le b, \ \ y = 0 \ \text{ if } \ x \in [a,b] \cap \mathbb{Q}, \ \ 0 \le y \le 1 \ \text{ if } \ x \in [a,b] \setminus \mathbb{Q} \,\}, \\ U_\beta &= \{\, (x,y) \mid \ a \le x \le b, \ \ f(x) \le y \le 1 \,\} \\ &= \{\, (x,y) \mid \ a \le x \le b, \ \ 0 \le y \le 1 \ \text{ if } \ x \in [a,b] \cap \mathbb{Q}, \ \ y = 1 \ \text{ if } \ x \in [a,b] \setminus \mathbb{Q} \,\}. \end{aligned}$$

By the same reasoning as used in Example 13.3, we find

$$\text{area}\, L + \text{area}\, U_\beta = 0 < \text{area}\, R_\beta = b - a,$$

and hence L and U_β are not area compatible and the integrability condition (14.4) is *not* satisfied.

Let us summarize further key properties of the integral. Some proofs will be deferred until later in this section. We begin with a basic inequality.

Proposition 14.6. *If $f(x) \le g(x)$ for all $x \in [a,b]$ and both are integrable, then*

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx. \tag{14.16}$$

Remark: The result also holds for the strict inequalities, but is harder to prove; see Theorem 14.13 below.

Proof: Assuming $0 \le f(x) \le g(x)$, the lower regions are

$$\begin{aligned} \int_a^b f(x)\,dx &= \text{area}\, L_f, \\ \int_a^b g(x)\,dx &= \text{area}\, L_g, \end{aligned} \qquad \text{where} \qquad \begin{aligned} L_f &= \{\, (x,y) \mid \ a \le x \le b, \ 0 \le y \le f(x) \,\}, \\ L_g &= \{\, (x,y) \mid \ a \le x \le b, \ 0 \le y \le g(x) \,\}. \end{aligned}$$

The inequality implies $L_f \subset L_g$, and hence, according to (13.3), area $L_f \le$ area L_g, which is the same as (14.16). More generally, we add the same suitably large constant to f and g to make them both positive while preserving the inequality, using (14.11) to complete the proof. *Q.E.D.*

In particular, if $\alpha \le f(x) \le \beta$ for all $x \in [a,b]$, then in view of (14.6),

$$\alpha\,(b-a) \le \int_a^b f(x)\,dx \le \beta\,(b-a). \tag{14.17}$$

A second consequence is the absolute value integral inequality

$$\left| \int_a^b f(x)\,dx \right| \le \int_a^b |\, f(x) \,|\,dx. \tag{14.18}$$

This follows from (14.10) and (14.16) using the inequalities $-|\, f(x)\,| \le f(x) \le |\, f(x)\,|$.

The next result says that we can evaluate an integral by splitting up its domain into subintervals.

Proposition 14.7. *Suppose $a < b < c$. Then f is integrable on the intervals $[a,b]$ and $[b,c]$ if and only if it is integrable on the combined interval $[a,c]$, in which case*

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx. \tag{14.19}$$

Proof: As before, we prove this assuming $0 \le f(x) \le \beta$ for $a \le x \le c$, and then extend the formula to general functions by adding in suitable constant. Under this assumption, the full lower and upper regions

$$L = \{\,(x,y) \mid a \le x \le c,\ 0 \le y \le f(x)\,\}, \qquad U_\beta = \{\,(x,y) \mid a \le x \le c,\ f(x) \le y \le \beta\,\},$$

are unions of their counterparts

$$\widetilde{L} = \{\,(x,y) \mid a \le x \le b,\ 0 \le y \le f(x)\,\}, \quad \widetilde{U}_\beta = \{\,(x,y) \mid a \le x \le b,\ f(x) \le y \le \beta\,\},$$
$$\widehat{L} = \{\,(x,y) \mid b \le x \le c,\ 0 \le y \le f(x)\,\}, \quad \widehat{U}_\beta = \{\,(x,y) \mid b \le x \le c,\ f(x) \le y \le \beta\,\},$$

Since $\widetilde{L},\ \widetilde{U}_\beta \subset \{\,(x,y) \mid x \le b\,\},\ \widehat{L},\ \widehat{U}_\beta \subset \{\,(x,y) \mid x \ge b\,\}$, Lemma 13.4 implies

$$\text{area } L = \text{area } \widetilde{L} + \text{area } \widehat{L}, \qquad \text{area } U_\beta = \text{area } \widetilde{U}_\beta + \text{area } \widehat{U}_\beta. \tag{14.20}$$

On the other hand, (13.12) implies

$$\begin{gathered} \text{area } L + \text{area } U_\beta \le \text{area } \overline{R}^{c,\beta}_{a,0}, \\ \text{area } \widetilde{L} + \text{area } \widetilde{U}_\beta \le \text{area } \overline{R}^{b,\beta}_{a,0}, \qquad \text{area } \widehat{L} + +\text{area } \widehat{U}_\beta \le \text{area } \overline{R}^{c,\beta}_{b,0}. \end{gathered} \tag{14.21}$$

If f is integrable on $[a,c]$, the first inequality is an equality. Combined with (14.20) and the fact that area $\overline{R}^{c,\beta}_{a,0} =$ area $\overline{R}^{b,\beta}_{a,0} +$ area $\overline{R}^{c,\beta}_{b,0}$ proves that the second and third are also equalities, thus proving integrability of f on $[a,b]$ and $[b,c]$. The converse is proved in a similar manner. With integrability of f on all three intervals established, the first equation in (14.20) yields (14.19). *Q.E.D.*

Let us drop the assumption that $c > b$ in (14.19), but still require that the equation as written holds. In particular, if $c = a$, then, by (14.8),

$$0 = \int_a^a f(x)\,dx = \int_a^b f(x)\,dx + \int_b^a f(x)\,dx.$$

Thus,

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx. \tag{14.22}$$

In other words, assuming $f(x) \ge 0$, integrating in the "wrong" direction, from right to left, produces the negative of the area of the lower region lying between the graph of f and the x axis, whereas integrating in the "correct" direction, from left to right, gives the area. On the other hand, if $f(x) \le 0$, then integration in the "wrong" direction produces the area of the region lying above the graph of f and below the x axis, while integrating in the "correct" direction yields its negative.

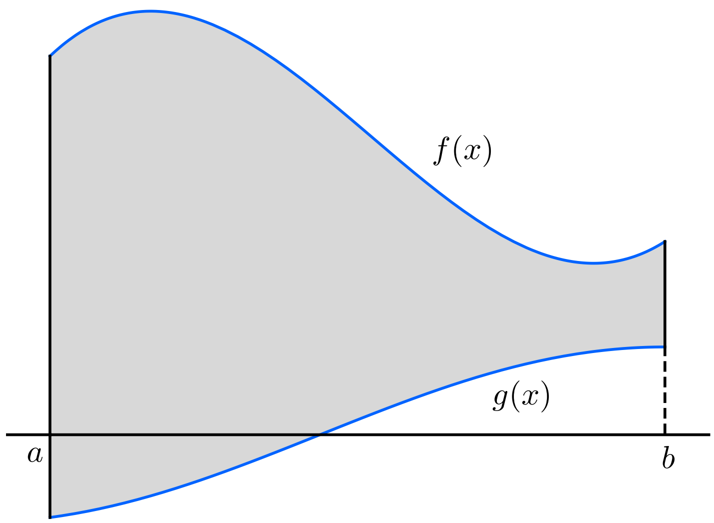


Figure 30. Region Between Functions.

The next result says that the integral of the sum of two functions is the sum of their integrals. A proof will appear later in this section.

Theorem 14.8. *If* $f:[a,b] \to [\alpha,\beta]$ *and* $g:[a,b] \to [\gamma,\delta]$ *are both bounded and integrable, then their sum* $f+g:[a,b] \to [\alpha+\gamma,\beta+\delta]$ *is also bounded and integrable, with*

$$\int_a^b \big[\, f(x)+g(x) \,\big]\, dx = \int_a^b f(x)\, dx + \int_a^b g(x)\, dx. \tag{14.23}$$

Warning: The integral of the product of two functions is (usually) *not* the product of their integrals!

In particular, if we combine (14.23) with (14.10), and assume that $f(x) \geq g(x)$ for all $a \leq x \leq b$, we find the integral of their difference can be used to compute the area of the region between their graphs, as illustrated in Figure 30:

$$\int_a^b \big[\, f(x)-g(x) \,\big]\, dx = \int_a^b f(x)\, dx - \int_a^b g(x)\, dx = \text{area} \left\{ (x,y) \;\middle|\; \begin{array}{c} a \leq x \leq b, \\ g(x) \leq y \leq f(x) \end{array} \right\}. \tag{14.24}$$

Indeed, if $0 \leq g(x) \leq f(x)$, then the region between their graphs is the set theoretic difference between the lower region for f and the lower region for g, whose areas are represented by the two individual integrals. If we drop the assumption that f is everywhere greater than g, then the integral of their difference represents the "signed area" between their graphs, where sections with $f(x) \geq g(x)$ are counted with positive area while those where $f(x) \leq g(x)$ have negative area. This generalizes our earlier observations concerning the integration of functions $f(x)$ that are both positive and negative, which corresponds to the case $g(x) \equiv 0$, whose graph is the x axis.

With the basic properties in hand, let us return to general considerations. Suppose we wish to prove that a given bounded non-negative function $f\colon [a,b] \to [0,\beta]$ is integrable, i.e., satisfies (14.4) where L, U_β are its lower and upper regions, whose union is the entire rectangle of width β:

$$L \cup U_\beta = R_\beta = \overline{R}^{b,\beta}_{a,0}.$$

Definition 14.9. An essentially disjoint collection of closed rectangles

$$\mathcal{V} = \{ S_1, \ldots, S_n, T_1, \ldots, T_m \}$$

is called f-*adapted* if $S_1, \ldots, S_n \subset L$ and $T_1, \ldots, T_m \subset U_\beta$.

Given an f-adapted collection, we let

$$S^* = \bigcup_{i=1}^{n} S_i, \qquad T^* = \bigcup_{i=1}^{m} T_i, \qquad V^* = S^* \cup T^* \subset R_\beta. \tag{14.25}$$

Further, let

$$Z^* = \overline{R_\beta \setminus V^*}, \tag{14.26}$$

denote the closure of the complementary subset, which, as in (13.9), is also the union of a set of rectangles $\mathcal{Z} = \{Z_1, \ldots, Z_l\}$. Since we are using the rectangles in V to approximate the areas of the lower and upper regions, they will fill up a large fraction of the overall rectangle R_β. Thus, we expect that Z^* is the union of a collection of correspondingly small rectangles. Indeed, according to (13.9),

$$\text{area } S^* + \text{area } T^* + \text{area } Z^* = \text{area } V^* + \text{area } Z^* = \text{area } R_\beta = \beta\,(b-a). \tag{14.27}$$

Moreover, by definition,

$$\text{area } L = \sup\{\text{area } S^*\}, \qquad \text{area } U_\beta = \sup\{\text{area } T^*\}, \tag{14.28}$$

where the suprema are taken over all possible f-adapted collections V. Thus,

$$\begin{aligned} \text{area } L + \text{area } U_\beta &= \sup\{\text{area } S^*\} + \sup\{\text{area } T^*\} = \sup\{\text{area } S^* + \text{area } T^*\} \\ &= \sup\{\text{area } R_\beta - \text{area } Z^*\} = \text{area } R_\beta - \inf\{\text{area } Z^*\}. \end{aligned}$$

We thus arrive at the following criterion for integrability.

Theorem 14.10. *The nonnegative bounded function* $f\colon [a,b] \to [0,\beta]$ *is* integrable *if and only if*

$$\inf\{\text{area } Z^*\} = 0, \tag{14.29}$$

where the infimum is over all possible f*-adapted collections of rectangles* $\mathcal{V}$ *with complement* Z^* *constructed as above in* (14.25, 26).

Lemma 14.11. *Let* $0 \le f(x) \le \beta$ *be integrable and bounded on* $[a,b]$*. If* $a \le c < d \le b$*, then* $f(x)$ *is integrable on* $[c,d]$ *and*

$$\int_c^d f(x)\,dx \le \int_a^b f(x)\,dx. \tag{14.30}$$

Proof: Assuming integrability, the inequality follows immediately because the two lower regions satisfy

$$\{\,(x,y) \mid\ a \leq x \leq b,\ 0 \leq y \leq f(x)\,\} \subset \{\,(x,y) \mid\ c \leq x \leq d,\ 0 \leq y \leq f(x)\,\},$$

and hence the area of the first is $\leq$ the area of the second. To prove integrability, given Z^* as above, its intersection $\widetilde{Z}^* = Z^* \cap X_{cd}$ with the vertical strip $X_{cd} = \{\,(x,y) \mid c \leq x \leq d\,\}$ is a complementary region for the integral on $[c,d]$, associated with the rectangles $S_i \cap X_{cd}$ and $T_i \cap X_{cd}$. Since $\widetilde{Z}^* \subset Z^*$, we have area $\widetilde{Z}^* \leq$ area Z^*. Thus, (14.29) implies $\inf\{\,\text{area}\ \widetilde{Z}^*\,\} = 0$, thereby proving integrability of $f(x)$ on the subinterval $[c,d]$. *Q.E.D.*

It is often impractical to consider all such collections of rectangles. Rather, we typically concentrate on those of a certain relatively simple type. For this purpose, we introduce a refinement of the integrability condition of Theorem 14.10.

Theorem 14.12. *Suppose that $f\colon [a,b] \to [0,\beta]$ is a nonnegative bounded function. Let $\mathcal{L}$ be a set. Each $\lambda \in \mathcal{L}$ indexes a collection of f-adapted rectangles $V_\lambda = \{\,S_{\lambda,1}, \ldots, S_{\lambda,n_\lambda}, T_{\lambda,1}, \ldots, T_{\lambda,m_\lambda}\,\}$ for some $n_\lambda, m_\lambda \in \mathbb{N}$, with*

$$S^*_\lambda = \bigcup_{i=1}^{n_\lambda} S_{\lambda,i} \subset L, \qquad T^*_\lambda = \bigcup_{j=1}^{m_\lambda} T_{\lambda,j} \subset U_\beta, \qquad V^*_\lambda = S^*_\lambda \cup T^*_\lambda \subset R_\beta.$$

Let $Z^*_\lambda = \overline{R_\beta \setminus V^*_\lambda}$ be the corresponding complementary set. If

$$\inf\{\,\text{area}\ Z^*_\lambda \mid\ \lambda \in \mathcal{L}\,\} = 0, \tag{14.31}$$

then the function $f(x)$ is integrable and, moreover,

$$\int_a^b f(x)\,dx = \text{area}\ L = \sup\{\,\text{area}\ S^*_\lambda \mid\ \lambda \in \mathcal{L}\,\}. \tag{14.32}$$

Intuitively, Theorem 14.12 says that if we are able to cover the graph of f by collections of closed rectangles — namely those forming the sets Z^* — that have arbitrarily small total area, then integrability of the function follows; see the right-most image in Figure 31. Except in very special cases, such as constant functions and step functions, the set $\mathcal{L}$ indexing the collections in Theorem 14.12 is infinite, and typically contains one or more real parameters.

Proof: According to (13.12), (14.27),

$$\text{area}\ R_\beta \geq \text{area}\ L + \text{area}\ U_\beta \geq \text{area}\ S^*_\lambda + \text{area}\ T^*_\lambda = \text{area}\ R_\beta - \text{area}\ Z^*_\lambda.$$

By our assumption, the infimum of the final term over all $\lambda \in \mathcal{L}$ is zero, and hence

$$\text{area}\ R_\beta = \text{area}\ L + \text{area}\ U_\beta = \sup\{\,\text{area}\ S^*_\lambda\,\} + \sup\{\,\text{area}\ T^*_\lambda\,\}.$$

The first equality proves integrability; the second, combined with the facts that

$$\sup\{\,\text{area}\ S^*_\lambda\,\} \leq \text{area}\ L, \qquad \sup\{\,\text{area}\ T^*_\lambda\,\} \leq \text{area}\ U_\beta,$$

establishes the formula (14.32) for the integral. *Q.E.D.*

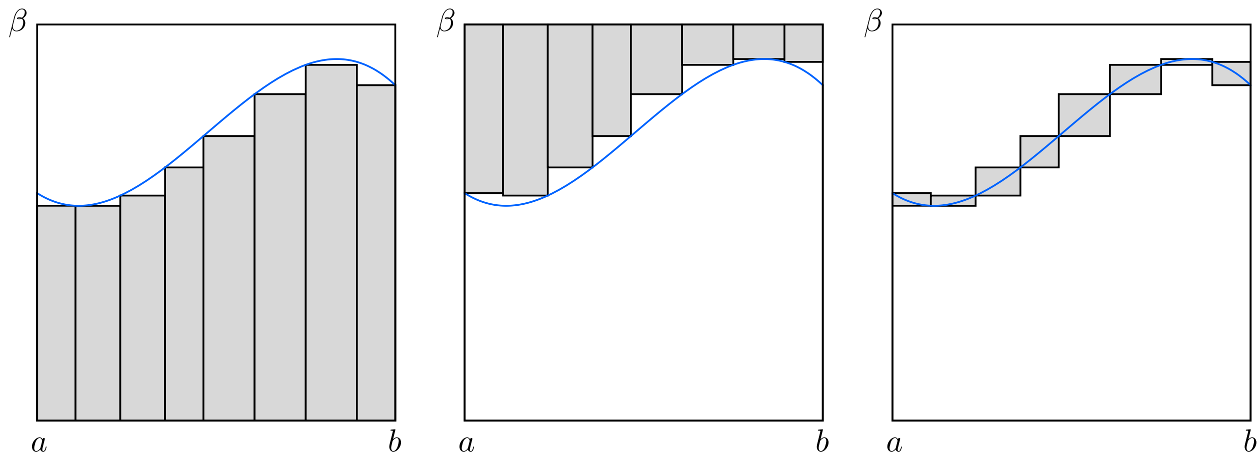


Figure 31. Lower, Upper, and Intermediate Rectangles.

Perhaps the most useful collections of open subrectangles are those constructed as follows. Given $n \in \mathbb{N}$, let us choose a partition

$$a = x_0 < x_1 < x_2 < \ \cdots \ < x_{n-1} < x_n = b \tag{14.33}$$

of the interval $[a, b]$ into subintervals $[x_{i-1}, x_i]$ whose lengths, $x_i - x_{i-1}$, are all deemed to be "small", although we do not impose any a priori restriction on their size. Choose $0 \leq \alpha_i \leq \beta_i \leq \beta$ that are lower and upper bounds for $f(x)$ on the i^{th} subinterval, so

$$\alpha_i \leq f(x) \leq \beta_i \qquad \text{for all} \qquad x_{i-1} \leq x \leq x_i.$$

Define the *upper* and *lower rectangles* associated with the partition and bounds by, respectively,

$$\begin{aligned} S_i &= \overline{R}{}^{x_i,\alpha_i}_{x_{i-1},0} = \left\{ \, (x,y) \;\middle|\; x_{i-1} \leq x \leq x_i, \quad 0 \leq y \leq \alpha_i \, \right\} \subset L, \\ T_i &= \overline{R}{}^{x_i,\beta}_{x_{i-1},\beta_i} = \left\{ \, (x,y) \;\middle|\; x_{i-1} \leq x \leq x_i, \quad \beta_i \leq y \leq \beta \, \right\} \subset U_\beta. \end{aligned} \tag{14.34}$$

These clearly satisfy the properties for forming an f-adapted system of rectangles. In particular,

$$S^* = \bigcup_{i=1}^{n} S_i \subset L, \quad \text{and hence} \quad \text{area}\; S^* = \sum_{i=1}^{n} \alpha_i \, (x_i - x_{i-1}) \leq \text{area}\; L. \tag{14.35}$$

We will therefore call area S^* a *lower sum* for the integral of f. The union of the *intermediate rectangles*

$$Z_i = \overline{R}{}^{x_i,\beta_i}_{x_{i-1},\alpha_i} = \left\{ \, (x,y) \;\middle|\; x_{i-1} \leq x \leq x_i, \quad \alpha_i \leq y \leq \beta_i \, \right\} \tag{14.36}$$

forms the *complementary set*† $Z^* = \displaystyle\bigcup_{i=1}^{n} Z_i$. Since $\alpha_i \leq f(x) \leq \beta_i$ for $x_{i-1} \leq x \leq x_i$, the

† If $\alpha_i = \beta_i$, which implies that $f(x) = \alpha_i = \beta_i$ is constant on the interval $[x_{i-1}, x_i]$, then the

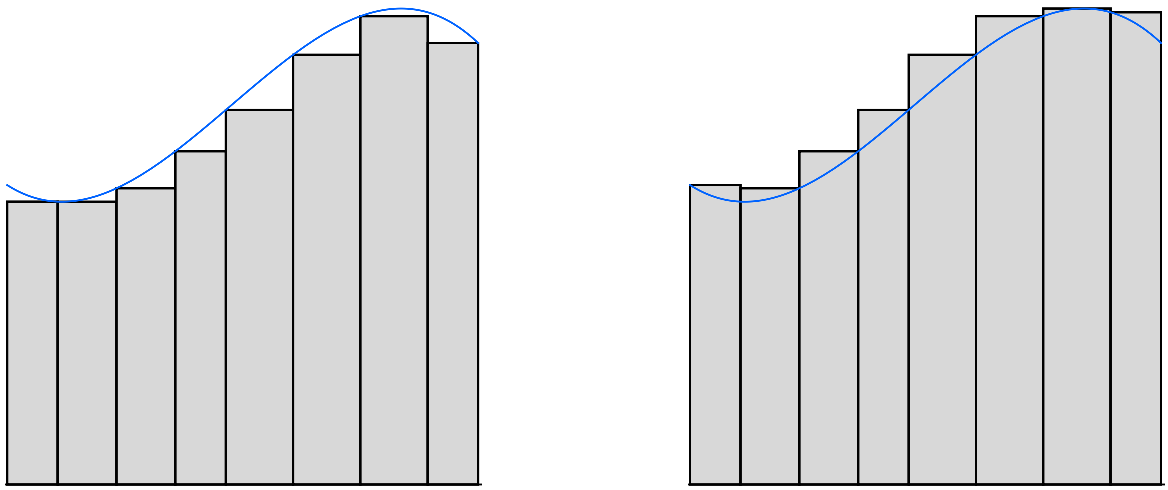


Figure 32. Lower and Upper Sums.

graph of f is contained in the complementary set:

$$G = \{ \, (x, f(x)) \mid \; a \leq x \leq b \, \} \subset Z^*,$$

which thus forms a system of small rectangles covering its graph; see Figure 31.

On the other hand, if we set

$$Y_i = S_i \cup Z_i = \overline{R}^{x_i, \beta_i}_{x_{i-1}, 0},$$

then, since $0 \leq f(x) \leq \beta_i$ when $x_{i-1} \leq x \leq x_i$,

$$Y^* = \bigcup_{i=1}^{n} Y_i \supset L, \quad \text{and hence} \quad \text{area } Y^* = \sum_{i=1}^{n} \beta_i \, (x_i - x_{i-1}) \geq \text{area } L. \tag{14.37}$$

The latter expression is referred to as an *upper sum*; see Figure 32. The area of the lower region L is thus bounded between the lower and upper sums:

$$\text{area } S^* = \sum_{i=1}^{n} \alpha_i \, (x_i - x_{i-1}) \leq \text{area } L \leq \sum_{i=1}^{n} \beta_i \, (x_i - x_{i-1}) = \text{area } Y^* = \text{area } S^* + \text{area } Z^*. \tag{14.38}$$

According to Theorem 14.12, the function $f(x)$ is integrable on $[a, b]$ provided the areas of the complementary sets Z^* have zero infimum over all partitions of $[a, b]$ and all choices of upper and lower bounds on the subintervals:

$$\inf \{ \text{area } Z^* \} = 0 \quad \text{where} \quad \text{area } Z^* = \sum_{i=1}^{n} \text{area } Z_i = \sum_{i=1}^{n} (\beta_i - \alpha_i) \, (x_i - x_{i-1}). \tag{14.39}$$

---

degenerate rectangle Z_i is not included in the complementary region Z^* as defined earlier, (14.26). However, since it has zero area, it can be included here without affecting the overall calculation of areas.

Once integrability is established, the value of the integral is the supremum of the lower sums (14.35) and the infimum of the upper sums (14.37):

$$\int_a^b f(x)\,dx = \sup\left\{\,\sum_{i=1}^n \alpha_i\,(x_i - x_{i-1})\,\right\} = \inf\left\{\,\sum_{i=1}^n \beta_i\,(x_i - x_{i-1})\,\right\}. \tag{14.40}$$

In fact, this condition is both necessary and sufficient for integrability. This follows from the fact that one can subdivide any f adapted system of rectangles to obtain one with the same lower and upper sums as the above. One can even restrict the allowed partitions in some suitable manner. For example, a common choice is to require that they be *regular* meaning that the lengths of the subintervals are the same, and so

$$x_i = a + i\,h, \quad i = 0,\ldots,n, \qquad \text{where} \qquad h = \frac{b-a}{n} \quad \text{for} \quad n \in \mathbb{N}. \tag{14.41}$$

When the value of an integral cannot be computed analytically, it can be approximated numerically through the upper and lower sums. (Although there are much better numerical integration techniques available, [**3**].)

We derived Theorem 14.12 and the integrability criterion (14.40) under the assumption that f is nonnegative, but by the usual device of adding in a constant, they also apply as stated to general functions. Let us use this construction to prove the fundamental results that all monotone and all continuous functions are integrable.

Proof of Theorem 14.4: We prove integrability of monotone increasing functions, leaving the decreasing case as a simple adaptation. Thus we assume $f(x) \le f(y)$ whenever $a \le x \le y \le b$. We use the regular partitions (14.41) to subdivide the interval into $n \in \mathbb{N}$ equally spaced subintervals; we thus view $\mathcal{L} = \mathbb{N}$ as our index set for the integrability criterion (14.31). Because f is increasing on any subinterval, its values at the endpoints provide lower and upper bounds:

$$\alpha_i = f(x_{i-1}) \le f(x) \le f(x_i) = \beta_i, \quad \text{when} \quad x_{i-1} \le x \le x_i.$$

Thus, by regularity, the area of the intermediate region (14.39) is

$$\begin{aligned}\text{area } Z^* &= \sum_{i=1}^n (\beta_i - \alpha_i)\,(x_i - x_{i-1}) = \sum_{i=1}^n \big[\, f(x_i) - f(x_{i-1})\,\big]\,\frac{b-a}{n} \\ &= \Big\{\,\big[\, f(x_1) - f(x_0)\,\big] + \big[\, f(x_2) - f(x_1)\,\big] + \;\cdots\; + \big[\, f(x_n) - f(x_{n-1})\,\big]\,\Big\}\,\frac{b-a}{n} \\ &= \big[\, f(x_n) - f(x_0)\,\big]\,\frac{b-a}{n} = \frac{\big[\, f(b) - f(a)\,\big]\,(b-a)}{n},\end{aligned}$$

where we observe that all the terms in the summation cancel except the first and the last. Since this holds for all $n \in \mathbb{N}$, the infimum of these intermediate areas is zero, and hence the integrability criterion (14.31) is verified. *Q.E.D.*

Proof of Theorem 14.5: Since f is continuous, according to Theorem 5.16, it is bounded: $\alpha \le f(x) \le \beta$ for all $a \le x \le b$. Let $s > 0$. The Uniform Continuity Theorem 10.14 implies that there exists $r > 0$ such that $|\, f(x) - f(y) \,| < s$ whenever $x, y \in [a,b]$ and $|\, x - y \,| < r$. Let us then choose a partition

$$a = x_0 < x_1 < x_2 < \;\cdots\; < x_{n-1} < x_n = b,$$

such that each subinterval has length $0 < x_i - x_{i-1} < r$; for example, we could choose a regular partition (14.41) with $n > (b-a)/r$. Let

$$\alpha_i = \inf \left\{ \, f(x) \;\middle|\; x_{i-1} \le x \le x_i \, \right\}, \qquad \beta_i = \sup \left\{ \, f(x) \;\middle|\; x_{i-1} \le x \le x_i \, \right\}.$$

By the continuity of f, according to Theorem 5.17, there exist $y_i, z_i \in [x_{i-1}, x_i]$ such that $f(y_i) = \alpha_i$, $f(z_i) = \beta_i$. Since

$$|\, y_i - z_i \,| \le x_i - x_{i-1} < r, \qquad \text{we conclude that} \qquad 0 \le \beta_i - \alpha_i = f(z_i) - f(y_i) < s.$$

Thus, in view of (14.36), we can bound the areas of the intermediate rectangles:

$$\operatorname{area} Z_i = (\beta_i - \alpha_i)\,(x_i - x_{i-1}) < s\,(x_i - x_{i-1}),$$

and hence their total area is bounded by

$$\operatorname{area} Z^* = \sum_{i=1}^{n} \operatorname{area} Z_i < s\,(b-a).$$

Since this holds for all $s > 0$, the infimum over all such Z^* is zero, and hence the integrability criterion of Theorem 14.12 is satisfied. *Q.E.D.*

Let us next prove the formula (14.23) for the integral of the sum of two functions.

Proof of Theorem 14.8: We will assume $0 \le f(x) \le \beta$ and $0 \le g(x) \le \delta$, whereby $0 \le f(x) + g(x) \le \beta + \delta$, for all $a \le x \le b$, the general case following by our usual device of adding in a constant to both functions. Given a partition

$$a = x_0 < x_1 < x_2 < \;\cdots\; < x_{n-1} < x_n = b,$$

choose $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n, \gamma_1, \ldots, \gamma_n, \delta_1, \ldots, \delta_n$ such that

$$0 \le \alpha_i \le f(x) \le \beta_i \le \beta, \qquad 0 \le \gamma_i \le g(x) \le \delta_i \le \delta, \quad \text{for all} \quad x_{i-1} \le x \le x_i.$$

This implies

$$0 \le \alpha_i + \gamma_i \le f(x) + g(x) \le \beta_i + \delta_i \le \beta + \delta, \qquad \text{for all} \quad x_{i-1} \le x \le x_i. \tag{14.42}$$

The corresponding intermediate rectangles for f and g are, respectively,

$$Z_i = \overline{R}^{x_i, \beta_i}_{x_{i-1}, \alpha_i}, \qquad \widehat{Z}_i = R^{x_i, \delta_i}_{x_{i-1}, \gamma_i},$$

Since we are assuming integrability of f and g, their combined areas

$$\operatorname{area} Z^* = \operatorname{area} \bigcup_{i=1}^{n} Z_i, \qquad \operatorname{area} \widehat{Z}^* = \operatorname{area} \bigcup_{i=1}^{n} \widehat{Z}_i,$$

both have zero infimum as we vary over all partitions and upper and lower bounds. Thus, in view of (14.42), the "summed" intermediate rectangles

$$\widetilde{Z}_i = \overline{R}^{x_i, \beta_i + \delta_i}_{x_{i-1}, \alpha_i + \gamma_i}$$

play the same role for $f(x)+g(x)$, and, moreover, area $\widetilde{Z}_i =$ area $Z_i +$ area $\widehat{Z}_i$. We conclude that their combined areas

$$\text{area } \widetilde{Z}^* = \text{area } \bigcup_{i=1}^{n} \widetilde{Z}_i = \text{area } Z^* + \text{area } \widehat{Z}^*,$$

also have zero infimum, thus proving the integrability of $f(x)+g(x)$.

Finally, to prove the integration formula (14.23), we note that the corresponding lower sums for, respectively, $f+g, f, g$, also add up:

$$\sum_{i=1}^{n} (\alpha_i + \gamma_i)\,(x_i - x_{i-1}) = \sum_{i=1}^{n} \alpha_i\,(x_i - x_{i-1}) + \sum_{i=1}^{n} \gamma_i\,(x_i - x_{i-1}),$$

which implies that the same holds for their integrals, thus proving (14.23). *Q.E.D.*

Finally, with some additional work, one can replace the inequality in (14.16) by a strict inequality.

Theorem 14.13. *Let $a < b$. If f, g are integrable and $f(x) < g(x)$ for all $x \in [a,b]$, then*

$$\int_a^b f(x)\,dx < \int_a^b g(x)\,dx. \tag{14.43}$$

Proof: We can use (14.24) to rewrite (14.43) as

$$\int_a^b \big[\, g(x) - f(x) \,\big]\,dx > 0 \quad \text{provided} \quad g(x) - f(x) > 0 \quad \text{for all} \quad a \le x \le b.$$

Thus it suffices to prove that if f is integrable and $f(x) > 0$ for all $a \le x \le b$, then $\displaystyle\int_a^b f(x)\,dx > 0$. The proof of that the integral is strictly positive is easy if f is continuous. Indeed, set $\alpha = \inf f[a,b]$. Theorem 5.17 tells us that there exists $x_* \in [a,b]$ such that $f(x_*) = \alpha$. This implies $\alpha > 0$, and hence, using (14.17), $\displaystyle\int_a^b f(x)\,dx \ge \alpha\,(b-a) > 0$.

If f is merely integrable, we argue by contradiction. Thus, suppose

$$f(x) > 0 \quad \text{for all} \quad a \le x \le b, \quad \text{but} \quad \int_a^b f(x)\,dx = 0.$$

We show that this implies, given any $r > 0$, the existence of a subinterval $I_1 = [a_1, b_1] \subset I = [a,b]$ with $a \le a_1 < b_1 \le b$ such that $0 < f(x) < r$ for all $x \in I_1$. Given this, we complete the proof by first choosing $I_1 = [a_1, b_1]$ so that $0 < f(x) < 1$ for $a_1 \le x \le b_1$. Observe that $\displaystyle\int_{a_1}^{b_1} f(x)\,dx = 0$ also, because if it were positive, Proposition 14.7 would imply that the integral over the entire interval $[a,b]$ must also be positive. Thus, by adapting the preceding claim to the interval I_1, we can find $I_2 = [a_2, b_2] \subset I_1$ so that $0 < f(x) < \frac{1}{2}$ for $x \in I_2$. Proceeding by induction, we construct a decreasing system of intervals $I \supset I_1 \supset I_2 \supset I_3 \supset \cdots$ such that $0 < f(x) < 1/k$ for $x \in I_k$. The endpoints of these intervals satisfy $a \le a_1 \le a_2 \le a_3 \le \cdots \le b_3 \le b_2 \le b_1 \le b$. Thus, $x_* = \sup\{ a_k \}$ satisfies $a_k \le x_* \le b_k$, and hence $x_* \in I_k$ for all $k \in \mathbb{N}$. But this implies $f(x_*) \le 1/k$ for all k, and hence $f(x_*) \le 0$, in contradiction to our assumption that f is everywhere positive.

To prove the initial claim, condition (14.40) implies that the upper sums for the integral of f on $[a,b]$ have zero infimum:

$$\inf \left\{ \sum_{i=1}^{n} \beta_i \, (x_i - x_{i-1}) \right\} = 0. \tag{14.44}$$

Thus, for every $r > 0$ we can find a partition and a set of upper bounds $\beta_i > 0$ such that the corresponding upper sum is strictly less than $r\,(b-a)$. Let $\beta_* = \min\{\beta_1, \ldots, \beta_n\}$ be the minimum of these upper bounds. Then

$$r\,(b-a) > \sum_{i=1}^{n} \beta_i \, (x_i - x_{i-1}) \ge \beta_* \sum_{i=1}^{n} (x_i - x_{i-1}) = \beta_* \, (b-a), \quad \text{and hence} \quad \beta_* < r.$$

We thus can set $I_1 = [x_{i-1}, x_i]$ to be the interval where $\beta_i = \beta_*$. (If there are several with this property, choose any one.) Then, by the above inequality, $f(x) \le \beta_i = \beta_* < r$ whenever $x \in I_1$. *Q.E.D.*

15. Indefinite Integration and the Fundamental Theorem of Calculus.

The next step is to look at what happens as we vary the endpoints of the interval of integration. Let us replace the upper limit of integration by a variable point x. The resulting function

$$I(x) = \int_a^x f(t)\,dt \tag{15.1}$$

is known as the *indefinite integral* of $f(x)$ since its upper endpoint x is no longer viewed as being fixed. As noted earlier, we should use a different "dummy" variable t inside the integral sign so as not to confuse the integration variable with the limits of integration.

The key result is the continuity of the indefinite integral, irrespective of whether of not the function being integrated is continuous.

Theorem 15.1. *Let $a \le c \le b$. If $f(x)$ is bounded and integrable on $[a,b]$, then its indefinite integral*

$$I(x) = \int_c^x f(t)\,dt$$

is a continuous function of $x \in [a,b]$.

Proof: If $a \le x, y \le b$, then, using (14.19, 22),

$$I(x) - I(y) = \int_c^x f(t)\,dt - \int_c^y f(t)\,dt = \int_y^x f(t)\,dt.$$

Thus, given a bound $|\, f(x)\,| \leq \beta$, by (14.17, 18),

$$|\, I(x) - I(y)\,| = \left| \int_y^x f(t)\, dt \right| \leq \int_y^x |\, f(t)\,|\, dt \leq \int_y^x \beta\, dt = \beta\, |\, y - x\,|.$$

Thus, $I(x)$ satisfies the Lipschitz continuity condition (4.2) with $\lambda = \beta$, and hence Theorem 4.4 implies that I is continuous. *Q.E.D.*

Proposition 15.2. *If $f(x)$ is integrable and > 0 then its indefinite integral* (15.1) *is a strictly increasing monotone function of x.*

Proof: If $x < y$, then, using (14.43),

$$I(y) - I(x) = \int_a^y f(t)\, dt - \int_a^x f(t)\, dt = \int_x^y f(t)\, dt > 0. \qquad Q.E.D.$$

Remark: The conclusion of Proposition 15.2 remains valid even when $f(x)$ has isolated zeros. For example,

$$\int_0^x t^2\, dt = \tfrac{1}{3}\, x^3$$

is strictly monotone. On the other hand, if $f(x) \equiv 0$ on some open subinterval, then its integral is no longer strictly monotone.

Finally, let us state and prove the Fundamental Theorem of Calculus, which is the remarkable fact that differentiation is the opposite of integration. It was this amazing discovery that sparked the calculus revolution in mathematics.

Theorem 15.3. *Suppose $f: \mathbb{R} \to \mathbb{R}$ is continuous. Then its indefinite integral*

$$I(x) = \int_a^x f(t)\, dt.$$

is differentiable and its derivative is the original function: $I'(x) = f(x)$.

Proof: As in Definition 7.3, given $b \in \mathbb{R}$, the differentiability of $I(x)$ at $x = b$ follows from the continuity of the associated difference quotient function

$$Q(x) = \begin{cases} \dfrac{I(x) - I(b)}{x - b}, & x \neq b, \\ f(b), & x = b, \end{cases}$$

whereby $I'(b) = f(b)$. If we can show that this holds for all b, the Fundamental Theorem follows.

To prove continuity, note first that, when $x \neq b$,

$$Q(x) = \frac{I(x) - I(b)}{x - b} = \frac{1}{x - b} \left(\int_a^x f(t)\, dt - \int_a^b f(t)\, dt \right) = \frac{1}{x - b} \int_b^x f(t)\, dt,$$

where we applied (14.19, 22). We further note that, again using (14.22),

$$Q(x) = \frac{1}{x - b} \int_b^x f(t)\, dt = -\, \frac{1}{x - b} \int_x^b f(t)\, dt = \frac{1}{b - x} \int_x^b f(t)\, dt. \tag{15.2}$$

Now, suppose $\alpha < f(b) < \beta$. Since f is continuous and $b \in f^{-1}(\alpha, \beta)$, there exists an open interval (c, d) containing b such that $(c, d) \subset f^{-1}(\alpha, \beta)$. Thus, if $c < x < d$, then $\alpha < f(x) < \beta$. Thus, as a consequence of Theorem 14.13, if $b < x < d$,

$$\alpha\,(x-b) < \int_b^x f(t)\,dt < \beta\,(x-b),$$

whereas if $c < x < b$, then

$$\alpha\,(b-x) < \int_x^b f(t)\,dt < \beta\,(b-x).$$

In view of (15.2), this implies $\alpha < Q(x) < \beta$ whenever $c < x < d$, including $x = b$. We deduce that $(c, d) \subset Q^{-1}(\alpha, \beta)$, which implies that the latter set is open. Since this holds for all such intervals, we deduce that Q is continuous, which establishes the Fundamental Theorem. *Q.E.D.*

Thus, to integrate a function $f(x)$, if we know an *antiderivative*, meaning a function $F(x)$ such that $F'(x) = f(x)$, then

$$\int_a^b f(x)\,dx = F(b) - F(a). \tag{15.3}$$

Indeed, F and I have the same derivative,

$$F'(x) - I'(x) = f(x) - f(x) \equiv 0,$$

and hence Corollary 7.15 implies that

$$F(x) - I(x) = c$$

is constant. Since $I(a) = 0$, we have $F(a) = c$ and hence

$$\int_a^x f(t)\,dt = I(x) = F(x) - c = F(x) - F(a).$$

Setting $x = b$ proves (15.3).

Example 15.4. We already know that if

$$F(x) = x^{n+1} \qquad \text{then} \qquad F'(x) = (n+1)\,x^n. \tag{15.4}$$

Thus, using (15.3), we can evaluate

$$\int_a^b x^n\,dx = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}, \tag{15.5}$$

where we divided both sides by $n + 1$. The case $n = 0$ reproduces our formula (14.6) for the integral of the constant function $x^0 = 1$, while the case $n = 1$ reproduces our formula (14.7) for the integral of $x^1 = x$, namely the area of the trapezoidal region under its graph. Formula (15.4) is also valid when $n < 0$, and hence so is the integration formula (15.5) provided $n \neq -1$.

The one case not covered by the preceding Example 15.4 is when $n = -1$. Indeed, there is no algebraic function $f(x)$ whose derivative is $1/x$. Its indefinite integral will define the *natural logarithm*

$$\log x = \int_1^x \frac{dx}{x} \qquad \text{for} \qquad x > 0. \tag{15.6}$$

We will usually drop the qualifier "natural" since it is essentially the only logarithm we discuss in these notes. Theorems 15.1 and 15.3 imply that, when $x > 0$, the logarithm is a continuous function whose derivative $\log'(x) = 1/x$. Corollary 7.15 says that the only other functions with this property are $\log x + c$ for a constant $c \in \mathbb{R}$. Proposition 15.2 implies that $\log x$ is a strictly increasing function on $\mathbb{R}^+$. Since $\log 1 = 0$, and $1/x > 0$ for $x > 0$, we have $\log x > 0$ when $x > 1$, while $\log x < 0$ when $x < 1$.

It can be shown that $\log x$ is unbounded, from both above and below, on the entire positive real line $\mathbb{R}^+$. Indeed, consider the region

$$L_\infty = \{\, (x,y) \mid \ x > 1,\ 0 < y < 1/x \,\}.$$

It contains the essentially disjoint rectangles

$$R_n = \overline{R}{}_{n-1,0}^{n,1/n} = \{\, (x,y) \mid \ n-1 \le x \le n, \ \ 0 \le y \le 1/n \,\} \subset L_\infty \quad \text{for} \quad n \in \mathbb{N}.$$

The combined area of these rectangles is

$$\sum_{n=1}^\infty \text{area } R_n = \sum_{n=1}^\infty \frac{1}{n} = \infty,$$

being the well known divergent harmonic series, [**28**]. Since

$$\int_1^\infty \frac{dx}{x} = \text{area } L_\infty \ge \sum_{n=1}^\infty \text{area } R_n,$$

this proves that $\log x$ is unbounded from above as $x \to \infty$. Similarly, the region

$$\widehat{L}_\infty = \{\, (x,y) \mid \ 0 < x < 1,\ 0 < y < 1/x \,\}$$

contains the essentially disjoint rectangles

$$\widehat{R}_n = \overline{R}{}_{1/n,0}^{1/(n-1),n-1} = \{\, (x,y) \mid \ 1/n \le x \le 1/(n-1), \ \ 0 \le y \le n-1 \,\} \subset \widehat{L}_\infty \quad \text{for} \quad n \in \mathbb{N},$$

that have combined area

$$\sum_{n=1}^\infty \text{area } \widehat{R}_n = \sum_{n=1}^\infty (n-1) \left(\frac{1}{n-1} - \frac{1}{n} \right) = \sum_{n=1}^\infty \frac{1}{n} = \infty.$$

Thus,

$$\int_0^1 \frac{dx}{x} = \text{area } \widehat{L}_\infty \ge \sum_{n=1}^\infty \text{area } \widehat{R}_n,$$

which shows that $\log x$ is not bounded from below as $x \to 0^+$.

The inverse function of the natural logarithm is the *exponential function* $\exp(x) = e^x$, which satisfies $\exp(x) = \log^{-1}(x)$. Thus,

$$e^{\log x} = x \qquad \text{or, equivalently,} \qquad \log(e^x) = x.$$

In other words,

$$y = e^x \qquad \text{if and only if} \qquad x = \log y.$$

In particular, the base e of the exponential is defined by†

$$y = e = e^1 \approx 2.71287\ \ldots\ , \qquad \text{which satisfies} \qquad 1 = \log e = \int_1^e \frac{dx}{x}.$$

In other words, $e > 1$ is the real number with the property that the area of the region lying below the graph of $1/x$ between $1 \le x \le e$ is exactly 1. Since $\log\colon \mathbb{R}^+ \to \mathbb{R}^+$ is continuous and strictly increasing, the same is true for $\exp\colon \mathbb{R}^+ \to \mathbb{R}^+$.

Let us compute the derivative of the exponential. Using (7.14), since $\log'(y) = 1/y$, we have

$$\exp'(x) = \frac{1}{\log'(\exp(x))} = \exp(x). \tag{15.7}$$

Thus, the exponential function is its own derivative. In fact, we can completely characterize such functions.

Proposition 15.5. *If $f'(x) = f(x)$ then $f(x) = c\, e^x$ for some constant $c \in \mathbb{R}$.*

Proof: Suppose first that $f(x) > 0$. Define $g(x) = \log f(x)$. Then, applying the Chain Rule (7.11),

$$g'(x) = f'(x)\, \log'\big[\, f(x)\, \big] = \frac{1}{f(x)}\, f(x) = 1.$$

Thus, $g(x) = x + a$ for some constant $a \in \mathbb{R}$, and hence

$$f(x) = e^{g(x)} = e^{x+a} = c\, e^x, \qquad \text{where} \qquad c = e^a > 0.$$

If $f(x) < 0$, then the same proof using $g(x) = \log\big[-f(x) \big]$ implies $f(x) = c\, e^x$ for some $c < 0$. Finally, if $f(a) = 0$ for some $a \in \mathbb{R}$, then we claim that $f(x) \equiv 0$. Indeed, if $f(b) \ne 0$, then, by the preceding argument, $f(x) = c\, e^x$ for some $c \ne 0$. But then $f(x) \ne 0$ for all $x \in \mathbb{R}$, contradicting our original assumption. *Q.E.D.*

We close this section with the Mean Value Theorem for integrals.

.

Proposition 15.6. *Let $a < b$ and suppose $f\colon [a, b] \to \mathbb{R}$ is continuous. Then there exists $a < c < b$ such that*

$$\int_a^b f(x)\, dx = f(c)\, (b - a). \tag{15.8}$$

† The decimal expansion of e can be found by numerical integration.

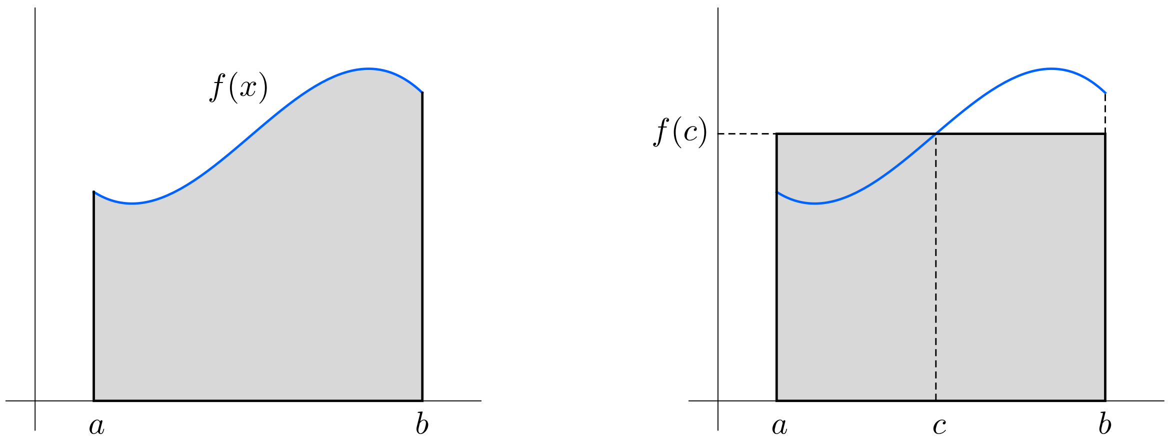


Figure 33. The Mean Value Theorem for Integrals.

The proof is an immediate application of the Mean Value Theorem 7.13 to the indefinite integral $I(x)$, using the Fundamental Theorem 15.3 to evaluate its derivative. Equation (15.8) can be interpreted as saying that the area under the graph of f equals the area of a rectangle of width $f(c)$ for some $a < c < b$.

16. Volume.

We now turn to the integration of a function $F(x,y)$ depending on two variables, leading to the notion of double integral, which, when F is non-negative, will measure the volume of the solid region lying under its graph. Thus the first step is to define what is meant by volume of a subset $A \subset \mathbb{R}^3$. We work in direct analogy with our earlier treatment of area.

We begin with the analog of a bounded closed rectangle, which is a *closed rectangular solid* or, for short, a (closed) *box*

$$\overline{B}_{a,c,e}^{b,d,f} = \left\{ (x,y,z) \in \mathbb{R}^3 \;\middle|\; a \le x \le b,\ c \le y \le d,\ e \le z \le f \right\}. \tag{16.1}$$

The volume of the box (16.1) is the product of its length, width, and height:

$$\operatorname{vol} \overline{B}_{a,c,e}^{b,d,f} = (b-a)\,(d-c)\,(f-e). \tag{16.2}$$

We also note that its volume is the product of its height, which is $f - e$, times the area of its projection onto the xy plane, which is the closed rectangle

$$\overline{R}_{a,c}^{b,d} = \left\{ (x,y,0) \in \mathbb{R}^3 \;\middle|\; a \le x \le b,\ c \le y \le d \right\}, \quad \text{with area } \overline{R}_{a,c}^{b,d} = (b-a)\,(d-c). \tag{16.3}$$

In particular, the volume of $\overline{B}_{a,c,e}^{b,d,f} > 0$ if and only if $a < b$, $c < d$, $e < f$, in which case its *interior* is the nonempty *open box*

$$B_{a,c,e}^{b,d,f} = \left\{ (x,y,z) \in \mathbb{R}^3 \;\middle|\; a < x < b,\ c < y < d,\ e < z < f \right\}. \tag{16.4}$$

On the other hand, if $a = b$ and/or $c = d$ and/or $e = f$, the box collapses to either a closed rectangle, or a closed line segment, or a point, each of which has zero volume and empty interior.

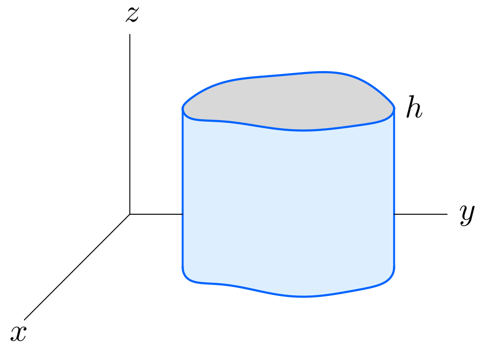


Figure 34. Generalized Cylinder.

The boundary of the box (16.1) consists of 6 closed rectangles, obtained by setting $x = a$ or $x = b$ or $y = c$ or $y = d$ or $z = e$ of $z = f$, with zero volume. A *half open box* consists of the union of an open box and a part of its boundary, typically one or more of its rectangular sides. Two boxes are said to be *essentially disjoint* if their intersection is either empty or is contained in one of their boundary rectangles and hence of zero volume. The volume of an essentially disjoint union of boxes equals the sum of their individual volumes. With this in hand, the general definition of the volume of a subset of $\mathbb{R}^3$ follows.

Definition 16.1. Let $A \subset \mathbb{R}^3$. Given any essentially disjoint collection of boxes $B_1, \ldots, B_n$, their combined volume equals $\operatorname{vol} B_1 + \cdots + \operatorname{vol} B_n$, and we let $V_A \subset [0, \infty)$ be the set of all possible such combined volumes. We then define the *volume* of A as the associated least upper bound:

$$\operatorname{vol} A = \begin{cases} 0, & A = \varnothing, \\ \sup V_A, & V_A \subset \mathbb{R}^+ \text{ is bounded}, \\ \infty, & V_A \subset \mathbb{R}^+ \text{ is unbounded}. \end{cases} \tag{16.5}$$

Thus, A has zero volume if and only if it contains no open boxes. If $A \subset B$ then $\operatorname{vol} A \leq \operatorname{vol} B$. Thus, if A is bounded, meaning that $A \subset B$ is contained in some bounded box B, then its volume is finite, $0 \leq \operatorname{vol} A \leq \operatorname{vol} B < \infty$. There are unbounded sets with finite volume, including planes and lines that have zero volume. Further,

$$\operatorname{vol} A + \operatorname{vol} B \leq \operatorname{vol}(A \cup B) \qquad \text{when} \qquad \operatorname{vol}(A \cap B) = 0. \tag{16.6}$$

The proof of this inequality mimics that of the corresponding statement (13.12) for areas. We call A and B *volume compatible* if the above inequality is an equality. In particular, as in Lemma 13.4, if $A \subset \{x \leq a\}$ and $B \subset \{x \geq a\}$, then they are volume compatible. One can replace x by y or z and the result remains valid.

The proof that the preceding definition of volume is consistent, i.e., the volume of a box equals the least upper bound of the volumes of any essentially disjoint collection of sub-boxes, proceeds via subdivision in analogy with the areas of rectangles; details are left to the reader. In particular, closed, open, and half open boxes having the same boundary have the same volume. Another useful property is that the volume of a (generalized) *cylinder*[†]

$$C = \{\, (x,y,z) \mid (x,y) \in D, \quad 0 \le z \le h \,\} \tag{16.7}$$

of height h lying over a planar region $D \subset \mathbb{R}^2$, as sketched in Figure 34, is equal to the area of its base times its height:

$$\operatorname{vol} C = h \ \operatorname{area} D. \tag{16.8}$$

Before proving this, we first note that we can identify the cylinder as a Cartesian product of D with a closed interval of length h, so $C = D \times [0,h]$. In general, mimicking the planar case (9.3), given subsets $D \subset \mathbb{R}^2$ and $U \subset \mathbb{R}$, define their *Cartesian product* to be

$$D \times U = \{\, (x,y,z) \mid (x,y) \in D,\ z \in U \,\} \subset \mathbb{R}^3. \tag{16.9}$$

The first set might itself be a Cartesian product, $D = S \times T$ for $S, T \subset \mathbb{R}$, in which case $D \times U = S \times T \times U$ is a *triple Cartesian product*. For example, a box is the Cartesian product of a closed rectangle and a closed interval, which is turn is the Cartesian product of three closed intervals:

$$\overline{B}^{b,d,f}_{a,c,e} = \overline{R}^{b,d}_{a,c} \times [e,f] = [a,b] \times [c,d] \times [e,f]. \tag{16.10}$$

To prove the formula (16.8) for the cylinder's volume, given any collection $R_1, \ldots, R_n$ of essentially disjoint closed rectangles $R_i \subset D$, we can form a corresponding essentially disjoint collection of boxes $B_i = R_i \times [0,h] \subset C$ of height h. Using (16.3), the volume of each B_i is equal to h times the area of the corresponding R_i. Since the supremum of the total areas over all possible choices of rectangles equals the area of D, the same result holds for the corresponding boxes:

$$\sup \left\{ \sum_{i=1}^{n} \operatorname{vol} B_i \right\} = h \ \sup \left\{ \sum_{i=1}^{n} \operatorname{area} R_i \right\} = h \ \operatorname{area} D.$$

Technically, we should further check that no other disjoint collection of boxes contained in C can have a larger volume, but one can always use subdivision to produce a collection of the above form with the same or larger volume, and hence the supremum is as advertised.

As with areas, volumes are unchanged under certain basic transformations of $\mathbb{R}^3$, including translations, reflections, rotations, and shears. Scaling in one of the coordinate directions multiplies the volume by the absolute value of the scaling factor. One proves these results by simply establishing them for boxes, and the details are again assigned to the reader.

17. Double Integrals.

The basic theory of integration for functions of several variables proceeds in direct analogy with the univariate case developed in Section 14, replacing area by volume throughout.

[†] If $D \subset \mathbb{R}^2$ is a circular disk, then C is a circular cylinder.

Integration of a positive function $F: D \to \mathbb{R}^+$ over a closed, bounded region $D \subset \mathbb{R}^2$ will measure the volume of the solid region lying between the graph of F and the xy plane $\{ z = 0 \} = \{ (x, y, 0) \mid x, y \in \mathbb{R} \} \subset \mathbb{R}^3$.

Let $D \subset \mathbb{R}^2$ be a bounded closed subset. Suppose $F: D \to [0, \beta]$ is a bounded non-negative function, so $0 \leq F(x, y) \leq \beta$ for all $(x, y) \in D$. Define the *lower* and *upper regions*

$$\begin{aligned} L &= \{ (x, y, z) \mid (x, y) \in D, \quad 0 \leq z \leq F(x, y) \}, \\ U_\beta &= \{ (x, y, z) \mid (x, y) \in D, \quad F(x, y) \leq z \leq \beta \}. \end{aligned} \tag{17.1}$$

Their union

$$L \cup U_\beta = C_\beta = \{ (x, y, z) \mid (x, y) \in D, \quad 0 \leq z \leq \beta \} \tag{17.2}$$

is the cylinder of height β with base D. On the other hand,

$$L \cap U_\beta = G = \{ (x, y, F(x, y)) \mid (x, y) \in D \} \tag{17.3}$$

is a surface representing the graph of F over $D \subset \mathbb{R}^2$, which has zero volume because it contains no boxes with nonzero volume.

Definition 17.1. Let $D \subset \mathbb{R}^2$ be a bounded closed subset. A bounded nonnegative function $F: D \to [0, \beta]$ is *integrable* if its lower and upper regions are volume compatible, meaning

$$\operatorname{vol} L + \operatorname{vol} U_\beta = \operatorname{vol} C_\beta = \beta \text{ area } D. \tag{17.4}$$

In this case, we define the *double integral* of F on D to be the volume of its lower region:

$$\iint_D F(x, y)\, dx\, dy = \operatorname{vol} L = \operatorname{vol} \{ (x, y, z) \mid (x, y) \in D, \ 0 \leq z \leq F(x, y) \}. \tag{17.5}$$

In particular, if $F(x, y) \equiv c \geq 0$ is constant, then its lower region is the cylinder $D \times [0, c]$, and hence

$$\iint_D c\ dx\, dy = c \text{ area } D.$$

The confirmation that constant functions satisfy the integrability criterion (17.4) for any choice of $\beta \geq c$ is left to the reader. In particular, one can compute the area of a domain by integrating the constant function 1 over it:

$$\text{area } D = \iint_D \ dx\, dy. \tag{17.6}$$

As in the single variable case, one extends double integration to functions that are not everywhere positive by adding in a constant. Specifically, if $F(x, y) \geq c$ for some $c \in \mathbb{R}$ and all $(x, y) \in \overline{D}$, then $\widetilde{F}(x, y) = F(x, y) - c \geq 0$, and hence, assuming $\widetilde{F}$ is integrable,

$$\iint_D F(x, y)\, dx\, dy \ = \iint_D \left[\widetilde{F}(x, y) + c \right] dx\, dy \ = \iint_D \widetilde{F}(x, y)\, dx\, dy + c \text{ area } D. \tag{17.7}$$

It is not difficult to show that this gives a consistent value for the double integral of F, independent of the choice of the lower bound α. In general, the double integral counts volumes of subregions lying above the $x\,y$ plane with a positive sign, and those lying below it with a negative sign.

The basic properties of double integrals parallel those of single integrals. The proofs are analogous, and thus left to the motivated reader.

- If
$$\text{area } D = 0, \qquad \text{then} \qquad \iint_D F(x,y)\,dx\,dy = 0, \tag{17.8}$$
for any function $F(x,y)$. This is because there are no open rectangles contained in D and hence no open boxes contained in either the lower or upper regions, which thus both have zero volume. More generally,
$$\text{if } F(x,y) = 0 \text{ for } (x,y) \in A \setminus D \text{ and area } D = 0, \text{ then } \iint_A F(x,y)\,dx\,dy = 0. \tag{17.9}$$
- If $F(x,y) \leq G(x,y)$ or $F(x,y) < G(x,y)$ for all $(x,y) \in D$, then the corresponding double integral inequality holds:
$$\iint_D F(x,y)\,dx\,dy \;\leq \iint_D G(x,y)\,dx\,dy, \qquad \iint_D F(x,y)\,dx\,dy \;< \iint_D G(x,y)\,dx\,dy. \tag{17.10}$$
As with single integrals, the first result is immediate, while the second requires work.
- The double integral of the sum of two functions is the sum of the individual double integrals:
$$\iint_D \big[\,F(x,y) + G(x,y)\,\big]\,dx\,dy \;= \iint_D F(x,y)\,dx\,dy \;+ \iint_D G(x,y)\,dx\,dy. \tag{17.11}$$
- Multiplying the function by a constant $c \in \mathbb{R}$ multiplies the double integral by the same constant:
$$\iint_D \big[\,c\,F(x,y)\,\big]\,dx\,dy = c \iint_D F(x,y)\,dx\,dy. \tag{17.12}$$
- If $D \subset E$, and $F(x,y) \geq 0$ for all $(x,y) \in E$, then
$$\iint_D F(x,y)\,dx\,dy \;\leq \iint_E F(x,y)\,dx\,dy. \tag{17.13}$$
- The integral over the union $D \cup E$ of two regions whose intersection has zero volume, $\mathrm{vol}(D \cap E) = 0$, is the sum of the individual double integrals:
$$\iint_{D\cup E} F(x,y)\,dx\,dy \;= \iint_D F(x,y)\,dx\,dy \;+ \iint_E F(x,y)\,dx\,dy. \tag{17.14}$$

For example, if we split the unit disk
$$D = \{\,x^2 + y^2 \leq 1\,\} \qquad \text{into two half disks} \qquad \begin{aligned} H^+ &= \{\,x^2 + y^2 \leq 1,\ y \geq 0\,\}, \\ H^- &= \{\,x^2 + y^2 \leq 1,\ y \geq 0\,\}, \end{aligned}$$
then
$$\iint_D F(x,y)\,dx\,dy \;= \iint_{H^+} F(x,y)\,dx\,dy \;+ \iint_{H^-} F(x,y)\,dx\,dy.$$

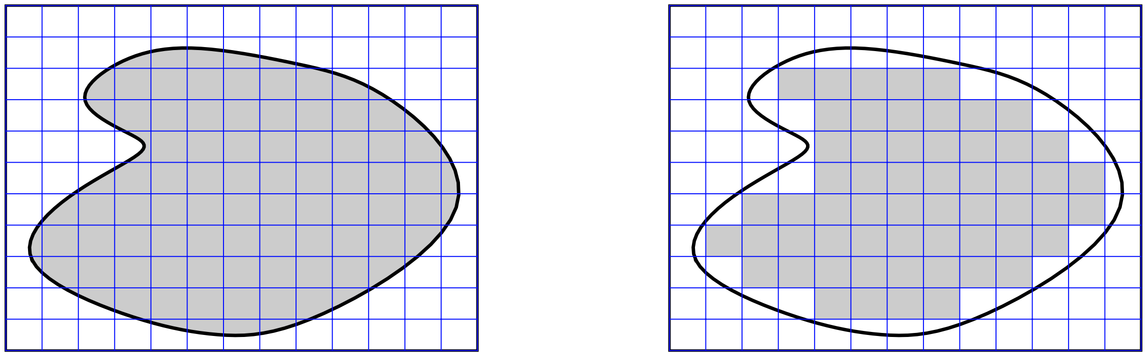


Figure 35. Partition of Integration Domain into Rectangles.

In this case, their intersection $D = H^+ \cap H^- = \{ (x,0) \mid -1 < x < 1 \}$ is an open line segment with zero area.

Let us establish an integrability criterion for a bounded non-negative function defined on a closed bounded domain $D \subset \mathbb{R}^2$, so that $F \colon D \to [0, \beta]$. More general functions are treated by the usual device of adding in a suitably large positive constant, using (17.7). There are analogs the general integrability criteria of Theorems 14.10 and 14.12 that the reader is invited to formulate. Instead, we skip directly to the analog of the integrability criterion (14.40) based on lower and upper sums.

Let $R_1, \ldots, R_n \subset D$ be an essentially disjoint collection of closed rectangles. We choose upper and lower bounds on each rectangle, so

$$0 \leq \alpha_m \leq F(x,y) \leq \beta_m \leq \beta \qquad \text{for} \qquad (x,y) \in R_m.$$

The corresponding lower, upper, and intermediate boxes are thus given by

$$S_m = R_m \times [0, \alpha_m] \subset L, \qquad T_m = R_m \times [\beta_m, \beta] \subset U_\beta, \qquad Z_m = R_m \times [\alpha_m, \beta_m]. \tag{17.15}$$

As in Theorem 14.12, the function F is integrable over D if and only if the total volumes of the intermediate boxes has zero infimum:

$$0 = \inf \left\{ \operatorname{vol} \left(\bigcup_{m=1}^{n} Z_m \right) \right\} = \inf \left\{ \sum_{m=1}^{n} \operatorname{vol} Z_m \right\} = \inf \left\{ \sum_{m=1}^{n} (\beta_m - \alpha_m) \operatorname{area} R_m \right\}. \tag{17.16}$$

In this case, the double integral equals the supremum of the corresponding *lower sums*, which are the total volumes of the boxes S_m, or, equivalently, the infimum of the *upper sums*, which are the total volumes of the boxes $Y_m = S_m \cup Z_m$:

$$\iint_D F(x,y)\, dx\, dy = \sup \left\{ \sum_{m=1}^{n} \alpha_m \operatorname{area} R_m \right\} = \inf \left\{ \sum_{m=1}^{n} \beta_m \operatorname{area} R_m \right\}. \tag{17.17}$$

Assuming $D \subset \overline{R}^{b,d}_{a,c}$, the simplest system of rectangles comes from choosing partitions

$$a = x_0 < x_1 < x_2 < \ \cdots \ < x_{k-1} < x_k = b, \qquad c = y_0 < y_1 < y_2 < \ \cdots \ < y_{l-1} < y_l = d, \tag{17.18}$$

and letting $\mathcal{R} = \{R_1, \ldots, R_n\}$ contain all the associated rectangles $R_{i,j}$ that satisfy

$$R_{i,j} = R^{x_i, y_j}_{x_{i-1}, y_{j-1}} \subset D \qquad \text{for} \qquad 1 \leq i \leq k, \quad 1 \leq j \leq l. \tag{17.19}$$

See Figure 35; the integration domain D is shaded in the left hand plot, and its partition into subrectangles is on the right.

The proof that any continuous function is integrable follows that of Theorem 14.5 using the Uniform Continuity Theorem 10.15.

Theorem 17.2. *If $D \subset \mathbb{R}^2$ is closed and bounded, and $F\colon \overline{D} \to \mathbb{R}$ is continuous, then F is integrable.*

As for the integrability of monotone functions, we define $F(x,y)$ to be *monotone non-decreasing* if

$$F(x,y) \leq F(z,w) \qquad \text{whenever} \qquad x \leq z, \quad y \leq w. \tag{17.20}$$

Reversing one or both of the latter inequalities produces three other forms of monotonicity for functions of two variables.

Theorem 17.3. *If $R \subset \mathbb{R}^2$ is a bounded, closed rectangle, and $F\colon \overline{D} \to \mathbb{R}$ is monotone and bounded, then F is integrable on R.*

Proof: Suppose $\alpha \leq F(x,y) \leq \beta$ for $(x,y) \in R$, and hence

$$F(z,w) - F(x,y) \leq \beta - \alpha \qquad \text{for all} \qquad (x,y), (z,w) \in R.$$

We choose uniform partitions

$$x_i = i\,\frac{b-a}{n}, \qquad y_j = j\,\frac{d-c}{n}, \qquad i,j = 0, \ldots, n,$$

for $n \in \mathbb{N}$, and subdivide $R = \bigcup_{i,j} R_{i,j}$ as in (17.18–19). Assuming monotone nondecreasing as in (17.20), we can choose lower and upper bounds to be the values of F at the bottom left and top right corners of each rectangle, respectively:

$$\alpha_{i,j} = F(x_{i-1}, y_{j-1}) \leq F(x,y) \leq F(x_i, y_j) = \beta_{i,j}, \quad x_{i-1} \leq x \leq x_i, \quad y_{j-1} \leq y \leq y_j.$$

Thus the intermediate volumes are given by

$$
\begin{aligned}
\operatorname{vol} Z^* &= \sum_{i=1}^{n} \sum_{j=1}^{n} (\beta_{i,j} - \alpha_{i,j})\,(x_i - x_{i-1})\,(y_j - y_{j-1}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \big[\, F(x_i, y_j) - F(x_{i-1}, y_{j-1}) \,\big]\, \frac{(b-a)\,(d-c)}{n^2} \\
&= \left[F(b,d) - F(a,c) + \sum_{i=1}^{n-1} \big[\, F(b,y_i) - F(a,y_i) + F(x_i,d) - F(x_i,c) \,\big] \right] \frac{(b-a)\,(d-c)}{n^2} \\
&\le 2n\,(\beta - \alpha)\,\frac{(b-a)\,(d-c)}{n^2} = \frac{2\,(\beta-\alpha)\,(b-a)\,(d-c)}{n}\,,
\end{aligned}
$$

noting that most of the terms in the intermediate summation cancel each other. (Compare the proof of the one variable version Theorem 14.4.) The latter quantity has zero infimum for $n \in \mathbb{N}$, thus proving integrability. *Q.E.D.*

Unfortunately, there is no direct analog of the Fundamental Theorem of Calculus for double integrals. The most important counterpart is Green's Theorem, [**1**, **18**], but this involves line integrals that we will not cover here. In the absence of such a result, an effective way to evaluate a double integral is to reduce it to a pair of ordinary single integrals. This result is known as Fubini's Theorem, named after the early twentieth century Italian mathematician Guido Fubini, although simpler versions, e.g., for bounded continuous functions, were known much earlier.

Theorem 17.4. *Let $R = \overline{R}{}^{b,d}_{a,c} \subset \mathbb{R}^2$ be a closed rectangle. Let $F\colon R \to \mathbb{R}$ be integrable. Assume further that, for each fixed $c \le y \le d$, the functions $f_y(x) = F(x,y)$ are integrable on the interval $a \le x \le b$. Then the function*

$$
h(y) = \int_a^b f_y(x)\,dx = \int_a^b F(x,y)\,dx
$$

is integrable on the interval $c \le y \le d$, and

$$
\iint_R F(x,y)\,dx\,dy = \int_c^d h(y)\,dy = \int_c^d \left(\int_a^b F(x,y)\,dx \right) dy. \tag{17.21}
$$

Similarly, if, for each fixed $a \le x \le b$, the function $g_x(y) = F(x,y)$ is integrable on the interval $c \le y \le d$, then

$$
k(x) = \int_c^d g_x(y)\,dy = \int_c^d F(x,y)\,dy
$$

is integrable on the interval $a \le x \le b$, and

$$
\iint_R F(x,y)\,dx\,dy = \int_a^b k(x)\,dx = \int_a^b \left(\int_c^d F(x,y)\,dy \right) dx. \tag{17.22}
$$

In the inner expressions in the Fubini formulae, the variable not being integrated is to be treated as a constant parameter. For example, let $R = \overline{R}_{0,0}^{2,1} = \{\,0 \le x \le 2,\ 0 \le y \le 1\,\}$. Then

$$\iint_R \left(x\,y + y^2 \right) dx\,dy = \int_0^2 \left[\int_0^1 \left(x\,y + y^2 \right) dy \right] dx = \int_0^2 \left(\tfrac12 x + \tfrac13 \right) dx = \tfrac53,$$

or, alternatively,

$$\iint_R \left(x\,y + y^2 \right) dx\,dy = \int_0^1 \left[\int_0^2 \left(x\,y + y^2 \right) dx \right] dy = \int_0^1 \left(2\,y + 2\,y^2 \right) dy = \tfrac53,$$

which give the same answer, as Theorem 17.4 says they must.

Remark: According to Proposition 10.3, continuity of $F(x,y)$ implies continuity, and hence integrability, of $f_y(x)$ and $g_x(y)$. However, integrability of $F(x,y)$ does not necessarily imply their integrability. For example, the function

$$F(x,y) = \begin{cases} 1, & x \in \mathbb{Q} \ \text{ and } \ y = \frac12, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{for} \qquad 0 \le x, y \le 1,$$

is integrable over the unit square $\overline{R}_{0,0}^{1,1}$ with double integral zero. This is because $F(x,y) \ne 0$ only on the line segment $\{\,\left(x, \frac12\right) \mid 0 \le x \le 1\,\}$, which has zero area; see (17.9). On the other hand, $f_{1/2}(x) = F\left(x, \frac12\right)$ is the non-integrable function (14.15). Intuitively, the fact that $f_y(x)$ is not integrable for only one value of x does not affect the overall integrability of $F(x,y)$. On the other hand, if we fix $0 \le x \le 1$, every one of the functions $g_x(y) = F(x,y)$ is zero at every point except possibly $y = \frac12$, and hence is integrable with zero integral. Thus, while in this case we cannot justify the first Fubini formula (17.21), we can apply the second Fubini formula (17.22) to compute the double integral.

Proof of Theorem 17.4: We will prove the first Fubini formula (17.21) in which one first integrates with respect to x and then with respect to y; proof of the second simply amounts to interchanging the roles of x and y. We will employ the scheme based on partitions (17.18) of the intervals $a \le x \le b$ and $c \le y \le d$, their associated rectangles (17.19) and lower and upper bounds $\alpha_{i,j} \le F(x,y) \le \beta_{i,j}$ for $(x,y) \in R_{i,j}$. In this case, since the domain $D = \overline{R}_{a,c}^{b,d}$ is a closed rectangle, all of the corresponding rectangles $R_{i,j}$ are subsets thereof, and their total area equals the area of D.

Let us fix $y_{j-1} \le y \le y_j$, and consider the function $f_y(x) = F(x,y)$ for $a \le x \le b$. The lower and upper bounds for $F(x,y)$ on $R_{i,j}$ also serve as lower and upper bounds for $f_y(x)$:

$$\alpha_{i,j} \le f_y(x) = F(x,y) \le \beta_{i,j} \quad \text{for} \quad x_{i-1} \le x \le x_i.$$

Because we are assuming integrability of $f_y(x)$, in view of (14.35, 37), the corresponding lower and upper sums, which we denote by λ_j, μ_j, respectively, bound its integral:

$$\lambda_j = \sum_{i=1}^k \alpha_{i,j}\,(x_i - x_{i-1}) \le h(y) = \int_a^b f_y(x)\,dx \le \sum_{i=1}^k \beta_{i,j}\,(x_i - x_{i-1}) = \mu_j. \tag{17.23}$$

Warning: It is not necessarily true that the supremum of these lower sums or the infimum of the upper sums equals the integral of $f_y(x)$. This is because the lower and upper bounds $\alpha_{i,j}, \beta_{i,j}$ are not completely general, since they are required to hold on the entire rectangle $R_{i,j}$, not just on the interval $[x_{i-1}, x_i]$.

Now let us form the corresponding lower and upper sums for the function $h(y)$, which similarly bound its integral:

$$\sum_{j=1}^{l} \lambda_j \,(y_j - y_{j-1}) \le \int_c^d h(y)\,dy \le \sum_{j=1}^{l} \mu_j \,(y_j - y_{j-1}).$$

Using (17.23), these lower and upper sums are

$$\begin{aligned} \sum_{j=1}^{l} \lambda_j \,(y_j - y_{j-1}) &= \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_{i,j}\,(x_i - x_{i-1})\,(y_j - y_{j-1}), \\ \sum_{j=1}^{l} \mu_j \,(y_j - y_{j-1}) &= \sum_{i=1}^{k} \sum_{j=1}^{l} \beta_{i,j}\,(x_i - x_{i-1})\,(y_j - y_{j-1}), \end{aligned} \tag{17.24}$$

and we recognize the right hand sides as the lower and upper sums (17.17) for the double integral of $F(x,y)$. Their difference is

$$\sum_{j=1}^{l} (\mu_j - \lambda_j)\,(y_j - y_{j-1}) = \sum_{i=1}^{k} \sum_{j=1}^{l} (\beta_{i,j} - \alpha_{i,j})\,(x_i - x_{i-1})\,(y_j - y_{j-1}). \tag{17.25}$$

The left hand side of (17.25) is the total area of the intermediate rectangles associated with the given lower and upper bounds for $h(y)$. The right hand side of (17.25) is the total volume of the intermediate boxes for the double integral. Since $F(x,y)$ is integrable, the infimum of the latter volumes is zero, and hence its double integral is the supremum of the lower sums, which also equals the infimum of the upper sums on the right hand side of (17.24). We conclude that the intermediate area sums on the left hand side of (17.25) also have zero infimum, which, using the criterion in Theorem 14.12, implies that $h(y)$ is integrable, and its integral equals the supremum of the lower sums appearing on the left hand side of (17.24). Since these quantities agree, we conclude that

$$\iint_R F(x,y)\,dx\,dy = \int_c^d h(y)\,dy,$$

which is equivalent to the first Fubini formula (17.21). *Q.E.D.*

Example 17.5. Here is a cautionary example. Suppose we try to integrate the function $F(x,y) = (x^2 - y^2)/(x^2+y^2)^2$ over the unit square $Q = \{\, 0 \le x, y \le 1 \,\}$. If we use the first Fubini formula (17.21), we obtain

$$\iint_Q \frac{(x^2 - y^2)\,dx\,dy}{(x^2+y^2)^2} = \int_0^1 \left(\int_0^1 \frac{(x^2-y^2)\,dx}{(x^2+y^2)^2} \right) dy = -\int_0^1 \frac{dy}{1+y^2} = -\frac{\pi}{4}.$$

The second equality is a consequence of the Fundamental Theorem 15.3 using the fact that the integrand is the partial derivative with respect to x of the function $-x/(x^2+y^2)$; the final integral involves the inverse tangent function and can be found in any standard calculus text. On the other hand, if we use the second Fubini formula (17.22) we obtain

$$\iint_Q \frac{(x^2-y^2)\,dx\,dy}{(x^2+y^2)^2} = \int_0^1 \left(\int_0^1 \frac{(x^2-y^2)\,dy}{(x^2+y^2)^2} \right) dx = \int_0^1 \frac{dx}{1+x^2} = +\frac{\pi}{4}.$$

Thus, the two iterated integrals give different values! The problem is that $F(x,y)$ is not bounded on Q — indeed, $F(x,0) = 1/x^2$, while $F(0,y) = -1/y^2$ — and hence is not integrable, so Theorem 17.4 does not apply. On the other hand, one can use the Fubini formulas to integrate F on any rectangle that does not contain the origin.

One can generalize Fubini's formula (17.22) to a region lying between the graphs of two scalar functions $f, g\colon [a,b] \to \mathbb{R}$ such that $f(x) < g(x)$ for all $x \in [a,b]$. Let

$$D = \{\, (x,y) \mid \ a < x < b,\ f(x) < y < g(x) \,\}.$$

Then, under the appropriate integrability assumptions,

$$\iint_D F(x,y)\,dx\,dy \ = \int_a^b \left(\int_{f(x)}^{g(x)} F(x,y)\,dy \right) dx. \tag{17.26}$$

In this case one cannot immediately reverse the order of integration, unless the region can be alternatively written as

$$D = \{\, (x,y) \mid \ c < y < d,\ h(y) < x < k(y) \,\},$$

in which case

$$\iint_D F(x,y)\,dx\,dy \ = \int_c^d \left(\int_{h(y)}^{k(y)} F(x,y)\,dx \right) dy. \tag{17.27}$$

In particular, one can compute the area of D using (17.6):

$$\text{area } D \ = \iint_D \ dx\,dy \ = \int_a^b \big[\, g(x) - f(x) \,\big]\,dx,$$

thus reproducing a standard univariate calculus formula (14.24).

With the foundations in place, further development of vector calculus, including extensions to three-dimensional space, change of variables, line and surface integrals, Green's Theorem, Stokes Theorem, Gauss' Theorem, proceeds in essentially the same manner as in traditional texts, e.g., [**1**, **6**, **16**]. Also see my lecture notes [**18**, **19**] on Vector Calculus in Two and Three Dimensions, available on my website. We thus view this as an appropriate point to close off our development of "Continuous Calculus".

18. Final Remarks.

For the reader who is already familiar with calculus in its usual limit-based incarnation, a few remarks are in order. First of all, after the initial idea crossed my mind, I was pleasantly surprised how easily all the parts ended up fitting together and, in many cases, simplified the proofs of many of the important theorems, especially those involving differentiation. Of course, one needs to spend a little time in advance discussing the topology of the real line and, later, the plane, but most of the results are not especially difficult, and are, of course, well worth learning in their own right. Initially I was only going to cover basic differential calculus, but as I gradually realized that all of the fundamental results, including Cauchy sequences and uniform continuity and convergence, could be handled by this framework, the more these notes expanded in scope and sophistication. Indeed, I could not resist including the simple proof of the Fundamental Theorem 15.3, and hence ended up including a somewhat innovative approach to Riemann integration, in which the topological approach to continuity also plays a role, albeit less prominent. I am also surprised that, as noted in the introduction, there appears to be only one precedent for taking this approach, [**2**], although perhaps there are others that I am simply unaware of.

Pedagogically, it would probably be better to initially skip over a number of the topological results collected together in the second and fifth sections, and get to the heart of the matter sooner, referring back as they arise. On the other hand, the distractions and complications of epsilons and deltas completely disappear without any sacrifice in rigor or generality! Moreover, while the language of limits proves to still be useful, they no longer play a central role and, in fact, could be done away with entirely.

The more accurate title of these notes, "Topological Calculus", confirms that this is an essentially topological theory, and independent of any choice of metric†. Thus, like the proverbial topologist who cannot distinguish between a doughnut and a coffee cup, topological calculus does not distinguish between calculus on doughnuts and calculus on a coffee cups! As such, I believe it is the right approach to calculus on manifolds, [**25**], before one introduces any sort of metric — Riemannian, Finsler, etc. In this approach, differential forms and the deRham theory, [**30**], can now be formulated in a topological (continuous) and, I would argue, more natural framework.

In any event, I am very interested in your reactions to these notes. Please send me email with comments, concerns, criticisms, corrections, etc.

Acknowledgments: I would like to thank Ben Brubaker, Jeremy Butterfield, Melvin Leok, Joseph Malkoun, Sheehan Olver, and Beatrice Pelloni for comments, references, and encouragement. I particularly thank Alexander Heaton for his careful reading of the manuscript resulting in many comments, corrections, and suggestions.

---

† Actually, this is not 100% correct, since we implicitly use the standard $|\,x-y\,|$ distance on $\mathbb{R}$ and the corresponding L^∞ norm on $\mathbb{R}^2$, cf. [**22**], at verious points in the development. Nevertheless, the underlying theme of our approach is topological.

References

[**1**] Apostol, T.M., *Calculus*, Blaisdell Publ., Waltham, Mass., 1967–69.

[**2**] Bernstein, D.J., *Calculus for Mathematicians*, University of Illinois at Chicago, 1997. `https://cr.yp.to/papers.html#calculus`

[**3**] Bradie, B., *A Friendly Introduction to Numerical Analysis*, Prentice–Hall, Upper Saddle River, N.J., 2006.

[**4**] Carathéodory, C., *Theory of Functions of a Complex Variable*, vol. 1, Chelsea Publ., New York, 1954.

[**5**] Carathéodory, C., *Vorlesungen über Reele Functionen*, Chelsea Publ., New York, 1968.

[**6**] Courant, R., *Differential and Integral Calculus*, Interscience Publ., New York, 1936–37.

[**7**] Edwards, C.H., and Penney, D.E., *Calculus*, 6th ed., Pearson Education, Boston, Mass., 2002.

[**8**] Hales, T.C., Jordan's proof of the Jordan Curve Theorem, *Studies Logic Grammar Rhetoric* **10** (23) (2007), 45–60.

[**9**] Hurewicz, W., and Wallman, H., *Dimension Theory*, Princeton University Press, Princeton, N.J., 2015.

[**10**] Johnson, D.M., The problem of the invariance of dimension in the growth of modern topology. I, *Arch. Hist. Exact Sci.* **20** (1979), 97–188.

[**11**] Johnson, D.M., The problem of the invariance of dimension in the growth of modern topology. II, *Arch. Hist. Exact Sci.* **25** (1981), 85–267.

[**12**] Kelley, J.L., *General Topology*, Graduate Texts in Mathematics, vol. 27, Springer–Verlag, New York, 1975.

[**13**] Kuhn, S., The derivative á la Carathéodory, *Amer. Math. Monthly* **98** (1991), 40–44.

[**14**] LeCun, Y., Bengio, Y., and Hinton, G., Deep learning, *Nature* **521** (7553) (2015), 436–444.

[**15**] Mandelbrot, B.B., *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.

[**16**] Marsden, J.E., and Tromba, A.J., *Vector Calculus*, 6th ed., W.H. Freeman, New York, 2012.

[**17**] Marsden, J., and Weinstein, A.J., *Calculus Unlimited*, Benjamin/Cummings, Menlo Park. Calif., 1981.

[**18**] Olver, P.J., *Vector Calculus in Two Dimensions*, Lecture Notes, University of Minnesota, 2013. `http://www-users.math.umn.edu/~olver/ln_/vc2.pdf`

[**19**] Olver, P.J., *Vector Calculus in Three Dimensions*, Lecture Notes, University of Minnesota, 2013. `http://www-users.math.umn.edu/~olver/ln_/vc3.pdf`

[**20**] Olver, P.J., *Complex Analysis and Conformal Mapping*, Lecture Notes, University of Minnesota, 2020. `http://www-users.math.umn.edu/~olver/ln_/cml.pdf`

[**21**] Olver, P.J., Motion and continuity, *Math. Intelligencer* **44** (2022), 241-249.

[**22**] Olver, P.J., and Shakiban, C., *Applied Linear Algebra*, Second Edition, Undergraduate Texts in Mathematics, Springer, New York, 2018.

[**23**] Robinson, A., *Non-standard Analysis*, Princeton University Press, Princeton, N.J., 1974.

[**24**] Rudin, W., *Principles of Mathematical Analysis*, 3rd ed., McGraw–Hill, New York, 1976.

[**25**] Spivak, M., *Calculus on Manifolds*, W.A. Benjamin, Menlo Park, Calif., 1965.

[**26**] Spivak, M., *Calculus*, W.A. Benjamin, New York, 1967.

[**27**] Stewart, J., *Calculus*: *Early Transcendentals*, 7th ed., Cengage Learning, Mason, Ohio, 2012.

[**28**] Strang, G., *Calculus*, Wellesley Cambridge Press, Wellesley, Mass., 1991.

[**29**] Vick, J.W., *Homology Theory*: *an Introduction to Algebraic Topology*, 2nd ed., Springer–Verlag, New York, 1994.

[**30**] Warner, F.W., *Foundations of Differentiable Manifolds and Lie Groups*, Springer–Verlag, New York, 1983.

