Mathematics 1

What's this? This is a textbook covering the course *Matematika 1 za fizike*, a course in real analysis given to first year physics students at the Faculty of Mathematics and Physics at the University of Ljubljana and taught by professor Janez Mrčun and Jure Kališnik since 2008.

About the document: This document was translated from Slovene, typeset, and supplemented with additional explanations and formatting by Elijan Mastnak in 2022, based on the original Slovene lecture notes for this course, which can be found on professor Mrčun's website at the link https://www.fmf.uni-lj.si/~mrcun/preprints/mat1.pdf.

Material includes: the fundamentals of set theory, the axiomatic construction of the natural, real, and complex numbers, vectors in $\mathbb{R}^3$, sequences and series of numbers, power series, function sequences, function series, and the continuity, differentiability, and integrability of real-valued single-variable functions. The course concludes with a brief introduction to the continuity and differentiability of multivariable functions, including the chain rule for partial derivatives, Taylor's formula for multivariable functions, and a sufficient condition for the extrema of a scalar-valued function of two variables.

Last update: April 13, 2023

Contents

1 Fundamentals of set theory

1.1 Formulating and describing sets

Intuitively, a set is a group of items that (generally) share some common property; items in a set are called *elements* or *members* of the set. Sets are conventionally denoted by uppercase letters, e.g. A, B, X, Y, while their elements are usually denoted by lowercase letters, e.g. a, b, x, y.

Set theory is based on the concept of membership, i.e. of some item "being an element of" a set. The notation $a \in A$ means that a is an element of the set A, while the notation $a \notin A$ means that a is not an element of the set A.

A set is fully described by specifying all of its elements. Two sets are said to be equal when they have identical elements; more precisely, two sets A and B are equal when every element of A is also an element of B, and every element of B is also an element of A. The notation $A = B$ means that the sets A and B are equal. Order and repetition of elements do not affect set equality, for example

$$\{1, 2, 1\} = \{1, 2\} = \{2, 1\} \qquad \text{(these three sets are all equal).}$$

Definition 1.1: Subsets

A set B is said to be a *subset* of a set A if every element of B is also an element of A; the notation $B \subset A$ means that B is a subset of A.

The definition of set equality implies that

$$A = B \text{ if, and only if, } B \subset A \text{ and } A \subset B.$$

Topic 1.1: Ways of defining sets

A set can be explicitly defined by writing its elements, separated by commas, between two braces; for example, the notation

$$A = \{1, 2, 3, 4\}$$

defines a set A whose elements are the first four natural numbers, 1, 2, 3, and 4.

When a set's elements follow a well-defined pattern, the set can also be defined, albeit informally, by listing only the first few elements followed by an ellipsis; the remaining elements are implicitly deduced from the pattern given by the first few elements. For example, the natural numbers, denoted by $\mathbb{N}$, can be informally defined in the form

$$\mathbb{N} = \{1, 2, 3, 4, 5, \ldots\} \qquad \text{(the natural numbers, defined informally).}$$

Although the above definition obviously does not explicitly state all of the natural numbers, the pattern in the first few elements makes it reasonably clear that the set $\mathbb{N}$ also contains the numbers $6, 7, 8$, and so on. We will define the natural numbers more formally in Chapter 2.

A set, say B, may also be defined as the subset of an already-defined set A based on some property $\mathcal{P}$ in the form

$$B = \{a \in A;\ a \text{ has the property } \mathcal{P}\};$$

in this case B contains all $a \in A$ for which a has the property $\mathcal{P}$. As an example, the set of all even natural numbers, denoted by $2\mathbb{N}$, can be defined as

$$2\mathbb{N} = \{n \in \mathbb{N};\, n \text{ is even}\} \subset \mathbb{N};$$

the first few elements of this set are $2, 4, 6$, and so on.

Definition 1.2: The empty set

The set containing no elements at all is called the *empty set*, is denoted by $\emptyset$, and can be written as

$$\emptyset = \{\} \qquad \text{(the empty set)}.$$

1.1.1 Basic operations on sets

This section defines some common operations on sets that again yield a set.

Definition 1.3: Intersection of sets and disjoint sets

The *intersection* of two sets A and B, denoted by $A \cap B$, is the set of all items that are elements of *both* A and B. The intersection $A \cap B$ can be defined in the following equivalent forms:

$$A \cap B = \{a \in A;\, a \in B\} = \{b \in B;\, b \in A\} = \{a \in A;\, a \in A \text{ and } a \in B\}.$$

Two sets are said to be *disjoint* if their intersection is the empty set, i.e. the sets A and B are disjoint if $A \cap B = \emptyset$.

Because $A \cap B$ is the set of all elements that lie in both A and B, the intersection $A \cap B$ is necessarily a subset of both A and B. Here is a simple example of a set intersection:

$$\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}.$$

Set intersections are commutative and associative, i.e.

$$\begin{aligned} A \cap B &= B \cap A && \text{(intersections are commutative)} \\ A \cap (B \cap C) &= (A \cap B) \cap C && \text{(intersections are associative)}; \end{aligned}$$

these properties follow from two sets being equal when they contain the same elements, regardless of the order and repetition of these elements. The associativity of set intersection motivates the definition of the intersection of $n \in \mathbb{N}$ sets as

$$\begin{aligned} \bigcap_{k=1}^{n} A_k &= A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap A_n \\ &= A_1 \cap \big(A_2 \cap \cdots (A_{n-1} \cap A_n) \cdots\big). \end{aligned}$$

Definition 1.4: Difference of sets

The *difference* of the sets A and B, denoted by $A \setminus B$, is the set of all elements of A

that are not elements of B. The set difference $A \setminus B$ is defined as

$$A \setminus B = \{a \in A;\, a \notin B\}.$$

The difference $A \setminus B$ is the set of all elements of A that are not also elements of B, so $A \setminus B$ is a subset of A, but not necessarily a subset of B. Here is a simple example of a set difference:

$$\{1, 2, 3\} \setminus \{2, 3, 4\} = \{1\}.$$

Definition 1.5: Union of sets

The *union* of the sets A and B, denoted by $A \cup B$, is the set of all items that are elements of both A and B. The set union $A \cup B$ is defined as

$$A \cup B = \{a;\, a \in A \text{ or } a \in B\}.$$

The union $A \cup B$ is the set of all elements either in A or in B, so A and B are both subsets of $A \cup B$. Here is a simple example of a set union:

$$\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}.$$

Like the set intersection, the set union is both commutative and associative, i.e.

$$\begin{aligned} A \cup B &= B \cap A && \text{(unions are commutative)} \\ A \cup (B \cap C) &= (A \cap B) \cap C && \text{(unions are associative).} \end{aligned}$$

The associativity of set unions motivates the definition of the union of $n \in \mathbb{N}$ sets as.

$$\begin{aligned} \bigcup_{k=1}^{n} A_k &= A_1 \cup A_2 \cup \cdots \cup A_{n-1} \cup A_n \\ &= A_1 \cup \big(A_2 \cup \cdots (A_{n-1} \cup A_n) \cdots\big). \end{aligned}$$

Definition 1.6: Cartesian product of sets

The *Cartesian product* of the sets A and B is denoted by $A \times B$ and defined as

$$A \times B = \big\{(a, b);\, a \in A, b \in B\big\}.$$

In words, the Cartesian product $A \times B$ is the set of all ordered pairs (a, b) satisfying the conditions:

- $a \in A$
- $b \in B$
- $(a, b) = (a', b')$ if, and only if, $a = a'$ and $b = b'$.

Here is a simple example of a Cartesian product:

$$\{1, 2\} \times \{2, 3\} = \{(1, 2), (1, 3), (2, 2), (2, 3)\}.$$

The Cartesian product of $n \in \mathbb{N}$ sets $A_1, A_2, \ldots, A_n$ is defined as

$$\begin{aligned}\prod_{k=1}^{n} A_k &= A_1 \times A_2 \times \cdots \times A_n \\ &= \big\{(a_1, a_2, \ldots, a_n);\, a_1 \in A_1, \ldots, a_n \in A_n\big\},\end{aligned}$$

where $(a_1, \ldots, a_n) = (a'_1, \ldots, a'_n)$ if, and only if, $a_k = a'_k$ for all $k = 1, \ldots, n$.

Definition 1.7: Complement of a set

Let A be a subset of some universal set U, i.e. $A \subset U$ (this sounds abstract here, but the universal set is usually clear from context). The *complement* of A with respect to the universal set U is denoted by A^{c} and defined as

$$A^{\mathsf{c}} = U \setminus A = \{a \in U\,;\, a \notin A\}.$$

The set A^{c} contains all elements of U that are not elements of A.

Note that a set complement is well-defined only with respect to a given universal set; i.e. the expression A^{c} is not well-defined without specifying the universal set with respect to which the complement is taken. In practice, the universal set will be either explicitly stated or clear from context.

As a simple example, let $A = \{1, 2\}$ and $U = \{1, 2, 3, 4\}$; then

$$A^{\mathsf{c}} = \{3, 4\} \qquad \text{(with respect to the set } U\text{)};$$

however, if $U' = \{1, 2, 3, 4, 5\}$, then

$$A^{\mathsf{c}} = \{3, 4, 5\} \qquad \text{(with respect to the set } U'\text{)}.$$

1.1.2 Indexed families of sets

Let J be an arbitrary nonempty set, and for each element $j \in J$ define the set A_j. The collection of all the sets A_j for all indices $j \in J$ is called an *indexed family of sets* and is written

$$\big(A_j\big)_{j \in J} \qquad \text{(notation: indexed family of sets)}.$$

In this context J is called the *index set* of the indexed family $\big(A_j\big)_{j \in J}$, and the elements $j \in J$ are called *indices*. The intersection and union of an indexed family of sets $\big(A_j\big)_{j \in J}$ are defined as

$$\bigcap_{j \in J} A_j = \{a\,;\, a \in A_j \text{ for all } j \in J\} \qquad \text{(intersection of an indexed family of sets)}$$

$$\bigcup_{j \in J} A_j = \{a\,;\, a \in A_j \text{ for some } j \in J\} \qquad \text{(union of an indexed family of sets)}.$$

The set of all sets in an indexed family of sets is called a *family of sets*; if $\big(A_j\big)_{j \in J}$ is an indexed family of sets indexed by the set J, then the corresponding family of sets, which we will denote by $\mathcal{A}$, is defined as

$$\mathcal{A} = \{A_j\,;\, j \in J\}.$$

The intersection and union of a family of sets $\mathcal{A}$ are defined as

$$\cap\mathcal{A} = \bigcap_{j\in J} A_j \qquad \text{(intersection of a family of sets)}$$
$$\cup\mathcal{A} = \bigcup_{j\in J} A_j \qquad \text{(union of a family of sets).}$$

Example 1.8: Families of sets and the natural numbers

Let $n \in \mathbb{N}$ be a natural number, let $J = \{1, 2, \ldots, n\}$ be an indexing set, and let $A_j = \{n \in \mathbb{N};\, n \geq j\} \subset \mathbb{N}$ for all $j \in J$. (The set A_j holds all natural numbers greater than or equal to j, for example $A_3 = \{3, 4, 5, \ldots\}$.) In this case:

- $\mathbb{N} = A_1 = \{1, 2, 3, \ldots\}$ by the definition of A_j and $\mathbb{N}$.
- $A_1 \supset A_2 \supset A_3 \supset \cdots$, i.e. each set A_j is a subset of the previous set A_{j-1}.
- $\bigcap\limits_{j\in J} A_j = \emptyset$, i.e no natural number $n \in \mathbb{N}$ occurs in *all* A_j.
- $\bigcup\limits_{j\in J} A_j = \mathbb{N}$, i.e. every natural number $n \in \mathbb{N}$ occurs in *at least one* A_j.

We conclude this section with de Morgan's laws, which are a pair of rules important in logic and Boolean algebra.

Proposition 1.9: De Morgan's laws

Let $\left(A_j\right)_{j\in J}$ be an indexed family of sets indexed by the set J, and for all $j \in J$ let A_j be a subset of the same universal set U. In this case

$$\bigcup_{j\in J} A_j^{\mathsf{c}} = \Big(\bigcap_{j\in J} A_j\Big)^{\mathsf{c}} \qquad \text{and} \qquad \bigcap_{j\in J} A_j^{\mathsf{c}} = \Big(\bigcup_{j\in J} A_j\Big)^{\mathsf{c}}.$$

Proof. We begin with LHS of the first equality and consider an arbitrary element $a \in \bigcup_{j\in J} A_j^{\mathsf{c}}$. The element a necessarily occurs in at least one of the A_j^{c}, so there exists some index $j_0 \in J$ for which $a \in A_{j_0}^{\mathsf{c}}$. By the definition of the set complement, it follows that $a \notin A_{j_0}$, and so $a \notin \bigcap_{j\in J} A_j$, which in turn implies that

$$a \in \Big(\bigcap_{j\in J} A_j\Big)^{\mathsf{c}}$$

More so, $a \notin \bigcap_{j\in J} A_j$ by the definition of the set complement, so there exists some index $j_0 \in J$ for which $a \notin A_{j_0}$. The result $a \notin A_{j_0}$ means that $a \in A_{j_0}^{\mathsf{c}}$, which in turn implies that $a \in \bigcup_{j\in J} A_j^{\mathsf{c}}$, proving the first equality.

The proof of the second equality is analogous. □

1.2 Functions

1.2.1 Basic concepts related to functions

Definition 1.10: Function, domain, codomain, and image

A *function* f *from a set* A *to a set* B, denoted by $f\colon A \to B$, is a mapping that assigns a *single* element $b = f(a) \in B$ to each element $a \in A$. Phrased equivalently, for each element $a \in A$, a function $f\colon A \to B$ associates a with a single element $b \in B$, denoted by $b = f(a)$.

We define the following additional terminology and notation for a function $f\colon A \to B$:

- The set A is called the function's *domain*.
- The set B is called the function's *codomain*.
- For each $a \in A$, the function f is said to *map* the element $a \in A$ to the value $f(a) \in B$; this relationship is sometimes denoted by $f\colon a \mapsto f(a)$. The value $f(a)$ is called the *image* of the element a under the function f.

Remark 1.11: Specifying a function

A function is fully specified by its domain, codomain, and mapping formula, i.e. the formula specifying how elements of the domain are mapped to elements of the codomain. Two functions are equal only if they share the same mapping formula *and* their domains and codomains are equal.

Definition 1.12: Graph of a function

The *graph* of a function $f\colon A \to B$ is denoted by $G(f)$ and defined as

$$G(f) = \big\{(a, b) \in A \times B\,;\, b = f(a)\big\}.$$

In words, the graph $G(f)$ is the set of all ordered pairs (a, b) for which $a \in A$ and $b = f(a)$ for some $a \in A$.

Note that, in general, $G(f) \subset A \times B$, i.e. $G(f)$ is a *subset* of $A \times B$ (even though one might intuitively expect $G(f) = A \times B$). The relationship $G(f) \subset A \times B$ stems from the fact that a function does not necessarily have to map to every element in its codomain. As a result, the b in $(a, b) \in G(f)$ in general don't encompass all $b \in B$, but only those $b \in B$ for which $b = f(a)$ for some $a \in A$, and so $G(f)$ is a subset of $A \times B$.

Definition 1.13: Image of a function

The image of a *function* is the set of all elements in the function's codomain that are images of at least one element in the function's domain. In symbols, the image of a function $f\colon A \to B$, denoted by $\operatorname{image}(f)$, is the set

$$\operatorname{image}(f) = \big\{b \in B\,;\, b = f(a) \text{ for at least one } a \in A\big\}.$$

In general, $\operatorname{image}(f) \subset B$, i.e. a function's image is a *subset* of its codomain; this is for the same reason mentioned above in the discussion of a function's graph—a function does

not in general map to every element in its codomain. As a simple example, consider the function $f\colon \mathbb{N} \to \mathbb{N}$ with the mapping formula $f(n) = 2n$. This function is defined to have the codomain $B = \mathbb{N}$, but the function's image is

$$\text{image}(f) = 2\mathbb{N} = \{2, 4, 6, 8, \ldots\} \neq B.$$

Topic 1.2: Image of a subset of a function's domain

Consider a function $f\colon A \to B$ mapping from the set A to the set B, and let $X \subset A$ denote an arbitrary subset of f's domain. The *image of the subset* $X \subset A$ under the function f is the set

$$f(X) = \big\{b \in B;\, b = f(a) \text{ for at least one } a \in X\big\} \subset B.$$

It follows immediately that $f(A) = \text{image}(f)$ in the special case when $X = A$.

The notation for defining $f(X)$ is admittedly a bit verbose; we will often use the following shorter notation instead:

$$\begin{aligned} f(X) &= \big\{b \in B;\, b = f(a) \text{ for at least one } a \in X\big\} \\ &= \big\{f(a);\, a \in X\big\}. \end{aligned}$$

More generally, letting $\mathcal{P}$ denote any property that can be unambiguously ascribed to the elements of the domain A, we can extend the above shorthand notation to

$$f\big(\{a \in X;\, a \text{ obeys property } \mathcal{P}\}\big) = \big\{f(a);\, a \in X \text{ and } a \text{ obeys property } \mathcal{P}\big\}.$$

Definition 1.14: Inverse image

Consider a function $f\colon A \to B$ mapping from the set A to the set B, and let $Y \subset B$ denote an arbitrary subset of f's codomain. The *inverse image* or *preimage* of the subset $Y \subset B$ under the function f, denoted by $f^{-1}(Y)$, is the set

$$f^{-1}(Y) = \big\{a \in A;\, f(a) \in Y\big\}.$$

In words, the inverse image $f^{-1}(Y)$ holds those elements in f's domain that are mapped to elements of Y. It directly follows that $f^{-1}(Y) \subset A$ and, in the special case when $Y = B$, that $f^{-1}(B) = A$.

Definition 1.15: Fiber of a single subset of the codomain

Consider a function $f\colon A \to B$ mapping from the set A to the set B, let $b \in B$ be an arbitrary point in f's codomain, and let $\{b\} \subset B$ denote the single-element (singleton) set containing b. The inverse image $f^{-1}(\{b\})$, i.e. the set of all elements of f's domain that are mapped to b, is called the *fiber* of the function f over the point $b \subset B$.

By the definition of the fiber,

- $f^{-1}(\{b\}) \subset A$ if there exists at least one element $a \in A$ for which $b = f(a)$,
- $f^{-1}(\{b\}) = \emptyset$ if there does not exist any $a \in A$ for which $b = f(a)$.

- The function image $\text{image}(f)$ contains all points $b \in B$ whose fibers $f^{-1}(\{b\})$ are nonempty.

Example 1.16: Some important functions

(i) (*The identity function*) The *identity function* (for an arbitrary set A) is the function $\text{id}_A \colon A \to A$ with the mapping formula

$$\text{id}_A(a) = a \text{ for all } a \in A.$$

In words, the identity function id_A maps every element in A to itself.

(ii) (*Restriction of a function to a subset*) Consider a function $f \colon A \to B$ mapping from the set A to the set B, and let $X \subset A$ be an arbitrary subset of f's domain. The *restriction* of the function f to the subset X, denoted by $f|_X$, is the function $f|_X \colon X \to B$ with the mapping formula

$$(f|_X)(a) = f(a) \text{ for all } a \in X.$$

If the restriction $f\big|_X$ has some property $\mathcal{P}$, then the parent function $f \colon A \to B$ is said to have the property $\mathcal{P}$ *on the subset* $X \subset A$.

(iii) (*Inclusion of a subset into a superset*) Consider an arbitrary set A and let $X \subset A$ be an arbitrary subset of A. The *inclusion map* of the subset X into the superset A, denoted by $\text{inc}_{X,A}$, is the function $\text{inc}_{X,A} \colon X \to A$ with the mapping formula

$$\text{inc}_{X,A}(a) = a \text{ for all } a \in X.$$

Interpretation: $\text{inc}_{X,A}(a)$ is similar to the identity function id_X, except that all a in $\text{image}(\text{inc}_{X,A})$ are treated as elements of the superset A (instead of elements of X, as for id_X).

(iv) (*Projection*) Let $A_1, A_2, \ldots, A_n$ be a collection of $n \in \mathbb{N}$ sets, let the tuple $(a_1, a_2, \ldots, a_n)$ denote an arbitrary element of the Cartesian product $A_1 \times A_2 \times \cdots \times A_n$, and define the integer $k \in \{1, 2, \ldots, n\}$.

A *projection* onto the k-th component of the Cartesian product $A_1 \times A_2 \times \cdots \times A_n$, denoted by pr_k, is a function of the form $\text{pr}_k \colon A_1 \times A_2 \times \cdots \times A_n \to A_k$ with the mapping formula

$$\text{pr}_k\big((a_1, \ldots, a_n)\big) = a_k \text{ for all } (a_1, \ldots, a_n) \in A_1 \times \cdots \times A_n.$$

In words, the projection pr_k takes as input n-tuples $(a_1, \ldots, a_n)$ from the entire Cartesian product $A_1 \times \cdots \times A_n$ and outputs only the tuple's k-th component, which is an element of A_k.

1.2.2 Composition of functions

Definition 1.17: Injective, surjective, and bijective functions

Consider a function $f\colon A \to B$ mapping from the set A to the set B.

(i) The function f is said to be *injective* if $f(a) \neq f(a')$ for any two elements $a, a' \in A$ for which $a \neq a'$. In words, f is injective if for any two unique elements a and a' in its domain, the images $f(a)$ and $f(a')$ of these elements under the function f are also unique.

(ii) The function f is said to be *surjective* if $f(A) = B$. In words, f is surjective if its image $f(A)$ equals its codomain B, i.e. if every element in its codomain is mapped to by some element in its domain.

(iii) The function f is said to be *bijective* if it is both injective and surjective.

Definition 1.18: Function composition

Let A, B, and C be three arbitrary sets and consider two functions $f\colon A \to B$ and $g\colon B \to C$; note that f is chosen to map into g's domain. The *composition* of f and g is the function $g \circ f\colon A \to C$ with the mapping formula

$$(g \circ f)(a) = g\big(f(a)\big) \text{ for all } a \in A.$$

The composition $g \circ f$ is defined only if f's codomain equals g's domain.

We now list two important properties of function composition; let A, B, C, and D be four arbitrary sets and consider three functions $f\colon A \to B$, $g\colon B \to C$, and $h\colon C \to D$. In this case:

(i) $(h \circ g) \circ f = h \circ (g \circ f)$ (composition is associative)

(ii) $f \circ \mathrm{id}_A = f$ and $\mathrm{id}_B \circ f = f$, which means that the identity function is a neutral element under composition.

Example 1.19: Composition

Consider the functions $f, g\colon \mathbb{R} \to \mathbb{R}$ given by the mapping formulae $f(x) = x^2$ and $g(x) = 2x$. In this case:

(i) $(f \circ g) = f\big(g(x)\big) = f(2x) = 4x^2$,

(ii) $(g \circ f) = g\big(f(x)\big) = g(x^2) = 2x^2$.

1.2.3 The inverse of a function

Definition 1.20: The inverse of a function

Consider a function $f\colon A \to B$ mapping from the set A to the set B.

(i) The function $g\colon B \to A$ is called a *left inverse* of f if

$$g \circ f = \mathrm{id}_A,$$

where id_A is the identity function on f's domain A, defined in Example 1.16 (i).

(ii) The function $h\colon B \to A$ is called a *right inverse* of f if

$$f \circ h = \mathrm{id}_B.$$

(iii) An arbitrary function is called the *inverse* of f if it is both f's left and right inverse. We will sometimes call a function's inverse its *two-sided inverse* to distinguish it from a left or right inverse.

Note that, in general, a function can have either multiple left inverses or multiple right inverses. As a simple example of a function with multiple left inverses, define the functions

(i) $f\colon \{0,1\} \to \{10,11,12\}$, $f(0) = 10$, $f(1) = 11$,

(ii) $g_0\colon \{10,11,12\} \to \{0,1\}$, $g_0(10) = 0$, $g_0(11) = 1$, $g_0(12) = 0$, and

(iii) $g_1\colon \{10,11,12\} \to \{0,1\}$, $g_1(10) = 0$, $g_1(11) = 1$, $g_1(12) = 1$.

In this case $g_0 \circ f = g_1 \circ f = \mathrm{id}_{\{0,1\}}$, so g_0 and g_1 are both left inverses of f.

As a simple example of a function with multiple right inverses, define the functions

(i) $\phi\colon \{0,1\} \to \{2\}$, $\phi(0) = \phi(1) = 2$,

(ii) $h_0\colon \{2\} \to \{0,1\}$, $h_0(2) = 0$, and

(iii) $h_1\colon \{2\} \to \{0,1\}$, $h_1(2) = 1$.

In this case $f \circ h_0 = \phi \circ h_1 = \mathrm{id}_{\{3\}}$, so both h_0 and h_1 are right inverses of ϕ. These two examples show that left and right inverses are not, in general, unique. However, a function's *inverse* (when it exists), *is* unique, as we will prove shortly.

Proposition 1.21: Left and right inverses must be equal

If a function has both a left and right inverse, the left and right inverses are identical.

Proof. Suppose a function $f\colon A \to B$ has both a left inverse $g\colon B \to A$ and a right inverse $h\colon B \to A$. We can then show that $h = g$ with the calculation

$$h \overset{(a)}{=} \mathrm{id}_A \circ h \overset{(b)}{=} (g \circ f) \circ h \overset{(c)}{=} g \circ (f \circ h) \overset{(d)}{=} g \circ \mathrm{id}_B \overset{(e)}{=} g,$$

where (a) and (e) follow from the definition of the identity function, (b) because g is the left inverse of f, (c) by the associativity of composition, and (d) because g is the left inverse of f. □

Corollary 1.22: A function's inverse is unique

A function can have at most one two-sided inverse. This statement follows directly from Proposition 1.21, which implies that for any function with one or more right inverses and one more left inverses, all of the left and right inverses are equal to the same two-sided inverse.

Definition 1.23: An invertible function

A function $f\colon A \to B$ mapping from the set A to the set B is said to be *invertible* if it has an inverse. In this case f's inverse is denoted by $f^{-1}\colon B \to A$.

We now list two basic properties of any two invertible functions $f\colon A \to B$ and $g\colon B \to C$:

(i) The inverse of an invertible function is again an invertible function and obeys

$$(f^{-1})^{-1} = f.$$

(ii) The composition of two invertible functions is again an invertible function and obeys

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proposition 1.24: Invertible and bijective functions

For any function $f\colon A \to B$:

(i) If f is injective and $A \neq \emptyset$, then f has a left inverse.

(ii) If f has a left inverse, then f is injective.

(iii) The function f is surjective if, and only if, f has a right inverse.

(iv) The function f is bijective if, and only if, f is invertible.

Proof.

(i) We will prove this point by constructing a function $g\colon B \to A$ that is a left inverse of f. Assuming f is injective, for each domain element $a \in A$ there exists exactly one corresponding image element $b = f(a) \in f(A)$, for which we define $g(b) = a$.

We must also consider elements in f's codomain that are not in f's image: because A is not empty, we may choose some element $a_0 \in A$ and, for all non-image elements $b' \in B \setminus f(A)$, define $a_0 = g(b')$. The thus-defined function $g\colon B \to A$ is a left inverse of f.

(ii) Let g be a left inverse of f. For any two elements $a, a' \in A$ for which $a \neq a'$, it holds that $g\big(f(a)\big) = a \neq a' = g\big(f(a')\big)$, from which follows that $f(a) \neq f(a')$, i.e. f is injective.

(iii) ($\Leftarrow$) Let h be a right inverse of f, in which case we may write each element $b \in B$ in the form $b = f\big(h(b)\big)$; this means that b is in f's image, i.e. that f is surjective.

($\Rightarrow$) Assume that f is surjective; in this case every element of f's codomain B is mapped to by some element in its domain A, and so for each element $b \in B$ the preimage $f^{-1}(\{b\}) \in A$ is nonempty. Let $a_b = f^{-1}(\{b\}) \in A$ denote any element in the preimage $f^{-1}(\{b\})$ (a_b always exists because the preimage is nonempty), and let the function $h\colon B \to A$ be defined by $h(b) = a_b$. The thus-defined function $h\colon B \to A$ is a right inverse of f.

(iv) This point follows directly from the above points and the definition of an invertible and bijective function. □

1.3 Set cardinality

Loosely, set cardinality refers to the number of elements in a set. We develop this concept more formally in this section.

Definition 1.25: Equipotence

The set A is *equipotent* to the set B if there exists a bijection from A to B. In this case we write

$$|A| = |B| \qquad \text{(notation: } A \text{ and } B \text{ are equipotent).}$$

Set equipotence obeys the following three properties for any three sets A, B, and C:

(i) (*Reflexivity*) Every set is equipotent to itself.

(ii) (*Symmetry*) If the set A is equipotent to the set B, then B is also equipotent to A.

(iii) (*Transitivity*) If the set A is equipotent to the set B and B is equipotent to the set C, then A is also equipotent to C.

Because it is reflexive, symmetric, and transitive, equipotence is an *equivalence relation*.

Definition 1.26: Classes of set cardinality

Consider an arbitrary set A and natural number $n \in \mathbb{N}$.

(i) The set A is said to *contain n elements* if A is equipotent to the set $\{1, 2, \ldots, n\}$.

(ii) The set A is said to be *finite* either if A is empty or if A has n elements for some natural number $n \in \mathbb{N}$.

(iii) The set A is said to be *infinite* if it is not finite.

(iv) The set A is said to be *countably infinite* if it is equipotent to the set of natural numbers $\mathbb{N}$, i.e. if there exists a bijection $f : \mathbb{N} \to A$.

(v) The set A is said to be *countable* either if A is finite or if A is countably infinite.

Here are two more properties related to set equipotence and cardinality:

(i) Two nonempty, finite sets are equipotent if, and only if, they have the same number of elements.

(ii) The empty set is equipotent only to itself.

Example 1.27: The set $2\mathbb{N}$ of even natural numbers

Perhaps counterintuitively, the set $2\mathbb{N} = \{2, 4, 6, \ldots\}$ of even natural numbers is equipotent to the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$, since there exists a bijection between the two sets, namely the function $f : \mathbb{N} \to 2\mathbb{N}$ defined by $f(n) = 2n$. The corresponding inverse function is $f^{-1} : 2\mathbb{N} \to \mathbb{N}$ with the mapping formula

$f^{-1}(n) = n/2$.

Example 1.28: The set $\mathbb{N} \times \mathbb{N}$ is countable

The Cartesian product $\mathbb{N} \times \mathbb{N}$ is countably infinite (and thus countable). To show this, we arrange the elements of $\mathbb{N} \times \mathbb{N}$ in the form

$$\begin{array}{cccc} (1,1) & (1,2) & (1,3) & \cdots \\ (2,1) & (2,2) & (2,3) & \cdots \\ (3,1) & (3,2) & (3,3) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

and construction a bijection $f\colon \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ by arranging the elements of $\mathbb{N} \times \mathbb{N}$ along the diagonals from top left to bottom right, i.e.

$$(1,1), (2,1), (1,2), (3,1), (2,2), (1,3), \ldots.$$

The corresponding inverse function $f^{-1}\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ has the mapping formula

$$f^{-1}\big((m,n)\big) = n + \frac{1}{2}(m+n-2)(m+n-1) \text{ for all } (m,n) \in \mathbb{N} \times \mathbb{N}.$$

Proposition 1.29: Subsets of countable sets are countable

Every subset of a countable set is itself countable.

Proof. Since every countable set can be related with a bijection to a subset of the natural numbers, it suffices to prove the theorem for subsets of the natural numbers.

Let $A \subset \mathbb{N}$ be an arbitrary *infinite* subset of $\mathbb{N}$. (If A were finite the proof would be trivial, since every finite set is by definition countable.) We aim to prove that A is *countably* infinite, i.e. that there exists a bijection $f\colon \mathbb{N} \to A$ from $\mathbb{N}$ and A.

We construct the bijection as follows: let $f(1)$ be the smallest natural number in A, then let $f(2)$ be the smallest natural number in $A \setminus \{f(1)\}$, and so on, so that $f(n+1)$ is the smallest natural number in the set $A \setminus \{f(1), f(2), \ldots, f(n)\}$; this smallest number is guaranteed to exist because A is an infinite subset of the natural numbers. The function $f\colon N \to A$ is injective by construction, since any two $n, n' \in \mathbb{N}$ with $n \neq n'$ map to two different $a, a' \in A$. More so, because each element in A is larger than only a finite number of natural numbers, every element in A is mapped to, and so f is also surjective. The function f is thus bijective, and so A is countably infinite and thus countable. □

Proposition 1.30: Relating surjective and injective functions to countable domains and codomains

Consider a function $f\colon A \to B$ mapping from the set A to the set B.

(i) If the function f is injective and its codomain B is countable, then its domain A is also countable.

(ii) If the function f is surjective and its domain A is countable, then its codomain B is also countable.

Proof.

(i) The map $A \to f(A)$ (from A to f's *image* $f(A)$) is necessarily surjective; more so, because f is injective, the map $A \to f(A)$ is also bijective. The function f's codomain B is countable, so by Proposition 1.29 its image $f(A) \subset B$ is also countable. Since $f(A)$ is countable and there exists a bijection between $f(A)$ and A, the domain A must also be countable.

(ii) Because f is surjective, f must have a right inverse $h\colon B \to A$ by Proposition 1.24 (iii). The function h thus has a left inverse f, and so h is injective by Proposition 1.24 (ii). Since h is injective and h's codomain A is countable, h's domain (and f's codomain) B must be countable by point (i). □

Corollary 1.31: Relating surjective and injective functions to countable sets

The natural numbers are by definition countable, so Proposition 1.30 directly implies:

(i) An arbitrary set A is countable if, and only if, there exists an injective function of the form $f\colon A \to \mathbb{N}$.

(ii) An arbitrary nonempty set A is countable if, and only if, there exists a surjective function $\mathbb{N} \to A$.

Proposition 1.32: Cartesian product of two countable sets

The Cartesian product of any two countable sets is also a countable set.

Proof. Let A and B be two countable sets and let a and b denote arbitrary elements of A and B, respectively. By Corollary 1.31 (i) there exist injective functions of the form $f\colon A \to \mathbb{N}$ and $g\colon B \to \mathbb{N}$. Using f and g, we then define the new function $f \times g\colon A \times B \to \mathbb{N} \times \mathbb{N}$ with mapping formula $(f \times g)(a, b) = \big(f(a), g(a)\big)$. The function $f \times g$ is injective by construction (since f and g are injective); more so, because the codomain $\mathbb{N} \times \mathbb{N}$ is countable by Example 1.28, the domain $A \times B$ is also countable by Proposition 1.30 (i). Because A and B were arbitrary, the proposition holds for arbitrary countable sets. □

Corollary 1.33: Cartesian product of n countable sets

The Cartesian product of any finite number of countable sets is also a countable set.

Proof. Let $A_1, A_2, \ldots, A_n$ be countable sets for some natural number $n \in \mathbb{N}$. We first define the bijection

$$A_1 \times A_2 \times \cdots \times A_{n-1} \times A_n \to A_1 \times \big(A_2 \times \cdots (A_{n-1} \times A_n) \cdots\big);$$
$$(a_1, a_2, \ldots, a_{n-1}, a_n) \mapsto (a_1, (a_2, \ldots (a_{n-1}, a_n) \ldots)),$$

and then successively apply Proposition 1.32 to the Cartesian products $A_1 \times A_2$, $(A_1 \times A_2) \times A_3$, and so on until $(A_1 \times A_2 \times \cdots \times A_{n-1}) \times A_n$. □

Proposition 1.34: Countable unions of countable sets are countable

The union of a countably many countable sets is itself a countable set.

Proof. Let $(A_j)_{j\in J}$ be an indexed family of countable sets indexed by a countable index set J. We aim to prove that the set $A = \bigcup_{j\in J} A_j$ is countable. Without loss of generality, we may assume that A and all A_j are nonempty (if they were empty, they would still be countable). Because the A_j are countable, by Corollary 1.31 (ii) there exist surjective functions $f_j \colon \mathbb{N} \to A_j$ for all $j \in J$. We then define the function $g \colon J \times \mathbb{N} \to A$ with the mapping formula

$$g(j, n) = f_j(n) \text{ for all } j \in J \text{ and } n \in \mathbb{N}.$$

The function g is surjective by construction, since all f_j are surjective. Additionally, the set $J \times \mathbb{N}$, a Cartesian product of two countable sets, is also countable by Proposition 1.32. Because g's domain $j \times \mathbb{N}$ is countable and g is surjective, g's codomain A is also countable by Proposition 1.30 (ii). □

Corollary 1.35: The integers are countable

The set of integers $\mathbb{Z}$ is countably infinite.

Proof. For review, the integers are (informally)[a] defined as the set

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}.$$

We prove the corollary by writing the integers as the union

$$\mathbb{Z} = -\mathbb{N} \cup \{0\} \cup \mathbb{N},$$

where $-\mathbb{N} = \{-1, -2, -3, \ldots\}$. The sets $-\mathbb{N}$, $\{0\}$, and $\mathbb{N}$ are all countable, so $\mathbb{Z}$ is also countable by Proposition 1.34. □

[a] We will define the integers formally in Chapter 2.

Corollary 1.36: The rational numbers are countable

The set of rational numbers $\mathbb{Q}$ is countably infinite.

Proof. For review, the rational numbers are defined as the set

$$\mathbb{Q} = \left\{ \frac{k}{n} \in \mathbb{R};\, k \in \mathbb{Z}, n \in \mathbb{N} \right\},$$

and can be loosely thought of as the set of all ratios of integer numbers. We prove the corollary by writing the rational numbers as the union

$$\mathbb{Q} = \bigcup_{n\in\mathbb{N}} A_n, \text{ where } A_n = \left\{ \frac{k}{n} \in \mathbb{R};\, k \in \mathbb{Z} \right\}.$$

This construction formulates $\mathbb{Q}$ as a countable union of countable sets, so $\mathbb{Q}$ is itself countable by Proposition 1.34. □

Proposition 1.37: The real numbers are not countable

The set of real numbers $\mathbb{R}$ is not countable.

Proof. We will prove the proposition by contradiction. Assume $\mathbb{R}$ is countable. In this case, by Proposition 1.29, the subset $A = \{x \in \mathbb{R};\, 0 \leq x < 1\} \subset \mathbb{R}$ would also be countable; there would then exist bijection between A and $\mathbb{N}$, and so could index A's elements in the sequential form

$$A = \{x_1, x_2, x_3, \ldots\}.$$

We will now use the fact that every real number in A can be uniquely written in the decimal form $0.c_1c_2c_3, \ldots$ where $c_k \in \{0, 1, 2, \ldots, 9\}$, under the condition that 9 does not repeat infinitely from some point onward.[a] All x_n could then be written in the decimal form

$$\begin{aligned} x_1 &= 0.a_{1,1}a_{1,2}a_{1,3}\ldots \\ x_2 &= 0.a_{2,1}a_{2,2}a_{2,3}\ldots \\ x_3 &= 0.a_{3,1}a_{3,2}a_{3,3}\ldots \\ &\vdots \\ x_n &= 0.a_{n,1}a_{n,2}a_{n,3}\ldots \\ &\vdots \end{aligned}$$

We then define, in decimal form, the number $y = 0.b_1b_2b_3\ldots \in A$, where

$$b_k = \begin{cases} 1 & a_{k,k} \neq 1 \\ 2 & a_{k,k} = 1 \end{cases} \text{ for all } k \in \mathbb{N}.$$

By construction, y differs from every x_n at least at the n-th decimal, and is thus different from all x_n. The existence of such a y contradicts the assertion that A could be written in the form $A = \{x_1, x_2, x_3, \ldots\}$, and so the assumption that $\mathbb{R}$ is countable is incorrect. □

[a] We will cover the decimal representation of real numbers in Chapter 2.

2 Numbers

2.1 The natural numbers

The natural numbers, denoted by $\mathbb{N}$, can be informally defined as the set

$$\mathbb{N} = \{1, 2, 3, \ldots\} \qquad \text{(the natural numbers)}.$$

Of course, this definition is somewhat lacking—it neither lists every natural number explicitly, nor does it rigorously formulate how the omitted members of the set are generated (relying instead on the intuitive pattern of "repeatedly adding one").

In this section we will adopt a different approach to defining the natural numbers, namely listing a set of fundamental properties that uniquely define the set of natural numbers (in the sense that only one set satisfies the properties, and we then call the thus-specified set the set of natural numbers). One possible set of these fundamental properties is the Peano axioms, which we state below.

Topic 2.1: The Peano axioms

(P1) The number 1 is a natural number.

(P2) Every natural number n has a unique successor, which we denote by n^+.

(P3) Any two different natural numbers have different successors.

(P4) The number 1 is not the successor of any natural number.

(P5) If A is a subset of the natural numbers, if $1 \in A$, and if $n^+ \in A$ for each $n \in A$, then A contains all the natural numbers.

Note: some formulations of the Peano axioms use 0 (instead of 1) as the "first" natural number, i.e. the only natural number that is not the successor of another natural number.

All other properties of the natural numbers follow from the Peano axioms, including the familiar operations of addition and multiplication, which we state below.

Topic 2.2: Addition and multiplication of natural numbers

Addition of natural numbers follows from the notion of the successor stated in the Peano axioms. We first introduce the following notation for successive natural numbers:

$$2 = 1^+, \quad 3 = 2^+, \quad 4 = 3^+, \text{ and so on;}$$

we are essentially assigning the numeric symbols (i.e. 1, 2, 3, 4, 5, and so on) familiar from everyday life to successive natural numbers defined using the Peano axioms. For an arbitrary natural number $n \in \mathbb{N}$ we then define addition using the rule

$$n + 1 = n^+.$$

The rule $n + 1 = n^+$ can be combined with the notion of a successor to define the sum of any two natural numbers as follows:

$$n + 1 = n^+, \quad n + 2 = (n + 1)^+, \quad n + 3 = (n + 2)^+, \text{ and so on.}$$

We can then use addition of natural numbers to define multiplication as follows:

$$n \cdot 1 = n, \quad n \cdot 2 = (n \cdot 1) + n, \quad n \cdot 3 = (n \cdot 2) + n, \text{ and so on.}$$

Topic 2.3: Proof by induction

The fifth Peano axiom (P5) is called the *induction axiom*, and can be used to prove that some property $\mathcal{P}(n)$ holds for all natural numbers $n \in \mathbb{N}$. This process is called *proof by induction* and follows two steps:

(1) (*The base case*) Show the number 1 has the property $\mathcal{P}$.

(2) (*The induction step*) Show that for every natural number n with the property $\mathcal{P}$, its successor n^+ also has the property $\mathcal{P}$.

Example 2.1: The sum of n numbers

Let us show by induction that

$$1 + 2 + \cdots + n = \frac{n}{2}(n+1) \text{ for all } n \in \mathbb{N}. \tag{2.1}$$

We first verify the base case, which holds because $1 = \frac{1}{2} \cdot (1+1) = 1$ when $n = 1$. We then perform the induction step, first assuming that

$$1 + 2 + \cdots + n = \frac{n}{2}(n+1) \text{ for some } n \in \mathbb{N}.$$

We then add $n+1$ to both sides and rearrange to get

$$\begin{aligned} 1 + 2 + \cdots + n + n + 1 &= \frac{n}{2}(n+1) + (n+1) \\ &= (n+1)\left(\frac{n}{2} + 1\right) \\ &= \frac{(n+1)}{2}(n+2). \end{aligned}$$

The result $1 + 2 + \cdots + n + n + 1 = \frac{(n+1)}{2}(n+2)$ means that if Equation 2.1 holds for n, it also holds for the successor $n^+ = n + 1$, completing the induction step.

2.2 The real numbers

We will define the real numbers in an axiomatic manner analogous to our definition of the natural numbers, i.e. we will list a set of axioms the uniquely specify the set of natural numbers. The end result will be Definition 2.18.

2.2.1 Axiomatic definition of the real numbers

The following axioms define the addition of real numbers.

Topic 2.4: Addition axioms for the real numbers

For any $x, y, z \in \mathbb{R}$, we assume the following axioms:

(A1) Addition of real numbers is associative, i.e.

$$x + (y + z) = (x + y) + z \text{ for all } x, y, z \in \mathbb{R}.$$

(A2) Addition of real numbers is commutative, i.e.

$$x + y = y + x \text{ for all } x, y, z \in \mathbb{R}.$$

(A3) There exists a real number $0 \in \mathbb{R}$, called the *additive identity*, for which

$$x + 0 = 0 + x = x \text{ for all } x \in \mathbb{R}.$$

(A4) For each $x \in \mathbb{R}$ there exists a number $-x \in R$, called the *additive inverse* to the number x, for which

$$x + (-x) = (-x) + x = 0.$$

Axioms (A1)–(A4) can be used to define the subtraction operation, which we define as

$$-: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto x - y = x + (-y) \in \mathbb{R}.$$

The result of the addition of two real numbers is called the numbers' *sum*, and the result of the subtraction of two real numbers is called the numbers' *difference*.

Remark 2.2: Abelian groups

Any set equipped with an operation[a] satisfying Axioms (A1)–(A4) is called an *Abelian group* under the given operation. The real numbers are thus an Abelian group under addition. Meanwhile (for example) the natural numbers are not an Abelian group under addition because they obey neither Axiom (A3) nor Axiom (A4)—no natural number has an additive inverse.

[a] More specifically, an internal binary operation.

Proposition 2.3: Properties of real addition

Addition of real numbers obeys the following properties:

(i) The number $0 \in \mathbb{R}$ is uniquely determined by the condition

$$x + 0 = 0 + x = x \text{ for all } x \in \mathbb{R}.$$

(ii) For all $x \in \mathbb{R}$, the number $-x \in \mathbb{R}$ is uniquely determined by the condition

$$x + (-x) = (-x) + x = 0.$$

(iii) The number 0 is its own additive inverse, i.e. $-0 = 0$.

(iv) Negation is distributive over addition, i.e.

$$-(x + y) = (-y) + (-x) \text{ for all } x, y \in \mathbb{R}.$$

(v) Negation is an involutory operation, i.e. $-(-x) = x$ for all $x \in \mathbb{R}$.

(vi) If $x + y = z + y$ for all $x, y, z \in R$ then $x = z$. This property is sometimes called the *the cancellation property for addition*.

(vii) For all $a, b \in \mathbb{R}$, the equation $x + a = b$ for the unknown real number $x \in \mathbb{R}$ has the unique solution $x = b - a$.

Proof. Each of these properties fundamentally follows from Axioms (A1)–(A4).

(i) Assume $0' \in \mathbb{R}$ were an additional additive identity element (in addition to $0 \in \mathbb{R}$) for which $x + 0' = 0' + x = x$ for all $x \in \mathbb{R}$. In this case, combining the additive identity equations for 0 and $0'$, we see that $0' = 0' + 0 = 0$. Evidently $0' = 0$ and the additive identity is unique.

(ii) Assume $\widetilde{x} \in \mathbb{R}$ were an additional additive inverse element (in addition to $-x \in \mathbb{R}$) for which $x + \widetilde{x} = \widetilde{x} + x = 0$ for all $x \in \mathbb{R}$. We then combine Axioms (A1), (A3), and (A4) to conclude that

$$\begin{aligned}\widetilde{x} &\overset{\text{(a)}}{=} \widetilde{x} + 0 \overset{\text{(b)}}{=} \widetilde{x} + \big(x + (-x)\big) \\ &\overset{\text{(c)}}{=} (\widetilde{x} + x) + (-x) \\ &\overset{\text{(d)}}{=} 0 + (-x) \overset{\text{(e)}}{=} -x,\end{aligned}$$

where (a) and (e) use (A3), (b) uses (A4), (c) uses (A1), and (d) holds by (A4) assuming $\widetilde{x}$ is x's additive inverse. Evidently $\widetilde{x} = -x$ and the additive inverse is unique.

(iii) We begin with the trivial equality $0 = 0$, then apply (A3) (with $x = 0$) to get $0 = 0 + 0$. By point (ii) above, the additive inverse $(-x)$ for which any real number x obeys $0 = x + (-x)$ is unique, so the RHS of the equality $0 = 0 + 0$ implies $0 = -0$.

(iv) We combine (A1), (A3), and (A4) to get

$$\begin{aligned}(x + y) + \big[(-y) + (-x)\big] &= x + \big[y + (-y)\big] + (-x) = x + 0 + (-x) \\ &= x + (-x) = 0 \overset{\text{(a)}}{=} (-y) + y = (-y) + 0 + y \\ &= (-y) + \big[(-x) + x\big] + y \\ &= \big[(-y) + (-x)\big] + (x + y),\end{aligned}$$

where the steps from (a) onward are simply a reversal of the steps so far. The result of this calculation is the equality

$$(x + y) + \big[(-y) + (-x)\big] = 0 = \big[(-y) + (-x)\big] + (x + y),$$

which we combine with the uniqueness of the additive inverse in point (ii) to conclude that $-(x + y) = (-y) + (-x)$.

(v) The property $-(-x) = x$ follows from the equality

$$x + (-x) = (-x) + x = 0 \text{ for all } x \in \mathbb{R}$$

(which implies that x plays the role of $-(-x)$) and the fact that the additive inverse is unique.

(vi) We begin with the assumed equality $x + y = z + y$ and then apply Axioms (A1), (A3), and (A4) to conclude that

$$\begin{aligned} x = x + 0 &= x + \big[y + (-y)\big] = (x + y) + (-y) \\ &\overset{\text{(a)}}{=} (z + y) + (-y) = z + \big[y + (-y)\big] \\ &= z + 0 = z, \end{aligned}$$

where (a) uses $x + y = z + y$.

(vii) We will first show that $b - a$ is a solution of $x + a = b$, then show it is the *only* solution. We begin with the calculation

$$(b - a) + a \overset{\text{(a)}}{=} \big[b + (-a)\big] + a = b + \big[(-a) + a\big] = b + 0 = b,$$

where (a) uses the definition of subtraction. The result $(b - a) + a = b$ means that $(b - a)$ is a solution of $x + a = b$.

We now show $b - a$ is the *only* solution $x + a = b$. Let $y \in \mathbb{R}$ be an arbitrary solution of the equation, so that

$$y + a = b \overset{\text{(a)}}{=} (b - a) + a,$$

where (a) holds because $(b - a)$ is also a solution of the equation. We then apply the cancellation property for addition (vi) to conclude that $y = b - a$, meaning $b - a$ is the only solution to $x + a = b$. □

The following axioms define the multiplication of real numbers.

Topic 2.5: Multiplication axioms for the real numbers

For any $x, y, z \in \mathbb{R}$ we assume the following axioms:

(A5) Multiplication of real numbers is associative, i.e.

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z \text{ for all } x, y, z \in \mathbb{R}.$$

(A6) Multiplication of real numbers is commutative, i.e.

$$x \cdot y = y \cdot x \text{ for all } x, y \in \mathbb{R}.$$

(A7) There exists a real number $1 \in \mathbb{R}$, called the *multiplicative identity*, for which

$$x \cdot 1 = 1 \cdot x = x \text{ for all } x \in \mathbb{R}.$$

(A8) For each $x \in \mathbb{R} \setminus \{0\}$, there exists a number $x^{-1} \in \mathbb{R}$, called the *multiplicative inverse* to the number x, for which

$$x \cdot x^{-1} = x^{-1} \cdot x = 1.$$

Axioms (A5)–(A8) can be used to define the subtraction operation, which we define, for all

$x \in \mathbb{R}$ and $y \in \mathbb{R} \setminus \{0\}$, as

$$\div : \mathbb{R} \times \mathbb{R} \setminus \{0\} \to \mathbb{R}, \quad (x, y) \mapsto x \div y = x \cdot y^{-1}.$$

Two other common notations for the division operation exist; these are

$$x \div y = x/y = \frac{x}{y} \qquad \text{(equivalent notations for division)}.$$

The result of the multiplication of two real numbers is called the numbers' *product*, and the result of the division of two real numbers is called the numbers' *quotient*.

Topic 2.6: Integer powers of real numbers

For each $x \in \mathbb{R}$ we define

$$\begin{aligned} x^1 &= x && (x \text{ raised to the first power}) \\ x^2 &= x^1 \cdot x && (x \text{ raised to the second power}) \\ x^3 &= x^2 \cdot x && (x \text{ raised to the third power}) \\ &\vdots && (\text{and so on...}); \end{aligned}$$

more generally, for each $n \in \mathbb{N}$ we recursively define the number x^{n+1} as

$$x^{n+1} = x^n \cdot x.$$

Similarly, now considering negative powers, for each $x \in \mathbb{R}$ we define

$$\begin{aligned} x^0 &= 1 && (x \text{ raised to the zeroth power}) \\ x^{-2} &= x^{-1} \cdot x^{-1} && (x \text{ raised to the negative first power}) \\ x^{-3} &= x^{-2} \cdot x^{-1} && (x \text{ raised to the negative second power}) \\ &\vdots && (\text{and so on...}). \end{aligned}$$

Note that we have not written x^{-1}, since this is already defined in (A8) as the multiplicative inverse to x. More generally, for each $n \in \mathbb{N}$ we recursively define

$$x^{-(n+1)} = x^{-n} \cdot x^{-1}.$$

Remark 2.4: Comparing the addition and multiplication axioms

Axioms (A1)–(A4) are similar to Axioms (A5)–(A8). There are two important differences, however:

- The number 0 does not have a multiplicative inverse, so the real numbers are not an Abelian group under multiplication (the number 0 *does* have an additive inverse).
- Division by zero is not defined for the real numbers (subtraction by zero *is* well-defined).

We now state two axioms that relate addition and multiplication of real numbers.

Topic 2.7: Axioms relating addition and multiplication

We assume the following axioms:

(A9) The numbers 1 and 0 are different real numbers, i.e. $1 \neq 0$.

(A10) Multiplication of real numbers is distributive over addition, i.e.

$$x \cdot (y + z) = (x \cdot y) + (x \cdot z) \text{ for all } x, y, z \in \mathbb{R}.$$

Proposition 2.5: Multiplication by zero

Multiplication by zero obeys the following two properties:

(i) $x \cdot 0 = 0$ for all $x \in \mathbb{R}$.

(ii) If $x \cdot y = 0$ holds for any two $x, y \in \mathbb{R}$, then either $x = 0$ or $y = 0$ (or both).

Proof.

(i) We begin by applying (A10) to the quantity $x \cdot 0 + x \cdot 1$ to get

$$x \cdot 0 + x \cdot 1 \overset{\text{(a)}}{=} x \cdot (0 + 1) \overset{\text{(b)}}{=} x \cdot 1 \overset{\text{(c)}}{=} 0 + x \cdot 1,$$

where (a) uses (A10) and (b) and (c) use (A3). We then apply the cancellation property Proposition 2.3 (vi) to the equation $x \cdot 0 + x \cdot 1 = 0 + x \cdot 1$ to get $x \cdot 0 = 0$.

(ii) Suppose $x \cdot y = 0$ and $x \neq 0$; we must then show that $y = 0$. Because $x \neq 0$, by (A8) there exists a multiplicative inverse x^{-1} for which $x^{-1} \cdot x = 1$. We then make the calculation

$$0 \overset{\text{(a)}}{=} x^{-1} \cdot 0 \overset{\text{(b)}}{=} x^{-1} \cdot (x \cdot y) = (x^{-1} \cdot x) \cdot y = 1 \cdot y = y,$$

where (a) uses point (i) above and (b) uses the assumed equality $x \cdot y = 0$. The proof is analogous if $x = 0$ and $y \neq 0$ and trivial if $x = y = 0$. □

Note: Proposition 2.5 together with Axioms (A5)–(A8) mean that the set $\mathbb{R} \setminus \{0\}$ is an Abelian group under multiplication (refer to Remark 2.2 for review).

We now list some properties of real multiplication analogous to the properties of real addition given in Proposition 2.3.

Proposition 2.6: Properties of real multiplication

Multiplication of real numbers obeys the following properties:

(i) The multiplicative identity 1 is uniquely defined by the condition

$$x \cdot 1 = 1 \cdot x = x \text{ for all } x \in \mathbb{R}.$$

(ii) For each $x \in \mathbb{R} \setminus \{0\}$, the multiplicative inverse x^{-1} is uniquely defined by the condition

$$x \cdot x^{-1} = x^{-1} \cdot x = 1.$$

(iii) The number 1 is its own multiplicative inverse, i.e. $1^{-1} = 1$.

(iv) Taking the multiplicative inverse is distributive over multiplication, i.e.

$$(x \cdot y)^{-1} = y^{-1} \cdot x^{-1} \text{ for all } x, y \in \mathbb{R} \setminus \{0\}.$$

This property relies on the fact that $x \cdot y \neq 0$ for all $x, y, \in \mathbb{R} \setminus \{0\}$, since the quantity $(x \cdot y)^{-1}$ would otherwise not be well-defined.

(v) Taking the multiplicative inverse is an involutory operation, i.e.

$$(x^{-1})^{-1} = x \text{ for all } x \in \mathbb{R} \setminus \{0\}.$$

(vi) If $x \cdot y = z \cdot y$ for $x, z \in \mathbb{R}$ and $y \in \mathbb{R} \setminus \{0\}$ then $x = z$. This property is sometimes called *the cancellation property for multiplication.*

(vii) For all $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$, the equation $x \cdot a = b$ for the unknown number $x \in \mathbb{R}$ has the unique solution $x = b/a$.

Proof. The proof is completely analogous to the proof of Proposition 2.3 (one can almost proceed by blindly replacing addition with multiplication, 0 with 1, and the additive inverse with the multiplicative inverse); we give the full proofs here only for the sake of completeness.

(i) Assume $1' \in \mathbb{R}$ were an additional multiplicative identity element (in addition to $1 \in \mathbb{R}$) for which $x \cdot 1' = 1' \cdot x = x$ for all $x \in \mathbb{R}$. In this case, combining the multiplicative identity equations for 1 and $1'$, we see that $1' = 1' \cdot 1 = 1$. Evidently $1' = 1$ and the multiplicative identity is unique.

(ii) Assume $\widetilde{x} \in \mathbb{R}$ were an additional multiplicative inverse element (in addition to $x^{-1} \in \mathbb{R}$) for which $x \cdot \widetilde{x} = \widetilde{x} \cdot x = 1$ for all $x \in \mathbb{R} \setminus \{0\}$. We then prove this point with the calculation

$$\begin{aligned} \widetilde{x} &\overset{\text{(a)}}{=} \widetilde{x} \cdot 1 \overset{\text{(b)}}{=} \widetilde{x} \cdot \left(x \cdot x^{-1}\right) \\ &\overset{\text{(c)}}{=} (\widetilde{x} \cdot x) \cdot x^{-1} \\ &\overset{\text{(d)}}{=} 1 \cdot x^{-1} \overset{\text{(e)}}{=} x^{-1}, \end{aligned}$$

where (a) and (e) use (A7), (b) uses (A8), (c) uses (A5), and (d) holds by (A8) assuming $\widetilde{x}$ is x's multiplicative inverse. Evidently $\widetilde{x} = -x$ and the multiplicative inverse is unique.

(iii) We begin with the trivial equality $1 = 1$, then apply (A7) (with $x = 1$) to get $1 = 1 \cdot 1$. By point (i) above, the multiplicative inverse x^{-1} for which any real number $x \in \backslash\mathbb{R} \setminus \{0\}$ obeys $1 = x \cdot x^{-1}$ is unique, so the RHS of the equality $1 = 1 \cdot 1$ implies $1 = 1^{-1}$.

(iv) We combine (A5), (A7), and (A8) to get

$$\begin{aligned} (x \cdot y) \cdot (y^{-1} \cdot x^{-1}) &= x \cdot (y \cdot y^{-1}) \cdot x^{-1} = x \cdot 1 \cdot x^{-1} \\ &= x \cdot x^{-1} = 1 \overset{\text{(a)}}{=} y^{-1} \cdot y = y^{-1} \cdot 1 \cdot y \\ &= y^{-1} \cdot (x^{-1} \cdot x) \cdot y \\ &= (y^{-1} \cdot x^{-1}) \cdot (x \cdot y), \end{aligned}$$

where the steps from (a) onward are simply a reversal of the steps so far. The result of this calculation is the equality

$$(x \cdot y) \cdot (y^{-1} \cdot x^{-1}) = 1 = (y^{-1} \cdot x^{-1}) \cdot (x \cdot y),$$

which we combine with the uniqueness of the multiplicative inverse in point (i) to conclude that $(x \cdot y)^{-1} = y^{-1} \cdot x^{-1}$.

(v) The property $(x^{-1})^{-1} = x$ follows from the equality

$$x \cdot x^{-1} = x^{-1} \cdot x = 1 \text{ for all } x \in \mathbb{R} \setminus \{0\}$$

(which implies that x plays the role of $(x^{-1})^{-1}$) and the fact that the multiplicative inverse is unique.

(vi) We begin with the assumed equality $x \cdot y = z \cdot y$, then apply Axioms (A5), (A7), and (A8) to conclude that

$$\begin{aligned} x = x \cdot 1 &= x \cdot (y \cdot y^{-1}) = (x \cdot y) \cdot y^{-1} \\ &\overset{\text{(a)}}{=} (z \cdot y) \cdot y^{-1} = z \cdot (y \cdot y^{-1}) \\ &= z \cdot 1 = z, \end{aligned}$$

where (a) uses $x \cdot y = z \cdot y$.

(vii) We will first show that b/a is a solution of $x \cdot a = b$, then show it is the *only* solution. We begin with the calculation

$$(b/a) \cdot a \overset{\text{(a)}}{=} (b \cdot a^{-1}) \cdot a = b \cdot (a^{-1} \cdot a) = b \cdot 1 = b,$$

where (a) uses the definition of division. The result $(b/a) \cdot a = b$ means that b/a is a solution of $x \cdot a = b$.

We now show b/a is the *only* solution $x \cdot a = b$. Let $y \in \mathbb{R}$ be an arbitrary solution of the equation, so that

$$y \cdot a = b \overset{\text{(a)}}{=} (b/a) \cdot a,$$

where (a) holds because b/a is also a solution of the equation. We then apply the cancellation property for multiplication (vi) to conclude that $y = b/a$, meaning b/a is the only solution to $x \cdot a = b$. □

Remark 2.7: Commutative rings

A set equipped with two operations[a] satisfying Axioms (A1)–(A10) is called a *commutative ring*. The real numbers are thus a commutative ring under real addition and multiplication. The integers are also a commutative ring under integer addition and multiplication.

[a] More specifically, two internal binary operations.

We now list some properties of the real numbers that follow as consequences of Axioms (A1)–(A10); in fact, the general forms of the following properties hold for any commutative ring,

since the properties fundamentally follow from Axioms (A1)–(A10).

Proposition 2.8: Consequences of Axioms (A1)–(A10)

The following properties hold for any $x, y, z \in \mathbb{R}$:

(i) Multiplication is distributive over subtraction, i.e.

$$x \cdot (y - z) = x \cdot y - x \cdot z.$$

(ii) The order of negation and multiplication is interchangeable, i.e.

$$x \cdot (-z) = -(x \cdot z).$$

(iii) Division is distributive over addition, i.e.

$$x/y + z/y = (x + z)/y \text{ if } y \in \mathbb{R} \setminus \{0\}.$$

Proof.

(i) We begin with the calculation

$$x \cdot (y - z) + x \cdot z \overset{\text{(a)}}{=} x \cdot \big[(y - z) + z\big] = x \cdot (y + 0) = x \cdot y,$$

where (a) uses (A10). We then apply Proposition 2.3 (vii) to the resulting equation $x \cdot (y - z) + x \cdot z = x \cdot y$ to conclude that $x \cdot (y - z) = x \cdot y - x \cdot z$.

(ii) This point follows directly from (i) if we choose $y = 0$.

(iii) This point follows directly from the calculation

$$x/y + z/y = x \cdot y^{-1} + z \cdot y^{-1} \overset{\text{(a)}}{=} (x + z) \cdot y^{-1} = (x + z)/y,$$

where (a) follows from (A10) and y^{-1} exists because $y \in \mathbb{R} \setminus \{0\}$. □

The following two axioms serve to define the positive real numbers.

Topic 2.8: Axioms for the positive real numbers

(A11) The real number $0 \in \mathbb{R}$ is not an element of $\mathbb{R}^+$; for any other real number $x \in \mathbb{R} \setminus \{0\}$, either $x \in \mathbb{R}^+$ or $-x \in \mathbb{R}^+$, but not both.

(A12) The positive real numbers $\mathbb{R}^+$ are closed under addition and multiplication, i.e.

$$x + y \in \mathbb{R}^+ \quad \text{and} \quad x \cdot y \in \mathbb{R}^+ \text{ for all } x, y \in \mathbb{R}^+.$$

The positive real numbers can be used to define the negative real numbers as follows.

Definition 2.9: The negative real numbers

The negative real numbers, denoted by $\mathbb{R}^-$, are defined as the set

$$\mathbb{R}^- = \{-x;\, x \in \mathbb{R}^+\}.$$

By Axiom (A11), the positive and negative real numbers are related as follows:

(i) $\mathbb{R} = \mathbb{R}^+ \cup \mathbb{R}^- \cup \{0\}$.

(ii) The sets $\mathbb{R}^+$, $\mathbb{R}^-$, and $\{0\}$ are all mutually disjoint.

In words other words, for every real number $x \in \mathbb{R}$, exactly one of the following options is possible: x is positive, negative, or zero. These properties in turn make it possible to define a natural ordering for the real numbers, which we formalize below.

Definition 2.10: Ordering the real numbers

Let $x, y \in \mathbb{R}$ be two arbitrary real numbers.

- The number x is said to be *less than* y if $y - x \in \mathbb{R}^+$. In this case we write

$$x < y \qquad \text{(notation: } x \text{ is less than } y\text{)}.$$

  Equivalently, we may also say that y is *greater than* x, denoted by $y > x$.

- The number x is said to be *less than or equal to* y if $y - x \in \mathbb{R}^+ \cup \{0\}$. In this case we write

$$x \leq y \qquad \text{(notation: } x \text{ is less than or equal to } y\text{)}.$$

  Equivalently, we may say that y is *greater than or equal to* x, denoted by $y \geq x$.

- A real number $z \in \mathbb{R}$ is said to be *positive* if $0 < z$ and *negative* if $z < 0$.

Remark 2.11: Ordered rings

A set equipped with two internal binary operations satisfying Axioms (A1)–(A12) is called an *ordered ring*. The real numbers are thus an ordered ring under real addition and multiplication. The integers are also an ordered ring under integer addition and multiplication.

We now list some properties of the real numbers that follow as consequences of Axioms (A1)–(A12); these axioms mostly concern the ordering of the real numbers. In fact, the general forms of the following properties hold for any commutative ring, since the properties fundamentally follow from Axioms (A1)–(A12).

Proposition 2.12: Consequences of Axioms (A1)–(A12)

The following properties hold for all $x, y, z, w \in \mathbb{R}$:

(i) The ordering of real numbers is a transitive relation, i.e.

$$\text{if } x < y \text{ and } y < z \text{ then } x < z.$$

(ii) (*The law of trichotomy*) Exactly one of the following three possibilities holds: either $x < y$, or $y < x$, or $x = y$.

(iii) If $x < y$, then $x + z < y + z$.

(iv) If $x < y$ and $z < w$, then $x + z < y + w$.

(v) If $x < y$ and $0 < z$, then $x \cdot z < y \cdot z$.

(vi) If $x < 0$ and $y < 0$, then $0 < x \cdot y$.

(vii) If $x < 0$ and $0 < y$, then $x \cdot y < 0$.

(viii) If $x \neq 0$, then $0 < x \cdot x$.

(ix) It holds that $0 < 1$.

Proof.

(i) Since $x < y$ and $y < z$, by definition $y - x \in \mathbb{R}^+$ and $z - y \in \mathbb{R}^+$. Axiom (A12) then implies that $(y - x) + (z - y) = z - x \in \mathbb{R}^+$, and so $x > z$.

(ii) This point follows directly from (A11) and the ordering of the real numbers in Definition 2.10.

(iii) This point follows directly from the equation

$$y - x = y + (z - z) - x = (y + z) - (x + z).$$

Since $x < y$ and thus $y - x \in \mathbb{R}^+$, it follows that $(y + z) - (x + z) \in \mathbb{R}^+$ and thus $x + z < y + z$.

(iv) Point (iii) implies $x + z < y + z$ and $y + z < y + w$; and point (i) in turn implies $x + z < y + w$.

(v) The inequalities $x < y$ and $0 < z$ mean $y - x \in \mathbb{R}^+$ and $z \in \mathbb{R}^+$. Axiom (A12) implies that $(y - x) \cdot z = y \cdot z - x \cdot z \in \mathbb{R}^+$, from which follows $x \cdot z < y \cdot z$.

(vi) The inequalities $x < 0$ and $y < 0$ mean that $x, y \in \mathbb{R}^-$, and so $(-x), (-y) \in \mathbb{R}^+$ by the definition of the negative real numbers. Axiom (A12) then implies that $(-x) \cdot (-y) \in \mathbb{R}^+$. We then conclude the proof with the calculation

$$x \cdot y \overset{\text{(a)}}{=} -\big[-(x \cdot y)\big] \overset{\text{(b)}}{=} -\big[(-x) \cdot y\big] \overset{\text{(c)}}{=} (-x) \cdot (-y) \in \mathbb{R}^+,$$

where (a) uses Proposition 2.3 (v) and (b) and (c) use Proposition 2.8 (ii).

(vii) The inequalities $x < 0$ and $0 < y$ mean that $(-x) \in \mathbb{R}^+$ and $y \in \mathbb{R}^+$, so $(-x) \cdot y \in \mathbb{R}^+$ by (A12). We then apply Proposition 2.8 (ii) to get

$$-(x \cdot y) = (-x) \cdot y \in \mathbb{R}^+;$$

the result $-(x \cdot y) \in \mathbb{R}^+$ means that $(x \cdot y) \in \mathbb{R}^-$.

(viii) If $0 < x$, then $0 < x \cdot x$ follows directly from Axiom (A12), while if $x < 0$, then $0 < x \cdot x$ follows from point (vi).

(ix) This point follows from the inequality

$$0 \overset{\text{(a)}}{<} 1 \cdot 1 = 1$$

where (a) follows from point (viii). □

We are now equipped to introduce the important concept of boundedness, which will feature

centrally throughout the remainder of this book.

Definition 2.13: Boundedness in the real numbers

Let $A \subset \mathbb{R}$ be an arbitrary subset of the real numbers.

- The subset A is said to be *bounded above* if there exists a number $M \in \mathbb{R}$ for which
$$x \leq M \text{ for all } x \in A.$$
If such a number M exists, it is called an *upper bound* of the set A.
- The subset A is said to be *bounded below* if there exists a number $m \in \mathbb{R}$ for which
$$m \leq x \text{ for all } x \in A.$$
If such a number m exists, it is called a *lower bound* of the set A.
- The subset A is said to be *bounded* if it is bounded both above and below.

The definition of boundedness leads naturally to the concept of the supremum and infimum of a set of real numbers, defined immediately below.

Definition 2.14: Supremum and infimum of a subset of $\mathbb{R}$

Let $A \subset \mathbb{R}$ be an arbitrary subset of the real numbers.

- The number $M \in \mathbb{R}$ is called the *least upper bound* or *supremum* of A if:
  (1) the number M is an upper bound of A, and
  (2) if $M' \in \mathbb{R}$ is also an upper bound of A, then $M \leq M'$.
  If it exists, A's supremum is denoted by $\sup(A)$.
- The number $m \in \mathbb{R}$ is called the *greatest lower bound* or *infimum* of A if:
  (1) the number m is a lower bound of A, and
  (2) if $m' \in \mathbb{R}$ is also a lower bound of A, then $m' \leq m$.
  If it exists, A's infimum is denoted by $\inf(A)$.

Directly from the definition, a set can have at most one supremum and at most one infimum.

Definition 2.15: Maximum and minimum of a subset of $\mathbb{R}$

Let $A \subset \mathbb{R}$ be an arbitrary subset of the real numbers.

- If A has a supremum and $\sup(A) \in A$, then $\sup(A)$ is called the *maximum* of the set A and is denoted by
$$\max(A) = \sup(A) \qquad \text{(maximum of } A\text{, if it exists)}.$$
- If A has an infimum and $\inf(A) \in A$, then $\inf(A)$ is called the *minimum* of the set A and is denoted by
$$\min(A) = \inf(A) \qquad \text{(minimum of } A\text{, if it exists)}.$$

The maximum and minimum of a set are thus closely related to the set's supremum and infimum; the important distinction is that a maximum or minimum must be *in* the set, while the supremum and infimum do not have this restriction. Thus every maximum is also a supremum and every minimum is an also an infimum, but not vice versa, and it is perfectly reasonable for a set to have a supremum (or infimum) but not a maximum (or minimum).

We now offer some examples illustrating concepts related to boundedness.

Example 2.16: Boundedness

(i) The set $A = \{x \in \mathbb{R}\,;\, 0 < x < 1\}$ is bounded—all real numbers greater than or equal to 1 are upper bounds of A, while all real numbers less than or equal to 0 are lower bounds of A. The set A's supremum and infimum are thus 1 and 0, respectively. Because 1 and 0 are not in A, A does not have a maximum or minimum.

However, 1 and 0 *are* the maximum and minimum of the slightly modified set $A' = \{x \in \mathbb{R}\,;\, 0 \leq x \leq 1\}$, since the numbers 0 and 1 are both members of A'.

(ii) The set $B = \{x \in \mathbb{R}\,;\, 0 \leq x\}$ is not bounded above but is bounded below; it thus has no supremum, while its infimum and minimum are $\inf(B) = \min(B) = 0$.

(iii) The empty set $\emptyset$ is bounded both above and below; every real number is both an upper and lower bound of $\emptyset$, so $\emptyset$ has neither a supremum or infimum.

(iv) For an arbitrary subset $A \subset \mathbb{R}$, let $-A = \{-x\,;\, x \in A\}$. The set A is bounded above (or below) if, and only if, the set $-A$ is bounded below (or above). If A has a supremum (or infimum), then $-A$ has an infimum (or supremum) for which $\inf(-A) = -\sup(-A)$ (or $\sup(-A) = -\inf(A)$).

Having introduced boundedness, we are not prepared to state the final axiom describing the real numbers.

Topic 2.9: The Dedekind axiom

(A13) Every nonempty, above-bounded subset of the real numbers has a supremum.

Axioms (A1)–(A13) are satisfied by only a single ring, which we choose to call the real numbers.

Corollary 2.17: Below-bounded sets have an infimum

Every nonempty, below-bounded subset of the real numbers has an infimum.

Proof. Let $A \in \mathbb{R}$ be a nonempty, below-bounded subset of the real numbers. In this case the set $-A = \{-x\,;\, x \in A\}$ is nonempty and bounded above and so has a least upper bound $\sup(-A)$ by Axiom (A13). Thus (see e.g. Example 2.16 (iv)) the original set A has a greatest lower bound equal to $\inf(A) = -\sup(-A)$. □

We conclude the section with the following definition:

Definition 2.18: The real numbers

The *real numbers* are the set $\mathbb{R}$ together with the subset $\mathbb{R}^+ \subset \mathbb{R}$, equipped with operation of real addition, defined as

$$+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto x + y,$$

the operation of real multiplication, defined as

$$\cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad (x, y) \mapsto x \cdot y = xy,$$

and characterized by Axioms (A1)–(A13) described in the sections below.

2.2.2 Important subsets of the real numbers

In this short section we will show that the real numbers contain the natural numbers, the integer numbers, and the rational numbers.

Topic 2.10: The natural, integer, and rational numbers in the real numbers

The natural numbers are constructed in the context of the real numbers as follows:

- The first natural number corresponds to the multiplicative identity $1 \in \mathbb{R}$.
- The successor operation for the natural numbers is defined as real addition by $1 \in \mathbb{R}$. This formulation leads to the sequence

  $$2 = 1 + 1, \quad 3 = 2 + 1, \quad 4 = 3 + 1, \quad \text{and so on,}$$

  which agrees with the construction $2 = 1^+$, $3 = 2^+$, etc. used in Topic 2.2.

The thus-defined set $\mathbb{N} = \{1, 2, 3, \ldots\} \subset \mathbb{R}$ satisfies the Peano axioms.

Having established the natural numbers as a subset of $\mathbb{R}$, we construct the integers in the context the real numbers as

$$\mathbb{Z} = -\mathbb{N} \cup \{0\} \cup \mathbb{N} = \{\ldots, -2, -1, 0, 1, 2, \ldots\} \subset \mathbb{R}.$$

Having established the integers as a subset of $\mathbb{R}$, we construct the rational numbers under the scope of the real numbers as

$$\mathbb{Q} = \left\{ \frac{k}{n} \in \mathbb{R}\,;\, k \in \mathbb{Z}, n \in \mathbb{N} \right\} \subset \mathbb{R}.$$

For integers $k, k' \in \mathbb{Z}$ and natural numbers $n, n' \in \mathbb{N}$, the fractions k/n and k'/n' correspond to the same rational number if, and only if, $k \cdot n' = k' \cdot n$.

2.2.3 Consequences of the Dedekind axiom

In this section we explore some properties of boundedness that follow from the Dedekind axiom.

Proposition 2.19: Boundedness of the natural numbers

The set of natural numbers $\mathbb{N} \subset \mathbb{R}$ is not bounded above.

Proof. We will prove the proposition by contradiction. Suppose $\mathbb{N}$ were bounded above. By Dedekind's axiom (A13), there would then exist a supremum $M \in \mathbb{R}$ of the set $\mathbb{N}$. Since, in the context of the real numbers, the natural numbers are separated by successive steps of the number 1, there would then exist a natural number $n \in \mathbb{N}$ for which $M - 1 < n$, and thus $M < n + 1$. But $n + 1$ is also a natural number, which contradicts the assumption that $\mathbb{N}$ is bounded above by M. □

Corollary 2.20: The integers are unbounded

The set of integer numbers $\mathbb{Z} \subset \mathbb{R}$ is neither bounded above nor bounded below.

Proof. By Proposition 2.19, the set $\mathbb{N}$ is not bounded above, and so the set $-\mathbb{N}$ is not bounded below (see the discussion in Example 2.16 (iv)). Thus the integers, which are constructed as $\mathbb{Z} = -\mathbb{N} \cup \{0\} \cup \mathbb{N}$, are bounded neither above nor below. □

Proposition 2.21: Relating the natural and real numbers

The following three properties help relate the natural and real numbers.

(i) For each $x \in \mathbb{R}$ there exists a natural number $n \in \mathbb{N}$ for which $x < n$.

(ii) For any two $a, b \in \mathbb{R}^+$ there exists a natural number $n \in \mathbb{N}$ for which $b < na$.

(iii) For each $a \in \mathbb{R}^+$ there exists a natural number $n \in \mathbb{N}$ for which $1/n < a$.

Proof.

(i) The set $\mathbb{N}$ is not bounded above by Proposition 2.19, so the number x is not an upper bound of $\mathbb{N}$. There thus exists some $n \in \mathbb{N}$ for which $x < n$.

(ii) By point (i), there exists $n \in \mathbb{N}$ for which $b/a < n$. Because $a, b \in \mathbb{R}^+$, it then follows that $b < na$.

(iii) This point is a special case of point (ii) if we let $b = 1$ and rearrange. □

We now offer some concrete examples and discussion of bounded sets.

Example 2.22: Bounded sets

(1) The set $A = \{1/n \in \mathbb{R};\, n \in \mathbb{N}\}$ is bounded both above and below, and its supremum is $\sup(A) = 1$. Evidently 0 is a lower bound of A, and, from the combination of Proposition 2.21 (iii) and the fact that 0 is the largest real number not in $\mathbb{R}^+$, 0 is in fact the greatest upper bound of A.

(2) For any real number $q > 1$, the set $B = \{q^n;\, n \in \mathbb{N}\}$ is not bounded above.

We can prove this claim by contradiction: suppose B were bounded above; it would then have a supremum $\sup(B) = M$. Because $M/q < M$, there would exist some $n \in \mathbb{N}$ for which $M/q < q^n \leq M$. Multiplying the inequality

through by q produces $M < q^{n+1}$, which contradicts $\sup(B) = M$. Thus B is not bounded above.

(3) For any positive real number $q \in \mathbb{R}^+$ for which $q < 1$, the set $C = \{q^n ; n \in \mathbb{N}\}$ is bounded below with infimum $\inf(C) = 0$.

Proof: assuming $q \in \mathbb{R}^+$ is less than one, then by point (2) the set $\{1/q^n ; n \in \mathbb{N}\}$ is not bounded above. Thus for each *positive* $\epsilon \in \mathbb{R}^+$ there exists some $n \in \mathbb{N}$ for which $1/\epsilon < 1/q^n$ and thus $q^n < \epsilon$. The set C's infimum must then be 0, i.e. the largest nonpositive real number.

(4) For any two sets $A, B \subset \mathbb{R}$, let us define the set $A + B$ as

$$A + B = \{a + b ; a \in A, b \in B\} \subset \mathbb{R}.$$

In this case:

(a) If both A and B are bounded above and nonempty, the set $A + B$ is also bounded above and nonempty, and obeys

$$\sup(A + B) = \sup(A) + \sup(B).$$

(b) If both A and B are bounded below and nonempty, the set $A + B$ is also bounded below and nonempty, and obeys

$$\inf(A + B) = \inf(A) + \inf(B).$$

(c) If A and B are both nonempty and $a < b$ for all $a \in A$ and all $b \in B$, then A is bounded above, B is bounded below, and $\sup(A) \leq \inf(B)$.

Proof.

(a) We first note that $a \leq \sup(A)$ and $b \leq \sup(B)$ for all $a \in A$ and all $b \in B$, so by Proposition 2.12 (iv)

$$a + b \leq \sup(A) + \sup(B) \text{ for all } a \in A \text{ and } b \in B,$$

and so $\sup(A) + \sup(B)$ is *an* upper bound of $A + B$.

To show that $\sup(A) + \sup(B)$ is $(A + B)$'s *least* upper bound, let $c \in \mathbb{R}$ be an arbitrary real number for which

$$c < \sup(A) + \sup(B).$$

Simple subtraction gives $c - \sup(A) < \sup(B)$, which by the definition of $\sup(B)$ means there exists some $b \in B$ for which $c - \sup(A) < b$ or $c - b < \sup(A)$.

Analogously, subtraction gives $c - \sup(B) < \sup(A)$, or, in terms of the above introduced b, $c - b < \sup(A)$. By the definition of $\sup(A)$ there exists some $a \in A$ for which $c - b < a$ or $c < a + b$. Because $c < a + b$, c is not an upper bound of $A + B$, so $\sup(A) + \sup(B)$ is indeed the least upper bound of the set $A + B$.

(b) Analogous to the proof of point (a).

(c) This point follows directly from the definition of boundedness, the supremum, and the infimum together with the stated conditions on the sets A and B. □

2.2.4 Representing the real numbers in decimal form

The decimal representation of a nonnegative real number $a \in \mathbb{R}^+ \cup \{0\}$ takes the form

$$a = a_0.a_1a_2a_3\ldots \qquad \text{(decimal representation of } a\text{)},$$

where $a_0 \in \mathbb{N} \cup \{0\}$ and $a_1, a_2, a_3, \ldots \in \{0, 1, \ldots, 9\}$. In this context, the numbers $\{a_1, a_2, a_3, \ldots\}$ are called *decimal digits*.

Topic 2.11: Every nonnegative real number has a decimal representation

Every decimal representation $a_0.a_1a_2a_3, \ldots$ corresponds to a nonnegative real number. To formulate this statement more precisely, for all $k \in \mathbb{N}$ we first define the rational numbers r_k and s_k as

$$r_k = a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_k}{10^k} \in \mathbb{Q}$$
$$s_k = a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_k + 1}{10^k} \in \mathbb{Q};$$

we also define the sets A and B as

$$A = \{r_k\,;\, k \in \mathbb{N}\} \qquad \text{and} \qquad B = \{s_k\,;\, k \in \mathbb{N}\}.$$

Because $r_k < s_k$ for all $k \in \mathbb{N}$; because $s_k - r_k = \frac{1}{10^k} = 10^{-k}$ for all $k \in \mathbb{N}$; and because $\inf\left(\{10^{-k}\,;\, k \in \mathbb{N}\}\right) = 0$, the sets A and B are bounded and obey $\sup(A) = \inf(B)$. As a result, the decimal representation $a_0.a_1a_2a_3\ldots$ defines the real number $a \in \mathbb{R}$ given by

$$a = \sup(A) = \inf(B) = a_0.a_1a_2a_3\ldots \in \mathbb{R}.$$

Every nonnegative real number can be written in the above decimal form. To show this, let $a \in \mathbb{R}$ be an arbitrary nonnegative number. There then exists some $a_0 \in \mathbb{N} \cup \{0\}$ for which

$$a_0 \leq a < a_0 + 1;$$

the number a_0 is the largest integer that is not larger than a. More so, there exists some $a_1 \in \{0, 1, \ldots, 9\}$ for which

$$a_0 + \frac{a_1}{10} \leq a < a_0 + \frac{a_1 + 1}{10};$$

this number a_1 is the first decimal in a's decimal notation. The remaining decimals $a_2, a_3, a_4, \ldots$ are found analogously: if the numbers $a_1, a_2, \ldots, a_{k-1} \in \{0, 1, \ldots, 9\}$, then $a_k \in \{0, 1, \ldots, 9\}$ is the uniquely defined number for which

$$a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_k}{10^k} \leq a < a_0 + \frac{a_1}{10} + \frac{a_2 + 1}{10^2} + \cdots + \frac{a_k + 1}{10^k}.$$

The procedure terminates if the sequence of decimal digits reaches a value of k for which

$$a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_k}{10^k} = a;$$

otherwise the procedure continues indefinitely to yield an infinite sequence of decimal digits $a_1, a_2, a_3, \ldots$; in either case, the sequence of decimal digits together with the integer a_0 define the decimal representation of the real number a as

$$a = a_0.a_1a_2a_3\ldots.$$

The decimal representation resulting from the above-described algorithm is called the *standard decimal representation* of the nonnegative number $a \in \mathbb{R}^+ \cup \{0\}$; in particular, the algorithm is such that the number 9 will never repeat indefinitely from some point onward in the sequence $a_1, a_2, a_3, \ldots$.

Every nonnegative real number has a unique standard decimal representation, and every standard decimal representation corresponds to a unique nonnegative real number.

Topic 2.12: Extensions of the decimal representation

(i) (*Decimal representation of negative numbers*) An arbitrary negative real number $b \in \mathbb{R}^-$ is represented in the decimal form

$$b = -b_0.b_1b_2b_3\ldots,$$

where $b_0.b_1b_2b_3$ is the decimal representation of the nonnegative number $(-b)$.

(ii) (*Decimal representation of integer numbers*) Integer numbers have a decimal representation in which a_0 is equal to the integer itself and all decimal digits $a_1, a_2, a_3, \ldots$ are exactly equal to zero.

(iii) (*Decimal representation of rational numbers*) The decimal representation of rational numbers is either finite, or infinitely repeating and periodic from some point onward. An example of a rational number with a finite decimal representation is $1/2 = 0.5$; two examples of a rational number with an infinite and periodically repeating decimal representation are

$$1/3 = 0.333\ldots \quad \text{and} \quad 1/11 = 0.090909\ldots.$$

The periodic repetition is abbreviated by drawing a line over the complete repeating portion, for example

$$1/3 = 0.333\ldots = 0.\overline{3} \quad \text{and} \quad 1/11 = 0.090909\ldots = 0.\overline{09}.$$

Remark 2.23: Bounded rational sets need not have a rational supremum

The sets A and B defined in the decimal representation are subsets of the rational numbers. If the number $a \in \mathbb{R}$ specified by A and B using the decimal construction is not a rational number, then a is defined as the supremum of the set $A \subset \mathbb{Q}$ or, equivalently, the infimum of the set $B \subset \mathbb{Q}$. The set A is thus an example of a

bounded set of rational numbers without a supremum in the rational numbers. As a result, Dedekind's axiom (A13) does not hold for the rational numbers.

2.2.5 Existence of the square root

Consider an arbitrary real number $a \in \mathbb{R}$ and the equation $x^2 = a$ for the unknown real number x.

(a) If $a < 0$, the equation $x^2 = a$ does not have any real solutions.

(b) If $a = 0$, the equation $x^2 = a$ has the single solution $x = 0$.

(c) If $a > 0$ and x_1 is a solution of $x^2 = a$, then $-x_1$ is also a solution of $x^2 = a$, and x_1 and $-x_1$ are the only two solutions.

Proposition 2.24: Square root of a nonnegative number

For every nonnegative real number $a \in \mathbb{R}^+ \cup \{0\}$, the equation $x^2 = a$ for the unknown variable x has exactly one *nonnegative* real solution, called the *square root* of a and denoted by $x = \sqrt{a}$. (We stress that if $a > 0$ the equation also has a negative solution $x_- = -\sqrt{a}$; this proposition concerns only the nonnegative solution.)

Proof. Assume a is positive, i.e. $a > 0$, and define the subset

$$S = \{x \in \mathbb{R}^+ ;\, x^2 < a\}.$$

We first show that in this case S is nonempty; we do so by rearranging the inequality $a^2 < (1 + a)^2 a$ to produce

$$a^2/(1 + a)^2 < a.$$

The number $a/(1 + a) \in \mathbb{R}^+$ then satisfies the conditions for membership of S (it is positive assuming $a > 0$ and also obeys $[a/(1 + a)]^2 < a$), so S is nonempty.

Because S is nonempty and bounded above (the number a, for example, is an upper bound of S), it must have supremum by the Dedekind axiom (A13). Let $b = \sup(S) \in \mathbb{R}$ denote S's supremum; our plan is to show that $b^2 = a$ and thus that b solves the equation $x = \sqrt{a}$ from this proposition's introduction. We will do this by showing that neither $b^2 > a$ nor $b^2 < a$ are correct, leaving $b^2 = a$ as the only remaining possibility.

(i) Assume $b^2 > a$ and consider the number

$$c = b - \frac{b^2 - a}{2b} = \frac{1}{2}\left(b + \frac{a}{b}\right).$$

By construction $0 < c < b$ (assuming $b^2 > a$); in addition

$$c^2 = b^2 - (b^2 - a) + \frac{(b^2 - a)^2}{4b^2} = a + \frac{(b^2 - a)^2}{4b^2} > a,$$

and so $c^2 > a > x^2$, and thus $c > x$, for each $x \in S$. Thus c is an upper bound of S, which, together with $c < b$, contradicts the assumption that b is S's least upper bound. The assumption $b^2 > a$ is thus incorrect.

(ii) Next, assume $b^2 < a$ and consider a number d for which

$$0 < d < b \quad \text{and} \quad d < \frac{a - b^2}{3b};$$

from this choice of d it follows that

$$(b+d)^2 = b^2 + d(2b+d) < b^2 + d(2b+b) < b^2 + (a - b^2) = a.$$

The resulting inequality $(b+d)^2 < a$ implies that $b + d \in S$. Since $d > 0$, the quantity $b + d \in \mathrm{S}$ is larger than b, which contradicts the assumption that b is an upper bound of S. Thus the assumption $b^2 < a$ is also incorrect.

(iii) Because both $b^2 > a$ and $b^2 < a$ are incorrect by points (i) and (ii), respectively, the only remaining possibility is that $b^2 = a$, completing the proof. □

The proof of Proposition 2.24 relied crucially on the fact that the set $S \subset \mathbb{R}$ has a real-valued supremum because of Dedekind's axiom. In the following proposition, we will show that quadratic equations of the form $x^2 = a$, where $a \in \mathbb{R}^+$, do not always have rational solutions (even though such equations do always have a *real* solution). This lack of rational solutions can be reverse engineered to imply that Dedekind's axiom does not hold for the rational numbers; it also means that, loosely, there are "more" real numbers than rational numbers.

Proposition 2.25: The square root of 2 is irrational

The equation $x^2 = 2$ for the number x does not have a rational solution.

Proof. We will prove the proposition by contradiction. Assume the equation $x^2 = 2$ did have a rational solution, which we write in the form

$$(m/n)^2 = 2,$$

where m/n is a fully reduced fraction of two natural numbers $m, n \in \mathbb{N}$. We rearrange to get $m^2 = 2n^2$, implying that m is an even number. Since m is even, we may write $m = 2k$ for some natural number k, producing

$$m^2 = 4k^2 = 2n^2 \implies n^2 = 2k^2.$$

The last equality implies that n is also an even number, which contradicts the assumption that the fraction m/n is reduced to lowest terms (since both m and n are divisible by 2) and proves, by contradiction, that $x^2 = 2$ does not have rational solutions for the number x. □

Topic 2.13: Terminology: Irrational, algebraic, and transcendental numbers

We are now equipped to state some important classes of numbers:

- *Irrational* numbers are real numbers that are not rational; $\sqrt{2}$ is an example of an irrational number.
- *Algebraic* numbers are numbers that are the solution of a polynomial equation with integer coefficients; the square roots of natural numbers are an example

of algebraic numbers.

- *Transcendental* numbers are those real numbers that are neither rational nor algebraic; π and e (Euler's number) are two well-known examples of transcendental numbers.

  Although only a few transcendental numbers are commonly used in practice, the vast majority of real numbers are transcendental. In fact, the set of all transcendental numbers is equipotent to the set of all real numbers and thus uncountably infinite, while the set of all algebraic numbers, like the set the set of all rational numbers, can be shown to be only countably infinite.

2.2.6 Intervals, neighborhoods, and absolute value

The sections in this chapter thus far have mostly concerned algebraic properties of the real numbers. However, the ordering concepts mentioned in the context of Remark 2.11 also make it possible to equip the real numbers with *topological* concepts like "closeness" and "farness". Namely, it is possible to formulate in a precise manner which real numbers are *closer to* and which are *farther from* any given real number. The topological structure of the real numbers is the foundation for the concepts of continuity and differentiability used in real analysis.

Definition 2.26: Intervals on the real line

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$. In this case, *bounded intervals* are subsets of $\mathbb{R}$ of the form:

- $(a, b) = \{x \in \mathbb{R};\, a < x < b\}$, called the *open interval* from a to b,
- $[a, b] = \{x \in \mathbb{R};\, a \leq x \leq b\}$, called the *closed interval* from a to b,
- $(a, b] = \{x \in \mathbb{R};\, a < x \leq b\}$, and
- $[a, b) = \{x \in \mathbb{R};\, a \leq x < b\}$.

In each of the four cases, a is the interval's infimum and b is the interval's supremum.

Meanwhile, *unbounded intervals* are subsets of $\mathbb{R}$ of the form:

- $(-\infty, b] = \{x \in \mathbb{R};\, x \leq b\}$
- $(a, \infty) = \{x \in \mathbb{R};\, a < x\}$
- $[a, \infty) = \{x \in \mathbb{R};\, a \leq x\}$
- $(-\infty, \infty) = \mathbb{R}$.

Definition 2.27: Absolute value of a real number

The *absolute value* of a real number $x \in \mathbb{R}$, denoted by $|x|$, is defined as

$$|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases} \in \mathbb{R}^+ \cup \{0\}.$$

The absolute value is thus a function of the form

$$|\cdot| : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}, \quad x \mapsto |x|.$$

Note that $|x| \geq 0$ for all $x \in \mathbb{R}$ and $|x| = 0$ if, and only if, $x = 0$.

The absolute value is used to straightforwardly define the distance between real numbers.

Definition 2.28: Distance between real numbers

The *distance* between any two real numbers $x, y \in \mathbb{R}$, denoted by $d(x, y)$, is the nonnegative real number

$$d(x, y) = |x - y| \in \mathbb{R}^+ \cup \{0\}.$$

Having defined intervals and absolute value, we are now equipped to introduce interior points, exterior points, and neighborhoods, which will feature centrally throughout the remainder of this book.

Definition 2.29: Interior, exterior, and boundary points on the real line

Let $A \subset \mathbb{R}$ be an arbitrary real set and let $a \in \mathbb{R}$ be an arbitrary real number.

$$(a - \epsilon, a + \epsilon) = .$$

(i) The point a is called an *interior point* of A if there exists some positive real number $\epsilon \in \mathbb{R}^+$ for which[a] $(a - \epsilon, a + \epsilon) \subset A$.

(ii) The point a is called an *exterior point* of A if a is an interior point of A's complement $A^{\mathsf{c}} = \mathbb{R} \setminus A$.

(iii) The point a is called a *boundary point* of $A \subset \mathbb{R}$ if it is neither an interior point nor an exterior point of A.

(iv) The set A is called a *neighborhood* of the point a if a is an interior point of A.

[a] The interval $(a-\epsilon, a+\epsilon)$ can be interpreted geometrically as the set of all real numbers separated from a by a distance less than ϵ; it is defined as

$$(a - \epsilon, a + \epsilon) = \{x \in \mathbb{R};\, a - \epsilon < x < a + \epsilon\} = \{x \in \mathbb{R};\, |x - a| < \epsilon\}.$$

We may now generalize the concept of open and closed intervals to arbitrary subsets of $\mathbb{R}$.

Definition 2.30: Open and closed real subsets

Let $A \subset \mathbb{R}$ be an arbitrary real subset.

(i) The set A is said to be *open* in $\mathbb{R}$ if A is a neighborhood of every point in A; equivalently, the set A is open if all of its points are interior points.

(ii) The set A is said to be *closed* in $\mathbb{R}$ if its complement $A^{\mathsf{c}} = \mathbb{R} \setminus A$ is open in $\mathbb{R}$.

Note that the definition of openness and closedness allows, perhaps counterintuitively on a first encounter, for subsets of $\mathbb{R}$ to be either (1) neither open nor closed or (2) *both* open and closed; we offer some examples below.

Example 2.31: Open and closed intervals

(i) The sets $\emptyset$ and $\mathbb{R}$ are both open and closed in $\mathbb{R}$. It should be straightforward to verify that $\mathbb{R}$ is open, which immediately implies that $\mathbb{R}^{\mathsf{c}} = \emptyset$ is closed. The justification that $\emptyset$ is open is less satisfactory; it is an example of what mathematicians and logicians call a *vacuous truth*. Loosely, because there are no points in the empty set in the first place, every one of the (nonexistent) points in $\emptyset$ can be interpreted as an interior point of $\emptyset$, and so the empty set is open. It then immediately follows that $\mathbb{R}$ is closed, since $\mathbb{R}^{\mathsf{c}} = \emptyset$ is open.

(ii) For all $c, d \in \mathbb{R}$ for which $c \leq d$, intervals of the form (c, d), $(-\infty, d)$, and (c, ∞) are open in $\mathbb{R}$.

(iii) For all $c, d \in \mathbb{R}$ for which $c \leq d$, intervals of the form $[c, d]$, $(-\infty, d]$, and $[c, \infty)$ are closed in $\mathbb{R}$ (since the complements of these sets, i.e. $(-\infty, c) \cup (d, \infty)$; (d, ∞); and $(-\infty, c)$ are all open in $\mathbb{R}$).

(iv) Intervals of the form $[c, d)$ and $(c, d]$ are neither open nor closed in $\mathbb{R}$; this is because the interval $[c, d)$ is not a neighborhood of $c \in [c, d)$, while the interval $(c, d]$ is not a neighborhood of $d \in (c, d]$. To see this, note that for all $\epsilon \in \mathbb{R}^+$, however small ϵ might be, the interval $(c - \epsilon, c + \epsilon)$ contains points not in $[c, d)$ and is thus not a subset of $[c, d)$. Similarly, the interval $(d - \epsilon, d + \epsilon)$ contains points not in $(c, d]$, and is thus not a subset of $(c, d]$.

Proposition 2.32: Important equalities and inequalities on the real line

The following statements hold for all $x, y \in \mathbb{R}$:

(i) $|xy| = |x||y|$,

(ii) $|x + y| \leq |x| + |y|$ (this inequality is called the *triangle inequality*), and

(iii) $||x| - |y|| \leq |x + y|$.

Proof.

(i) $|xy| = |x||y|$ follows directly from applying the definition of the absolute value to each of the possible cases $x, y \geq 0$; $x, y < 0$; $x \geq 0$, $y < 0$; and $x < 0$, $y \geq 0$.

(ii) We separately consider three possible cases:

(a) If $x = 0$ or $y = 0$, or if x and y are of the same sign, then evidently $|x + y| = |x| + |y|$ by the rules of real arithmetic and the definition of the absolute value.

(b) If $x < 0$ and $y > 0$, we consider two subcases:

i. If $x + y \geq 0$, then by the definition of the absolute value the inequality $|x + y| \leq |x| + |y|$ simplifies to $x + y \leq -x + y$, which holds under the assumption $x < 0$.

ii. If $x + y < 0$, then the inequality $|x + y| \leq |x| + |y|$ simplifies to $-x - y \leq -x + y$, which holds under the assumption $y > 0$.

(c) The case $x > 0$ and $y < 0$ is equivalent to the case $x < 0$ and $y > 0$, since the labels x and y are arbitrary.

(iii) We first apply the triangle inequality to get

$$|y| = |x + y - x| \leq |x + y| + |-x| = |x + y| + |x|,$$

then rearrange to get

$$-|x + y| \leq |x| - |y|.$$

Similarly, make the calculation

$$|x| = |x + y - y| \leq |x + y| + |-y| = |x + y| + |y|,$$

and rearrange to get

$$|x| - |y| \leq |x + y|.$$

The combination of $|x| - |y| \leq |x + y|$ and $-|x + y| \leq |x| - |y|$ implies

$$||x| - |y|| \leq |x + y|.$$ □

2.3 The complex numbers

The *complex* numbers, denoted by $\mathbb{C}$, are the set

$$\mathbb{C} = \mathbb{R} \times \mathbb{R} = \{(a, b)\,;\, a, b \in \mathbb{R}\}$$

equipped with the following addition and multiplication operations:

- (*Complex addition*) The sum of any two complex numbers $(a, b), (a', b') \in \mathbb{C}$ is

$$(a, b) + (a', b') = (a + a', b + b') \in \mathbb{C}.$$

- (*Complex multiplication*) The product of any two complex numbers $(a, b), (a', b') \in \mathbb{C}$ is

$$(a, b) \cdot (a', b') = (a, b)(a', b') = (aa' - bb', ab' + ba') \in \mathbb{C}.$$

Topic 2.14: Axioms for the complex numbers

The above-defined addition and multiplication operations satisfy a set of addition and multiplication axioms completely analogous in structure to Axioms (A1)–(A10), i.e. the addition and multiplication axioms for the real numbers. Because of this similarity, we will not state the analogous complex number axioms here, and instead only summarize the important modifications in generalizing Axioms (A1)–(A10) to the complex numbers.

- The complex additive identity is $(0, 0) \in \mathbb{C}$.
- The additive inverse to the complex number (a, b) is the complex number $-(a, b) = (-a, -b) \in \mathbb{C}$.
- The complex multiplicative identity is $(1, 0) \in \mathbb{C}$.

- The multiplicative inverse to any complex number $(a, b) \in \mathbb{C} \setminus \{(0, 0)\}$ is

$$(a, b)^{-1} = \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2} \right).$$

  The multiplicative inverse to $(0, 0) \in \mathbb{C}$ is undefined (just like the real multiplicative inverse to $0 \in \mathbb{R}$ is undefined).

Because they satisfy Axioms (A1)–(A10), the complex numbers form a commutative ring. All properties and operations derived for the real numbers as a consequence of Axioms (A1)–(A10) (including Props. 2.3, 2.5, 2.6, and 2.8) have natural extensions to the complex numbers (and in general to any commutative ring), and we will not restate them here.

However, the complex numbers do not satisfy Axioms (A11) or (A12) and thus do not form an ordered ring—there is no natural ordering relation for the complex numbers.

Topic 2.15: Real numbers under the complex numbers

In the scope of the complex numbers, the real numbers are represented as the subset

$$\{(a, 0)\,;\, a \in \mathbb{R}\} = \mathbb{R} \times \{0\} \subset \mathbb{C};$$

the real number $a \in \mathbb{R}$ corresponds to the complex number $(a, 0) \in \mathbb{C}$.

Complex addition and multiplication restricted to the set $\mathbb{R} \times \{0\} \subset \mathbb{C}$ agrees with real addition and multiplication, since

$$(a, 0) + (a', 0) = (a + a', 0) \quad \text{and} \quad (a, 0) \cdot (a', 0) = (aa', 0) \text{ for all } a, a' \in \mathbb{R}.$$

It is thus conventional to associate $\mathbb{R}$ with the set $\mathbb{R} \times \{0\}$, and write $\mathbb{R} = \mathbb{R} \times \{0\} \subset \mathbb{C}$ and $a = (a, 0)$ for every $a \in \mathbb{R}$.

Definition 2.33: Components of complex numbers

For shorthand, we conventionally denote an arbitrary complex number $(a, b) \in \mathbb{C}$ with the single letter $z = (a, b) \in \mathbb{C}$; any $z = (a, b) \in \mathbb{C}$ thus has two components:

- The first component, a, is called z's *real component* and is denoted by

$$\mathrm{Re}(z) = a \qquad \text{(real component of } z = (a, b) \in \mathbb{C}).$$

- The second component, b, is called z's *imaginary component* and is denoted by

$$\mathrm{Im}(z) = b \qquad \text{(imaginary component of } z = (a, b) \in \mathbb{C}).$$

Definition 2.34: The imaginary unit

The *imaginary unit*, denoted by i, is the complex number

$$\mathrm{i} = (0, 1) \in \mathbb{C}.$$

In terms of i, an arbitrary complex number $z = (a, b)$ may be uniquely written as

$$z = (a, b) = (a, 0) + (0, 1) \cdot (b, 0) = a + \mathrm{i}b.$$

By definition of complex multiplication, the imaginary unit i satisfies

$$\mathrm{i}^2 = (-1, 0) = -1 \implies \mathrm{i}^2 + 1 = 0.$$

More so, the imaginary unit obeys $\mathrm{i}^2 = 1$, $\mathrm{i}^3 = -\mathrm{i}$, $\mathrm{i}^4 = 1$, and so on.

In terms of i, the product of two complex numbers is

$$\begin{aligned}(a + \mathrm{i}b)(a' + \mathrm{i}b') &= aa' + \mathrm{i}ba \ + a\mathrm{i}b' + \mathrm{i}^2 bb' \\ &= (aa' - bb') + \mathrm{i}(ba + ab').\end{aligned}$$

Definition 2.35: The complex conjugate

The *complex conjugate* of a complex number $z = a + \mathrm{i}b$ with real part $a \in \mathbb{R}$ and imaginary part $b \in \mathbb{R}$ is the complex number

$$\overline{z} = a - \mathrm{i}b.$$

For all $z, w \in \mathbb{C}$ and all $x \in \mathbb{R}$, the complex conjugate obeys the following properties:

- (*Distributivity over addition*), i.e. $\overline{z + w} = \overline{z} + \overline{w}$,
- (*Commutativity over multiplication*), i.e. $\overline{zw} = \overline{wz}$,
- $\overline{\mathrm{i}} = -\mathrm{i}$, and
- $\overline{x} = x$ for all $x \in \mathbb{R}$.

For an arbitrary complex number $z = a + \mathrm{i}b$ with real part $a \in \mathbb{R}$ and imaginary part $b \in \mathbb{R}$ it holds that

$$z\overline{z} = a^2 + b^2 \geq 0.$$

The fact that $z|z|$ is always nonnegative motivates defining the absolute value of an arbitrary complex number z as in the following definition.

Definition 2.36: Absolute value of a complex number

The *absolute value* of the complex number $z = (a, b) \in \mathbb{C}$, denoted by $|z|$, is defined as

$$|z| = \sqrt{z\overline{z}} = \sqrt{a^2 + b^2}.$$

The complex absolute value, much like the real absolute value, obeys $|z| = 0$ if, and only if, $z = 0$. More so, for real numbers $z = a + \mathrm{i} \cdot 0 = a \in \mathbb{R}$ the complex absolute value obeys

$$|z| = \sqrt{a^2} \geq 0,$$

which agrees with the real absolute value defined in Definition 2.27.

Complex conjugation and absolute value are used to define complex division.

Definition 2.37: Complex division

The quotient of any two complex numbers $z, w \in \mathbb{C}$ for which $w \neq 0$ is defined as

$$\frac{z}{w} = \frac{z\bar{w}}{w\bar{w}} = \frac{z\bar{w}}{|w|^2} = \frac{\mathrm{Re}(z\bar{w})}{|w|^2} + \mathrm{i}\frac{\mathrm{Im}(z\bar{w})}{|w|^2}.$$

The above definition makes elegant use of the complex conjugate and absolute value: multiplying the fraction z/w above and below by $\bar{w}/\bar{w}$ transforms the original denominator from the complex number $w \in \mathbb{C}$ to the positive real number $|w|^2 \in \mathbb{R}^+$, thus formulating complex division as a straightforward extension of real division (or, equivalently, a simple scaling of a complex number by a positive real scalar), but operations we have already defined in this book.

Proposition 2.38: Properties of the complex absolute value

The following properties holds any two complex numbers $z, w \in \mathbb{C}$:

(i) $|zw| = |z||w|$,

(ii) $|z + w| \leq |z| + |w|$ (this inequality is called the *triangle inequality*),

(iii) $||z| - |w|| \leq |z + w|$, and

(iv) $||z| - |w|| \leq |z - w|$.

Proof.

(i) We begin with the calculation

$$|zw|^2 = zw\overline{zw} = zw\bar{z}\,\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2 = (|z||w|)^2,$$

then take the square root of both sides of the resulting equality $|zw|^2 = (|z||w|)^2$ to get $|zw| = |z||w|$.

(ii) We first define $z = x + \mathrm{i}y$ and $w = a + \mathrm{i}b$, where $x, y, a, b \in \mathbb{R}$, and then compute

$$\begin{aligned} |z + w| &= |(x + a) + \mathrm{i}(y + b)| = \sqrt{(x + a)^2 + (y + b)^2} \\ &= \sqrt{x^2 + 2xa + a^2 + y^2 + 2yb + b^2} \\ &= \sqrt{x^2 + 2xa + a^2 + y^2 + 2yb + b^2} \\ &= \sqrt{|z|^2 + |w|^2 + 2(ax + by)}. \end{aligned} \tag{2.2}$$

We also note that

$$|xa + yb| = \sqrt{(xa + yb)^2} = \sqrt{(x^2 + y^2)(a^2 + b^2) - (xb - ya)^2},$$

which leads to the inequality

$$|xa + yb| \leq \sqrt{(x^2 + y^2)(a^2 + b^2)} = |z||w|$$

and in turn

$$-|z||w| \leq xa + yb \leq |z||w|.$$

We then substitute $-|z||w| \leq xa + yb \leq |z||w|$ into Equation 2.2 to get

$$\sqrt{|z|^2 + |w|^2 - 2|z||w|} \leq |z + w| \leq \sqrt{|z|^2 + |w|^2 + 2|z||w|}.$$

We can then prove (ii) with the calculation

$$\begin{aligned} |z + w| &\leq \sqrt{|z|^2 + |w|^2 + 2|z||w|} = \sqrt{(|z| + |w|)^2} \\ &= ||z| + |w|| \overset{\text{(a)}}{=} |z| + |w|, \end{aligned}$$

where (a) holds because both $|z|$ and $|w|$ are necessarily nonnegative.

(iii) We prove (iii) with the calculation

$$\begin{aligned} |z + w| &\geq \sqrt{|z|^2 + |w|^2 - 2|z||w|} = \sqrt{(|z| - |w|)^2} \\ &= ||z| - |w||. \end{aligned}$$

(iv) This point follows directly from the previous point and the calculation

$$||z| - |w|| = ||z| - |-w|| \overset{\text{(a)}}{\leq} |z + (-w)| = |z - w|,$$

where (a) holds by point (iii). □

Note the similarities of points (i)–(iii) to their real-valued analogs in Proposition 2.32; this similarity comes from the fact that the real and complex absolute value, addition, and multiplication operations obey analogous properties.

We conclude this introductory section with a routine exercise in computing imaginary components, complex conjugates, and absolute values.

Example 2.39: Practice with complex numbers

Let us compute the real and imaginary components, complex conjugate, and absolute value of the complex number

$$z = \frac{a + \mathrm{i}}{a - \mathrm{i}}, \text{ where } a \in \mathbb{R}.$$

We first multiply z above and below by its denominator's complex conjugate and rearrange to get

$$z = \frac{a + \mathrm{i}}{a - \mathrm{i}} \cdot \frac{a + \mathrm{i}}{a + \mathrm{i}} = \frac{a^2 + 2a\mathrm{i} - 1}{a^2 + 1} = \frac{a^2 - 1}{a^2 + 1} + \frac{2a}{a^2 + 1}\mathrm{i}.$$

The number z's components are thus

$$\mathrm{Re}(z) = \frac{a^2 - 1}{a^2 + 1} \quad \text{and} \quad \mathrm{Im}(z) = \frac{2a}{a^2 + 1}$$

Meanwhile, by interchanging the sign of all i terms, z's complex conjugate is

$$\bar{z} = \frac{a - \mathrm{i}}{a + \mathrm{i}},$$

which in turn allows to straightforwardly compute z's absolute value:

$$z\bar{z} = \left(\frac{a+\mathrm{i}}{a-\mathrm{i}}\right)\left(\frac{a-\mathrm{i}}{a+\mathrm{i}}\right) = 1 \implies |z| = \sqrt{z\bar{z}} = 1.$$

2.3.1 Geometric representation of complex numbers

A complex number $z \in \mathbb{C}$ can be represented geometrically as a point or vector in the complex plane, i.e. as an arrow directed from the origin to the point z. In this vector representation:

- The arrow's length is the absolute value $|z|$.
- The angle ϕ between the arrow and the real axis, measured in the positive (counterclockwise) direction, is called z's *argument* $\arg(z) = \phi \in [0, 2\pi)$.

In terms of the trigonometric functions sine and cosine, any complex number can be written in the form

$$z = |z|(\cos\phi + \mathrm{i}\sin\phi).$$

Topic 2.16: Polar form of a complex number

The *polar form* of a complex number $z \in \mathbb{C}$ is

$$z = |z|\mathrm{e}^{\mathrm{i}\phi} \overset{\text{(a)}}{=} |z|(\cos\phi + \mathrm{i}\sin\phi);$$

equality (a) relies on *Euler's identity* $\mathrm{e}^{\mathrm{i}\phi} = \cos\phi + \mathrm{i}\sin\phi$, which we will prove in Chapter 10, after covering power series. Euler's identity holds for all $\phi \in \mathbb{R}$.

Topic 2.17: Geometric view of complex addition and multiplication

The sum of two complex numbers $z, w \in \mathbb{C}$ can be interpreted as the vector sum of z and w taken as vectors in complex plane.

To geometrically interpret the product of two complex numbers $z, w \in \mathbb{C}$, we first write z and w in the polar forms $z = |z|\mathrm{e}^{\mathrm{i}\phi}$ and $w = |w|\mathrm{e}^{\mathrm{i}\phi}$. We then apply Euler's identity and trigonometric identities for sine and cosine to produce

$$\begin{aligned} zw &= |z|\mathrm{e}^{\mathrm{i}\phi}|w|\mathrm{e}^{\mathrm{i}\psi} \\ &= |z||w|(\cos\phi + \mathrm{i}\sin\phi)(\cos\psi + \mathrm{i}\sin\psi) \\ &= |z||w|(\cos\phi\cos\psi - \sin\phi\sin\psi + \mathrm{i}\cos\phi\sin\psi + \mathrm{i}\sin\phi\cos\psi) \\ &= |z||w|\big[\cos(\phi+\psi) + \mathrm{i}\sin(\phi+\psi)\big] \\ &= |zw|\mathrm{e}^{\mathrm{i}(\phi+\psi)}. \end{aligned}$$

Thus one finds the vector representing the product zw, by multiplying the vector representing z by the scalar $|w|$ and rotating it by the angle ψ in the positive direction. In particular, this means that multiplying a complex number z by the imaginary unit i corresponds rotating the vector representing z by an angle of $\pi/2$ counterclockwise.

Topic 2.18: De Moivre's formula

Repeatedly applying the result $zw = |zw|e^{i(\phi+\psi)}$ produces

$$z^2 = zz = |z|e^{i\phi}|z|e^{i\phi} = |z|^2 e^{i2\phi}$$
$$z^3 = z^2 z = |z|^2 e^{i2\phi}|z|e^{i\phi} = |z|^3 e^{i3\phi}, \text{ and so on;}$$

by the principle of induction it then follows that

$$z^{k+1} = z^k z = |z|^k e^{ik\phi}|z|e^{i\phi} = |z|^{k+1} e^{i(k+1)\phi} \text{ for all } k \in \mathbb{N}.$$

Taking $k + 1 = n$ produces *de Moivre's formula*:

$$z^n = (|z|e^{i\phi})^n = |z|^n e^{in\phi};$$

substituting in Euler's identity $e^{i\phi} = \cos\phi + i\sin\phi$ gives the equivalent form

$$z^n = \big[|z|(\cos\phi + i\sin\phi)\big]^n = |z|^n\big[\cos(n\phi) + i\sin(n\phi)\big],$$

which holds for all $z = |z|e^{i\phi}$ with $\phi \in \mathbb{R}$ and all $n \in \mathbb{N}$.

De Moivre's formula is used to easily compute the n-th power of a complex number.

2.3.2 Roots of complex numbers

Consider a nonzero complex number $w \in \mathbb{C} \setminus \{0\}$ and natural number $n \in \mathbb{N}$. Our goal in this section is to solve equations of the form $z^n = w$ for the unknown complex number $z \in \mathbb{C}$.

We begin by first writing z and w in their polar forms

$$z = |z|e^{i\phi} \quad \text{and} \quad w = |w|e^{i\psi},$$

and then applying de Moivre's formula to get

$$|z|^n e^{in\phi} = |w|e^{i\psi} \iff z^n = w.$$

Two complex numbers are equal if they have equal absolute values and equal argument, and so the equation $|z|^n e^{in\phi} = |w|e^{i\psi}$ is equivalent to the two simultaneous conditions

$$|z|^n = |w| \quad \text{and} \quad n\phi = \psi + 2\pi k, \quad k \in \mathbb{Z}.$$

We then solve for $|z|$ and ϕ to get

$$|z| = \sqrt[n]{w} \quad \text{and} \quad \phi = \frac{\psi}{n} + \frac{2\pi k}{n}, \quad k \in \mathbb{Z}.$$

Unique solutions to $z^n = w$ occur for $k \in \{0, 1, \ldots, n-1\}$ and may be written in the form

$$z_k = \sqrt[n]{|w|}e^{i\left(\frac{\psi}{n} + \frac{2\pi k}{n}\right)}.$$

Interpreted geometrically, the solutions z_k to the equation $z^n = w$ lie on a circle centered at the origin with radius $\sqrt[n]{|w|}$. These solutions form the vertices of a regular n-gon (a polygon with n vertices) in the complex plane. The n solutions to $z^n = w$ are called *the n-th roots* of the complex number $w \in \mathbb{C} \setminus \{0\}$.

Topic 2.19: The complex roots of unity

The complex roots of the number 1 (a special case of the equation $z^n = w$ when $w = 1$) are called the *complex roots of unity*. The n-th complex roots of unity are

$$\left\{ e^{i\frac{2\pi k}{n}} ; k = 0, 1, \ldots, n-1 \right\} \subset \mathbb{C} \qquad (n\text{-th roots of unity}).$$

In passing, we mention that the n-th roots of unity form an Abelian group under multiplication.

Example 2.40: Roots of complex numbers

(1) Let us solve the complex equation $z^2 = 2i$.

We first write $z = x + iy$, where $x, y \in \mathbb{R}$, and multiply out to get

$$z^2 = x^2 - y^2 + 2ixy = 2i$$

We then equate real and imaginary parts on each side of the equation to get

$$x^2 - y^2 = 0 \quad \text{and} \quad 2xy = 2.$$

The equation $2xy = 2$ means that x and y are both nonzero and that $y = 1/x$; we then substitute $y = 1/x$ into $x^2 - y^2 = 0$ to get

$$x^4 = 1 \implies x = \pm 1.$$

We then combine $x = \pm 1$ with $y = 1/x$ to get $y = \pm 1$ and thus

$$z_{1,2} = \pm(1 + i) \qquad (\text{solutions to } z^2 = 2i).$$

(2) Let us solve the equation $z^2 = u + iv$, where $u, v \in \mathbb{R}$.

If $u = v = 0$ then $z = 0$. Otherwise, we again write $z = x + iy$ with $x, y \in \mathbb{R}$, multiply out z^2, and equate real and imaginary parts to get the system of equations

$$x^2 - y^2 = u \quad \text{and} \quad 2xy = v.$$

We then separately consider the following cases:

(a) If $u > 0$ and $v = 0$, then $y = 0$ and $x = \pm\sqrt{u} \neq 0$. The solutions of $z^2 = u + iv$ are then $z_{1,2} = \pm\sqrt{u}$.

(b) If $u < 0$ and $v = 0$, then $x = 0$ and $y = \pm\sqrt{-u} \neq 0$. The solutions of $z^2 = u + iv$ are then $z_{1,2} = \pm i\sqrt{-u}$.

(c) If $u \neq 0$ and $v \neq 0$, the equation $2xy = v$ implies that $x \neq 0$, $y \neq 0$, and that $y = v/(2x)$. We then substitute $y = v/(2x)$ into $x^2 - y^2 = u$ and rearrange to get

$$4x^4 - 4x^2u - v^2 = 0.$$

We then introduce the new variable $s = x^2 > 0$ to get the quadratic equation $4s^2 - 4us - v^2 = 0$, which (e.g. using the quadratic formula) has the solutions

$$s_{1,2} = \frac{u \pm \sqrt{u^2 + v^2}}{2}.$$

Because $\sqrt{u^2+v^2} > u$ and $s > 0$, only the positive solution is valid, so

$$s = \frac{u + \sqrt{u^2+v^2}}{2}.$$

We then apply $x = \pm\sqrt{s}$ and $y = v/(2x)$ to get

$$z_{1,2} = \pm\left(\frac{\sqrt{u+\sqrt{u^2+v^2}}}{\sqrt{2}} + \mathrm{i}\frac{v}{\sqrt{2}\sqrt{u+\sqrt{u^2+v^2}}}\right).$$

2.3.3 Complex disks and neighborhoods

This section generalizes some of the concepts in Section 2.2.6 to the complex plane; we explicitly reference the definition being generalized whenever appropriate.

Definition 2.41: Open and closed disks in the complex plane

Let $z \in \mathbb{C}$ be an arbitrary complex number and let R be a nonnegative real number. In this case:

- An *open disk* centered at z with radius R is the subset

$$\mathcal{K}(z,R) = \{w \in \mathbb{C};\ |z-w| < R\} \subset \mathbb{C}.$$

- A *closed disk* centered at z with radius R is the subset

$$\overline{\mathcal{K}}(z,R) = \{w \in \mathbb{C};\ |z-w| \leq R\} \subset \mathbb{C}.$$

Definition 2.41 is a generalization of Definition 2.26 to the complex plane. Geometrically, the open disk $\mathcal{K}(z,R)$ is the set of all points in the complex plane separated from z by a distance less that R; this set is a disk of radius R centered at z.

Definition 2.42: Boundedness of a subset of $\mathbb{C}$

A complex subset $A \subset \mathbb{C}$ is said to be *bounded* if there exists a positive real number $M \in \mathbb{R}^+$ for which

$$A \subset \overline{\mathcal{K}}(0,M),$$

where $\overline{\mathcal{K}}(0,M)$ is a closed disk of radius M centered at the origin. In other words, the set A is bounded if there exists a closed disk containing A.

The above definition means means that any point in a bounded set is no farther from the origin than M, and that any two points in a bounded set are no further from each other than $2M$.

Every subset of the real numbers is also a subset of the complex numbers, and is bounded as a subset of the real numbers if, and only if, it is bounded as a subset of the complex numbers.

Definition 2.43: Interior points and neighborhoods in $\mathbb{C}$

Let $A \subset \mathbb{C}$ be an arbitrary complex subset and let $z \in \mathbb{C}$ be an arbitrary complex number. In this case

(i) The point z is an *interior point* of A if there exists some $\epsilon \in \mathbb{R}^+$ for which $\mathcal{K}(z, \epsilon) \subset A$. In other words, z is an interior point of A if there exists an open disk centered at z fully inside A.

(ii) The set A is a *neighborhood* of the point $z \in \mathbb{C}$ if z is an interior point of A.

Definition 2.43 is a generalization of Definition 2.29 to the complex plane.

Definition 2.44: Open and closed subsets of $\mathbb{C}$

Let $A \subset \mathbb{C}$ be an arbitrary complex subset and let $z \in \mathbb{C}$ be an arbitrary complex number. In this case:

(i) The set A is said to be *open* in $\mathbb{C}$ if A is a neighborhood of all points in A.

(ii) The subset $A \subset \mathbb{C}$ is said to be *closed* in $\mathbb{C}$ if its complement $A^{\mathsf{c}} = \mathbb{C} \setminus A$ is open in $\mathbb{C}$.

Definition 2.44 is a generalization of Definition 2.30 to the complex plane. Like in Example 2.31:

- Both the sets $\emptyset$ and $\mathbb{C}$ are simultaneously open and closed in $\mathbb{C}$.
- Open disks are open subsets of $\mathbb{C}$.
- Closed disks are closed subsets of $\mathbb{C}$.

3 Three-dimensional Euclidean space

This chapter covers the basic operations between elements—called *vectors*—of the three-dimensional vector space $\mathbb{R}^3$, and also covers the description of lines and planes in $\mathbb{R}^3$. You will cover vector spaces in a more general and abstract fashion in a course on linear algebra.

Loosely, the vector space $\mathbb{R}^3$ is a mathematical description of the three-dimensional space of every human life, for which we already have a well-developed geometric intuition from everyday life. If one specifies a coordinate system and origin for $\mathbb{R}^3$, it is possible to represent every point in $\mathbb{R}^3$ with three real-valued coordinates as an ordered triplet taken from the set

$$\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R} = \big\{(x, y, z)\,;\, x, y, z \in \mathbb{R}\big\}.$$

The element $(x, y, z) \in \mathbb{R}^3$ can be interpreted as a point A in the space $\mathbb{R}^3$. We denote this point by $A(x, y, z)$ and call the real numbers x, y, and z the *Cartesian coordinates* of the point A. The *origin* in the space $\mathbb{R}^3$ is the point $O(0, 0, 0)$.

Topic 3.1: Vector representation of points in $\mathbb{R}^3$

The element (x, y, z) of $\mathbb{R}^3$ can be interpreted as a directed line segment pointing from the origin $O(0, 0, 0)$ to the point $A(x, y, z)$. This directed line segment is called the *position vector* of the point A and is denoted by

$$\boldsymbol{r}_A = \overrightarrow{OA} \qquad \text{(position vector of the point } A\text{)}.$$

Since the origin $O(0, 0, 0)$ is constant, any position vector $\boldsymbol{r}_A$ is fully specified by the final point $A(x, y, z)$, and so a point $A(x, y, z)$ is often identified with its position vector $\boldsymbol{r}_A$ using the notation

$$\boldsymbol{r}_A = \overrightarrow{OA} = (x, y, z) = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Because of the one-to-one correspondence between points in $\mathbb{R}^3$ and their position vectors, we will regularly refer to elements of the space $\mathbb{R}^3$ as *vectors*. In this context we will also refer to real numbers as *scalars*. We will commonly denote arbitrary vectors in $\mathbb{R}^3$ by

$$\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3 \qquad \text{(a generic vector in } \mathbb{R}^3\text{)},$$

and call the numbers $v_1, v_2, v_3 \in \mathbb{R}$ the *components* of the vector $\boldsymbol{v}$.

3.1 The vector space $\mathbb{R}^3$

We will begin by defining the vector addition and scalar multiplication of elements of the set $\mathbb{R}^3$, and then explain how the set $\mathbb{R}^3$ equipped with these two operations forms what mathematicians call a *vector space*.

Definition 3.1: Vector addition and scalar multiplication

We define the following two operations on the set $\mathbb{R}^3$.

(i) (*Vector addition*) The *vector sum* of any two vectors $\boldsymbol{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ and $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$, denoted by $\boldsymbol{u} + \boldsymbol{v}$, is defined as

$$\boldsymbol{u} + \boldsymbol{v} = (u_1 + v_1, u_2 + v_2, u_3 + v_3) \in \mathbb{R}^3.$$

(ii) (*Multiplication of a vector by a scalar*) For any vector $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ and any scalar $\alpha \in \mathbb{R}$, the *product* of α and $\boldsymbol{v}$, denoted by $\alpha\boldsymbol{v}$ or $\alpha \cdot \boldsymbol{v}$, is defined as

$$\alpha\boldsymbol{v} = \alpha \cdot \boldsymbol{v} = (\alpha v_1, \alpha v_2, \alpha v_3) \in \mathbb{R}^3.$$

Any vector space must also contain (1) a zero vector and (2) an additive inverse to every element in the vector space. In the vector space $\mathbb{R}^3$, these are defined as follows.

Definition 3.2: The zero vector and additive inverse

The *zero vector* in $\mathbb{R}^3$ is denoted by $\mathbf{0}$ and defined as

$$\mathbf{0} = (0, 0, 0) \in \mathbb{R}^3.$$

The *additive inverse* to the vector $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ is denoted by $-\boldsymbol{v}$ and defined as

$$-\boldsymbol{v} = (-1) \cdot \boldsymbol{v} = (-v_1, -v_2, -v_3) \in \mathbb{R}^3.$$

The zero vector has the important property that $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{R}^3$, while the additive inverse has the property that $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}$, for all $\boldsymbol{v} \in \mathbb{R}^3$. More generally, the following properties hold for the vector space $\mathbb{R}^3$.

Topic 3.2: Axiomatic properties of the vector space $\mathbb{R}^3$

The following properties hold for all vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$ and all scalars $\alpha, \beta \in \mathbb{R}$.

(i) $(\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} = \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w})$ (associativity of vector addition)

(ii) $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$ (commutativity of vector addition)

(iii) $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$ (property of the zero vector)

(iv) $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}$ (property of the additive inverse)

(v) Scalar multiplication is distributive over vector addition, i.e.

$$\alpha(\boldsymbol{u} + \boldsymbol{v}) = \alpha\boldsymbol{u} + \alpha\boldsymbol{v}.$$

(vi) Scalar multiplication is distributive over scalar addition, i.e.

$$(\alpha + \beta)\boldsymbol{v} = \alpha\boldsymbol{v} + \beta\boldsymbol{v}.$$

(vii) Multiplication of scalars and scalar multiplication of vectors is compatible, i.e.

$$\alpha(\beta\boldsymbol{v}) = (\alpha\beta)\boldsymbol{v}.$$

(viii) $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ (property of the multiplicative identity)

Because the operations of vector addition and scalar multiplication obey the axiomatic properties (i) through (viii), the set $\mathbb{R}^3$, equipped with the operations of vector addition and scalar multiplication, is called a *vector space over the scalar field* $\mathbb{R}$.

Note that we have not derived (and thus cannot and will not prove) properties (i)–(viii) from first principles. Instead, we take these properties as axioms characterizing the vector space $\mathbb{R}^3$, similarly to how we used Axioms (A1) through (A13) as a way of characterizing the real numbers. We will then use properties (i)–(viii) to derive all other properties of $\mathbb{R}^3$.

We now introduce the Cartesian unit vectors, which are a set of three vectors with the special property that any vector in $\mathbb{R}^3$ can be written as a simple sum of the unit vectors.

Definition 3.3: The Cartesian unit vectors

The *Cartesian unit vectors*, denoted by $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$, are defined as

$$\hat{\mathbf{i}} = (1,0,0) \in \mathbb{R}^3$$
$$\hat{\mathbf{j}} = (0,1,0) \in \mathbb{R}^3$$
$$\hat{\mathbf{k}} = (0,0,1) \in \mathbb{R}^3.$$

An arbitrary vector $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ can be uniquely written as a sum of scalars multiplied by the Cartesian unit vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ in the form

$$\boldsymbol{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}.$$

This representation of $\boldsymbol{v}$ is called the *expansion of* $\boldsymbol{v}$ *in terms of* $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, *and* $\hat{\mathbf{k}}$.

The Cartesian unit vectors are a special case of what mathematicians call a *basis* for the vector space $\mathbb{R}^3$; a basis for a (finite-dimensional) vector space has the important property that any vector in the space can be written as a linear combination of the basis vectors. The vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ point in the directions of the Cartesian unit axes and are called the *standard basis vectors* of the vector space $\mathbb{R}^3$. You will learn more about bases in a course on linear algebra.

Example 3.4: Translation of vectors to the origin

Let A and B be two points in $\mathbb{R}^3$ with associated position vectors

$$\boldsymbol{r}_A = (x, y, z) \quad \text{and} \quad \boldsymbol{r}_B = (x', y', z'),$$

and consider the directed line segment that points from A to point B.

If we translate (i.e. move) this line segment so that it starts at the origin O instead of at the point A, in such a way that preserves the line's direction in space, the result of the translation is the vector in $\mathbb{R}^3$ given by

$$\overrightarrow{AB} = \boldsymbol{r}_B - \boldsymbol{r}_A = (x' - x, y' - y, z' - z).$$

We will assume a convention in which a directed line segment from point A to point B is equivalent to the vector $\overrightarrow{AB} = \boldsymbol{r}_B - \boldsymbol{r}_A \in \mathbb{R}^3$. Every directed line segment between two points in $\mathbb{R}^3$ thus corresponds to a vector, and two such directed line segments correspond to the same vector if, and only if, one is a direction-preserving translation of the other.

3.2 The scalar, vector, and triple products

3.2.1 The scalar product in $\mathbb{R}^3$

Definition 3.5: The scalar product in $\mathbb{R}^3$

The *scalar product* of any two vectors $\boldsymbol{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ and $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ is a real number denoted by $\boldsymbol{u} \cdot \boldsymbol{v}$ and defined as

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + u_2 u_2 + u_3 v_3 \in \mathbb{R}.$$

The scalar product is sometimes also called the *dot product*; in more general vector spaces, an operation analogous to the scalar product is called the *inner product*.

Proposition 3.6: Properties of the scalar product

The following properties hold for all vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$ and all scalars $\alpha \in \mathbb{R}^3$

(i) $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$ (commutativity)

(ii) $\boldsymbol{u} \cdot (\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{u} \cdot \boldsymbol{w}$ (distributivity over vector addition)

(iii) $(\boldsymbol{u} + \boldsymbol{v}) \cdot \boldsymbol{w} = \boldsymbol{u} \cdot \boldsymbol{w} + \boldsymbol{v} \cdot \boldsymbol{w}$ (distributivity over vector addition)

(iv) $\alpha(\boldsymbol{u} \cdot \boldsymbol{v}) = (\alpha \boldsymbol{u}) \cdot \boldsymbol{v} = \boldsymbol{u} \cdot (\alpha \boldsymbol{v})$ (homogeneity)

(v) $\boldsymbol{v} \cdot \boldsymbol{v} \geq 0$ (positive definiteness)

(vi) $\boldsymbol{v} \cdot \boldsymbol{v} = 0$ if, and only if, $\boldsymbol{v} = \boldsymbol{0}$. (positive definiteness)

Proof. All properties can be directly verified by direct computation using the definition of the scalar product. □

Because it obeys properties (ii)–(iv), the scalar product is a *bilinear operation*.

The scalar product can be used to equip $\mathbb{R}^3$ with a number of useful geometric notions, including the length of a vector and the distances and angles between vectors. We explore length and distance immediately below.

Topic 3.3: Length of a vector in $\mathbb{R}^3$

The length of any vector $\boldsymbol{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ is denoted by $\|\boldsymbol{v}\|$ and defined as

$$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}} = \sqrt{v_1^2 + v_2^2 + v_3^3} \in \mathbb{R}.$$

The positive-definiteness of the scalar product (Proposition 3.6 (vi)), means that the length of any vector $\boldsymbol{v} \in \mathbb{R}^3$ obeys

$$\boldsymbol{v} \geq 0 \quad \text{and} \quad \|\boldsymbol{v}\| = 0 \iff \boldsymbol{v} = \boldsymbol{0}.$$

Additionally, for all $\alpha \in \mathbb{R}$ the length of any scaled vector obeys

$$\|\alpha \boldsymbol{v}\| = |\alpha| \|\boldsymbol{v}\| \text{ for all } \boldsymbol{v} \in \mathbb{R}^3.$$

Note: in $\mathbb{R}^3$ (and more generally in any Euclidean space) the length of a vector is commonly denoted by either $\|\boldsymbol{v}\|$ or $\|\boldsymbol{v}\|$; we will use $\|\boldsymbol{v}\|$ to avoid ambiguity with

the absolute value of a scalar. In more general vector spaces, an operation analogous to taking the length of a vector is called the *vector norm*, and is also denoted by $\|\boldsymbol{v}\|$.

Topic 3.4: Distance between vectors in $\mathbb{R}^3$

The distance between any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ is denoted by $d(\boldsymbol{u}, \boldsymbol{v})$ and defined as

$$d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{v} - \boldsymbol{u}\| = \sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2 + (v_3 - u_3)^2}.$$

Geometrically, viewing vectors as directed line segments in $\mathbb{R}^3$ as discussed in Topic 3.1, the distance between two vectors is the distance between their endpoints. In this view, for any two points A and B in $\mathbb{R}^3$ with associated position vectors $\boldsymbol{r}_A = (x, y, z)$ and $\boldsymbol{r}_B = (x', y', z')$, the distance between the points is

$$d(A, B) = \|\boldsymbol{r}_B - \boldsymbol{r}_A\| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}.$$

Proposition 3.7: The Cauchy-Schwartz inequality

Every two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ obey the Cauchy-Schwartz inequality

$$|\boldsymbol{u} \cdot \boldsymbol{v}| \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\|.$$

Proof. If either $\boldsymbol{u} = \boldsymbol{0}$, $\boldsymbol{v} = \boldsymbol{0}$, or $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{0}$, we can directly verify the inequality:

$$|\boldsymbol{v} \cdot \boldsymbol{u}| = 0 \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\| = 0.$$

Now assume both $\boldsymbol{u}$ and $\boldsymbol{v}$ are nonzero, and for shorthand define the three terms

$$\alpha = \|\boldsymbol{v}\|^2, \quad \beta = \boldsymbol{u} \cdot \boldsymbol{v}, \quad \boldsymbol{w} = \alpha\boldsymbol{u} - \beta\boldsymbol{v}.$$

We then make the calculation

$$\begin{aligned}
\boldsymbol{w} \cdot \boldsymbol{w} &\overset{(a)}{=} (\alpha\boldsymbol{u} - \beta\boldsymbol{v}) \cdot (\alpha\boldsymbol{u} - \beta\boldsymbol{v}) \\
&= \alpha^2 \boldsymbol{u} \cdot \boldsymbol{u} - \alpha\beta\boldsymbol{u} \cdot \boldsymbol{v} - \beta\alpha\boldsymbol{v} \cdot \boldsymbol{u} + \beta^2 \boldsymbol{v} \cdot \boldsymbol{v} \\
&\overset{(b)}{=} \alpha^2 \boldsymbol{u} \cdot \boldsymbol{u} - 2\alpha\beta\boldsymbol{u} \cdot \boldsymbol{v} + \beta^2 \boldsymbol{v} \cdot \boldsymbol{v} \\
&= \alpha^2 \|\boldsymbol{u}\|^2 - 2\alpha\beta\boldsymbol{u} \cdot \boldsymbol{v} + \beta^2 \|\boldsymbol{v}\|^2 \\
&\overset{(c)}{=} \|\boldsymbol{v}\|^4 \|\boldsymbol{u}\|^2 - 2\|\boldsymbol{v}\|^2 (\boldsymbol{u} \cdot \boldsymbol{v})^2 + (\boldsymbol{u} \cdot \boldsymbol{v})^2 \|\boldsymbol{v}\|^2 \\
&= \|\boldsymbol{v}\|^4 \|\boldsymbol{u}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2 \|\boldsymbol{v}\|^2 \overset{(d)}{\geq} 0,
\end{aligned}$$

where (a) follows from the definition of $\boldsymbol{w}$, (b) holds because of the dot product's commutativity, (c) follows from the definitions of α and β, and (d) holds because $\boldsymbol{w} \cdot \boldsymbol{w} \geq 0$ by Proposition 3.6 (v).

We then divide through by $\|\boldsymbol{v}\|^2$ (which is allowed because $\|\boldsymbol{v}\|^2 > 0$ under the assumption that $\boldsymbol{v} \neq \boldsymbol{0}$), rearrange, and take the square root of the inequality to get

$$\begin{aligned}
&\|\boldsymbol{v}\|^2 \|\boldsymbol{u}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2 \geq 0 && \text{(after dividing through by } \|\boldsymbol{v}\|^2\text{)} \\
&\|\boldsymbol{v}\|^2 \|\boldsymbol{u}\|^2 \geq (\boldsymbol{u} \cdot \boldsymbol{v})^2 && \text{(rearranging)} \\
&\sqrt{\|\boldsymbol{v}\|^2 \|\boldsymbol{u}\|^2} = \|\boldsymbol{v}\|\|\boldsymbol{u}\| \geq \sqrt{(\boldsymbol{u} \cdot \boldsymbol{v})^2} = |\boldsymbol{u} \cdot \boldsymbol{v}| && \text{(taking the square root).}
\end{aligned}$$

The result is the Cauchy-Schwartz inequality $|\boldsymbol{u} \cdot \boldsymbol{v}| \leq \|\boldsymbol{v}\|\|\boldsymbol{u}\|$. □

Strict equality, i.e. $|\boldsymbol{u} \cdot \boldsymbol{v}| = \|\boldsymbol{v}\| \cdot \|\boldsymbol{u}\|$, occurs when $\boldsymbol{w} = \boldsymbol{0}$. From the definition of $\boldsymbol{w}$, this occurs when $\alpha\boldsymbol{u} = \beta\boldsymbol{v}$, i.e. (assuming $\boldsymbol{u}$ and $\boldsymbol{v}$ are nonzero) when $\boldsymbol{u}$ and $\boldsymbol{v}$ point in the same direction in space (or, equivalently, when one vector is a scalar multiple of the other).

Proposition 3.8: The triangle inequality

Every two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ obey the triangle inequality

$$\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|.$$

Proof. We begin with the calculation

$$\|\boldsymbol{u} + \boldsymbol{v}\|^2 = (\boldsymbol{u} + \boldsymbol{v}) \cdot (\boldsymbol{u} + \boldsymbol{v}) = \boldsymbol{u} \cdot \boldsymbol{u} + 2\boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \boldsymbol{v},$$

then apply the Cauchy-Schwartz inequality to get

$$\begin{aligned} \|\boldsymbol{u} + \boldsymbol{v}\|^2 &= \boldsymbol{u} \cdot \boldsymbol{u} + 2\boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \boldsymbol{v} \\ &\overset{(a)}{\leq} \|\boldsymbol{u}\|\|\boldsymbol{u}\| + 2\|\boldsymbol{u}\|\|\boldsymbol{v}\| + \|\boldsymbol{v}\|\|\boldsymbol{v}\| \\ &= (\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)^2, \end{aligned}$$

where (a) uses the Cauchy-Schwartz inequality. We then take the square root of both sides of the resulting inequality to get $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$. □

The triangle equality gets its name from interpreting the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{u}+\boldsymbol{v}$ as directed line segments defining the sides of triangle, together with the fact that, from elementary geometry, the sum of the lengths of any two sides of a triangle is greater than or equal to the length of the remaining side.

Definition 3.9: Orthogonality

Two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ are said to be *orthogonal* when

$$\boldsymbol{v} \cdot \boldsymbol{u} = 0.$$

If $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal, we write $\boldsymbol{u} \perp \boldsymbol{v}$.

Orthogonality is generalization of the concept of perpendicularity to general vector spaces, and the definition of orthogonality agrees with the notion of perpendicularity from elementary geometry. To show this, we first make the auxiliary calculation

$$\|\boldsymbol{u} + \boldsymbol{v}\|^2 = (\boldsymbol{u} + \boldsymbol{v}) \cdot (\boldsymbol{u} + \boldsymbol{v}) = \|\boldsymbol{u}\|^2 + 2\boldsymbol{u} \cdot \boldsymbol{v} + \|\boldsymbol{v}\|^2. \tag{3.1}$$

We then note that any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ specify a triangle with sides $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{u} + \boldsymbol{v}$. The sides $\boldsymbol{u}$ and $\boldsymbol{v}$ are perpendicular when the triangle obeys the Pythagorean theorem $\|\boldsymbol{u} + \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$. But by Equation 3.1, this perpendicularity condition occurs if, and only if, $\boldsymbol{u} \cdot \boldsymbol{v} = 0$, i.e. when $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal.

Definition 3.10: Orthogonal projection

The *orthogonal projection* of the vector $\boldsymbol{u} \in \mathbb{R}^3$ onto the nonzero vector $\boldsymbol{v} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}$, denoted by $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$, is defined as the vector

$$\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u}) = \left(\frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{v} \cdot \boldsymbol{v}}\right) \boldsymbol{v} \in \mathbb{R}^3.$$

This definition, although perhaps intimidating at first sight, has a well-defined geometrical interpretation: the orthogonal projection $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$ of $\boldsymbol{u}$ onto $\boldsymbol{v}$ is the scaled multiple of $\boldsymbol{v}$ for which $\boldsymbol{u} - \mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$ is perpendicular to $\boldsymbol{v}$, i.e.

$$\left[\boldsymbol{u} - \mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})\right] \cdot \boldsymbol{v} = 0.$$

The formula for the orthogonal projection follows from denoting $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u}) = \alpha \boldsymbol{v}$ for some $\alpha \in \mathbb{R}$, and then applying the above orthogonality condition to compute

$$\left[\boldsymbol{u} - \mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})\right] \cdot \boldsymbol{v} = (\boldsymbol{u} - \alpha \boldsymbol{v}) \cdot \boldsymbol{v} = 0 \implies \alpha = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{v} \cdot \boldsymbol{v}}.$$

Having solved for α, it then follows that

$$\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u}) = \alpha \boldsymbol{v} = \left(\frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\boldsymbol{v} \cdot \boldsymbol{v}}\right) \boldsymbol{v}.$$

Note that $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u}) = \boldsymbol{0}$ if $\boldsymbol{u} = \boldsymbol{0}$.

Topic 3.5: Angle between vectors

The *angle* ϕ between any two nonzero vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R} \setminus \{\boldsymbol{0}\}$ is given by

$$\cos \phi = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}.$$

One can derive this formula directly from geometrical considerations: if one defines a triangle with side lengths $\boldsymbol{u}$, $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$, and $\boldsymbol{u} - \mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$ then by the definition of the cosine function the angle between the vectors $\boldsymbol{u}$ and $\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$, and thus also the angle between $\boldsymbol{u}$ and $\boldsymbol{v} \parallel \mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})$, is

$$\cos \phi = \frac{\left\|\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u})\right\|}{\|\boldsymbol{u}\|} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{v}\|^2} \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{u}\|} = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}.$$

The angle between two vectors is undefined if one (or both) of the vectors is $\boldsymbol{0}$.

The angle ϕ between any two nonzero vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R} \setminus \{\boldsymbol{0}\}$ is

- acute if $\boldsymbol{u} \cdot \boldsymbol{v} > 0$,
- obtuse if $\boldsymbol{u} \cdot \boldsymbol{v} < 0$, and
- a right angle if $\boldsymbol{u} \cdot \boldsymbol{v} = 0$.

The definition of the angle between two vectors can be used to define the notions of parallel and linearly dependent vectors—these concepts play an important role in the theory of vector spaces, but we will use them only briefly in this book.

Definition 3.11: Parallel and linearly dependent vectors

Two nonzero vectors in $\mathbb{R}^3$ are said to be *parallel* or *colinear* if the angle between then is equal to either 0 or π. The zero vector $\mathbf{0}$ is thus parallel to all vectors in $\mathbb{R}^3$.

Two vectors in $\mathbb{R}^3$ are said to be *linearly dependent* if one is a scalar multiple of the other. More formally (but equivalently) the vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ are linearly dependent if, and only if, there exist scalars $\alpha, \beta \in \mathbb{R}^3$ of which at least one is nonzero and for which

$$\alpha\boldsymbol{u} + \beta\boldsymbol{v} = \mathbf{0}.$$

Topic 3.6: Unit vectors and direction cosines

A vector $\boldsymbol{u} \in \mathbb{R}^3$ is called a *unit* or *normalized* vector if

$$\|\boldsymbol{u}\| = 1 \qquad \text{(condition that } \boldsymbol{u} \text{ is a unit vector)}.$$

A unit vector is typically written with a circumflex, for example $\hat{\mathbf{u}}$. An arbitrary vector $\boldsymbol{v} \in \mathbb{R}^3$ can be *normalized*, i.e. made into a unit vector, by dividing it by its magnitude:

$$\hat{\mathbf{v}} = \frac{1}{\|\boldsymbol{v}\|}\boldsymbol{v} \qquad \text{(normalized version of } \boldsymbol{v}).$$

The vector $\hat{\mathbf{v}}$ points in the same direction as $\boldsymbol{v}$ but has unit magnitude.

Any normalized vector $\hat{\mathbf{u}}$ can be written in the form

$$\hat{\mathbf{u}} = (\cos\alpha, \cos\beta, \cos\gamma) = \cos\alpha\,\hat{\mathbf{i}} + \cos\beta\,\hat{\mathbf{j}} + \cos\gamma\,\hat{\mathbf{k}},$$

where α, β, and γ are the angles between $\hat{\mathbf{u}}$ and the standard basis vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$, respectively. The components $\cos\alpha$, $\cos\beta$, and $\cos\gamma$ are called the *direction cosines* of the unit vector $\hat{\mathbf{u}}$.

We conclude this section with a routine exercise in computing vector lengths, dot products, angles, and orthogonal projections.

Example 3.12: Practice with vector computations

Define the vectors $\boldsymbol{u} = (1, 2, 3)$ and $\boldsymbol{v} = (1, 0, -1)$, and let us compute the lengths $\|\boldsymbol{u}\|$, $\|\boldsymbol{v}\|$, and $\|\boldsymbol{u} + \boldsymbol{v}\|$; the angle ϕ between the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and the orthogonal projection of $\boldsymbol{u}$ onto $\boldsymbol{v}$.

We first note that $\boldsymbol{u} + \boldsymbol{v} = (2, 2, 2)$, and then compute the lengths of each vector:

$$\begin{aligned}\|\boldsymbol{u}\| &= \sqrt{1+4+9} = \sqrt{14}\\ \|\boldsymbol{v}\| &= \sqrt{1+0+1} = \sqrt{2}\\ \|\boldsymbol{u}+\boldsymbol{v}\| &= \sqrt{4+4+4} = \sqrt{12}.\end{aligned}$$

We then compute the dot product

$$\boldsymbol{u} \cdot \boldsymbol{v} = 1 + 0 - 3 = -2,$$

which we then use to compute the angle ϕ:

$$\cos\phi = \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} = \frac{-2}{\sqrt{28}} = -\frac{1}{\sqrt{7}}.$$

Finally, the orthogonal projection of $\boldsymbol{u}$ onto $\boldsymbol{v}$ is

$$\mathrm{pr}_{\boldsymbol{v}}^{\perp}(\boldsymbol{u}) = \left(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\boldsymbol{v}\cdot\boldsymbol{v}}\right)\cdot\boldsymbol{v} = \left(\frac{-2}{2}\right)\boldsymbol{v} = (-1,0,1).$$

3.2.2 The vector product in $\mathbb{R}^3$

Definition 3.13: The vector product

The *vector product* of any two vectors $\boldsymbol{u} = (u_1,u_2,u_3)$ and $\boldsymbol{v} = (v_1,v_2,v_3)$ is a vector denoted by $\boldsymbol{u}\times\boldsymbol{v}$ and defined as

$$\boldsymbol{u}\times\boldsymbol{v} = (u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1) \in \mathbb{R}^3.$$

The vector product is also called the *cross product*, and can be written in the following equivalent forms:

$$\begin{aligned}\boldsymbol{u}\times\boldsymbol{v} &= (u_2v_3 - u_3v_2)\hat{\mathbf{i}} + (u_3v_1 - u_1v_3)\hat{\mathbf{j}} + (u_1v_2 - u_2v_1)\hat{\mathbf{k}} \\ &\overset{\text{(a)}}{=} \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}\hat{\mathbf{i}} + \begin{vmatrix} u_3 & u_1 \\ v_3 & v_1 \end{vmatrix}\hat{\mathbf{j}} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}\hat{\mathbf{k}} \\ &\overset{\text{(b)}}{=} \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},\end{aligned} \tag{3.2}$$

where (a) and (b) use a notation and concept from linear algebra called the *determinant* of a square matrix. We mention this determinant form of the vector product only for the sake of completeness—aside from a brief appearance in the calculation of the triple product in Section 3.2.3, will not use determinants elsewhere in this book.

Proposition 3.14: Properties of the vector product I

The following properties hold for all vectors $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^3$ and all scalars $\alpha\in\mathbb{R}$:

(i) $\boldsymbol{u}\times\boldsymbol{v} = -(\boldsymbol{v}\times\boldsymbol{u})$ (anticommutativity)

(ii) $\boldsymbol{u}\times(\boldsymbol{v}+\boldsymbol{w}) = \boldsymbol{u}\times\boldsymbol{v} + \boldsymbol{u}\times\boldsymbol{w}$ (distributivity over vector addition)

(iii) $(\boldsymbol{u}+\boldsymbol{v})\times\boldsymbol{w} = \boldsymbol{u}\times\boldsymbol{w} + \boldsymbol{v}\times\boldsymbol{w}$ (distributivity over vector addition)

(iv) $\alpha(\boldsymbol{u}\times\boldsymbol{v}) = (\alpha\boldsymbol{u})\times\boldsymbol{v} = \boldsymbol{u}\times(\alpha\boldsymbol{v})$ (homogeneity)

(v) The vector $\boldsymbol{u}\times\boldsymbol{v}$ is orthogonal to both $\boldsymbol{u}$ and $\boldsymbol{v}$, i.e.

$$\boldsymbol{u}\cdot(\boldsymbol{u}\times\boldsymbol{v}) = \boldsymbol{v}\cdot(\boldsymbol{u}\times\boldsymbol{v}) = 0$$

(vi) $\boldsymbol{v}\times\boldsymbol{v} = \boldsymbol{0}$.

Proof. Points (i)–(v) can be directly verified by direct computation using the definition of the vector product. Point (vi) is a direct consequence of the anticommutativity in point (i). □

Because it satisfies properties (ii)–(iv), the vector product is a bilinear operation.

The unit vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ obey the cross product identities

$$\begin{aligned} \hat{\mathbf{i}} \times \hat{\mathbf{j}} &= \hat{\mathbf{k}} & \hat{\mathbf{i}} \times \hat{\mathbf{i}} &= \mathbf{0} \\ \hat{\mathbf{j}} \times \hat{\mathbf{k}} &= \hat{\mathbf{i}} & \hat{\mathbf{j}} \times \hat{\mathbf{j}} &= \mathbf{0} \\ \hat{\mathbf{k}} \times \hat{\mathbf{i}} &= \hat{\mathbf{j}} & \hat{\mathbf{k}} \times \hat{\mathbf{k}} &= \mathbf{0}. \end{aligned}$$

In fact, because the vector product is a bilinear operation and every vector in $\mathbb{R}^3$ can be written as a linear combination of the basis vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$, the vector product is fully determined by properties (ii)–(iv) and the vector products of the unit vectors. Namely:

$$\begin{aligned} \boldsymbol{u} \times \boldsymbol{v} &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}) \times (v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}) \\ &= u_1v_2(\hat{\mathbf{i}} \times \hat{\mathbf{j}}) + u_1v_3(\hat{\mathbf{i}} \times \hat{\mathbf{k}}) + u_2v_1(\hat{\mathbf{j}} \times \hat{\mathbf{i}}) \\ &\quad + u_2v_3(\hat{\mathbf{j}} \times \hat{\mathbf{k}}) + u_3v_1(\hat{\mathbf{k}} \times \hat{\mathbf{i}}) + u_3v_2(\hat{\mathbf{k}} \times \hat{\mathbf{j}}) \\ &= u_1v_2\hat{\mathbf{k}} + u_1v_3(-\hat{\mathbf{j}}) + u_2v_1(-\hat{\mathbf{k}}) + u_2v_3\hat{\mathbf{i}} + u_3v_1\hat{\mathbf{j}} + u_3v_2(-\hat{\mathbf{i}}) \\ &= (u_2v_3 - u_2v_2)\hat{\mathbf{i}} - (u_1v_3 - u_3v_1)\hat{\mathbf{j}} + (u_1v_2 - u_2v_1)\hat{\mathbf{k}}, \end{aligned}$$

which agrees with the definition of the vector product in Definition 3.13.

Proposition 3.15: Properties of the vector product II

The following properties hold for all vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$:

(i) $\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}) = (\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w}$

(ii) $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) = (\boldsymbol{u} \cdot \boldsymbol{w})\boldsymbol{v} - (\boldsymbol{u} \cdot \boldsymbol{v})\boldsymbol{w}$

(iii) $\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) + \boldsymbol{v} \times (\boldsymbol{w} \times \boldsymbol{u}) + \boldsymbol{w} \times (\boldsymbol{u} \times \boldsymbol{v}) = \mathbf{0}$, which is called the *Jacobi identity* for the cross product.

(iv) If $\boldsymbol{u}$ and $\boldsymbol{v}$ are both nonzero and ϕ is the angle between $\boldsymbol{v}$ and $\boldsymbol{u}$, then

$$\|\boldsymbol{u} \times \boldsymbol{v}\| = \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin\phi.$$

Proof. Points (i) through (iii) can be verified by direct computation using the definition of the vector product. Before proving point (iv), we first quote the equality

$$\|\boldsymbol{u} \times \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2,$$

which can be verified by direct computation using the definition of the scalar and vector product. Point (iv) then follows from the calculation

$$\begin{aligned} \|\boldsymbol{u} \times \boldsymbol{v}\|^2 &= \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - (\boldsymbol{u} \cdot \boldsymbol{v})^2 \\ &\overset{\text{(a)}}{=} \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 - \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 \cos^2\phi \\ &= \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2(1 - \cos^2\phi) \\ &= \|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2 \sin^2\phi. \end{aligned}$$

where (a) follows from the definition of the angle between two vectors. We then take the square root of both sides to get $\|\boldsymbol{u} \times \boldsymbol{v}\| = \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin \phi$. □

Topic 3.7: Geometric interpretation of the vector product

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two nonzero vectors in $\mathbb{R}^3$ and let ϕ denote the angle between them. We first note that the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ define a parallelogram with a base of length $\|\boldsymbol{u}\|$ and a height $\|\boldsymbol{v}\| \sin \phi$. The area of this parallelogram is

$$A \overset{\text{(a)}}{=} \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin \phi \overset{\text{(b)}}{=} \|\boldsymbol{u} \times \boldsymbol{v}\|,$$

where (a) follows from elementary geometry and (b) from Proposition 3.15 (iv). Note that this area formula makes sense even if one of the vectors is zero, in which case the area of the parallelogram is also zero.

The cross product $\boldsymbol{u} \times \boldsymbol{v}$ of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$ is thus a third vector orthogonal to both $\boldsymbol{u}$ and $\boldsymbol{v}$ with length equal to the area of the parallelogram spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$.

Assuming $\boldsymbol{u} \times \boldsymbol{v}$ is nonzero, the direction of $\boldsymbol{u} \times \boldsymbol{v}$ is such that the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{u} \times \boldsymbol{v}$, in that order, satisfy the right-hand rule. In particular, this means that the mutual orientation of $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{u} \times \boldsymbol{v}$, in that order, is the same as the mutual orientation of the standard basis vectors $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$, in that order.

The vector product of two vectors is closely related to the vectors' linear dependence (or lack of it), as explored in the following proposition.

Proposition 3.16: Linear dependence in $\mathbb{R}^3$

The following properties are equivalent for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$.

(i) The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel.

(ii) The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent.

(iii) The vectors satisfy the Cauchy-Schwartz equality $|\boldsymbol{u} \cdot \boldsymbol{v}| = \|\boldsymbol{u}\|\|\boldsymbol{v}\|$

(iv) The vectors' cross product is zero, i.e. $\boldsymbol{u} \times \boldsymbol{v} = \mathbf{0}$.

Proof. If either $\boldsymbol{u}$, $\boldsymbol{v}$, or both are zero, then all properties follow directly from the definitions of parallel vectors, linear independence, etc.

Now assume $\boldsymbol{u}$ and $\boldsymbol{v}$ are both nonzero, and let ϕ denote the angle between them. In this case:

- The equivalence of (ii) and (iii) follows from the definition of linear dependence in Definition 3.11 and the fact that two vectors satisfy the Cauchy-Schwartz equality only if one is a scalar multiple of the other (recall the comment at the end of the proof of Proposition 3.7).
- Point (iv) follows from point (ii) and the calculation

$$\boldsymbol{u} \times \boldsymbol{v} \overset{\text{(a)}}{=} (\gamma \boldsymbol{v}) \times \boldsymbol{v} \overset{\text{(b)}}{=} \gamma(\boldsymbol{v} \times \boldsymbol{v}) \overset{\text{(c)}}{=} \mathbf{0},$$

where (a) uses point (ii) and the definition of linear dependence to write $\boldsymbol{u} = \gamma \boldsymbol{v}$, (b) uses the homogeneity of the vector product, and (c) follows from the vector product identity $\boldsymbol{v} \times \boldsymbol{v} = \boldsymbol{0}$.

- Point (i) follows from (iv) and the calculation

$$\boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{0} \implies 0 = \|\boldsymbol{u} \times \boldsymbol{v}\| \overset{\text{(a)}}{=} \|\boldsymbol{u}\|\|\boldsymbol{v}\| \sin\phi \overset{\text{(b)}}{\implies} \sin\phi = 0,$$

where (a) uses Proposition 3.15 (iv) and (b) follows from the assumption that both $\boldsymbol{u}$ and $\boldsymbol{v}$ are nonzero, requiring $\sin\phi = 0$. The result $\sin\phi = 0$ means that ϕ equals either 0 or π, and so $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel by Definition 3.11.

- Point (iii) follows from point (i); namely, if $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel, i.e. if ϕ equals 0 or π, then $\cos\phi = \pm 1$ and so

$$|\boldsymbol{u} \cdot \boldsymbol{v}| \overset{\text{(a)}}{=} \|\boldsymbol{u}\|\|\boldsymbol{v}\||\cos\phi| \overset{\text{(b)}}{=} \|\boldsymbol{u}\|\|\boldsymbol{v}\|,$$

where (a) follows from the definition of the angle between two vectors and (b) from $\cos\phi = \pm 1$. □

3.2.3 The triple product in $\mathbb{R}^3$

Definition 3.17: The triple product

The *triple product* of the vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$ is a scalar denoted by $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ and defined as

$$(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w} \in \mathbb{R}.$$

Because the scalar product is commutative, because the vector product is anticommutative, and because $\boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}) = (\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w}$ by Proposition 3.15 (i), the triple product obeys

$$\begin{aligned}(\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w} &= (\boldsymbol{v} \times \boldsymbol{w}) \cdot \boldsymbol{u} = (\boldsymbol{w} \times \boldsymbol{u}) \cdot \boldsymbol{v} \\ &= -(\boldsymbol{v} \times \boldsymbol{u}) \cdot \boldsymbol{w} = -(\boldsymbol{w} \times \boldsymbol{v}) \cdot \boldsymbol{u} = -(\boldsymbol{u} \times \boldsymbol{w}) \cdot \boldsymbol{v},\end{aligned}$$

and so the triple product is invariant under cyclic permutation of its arguments.

The triple product can be computed in any of the following forms:

$$\begin{aligned}(\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w} &= \boldsymbol{u} \cdot (\boldsymbol{v} \times \boldsymbol{w}) \\ &\overset{\text{(a)}}{=} \boldsymbol{u} \cdot \left(\begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix} \hat{\mathbf{i}} - \begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix} \hat{\mathbf{j}} + \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix} \hat{\mathbf{k}} \right) \\ &\overset{\text{(b)}}{=} u_1 \begin{vmatrix} v_2 & v_3 \\ w_2 & w_3 \end{vmatrix} - u_2 \begin{vmatrix} v_1 & v_3 \\ w_1 & w_3 \end{vmatrix} + u_3 \begin{vmatrix} v_1 & v_2 \\ w_1 & w_2 \end{vmatrix} \\ &\overset{\text{(c)}}{=} \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \\ &= u_1 v_2 w_3 + u_2 v_3 w_1 + u_3 v_1 w_2 \\ &\quad - u_1 v_3 w_2 - u_2 v_1 w_3 - u_3 v_2 w_1,\end{aligned}$$

where (a), (b), and (c), like in Equation 3.2 for the vector product, are expressed in terms of the determinant of a square matrix.

Topic 3.8: Geometric interpretation of the triple product

Consider the parallelepiped spanned by the vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^3$. If the vectors $\boldsymbol{u} \times \boldsymbol{v}$ and $\boldsymbol{w}$ are both nonzero, let ϕ denote the angle between them; otherwise, let ϕ be an arbitrary real number. The base of the parallelepiped is the parallelogram spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$ with area A and height h equal to

$$A = \|\boldsymbol{u} \times \boldsymbol{v}\| \quad \text{and} \quad h = \|\boldsymbol{w}\||\cos\phi|.$$

We then make the calculation

$$|(\boldsymbol{u} \times \boldsymbol{v}) \cdot \boldsymbol{w}| = \|\boldsymbol{u} \times \boldsymbol{v}\|\|\boldsymbol{w}\||\cos\phi| = Ah = V,$$

which implies that the absolute value of the triple product $|(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})|$ equals the volume $V = Ah$ of the parallelepiped spanned by $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$. Meanwhile, if $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = 0$, the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ are said to be *coplanar*. Interpreted geometrically, this means that $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ are all parallel to the same plane—we will discuss planes in $\mathbb{R}^3$ in the next section.

We conclude this section by noting that the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$, in that order, satisfy the right-hand rule if, and only if, $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) > 0$.

3.3 Lines and planes in $\mathbb{R}^3$

This section introduces the equation-based formulation of lines and planes in $\mathbb{R}^3$ and covers the calculation of the their mutual positions and orientations.

3.3.1 Equation of a plane in $\mathbb{R}^3$

Definition 3.18: Planes in $\mathbb{R}^3$

Let $\boldsymbol{r}_0$ and $\boldsymbol{n}$ be two vectors in $\mathbb{R}^3$ and assume that $\boldsymbol{n} \neq \boldsymbol{0}$. In this case the *plane* in $\mathbb{R}^3$ perpendicular to $\boldsymbol{n}$ and containing the point with position vector $\boldsymbol{r}_0$ is the set of points $\Sigma \subset \mathbb{R}^3$ defined as

$$\Sigma = \{\boldsymbol{r} \in \mathbb{R}^3;\, (\boldsymbol{r} - \boldsymbol{r}_0) \perp \boldsymbol{n}\} \subset \mathbb{R}^3.$$

The plane Σ is thus the set of all points $\boldsymbol{r} \in \mathbb{R}^3$ solving the equation

$$(\boldsymbol{r} - \boldsymbol{r}_0) \cdot \boldsymbol{n} = 0;$$

this equation is called the *equation of the plane* Σ. The nonzero vector $\boldsymbol{n}$ is called the *normal vector* to the plane Σ; two planes are said to be *parallel* if their normal vectors are parallel.

Topic 3.9: Point-normal and standard equation of a plane

Let $\boldsymbol{r} = (x, y, z)$, $\boldsymbol{r}_0 = (x_0, y_0, z_0)$ and $\boldsymbol{n} = (a, b, c)$. Using this notation, the equation $(\boldsymbol{r} - \boldsymbol{r}_0) \cdot \boldsymbol{n} = 0$, i.e. the equation of the plane with normal $\boldsymbol{n}$ passing through the point $\boldsymbol{r}_0$, can be written in following two equivalent forms

(i) *Point normal form*:

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

(ii) *Standard* or *general form*:

$$ax + by + cz = d.$$

At least one of the constants a, b, c must be nonzero for the plane to be well-defined, in which case the two forms are related by the identity

$$d = ax_0 + by_0 + cz_0.$$

Topic 3.10: Defining a plane with three points

Let $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ be the position vectors of three points in $\mathbb{R}^3$. If the vectors $\boldsymbol{r}_2 - \boldsymbol{r}_1$ and $\boldsymbol{r}_3 - \boldsymbol{r}_1$ are not parallel, there exists exactly one plane $\Sigma \subset \mathbb{R}^3$ that passes through all three points. The normal vector to this plane is

$$\boldsymbol{n} = (\boldsymbol{r}_2 - \boldsymbol{r}_1) \times (\boldsymbol{r}_3 - \boldsymbol{r}_1), \tag{3.3}$$

and the plane itself is the set of all $\boldsymbol{r} \in \mathbb{R}^3$ solving the equation

$$(\boldsymbol{r} - \boldsymbol{r}_1) \cdot \big[(\boldsymbol{r}_2 - \boldsymbol{r}_1) \times (\boldsymbol{r}_3 - \boldsymbol{r}_1)\big] = 0, \tag{3.4}$$

which is just a special case of the equation $(\boldsymbol{r} - \boldsymbol{r}_0) \cdot \boldsymbol{n} = 0$ with normal vector $\boldsymbol{n}$ given by Equation 3.3.

Note that Equation 3.4 is a triple product of the vectors $(\boldsymbol{r} - \boldsymbol{r}_1)$, $(\boldsymbol{r}_2 - \boldsymbol{r}_1)$, and $(\boldsymbol{r}_3 - \boldsymbol{r}_1)$, and so the equation of a plane defined by three points can be written in the determinant form.

$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} = 0.$$

3.3.2 Equation of a line in $\mathbb{R}^3$

Definition 3.19: Lines in $\mathbb{R}^3$

Let $\boldsymbol{r}_0$ and $\boldsymbol{s}$ be two vectors in $\mathbb{R}^3$ and assume $\boldsymbol{s} \neq \boldsymbol{0}$. In this case the *line* in $\mathbb{R}^3$ parallel to the vector $\boldsymbol{s}$ and passing through the point with position vector $\boldsymbol{r}_0$ is the set of points $\ell \subset \mathbb{R}^3$ defined as

$$\ell = \{\boldsymbol{r}_0 + t\boldsymbol{s} \in \mathbb{R}^3 \,;\, t \in \mathbb{R}\} \subset \mathbb{R}^3. \tag{3.5}$$

The line ℓ thus consists of vectors of the form

$$\boldsymbol{r} = \boldsymbol{r}_0 + t\boldsymbol{s}, \qquad t \in \mathbb{R};$$

The formulation of a line in Equation 3.5 in terms of the quantity $\boldsymbol{r}_0 + t\boldsymbol{s}$ is called the

parametric form of a line, and t is called the line's *parameter*. The parameter t determines the position along the line of a point with position vector $\boldsymbol{r} = \boldsymbol{r}_0 + t\boldsymbol{s}$. The vector $\boldsymbol{s}$ is called the line's *direction vector*.

Topic 3.11: Standard equation of a line

Let $\boldsymbol{r} = (x, y, z)$, $\boldsymbol{r}_0 = (x_0, y_0, z_0)$ and $\boldsymbol{s} = (\alpha, \beta, \gamma)$, and assume at least one of the constants α, β, γ must be nonzero. Using this notation, the vector equation $\boldsymbol{r} = \boldsymbol{r}_0 + t\boldsymbol{s}$ can be written as the system of scalar equations

$$\begin{aligned} x &= x_0 + t\alpha \\ y &= y_0 + t\beta \\ z &= z_0 + t\gamma. \end{aligned}$$

Solving the equations for the parameter t and equating the results produces the *standard* form of the equation of a line, which we divide into the following subcases depending on the values of the components α, β, and γ:

(i) If α, β, and γ are all nonzero, the standard equation of a line is

$$\frac{x - x_0}{\alpha} = \frac{y - y_0}{\beta} = \frac{z - z_0}{\gamma}.$$

(ii) If $\alpha = 0$ and $\beta, \gamma \neq 0$, then the standard equation of a line is

$$x = x_0, \quad \frac{y - y_0}{\beta} = \frac{z - z_0}{\gamma}.$$

(iii) If $\alpha = \beta = 0$ and $\gamma \neq 0$, the standard equation of a line is

$$x = x_0, \quad y = y_0, \quad z = z_0 + t\gamma.$$

Analogous equations result if different combinations of α, β, and γ are zero/nonzero, but at least one of these three constants must be nonzero for the line to be well-defined.

Topic 3.12: Specifying a line as an intersection of planes

A line in $\mathbb{R}^3$ can be specified as the intersection of two nonparallel planes Σ_1 and Σ_2. Assume planes Σ_1 and Σ_2 are respectively given in standard form by the equations

$$a_1x + b_1y + c_1z = d_1 \quad \text{and} \quad a_2x + b_2y + c_2z = d_2;$$

and let the planes' normal vectors be given respectively by

$$\boldsymbol{n}_1 = (a_1, b_1, c_1) \quad \text{and} \quad \boldsymbol{n}_2 = (a_2, b_2, c_2).$$

The intersection of the two planes then uniquely specifies a line $\ell = \Sigma_1 \cap \Sigma_2$ consisting of all points $\boldsymbol{r} = (x, y, z) \in \mathbb{R}^3$ solving the system of equations

$$\begin{aligned} a_1x + b_1y + c_1z &= d_1 \\ a_2x + b_2y + c_2z &= d_2. \end{aligned}$$

The point $\boldsymbol{r}_0$ used to specify the line $\ell = \Sigma_1 \cap \Sigma_2$ can be any solution to the above equations, while the line's direction vector can be taken to be

$$\boldsymbol{s} = \boldsymbol{n}_1 \times \boldsymbol{n}_2,$$

which is perpendicular to both $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ and nonzero under the assumption that Σ_1 and Σ_2 are not parallel.

Topic 3.13: Reparameterizing a line in $\mathbb{R}^3$

We have mentioned that a line $\ell \subset \mathbb{R}^3$ parallel to the vector $\boldsymbol{s}$ and containing the point $\boldsymbol{r}_0$ is the set of all points

$$\ell = \{\boldsymbol{r}_0 + t\boldsymbol{s} \in \mathbb{R}^3 ; t \in \mathbb{R}\}.$$

However, it is perfectly reasonable to specify the *same* line with a different direction vector and point—any point $\boldsymbol{r}_0'$ in the line ℓ and any direction vector $\boldsymbol{s}'$ parallel to $\boldsymbol{s}$ specify the same line ℓ.

Because $\boldsymbol{r}_0' \in \ell$, there exists some $t_0 \in \mathbb{R}$ for which $\boldsymbol{r}_0' = \boldsymbol{r}_0 + t_0\boldsymbol{s}$, and because $\boldsymbol{s}'$ is parallel to $\boldsymbol{s}$, there exists some nonzero scalar $\theta \in \mathbb{R}$ for which $\boldsymbol{s}' = \theta\boldsymbol{s}$. The new point $\boldsymbol{r}_0'$ and direction vector $\boldsymbol{s}'$ specify a new parameterization of the line ℓ, namely

$$\ell = \{\boldsymbol{r}_0' + t'\boldsymbol{s}' \in \mathbb{R}^3 ; t' \in \mathbb{R}\}.$$

The process of changing initial point and direction vector used to specify a given line is called *reparameterization*. The original and new parameterizations are related by the equality

$$\boldsymbol{r} = \boldsymbol{r}_0' + t's \overset{\text{(a)}}{=} (\boldsymbol{r}_0 + t_0\boldsymbol{s}) + t'\theta\boldsymbol{s} = \boldsymbol{r}_0 + (t_0 + t'\theta)\boldsymbol{s} = \boldsymbol{r}_0 + t\boldsymbol{s},$$

where (a) follows from the above expressions for $\boldsymbol{r}_0'$ and $\boldsymbol{s}'$; the original and new parameters are thus related by the equation

$$t = t_0 + t'\theta.$$

After reparameterization, the standard equation of a line in Topic 3.11 (i) becomes

$$\frac{x - (x_0 + t_0\alpha)}{\theta\alpha} = \frac{y - (y_0 + t_0\beta)}{\theta\beta} = \frac{z - (z_0 + t_0\gamma)}{\theta\gamma},$$

or, after rearranging into a more interpretable form,

$$\frac{1}{\theta}\left(\frac{x - x_0}{\alpha} - t_0\right) = \frac{1}{\theta}\left(\frac{y - y_0}{\beta} - t_0\right) = \frac{1}{\theta}\left(\frac{z - z_0}{\gamma} - t_0\right). \tag{3.6}$$

Equation 3.6 means that the new standard-form equation of the line ℓ can be obtained from the original equation by a translation by t_0 and a scaling by $1/\theta$. Thus two standard-form equations correspond to the same line if one set of equations can be obtained from the other by translation and scaling of the original.

We illustrate how the same line can have multiple standard-form equations in the following

example.

Example 3.20: Two standard-form equations of the same line

Let the lines $\ell_1 \in \mathbb{R}^3$ and $\ell_2 \in \mathbb{R}^3$ be given respectively by the standard-form equations

$$\frac{x-5}{2} = \frac{y-7}{3} = \frac{z}{5} \quad \text{and} \quad \frac{x-1}{4} = \frac{y-1}{6} = \frac{z+10}{10}.$$

From observing the fractions' denominators, the two lines have direction vectors

$$\boldsymbol{s}_1 = (2, 3, 5) \quad \text{and} \quad \boldsymbol{s}_2 = (4, 6, 10);$$

since $\boldsymbol{s}_1$ is a scalar multiple of $\boldsymbol{s}_2$, the vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, and thus the lines ℓ_1 and ℓ_2, are parallel. The lines being parallel leaves two possibilities: either the two lines are equal or they do not intersect at all, so it suffices to show that the lines contain a single mutual point to show that the lines are in fact equal. By observing the fractions' numerators, the first line contains the point $(5, 7, 0)$; substituting this point into the equation of ℓ_2 gives

$$\frac{5-1}{4} = \frac{7-1}{6} = \frac{0+10}{10} = 1 = 1 = 1.$$

The point $(5, 7, 0)$ thus solves the equation of, and thus lies on, the line ℓ_2, implying that ℓ_1 and ℓ_2 are the same line.

3.3.3 The mutual orientation of points, lines, and planes in $\mathbb{R}^3$

We conclude this chapter with a survey of techniques for computing the distances and angles between points, lines, and planes.

Topic 3.14: Distance between two points

The distance between any two points with position vectors $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathbb{R}^3$ is

$$d(\boldsymbol{r}_1, \boldsymbol{r}_2) = \|\boldsymbol{r}_2 - \boldsymbol{r}_1\|.$$

Topic 3.15: Distance between a point and a line

Let $\boldsymbol{r}_1$ be any point in $\mathbb{R}^3$ and let $\ell \subset \mathbb{R}^3$ be a line with the parameterization

$$\boldsymbol{r} = \boldsymbol{r}_0 + t\boldsymbol{s}, \quad \boldsymbol{r}_0 \in \mathbb{R}^3 \text{ and } \boldsymbol{s} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}.$$

The distance between the point with position vector $\boldsymbol{r}_1$ and the line ℓ is defined as the distance between $\boldsymbol{r}_1$ and the point on ℓ closest to $\boldsymbol{r}_1$. From geometric arguments, this distance is the height d of a parallelogram with base $\boldsymbol{s}$ and side $\boldsymbol{r}_1 - \boldsymbol{r}_0$. Two equivalent expressions for the area of the this parallelogram are

$$A = \|\boldsymbol{s}\|d = \|\boldsymbol{s} \times (\boldsymbol{r}_1 - \boldsymbol{r}_0)\| = A,$$

which we then rearrange to solve for the distance between $\boldsymbol{r}_1$ and ℓ

$$d(\boldsymbol{r}_1, \ell) = \frac{\|\boldsymbol{s} \times (\boldsymbol{r}_1 - \boldsymbol{r}_0)\|}{\|\boldsymbol{s}\|}.$$

Topic 3.16: Distance between a point and a plane

Let $\boldsymbol{r}_1$ be any point in $\mathbb{R}^3$ and let $\Sigma \subset \mathbb{R}^3$ be the plane given by

$$(\boldsymbol{r} - \boldsymbol{r}_0) \cdot \boldsymbol{n} = 0, \quad \boldsymbol{r}_0 \in \mathbb{R}^3 \text{ and } \boldsymbol{n} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}.$$

The distance between the point with position vector $\boldsymbol{r}_1$ and the plane Σ is defined as the distance between $\boldsymbol{r}_1$ and the point in Σ closest to $\boldsymbol{r}_1$. This distance equals the magnitude of the orthogonal projection of the vector $\boldsymbol{r}_1 - \boldsymbol{r}_0$ onto the plane's normal vector $\boldsymbol{n}$, i.e.

$$\begin{aligned} d(\boldsymbol{r}_1, \Sigma) &= \left\| \mathrm{pr}_{\boldsymbol{n}}^{\perp}(\boldsymbol{r}_1 - \boldsymbol{r}_0) \right\| = \left\| \left(\frac{(\boldsymbol{r}_1 - \boldsymbol{r}_0) \cdot \boldsymbol{n}}{\boldsymbol{n} \cdot \boldsymbol{n}} \right) \boldsymbol{n} \right\| \\ &= \frac{|(\boldsymbol{r}_1 - \boldsymbol{r}_0) \cdot \boldsymbol{n}|}{\|\boldsymbol{n}\|}. \end{aligned}$$

Topic 3.17: Distance between parallel lines

Let ℓ_1 and ℓ_2 be two parallel lines in $\mathbb{R}^3$ given by the parameterizations

$$\boldsymbol{r} = \boldsymbol{r}_1 + t\boldsymbol{s}_1 \quad \text{and} \quad \boldsymbol{r} = \boldsymbol{r}_2 + t\boldsymbol{s}_2.$$

Recall that two lines are parallel if $\boldsymbol{s}_1 \times \boldsymbol{s}_2 = \boldsymbol{0}$. In this case, the distance between the lines equals the distance between the first line and any point on the second line, or vice versa; using the formula in Topic 3.15 for the distance between a point and a line, the distance between two parallel is thus

$$d(\ell_1, \ell_2) = \frac{\|\boldsymbol{s}_2 \times (\boldsymbol{r}_1 - \boldsymbol{r}_2)\|}{\|\boldsymbol{s}_2\|}.$$

Topic 3.18: Distance between nonparallel lines

Let ℓ_1 and ℓ_2 be two nonparallel lines in $\mathbb{R}^3$ given by the parameterizations

$$\boldsymbol{r} = \boldsymbol{r}_1 + t\boldsymbol{s}_1 \quad \text{and} \quad \boldsymbol{r} = \boldsymbol{r}_2 + t\boldsymbol{s}_2;$$

recall that two lines are nonparallel if $\boldsymbol{s}_1 \times \boldsymbol{s}_2 \neq \boldsymbol{0}$. In this case, the distance between the lines equals the height d of the parallelepiped with side $\boldsymbol{r}_2 - \boldsymbol{r}_1$ and whose base is the parallelogram spanned by the direction vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. The volume of the parallelepiped is on the one hand equal to the absolute value of the triple product of $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, and $\boldsymbol{r}_2 - \boldsymbol{r}_1$ and on the other hand equal to the base-height product $A \cdot h = \|\boldsymbol{s}_1 \times \boldsymbol{s}_2\| \cdot d$. We first equate these two volume expressions:

$$V = \|\boldsymbol{s}_1 \times \boldsymbol{s}_2\| \cdot d = |(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{r}_2 - \boldsymbol{r}_1)| = V,$$

then solve for the distance d to get

$$d(\ell_1, \ell_2) = \frac{|(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{r}_2 - \boldsymbol{r}_1)|}{\|\boldsymbol{s}_1 \times \boldsymbol{s}_2\|} = \frac{|(\boldsymbol{s}_1 \times \boldsymbol{s}_2) \cdot (\boldsymbol{r}_2 - \boldsymbol{r}_1)|}{\|\boldsymbol{s}_1 \times \boldsymbol{s}_2\|}.$$

Topic 3.19: Distance between a line and a plane

If a line and plane intersect, the distance between them is zero. If a line and plane do not intersect, the distance between them is distance between the plane and any point on the line as covered in Topic 3.16.

Topic 3.20: Distance between two planes

If two planes intersect, the distance between them is zero. If two planes do not intersect, the distance between them is distance between the first plane and any point on the second plane (or vice versa) as covered in Topic 3.16.

Topic 3.21: Angle between two lines

Let ℓ_1 and ℓ_2 be two lines in $\mathbb{R}^3$ given by the parameterizations

$$\boldsymbol{r} = \boldsymbol{r}_1 + t\boldsymbol{s}_1 \quad \text{and} \quad \boldsymbol{r} = \boldsymbol{r}_2 + t\boldsymbol{s}_2.$$

The angle ϕ between the lines falls in the range $\phi \in [0, \pi/2]$ and is determined by the mutual orientation of the lines' direction vectors. If the angle between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ is acute or a right angle, the angle between ℓ_1 and ℓ_2 equals the angle between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. If the angle between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ is obtuse, the angle between ℓ_1 and ℓ_2 equals the angle between $\boldsymbol{s}_1$ and $-\boldsymbol{s}_2$. Using the expression for the angle between two vectors from Topic 3.5, a single formula for the angle between two lines is

$$\cos\phi = \frac{|\boldsymbol{s}_1 \cdot \boldsymbol{s}_2|}{\|\boldsymbol{s}_1\|\|\boldsymbol{s}_2\|}.$$

Topic 3.22: Angle between a line and a plane

Let ℓ be a line in $\mathbb{R}^3$ with the parameterization

$$\boldsymbol{r} = \boldsymbol{r}_1 + t\boldsymbol{s}, \quad \boldsymbol{r}_1 \in \mathbb{R}^3 \text{ and } \boldsymbol{s} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}.$$

and let Σ be the plane in $\mathbb{R}^3$ given by

$$(\boldsymbol{r} - \boldsymbol{r}_2) \cdot \boldsymbol{n} = 0, \quad \boldsymbol{r}_2 \in \mathbb{R}^3 \text{ and } \boldsymbol{n} \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}.$$

The angle between the line ℓ and plane Σ falls in the range $\phi \in [0, \pi/2]$ and equals the angle between the line ℓ and ℓ's orthogonal projection onto the plane Σ. From geometrical arguments, the angle between ℓ and a hypothetical line with direction vector $\boldsymbol{n}$ equals $\pi/2 - \phi$. Using the trigonometric identity $\sin\phi = \cos(\pi/2 - \phi)$ and the formula for the distance between two lines, the angle between ℓ and Σ is then

$$\sin\phi = \frac{|\boldsymbol{s} \cdot \boldsymbol{n}|}{\|\boldsymbol{s}\|\|\boldsymbol{n}\|}.$$

Topic 3.23: Angle between two planes

Let Σ_1 and Σ_2 be two plane in $\mathbb{R}^3$ defined by the equations

$$(\boldsymbol{r} - \boldsymbol{r}_1) \cdot \boldsymbol{n}_1 = 0 \quad \text{and} \quad (\boldsymbol{r} - \boldsymbol{r}_2) \cdot \boldsymbol{n}_2 = 0,$$

where $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathbb{R}^3$ and $\boldsymbol{n}_1, \boldsymbol{n}_2 \in \mathbb{R}^3 \setminus \{\boldsymbol{0}\}$. The angle between the planes falls in the range $\phi \in [0, \pi/2]$ and equals the angle between two lines with direction vectors equal to $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$; referring to Topic 3.21, the angle between the planes is

$$\cos\phi = \frac{|\boldsymbol{n}_1 \cdot \boldsymbol{n}_2|}{\|\boldsymbol{n}_1\|\|\boldsymbol{n}_2\|}.$$

4 Sequences of numbers

4.1 Basic concepts

Definition 4.1: Sequences and elements

- A *sequence of complex numbers*, or simply a *complex sequence*, is a function of the form $\mathbb{N} \to \mathbb{C}$ mapping $n \mapsto a_n$ and is usually written with one of the following notations

$$(a_1, a_2, a_3, \ldots) = (a_n)_{n\in\mathbb{N}} = (a_n)_n = (a_n).$$

  The individual numbers in the sequence (e.g. a_1, a_2, etc.) are called the sequence's *elements*. The n-th element of the complex sequence $(a_n)_{n\in\mathbb{N}}$ is the number $a_n \in \mathbb{C}$.

- The *image* of a complex sequence $(a_n)_{n\in N}$ is the set of the sequence's elements:

$$\text{image}\,(a_n)_{n\in N} = \{a_1, a_2, a_3, \ldots\} = \{a_n\,;\, n \in \mathbb{N}\} \subset \mathbb{C}.$$

If all elements in the sequence $(a_n)_{n\in\mathbb{N}}$ are real (or rational, integer, or natural) numbers, then $(a_n)_{n\in\mathbb{N}}$ is called a sequence of real (or rational, integer, or natural) numbers and its image is a subset of $\mathbb{R}$ (or $\mathbb{Q}$, $\mathbb{Z}$, or $\mathbb{N}$). This section primarily concerns sequences of real numbers.

We now extend the concepts of boundedness introduced in Section 2 to the description of sequences.

Definition 4.2: Boundedness of real sequences

Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. In this case

(i) The real sequence $(a_n)_{n\in\mathbb{N}}$ is *bounded* if its image is a bounded subset of $\mathbb{R}$.

(ii) The real sequence $(a_n)_{n\in\mathbb{N}}$ is *bounded above* if its image is an above-bounded subset of $\mathbb{R}$. In this case the sequence's *supremum* is the number

$$\sup\,(a_n)_{n\in\mathbb{N}} = \sup\{a_n\,;\, n \in \mathbb{N}\}.$$

(iii) The real sequence $(a_n)_{n\in\mathbb{N}}$ is *bounded below* if its image is a below-bounded subset of $\mathbb{R}$. In this case the sequence's *infimum* is the number

$$\inf\,(a_n)_{n\in\mathbb{N}} = \inf\{a_n\,;\, n \in \mathbb{N}\}.$$

The concept of general boundedness also applies to complex-valued sequences: a complex-valued sequence is bounded if its image is a bounded subset of $\mathbb{C}$. However, boundedness above and below, and thus the supremum and infimum, cannot be naturally extended to complex sequences because, as mentioned in the comments following Topic 2.14, the complex numbers lack a natural ordering relation.

Definition 4.3: Growth of real sequences

Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. In this case

(i) The sequence $(a_n)_{n\in\mathbb{N}}$ is said to be *increasing* if

$$a_{n+1} \geq a_n \text{ for all } n \in \mathbb{N}.$$

(ii) The sequence $(a_n)_{n\in\mathbb{N}}$ is said to be *decreasing* if

$$a_{n+1} \leq a_n \text{ for all } n \in \mathbb{N}.$$

(iii) The sequence $(a_n)_{n\in\mathbb{N}}$ is *monotonic* if it is either increasing or decreasing.

(iv) The sequence $(a_n)_{n\in\mathbb{N}}$ is said to be *strictly increasing* if

$$a_{n+1} > a_n \text{ for all } n \in \mathbb{N}.$$

(v) The sequence $(a_n)_{n\in\mathbb{N}}$ is said to be *strictly decreasing* if

$$a_{n+1} < a_n \text{ for all } n \in \mathbb{N}.$$

(vi) The sequence $(a_n)_{n\in\mathbb{N}}$ is *strictly monotonic* if it is either strictly increasing or strictly decreasing.

Remark 4.4: Indexing sequences

We have defined sequences as functions with the natural numbers $\mathbb{N}$ as the domain—such sequences are said to be *indexed by the natural numbers*.

More generally, the set of indices can be any nonempty subset $J \subset \mathbb{Z}$. A subset of complex numbers *indexed by the subset* J is a function $J \to \mathbb{C}$ mapping $j \to u_j$ and written in the form $(u_j)_{j\in J}$. Common examples of indexing sets are $\mathbb{N}$, $\mathbb{N} \cup \{0\}$, and $\mathbb{Z} \cap [p, \infty)$ for some $p \in \mathbb{Z}$. We will consider only sequences indexed by sets are infinite and below-bounded subsets of the integers $\mathbb{Z}$; the analysis of such sequences is very similar to the analysis of sequences indexed by $\mathbb{N}$.

We now offer a few examples of important sequences.

Example 4.5: Common sequences

(i) For all $a \in \mathbb{C}$ the *constant sequence* with value a is the sequence

$$(a, a, a, \ldots);$$

every element in the constant sequence is equal to a.

(ii) The sequence of natural numbers $(1, 2, 3, 4, 5, \ldots)$ is given by the general mapping $a_n = n$ for all $n \in \mathbb{N}$. This sequence is strictly increasing and bounded below, but is not bounded above.

(iii) Let $a, d \in \mathbb{C}$ be two complex numbers. The *arithmetic sequence* with initial

element a and difference d is the sequence

$$(a, a+d, a+2d, a+3d, \ldots), \quad a_n = a + (n-1)d, \quad n \in \mathbb{N}.$$

The arithmetic sequence may also be defined recursively by specifying the initial element $a_1 = a$ and formula $a_{n+1} = a_n + d$ for all $n \in \mathbb{N}$.

(iv) Let $a, q \in \mathbb{C}$ be two complex numbers. The *geometric sequence* with initial element a and ratio q is the sequence

$$(a, aq, aq^2, aq^3, \ldots), \quad a_n = aq^{n-1}.$$

The geometric sequence may also be defined recursively by specifying the initial element $a_1 = a$ and formula $a_{n+1} = a_n q$ for all $n \in \mathbb{N}$. As a concrete example, the sequence

$$(1, \mathrm{i}, -1, -\mathrm{i}, 1, \mathrm{i}, -1, -\mathrm{i}, 1, \ldots)$$

with the generic element $a_n = \mathrm{i}^{n-1}$ is a bounded geometric sequence with initial element 1 and ratio i.

(v) For any nonnegative real number a, the sequence $(a/n)_n$ is a decreasing sequence with infimum 0.

(vi) For any nonnegative real number $q \in (0, 1)$, the sequence $(aq^{n-1})_n$ is also a decreasing sequence with infimum 0.

4.2 Cluster points of sequences

Intuitively, cluster points are the points in the complex plane around which large numbers of a sequence's elements tend to accumulate. While cluster points themselves need not be elements of a sequence, a sequence's cluster points can be arbitrarily well-approximated with an infinite number of elements from the sequence.

The definition of cluster points relies on the concepts of neighborhoods and interior points introduced in Section 2.3.3; you might wish to review those concepts now.

Definition 4.6: Cluster point

The point $a \in \mathbb{C}$ is a *cluster point* of the complex sequence (a_n) if for every neighborhood $V \subset \mathbb{C}$ of the point a, the set $\{n \in \mathbb{N};\, a_n \in V\}$ is infinite, i.e. if every neighborhood of a contains infinitely many elements of the sequence.

In practice, one often uses the following condition to determine if a number is a cluster point of a complex sequence.

Proposition 4.7: Condition for cluster points

The point $a \in \mathbb{C}$ is a cluster point of the sequence (a_n) if, and only if, for all $\epsilon \in \mathbb{R}^+$ and for all $N \in \mathbb{N}$ there exists a natural number for which $n \geq N$ and

$$|a_n - a| < \epsilon \quad \text{or, equivalently,} \quad a_n \in \mathcal{K}(a, \epsilon),$$

where $\mathcal{K}(a, \epsilon)$ is the complex disk of radius ϵ centered at $a \in \mathbb{C}$.

Intuitively, this condition states that a is a cluster point of (a_n) if, and only if, it is possible to make an arbitrarily good approximation of a with elements of (a_n). In other words, for any ϵ, however small, there must exist some a_n for which a_n and a are separated by less than ϵ.

Proof. ($\Rightarrow$) Let a be a cluster point of (a_n), and choose arbitrary $\epsilon \in \mathbb{R}^+$ and $N \in \mathbb{N}$. The open disk $\mathcal{K}(a, \epsilon)$ is a neighborhood of a by construction, so under the assumption that a is a cluster point of (a_n), the set $\{n \in \mathbb{N};\, a_n \in \mathcal{K}(a, \epsilon)\}$ is infinite. Because the set is infinite, it necessarily contains elements a_n with index $n \geq N$ for any finite $N \in \mathbb{N}$, i.e. there exists $n \in \mathbb{N}$ for which $n \geq \mathbb{N}$ and $a_n \in \mathcal{K}(a, \epsilon)$.

($\Leftarrow$) Let $V \in \mathbb{C}$ be an arbitrary neighborhood of the point a, meaning that there exists $\epsilon \in \mathbb{R}^+$ for which $\mathcal{K}(a, \epsilon) \subset V$. By the proposition's assumptions, for all $N \in \mathbb{N}$ there exists $n \in \mathbb{N}$ for which $n \geq N$ and $a_n \in \mathcal{K}(a, \epsilon)$. The set $\mathcal{K}(a, \epsilon)$ thus contains infinitely many elements of the sequence (a_n), and so the set V also contains infinitely many elements of (a_n), meaning that a is a cluster point of (a_n). □

We now make the important observation that cluster points of sequences of real numbers can only be real numbers. To verify this, we first note that $\mathcal{K}(a, \epsilon) \cap \mathbb{R} = (a - \epsilon, a + \epsilon)$ for all $a \in \mathbb{R}$ and all $\epsilon \in \mathbb{R}^+$, i.e. the real component of an open disk of radius ϵ centered at a equals the open interval $(a - \epsilon, a + \epsilon)$.

Thus, when verifying if a point $a \in \mathbb{R}$ is a cluster point of some sequence, one should only consider neighborhoods of a that are open real intervals centered a. In other words, the point $a \in \mathbb{R}$ is a cluster point of a sequence of real numbers if, and only if, for all $\epsilon \in \mathbb{R}^+$ infinitely many elements of the sequence occur in the interval $(a - \epsilon, a + \epsilon)$.

Example 4.8: Cluster points

(1) The sequence $(1, 2, 3, 4, \ldots)$ with general term $a_n = n$ has no cluster points, since any bounded interval on the real line contains only finitely many elements of the sequence.

(2) The constant sequence $(a, a, a, \ldots)$ with general term $a_n = a, a \in \mathbb{C}$ has one cluster point—the point a.

(3) The geometric sequence $(1, \mathrm{i}, -1, -\mathrm{i}, 1, \mathrm{i}, -1, -\mathrm{i}, 1, \ldots)$ with general term $a_n = \mathrm{i}^{n-1}$ has four cluster points—the points 1, i, -1, and $-\mathrm{i}$.

(4) If (a_n) is a sequence of rational numbers whose image is the entire set $\mathbb{Q}$, then every real number is a cluster point of (a_n), since an arbitrary interval of any real number contains infinitely many rational numbers.

We now state the very important Bolzano-Weierstrass theorem, which will play a central role in the analysis of sequences for the remainder of this book.

Theorem 4.9: Bolzano-Weierstrass theorem for real sequences

Every bounded sequence of real numbers has at least one cluster point.

Proof. Let (a_n) be a bounded sequence of real numbers. Since (a_n) is bounded, it

must have a supremum and infimum—we denote these by

$$m = \inf(a_n) \quad \text{and} \quad M = \sup(a_n).$$

We then define the set U as follows:

$$U = \{x \in \mathbb{R};\, x > a_n \text{ for only finitely many } n \in \mathbb{N}\}.$$

Geometrically, U is the set of all real numbers x for which there are only finitely many elements of (a_n) to the left of x on the real line.

By its definition, the set U necessarily contains the sequence's infimum m, since no elements of (a_n) occur to the left of m. As a consequence, U is a nonempty set. Similarly, U necessarily does not contain any points to the right of the sequence's supremum M, since all of (a_n)'s (infinitely many) elements occur to the left of M. The set U is thus bounded above with M as an upper bound.

For shorthand, let $a = \sup(U)$. We will complete the proof by showing that a is a cluster point of the sequence (a_n). By Proposition 4.7, it suffices to show that for all $\epsilon \in \mathbb{R}^+$, infinitely many elements of the sequence (a_n) lie in the interval $(a-\epsilon, a+\epsilon)$.

Because the point a is the supremum of U, there exists some $x \in U$ for which $a - \epsilon < x \leq a$. By the definition of U and because $x \in U$, only finitely many elements of (a_n) occur to the left of x on the real line. Because a is the supremum of U, the point $a + \epsilon$ is certainly not in U, which again by the definition of U (and the fact that (a_n) has infinitely many elements) means that infinitely many elements of (a_n) occur to the left of $a + \epsilon$ on the real line. Of the infinitely many elements of (a_n) to the left of $a + \epsilon$, only finitely many occur to the left of x, so infinitely many must lie in the interval $[x, a + \epsilon)$. And because $[x, a + \epsilon)$ is a subset of $(a - \epsilon, a + \epsilon)$, the interval $(a - \epsilon, a + \epsilon)$ also contains infinitely many elements of (a_n), so a is a cluster point of (a_n) by Proposition 4.7. □

It is also possible to prove the Bolzano-Weierstrass theorem with the so-called bisection method; we offer the proof in the following remark.

Remark 4.10: Alternate proof of the Bolzano-Weierstrass theorem

As in Theorem 4.9, let (a_n) be a bounded sequence of real numbers with infimum $m = \inf(a_n)$ and supremum $M = \sup(a_n)$. We then define the real sequences (m_n) and (M_n) as follows:

(i) The sequence (m_n) is increasing and the sequence (M_n) is decreasing.

(ii) $M_n - m_n = (1/2)^{n-1}(M - m)$ for all $n \in \mathbb{N}$. Geometrically, this condition means that the distance between M_n and m_n on the real line begins at $(M - m)$ and decreases by a factor of $1/2$ for each additional $n \in \mathbb{N}$.

(iii) For all $n \in \mathbb{N}$, the closed interval $[m_n, M_n]$ contains infinitely many elements of the sequence (a_n).

We will now construct the sequences (m_n) and (M_n) recursively. The sequences begin with $m_1 = m$ and $M_1 = M$. Suppose we have already constructed the first k

terms $m_1, m_2, \ldots, m_k$ and $M_1, M_2, \ldots, M_k$ with the required properties, and let

$$c_k = (M_k - m_k)/2$$

denote the midpoint of M_k and m_k on the real line. Then:

- If the interval $[m_k, c_k]$ contains infinitely many elements of (a_n), we define $m_{k+1} = m_k$ and $M_{k+1} = c_k$.
- Otherwise, i.e. if the interval $[c_k, M_k]$ contains finitely many elements of (a_n), we define $M_{k+1} = M_k$ and $m_{k+1} = c_k$.

The thus-constructed sequences (m_n) and (M_n) satisfy items (i) through (iii) above.

For shorthand, let $a = \sup(m_n)$ and $b = \inf(M_n)$. By construction, the sequence (m_n) is increasing and bounded above by every element of (M_n), and, reciprocally, the sequence (M_n) is decreasing and bounded below by every element of (m_n)—these two properties imply $a \leq b$. Particularly, $m_k \leq a < b \leq M_k$ for all $k \in \mathbb{N}$, which we then rearrange to get

$$b - a \leq M_k - m_k \overset{\text{(a)}}{=} (1/2)^{k-1} \cdot (M - m),$$

where (a) uses point (ii) in the construction of (m_n) and (M_n). By its construction, the sequence

$$\big((1/2)^{k-1} \cdot (M - m)\big)_k$$

is a decreasing sequence of nonnegative real numbers with infimum 0, which together with the definition of (m_n) and (M_n) implies that $a = b$.

We will now show that the point a is a cluster point of the sequence (a_n). We first define arbitrary $\epsilon \in \mathbb{R}^+$. Because $a = \sup(m_n) = \inf(M_n) = b$, the sequence (m_n) will fall arbitrarily close to a and (M_n) will grow arbitrarily close to b, so there exists an index $n \in \mathbb{N}$ for which

$$a - \epsilon < m_n \leq M_n < a + \epsilon.$$

The interval $(a - \epsilon, a + \epsilon)$ thus contains all elements of (a_n) that fall in the interval $[m_n, M_n]$, of which there are infinitely many by point (iii) in the construction of (m_n) and (M_n). Because infinitely many elements of (a_n) fall in the interval $(a - \epsilon, a + \epsilon)$, a is a cluster point of the sequence (a_n). □

4.3 The limit of a sequence

Definition 4.11: Limit of a sequence

The point $a \in \mathbb{C}$ is the *limit* of the sequence of complex numbers (a_n) if for every neighborhood $V \subset C$ of the point a, the set $\{n \in \mathbb{N};\ a_n \notin V\}$ is finite.

In other words, the point a is the limit of the sequence (a_n) if, and only if, every neighborhood V of the point a contains all of (a_n)'s elements, with the possible exception of *finitely* many.

Analogously to testing if some point is a cluster point of a sequence, when testing if a point

a is the limit of a sequence, it is sufficient to consider only neighborhoods of a that are open disks centered at a. In other words, a is the limit of a sequence if, and only if, for all $\epsilon \in \mathbb{R}^+$ the open disk $\mathcal{K}(a, \epsilon)$ contains all of the sequence's elements with the possible exception of finitely many.

We now state an analog of Proposition 4.7 applicable to limits of sequences.

Proposition 4.12: Condition for the limit of sequence

The point $a \in \mathbb{C}$ is the limit of the complex sequence (a_n) if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists a natural number $N \in \mathbb{N}$ for which

$$|a_n - a| < \epsilon \quad \text{or, equivalently,} \quad a_n \in \mathcal{K}(a, \epsilon),$$

for all natural numbers $n \in \mathbb{N}$ for which $n \geq N$. Loosely, this condition means that *all* of the sequence's elements—from some element onward—must grow arbitrarily arbitrarily close the limit a.

Proof. ($\Rightarrow$) If a is the limit of the sequence (a_n), then for all $\epsilon \in \mathbb{R}^+$ the open disk $\mathcal{K}(a, \epsilon)$ contains all of (a_n)'s elements with the possible exception of finitely many, so there must exist some $N \in \mathbb{N}$ for which $a_n \in \mathcal{K}(a, \epsilon)$ for all natural numbers $n \geq N$.

($\Leftarrow$) For any neighborhood $V \subset \mathbb{C}$ of the point a there exists some $\epsilon \in \mathbb{R}^+$ for which the open disk $\mathcal{K}(a, \epsilon)$ is a subset of V. As in the proposition's assumptions, assume that for all $\epsilon \in \mathbb{R}^+$ there exists $N \in \mathbb{N}$ for which

$$a_n \in \mathcal{K}(a, \epsilon) \text{ for all } n \in \mathbb{N} \text{ for which } n \geq N.$$

This means that infinitely many elements of (a_n) lie in the open disk $\mathcal{K}(a, \epsilon)$. Since $\mathcal{K}$ is a subset of V, infinitely many elements of (a_n) must also lie in the V, and so a is the limit of (a_n). □

Proposition 4.13: A sequence's limit is its only cluster point

Any complex sequence has at most one limit, and if $a \in \mathbb{C}$ is the limit of the complex sequence (a_n), then a is (a_n)'s only cluster point.

Proof. That a complex sequence can have at most one limit follows directly from the definition of the limit of a sequence—it would be impossible for *all* of a sequence's elements, with the possible exception of *finitely* many, to occur in every neighborhood of two distinct points.

Continuing on, by definition, the limit a of a sequence (a_n) is also a cluster point of the sequence, since infinitely many elements of (a_n) occur in an arbitrary neighborhood of a. Assume that $b \in \mathbb{C}$ is an arbitrary point different from a, and let

$$\epsilon = |b - a|/2 > 0.$$

Since a is the limit of (a_n), the open disk $\mathcal{K}(a, \epsilon)$ contains all elements of (a_n) possibly except for finitely many. Because of the condition $\epsilon = |b - a|/2$, the open disks $\mathcal{K}(a, \epsilon)$ and $\mathcal{K}(b, \epsilon)$ are disjoint, and so the disk $\mathcal{K}(b, \epsilon)$ necessarily contains only finitely many elements of (a_n), implying that b is not a cluster point of (a_n). Since b

was an arbitrary point other than a, it follows that no point other than a can be a cluster point of (a_n). □

Definition 4.14: Convergence and divergence of sequences

A complex sequence (a_n) is called *convergent* if it has a limit, and if $a \in \mathbb{C}$ is the limit of the sequence (a_n), then (a_n) is said to *converge to* the point a. We denote the limit of a convergent sequence (a_n) by

$$\lim_{n\to\infty}(a_n) = \lim_{n\to\infty} a_n \in \mathbb{C}.$$

A sequence is said to be *divergent* if it is not convergent.

Proposition 4.15: Monotone convergence theorem

(i) Every increasing, above-bounded sequence of real numbers (a_n) converges to the limit

$$\lim_{n\to\infty} a_n = \sup\{a_n\,;\, n \in \mathbb{N}\}.$$

(ii) Every decreasing, below-bounded sequence of real numbers (a_n) converges to the limit

$$\lim_{n\to\infty} b_n = \inf\{b_n\,;\, n \in \mathbb{N}\}.$$

Proof.

(i) Let (a_n) be an increasing, above-bounded real sequence, let $a = \sup(a_n)$, and choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the point a is the supremum of the sequence (a_n), there exists $N \in \mathbb{N}$ for which

$$a - \epsilon < a_N \leq a.$$

Because (a_n) is increasing, the condition $a - \epsilon < a_N \leq a_n \leq a$ holds for all $n \geq N$, and the inequality $a - \epsilon < a_n \leq a$ for all $n \geq N$ means that (a_n) grows arbitrarily close to a with increasing n, and thus converges to a.

(ii) Analogous to the proof of (i). □

Example 4.16: Limits of common sequences

(1) The constant sequence $(a, a, a, \ldots)$ with general term $a_n = a \in \mathbb{C}$ has the limit a.

(2) The sequence with general term $a_n = 1/n$ is strictly decreasing with infimum 0, and so it converges with limit $\lim_{n\to\infty} 1/n = 0$ by Proposition 4.15.

(3) For all positive real numbers $q \in (0, 1)$, the sequence (q^n) is strictly decreasing with infimum 0, and so it converges with limit $\lim_{n\to\infty} q^n = 0$.

(4) The sequence with general term $a_n = (-1)^n$ has two cluster points, -1 and 1, which implies that the sequence does not have a limit and is thus divergent.

(5) Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of complex numbers and let $k \in \mathbb{N}$ be an arbitrary natural number. By the definition of a limit, the number $a \in \mathbb{C}$ is the limit of the sequence $(a_n)_{n\in\mathbb{N}}$ if, and only if, a is also the limit of the shifted sequence $(a_{n+k})_{n\in\mathbb{N}}$ for all $k \in \mathbb{N}$. In other words, a finite number of elements has no effect on the convergence and limit of a sequence.

The following rules simplify the calculation of the limits of combinations of sequences.

Proposition 4.17: Rules for computing limits of sequences

Let (a_n) and (b_n) be two convergent complex sequences and define the scalar $\alpha \in \mathbb{C}$. In this case:

(i) The sequence $(a_n + b_n)_n$ converges to the limit

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty}(a_n) + \lim_{n\to\infty}(b_n).$$

(ii) The sequence $(\alpha\alpha_n)_n$ converges to the limit

$$\lim_{n\to\infty}(\alpha a_n) = \alpha \lim_{n\to\infty}(a_n).$$

(iii) The sequence $(a_n b_n)_n$ converges to the limit

$$\lim_{n\to\infty}(a_n b_n) = \left[\lim_{n\to\infty}(a_n)\right] \cdot \left[\lim_{n\to\infty}(b_n)\right].$$

(iv) If $b_n \neq 0$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} b_n \neq 0$, the sequence $(a_n/b_n)_n$ converges to

$$\lim_{n\to\infty}(a_n/b_n) = \left[\lim_{n\to\infty}(a_n)\right] / \left[\lim_{n\to\infty}(b_n)\right].$$

(v) The sequence $(\overline{a_n})_n$ converges to the limit

$$\lim_{n\to\infty}(\overline{a_n}) = \overline{\lim_{n\to\infty}(a_n)},$$

where $\overline{a}$ denotes the complex conjugate of the number $a \in \mathbb{C}$.

(vi) The sequence $(|a_n|)_n$ converges to the limit

$$\lim_{n\to\infty}(|a_n|) = \left|\lim_{n\to\infty}(a_n)\right|.$$

Proof. For shorthand, let $a = \lim_{n\to\infty}(a_n)$ and $b = \lim_{n\to\infty}(b_n)$.

(i) We aim to show that the number $a + b$ is the limit of the sequence $(a_n + b_n)_n$. Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because a is the limit of (a_n), there exists $N_a \in \mathbb{N}$ for which $|a_n - a| < \epsilon/2$ for all $n \geq N_a$. Similarly, because b is the limit of (b_n), there exists $N_b \in \mathbb{N}$ for which $|b_n - b| < \epsilon/2$ for all $n \geq N_b$. Let N be the larger of N_a and N_b, i.e. $N = \max\{N_a, N_b\}$. In this case, for all $n \geq N$,

$$\begin{aligned} |(a_n + b_n) - (a + b)| &= |a_n - a + b_n - b| \\ &\leq |a_n - a| + |b_n - b| \\ &< \epsilon/2 + \epsilon/2 = \epsilon, \end{aligned}$$

meaning that $(a + b)$ is the limit of $(a_n + b_n)_n$.

(iii) Choose arbitrary $\epsilon \in \mathbb{R}^+$ and let δ be smaller of the numbers 1 and $\epsilon/(1+|a|+|b|)$, i.e.

$$\delta = \min\{1, \epsilon/(1 + |a| + |b|)\};$$

note that $\delta > 0$ by construction. As in the proof of item (i), there exists $N \in \mathbb{N}$ for which $|a_n - a| < \delta$ and $|b_n - b| < \delta$ for all $n \geq N$, from which follows

$$\begin{aligned} |a_n b_n - ab| &= |(a_n - a)(b_n - b) + a(b_n - b) + b(a_n - a)| \\ &\leq |a_n - a||b_n - b| + |a||b_n - b| + |b||a_n - a| \\ &< \delta^2 + |a|\delta + |b|\delta \\ &\overset{\text{(a)}}{\leq} \delta + |a|\delta + |b|\delta \\ &= \delta(1 + |a| + |b|) \leq \epsilon \end{aligned}$$

where (a) holds because $\delta \in (0, 1) \implies \delta^2 < \delta$. The number ab is thus the limit of the sequence $(a_n b_n)_n$.

(ii) Point (ii) follows from point (iii) if for (b_n) we choose the constant sequence with general term $b_n = \alpha$.

(v) Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because a is the limit of (a_n), there exists $N \in \mathbb{N}$ for which $|a_n - a| < \epsilon$ for all natural numbers $n \geq N$. Using the elementary properties of the complex conjugate, for all $n \geq N$ it then holds that

$$|\overline{a_n} - \overline{a}| = \left|\overline{a_n - a}\right| = |a_n - a| < \epsilon,$$

and so $\overline{a}$ is the limit of the conjugated sequence $(\overline{a_n})_n$.

(vi) Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because a is the limit of (a_n), there exists $N \in \mathbb{N}$ for which $|a_n - a| < \epsilon$ for all natural numbers $n \geq N$. For all $n \geq N$ it then holds that

$$||a_n| - |a|| \leq |a_n - a| < \epsilon,$$

and so $|a|$ is the limit of the sequence $(|a_n|)_n$.

(iv) Let $m = \inf(|b_n|)$; because of the absolute value and the assumptions that $\lim_{n\to\infty} b_n \neq 0$ and $b_n \neq 0$ for all $n \in \mathbb{N}$, the infimum m must be a positive number, and by item (vi) above, the positive number $|b|$ is the limit of the sequence of positive numbers $(|b_n|)_n$. There thus exists natural number $l \in \mathbb{N}$ for which

$$|b|/2 < |b_n| \text{ for all } n \geq l,$$

which implies the inequality

$$0 < \min\{|b_1|, |b_2|, \ldots, |b_l|, |b|/2\} \leq m.$$

We then choose arbitrary $\epsilon \in \mathbb{R}^+$ and define $\delta = \epsilon|b|m$. There then exists $N \in \mathbb{N}$ for which $|b_n - b| < \delta$ for all natural numbers $n \geq N$, and so

$$\left|\frac{1}{b} - \frac{1}{b_n}\right| = \left|\frac{b_n - b}{bb_n}\right| < \frac{\delta}{|b||b_n|} \leq \frac{\delta}{|b|m} = \epsilon,$$

which means that $1/b$ is the limit of the sequence $(1/b_n)$.

Since $(1/b_n)$ converges, by point (iii) the product of (a_n) and $(1/b_n)$ also converges and obeys

$$\lim_{n\to\infty}\left(\frac{a_n}{b_n}\right) = \lim_{n\to\infty}\left(a_n\cdot\frac{1}{b_n}\right) \overset{\text{(a)}}{=} \lim_{n\to\infty}(a_n)\cdot\lim_{n\to\infty}\left(\frac{1}{b_n}\right) = \frac{a}{b},$$

where (a) follows from point (iii). □

As a direct implication of the above rules for computing the limits of complex sequences, convergent complex sequences may be interpreted as pairs of convergent real sequences.

Corollary 4.18: Real and complex parts of a convergent sequence

A complex sequence (z_n) converges if, and only if, the sequences $\big(\mathrm{Re}(z_n)\big)_n$ and $\big(\mathrm{Im}(z_n)\big)_n$ are both convergent. In this case

$$\lim_{n\to\infty}(z_n) = \lim_{n\to\infty}\big(\mathrm{Re}(z_n)\big)_n + \mathrm{i}\lim_{n\to\infty}\big(\mathrm{Im}(z_n)\big)_n.$$

Proof. ($\Leftarrow$) If the sequences $\big(\mathrm{Re}(z_n)\big)_n$ and $\big(\mathrm{Im}(z_n)\big)_n$ are both convergent, then $\big(z_n = \mathrm{Re}(z_n) + \mathrm{i}\,\mathrm{Im}(z_n)\big)_n$ also converges by Proposition 4.17 (i) and (ii).

($\Rightarrow$) Assume that the sequence z_n converges. In this case the sequence $(\overline{z_n})$ converges by Proposition 4.17 (v), while Proposition 4.17 (i) and (ii) implies the convergence of

$$\big((z_n+\overline{z_n})/2\big)_n = \big(\mathrm{Re}(z_n)\big)_n \quad\text{and}\quad \big((z_n-\overline{z_n})/2\mathrm{i}\big)_n = \big(\mathrm{Im}(z_n)\big)_n.$$ □

Note that if a sequence (a_n) converges, then by Proposition 4.17 (vi) the sequence $(|a_n|)$ also converges. The inverse statement, i.e. that the convergence of $(|a_n|)$ implies the convergence of (a_n), is not in general true. This inverse statement *is* true, however, in the case of a sequence converging to 0, as stated in the following proposition.

Proposition 4.19: Sequence with a limit of zero

The number 0 is the limit of a complex sequence (a_n) if, and only if, the number 0 is the limit of the sequence of absolute values $(|a_n|)$.

Proof. ($\Rightarrow$) If 0 is the limit of (a_n), then 0 is also the limit of $(|a_n|)$ by Proposition 4.17 (vi).

($\Leftarrow$) Assume 0 is the limit of $(|a_n|)$. Then for all $\epsilon\in\mathbb{R}^+$ there exists $N\in\mathbb{N}$ for which

$$|0-|a_n|| = |a_n| = |0-a_n| < \epsilon \text{ for all } n\in\mathbb{N} \text{ for which } n\geq N,$$

and so 0 is also the limit of the sequence (a_n). □

We will now cover the important geometric sequence in some detail.

Example 4.20: The geometric sequence

Consider the geometric sequence with general term aq^{n-1}, where $a, q \in \mathbb{C}$ are arbitrary complex numbers. The geometric sequence obeys the recursive relation

$$a_{n+1} = a_n q \text{ for all } n \in \mathbb{N}.$$

If the geometric sequence (a_n) converges, then the shifted sequence $(a_{n+1})_n$ also converges, and, via the geometric recursive relation, obeys

$$\lim_{n\to\infty}(a_n) = \lim_{n\to\infty}(a_{n+1}) = \lim_{n\to\infty}(a_n q) = q\lim_{n\to\infty}(a_n).$$

We then rearrange the above equation to get

$$(1-q)\lim_{n\to\infty}(a_n) = 0. \tag{4.1}$$

As a result, any convergent geometric sequence $(aq^{n-1})_n$ with ratio $q \neq 1$ must necessarily converge to 0 (this is the only way for Equation 4.1 to hold).

We now consider the possible values of a and q:

(i) If $a = 0$, the sequence $(aq^{n-1})_n$ is constant, with all elements equal to 0. The sequence thus converges to 0.

(ii) If $a \neq 0$ and $q = 1$, then $(aq^{n-1})_n$ is constant, with all elements equal to a. The sequence thus converges to a.

(iii) If $a \neq 0$, $q \neq 1$, and $|q| \geq 1$, then the sequence $(aq^{n-1})_n$ is divergent. To verify this, first note that

$$\left|aq^{n-1}\right| = |a||q|^{n-1} \geq |a| > 0 \text{ for all } n \in \mathbb{N}.$$

It follows that $\left(\left|aq^{n-1}\right|\right)_n$ does not converge to 0, and so $(aq^{n-1})_n$ also does not converge to 0 by Proposition 4.19. Since, by Equation 4.1, $(aq^{n-1})_n$ cannot converge to any point other than 0, it must be divergent.

(iv) If $a \neq 0$ and $|q| < 1$, then the sequence $(aq^{n-1})_n$ converges to the limit

$$\lim_{n\to\infty}\left(aq^{n-1}\right) = 0.$$

To verify this, first note that

$$|aq^n| = |a||q^n| < |a||q|^{n-1} = \left|aq^{n-1}\right| \text{ for all } n \in \mathbb{N}.$$

The sequence $\left(\left|aq^{n-1}\right|\right)_n$ is thus a strictly decreasing sequence of positive real numbers and must be convergent (since it is bounded below by 0) by Proposition 4.15. Because $\left(\left|aq^{n-1}\right|\right)_n$ is a geometric sequence with $|q| \neq 1$, by Equation 4.1 the limit of the sequence must be 0, and so the limit of the sequence (aq^{n-1}) must also be 0 by Proposition 4.19.

Example 4.21: Computing limits analytically

In this exercise we will analytically compute two limits; the first limit is

$$\begin{aligned}\lim_{n\to\infty}\left(\frac{n}{n+1}-\frac{n+1}{n}\right) &= \lim_{n\to\infty}\left(\frac{1}{1+\frac{1}{n}}-\frac{1+\frac{1}{n}}{1}\right)\\ &= \lim_{n\to\infty}\frac{1}{1+(1/n)}-\lim_{n\to\infty}\left(1+\frac{1}{n}\right)\\ &= \frac{1}{1+\lim\limits_{n\to\infty}(1/n)}-1-\lim_{n\to\infty}\frac{1}{n}\\ &= \frac{1}{1+0}-1-0=0.\end{aligned}$$

And the second limit is

$$\begin{aligned}\lim_{n\to\infty}\frac{n^2+3n-2}{5n^2} &= \lim_{n\to\infty}\frac{1+(3/n)-(2/n^2)}{5}\\ &= \frac{1}{5}\left(1+3\lim_{n\to\infty}\frac{1}{n}-2\lim_{n\to\infty}\frac{1}{n^2}\right)\\ &= \frac{1}{5}\left(1+3\cdot 0-2\left(\lim_{n\to\infty}\frac{1}{n}\right)^2\right)\\ &= \frac{1}{5}\left(1-2\cdot 0^2\right)=\frac{1}{5}.\end{aligned}$$

Remark 4.22: Real sequences diverging without bound

A special example of a divergent sequence of real numbers is a sequence in which the elements grow either arbitrarily positive or arbitrarily negative from some element onward. In such cases, letting (a_n) denote a sequence of real numbers, the notation

$$\lim_{n\to\infty}(a_n)=\infty \tag{4.2}$$

means that for all $M\in\mathbb{R}$ there exists some natural number $N\in\mathbb{N}$ for which $a_n\geq M$ for all indices $n\in\mathbb{N}$ for which $n\geq N$, i.e. that the sequence's elements are unbounded above from some index onward. In this case we may say the sequence *diverges to positive infinity.*

Analogously, the notation

$$\lim_{n\to\infty}(a_n)=-\infty \tag{4.3}$$

means that for all $m\in\mathbb{R}$ there exists some natural number $N\in\mathbb{N}$ for which $a_n\leq m$ for all indices $n\in\mathbb{N}$ for which $n\geq N$. In this case we may say the sequence *diverges to negative infinity.*

Here are two example uses of this notation:

- $\lim\limits_{n\to\infty}(n^3-n+2)=\infty$
- $\lim\limits_{n\to\infty}\left(\frac{1-n^2}{1+n}\right)=-\infty$.

This notation makes it tempting to say that a sequence's limit "equals" infinity or negative infinity, but it is more precise to recall that Equations 4.2 and 4.3 are simply shorthand ways of stating that a sequence is unbounded from some index onward.

We include this section with a detail treatment of Euler's number, which is among the most important transcendental numbers in mathematics.

Example 4.23: Euler's number

Consider the sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ with general terms given by

$$a_n = \left(1 + \frac{1}{n}\right)^n \quad \text{and} \quad b_n = \left(1 - \frac{1}{n}\right)^{-n}.$$

We will show that $(a_n)_{n=1}^{\infty}$ is strictly increasing, that $(b_n)_{n=1}^{\infty}$ is strictly decreasing, and that both sequence are convergent to the same limit, a transcendental number called *Euler's number*, denoted by e, and defined as

$$\mathrm{e} = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^{-n}.$$

We divide the derivation into multiple parts.

(i) We will first use induction show to that

$$(1+x)^k > 1 + kx \tag{4.4}$$

for all $x \in \mathbb{R}$ for which $0 < |x| < 1$ and for all integers $k \in \mathbb{Z}$ for which $|k| \geq 2$.

The base case $k = 2$ evidently holds, since $1+2x+x^2 > 1+2x$ for all $|x| \in (0, 1)$. In the induction step, if Equation 4.4 holds for $k \geq 2$, then it also holds that

$$\begin{aligned}(1+x)^{k+1} &= (1+x)^k(1+x) \overset{\text{(a)}}{>} (1+kx)(1+x)\\ &= 1 + (k+1)x + kx^2\\ &\overset{\text{(b)}}{>} 1 + (k+1)x,\end{aligned}$$

where (a) uses the induction step and (b) holds under the assumption $k \geq 2$. The inequality in Equation 4.4 thus holds for all integers $k \geq 2$ and all real numbers x obeying $0 < |x| < 1$.

Now let k be a negative integer for which $k \leq -2$. For all $x \in \mathbb{R}$ for which $0 < |x| < 1$, it then holds that

$$(1+x)(1-x) = 1 - x^2 < 1 \implies \frac{1}{1+x} > 1 - x,$$

from which follows that

$$\begin{aligned}(1+x)^k &= \left(\frac{1}{1+x}\right)^{-k} > (1-x)^{-k}\\ &\overset{\text{(a)}}{=} 1 + (-k)(-x) = 1 + kx.\end{aligned}$$

where (a) follows from Equation 4.4, which we have already proven for $-k \geq 2$. We have thus shown that Equation 4.4, holds for all integers k for which $k \leq -2$ and all x for which $0 < |x| < 1$.

(ii) From the inequality in Equation 4.4, it follows that

$$\left(1+\frac{1}{k}\right)^k\left(1-\frac{1}{k}\right)^k = \left(1-\frac{1}{k^2}\right)^k \overset{(a)}{>} 1+k\left(-\frac{1}{k^2}\right) = 1-\frac{1}{k}.$$

where (a) follows from Equation 4.4. We then divide the above result through by $\left[1-(1/k)\right]^k$ to get the inequality

$$\left(1+\frac{1}{k}\right)^k > \left(1-\frac{1}{k}\right)^{1-k}, \tag{4.5}$$

which holds for all $k \in \mathbb{Z}$ for which $|k| \geq 2$.

(iii) By the inequality in Equation 4.5, for all natural numbers $n \geq 2$ it holds that

$$\begin{aligned} a_n &= \left(1+\frac{1}{n}\right)^n \overset{(a)}{>} \left(1-\frac{1}{n}\right)^{1-n} \\ &= \left(\frac{n-1}{n}\right)^{1-n} = \left(\frac{n}{n-1}\right)^{n-1} \\ &= \left(1+\frac{1}{n-1}\right)^{n-1} = a_{n-1}, \end{aligned}$$

where (a) follows from Equation 4.5. The result $a_n > a_{n-1}$ means that the sequence $(a_n)_{n=1}^{\infty}$ is strictly increasing.

(iv) Similarly, again using Equation 4.5, for all natural numbers $n \geq 2$ it holds that

$$\begin{aligned} b_n &= \left(1-\frac{1}{n}\right)^{-n} = \left(1+\frac{1}{(-n)}\right)^{-n} \\ &\overset{(a)}{>} \left(1-\frac{1}{(-n)}\right)^{1-(-n)} = \left(1+\frac{1}{n}\right)^{1+n} \\ &= \left(\frac{n+1}{n}\right)^{1+n} = \left(\frac{n}{n+1}\right)^{-(1+n)} \\ &= \left(1-\frac{1}{n+1}\right)^{-(n+1)} = b_{n+1}. \end{aligned}$$

where (a) follows from Equation 4.5. The result $b_n > b_{n+1}$ means that the sequence $(b_n)_{n=2}^{\infty}$ is strictly decreasing.

(v) For each natural number $n \in \mathbb{N}$, the sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=2}^{\infty}$ are related according to

$$\begin{aligned} b_{n+1} &= \left(1-\frac{1}{n+1}\right)^{-(n+1)} = \left(\frac{n}{n+1}\right)^{-(n+1)} \\ &= \left(\frac{n+1}{n}\right)^{n+1} = \left(1+\frac{1}{n}\right)^{n+1} \\ &= \left(1+\frac{1}{n}\right) a_n, \end{aligned} \tag{4.6}$$

and so $a_n < b_{n+1}$. From this, and from points (iii) and (iv)

$$2 = a_1 < a_2 < \ldots < a_n < b_{n+1} < b_n < \ldots < b_2 = 4 \text{ for all } n \in \mathbb{N}.$$

The sequence $(a_n)_{n=1}^{\infty}$ is strictly increasing and bounded, and thus convergent. Similarly, the sequence $(b_n)_{n=2}^{\infty}$ is strictly decreasing and bounded, and thus also convergent. We complete the derivation with the calculation

$$\begin{aligned}\lim_{n\to\infty}(b_n) = \lim_{n\to\infty}(b_{n+1}) &\overset{\text{(a)}}{=} \lim_{n\to\infty}\left[\left(1+\frac{1}{n}\right)a_n\right] \\ &= \lim_{n\to\infty}\left(1+\frac{1}{n}\right)\cdot\lim_{n\to\infty}(a_n) = \lim_{n\to\infty}(a_n),\end{aligned}$$

where (a) follows from Equation 4.6.

The sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=2}^{\infty}$ thus converge to the same limit, called *Euler's number* and given by

$$\mathrm{e} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{-n}.$$

Euler's number e is irrational, and its approximation to 12 decimal points is

$$\mathrm{e} \approx 2.718281828459.$$

4.4 Subsequences

A subsequence of a given sequence $(a_n)_{n\in\mathbb{N}}$ results from keeping infinitely many elements of $(a_n)_{n\in\mathbb{N}}$, but removing all other elements. We define this more precisely as follows:

Definition 4.24: Subsequence of a complex sequence

Let $(a_n)_{n\in\mathbb{N}}$ be an arbitrary complex sequence and let $(n_k)_{k\in\mathbb{N}}$ be an arbitrary strictly increasing sequence of natural numbers. The *subsequence* of $(a_n)_{n\in\mathbb{N}}$ determined by the index sequence $(n_k)_{k\in\mathbb{N}}$ is the sequence $(a_{n_k})_{k\in\mathbb{N}}$. Different subsequences of $(a_n)_{n\in\mathbb{N}}$ are determined by different indexing sequences $(n_k)_{k\in\mathbb{N}}$.

Example 4.25: Subsequences

Let (a_n) be the real-valued sequence with general term $a_n = 1/n$, i.e.

$$(a_n) = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots\right).$$

The subsequence of (a_n) determined by the strictly increasing sequence of natural numbers $(2k)_{k\in\mathbb{N}}$ is the sequence

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \ldots\right).$$

Meanwhile, the subsequence of (a_n) determined by the strictly increasing sequence of natural numbers $(2^{k-1})_{k\in\mathbb{N}}$ is the sequence

$$\left(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16} \dots\right).$$

Finally, the sequence $\left(\frac{1}{2}, 1, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots\right)$ formed by rearranging the first two terms of (a_n) is not a subsequence of (a_n), since the indexing sequence used to generate it is not strictly increasing.

If (a_n) be a sequence of complex numbers, then $(a_{2k})_{k\in\mathbb{N}}$ is the subsequence of (a_n) containing all elements of (a_n) with even indices, while $(a_{2k-1})_{k\in\mathbb{N}}$ is the subsequence of (a_n) containing all elements of (a_n) with odd indices. From the definition of the limit of a sequence it follows that $a \in \mathbb{C}$ is the limit of (a_n) if, and only if, a is the limit of both $(a_{2k})_{k\in\mathbb{N}}$ and $(a_{2k-1})_{k\in\mathbb{N}}$. We state this idea more formally in the following proposition.

Proposition 4.26: Limits and boundedness of subsequences

Let (a_n) be a sequence of complex numbers.

(i) If $a \in \mathbb{C}$ is a cluster point of any subsequence of (a_n), then a is also a cluster point of (a_n).

(ii) If $a \in \mathbb{C}$ is the limit of the sequence (a_n), then a is also the limit of all of (a_n)'s subsequences. In particular, this means that every subsequence of a convergent sequence is itself convergent.

(iii) If the sequence (a_n) is bounded, then every subsequence of (a_n) is also bounded.

Proof. The proposition follows directly from the definition of a sequence's cluster point, limit, and boundedness. □

Proposition 4.27: Cluster points and subsequences

Let (a_n) be a sequence of complex numbers. The number $b \in \mathbb{C}$ is a cluster point of (a_n) if, and only if, there exists a subsequence of (a_n) that converges to b.

Proof. ($\Leftarrow$) If there exists a subsequence of (a_n) that converges to b, then every neighborhood of b contains all (specifically, infinitely many) elements of the subsequence (with the possible exception of finitely many, which is irrelevant in this context). The original sequence (a_n) includes all elements of the subsequence, and so every neighborhood of b must also contain infinitely many elements of (a_n), meaning that b is a cluster point of a_n.

($\Rightarrow$) Let b be a cluster point of the sequence (a_n). We will recursively construct a strictly increasing sequence of natural numbers $(n_k)_{k\in\mathbb{N}}$ for which $a_{n_k} \in \mathcal{K}(b, 1/k)$ for all $k \in \mathbb{N}$, meaning that the subsequence $(a_{n_k})_k$ converges to b.

To begin, we choose any natural number $n_1 \in \mathbb{N}$ for which

$$a_{n_1} \in \mathcal{K}(b, 1).$$

Such a number n_1 certainly exists, since, because b is a cluster point of (a_n), every open disk $\mathcal{K}(b,1)$ centered on b contains infinitely many elements of (a_n). Now assume we have already found indices $n_1, n_2, \ldots, n_k$ with the desired property that

$$a_{n_j} \in \mathcal{K}(b, 1/j) \text{ for } j = 1, 2, \ldots, k.$$

Because the open disk $\mathcal{K}(b, 1/(k+1))$ contains infinitely many elements of (a_n), there certainly exists some $n_{k+1} \in \mathbb{N}$ for which

$$n_{k+1} > n_k \quad \text{and} \quad a_{n_{k+1}} \in \mathcal{K}\big(b, 1/(k+1)\big).$$

In this way we have construct the desired subsequence $(a_{n_k})_{k\in\mathbb{N}}$ of the sequence (a_n).

Now choose arbitrary $\epsilon \in \mathbb{R}^+$, and let $l \in \mathbb{N}$ be a natural number for which $1/l < \epsilon$. For all natural numbers $k \geq l$ it holds that

$$a_{n_k} \in \mathcal{K}(b, 1/k) \subset \mathcal{K}(b, 1/l) \subset \mathcal{K}(b, \epsilon),$$

which, by Proposition 4.12, means that the subsequence $(a_{n_k})_{k\in\mathbb{N}}$ converges to b. □

We are now prepared to extend the Bolzano-Weierstrass theorem (Theorem 4.9) to complex-valued sequences.

Theorem 4.28: Bolzano-Weierstrass theorem for complex sequences

Every bounded sequence of complex numbers has at least one cluster point.

Proof. Let (z_n) be a bounded sequence of complex numbers. For shorthand, let

$$a_n = \mathrm{Re}(z_n) \quad \text{and} \quad b_n = \mathrm{Im}(z_n) \text{ for all } n \in \mathbb{N}.$$

Since (z_n) is bounded, the real sequences (a_n) and (b_n) are also bounded, since $|a_n| \leq |z_n|$ and $|b_n| \leq |z_n|$ for all $n \in \mathbb{N}$. We will now prove the theorem with two successive applications of the Bolzano-Weierstrass theorem for real sequences (Theorem 4.9).

By Theorem 4.9, the bounded real sequences (a_n) has at least one cluster point, which we denote by a. By Proposition 4.27, there then exists a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ of the sequence (a_n) that converges to a.

By Proposition 4.26 (iii), the subsequence $(b_{n_k})_{k\in\mathbb{N}}$ of the bounded sequence (b_n) is itself bounded, so we can apply Theorem 4.9 to the *subsequence* $(b_{n_k})_{k\in\mathbb{N}}$ to conclude that $(b_{n_k})_{k\in\mathbb{N}}$ has at least one cluster point, which we denote by b. By Proposition 4.27, there then exists a subsequence $(b_{n_{k_l}})_{l\in\mathbb{N}}$ of the subsequence $(b_{n_k})_{k\in\mathbb{N}}$ that converges to b. In other words, there exists some subsequence $(n_{k_l})_{l\in\mathbb{N}}$ of the strictly increasing sequence of natural numbers $(n_k)_{k\in\mathbb{N}}$ for which the subsequence $(b_{n_{k_l}})_{l\in\mathbb{N}}$ of the subsequence $(b_{n_k})_{k\in\mathbb{N}}$ converges to b.

We then note that, because $(a_{n_{k_l}})_{l\in\mathbb{N}}$ is a subsequence of $(a_{n_k})_{k\in\mathbb{N}}$ and because $(a_{n_k})_{k\in\mathbb{N}}$ converges to a, the subsequence $(a_{n_{k_l}})_{l\in\mathbb{N}}$ must also converges to a by Prop 4.26 (ii). We have thus found a strictly increasing sequence of natural numbers

$(n_{k_l})_{l\in\mathbb{N}}$ for which

$$\lim_{n\to\infty}(a_{n_{k_l}}) = a \quad \text{and} \quad \lim_{n\to\infty}(b_{n_{k_l}}) = b,$$

and so by the limit computation rules in Proposition 4.17, the subsequence

$$\left(z_{n_{k_l}} = a_{n_{k_l}} + \mathrm{i}b_{n_{k_l}}\right)_{l\in\mathbb{N}}$$

of the sequence (z_n) converges to $a + \mathrm{i}b$. By Prop 4.27, $a + \mathrm{i}b$ is then a cluster point of (z_n), and because the complex sequence (z_n) was arbitrary, it follows that every bounded complex sequence has at least one cluster point. □

Proposition 4.29: Relating convergence and boundedness

(i) Every convergent sequence of complex numbers is bounded.

(ii) Every bounded sequence of complex numbers with exactly one cluster point converges to the cluster point.

Proof.

(i) Let (a_n) be a complex sequence that converges to the point $a \in \mathbb{C}$, meaning there exists some $N \in \mathbb{N}$ for which

$$a_n \in \mathcal{K}(a, 1) \text{ for all natural numbers } n \geq N.$$

Using the triangle inequality for complex numbers, the open disk $\mathcal{K}(a, 1)$ is a subset of the disk $\mathcal{K}(0, |a| + 1)$, and so $a_n \in \mathcal{K}(0, |a| + 1)$ for all $n \geq N$. We must also consider elements of (a_n) for $n < N$; to this end, define

$$M = \max\{|a_1|, |a_2|, \ldots, |a_N|, |a| + 1\}.$$

In this case all elements of (a_n) lie in the closed disk $\overline{\mathcal{K}}(0, M)$, which means that (a_n) is bounded by M.

(ii) Let (a_n) be a bounded complex sequence with exactly one cluster point $a \in \mathbb{C}$. We will show, by contradiction, that a is the limit of the sequence (a_n).

Assume that a is not the limit of (a_n). In this case there would exist some $\epsilon \in \mathbb{R}^+$ for which infinitely many elements of (a_n) would lie outside the open disk $\mathcal{K}(a, \epsilon)$. There would then exist a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ of (a_n) for which all elements of the subsequence would lie outside the disk $\mathcal{K}(a, \epsilon)$, which, by Proposition 4.27, would imply that a is not a cluster point of the subsequence $(a_{n_k})_{k\in\mathbb{N}}$. But because all cluster points of the subsequence $(a_{n_k})_{k\in\mathbb{N}}$ are necessarily also cluster points of (a_n) by Proposition 4.26, and because a is assumed to be the only cluster point of (a_n), we would conclude that the subsequence $(a_{n_k})_{k\in\mathbb{N}}$ does not have any cluster points. The bounded subsequence $(a_{n_k})_{k\in\mathbb{N}}$ not having cluster points contradicts the Bolzano-Weierstrass theorem, and so we conclude that a must the limit of the sequence (a_n). □

4.5 The Cauchy criterion for convergence of sequences

Definition 4.30: Cauchy sequences

A sequence of complex numbers (a_n) is a *Cauchy sequence* if for all $\epsilon \in \mathbb{R}^+$ there exists some natural number $N \in \mathbb{N}$ for which

$$|a_m - a_n| < \epsilon \text{ for any two natural numbers } m, n \geq N.$$

In other words, a sequence is a Cauchy sequence if all elements of the sequence grow arbitrarily close to each other from some element onward.

Proposition 4.31: Convergence of Cauchy sequences

A sequence of complex numbers converges if, and only if, it is a Cauchy sequence.

Proof. ($\Rightarrow$) Let (a_n) be a complex sequence that converges to $a \in \mathbb{C}$, and so for all $\epsilon \in \mathbb{R}^+$ there exists some $N \in \mathbb{N}$ for which

$$|a_n - a| < \epsilon/2 \text{ for all natural numbers } n \geq N.$$

For any two natural numbers $m, n \geq \mathbb{N}$ it then holds that

$$|a_m - a_n| = |a_m - a + a - a_n| \leq |a_m - a| + |a - a_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

and so (a_n) is a Cauchy sequence.

($\Leftarrow$) Assume (a_n) is a Cauchy sequence of complex numbers—we must show then show that (a_n) converges.

We will first show that (a_n) is bounded. Because (a_n) is a Cauchy sequence, there exists natural number $l \in \mathbb{N}$ for which

$$|a_m - a_n| < 1 \text{ for all natural numbers } n, m \geq l,$$

which implies that

$$a_n \in \mathcal{K}(a_l, 1) \text{ for all natural numbers } n \geq l.$$

Similarly to the proof of Proposition 4.29, we note that the open disk $\mathcal{K}(a, 1)$ is a subset of the open disk $\mathcal{K}(0, |a| + 1)$ as a consequence of the triangle inequality, and also define

$$M = \max\{|a_1|, |a_2|, \ldots, |a_N|, |a| + 1\}.$$

In this case all elements of (a_n) lie in the closed disk $\bar{\mathcal{K}}(0, M)$, which means that (a_n) is bounded. Because (a_n) is bounded, it has at least one cluster point by the Bolzano-Weierstrass theorem (Theorem 4.28).

We will now show, by contradiction, that (a_n) has exactly one cluster point, and is thus convergent by Proposition 4.29 (ii). Assume (a_n) has multiple cluster points, let a and b denote two of (a_n)'s cluster points, and define

$$\epsilon = \frac{|b - a|}{3} > 0.$$

Because (a_n) is a Cauchy sequence, there exists $N \in \mathbb{N}$ for which

$$|a_m - a_n| < \epsilon \text{ for all natural numbers } m, n \geq N.$$

However, because a and b are cluster points of (a_n), there exists a natural number $n \geq N$ for which $|a_n - a| < \epsilon$ and a natural number $m \geq N$ for which $|a_m - b| < \epsilon$. This implies that

$$\begin{aligned} |b - a| &= |b - a_m + a_m - a_n + a_n - a| \\ &\leq |b - a_m| + |a_m - a_n| + |a_n - a| \\ &< \epsilon + \epsilon + \epsilon = 3\epsilon \\ &\overset{\text{(a)}}{=} |b - a|, \end{aligned}$$

where (a) follows from the definition of ϵ. The result $|b - a| < |b - a|$ is impossible, and so the assumption that (a_n) has multiple cluster points is incorrect. Thus (a_n) has exactly one cluster point and thus converges by Proposition 4.29 (ii). □

5 Series of numbers

5.1 The sum of a series of numbers

Definition 5.1: Series, series elements, and partial sums

A *series of complex numbers* is a function of the form $\mathbb{N} \to \mathbb{C}$ mapping $n \mapsto a_n$ and is usually written using one of the following notations:

$$\sum (a_n)_{n=1}^{\infty} = \sum (a_n)_{n\in\mathbb{N}}.$$

For all $n \in \mathbb{N}$, the number a_n is called the *n-th element* of the series $\sum (a_n)_{n=1}^{\infty}$, and a arbitrary series of numbers is fully specified by the sequence of its elements, $(a_n)_{n\in\mathbb{N}}$.

For all series $\sum (a_n)_{n\in\mathbb{N}}$ and all natural numbers $k \in \mathbb{N}$, the *k-th partial sum* of the series $\sum (a_n)_{n\in\mathbb{N}}$ is the complex number

$$\sum_{n=1}^{k} a_n = a_1 + a_2 + \cdots + a_k \in \mathbb{C}.$$

Finally, the *sequence of partial sums* associated with a complex series $\sum (a_n)_{n\in\mathbb{N}}$ is the complex sequence

$$\left(\sum_{n=1}^{k} a_n \right)_{k\in\mathbb{N}}.$$

Definition 5.2: Convergent and divergent series

A series of complex numbers $\sum (a_n)_{n\in\mathbb{N}}$ said to *converge* if its associated sequence of partial sums $\left(\sum_{n=1}^{k} a_n \right)_{k\in\mathbb{N}}$ converges. If a series converges, the limit of the series's sequence of partial sums is called the series's *sum* and is denoted by

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + \cdots = \lim_{k\to\infty} \left(\sum_{n=1}^{k} a_n \right) \in \mathbb{C}.$$

A series of complex numbers that does not converge is said to *diverge*.

Note the subtle but important difference in notation between a series and its sum:

- We will write a series, e.g. $\sum (a_n)_{n=1}^{\infty}$, by writing summation symbol to the left of the sequence of series elements $(a_n)_{n=1}^{\infty}$ specifying the series; a series will always appear in this book without subscripts or superscripts on the summation symbol.
- We will write the sum of a series, e.g. $\sum_{n=1}^{\infty} a_n$, with subscripts (and possibly superscripts) on the summation symbol.

There is one more important subtlety: a sum is well-defined only for a *convergent* series. However, in the wider mathematical and scientific literature (but not in this book) one regularly sees the series itself (and not just its sum) denoted by $\sum_{n=1}^{\infty} a_n$, even in cases when the series itself diverges. In such cases, it must be determined from context if the notation $\sum_{n=1}^{\infty} a_n$ refers to a series of numbers or just the series's sum.

Remark 5.3: Notation for divergent real series

If the sequence of partial sums of a series of real number $\sum(a_n)_{n=1}^{\infty}$ diverges to positive infinity (in the sense of Remark 4.22), i.e. if

$$\lim_{k\to\infty}\left(\sum_{n=1}^{k} a_n\right) = \infty,$$

we abbreviate this with the shorthand notation

$$\sum_{n=1}^{\infty} = \infty.$$

Similarly, if the series's sequence of partial sums diverges to negative infinity, we may write

$$\underbrace{\lim_{k\to\infty}\left(\sum_{n=1}^{k} a_n\right) = -\infty}_{\text{full notation}} \quad \Longleftrightarrow \quad \underbrace{\sum_{n=1}^{\infty} = -\infty}_{\text{shorthand notation}} .$$

Example 5.4: The geometric series

Let a and q be two arbitrary complex numbers. The *geometric series* with initial element a and quotient q is the series

$$\sum(aq^{n-1})_{n=1}^{\infty}.$$

For all $k \in \mathbb{N}$, the k-th partial sum of the geometric series is

$$\begin{aligned}\sum_{n=1}^{k} aq^{n-1} &= a + aq + aq^2 + \cdots + aq^{k-1} \\ &= a(1 + q + q^2 + \cdots + q^{k-1}),\end{aligned}$$

and if $q \neq 1$, the series's k-th partial sum may also be written in the form

$$\sum_{n=1}^{k} aq^{n-1} = \frac{a(1-q^k)}{1-q}.$$

The geometric series's convergence depends on the values of a and q:

(i) If $a = 0$, then all elements of the geometric series are zero, and so all of the series's partial sums are also zero. In this case the geometric series converges, and its sum is zero.

(ii) If $a \neq 0$ and $q = 1$, then all elements of the geometric series are a, and so the series's k-th partial sum is ka for all $k \in \mathbb{N}$. In this case the sequence of partial sums diverges, and so the series diverges.

(iii) If $a \neq 0$, $q \neq 1$, and $|q| \geq 1$, we first note that the sequence $(q^k)_k$ diverges, and so in this case the sequence of partial sums

$$\left(\frac{a(1-q^k)}{1-q}\right)_k,$$

and thus the geometric series, also diverge.

(iv) If $|q| < 1$, the geometric series converges to the sum

$$\sum_{n=1}^{\infty} aq^{n-1} = \lim_{k\to\infty} \sum_{n=1}^{k} aq^{n-1} = \lim_{k\to\infty} \frac{a(1-q^k)}{1-q} = \frac{a}{1-q}.$$

Proposition 5.5: Cauchy criterion for convergence of series

A series of complex numbers $\sum(a_n)_{n=1}^{\infty}$ converges if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists a natural number $N \in \mathbb{N}$ for which

$$|a_n + a_{n+1} + \cdots + a_m| < \epsilon$$

for all natural numbers $m, n \in \mathbb{N}$ for which $m \geq n \geq N$.

Loosely, the Cauchy criterion states that a series converges when any contiguous, finite-length sum of the series elements is arbitrarily small from some index onwards.

Proof. Let $(S_k)_{k\in\mathbb{N}}$ be the sequence of partial sums of the series $\sum(a_n)_{n=1}^{\infty}$. By the definition of a series's convergence, the series converges if, and only if, the sequence (S_k) converges, which (by Proposition 4.31) occurs if, and only if, (S_k) is a Cauchy sequence.
For review from Definition 4.30, the sequence (S_k) is a Cauchy sequence if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists $N' \in \mathbb{N}$ for which

$$|S_k - S_l| < \epsilon \text{ for all natural numbers } k, l \geq N'.$$

We assume without loss of generality that $k > l$, and then introduce the new notation

$$m = k, \quad n = l + 1, \quad N = N' + 1.$$

In terms of this notation, (S_k) is a Cauchy sequence if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists a natural number $N \geq 2$ for which

$$|S_m - S_{n-1}| < \epsilon \text{ for all natural numbers } m \geq n \geq N.$$

By the definition of a partial sum, it holds that

$$S_m - S_{n-1} = a_n + a_{n+1} + \cdots + a_m,$$

and so the Cauchy criterion for the convergence of series in this proposition is equivalent to the sequence of partial sums (S_k) being a Cauchy sequence and thus being convergent, proving the proposition. □

Corollary 5.6: Convergence of a series's elements

If a series of complex numbers $\sum(a_n)_{n=1}^{\infty}$ converges, then the sequence of its elements must converge to zero, i.e.

$$\lim_{n\to\infty} a_n = 0.$$

Proof. Choose arbitrary $\epsilon \in \mathbb{R}^+$. Since the series $\sum(a_n)_{n=1}^{\infty}$ converges, by the Cauchy criterion for series convergence there exists $N \in \mathbb{N}$ for which

$$|a_n + a_{n+1} + \cdots + a_m| < \epsilon \text{ for all natural numbers } m \geq n \geq N.$$

In particular, we may choose $m = n$, and write

$$|a_n + \cdots + a_m| = |a_n| = |a_n - 0| < \epsilon \text{ for all } n \geq N,$$

and so the sequence of elements (a_n) converges to zero by Proposition 4.12. □

The sequence of a series's elements converging to zero is thus a necessary condition for the series itself to converge. It is not a sufficient condition, however—a series whose sequence of elements converges to zero may still diverge, as illustrated by the following example.

Example 5.7: The harmonic series diverges

The *harmonic series* is the series

$$\sum \left(\frac{1}{n}\right)_{n=1}^{\infty}.$$

The series's elements converge to zero, i.e. $\lim_{n\to\infty} 1/n = 0$, but the series itself diverges. To show that the harmonic series diverges, we first define the number c_k as

$$c_k = \frac{1}{2^k + 1} + \frac{1}{2^k + 2} + \cdots + \frac{1}{2^{k+1}} \in \mathbb{R} \text{ for all } k \in \mathbb{N}.$$

The number c_k is the sum of 2^k successive elements of the harmonic series. Each of these elements can be bounded above by $1/(2^{k+1})$, from which follows the bound

$$c_k > \frac{1}{2^{k+1}} + \frac{1}{2^{k+1}} + \cdots + \frac{1}{2^{k+1}} = 2^k \frac{1}{2^{k+1}} = \frac{1}{2}.$$

Because for all $N \in \mathbb{N}$ there exists sufficiently large $k \in \mathbb{N}$ such that $2^{k+1} \geq N$, and because c_k, the sum of the series's large-index elements, does not grow arbitrarily small for any k (but instead obeys $c_k > 1/2$), the harmonic series diverges by the Cauchy criterion for series convergence.

Definition 5.8: Absolute and conditional convergence

A series of complex numbers $\sum(a_n)_{n=1}^{\infty}$ said to *converge absolutely* if its associated series of absolute value elements $\sum(|a_n|)_{n=1}^{\infty}$ converges. A series is said to *converge conditionally* if it converges but is not absolutely convergent.

Proposition 5.9: Absolutely convergence implies convergence

If the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely, it converges and obeys

$$\left|\sum_{n=1}^{\infty} a_n\right| \leq \sum_{n=1}^{\infty} |a_n|.$$

Proof. Let $\sum(a_n)_{n=1}^{\infty}$ be an absolutely convergent series of complex numbers and choose arbitrary $\epsilon \in \mathbb{R}^+$. By the Cauchy criterion for series convergence there exists $N \in \mathbb{N}$ for which

$$\big||a_n| + |a_{n+1}| + \cdots + |a_m|\big| \overset{\text{(a)}}{=} |a_n| + |a_{n+1}| + \cdots + |a_m| < \epsilon$$

for all natural numbers $m, n \in \mathbb{N}$ for which $m \geq n \geq N$, where (a) holds because each absolute value term is nonnegative. We then apply the triangle inequality to get

$$|a_n + a_{n+1} + \cdots + a_m| \leq |a_n| + |a_{n+1}| + \cdots + |a_m| < \epsilon,$$

and so the series $\sum(a_n)_{n=1}^{\infty}$ converges by the Cauchy criterion.

To prove proposition's inequality, for all $k \in \mathbb{N}$ let S_k denote the k-th partial sum of the series $\sum(a_n)_{n=1}^{\infty}$ and let T_k denote the k-th partial sum of the series $\sum(|a_n|)_{n=1}^{\infty}$. By the triangle inequality and the definition of a partial sum, it holds that

$$|S_k| \leq T_k \text{ for all } k \in \mathbb{N},$$

which in turn implies that

$$\left|\sum_{n=1}^{\infty} a_n\right| = \left|\lim_{k\to\infty} S_k\right| = \lim_{k\to\infty} |S_k| \leq \lim_{k\to\infty} T_k = \sum_{n=1}^{\infty} |a_n|. \qquad \square$$

Proposition 5.10: Properties of convergent series

Let $\sum(a_n)_{n=1}^{\infty}$ and $\sum(b_n)_{n=1}^{\infty}$ be two convergent series of complex numbers and let $\alpha \in \mathbb{C}$ be an arbitrary scalar. In this case:

(i) The series $\sum(a_n + b_n)_{n=1}^{\infty}$ converges to the sum

$$\sum_{n=1}^{\infty}(a_n + b_n) = \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n.$$

(ii) The scaled series $\sum(\alpha a_n)_{n=1}^{\infty}$ converges to the sum

$$\sum_{n=1}^{\infty}(\alpha a_n) = \alpha \sum_{n=1}^{\infty} a_n.$$

(iii) The complex-conjugated series $\sum(\overline{a_n})_{n=1}^{\infty}$ converges to the sum

$$\sum_{n=1}^{\infty}(\overline{a_n}) = \overline{\sum_{n=1}^{\infty} a_n}.$$

(iv) If $\sum(a_n)_{n=1}^{\infty}$ and $\sum(b_n)_{n=1}^{\infty}$ both converge absolutely, then the three series $\sum(a_n + b_n)_{n=1}^{\infty}$, $\sum(\alpha a_n)_{n=1}^{\infty}$, and $\sum(\overline{a_n})_{n=1}^{\infty}$ converge absolutely as well.

Proof. To prove points (i)–(iii), let (S_k) and (R_k) denote the sequences of partial sums of the series $\sum(a_n)_{n=1}^{\infty}$ and $\sum(b_n)_{n=1}^{\infty}$, respectively. In this case $(S_k + R_k)$ is

the sequences of partial sums of the series $\sum(a_n + b_n)_{n=1}^{\infty}$, (αS_k) is the sequences of partial sums of the series $\sum(\alpha a_n)_{n=1}^{\infty}$, and $(\overline{S_k})$ is the sequences of partial sums of the series $\sum(\overline{a_n})_{n=1}^{\infty}$.

Because the sequences (S_k) and (R_k) are assumed to be convergent, the sequences $(S_k + R_k)$, (αS_k) and $(\overline{S_k})$ are also convergent by Proposition 4.17, and their limits are

$$\begin{aligned}
\sum_{n=1}^{\infty}(a_n + b_n) &= \lim_{k\to\infty}(S_k + R_k) = \lim_{k\to\infty} S_k + \lim_{k\to\infty} R_k \\
&= \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n, \\
\sum_{n=1}^{\infty} \alpha a_n &= \lim_{k\to\infty} \alpha S_k = \alpha \lim_{k\to\infty} S_k = \alpha \sum_{n=1}^{\infty} a_n, \\
\sum_{n=1}^{\infty} \overline{a_n} &= \lim_{k\to\infty} \overline{S_k} = \overline{\lim_{k\to\infty} S_k} = \overline{\sum_{n=1}^{\infty} a_n}.
\end{aligned}$$

To prove point (iv), assume the series $\sum(a_n)_{n=1}^{\infty}$ and $\sum(b_n)_{n=1}^{\infty}$ both converge absolutely. If $\alpha = 0$, the series $\sum(\alpha a_n)_{n=1}^{\infty}$ evidently also converges absolutely to the sum 0, so we will assume that $\alpha \neq 0$. Choose arbitrary $\epsilon \in \mathbb{R}^+$ and define

$$\delta = \min\left\{\frac{\epsilon}{2}, \frac{\epsilon}{|\alpha|}\right\}.$$

Because $\sum(a_n)_{n=1}^{\infty}$ and $\sum(b_n)_{n=1}^{\infty}$ both converge, by the Cauchy criterion for series convergence there exists $N \in \mathbb{N}$ for which

$$|a_n + a_{n+1} + \cdots + a_m| \leq |a_n| + |a_{n+1}| + \cdots + |a_m| < \delta \tag{5.1a}$$
$$|b_n + b_{n+1} + \cdots + b_m| \leq |b_n| + |b_{n+1}| + \cdots + |b_m| < \delta \tag{5.1b}$$

for all $m, n \in \mathbb{N}$ for which $m \geq n \geq N$. Equations 5.1 in turn imply:

(i) The absolute convergence of the series $\sum(a_n + b_n)_{n=1}^{\infty}$ from

$$\begin{aligned}
&|a_n + b_n| + |a_{n+1} + b_{n+1}| + \cdots + |a_m + b_m| \\
&\qquad \leq |a_n| + |b_n| + |a_{n+1}| + |b_{n+1}| + \cdots + |a_m| + |b_m| \\
&\qquad \overset{(a)}{<} \delta + \delta \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where (a) follows from Equation 5.1.

(ii) The absolute convergence of the series $\sum(\alpha a_n)_{n=1}^{\infty}$ from

$$|\alpha a_n| + |\alpha a_{n+1}| + \cdots + |\alpha a_m| = |a|(|a_n| + |a_{n+1}| + \cdots + |a_m|) < |\alpha|\delta \leq \epsilon.$$

(iii) The absolute convergence of the series $\sum(\overline{a_n})_{n=1}^{\infty}$ from

$$|\overline{a_n}| + |\overline{a_{n+1}}| + \cdots + |\overline{a_m}| = |a_n| + |a_{n+1}| + \cdots + |a_m| < \delta < \epsilon. \qquad \square$$

Proposition 5.11: Summation of absolutely convergent series

Let $\sum(a_n)_{n=1}^{\infty}$ be an absolutely convergent series of complex numbers and let the function $\sigma : \mathbb{N} \to \mathbb{N}$ be any bijection between the natural numbers. In this case the rearrange series $\sum(a_{\sigma(n)})_{n=1}^{\infty}$ also converges absolutely and obeys

$$\sum_{n=1}^{\infty} a_{\sigma(n)} = \sum_{n=1}^{\infty} a_n.$$

In other words, the terms in an absolutely convergent series can be summed in any order without changing the result.

Proof. For all $k \in \mathbb{N}$ define the following partial sums:

$$\begin{aligned} S_k &= a_1 + a_2 + \cdots + a_k, \\ S_k^{\sigma} &= a_{\sigma(1)} + a_{\sigma(2)} + \cdots + a_{\sigma(k)}, \\ R_k &= |a_1| + |a_2| + \cdots + |a_k|, \\ R_k^{\sigma} &= \left|a_{\sigma(1)}\right| + \left|a_{\sigma(2)}\right| + \cdots + \left|a_{\sigma(k)}\right|. \end{aligned}$$

Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the series $\sum(a_n)_{n=1}^{\infty}$ converges, by the Cauchy criterion for series convergence there exists $N \in \mathbb{N}$ for which

$$|a_n| + |a_{n+1}| + \cdots + |a_m| < \epsilon \text{ for all } m, n \in \mathbb{N} \text{ for which } m \geq n \geq N.$$

Because the set $\{1, 2, \ldots, N\}$ is finite, there exists a natural number $p > N$ for which

$$\{1, 2, \ldots, N\} \subset \{\sigma(1), \sigma(2), \ldots, \sigma(p)\}.$$

By the definitions of S_k and S_k^{σ}, for all natural numbers $k \geq p$ it holds that

$$|S_k - S_k^{\sigma}| = \left|(a_1 + \cdots + a_N + \cdots + a_k) - (a_{\sigma(1)} + \cdots + a_{\sigma(k)})\right| \tag{5.2}$$

For all $k \geq p$, the set $\{\sigma(1), \ldots, \sigma(k)\}$ contains all indices in the set $\{1, \ldots, N\}$, so the terms $a_1, \ldots, a_N$ cancel from the right-hand side of Equation 5.2. The remaining terms in the right-hand side of Equation 5.2 have indices greater than N and less than or equal to the number $M = \max\{\sigma(1), \sigma(2), \ldots, \sigma(k)\}$. By the triangle inequality, we can then make the bound

$$|S_k - S_k^{\sigma}| \leq \sum_{n=N+1}^{M} |a_n| < \epsilon, \tag{5.3}$$

which implies that the sequence $(S_k - S_k^{\sigma})_k$ converges to zero. Because the sequence (S_k) converges, the sequence $(S_k^{\sigma})_k$ also converges to the same limit as (S_k), and so the series $\sum(a_{\sigma(n)})_{n=1}^{\infty}$ converges to the same sum as the series $\sum(a_n)_{n=1}^{\infty}$. Simultaneously, it holds that

$$|R_k - R_k^{\sigma}| = \left|(|a_1| + \cdots + |a_N| + \cdots |a_k|) - (|a_{\sigma(1)}| + \cdots + |a_{\sigma(k)}|)\right|,$$

and an analogous argument to that applied above produces the bound

$$|R_k - R_k^{\sigma}| \leq \sum_{n=N+1}^{M} |a_n| < \epsilon, \tag{5.4}$$

which shows that the sequence $(R_k - R_k^{\sigma})_k$ converges to zero. Because the sequence (R_k) converges, the sequence $(R_k^{\sigma})_k$ also converges, and so the series $\sum(a_{\sigma(n)})_{n=1}^{\infty}$ is absolutely convergent. □

Remark 5.12: Summation of conditionally convergent series

Conditionally convergent series behave quite differently from absolutely convergent series when the order in which their terms are summed is rearranged. Namely, if $\sum(a_n)_{n=1}^{\infty}$ is a conditionally convergent series of real numbers and $\alpha \in \mathbb{R}$ is an arbitrary real number, it is possible to find a bijection $\sigma : \mathbb{N} \to \mathbb{N}$ for which $\sum_{n=1}^{\infty} a_{\sigma(n)} = \alpha$.

In other words, it is possible to rearrange the terms in a conditionally convergent series such that the series's sum comes out to any real number. However, for the rearrangement σ to change the sum of a series, σ must rearrange an infinite number of the series's terms; rearranging only a finite number of a conditionally convergent series's terms does not change the series's sum, since the addition of a finite number of real or complex numbers is commutative.

5.2 Tests for convergence of series

In this section we survey a collection of tests that can be used to confirm if a series of complex numbers converges or diverges. These tests do not, however, provide a means of determining a convergent series's sum, which is generally a much more difficult task than simply verifying convergence.

Proposition 5.13: Comparison test

Let $\sum(a_n)_{n=1}^{\infty}$ be a series of complex numbers. If there exists a convergent series of nonnegative real numbers $\sum(c_n)_{n=1}^{\infty}$ and a natural number $N \in \mathbb{N}$ for which

$$|a_n| \leq c_n \text{ for all } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely.

Proof. Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the series $\sum(c_n)_{n=1}^{\infty}$ converges, by the Cauchy criterion there exists natural number $N' \geq N$ for which

$$|c_n + c_{n+1} + \cdots + c_m| < \epsilon \text{ for all } m, n \in \mathbb{N} \text{ for which } m \geq n \geq N'.$$

Under the assumption that $|a_n| \leq c_n$ for all $n \geq N$, it follows that

$$\begin{aligned} |a_n| + |a_{n+1}| + \cdots + |a_m| &\leq c_n + c_{n+1} + \cdots + c_m \\ &\overset{\text{(a)}}{=} |c_n + c_{n+1} + \cdots + c_m| < \epsilon, \end{aligned}$$

where (a) holds under the assumption that $\sum(c_n)_{n=1}^{\infty}$ is nonnegative, proving the series $\sum(|a_n|)_{n=1}^{\infty}$ converges by the Cauchy criterion (Proposition 5.5). □

The comparison test formalizes the intuitive notion that, if a given series converges, then any other series with element-wise smaller terms should also converge. We stress that the convergent series used in the comparison test must be nonnegative and real-valued.

The comparison test can also be used, in reverse, to verify divergence.

Proposition 5.14: Comparison test for divergence

Let $\sum(c_n)_{n=1}^{\infty}$ be a series of nonnegative real numbers. If there exists a divergent series of complex numbers $\sum(a_n)_{n=1}^{\infty}$ and a natural number $N \in \mathbb{N}$ for which

$$|a_n| \leq c_n \text{ for all } n \geq N,$$

then the series $\sum(c_n)_{n=1}^{\infty}$ diverges.

Proof. We will prove the proposition by contradiction. Choose arbitrary $\epsilon \in \mathbb{R}^+$ and assume that $\sum(c_n)_{n=1}^{\infty}$ converges. By the Cauchy criterion for series convergence, there would then exists $N' \geq N$ for which

$$|c_n + c_{n+1} + \cdots + c_m| < \epsilon \text{ for all } m, n \in \mathbb{N} \text{ for which } m \geq n \geq N'.$$

We then make the computation

$$\begin{aligned} |a_n + a_{n+1} + \cdots + a_m| &\overset{\text{(a)}}{\leq} |a_n| + |a_{n+1}| + \cdots + |a_m| \leq c_n + c_{n+1} + \cdots + c_m \\ &\overset{\text{(b)}}{=} |c_n + c_{n+1} + \cdots + c_m| < \epsilon, \end{aligned}$$

where (a) follow from the triangle inequality and (b) holds under the assumption that $\sum(c_n)_{n=1}^{\infty}$ is nonnegative. The above result implies that $\sum(a_n)_{n=1}^{\infty}$ converges, which is a contradiction; thus $\sum(c_n)_{n=1}^{\infty}$ diverges. □

Definition 5.15: Majorant and minorant of a series

Consider a series of complex numbers $\sum(a_n)_{n=1}^{\infty}$ and a series of nonnegative real numbers $\sum(c_n)_{n=1}^{\infty}$. The series $\sum(c_n)_{n=1}^{\infty}$ is called a *majorant* of the series $\sum(a_n)_{n=1}^{\infty}$ *from the N-th term onward* if there exists natural number $N \in \mathbb{N}$

$$|a_n| \leq c_n \text{ for all } n \geq N.$$

Equivalently, the series $\sum(c_n)_{n=1}^{\infty}$ may be said to *dominate* the series $\sum(a_n)_{n=1}^{\infty}$ *from the N-th term onward*. In this case the series $\sum(a_n)_{n=1}^{\infty}$ is called the *minorant* of the series $\sum(c_n)_{n=1}^{\infty}$ *from the N-th term onward*.

Using this terminology, proving a series converges using the comparison test amounts to finding a convergent majorant of the series. Inversely, again by the comparison test, proving a nonnegative real series diverges amounts to finding a divergent minorant for the series.

We now demonstrate the use of the comparison test with two concrete examples.

Example 5.16: Using the comparison test

(1) We have already shown in Example 5.7 that the harmonic series $\sum(1/n)_{n=1}^{\infty}$ diverges. Using this fact, we can apply the comparison test to show that the series

$$\sum\left(\frac{1}{n^{\alpha}}\right)_{n=1}^{\infty}, \quad \alpha \in \mathbb{R}$$

diverges for all $\alpha \leq 1$. To verify this, note that

$$\frac{1}{n} \leq \frac{1}{n^\alpha} \text{ for all } n \in \mathbb{N} \text{ and } \alpha \leq 1,$$

meaning that the divergent harmonic series $\sum(1/n^\alpha)_{n=1}^{\infty}$ is a minorant of $\sum(1/n^\alpha)_{n=1}^{\infty}$, which thus itself diverges by the comparison test.

(2) By Example 5.4, the geometric series $\sum\left(\frac{1}{5^{n-1}}\right)_{n=1}^{\infty}$ converges. The series

$$\sum\left(\frac{1}{n5^{n-1}}\right)_{n=1}^{\infty}$$

also converges by comparison with the above geometric series, since

$$\frac{1}{n5^{n-1}} \leq \frac{1}{5^{n-1}} \text{ for all } n \in \mathbb{N}.$$

Proposition 5.17: The root test

Let $\sum(a_n)_{n=1}^{\infty}$ be a series of complex numbers. In this case:

(1) If there exists real number $q \in [0, 1)$ and natural number $N \in \mathbb{N}$ for which

$$\sqrt[n]{|a_n|} \leq q \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely.

(2) If there exists natural number $N \in \mathbb{N}$ for which

$$\sqrt[n]{|a_n|} \geq 1 \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ diverges.

(3) If the limit $L = \lim_{n\to\infty} \sqrt[n]{|a_n|}$ exists, then:

(i) if $L < 1$, the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely,

(ii) if $L > 1$, the series $\sum(a_n)_{n=1}^{\infty}$ diverges, and

(iii) if $L = 1$, the root test cannot determine the convergence of $\sum(a_n)_{n=1}^{\infty}$.

Proof.

(1) By the proposition's assumptions,

$$\sqrt[n]{|a_n|} \leq q \text{ for all } n \geq N.$$

We then raise both sides of the inequality to the n-th power to get the relation $|a_n| \leq q^n$, which means that the geometric series $\sum(q^n)_{n=1}^{\infty}$ is a majorant of the series $\sum(a_n)_{n=1}^{\infty}$. Because of the assumption $q < 1$, the geometric series converges by Example 5.4, and so $\sum(a_n)_{n=1}^{\infty}$ converges by the comparison test (Proposition 5.13).

(2) By the proposition's assumptions,

$$\sqrt[n]{|a_n|} \geq 1, \text{ and thus } |a_n| \geq 1, \text{ for all } n \geq N.$$

The sequence (a_n) thus does not converge to zero, and so the series $\sum(a_n)_{n=1}^{\infty}$ diverges by Corollary 5.6.

(3) (i) Assume $L < 1$ and let $q = (L+1)/2$. Then

$$L = \lim_{n\to\infty} \sqrt[n]{|a_n|} < q < 1.$$

Since L is the limit of the sequence $\left(\sqrt[n]{|a_n|}\right)_n$, by Proposition 4.12 for all $\epsilon \in \mathbb{R}^+$ there exists $N \in \mathbb{N}$ for which

$$\left|\sqrt[n]{|a_n|} - L\right| < \epsilon \text{ for all natural numbers } n \geq N,$$

Since $L < 1$ and because $\sqrt[n]{|a_n|}$ and L grow arbitrarily close for large n, there must exist $N \in \mathbb{N}$ for which

$$\sqrt[n]{|a_n|} < q < 1 \text{ for all } n \geq N.$$

The series $\sum(a_n)_{n=1}^{\infty}$ then converges absolutely by point (1).

(ii) Assuming $L = \lim_{n\to\infty} > 1$, by Proposition 4.12 there exists $N \in \mathbb{N}$ for which

$$\sqrt[n]{|a_n|} > 1 \text{ for all } n \geq N.$$

The series $\sum(a_n)_{n=1}^{\infty}$ then diverges by point (2). □

Proposition 5.18: The ratio test

Let $\sum(a_n)_{n=1}^{\infty}$ be a series of nonzero complex numbers. In this case:

(1) If there exists a real number $q \in [0,1)$ and natural number $N \in \mathbb{N}$ for which

$$\left|\frac{a_{n+1}}{a_n}\right| \leq q \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely.

(2) If there exists natural number $N \in \mathbb{N}$ for which

$$\left|\frac{a_{n+1}}{a_n}\right| \geq 1 \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ diverges.

(3) If the limit $L = \lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|$ exists, then:

(i) if $L < 1$, the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely,

(ii) if $L > 1$, the series $\sum(a_n)_{n=1}^{\infty}$ diverges, and

(iii) if $L = 1$, the ratio test cannot determine the convergence of $\sum(a_n)_{n=1}^{\infty}$.

Proof.

(1) Because finitely many terms in a series do not affect the series's convergence, we can assume without loss of generality that $N = 1$. By the proposition's assumptions, it then holds that

$$\left|\frac{a_{n+1}}{a_n}\right| \leq q \implies |a_{n+1}| \leq q|a_n| \text{ for all } n \in \mathbb{N}.$$

By repeatedly shifting the index n, this relationship implies that

$$|a_{n+1}| \leq q|a_n| \leq q^2|a_{n-1}| \leq \cdots \leq q^n|a_1|.$$

Assuming $q \in (0, 1]$, the series $\sum(a_n)_{n=1}^{\infty}$ is thus dominated by the convergent geometric series $\sum(|a_1|q^{n-1})_{n=1}^{\infty}$ and is itself convergent by the comparison test.

(2) By the proposition's assumptions, it holds that

$$\left|\frac{a_{n+1}}{a_n}\right| \geq 1 \implies |a_{n+1}| \geq |a_n| \text{ for all natural numbers } n \geq N.$$

The inequality $|a_{n+1}| \geq |a_n|$ for all $n \geq N$ means the sequence $(|a_n|)_n$ increases in absolute value with increasing n, thus cannot converge to zero, and in turn diverges by Corollary 5.6.

(3) (i) Assume $L < 1$ and let $q = (L+1)/2$. Analogously to the proof of point (3) in the root test (Proposition 5.17), it then holds that $L < q < 1$, and in turn, by the sequence limit condition in Proposition 4.12, that there exists $N \in \mathbb{N}$ for which

$$\left|\frac{a_{n+1}}{a_n}\right| < q \text{ for all } n \geq N.$$

The series $\sum(a_n)_{n=1}^{\infty}$ then converges absolutely by point (1).

(ii) Assume $L > 1$. Analogously to the proof of point (3) in the root test (Proposition 5.17), by the sequence limit condition in Proposition 4.12 there exists $N \in \mathbb{N}$ for which

$$\left|\frac{a_{n+1}}{a_n}\right| > 1 \text{ for all } n \geq N.$$

The series $\sum(a_n)_{n=1}^{\infty}$ then diverges by point (2). □

Proposition 5.19: Raabe's test

Let $\sum(a_n)_{n=1}^{\infty}$ be a series of nonzero complex numbers. In this case:

(1) If there exists a real number $q > 1$ and natural number $N \in \mathbb{N}$ for which

$$n \cdot \left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right) \geq q \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely.

(2) If there exists natural number $N \in \mathbb{N}$ for which

$$n \cdot \left(\left| \frac{a_n}{a_{n+1}} \right| - 1 \right) \leq 1 \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ does not converge absolutely.

(3) If there exists natural number $N \in \mathbb{N}$ for which

$$n \cdot \left(\left| \frac{a_n}{a_{n+1}} \right| - 1 \right) \leq 0 \text{ for all natural numbers } n \geq N,$$

then the series $\sum(a_n)_{n=1}^{\infty}$ diverges.

(4) If the limit $L = \lim\limits_{n\to\infty} n \left(\left| \frac{a_n}{a_{n+1}} \right| - 1 \right)$ exists, then:

(i) if $L > 1$, the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely,

(ii) if $0 \leq L < 1$, the series $\sum(a_n)_{n=1}^{\infty}$ does not converge absolutely, and

(iii) if $L < 0$, the series $\sum(a_n)_{n=1}^{\infty}$ diverges.

Proof.

(1) Because finitely many terms in a series do not affect the series's convergence, we can assume without loss of generality that $N = 1$. By the proposition's assumptions, it then holds that

$$n \left(\frac{a_n}{a_{n+1}} - 1 \right) \geq q \text{ for all } n \in \mathbb{N}. \tag{5.5}$$

Let $r = q - 1$. By Equation 5.5, it then follows that

$$n|a_n| - n|a_{n+1}| \geq (1 + r)|a_{n+1}|,$$

or, after rearranging

$$n|a_n| - (n + 1)|a_{n+1}| \geq r|a_{n+1}|.$$

For all $n \in \mathbb{N}$ we then define the terms

$$b_1 = 0 \quad \text{and} \quad b_{n+1} = n|a_n| - (n + 1)|a_{n+1}|,$$

which implies

$$b_{n+1} = n|a_n| - (n + 1)|a_{n+1}| \geq r|a_{n+1}| \overset{\text{(a)}}{>} 0,$$

where (a) holds under the assumption that $q > 1$ in point (1).

We then note that the series of real numbers $\sum(b_n)_{n=1}^{\infty}$ is a majorant for $\sum(a_n)_{n=1}^{\infty}$. Thus, by the comparison test, proving that $\sum(b_n)_{n=1}^{\infty}$ converges will also prove that $\sum(a_n)_{n=1}^{\infty}$ converges absolutely. To show that the series

$\sum(b_n)_{n=1}^{\infty}$ converges, we first let S_k let the series's k-th partial sum, then make the calculation

$$\begin{aligned} S_k &= b_1 + b_2 + \cdots + b_k \\ &= 0 + (|a_1| - 2|a_2|) + (2|a_2| - 3|a_3|) + \big[(k-1)|a_{k-1}| - k|a_k|\big] \\ &= |a_1| - k|a_k| \leq |a_1|. \end{aligned}$$

The sequence of partial sums (S_k) is thus bounded above by $|a_1|$. Because the sequence (S_k) is increasing (since by construction $\sum(b_n)_{n=1}^{\infty}$ is a series of nonnegative real numbers) in addition to being bounded above, the it converges by Proposition 4.15 (i). The series $\sum(b_n)_{n=1}^{\infty}$ thus converges because its sequence of partial sums converges, and so, as noted above, $\sum(a_n)_{n=1}^{\infty}$ converges by the comparison test.

(2) As point (1), we assume $N = 1$ without loss of generality. By the proposition's assumptions, it then holds that

$$n\left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right) \leq 1 \text{ for all } n \geq N,$$

which we then rearrange to get

$$n|a_n| \leq (n+1)|a_{n=1}|.$$

By repeatedly shifting the index n, it follows that

$$|a_1| \leq 2|a_2| \leq \cdots \leq n|a_n| \leq (n+1)|a_{n+1}| \leq \cdots,$$

which in particular implies that $|a_1|/n \leq |a_n|$.

The inequality $|a_1|/n \leq |a_n|$ means that the series $\sum(|a_1|/n)_{n=1}^{\infty}$ (i.e. a harmonic series scaled by the constant factor $|a_1|$) is a minorant of the series $\sum(|a_n|)_{n=1}^{\infty}$. Because the harmonic series diverges, the series $\sum(|a_n|)_{n=1}^{\infty}$ also diverges by the comparison test, and so $\sum(a_n)_{n=1}^{\infty}$ does not converge absolutely.

(3) We again assume $N = 1$ without loss of generality. By the proposition's assumptions, it then holds that

$$\left|\frac{a_n}{a_{n+1}}\right| \leq 1 \implies \left|\frac{a_{n+1}}{a_n}\right| \geq 1 \text{ for all } n \in \mathbb{N},$$

and so $\sum(a_n)_{n=1}^{\infty}$ diverges by the ratio test (Proposition 5.18).

(4) (i) Assume $L > 1$ and let $q = (L+1)/2$, which leads to the inequality

$$1 < q < L.$$

Following the same logic as in the proof of point (3) of the root test (Proposition 5.17), because L is the limit of the sequence with general term $n\big[|(a_n)/(a_{n+1})| - 1\big]$, by Proposition 4.12 there must exist $N \in \mathbb{N}$ for which

$$n\left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right) > q \text{ for all } n \geq N.$$

Since $q > 1$, the series $\sum(a_n)_{n=1}^{\infty}$ converges absolutely by point (1).

(ii) Assume $0 \leq L \leq 1$. Because L is the limit of the sequence with general term $n\left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right)$, by Proposition 4.12 there must exist $N \in \mathbb{N}$ for which
$n\left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right) < 1$ for all $n \geq N$. Thus the series $\sum(a_n)_{n=1}^{\infty}$ does not converge absolutely by point (2).

(iii) Assume $L < 1$. Because L is the limit of the sequence with general term $n\big[|(a_n)/(a_{n+1})| - 1\big]$, by Proposition 4.12 there exists $N \in \mathbb{N}$ for which

$$n\left(\left|\frac{a_n}{a_{n+1}}\right| - 1\right) < 0 \text{ for all } n \geq N.$$

The series $\sum(a_n)_{n=1}^{\infty}$ thus diverges by point (3). □

We now give some examples of using the above tests in practice.

Example 5.20: Using the root, ratio, and Raabe tests

(1) By the root test (Proposition 5.17), the series of positive real numbers

$$\sum \left(\mathrm{e}^{-n^2}\right)_{n=1}^{\infty}$$

converges, which follows from the calculation

$$\lim_{n\to\infty} \sqrt[n]{\mathrm{e}^{-n^2}} = \lim_{n\to\infty} \mathrm{e}^{-n} = \lim_{n\to\infty} \left(\frac{1}{\mathrm{e}}\right)^n = 0 < 1.$$

(2) By the ratio test (Proposition 5.18) the series of positive real numbers

$$\sum \left(\frac{n!}{n^n}\right)_{n=1}^{\infty}$$

converges, which follows from the calculation

$$\begin{aligned}
\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| &= \lim_{n\to\infty} \frac{(n+1)!}{(n+1)^{n+1}} \cdot \frac{n^n}{n!} \\
&\overset{\text{(a)}}{=} \lim_{n\to\infty} \frac{(n+1)n!}{(n+1)^n(n+1)} \cdot \frac{n^n}{n!} \\
&= \lim_{n\to\infty} \frac{n^n}{(n+1)^n} \overset{\text{(b)}}{=} \lim_{n\to\infty} \left(\frac{1}{1+\frac{1}{n}}\right)^n \\
&= \lim_{n\to\infty} \frac{1}{\left(1+\frac{1}{n}\right)^n} \overset{\text{(c)}}{=} \frac{1}{\mathrm{e}} < 1,
\end{aligned}$$

where (a) follows from the definition of the factorial operation, (b) from dividing the numerator and denominator by n, and (c) from the definition of Euler's number in Example 4.23.

(3) By the ratio test the sequence of positive real numbers

$$\sum \left(\frac{(n!)^2}{(2n)!}\right)_{n=1}^{\infty},$$

converges, which follows from the calculation

$$\begin{aligned} L &= \lim_{n\to\infty} \frac{\big[(n+1)!\big]^2}{(2n+2)!} \cdot \frac{(2n)!}{(n!)^2} \\ &= \lim_{n\to\infty} \frac{(n+1)! \cdot (n+1)!}{(2n)!(2n+1)(2n+2)} \frac{(2n)!}{n! \cdot n!} \\ &= \lim_{n\to\infty} \frac{(n+1)^2}{(2n+1)(2n+2)} = \lim_{n\to\infty} \frac{n+1}{4n+2} \\ &= \lim_{n\to\infty} \frac{1+\frac{1}{n}}{4+\frac{2}{n}} = \frac{1}{4} < 1. \end{aligned}$$

(4) Let $\alpha \in \mathbb{R}$ be an arbitrary real number, and let us use Raabe's test (Proposition 5.19) to analyze the convergence of the series of positive real numbers

$$\sum \left(\frac{1}{n^\alpha}\right)_{n=1}^{\infty}.$$

First define a function $f : \mathbb{R}^+ \to \mathbb{R}$ with the mapping formula $f(x) = x^\alpha$. Although it predates our definition of the derivative (which we will study in Chapter 7), we note that the derivative f' of the above function f is

$$f'(x) = \alpha x^{\alpha-1}.$$

We then make the calculation

$$\begin{aligned} L &= \lim_{n\to\infty} n \left(\frac{(1/n^\alpha)}{1/\big[(n+1)^\alpha\big]} - 1 \right) \\ &= \lim_{n\to\infty} n \left[\left(\tfrac{n+1}{n}\right)^\alpha - 1 \right] \\ &= \lim_{n\to\infty} n \left[\left(1+\tfrac{1}{n}\right)^\alpha - 1 \right] \\ &\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{1}{h} \big[(1+h)^\alpha - 1\big] \\ &\overset{\text{(b)}}{=} \lim_{h\to 0} \frac{f(1+h) - f(1)}{h} \\ &\overset{\text{(c)}}{=} f'(1) = \alpha 1^{\alpha-1} = \alpha. \end{aligned}$$

where in (a) we have defined $h = 1/n$, (b) follows from the definition $f(x) = x^\alpha$, and (c) follows from the definition of the derivative (Definition 7.1).

By point (4) of Raabe's test, the series $\sum \left(\frac{1}{n^\alpha}\right)_{n=1}^{\infty}$ thus converges if $\alpha > 1$ and diverges if $\alpha < 1$. Point (4) of Raabe's test cannot conclusively determine the series's convergence in the case $\alpha = 1$, but point (2) can. Namely, if $\alpha = 1$ (which satisfies point (2) of Raabe's test), the series $\sum \left(\frac{1}{n^\alpha}\right)_{n=1}^{\infty}$ does not converge absolutely; however, because all elements in the series are positive, which means that

$$\frac{1}{n^\alpha} = \left|\frac{1}{n^\alpha}\right| \text{ for all } n \in \mathbb{N},$$

if the series does not converge absolutely it does not converge at all, and thus diverges if $\alpha = 1$. (Of course, if $\alpha = 1$ the series is the harmonic series $\sum(\frac{1}{n})_{n=1}^{\infty}$, which we already know diverges from Example 5.7.)

Note that each of the series in the above example was real-valued with positive terms, in which case convergence and absolute convergence are equivalent. More generally, series can have negative terms, and an important case of a real-valued series with negative terms is an alternating series, which we define below.

Definition 5.21: Alternating series

An *alternating series* is a real-valued series of the form

$$\sum \left((-1)^{n+1} a_n\right)_{n=1}^{\infty} \quad \text{or} \quad \sum \left((-1)^{n} a_n\right)_{n=1}^{\infty},$$

where $a_n \geq 0$ for all $n \in \mathbb{N}$.

Leibniz's test, which we state below, is commonly used to test the convergence of alternating series.

Proposition 5.22: Leibniz's alternating series test

For any alternating series of real numbers

$$\sum \left((-1)^{n+1} a_n\right)_{n=1}^{\infty},$$

if the sequence (a_n) is decreasing and converges to the number zero, then the series $\sum \left((-1)^{n+1} a_n\right)_{n=1}^{\infty}$ also converges and obeys the inequality

$$\left| \sum_{n=1}^{\infty} (-1)^{n+1} a_n - \sum_{n=1}^{k} (-1)^{n+1} a_n \right| \leq a_{k+1} \text{ for all } k \in \mathbb{N}.$$

Proof. For all $k \in \mathbb{N}$ we first define the partial sum

$$S_k = \sum_{n=1}^{k} (-1)^{n+1} a_n.$$

The sequence (a_n) is decreasing and positive by the proposition's assumptions, so for each $m \in \mathbb{N}$ we can make the bound

$$\begin{aligned} S_{2m} &\leq S_{2m} + (a_{2m+1} - a_{2m+2}) = S_{2m+2} \\ &= S_{2m+1} - a_{2m+2} \leq S_{2m+1} \\ &= S_{2m-1} - (a_{2m} - a_{2m+1}) \leq S_{2m-1}. \end{aligned}$$

The result $S_{2m} \leq S_{2m+2}$ means that the sequence $(S_{2m})_m$ is increasing, while $S_{2m} \leq S_{2m-1}$ means that the sequence $(S_{2m-1})_m$ is decreasing. Because (a_n) converges to zero, both sequences must also be bounded, so both the sequences $(S_{2m})_m$ and $(S_{2m-1})_m$ converge by Proposition 4.15.

More so, the sequences $(S_{2m})_m$ and $(S_{2m-1})_m$ also converge to the same limit, since

$$\begin{aligned} \lim_{m\to\infty} S_{2m} &= \lim_{m\to\infty} (S_{2m-1} - a_{2m}) \\ &\overset{(a)}{=} \lim_{m\to\infty} S_{2m-1} - \lim_{m\to\infty} a_{2m} \\ &\overset{(b)}{=} \lim_{m\to\infty} S_{2m-1}, \end{aligned}$$

where (a) follows from the limit operation's distributivity over addition (Proposition 4.17 (i)) and (b) from the assumption that (a_n) converges to zero. Because both the even-indexed and odd-indexed subsequences of (S_k) converge to the same limit, the full sequence of partial sums (S_k) must also converge to that limit by a reverse-engineered version of Proposition 4.26. The alternating series $\sum\left((-1)^{n+1}a_n\right)_{n=1}^{\infty}$ thus converges because its sequence partial sums (S_k) converges.

To prove the proposition's inequality, let S denote the sum of the series

$$\sum\left((-1)^{n+1}a_n\right)_{n=1}^{\infty}.$$

Because the subsequence of partial sums with even indices $(S_{2m})_m$ increases towards S and the subsequence of partial sums with odd indices $(S_{2m-1})_m$ decreases towards S, the partials sums obey the inequality

$$S_{2m} \leq S \leq S_{2m+1} \leq S_{2m-1} \text{ for all } m \in \mathbb{N}.$$

This inequality in turn implies that

$$\left|\sum_{n=1}^{\infty}(-1)^{n+1}a_n - \sum_{n=1}^{k}(-1)^{n+1}a_n\right| = |S - S_k| \leq |S_{k+1} - S_k| = a_{k+1}$$

for all $k \in \mathbb{N}$, which proves the second part of the proposition. □

We conclude this chapter with two examples of using Leibniz's test to verify the convergence of alternating series.

Example 5.23: Alternating series

(1) Let us verify the convergence of the series

$$\sum\left((-1)^{n+1}\frac{1}{n}\right)_{n=1}^{\infty}.$$

The series is an alternating series, and because its terms monotonically decrease and converge to zero by absolute value, the series converges by Leibniz's test (Proposition 5.22). Later, we will show the series's sum equals the natural logarithm of the number 2.

Note that the series formed from the above series's absolute value elements is the divergent harmonic series

$$\sum(1/n)_{n=1}^{\infty},$$

so the alternating series $\sum\left((-1)^{n+1}/n\right)_{n=1}^{\infty}$ converges only conditionally.

(2) The series in point (1) is a special case of the alternating series

$$\sum\left((-1)^{n+1}\frac{1}{n^{\alpha}}\right)_{n=1}^{\infty}, \quad \alpha \in \mathbb{R}.$$

If $\alpha > 0$, the series's elements decrease monotonically to zero, and so the series converges by Leibniz's test. More specifically, the series converges absolutely if

$\alpha > 1$ and converges conditionally if $0 < \alpha \leq 1$. If $\alpha \leq 0$, the series's terms do no converge to zero, and so the series diverges by Corollary 5.6.

6 Functions of a single real variable

The functions used most often to describe physical quantities are scalar-valued and vector-valued functions of one or more real variables. In this chapter we will study the simplest case: scalar-valued functions of a single real variable.

Definition 6.1: Real-valued function of one real variable

A *real-valued function of a single real variable* is a function of the form

$$f\colon U \to \mathbb{R}, \quad x \mapsto f(x),$$

whose domain $U \subset \mathbb{R}$ is some subset of the real numbers. For shorthand, real-valued functions of a single real variable are often called *scalar* functions, since the elements of such a function's domain and codomain are scalar numbers. We will regularly use the term "scalar function" throughout the remainder of this book. The graph (Definition 1.12) of a scalar function is a subset of the plane $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

Recall, from Remark 1.11, that a function is fully specified by the combination of its domain, codomain, and mapping formula. However, it is common practice to state scalar functions with explicitly stating their mapping formulae. When the domain and codomain of a scalar function are not stated explicitly, by convention one implicitly assumes the codomain is $\mathbb{R}$ and the domain is the set of all real numbers for which the function's mapping formula is well defined, i.e. those $x \in \mathbb{R}$ for which it is possible to compute the function value $f(x)$ using the given mapping formula.

Definition 6.2: Zero of a scalar function

A *zero* of a scalar function $f\colon U \to \mathbb{R}$ defined on the real subset $U \subset \mathbb{R}$ is any point $\xi \in U$ for which

$$f(\xi) = 0,$$

i.e. any point in the function's domain that is mapped to zero.

We describe the growth of scalar functions using analogous terminology and definitions to those used to describe the growth of real sequences in Definition 4.3.

Definition 6.3: Describing the growth of functions

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$.

(i) The function f is said to be *increasing* if

$$f(y) \geq f(x) \text{ for all } x, y \in U \text{ for which } y > x.$$

(ii) The function f is said to be *decreasing* if

$$f(y) \leq f(x) \text{ for all } x, y \in U \text{ for which } y > x.$$

(iii) The function f is *monotonic* if it is either increasing or decreasing.

(iv) The function f is said to be *strictly increasing*

$$f(y) > f(x) \text{ for all } x, y \in U \text{ for which } y > x.$$

(v) The function f is said to be *strictly decreasing*

$$f(y) < f(x) \text{ for all } x, y \in U \text{ for which } y > x.$$

(vi) The function f is said to be *strictly monotonic* if it is either strictly increasing or strictly decreasing.

The terminology and definitions used to describe the boundedness of scalar functions are again analogous to their scalar analogs, in this case from Definition 4.2.

Definition 6.4: Describing the boundedness of functions

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$.

(i) The function f is said to be *bounded above* if its image $f(U)$ is an above-bounded subset of $\mathbb{R}$.

(ii) The function f is said to be *bounded below* if its image $f(U)$ is a below-bounded subset of $\mathbb{R}$.

(iii) The function f is *bounded* if it is both bounded above and bounded below.

Definition 6.5: The supremum, infimum, and extrema of a function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on a nonempty real subset $U \subset \mathbb{R}$.

(i) If the function f is bounded above, then f's *supremum* is defined as the supremum of its image $f(U) \subset \mathbb{R}$ and is denoted by

$$\sup(f) = \sup\big(f(U)\big).$$

If in addition f's image contains its supremum, i.e. if $\sup(f) \in f(U)$, then the *maximum* of the function f, denoted by $\max(f)$, is defined as

$$\max(f) = \sup(f) \qquad (\text{if } \sup(f) \in f(U)).$$

(ii) If the function f is bounded below, then f's *infimum* is defined as the infimum of its image $f(U) \subset \mathbb{R}$ and is denoted by

$$\inf(f) = \inf\big(f(U)\big).$$

If in addition f's image contains its infimum, i.e. if $\inf(f) \in f(U)$, then the *minimum* of the function f, denoted by $\min(f)$, is defined as

$$\min(f) = \inf(f) \qquad (\text{if } \inf(f) \in f(U)).$$

Example 6.6: Basic concepts related to scalar functions

(1) The identity function (see Definition (i)) on $\mathbb{R}$, defined as $\mathrm{id}_{\mathbb{R}}\colon \mathbb{R} \to \mathbb{R}$ with the mapping formula $x \mapsto x$, is an example of a scalar function. The identify function is strictly increasing and is neither bounded above nor bounded below.

(2) Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$. In this case the restriction (see Definition (ii)) $f|_V$ of the function f to an arbitrary subset $V \subset U$ of f's domain is also a scalar function.

(3) Let $U \in \mathbb{R}$ be a subset of $\mathbb{R}$. The function $f\colon U \to \mathbb{R}$ is called the *constant function with value* $c \in \mathbb{R}$ if

$$f(x) = c \text{ for all } x \in U.$$

Every constant scalar function is bounded and simultaneously increasing and decreasing. However, if U has at least two elements, then any constant scalar function is neither strictly increasing nor strictly decreasing.

(4) A *linear* function of a single real variable is a function of the form

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = kx + n$$

for any two constants $k, n \in \mathbb{R}$. A linear function is strictly increasing if $k > 0$ and strictly decreasing if $k < 0$. A linear function f is bounded if, and only if, $k = 0$, in which case f is a constant function with value n.

As background for defining even and odd functions, we first introduce the concept of symmetry about the origin. For any real subset $U \subset \mathbb{R}$, the subset $-U$ is defined as

$$-U = \{-x;\, x \in U\}.$$

If $U = -U$, then the set U is said to be *symmetric about the origin*.

Definition 6.7: Even and odd functions

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on a real subset $U \subset \mathbb{R}$ that is symmetric about the origin. In this case:

(i) The function f is *even* if $f(-x) = f(x)$ for all $x \in U$.

(ii) The function f is *odd* if $f(-x) = -f(x)$ for all $x \in U$.

The graph of an even function is geometrically symmetric about the ordinate axis, while the graph of an odd function is geometrically symmetric about the origin.

We stress that the properties of evenness and oddness are well-defined only for functions whose domains are symmetric about the origin. However, because the set of real numbers $\mathbb{R}$ is symmetric about the origin, and because $\mathbb{R}$ is a common domain for scalar functions, this restriction is rarely a problem in practice.

Example 6.8: Sums of even and odd functions

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the origin-symmetric subset $U = -U$, and define the two functions $f_\mathrm{e}, f_\mathrm{o}\colon U \to \mathbb{R}$ as

$$f_\mathrm{e}(x) = \frac{1}{2}\big[f(x) + f(-x)\big] \quad \text{and} \quad f_\mathrm{o}(x) = \frac{1}{2}\big[f(x) - f(-x)\big];$$

Note that f_e is even by construction, that f_o is odd by construction, and that

$$f = f_\mathrm{e} + f_\mathrm{o}.$$

Because f was arbitrary except for the symmetric-domain condition $U = -U$, this example is a proof by construction that any scalar function defined on a domain with origin symmetry can be written as the sum of an even and odd function.

We now state some definitions of how functions can be combined under common operations. We will regularly use these definitions throughout the remainder of this book.

Definition 6.9: Operations involving scalar functions

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the real subsets $U, V \subset \mathbb{R}$. In this case:

(i) The *sum* of f and g is denoted by $f + g$ and defined as

$$f + g\colon U \cap V \to \mathbb{R}, \quad x \mapsto (f + g)(x) = f(x) + g(x).$$

(ii) The *product* of f and g is denoted by either $f \cdot g$ or fg and is defined as

$$fg\colon U \cap V \to \mathbb{R}, \quad x \mapsto (fg)(x) = f(x)g(x).$$

(iii) The *quotient* of f and g is denoted by f/g and defined as

$$f/g\colon U \cap g^{-1}(\mathbb{R} \setminus \{0\}) \to \mathbb{R}, \quad x \mapsto (f/g)(x) = \frac{f(x)}{g(x)}.$$

The function f/g is defined on the intersection of f and the set of all x in g's domain V that are not mapped to zero.

(iv) The composition of f and g is denoted by $f \circ g$ and defined as

$$f \circ g\colon g^{-1}(U) \to \mathbb{R}, \quad x \mapsto (f \circ g)(x) = f(g(x)).$$

The function $f \circ g$ is defined on the set of all x in g's domain V that map to f's domain U. You may also be interested in review Definition 1.18, which defines the composition of arbitrary functions (not just scalar functions).

One must proceed with some caution when specifying the domains of the above functions. In each of the above points, the domain of the defined function is the set of all points in which the corresponding mapping formula is possible to calculate.

Remark 6.10: Scalar functions and group theory

From Definition 6.9 (i) and (ii), if two scalar functions f and g are defined on the same real subset $U \subset \mathbb{R}$, then their sum $f + g$ and product fg are also scalar functions defined on U; the set of all scalar functions mapping from U to $\mathbb{R}$ is thus closed under addition and multiplication. This set is in fact an Abelian group under addition (recall Remark 2.2 in the context of the real numbers) for which the additive identity is the zero function, and the additive inverse for the function $f\colon U \to \mathbb{R}$ is the function $-f\colon U \to \mathbb{R}, (-f)(x) = -f(x)$.

We defined the inverse of a general function in Definition 1.20 of Section 1.2.3; we now give

the more specific definition of the inverse of a scalar function.

Remark 6.11: Inverse of a scalar function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the nonempty subset $U \subset \mathbb{R}$. If f is injective, then we define its inverse f^{-1} as

$$f^{-1}\colon f(U) \to \mathbb{R}, \quad f^{-1}(y) \in f^{-1}(\{y\}),$$

where $f^{-1}(\{y\})$ is the preimage of the point $y \in f(U)$; this preimage has exactly one point for every $y \in f(U)$ under the assumption that f is injective, and so $f^{-1}(y)$ is well-defined for all $y \in f(U)$.

We must be careful, though. If $f(U) \neq \mathbb{R}$ and $f(U)$ does not contain all elements in $\mathbb{R}$, then $f\colon f(U) \to \mathbb{R}$ is not surjective and does not have a right inverse (and thus does not have an inverse). In this case by f^{-1} we implicitly denote the inverse of the function

$$f|^{f(U)}\colon U \to f(U), \quad x \mapsto f(x),$$

which is simply the function f with the original codomain $\mathbb{R}$ replaced by $f(U)$. The function $f|^{f(U)}$ is surjective by construction, injective because f is injective, and thus bijective; we denote its inverse by $\left(f|^{f(U)}\right)^{-1}\colon f(U) \to U$.

In this case the notation f^{-1} used loosely formally refers to the function

$$\mathrm{id}_{\mathbb{R}}|_U \circ \left(f|^{f(U)}\right)^{-1}.$$

Proposition 6.12: Growth and injectivity of scalar functions

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$. In this case:

(i) If f is strictly increasing then f is also injective, and its inverse function $f^{-1}\colon f(U) \to \mathbb{R}$ is also a strictly increasing function.

(ii) If f is strictly decreasing then f is also injective, and its inverse function $f^{-1}\colon f(U) \to \mathbb{R}$ is also a strictly decreasing function.

Proof.

(i) Assume f is strictly increasing, and let $x, x' \in U$ be any two elements in f's domain for which $x \neq x'$. Because f is strictly increasing, if $x > x'$ then $f(x) > f(x')$; similarly if $x < x'$ then $f(x) < f(x')$. In either case $x' \neq x$ implies $f(x) \neq f(x')$, and so f is injective.

Now consider any two function values $u, v \in f(U)$, and note by the definition of the inverse function in Definition 1.20 that

$$u = f(f^{-1}(u)) \quad \text{and} \quad v = f(f^{-1}(v)).$$

Because f is strictly increasing, if $f^{-1}(u) > f^{-1}(v)$ then

$$u = f(f^{-1}(u)) > f(f^{-1}(v)) = v.$$

Reversing this result then means that if $u > v$, then $f^{-1}(u) > f^{-1}(v)$, and so f^{-1} is increasing.

(ii) Analogous to the proof of item (i), just with inequalities reversed. □

6.1 The limit of a scalar function

Definition 6.13: Cluster point and isolated point of a real subset

The point $a \in \mathbb{R}$ is a *cluster point* of the real subset $A \subset \mathbb{R}$ if

$$(a-\delta, a+\delta) \cap (A \setminus \{a\}) \neq \emptyset \text{ for all } \delta \in \mathbb{R}^+.$$

In words, a is a cluster point of A if any arbitrarily small real interval centered at a contains elements of A other than a.

Meanwhile, a point $b \in \mathbb{R}$ is said to be an *isolated point* of A if b is an element of A but not a cluster point of A.

Any interior point (Definition 2.29 (i)) of a real subset is necessarily a cluster point of the set, while any exterior point (Definition 2.29 (ii)) is necessarily not a cluster point of the set. Each of a real subset's boundary points (Definition 2.29 (iii)) is either a cluster point or an isolated point of the set.

In practice, one often uses the following condition to verify if a given point is a cluster point of a real subset.

Proposition 6.14: Condition for a cluster point of a subset

The point $a \in \mathbb{R}$ is a cluster point of the real subset $A \subset \mathbb{R}$ if, and only if, there exists a sequence of real numbers (a_n) for which $a_n \in A \setminus \{a\}$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} a_n = a.$$

In words, $a \in \mathbb{R}$ is a cluster point of A if it is possible to approximate a arbitrarily closely using a sequence of points taken from the set $A \setminus \{a\}$.

Proof. ($\Rightarrow$) Assume a is a cluster point of A, which means that the set

$$\left(a-\frac{1}{n}, a+\frac{1}{n}\right) \cap (A \setminus \{a\})$$

is nonempty for all $n \in \mathbb{N}$. Because this set is nonempty, for every $n \in \mathbb{N}$ we can choose from it a point a_n to form a sequence (a_n) that converges to (a) with that property that $a_n \in A \setminus \{a\}$ for all $n \in \mathbb{N}$.

($\Leftarrow$) Let (a_n) be a sequence of real numbers for which

$$\lim_{n\to\infty} a_n = a \quad \text{and} \quad a_n \in A \setminus \{a\} \text{ for all } n \in \mathbb{N}.$$

By the definition of the limit of a sequence (Definition 4.11), for all $\delta \in \mathbb{R}^+$ the neighborhood $(a-\delta, a+\delta)$ of the point a contains all elements of (a_n) with the possible exception of finitely many. Since (a_n)'s elements are elements of A, the set $(a-\delta, a+\delta) \cap (A \setminus \{a\})$ is nonempty, and so a is a cluster point of A. □

Example 6.15: Interior, exterior, and cluster points

(i) The real subset $A = \{1/n\,;\, n \in \mathbb{N}\} \subset \mathbb{R}$ does not contain any nonempty open intervals and thus does not have any interior points. All points in $\mathbb{R} \setminus A$ are exterior points of A except for 0; in fact 0 is the only cluster point of A, while all elements of A are isolated points.

(ii) Consider the subset $A = [0, 1) \cup \{2\} \subset \mathbb{R}$. The points in the open interval $(0, 1)$ are A's interior points, and A's boundary points are the numbers 0, 1, and 2. All points in the closed interval $[0, 1]$ are A's cluster points (included 1, which itself is not an element of A). The number 2 is the only isolated point of the subset A.

We are now prepared to define the limit of a scalar function, which will feature centrally throughout the remainder of this book.

Definition 6.16: Limit of a scalar function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$ and let $a \in \mathbb{R}$ be a cluster point of the subset U. The number $L \in \mathbb{R}$ is the *limit* of the function f *at the point* a if for all $\epsilon \in \mathbb{R}^+$ there exists sufficiently small $\delta \in \mathbb{R}^+$ for which

$$|f(x) - L| < \epsilon \text{ for all } x \in (a - \delta, a + \delta) \cap (U \setminus \{a\});$$

or, equivalently, for which

$$f(x) \in (L - \epsilon, L + \epsilon) \text{ for all } x \in (a - \delta, a + \delta) \cap (U \setminus \{a\}).$$

In either case, the limit of f is said to *exist* at a and is written

$$L = \lim_{x \to a} f(x).$$

Intuitively, L is the limit of the function f at the point a if it is possible to approximate L arbitrarily closely with function values $f(x)$ as the function argument x comes arbitrarily close to a. Note that f need not be defined at a for the f's limit to exist at a; in fact, the value of f at a is completely irrelevant to the definition of the limit of f at a.

Proposition 6.17: A function's limit is unique

A function $f\colon U \to \mathbb{R}$ can have at most one limit at any cluster point a of the subset $U \subset \mathbb{R}$.

Proof. Let L be the limit of f at a and let L' be any real number different from L. Next, define $\epsilon \in \mathbb{R}^+$ for which

$$\epsilon < \frac{|L - L'|}{2},$$

in which case, by the triangle inequality, the intervals $(L-\epsilon, L+\epsilon)$ and $(L'-\epsilon, L'+\epsilon)$ are disjoint. Since L is the limit of f at a, there exists $\delta \in \mathbb{R}^+$ for which

$$|f(x) - L| < \epsilon \text{ for all } x \in (a - \delta, a + \delta) \cap (U \setminus \{a\}).$$

Since $(L-\epsilon, L+\epsilon)$ and $(L'-\epsilon, L'+\epsilon)$ are disjoint, for each such x it also holds that

$$\left|f(x) - L'\right| \geq \epsilon.$$

It then follows that for arbitrary positive real number $\delta' < \delta$ there exists some

$$x \in (a-\delta', a+\delta') \cap (U \setminus \{a\})$$

for which $|f(x) - L'| \geq \epsilon$, and so L' cannot be the limit of f at the point a. Thus no real number than L can be f's limit at a. □

Definition 6.18: Left- and right-sided limit

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$.

- If $a \in \mathbb{R}$ is a cluster point of the subset $U \cap (a, \infty)$ and if the limit of the restricted function $f|_{U\cap(a,\infty)}$ exists at the point a, this limit is called the *right-sided limit* of f at the point a and is variously denoted by

$$\lim_{x\to a^+} f(x) = \lim_{x\downarrow a} f(x) = \lim_{x\to a}(f_{U\cap\{(a,\infty)\}})(x).$$

- If $a \in \mathbb{R}$ is a cluster point of the subset $U \cap (-\infty, a)$ and if the limit of the restricted function $f|_{U\cap(-\infty,a)}$ exists at the point a, this limit is called the *left-sided limit* of f at the point a and is variously denoted by

$$\lim_{x\to a^-} f(x) = \lim_{x\uparrow a} f(x) = \lim_{x\to a}(f_{U\cap\{(-\infty,a)\}})(x).$$

From the above definition, if $a \in \mathbb{R}$ is a cluster point of both $U \cap (a, \infty)$ and $U \cap (-\infty, a)$, then $L \in \mathbb{R}$ is the limit of the function f at a if, and only if, L is both the left-sided and right-sided limit of f and a.

In addition to considering a function's limit behavior a particular point $a \in \mathbb{R}$, one is often interested in the function's behavior as its argument increases without bound—we explore this situation in the following pages.

Definition 6.19: Limits at infinity

- Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the above-unbounded real subset $U \subset \mathbb{R}$. The real number $L \in \mathbb{R}$ is called the *limit* of f *at infinity* if for all $\epsilon \in \mathbb{R}^+$ there exists sufficiently large $B \in \mathbb{R}$ for which

$$|f(x) - L| < \epsilon \text{ for all } x \in (B, \infty) \cap U.$$

  In this case the limit of f at infinity is said to *exist* and is denoted by

$$L = \lim_{x\to\infty} f(x) \qquad \text{(notation: limit of } f \text{ at infinity).}$$

- Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the below-unbounded real subset $U \subset \mathbb{R}$. The real number $L \in \mathbb{R}$ is called the *limit* of f *at negative infinity* if

for all $\epsilon \in \mathbb{R}^+$ there exists sufficiently negative $b \in \mathbb{R}$ for which

$$|f(x) - L| < \epsilon \text{ for all } x \in (-\infty, b) \cap U.$$

In this case the limit of f at negative infinity is said to *exist* and is denoted by

$$L = \lim_{x \to -\infty} f(x) \qquad \text{(notation: limit of } f \text{ at negative infinity).}$$

If the limit $L = \lim_{x\to\infty} f$ of a scalar function $f\colon U \to \mathbb{R}$ exists at infinity, then for any (ϵ, B) pair the entire graph of f on the set $(B, \infty) \cap U$ lies between the two horizontal lines with heights $L - \epsilon$ and $L + \epsilon$. The horizontal line with height L is called the *horizontal asymptote* of f *at infinity*. Similarly, if the limit $L = \lim_{x\to-\infty}$ of f at negative infinity exists, then for any (ϵ, b) pair the entire graph of f on the set $(-\infty, b) \cap U$ lies between the two horizontal lines with heights $L - \epsilon$ and $L + \epsilon$. The horizontal line with height L is called the *horizontal asymptote* of f *at negative infinity*.

Remark 6.20: Limits at infinity as limits at zero

By defining a new variable $t = 1/x$, the limits at positive infinity and negative infinity of a function $f\colon U \to \mathbb{R}$ can be interpreted as a special case of a function at the point zero, since

$$\lim_{x\to\infty} f(x) = \lim_{t\to 0^+} f(1/t) \qquad \text{(at positive infinity)}$$
$$\lim_{x\to-\infty} f(x) = \lim_{t\to 0^-} f(1/t) \qquad \text{(at negative infinity).}$$

For this reason, limits at infinity obey the same properties as limits at a finite point.

Example 6.21: Limits of monotonic functions at interval endpoints

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and define the constant $c \in U$.

(i) If f is increasing and c is a cluster point of the subset $U \cap (-\infty, c)$, then $f(c)$ is an upper bound of the set $\{f(x)\,;\, x \in U, x < c\}$, and this set's supremum is the left-sided limit of f at c, i.e.

$$\lim_{x\to c^-} f(x) = \sup\{f(x)\,;\, x \in U, x < c\} \leq f(c).$$

(ii) If f is increasing and c is a cluster point of the subset $U \cap (c, \infty)$, then $f(c)$ is a lower bound of the set $\{f(x)\,;\, x \in U, x > c\}$, and this set's infimum is the right-sided limit of f at c, i.e.

$$\lim_{x\to c^+} f(x) = \inf\{f(x)\,;\, x \in U, x > c\} \geq f(c).$$

(iii) If f is decreasing and c is a cluster point of the subset $U \cap (-\infty, c)$, then $f(c)$ is a lower bound of the set $\{f(x)\,;\, x \in U, x < c\}$, while the set's infimum is the left-sided limit of f at c, i.e.

$$\lim_{x\to c^-} f(x) = \inf\{f(x)\,;\, x \in U, x < c\} \geq f(c).$$

(iv) If f is decreasing and c is a cluster point of the subset $U \cap (c, \infty)$, then $f(c)$ is an upper bound of the set $\{f(x)\,;\, x \in U, x > c\}$, while the set's supremum is the right-sided limit of f at c, i.e.

$$\lim_{x \to c^+} f(x) = \sup\{f(x)\,;\, x \in U, x > c\} \leq f(c).$$

The following proposition characterizes limits of functions in terms of limits of sequences.

Proposition 6.22: Relating limits of functions and sequences

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and let $a \in \mathbb{R}$ be a cluster point of U. In this case the number $L \in \mathbb{R}$ is the limit of f at the point a if, and only if, for all sequences (x_n) from the set $U \setminus \{a\}$ that converge to a, the corresponding sequence of function values $(f(x_n))_n$ converges to L, i.e. if

$$L = \lim_{n \to \infty} f(x_n).$$

Proof. ($\Rightarrow$) Let L be the limit of f at the point a, and let (x_n) be an arbitrary sequence from the set $U \setminus \{a\}$ that converges to a; we must show that the sequence of function values $(f(x_n))_n$ converges to L. To do so, first choose arbitrary $\epsilon \in \mathbb{R}^+$. By the definition of the limit of a function, there exists $\delta \in \mathbb{R}^+$ for which

$$|f(x) - L| < \epsilon \text{ for all } x \in (a - \delta, a + \delta) \cap (U \setminus \{a\}). \tag{6.1}$$

Because the sequence (x_n) converges to a, by Proposition 4.12 there exists $N \in \mathbb{N}$ for which

$$x_n \in (a - \delta, a + \delta) \cap (U \setminus \{a\}) \text{ for all } n \geq N.$$

Thus for all such sufficiently-large n, the sequence elements x_n fall in the subset $x \in (a - \delta, a + \delta) \cap (U \setminus \{a\})$ for which Equation 6.1 is valid; this means

$$|f(x_n) - L| < \epsilon \text{ for all } n \geq N,$$

and so the sequence of function values $(f(x_n))_n$ converges to L by Proposition 4.12.

($\Leftarrow$) Assume that $L = \lim_{n\to\infty} f(x_n)$ for all sequences (x_n) from the set $U \setminus \{a\}$ that converge to a. We will then show, by contradiction, that L must be the limit of the function f at the point a.

To do so, assume that L is not the limit of f at a. Since L is not the limit of f at a, for any $\delta \in \mathbb{R}^+$ there exists some $\epsilon \in \mathbb{R}^+$ for which

$$|f(x) - L| \geq \epsilon \text{ for some } x \in (a - \delta, a + \delta) \cap (U \setminus \{a\}).$$

Thus, for all $n \in \mathbb{N}$ there exists a point

$$x_n \in \left(a - \frac{1}{n}, a + \frac{1}{n}\right) \cap (U \setminus \{a\})$$

for which $|f(x_n) - L| \geq \epsilon$. We have thus constructed a sequence (x_n) from $U \setminus \{a\}$ that converges to a, but for which L is not the limit of the sequence $(f(x_n))_n$. This contradicts the assumption $L = \lim_{n\to\infty} f(x_n)$ for all sequences (x_n) from the set $U \setminus \{a\}$ that converge to a, and so L must be the limit of f at a. □

By relating limits of sequences to limits of functions, Proposition 6.22 makes it possible to use sequence-related propositions from Chapter 4 to prove analogous propositions applied to the limits of scalar functions. We will make heavy use of this convenient fact in the following pages.

Remark 6.23: Limits of functions at infinity

Because, as mentioned in Remark 6.20, limits of functions at infinity are a special case of limits of functions at the point zero and thus obey the same properties as limits of functions at points in $\mathbb{R}$, a similar proposition to Proposition 6.22 holds for limits at infinity, i.e.:

(i) If $f\colon U \to \mathbb{R}$ is a scalar function defined on an above-unbounded subset $U \subset \mathbb{R}$, then the number $L \in \mathbb{R}$ is the limit of f at infinity if, and only if,

$$L = \lim_{n\to\infty} f(x_n)$$

for all sequences (x_n) from the subset U for which $\lim_{n\to\infty} x_n = \infty$.

(ii) If $f\colon U \to \mathbb{R}$ is a scalar function defined on an below-unbounded subset $U \subset \mathbb{R}$, then the number $L \in \mathbb{R}$ is the limit of f at negative infinity if, and only if,

$$L = \lim_{n\to\infty} f(x_n)$$

for all sequences (x_n) from the subset U for which $\lim_{n\to\infty} x_n = -\infty$.

The following rules are useful in practice for computing the limits of combinations of functions; these rules are analogs of the rules given in Proposition 4.17 for computing the limits of functions of sequences.

Proposition 6.24: Rules for computing limits of functions

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the real subsets $U, V \subset \mathbb{R}$, let $a \in \mathbb{R}$ be a cluster point of the subset $U \cap V \subset \mathbb{R}$, and assume both f and g have a limit at the point a. In this case:

(i) The function $f + g$ has a limit at a equal to

$$\lim_{x\to a} \big(f(x) + g(x)\big) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x).$$

(ii) The function $f \cdot g$ has a limit at a equal to

$$\lim_{x\to a} \big[f(x)g(x)\big] = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x).$$

(iii) If the limit of g at the point $a \in \mathbb{R}$ is nonzero, then a is a cluster point of the subset $U \cap g^{-1}(R \setminus \{0\}) \subset \mathbb{R}$, and the function f/g has a limit at a equal to

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim\limits_{x\to a} f(x)}{\lim\limits_{x\to a} g(x)}.$$

Proof. The proof follows from the rules for computing limits of sequences in Proposition 4.17, and from Proposition 6.22, which allows one to formulate the limit of a function in terms of the limit of a sequence.

(iii) Under the assumption $\lim_{x\to a} g(x) \neq 0$, by the definition of the limit of a function (Definition 6.16) there exists $\delta \in \mathbb{R}^+$ for which

$$g(x) \neq 0 \text{ for all } x \in (a-\delta, a+\delta) \cap (V \setminus \{a\}).$$

Because a is assumed to be a cluster point of $U \cap V$, it must also a cluster point of $U \cap g^{-1}(\mathbb{R} \setminus \{0\})$. We then choose an arbitrary sequence (x_n) from the set $\big(U \cap g^{-1}(\mathbb{R} \setminus \{0\})\big) \setminus \{a\}$ that converges to a. By Proposition 6.22 it holds that

$$\lim_{n\to\infty} f(x_n) = \lim_{x\to a} f(x) \quad \text{and} \quad \lim_{n\to\infty} g(x_n) = \lim_{x\to a} g(x).$$

The rules for computing limits of sequences in Proposition 4.17 then imply that

$$\lim_{n\to\infty} \frac{f(x_n)}{g(x_n)} = \frac{\lim_{n\to\infty} f(x_n)}{\lim_{n\to\infty} g(x_n)} \overset{\text{(a)}}{=} \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)},$$

where (a) follows from Proposition 6.22, for any sequence (x_n) from the set $\big(U \cap g^{-1}(\mathbb{R} \setminus \{0\})\big) \setminus \{a\}$ that converges to a.

Items (i) and (ii) are proved analogously but in a simpler manner, since one need not worry about technicalities arising from division by zero. □

Since left-sided limits, right-sided limits, and limits at positive and negative infinity are just special cases of the limit of a function at a point (see e.g. Remark 6.20), the rules for computing limits in Proposition 6.24 also hold for one-sided limits and limits at infinity.

In certain cases, a scalar function does not have a limit at a given point $a \in \mathbb{R}$, but the function's value increases or decreases without bound in a predictable manner as the argument approaches a. We denote such cases with the notation in the following remark.

Remark 6.25: Notation for divergent limits

For any scalar function $f\colon U \to \mathbb{R}$ defined on the real subset $U \subset \mathbb{R}$ and point $a \in \mathbb{R}$, we introduce the following notation:

(i) If a is a cluster point of U and if for all $M \in \mathbb{R}$ there exists $\delta \in \mathbb{R}^+$ for which $M \leq f(x)$ for all $x \in (a-\delta, a+\delta) \cap (U \setminus \{a\})$, we write

$$\lim_{x\to a} f(x) = \infty.$$

For analogous situations involving one-sided limits, we use the notation

$$\lim_{x\to a^+} f(x) = \infty \quad \text{and} \quad \lim_{x\to a^-} f(x) = \infty.$$

(ii) If $a \in \mathbb{R}$ is a cluster point of U and if for all $m \in \mathbb{R}$ there exists $\delta \in \mathbb{R}^+$ for which $f(x) \leq m$ for all $x \in (a-\delta, a+\delta) \cap (U \setminus \{a\})$, we write

$$\lim_{x\to a} f(x) = -\infty,$$

or, for analogous cases involving one-sided limits,

$$\lim_{x\to a^+} f(x) = -\infty \quad \text{and} \quad \lim_{x\to a^-} f(x) = -\infty.$$

(iii) If U is unbounded above and if for all $M \in \mathbb{R}$ there exists $B \in \mathbb{R}$ for which $M \leq f(x)$ for all $x \in (B, \infty) \cap U$, we write

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

(iv) If U is unbounded above and if for all $m \in \mathbb{R}$ there exists $B \in \mathbb{R}$ for which $f(x) \leq m$ for all $x \in (B, \infty) \cap U$, we write

$$\lim_{x \to \infty} f(x) = -\infty.$$

(v) If U is unbounded below and if for all $M \in \mathbb{R}$ there exists $b \in \mathbb{R}$ for which $M \leq f(x)$ for all $x \in (-\infty, b) \cap U$, we write

$$\lim_{x \to -\infty} f(x) = \infty.$$

(vi) If U is unbounded below and if for all $m \in \mathbb{R}$ there exists $b \in \mathbb{R}$ for which $f(x) \leq m$ for all $x \in (-\infty, b) \cap U$, we write

$$\lim_{x \to -\infty} f(x) = -\infty.$$

6.2 Continuity

Definition 6.26: Continuity of a scalar function

Let $f \colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$. The function f is said to be *continuous at the point* $a \in U$ if for all $\epsilon \in \mathbb{R}^+$ there exists corresponding $\delta \in \mathbb{R}^+$ for which

$$|f(x) - f(a)| < \epsilon \text{ for all } x \in (a - \delta, a + \delta) \cap U;$$

or, equivalently, for which

$$f(x) \in \big(f(a) - \epsilon, f(a) + \epsilon\big) \text{ for all } x \in (a - \delta, a + \delta) \cap U.$$

The function f is said to be *continuous* if it is continuous at all points in its domain.

The definition of the continuity of a function $f \colon U \to \mathbb{R}$ at a point $a \in \mathbb{R}$ is very similar to the definition of the limit of a function at a point—the main difference is that the definition of continuity depends crucially on f's value at a, while f's value at a is completely irrelevant to the definition of a limit. In fact, if $a \in U$ is a cluster point of the set $U \subset \mathbb{R}$, if follows directly from Definitions 6.16 and 6.26 that a function $f \colon U \to \mathbb{R}$ is continuous at a if, and only if, the function value $f(a)$ equals the function's limit at a.

Note that a function's continuity at a point is well-defined only at points at which the function itself is defined—it does not make sense to speak of a function's continuity at points outside its domain. We illustrate this technicality in the following example.

Example 6.27: Continuity of the sign function

Let us consider the sign (or signum) function $\operatorname{sgn}\colon \mathbb{R} \to \mathbb{R}$, defined as

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0, \\ 0 & x = 0, \\ -1 & x < 0. \end{cases}$$

The sign function is defined at all points on the real line and is continuous at all points except at zero, where its graph has a jump discontinuity.

However, the restricted function $\operatorname{sgn}|_{\mathbb{R}\setminus\{0\}}$ is continuous on its entire domain. Although this restricted function's graph consists of two distinct parts, the function satisfies the definition of continuity at all points in its domain, and hence is continuous; the behavior at zero, at which the restricted function is not defined, is immaterial.

Proposition 6.28: Conditions for continuity

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and let $a \in U$ be an arbitrary point in f's domain.

(i) If a is an isolated point of U, then f is continuous at a.

(ii) If a is a cluster point of U, then f is continuous at a if, and only if

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

(iii) The function f is continuous at a if, and only if,

$$f(a) = \lim_{x \to \infty} f(x_n)$$

for all sequences (x_n) from U that converge to a.

Proof. Although these conditions follow almost directly from the definition of the limit and continuity, we prove them here for the sake of completeness.

(i) If $a \in U$ is an isolated point of the subset U, the condition for continuity of the function $f\colon U \to \mathbb{R}$ at a is fulfilled automatically, since with a small enough choice of δ the set $x \in (a-\delta, a+\delta) \cap U$ can be made to contain only a, and $|f(x) - f(a)| = 0 < \epsilon$ for all $\epsilon \in \mathbb{R}^+$ when $x = a$.

(ii) ($\Rightarrow$) Assume f is continuous at U, which means that for all $\epsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ for which

$$|f(x) - f(a)| < \epsilon \text{ for all } x \in (a-\delta, a+\delta) \cap U \tag{6.2}$$

The definition of a limit uses the same equation as Equation 6.2, but uses the more restricted subset $(a-\delta, a+\delta) \cap (U \setminus \{a\})$. But since $(a-\delta, a+\delta) \cap (U \setminus \{a\})$ is a subset of $(a-\delta, a+\delta) \cap U$, Equation 6.2 certainly holds for $(a-\delta, a+\delta) \cap (U \setminus \{a\})$ as well, and so $f(a)$ is the limit of f at a.

($\Leftarrow$) Now assume that $f(a) = \lim_{x\to a} f(x)$, which means that for all $\epsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ for which

$$|f(a) - f(x)| < \epsilon \text{ for all } x \in (a-\delta, a+\delta) \cap (U \setminus \{a\}).$$

We can then separately verify that, when $x = a$,

$$|f(a) - f(x)| = |f(a) - f(a)| = 0 < \epsilon,$$

which means that

$$|f(a) - f(x)| < \epsilon \text{ for all } x \in (a-\delta, a+\delta) \cap U,$$

and so f is continuous at a.

(iii) ($\Rightarrow$) Assume that f is continuous at a. By point (ii), $f(a)$ is the limit of f at a; point (iii) then follows immediately from Proposition 6.22.

($\Leftarrow$) Assume that $f(a) = \lim_{x\to\infty} f(x_n)$ for all sequences (x_n) from U that converge to a. By Proposition 6.22, $f(a)$ is then the limit of f at a, and so f is continuous at a by point (ii). □

We now make two important statements about interchanging the order of limit and function evaluation that we will regularly make use of in the remainder of this book.

Remark 6.29: Changing the order of limits

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, let $a \in U$ be an arbitrary point in f's domain, and let (x_n) be an arbitrary sequence of points from U that converges to a. The equality in Proposition 6.28 (iii) can then be written in the form

$$f(a) \overset{\text{(a)}}{=} f\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} f(x_n), \tag{6.3}$$

where (a) uses the assumption that $a = \lim_{x\to\infty} x_n$.

Equation 6.3 motivates the following important observation: the order in which a *continuous* function and the limit of a sequence are evaluated can be interchanged. A similar result holds for interchanging the evaluation of a continuous function and the limit of a function, as formalized in the following proposition.

Proposition 6.30: Changing the order of function and limit evaluation

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the subsets $U, V \subset \mathbb{R}$, let $a \in \mathbb{R}$ be a cluster point of the set $g^{-1}(U) \subset \mathbb{R}$, and assume the limit of g at a exists. If $\lim_{x\to a} g(x) \in U$ and if f is continuous at the point $\lim_{x\to a} g(x)$, then limit of the composite function $f \circ g$ also exists at a and obeys

$$\lim_{x\to a} f\big(g(x)\big) = f\left(\lim_{x\to a} g(x)\right).$$

Proof. Since the limit of g at a exists, by Proposition 6.22, for any sequence (x_n) from the set $g^{-1}(U) \setminus \{a\}$ that converges to a, the corresponding sequence of function

values $\big(g(x_n)\big)_n$ must converge to the limit $\lim_{x\to a} g(x)$.

Because the function f is continuous at the point $\lim_{x\to a} g(x) \in U$, and because we have already shown that the sequence $\big(g(x_n)\big)_n$ converges to the limit $\lim_{x\to a} g(x)$, by Proposition 6.28 (iii),

$$f\left(\lim_{x\to a} g(x)\right) = \lim_{n\to\infty} f\big(g(x_n)\big).$$

We have shown that for all sequences $\big(g(x_n)\big)_n$ from the set U that converge to $\lim_{x\to a} g(x)$, the corresponding sequence of function values $\big(f\big(g(x_n)\big)\big)_n$ converges to $f\big(\lim_{x\to a} g(x)\big)$. Thus, by Proposition 6.22, the number $f\big(\lim_{x\to a} g(x)\big)$ is the limit of the function $f(g(x))$ as x approaches a, i.e.

$$\lim_{x\to a} f\big(g(x)\big) = f\left(\lim_{x\to a} g(x)\right). \qquad \square$$

Example 6.31: Rearranging limits and function evaluation

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the subsets $U, V \subset \mathbb{R}$, let $a \in \mathbb{R}$ be a cluster point of the set $g^{-1}(U) \subset \mathbb{R}$, and let $b \in \mathbb{R} \setminus U$ be the limit of g at the point a.

In this case, $b = \lim_{x\to a} g(x)$ is a cluster point of f's domain U. To verify this, note that, because a is assumed to be a cluster point of $g^{-1}(U)$, by Proposition 6.14 there exists a sequence (x_n) from the set $g^{-1}(U) \setminus \{a\}$ that converges to a. Because b is limit of g at the point a, the sequence of function values $\big(g(x_n)\big)_n$ converges to b by Proposition 6.22. Because the sequence $\big(g(x_n)\big)_n$ converges to b, and because

$$g(x_n) \in U \quad \text{and} \quad g(x_n) \neq b \text{ for all } n \in \mathbb{N},$$

the point b is a cluster point of U by Proposition 6.14.

Next, we note that if the limit of f at b exists, the limit of the composite function $f \circ g$ also exists at a and obeys

$$\lim_{x\to a} f\big(g(x)\big) = \lim_{t\to b} f(t).$$

To verify this, we first define the function $h\colon U \cup \{b\} \to \mathbb{R}$ given by

$$h|_U = f \quad \text{and} \quad h(b) = \lim_{t\to b} f(t).$$

The thus-construction function h is continuous at b, and so the limit $\lim_{x\to a} h(g(x))$ exists by Proposition 6.30. More so, the number $\lim_{x\to a} h(g(x))$ is also the limit of the function $f \circ g$ at a, which follows from the calculation

$$\lim_{x\to a} f\big(g(x)\big) \overset{\text{(a)}}{=} \lim_{x\to a} h\big(g(x)\big) \overset{\text{(b)}}{=} h\left(\lim_{x\to a} g(x)\right) \overset{\text{(c)}}{=} h(b) \overset{\text{(d)}}{=} \lim_{t\to b} f(t),$$

where (a) follows from the definition of h, (b) from Proposition 6.30, (c) from the definition $b = \lim_{x\to a} g(x)$, and (d) again from the definition of h.

We will now examine the behavior of continuity under common function operations.

Corollary 6.32: Continuity under function addition, multiplication, and division

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the real subsets $U, V \subset \mathbb{R}$, and define the constant $a \in U \cap V$.

(i) If f and g are continuous at a, then the function $f + g$ is also continuous at a.

(ii) If f and g are continuous at a, then the function $f \cdot g$ is also continuous at a.

(iii) If f and g are continuous at a and $g(a) \neq 0$, then f/g is also continuous at a.

Proof. This corollary follows directly from the rules for computing limits of functions in Proposition 6.24 together with the conditions for continuity in Proposition 6.28. □

Corollary 6.33: Continuity under function composition

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the real subsets $U, V \subset \mathbb{R}$, and define the constant $a \in g^{-1}(U)$. If g is continuous at a and f is continuous at $g(a)$, then the composite function $f \circ g$ is also continuous at a.

Proof. Let (x_n) be an arbitrary sequence of points from the set $g^{-1}(U)$ that converges to a. Because g is assumed to be continuous at a, the sequence $\big(g(x_n)\big)_n$ converges to $g(a) \in U$ by Proposition 6.28 (iii). Similarly, because f is assumed to be continuous at $g(a)$, the sequence $\big(f(g(x_n))\big)_n$ converges to $f(g(a))$, and so the function $f \circ g$ is continuous at a, again by Proposition 6.28 (iii). □

We conclude this section with a discussion of continuity in the context of common functions.

Example 6.34: Common continuous functions

(1) The constant function $f\colon \mathbb{R} \to \mathbb{R}$ with constant value c (see Example 6.6 (3)) is continuous, since, for all $\epsilon \in \mathbb{R}^+$,

$$|f(x) - f(a)| = |c - c| = 0 < \epsilon \text{ for all } a, x \in \mathbb{R}.$$

(2) Any linear function $f\colon \mathbb{R} \to \mathbb{R}$ of the form $f(x) = kx + n$ (Example 6.6 (4)) is continuous for all $k, n \in \mathbb{R}$. This holds trivially for the case $k = 0$, in which case f is a constant function and is continuous by point (1). Now assume $k \neq 0$, choose arbitrary $\epsilon \in \mathbb{R}^+$, and define $\delta = \epsilon/|k|$. For all $a, x \in \mathbb{R}$ for which $|x - a| < \delta$, f's continuity follows from the calculation

$$\begin{aligned} |f(x) - f(a)| &= |kx + n - (ka + n)| = |kx - ka| \\ &= |k||x - a| < |k|\delta = \epsilon. \end{aligned}$$

(3) The restriction of a continuous function to an arbitrary subset of its domain is again a continuous function. In symbols, let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and let $V \subset U$ be a subset of U; if f is continuous at $a \in V$, then $f|_V$ is evidently also continuous at a by the definition of continuity (Definition 6.26) and function restriction (Definition (ii)).

(4) The power function $f\colon \mathbb{R} \to \mathbb{R}, f(x) = x^n$ is continuous for all $n \in \mathbb{N}$. To show this, we write the function in the form

$$f(x) = x \cdot x \cdot \underbrace{\cdots}_{n} \cdot x$$

and then note that the identity function $x \mapsto x$ is a linear function and continuous by point (2), and the product of continuous functions is again continuous by Corollary 6.32. The number n is called the power function's *power* or *exponent*.

(5) A real polynomial of degree $n \in \mathbb{N} \cup \{0\}$ is a scalar function for the form

$$P\colon \mathbb{R} \to \mathbb{R}, \quad P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

for some constants $a_0, a_1, \ldots, a_n \in \mathbb{R}$, $a_n \neq 0$. A polynomial is thus a sum of the products of power functions and constants, i.e. the sum of the products of continuous functions, and is thus continuous by Corollary 6.32.

(6) The rational function is quotient of two polynomials, and is thus continuous by point (5) and Corollary 6.32.

6.3 Properties of continuous functions

In this section we cover some important theorems describing the behavior of continuous scalar functions—many of these theorems form the foundation of real analysis.

Theorem 6.35: Zeros of continuous functions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a, b] \to \mathbb{R}$ be a continuous function. If $f(a)f(b) < 0$ (i.e. if $f(a)$ and $f(b)$ have opposite signs) there exists at least one point $\xi \in (a, b)$ for which $f(\xi) = 0$.

Loosely, the proposition states that if a continuous function's value changes in sign between two points on the x axis, the function's graph must have crossed the x axis at least once somewhere between those two points.

Proof. We will prove the proposition by construction, using the bisection method to find a zero of f between the points a and b. If the process produces a zero of f in finitely many iterations, the proof is complete; otherwise, we will inductively construct two real sequences (a_n) and (b_n) with the following properties:

(i) The sequence (a_n) is increasing and the sequence (b_n) is decreasing.

(ii) The general elements a_n and b_n satisfy the inequalities

$$a \leq a_n < b_n \leq b \quad \text{and} \quad b_n - a_n = (b - a)/2^{n-1} \text{ for all } n \in \mathbb{N}.$$

In words, a is a lower bound of (a_n), b is an upper bound of (b_n), all terms in (b_n) are greater than the corresponding term in (a_n), and the distance between a_n and b_n decreases by a factor of 2 with each increasing value of n.

(iii) The general terms a_n and b_n are such that

$$f(a_n)f(b_n) < 0 \text{ for all } n \in \mathbb{N},$$

i.e. f's values at a_n and b_n have opposite sign for all n.

We first define $a_1 = a$ and $b_1 = b$. Assume we have already constructed the terms $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ so as to satisfy the above properties, and define

$$c_n = (a_n + b_n)/2.$$

If $f(c_n) = 0$, we have found a zero $\xi = c_n$ of the function f in the interval (a, b), completing the proof. If $f(c_n) \neq 0$, we construct the next sequence terms as follows:

(1) If $f(a_n)f(c_n) < 0$, let $a_{n+1} = a_n$ and $b_{n+1} = c_n$, in which case

$$f(a_{n+1})f(b_{n+1}) = f(a_n)f(c_n) < 0.$$

(2) If $f(a_n)f(c_n) > 0$, let $a_{n+1} = c_n$ and $b_{n+1} = b_n$, in which case

$$f(a_n)^2 f(c_n)f(b_n) = \big[f(a_n)f(c_n)\big]\big[f(a_n)f(b_n)\big] < 0,$$

and so again $f(a_{n+1})f(b_{n+1}) = f(c_n)f(b_n) < 0$.

By construction, the sequence (a_n) is increasing and bounded above by b, and is thus convergent by Proposition 4.15. Similarly, (b_n) is decreasing and bounded below by a, and is thus also convergent by Proposition 4.15. We then make the calculation

$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n \overset{\text{(a)}}{=} \lim_{n\to\infty} (b_n - a_n) \overset{\text{(b)}}{=} \lim_{n\to\infty} \frac{b-a}{2^{n-1}} = 0,$$

where (a) follows from the limit rule in Proposition 4.17 (i) and (b) from the definition of the sequences (a_n) and (b_n). The result implies that $\lim_{n\to\infty} b_n = \lim_{n\to\infty} a_n$, i.e. that (a_n) and (b_n) have the same limit, which we denote by

$$\xi = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

It remains to show that ξ is a zero of the function f. Using the assumption that f is continuous, we make the calculation

$$\begin{aligned} f(\xi)^2 = f(\xi)f(\xi) &= f\left(\lim_{n\to\infty} a_n\right) \cdot f\left(\lim_{n\to\infty} b_n\right) \\ &\overset{\text{(a)}}{=} \lim_{n\to\infty} f(a_n) \cdot \lim_{n\to\infty} f(b_n) \\ &\overset{\text{(b)}}{=} \lim_{n\to\infty} \big[f(a_n)f(b_n)\big], \end{aligned}$$

where (a) follows from the continuity condition in Proposition 6.28 (iii) and (b) from the rules for computing limits in Proposition 6.24 (ii). Since $f(a_n)f(b_n) < 0$ for all $n \in \mathbb{N}$ by the definition of the sequences (a_n) and (b_n), it follows that

$$\lim_{n\to\infty} f(a_n)f(b_n) = f(\xi)^2 \leq 0.$$

Since f is a real-valued function, the result $f(\xi)^2 \leq 0$ implies that $f(\xi) = 0$. □

Theorem 6.36: Continuous functions are bounded on closed intervals

Every continuous scalar function $f\colon U \to \mathbb{R}$ defined on a bounded, closed subset $U \subset \mathbb{R}$ is itself bounded.

Proof. We will first prove by contradiction that f is bounded above. Assume f is not bounded above, which would mean that for every natural number $n \in \mathbb{N}$ there exists a point $x_n \in U$ for which

$$f(x_n) \geq n. \tag{6.4}$$

Because the set U is bounded, the sequence (x_n), which consists only of points from U, must also be bounded. Because the sequence (x_n) is bounded, it must have at least one cluster point by the Bolzano-Weierstrass theorem (Theorem 4.9); we denote this cluster point by c, and note that, because U is closed, c must be an element of U. Because the function f is continuous at c, there exists $\delta \in \mathbb{R}^+$ for which

$$|f(x) - f(c)| < 1 \text{ for all } x \in (c-\delta, c+\delta).$$

The number c is a cluster point of (x_n), so every neighborhood of c contains infinitely many elements of (x_n) by Definition 4.6. In particular, this means there exists a natural number $n \in \mathbb{N}$ for which

$$n > f(c) + 1 \quad \text{and} \quad x_n \in (c-\delta, c+\delta).$$

Because $x_n \in (c-\delta, c+\delta)$ and because f is continuous at c, it follows that

$$|f(x_n) - f(c)| < 1.$$

But at the same time, assuming f is not bounded above, it holds that

$$f(x_n) - f(c) \overset{\text{(a)}}{\geq} n - f(c) \overset{\text{(b)}}{>} f(c) + 1 - f(c) = 1,$$

where (a) follows from Equation 6.4 and (b) from the earlier conclusion that $n > f(c) + 1$; the result $f(x_n) - f(c) > 1$ contradicts the fact that $|f(x_n) - f(c)| < 1$ under the assumption that f is continuous at c. It follows that f must be bounded above.

It remains to show that f is bounded below. To do so, note that the function $-f$, as the product of a constant and a continuous function, is also continuous. Since $-f$ is continuous, we can repeat an analogous procedure to that used above to show that $-f$ is also bounded above, which means that f is bounded below. The function f is then bounded above and below, and is thus bounded. □

Theorem 6.37: Continuous functions attain a minimum and maximum on closed intervals

Let $f\colon U \to \mathbb{R}$ be a continuous scalar function defined on a nonempty, bounded, closed subset $U \subset \mathbb{R}$. In this case there exist points $u, v \in U$ for which

$$f(u) = \sup(f) = \max(f) \quad \text{and} \quad f(v) = \inf(f) = \min(f).$$

In words, every continuous scalar function attains a maximum and minimum value on a closed, bounded, nonempty real subset.

Proof. We first note that the function f is bounded by Theorem 6.36, and so it must have a supremum, which we denote by $M = \sup(f)$. We then define the continuous, nonnegative function

$$g\colon U \to \mathbb{R}, \quad g(x) = M - f(x).$$

Our plan is to show, by contradiction, that the function g has at least one zero $u \in U$; in this case $f(u) = M$, meaning that f attains its supremum and thus has a maximum on U.

Assume that g does not have any zeros. In this case, by Corollary 6.32, the function $1/g$ is continuous and defined on the entire set U, and is thus bounded by Theorem 6.36. Because g is bounded and nonnegative, there exists a positive real number $A \in \mathbb{R}$ for which

$$1/g(x) = 1/[M - f(x)] \leq A \text{ for all } x \in U,$$

which in turn implies the bound

$$f(x) \leq M - \frac{1}{A}.$$

This means that $M - 1/A$ is an upper bound of the function image $f(U)$, which contradicts the earlier assumption that M is the supremum (i.e. *least* upper bound) of $f(U)$. It follows by contradiction that g must have at least one zero $u \in U$, in which case

$$g(u) = 0 = M - f(u) \implies f(u) = M = \sup(f).$$

It follows that $f(x)$ attains its supremum on U, and thus has a maximum on U.

We could then repeat an analogous argument for the function $-f$, which is also continuous as the product of a constant and a continuous function. We would find that there exists $v \in U$ for which $-f(v) = \sup(-f) = -\inf(f)$, meaning that f also attains a minimum on U, completing the proof. □

Proposition 6.38: Continuous functions on closed intervals attain all values between their minimum and maximum

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, and let $f\colon [a,b] \to \mathbb{R}$ be a continuous scalar function. In this case

$$f([a,b]) = \big[\inf(f), \sup(f)\big].$$

In other words, f attains both a minimum and maximum value on $[a,b]$ and assumes precisely all values between its minimum and maximum.

Proof. We first note that $f([a,b]) \subset [\inf(f), \sup(f)]$ by the definition of a function's image and its infimum and supremum. It remains to show that $f([a,b])$ contains *all* values in the interval $[\inf(f), \sup(f)]$. By Theorem 6.37, there exist points $u, v \in [a,b]$ for which

$$f(u) = \sup(f) \quad \text{and} \quad f(v) = \inf(f),$$

which means that $\inf(f), \sup(f) \in f([a,b])$. We then choose an real number $A \in \mathbb{R}$ for which $\inf(f) < A < \sup(f)$ and define the continuous function

$$g\colon [a,b] \to \mathbb{R}, \quad g(x) = f(x) - A.$$

We then make the calculation

$$\begin{aligned} g(u)g(v) &\overset{\text{(a)}}{=} \big[f(u) - A\big]\big[f(v) - A\big] \\ &= \big[\sup(f) - A\big]\big[\inf(f) - A\big] \overset{\text{(b)}}{<} 0, \end{aligned}$$

where (a) follows from the definition of the function g and (b) from the requirement that $\inf(f) < A < \sup(f)$. Since g is continuous, by Theorem 6.35 there exists a point $x \in (u,v)$ for which $g(x) = 0$ and thus $f(x) = A$. But because A was chosen arbitrarily the interval $(\inf(f), \sup(f))$, it follows that f must attain all values in the open interval $(\inf(f), \sup(f))$. Since we have already shown that $\inf(f), \sup(f) \in f([a,b])$, it follows that f also attains all values in the *closed* interval $[\inf(f), \sup(f)]$, proving the proposition. □

Proposition 6.38 implies that for any continuous function $f\colon [a,b] \to \mathbb{R}$, where $a, b \in \mathbb{R}$ and $a \leq b$, the equation $f(x) = A$ has a solution if, and only if, $A \in [\inf(f), \sup(f)]$.

Corollary 6.39: Continuity conditions for monotonic functions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$. In this case:

(i) Any increasing function $f\colon [a,b] \to \mathbb{R}$ is continuous if, and only if,

$$f([a,b]) = [f(a), f(b)].$$

(ii) Any decreasing function $g\colon [a,b] \to \mathbb{R}$ is continuous if, and only if,

$$g([a,b]) = [g(b), g(a)].$$

Proof. Both points follow from Proposition 6.38 and Example 6.21. □

Corollary 6.40: The image of a continuous function is an interval

If $f\colon U \to \mathbb{R}$ is a continuous scalar function defined on a real interval $U \subset \mathbb{R}$, then f's image $f(U)$ is also a real interval.

Proof. Because f's domain U is an interval, for any two points $a, b \in U$ for which $a < b$ it holds that $[a,b] \subset U$. By Proposition 6.38, the continuous function f must take on all values between $f(a)$ and $f(b)$ on the interval $[a,b]$. It follows that for any two points in the image $f(U)$, the entire interval between the two points is contained in $f(U)$; because these two points were chosen arbitrarily, the image $f(U)$ must be an interval. □

Corollary 6.40 can be used to show the existence of the root of nonnegative real numbers.

Corollary 6.41: Roots of nonnegative real numbers

Let $n \in \mathbb{N}$ be an arbitrary natural number and let $A \in \mathbb{R}$ be any nonnegative real number. In this case there exists exactly one nonnegative real number $x \in \mathbb{R}$ for which

$$x^n = A;$$

we denote this number by $\sqrt[n]{A}$ and call it the *n-th root* of A.

Proof. The power function $f\colon [0,\infty) \to \mathbb{R}$ defined as

$$f(x) = x^n$$

is continuous (see Example 6.34 (4)), strictly increasing, and unbounded above. By Corollary 6.39, f's image is the interval $[0,\infty)$, so for each $A \in [0,\infty)$ there must exist some $x \in [0,\infty)$ for which $f(x) = A$. And because f is strictly increasing, it is injective by Proposition 6.12, and so the solution to the equation $f(x) = A$ is unique. □

Remark 6.42: Solutions of power equations

Let $A \in \mathbb{R}$ be an arbitrary real number.

(1) Let $n \in \mathbb{N}$ be an even natural number. If $A \geq 0$, the equation $x^n = A$ has two real solutions; these are $\sqrt[n]{A}$ and $-\sqrt[n]{A}$. (Note that this does not contradict the uniqueness statement in Corollary 6.41, which applies only to nonnegative solutions of $x^n = A$.)

If $A < 0$, the equation $x^n = A$ does not have any real solutions.

For even n, the n-th root function $[0,\infty) \to \mathbb{R}$ mapping $t \mapsto \sqrt[n]{t}$ is strictly increasing, and its image is the interval $[0,\infty) \to \mathbb{R}$.

(2) Let n be an odd natural number. In this case the equation $x^n = A$ has a single real solution. If $A \geq 0$, the solution is the nonnegative real number $\sqrt[n]{A}$; if $A < 0$, the solution is the negative real number $-\sqrt[n]{-A}$, in which case we write

$$-\sqrt[n]{-A} = \sqrt[n]{A}.$$

For odd n, the n-th root function $\mathbb{R} \to \mathbb{R}$ mapping $t \mapsto \sqrt[n]{t}$ is a strictly increasing bijection.

Proposition 6.43: Growth and invertibility of continuous, injective functions

If $f\colon U \to \mathbb{R}$ is a continuous scalar function defined on the real subset $U \to \mathbb{R}$, then:

(i) The function f is injective if, and only if, it is strictly monotonic.

(ii) If f is injective, then its inverse function $f^{-1}\colon f(U) \to \mathbb{R}$ is continuous.

Proof.

(i) If f is injective, it is also strictly monotonic by Proposition 6.38, since f being

monotonic is the only way it can be both injective and satisfy $f([a,b]) = [\inf(f), \sup(f)]$.

(ii) The function f must be strictly monotonic by item (i), and so its inverse f^{-1} is also a strictly monotonic function by Proposition 6.12. More so, the function image $f(U)$ is an interval by Corollary 6.40.

Assume that f is strictly increasing, and choose two arbitrary points $u, v \in f(U)$ for which $u < v$. Because f is continuous, Corollary 6.39 implies that

$$f\big(f^{-1}(u), f^{-1}(v)\big) = \Big[f\big(f^{-1}(u)\big), f\big(f^{-1}(v)\big)\Big] = [u, v],$$

which in turn implies that

$$f^{-1}\big([u,v]\big) = \big[f^{-1}(u), f^{-1}(v)\big].$$

Corollary 6.39 then implies that the function f^{-1} is continuous on the interval $[u, v]$. Since this result holds for all $u, v \in f(U)$ for which $u < v$, the inverse function $f^{-1}\colon f(U) \to \mathbb{R}$ must be continuous.

We could then apply an analogous argument to the case in which f is strictly decreasing, and again conclude that f^{-1} is continuous. □

6.4 Uniform continuity

Definition 6.44: Uniform continuity

A scalar function $f\colon U \to \mathbb{R}$ defined on the real subset $U \subset \mathbb{R}$ is said to be *uniformly continuous* if for all $\epsilon \in \mathbb{R}^+$ there exists corresponding $\delta \in \mathbb{R}^+$ for which

$$|f(x) - f(y)| < \epsilon \text{ for all } x, y \in U \text{ for which } |x - y| < \delta.$$

Note that continuity (Definition 6.26) is defined at individual points, while uniform continuity is defined for a function's entire domain. In the definition of continuity at a point $a \in \mathbb{R}$, the value of δ may depend on both the choice of ϵ and the choice of a. In the definition of uniform continuity, the value of δ can depend only on the choice of ϵ.

Every uniformly continuous function is also continuous, while the converse is not true in general—continuous functions are not necessarily uniformly continuous, as we show in the following example.

Example 6.45: A continuous function that is not uniformly continuous

The rational function $f(x) = 1/x$, defined on the domain $\mathbb{R} \setminus \{0\}$, is continuous but not uniformly continuous.

Specifically, the rational function is not uniformly continuous because of its two vertical asymptotes as x approaches zero from the right and left. As the function's graph becomes steeper and steeper, for any fixed ϵ, the definition of uniform continuity would require ever smaller values of δ as x approaches 0; in fact there is no $\epsilon \in \mathbb{R}^+$ for which there exists a fixed value of δ for which the graph of f above an interval of

length 2δ would fall in a rectangle of width 2ϵ.

The above example—of a continuous function that is not uniformly continuous—arose because the rational function was defined on an open domain. On closed and bounded domains, every continuous function is also uniformly continuous.

Proposition 6.46: Uniform continuity on closed, bounded domains

Every continuous scalar function $f\colon U \to \mathbb{R}$ defined on a closed, bounded subset $U \subset \mathbb{R}$ is uniformly continuous.

Proof. We will prove the proposition by contradiction. Assume that a continuous function f with a closed, bounded domain $U \subset \mathbb{R}$ is *not* uniformly continuous. In this case there would exist $\epsilon \in \mathbb{R}^+$ for which, for all $n \in \mathbb{N}$, there would exist points $x_n, y_n \in U$ for which

$$|x_n - y_n| < \frac{1}{n} \quad \text{and} \quad |f(x_n) - f(y_n)| \geq \epsilon. \tag{6.5}$$

Any such sequence of points (x_n) is bounded because it comes from the bounded set U, and so by Proposition 4.9 (x_n) has at least one cluster point; we denote this cluster point by c and note that, because U is closed and all elements of the sequence (x_n) lie in U, the cluster point c must also lie in U.

Because f is continuous at c, there exists $\delta \in \mathbb{R}^+$ for which

$$|f(x) - f(c)| < \frac{\epsilon}{2} \text{ for all } x \in (c-\delta, c+\delta) \cap U. \tag{6.6}$$

Because c is a cluster point of x_n, by Proposition 4.7 there exists $n \in \mathbb{N}$ for which

$$n \geq \frac{2}{\delta} \quad \text{and} \quad |x_n - c| < \frac{\delta}{2}. \tag{6.7}$$

We then make the calculation

$$\begin{aligned} |y_n - c| &= |y_n - x_n + x_n - c| \\ &\overset{\text{(a)}}{\leq} |y_n - x_n| + |x_n - c| \\ &\overset{\text{(b)}}{<} \frac{1}{n} + \frac{\delta}{2} \overset{\text{(c)}}{\leq} \delta. \end{aligned}$$

where (a) follows from the triangle inequality, (b) from Equations 6.5 and 6.7, and (c) from the bound $n \geq 2/\delta$. The points x_n and y_n are thus both separated from c by less than δ, so the continuity of f at c in Equation 6.6 implies that

$$|f(x_n) - f(c)| < \frac{\epsilon}{2} \quad \text{and} \quad |f(y_n) - f(c)| < \frac{\epsilon}{2}.$$

We then conclude the proof with the calculation

$$\begin{aligned} |f(x_n) - f(y_n)| &= |f(x_n) - f(c) + f(c) - f(y_n)| \\ &\leq |f(x_n) - f(c)| + |f(c) - f(y_n)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

The result $|f(x_n) - f(y_n)| < \epsilon$ contradicts Equation 6.5, and so f must be uniformly continuous. □

6.5 The exponential function

In this section we will formally define the exponential function and explore some of its important properties. We will proceed in small steps, beginning with the exponential function with natural-number powers and slowly progressing toward the exponential function with real-valued exponents.

6.5.1 The exponential function for rational exponents

In this section we will progress from natural number to integer to rational-value exponents.

Topic 6.1: The exponential function with natural-valued exponents

For review from Definition 2.6, for arbitrary $a \in \mathbb{R}$ and $n \in \mathbb{N}$ we define $a^1 = a$ and then recursively define $a^{n+1} = a^n a$. The number $a^n \in \mathbb{R}$ is defined as the product of n copies of a, i.e.

$$a^n = \underbrace{a \cdot a \cdot \cdots \cdot a}_{n}$$

It then follows from the properties of real multiplication that

$$a^n a^m = a^{n+m} \tag{6.8a}$$
$$a^n b^n = (ab)^n \tag{6.8b}$$
$$(a^n)^m = a^{nm} \tag{6.8c}$$

for all $a, b \in \mathbb{R}$ and all $n, m \in \mathbb{N}$.

Topic 6.2: The exponential function with integer-valued exponents

Again consider a real number $a \in \mathbb{R}$, and assume $a \neq 0$. We first must define the number a^0. Equation 6.8a holds for the zero exponent if $a^n a^0 = a^{n+0} = a^n$, which motivates the definition

$$a^0 = 1.$$

Note that defining $a^0 = 1$ also satisfies the remaining properties in Equations 6.8.

We now aim to define a^{-n} for $n \in \mathbb{N}$, i.e. a real number raised to a negative integer power. Equation 6.8a holds for the negative powers if

$$a^n a^{-n} = a^{n-n} = a^0 = 1,$$

which we solve for a^{-n} to get

$$a^{-n} = \frac{1}{a^n};$$

note that the definition $a^{-n} = a^n$ is also consistent with Property 6.8c.

We have now defined the number a^k for all $a \in \mathbb{R} \setminus \{0\}$ and all $k \in \mathbb{Z}$ such that the following properties hold for all $a, b \in \mathbb{R} \setminus \{0\}$ and all $k, l \in \mathbb{Z}$.

$$a^k a^m = a^{k+l} \tag{6.9a}$$
$$a^k b^k = (ab)^k \tag{6.9b}$$
$$(a^k)^l = a^{kl}. \tag{6.9c}$$

We then combine Equations 6.9 with the definition of the n-th root in Corollary 6.41 to deduce the following properties, which hold for all $a, b \in \mathbb{R}^+$, all $m, n \in \mathbb{N}$, and all $k \in \mathbb{Z}$:

$$\sqrt[n]{a}\,\sqrt[m]{a} = \sqrt[nm]{a^{n+m}} \tag{6.10a}$$

$$\sqrt[n]{ab} = \sqrt[n]{a}\,\sqrt[n]{b} \tag{6.10b}$$

$$\sqrt[n]{\sqrt[m]{a}} = \sqrt[nm]{a} \tag{6.10c}$$

$$\sqrt[n]{a^k} = \left(\sqrt[n]{a}\right)^k. \tag{6.10d}$$

We now aim to expand the definition of powers so that the properties in Equations 6.9 hold for rational exponents. Concretely, we wish, for arbitrary $a \in \mathbb{R}^+$ and $n \in \mathbb{N}$, to define the power $1/n$ and the number $a^{1/n}$ in a way that satisfies Equations 6.9.

Topic 6.3: The exponential function with rational-valued exponents

Let us choose any nonnegative real number $a \in \mathbb{R}^+$, and note that Equation 6.9c holds only if

$$\left(a^{1/n}\right)^n = a^{(1/n)n} = a^1 = a.$$

Recalling the definition of the n-th root, the result $(a^{1/n})^n = a$ motivates the definition

$$a^{1/n} = \sqrt[n]{a}.$$

New let $r \in \mathbb{Q}$ be an arbitrary rational number, which we will write in the form $r = p/q$ where $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. Equation 6.9c holds for the exponent p/q if

$$\left(a^{p/q}\right)^q = a^{(p/q)q} = a^p,$$

which we then solve for $a^{p/q}$ to get

$$a^{p/q} = \sqrt[q]{a^p} \overset{\text{(a)}}{=} \left(\sqrt[q]{a}\right)^p,$$

where (a) uses Equation 6.10d.

There is a slight complication: a rational number's fraction representation is not unique. Namely, the above derivation relied on writing the rational number r in the form $r = p/q$, but since this representation is not unique, we must confirm that the above definition of rational-valued exponents is well-defined. To verify this, suppose

$$r = \frac{p}{q} = \frac{p'}{q'}$$

for some $p' \in \mathbb{Z}$ and $q' \in \mathbb{N}$. We rearrange to get $pq' = p'q$, and then make the calculation

$$\begin{aligned}\left[\left(\sqrt[q]{a}\right)^p\right]^{qq'} &\overset{\text{(a)}}{=} \left(\sqrt[q]{a}\right)^{pqq'} = \left(\sqrt[q]{a}\right)^{qp'q} \overset{\text{(b)}}{=} a^{p'q} \\ &\overset{\text{(c)}}{=} \left(\sqrt[q']{a}\right)^{q'p'q} \\ &\overset{\text{(d)}}{=} \left[\left(\sqrt[q']{a}\right)^{p'}\right]^{qq'}.\end{aligned}$$

where (a) follows from Equation 6.9c, (b) and (c) from Equation 6.9c and the definition of the n-th root, and (d) again from Equation 6.9c. We then compare the beginning and end

of the equality to conclude that

$$\left(\sqrt[q]{a}\right)^p = \left(\sqrt[q']{a}\right)^{p'} \text{ if } \frac{p}{q} = \frac{p'}{q'},$$

which means the definition of a rational power as $a^r = a^{p/q}$ is well-defined for any $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ for which $r = p/q$.

We may now extend the properties in Equations 6.9 to apply to rational powers. Namely, by Equations 6.9 and 6.10 and the definition of the exponential function for rational powers, the following properties hold for all $a, b \in \mathbb{R}^+$ and all $r, s \in \mathbb{Q}$:

$$\begin{aligned} a^r a^s &= a^{r+s} \\ a^r b^r &= (ab)^r \\ (a^r)^s &= a^{rs}. \end{aligned}$$

Definition 6.47: The exponential function with rational exponents

For all $a \in \mathbb{R}^+$, an exponential function with power $r \in \mathbb{Q}$ is a function of the form

$$\phi_a \colon \mathbb{Q} \to \mathbb{R}, \quad \phi_a(r) = a^r.$$

This function is positive and monotonic for all $r \in \mathbb{Q}$, and obeys the following properties:

(i) If $a = 1$, then ϕ_a is constant with the value 1.

(ii) If $a > 1$, then ϕ_a is strictly increasing.

(iii) If $0 < a < 1$, then ϕ_a is strictly decreasing.

The function ϕ_a is also continuous, as we now show.

Proposition 6.48: Continuity of the rational exponential function

The exponential function $\phi_a \colon \mathbb{Q} \to \mathbb{R}$ is continuous for all $a \in \mathbb{R}^+$ and $r \in \mathbb{Q}$.

Proof. If $a = 1$, the function ϕ_a is constant and thus continuous by Example 6.34 (1).

Now assume that $a > 1$. We will first show that ϕ_a is continuous at the point 0. To do so, we choose arbitrary $\epsilon \in \mathbb{R}^+$ and note that

$$\lim_{n\to\infty} \sqrt[n]{a} = \lim_{n\to\infty} a^{1/n} = 1 = \lim_{n\to\infty} (1/a)^{1/n} = \lim_{n\to\infty} \sqrt[n]{1/a},$$

which means there exists $N \in \mathbb{N}$ for which

$$\left|\sqrt[n]{a} - 1\right| < \epsilon \quad \text{and} \quad \left|\sqrt[n]{1/a} - 1\right| < \epsilon \text{ for all } n \geq N.$$

The function ϕ_a is strictly increasing, so for any $h \in \mathbb{Q} \cap [0, 1/N]$ it holds that

$$\left|a^h - 1\right| \leq \left|a^{1/N} - 1\right| = \left|\sqrt[N]{a} - 1\right| < \epsilon. \tag{6.11}$$

Similarly, for any $h \in \mathbb{Q} \cap [-1/N, 0]$, it holds that

$$\left|a^h - 1\right| \leq \left|a^{-1/N} - 1\right| = \left|\sqrt[N]{1/a} - 1\right| < \epsilon. \tag{6.12}$$

We then combine the result of Equations 6.11 and 6.12 to conclude that

$$\left|a^h - 1\right| < \epsilon \text{ for all } h \in \mathbb{Q} \cap [-1/N, 1/N],$$

which means that ϕ_a is continuous at the point 0.

We will now show that when $a > 1$, the function ϕ_a is continuous at all points $s \in \mathbb{Q}$. Again choose arbitrary $\epsilon \in \mathbb{R}^+$. Because ϕ_a is continuous at 0, there exists $\delta \in \mathbb{R}^+$ for which

$$\left|a^h - 1\right| < \frac{\epsilon}{a^s} \text{ for all } h \in \mathbb{Q} \cap (-\delta, \delta),$$

which in turn implies that

$$\left|a^{s+h} - a^s\right| = a^s\left|a^h - 1\right| < a^s \frac{\epsilon}{a^s} = \epsilon.$$

In other words,

$$|a^r - a^s| < \epsilon \text{ for all } r \in \mathbb{Q} \cap (s - \delta, s + \delta),$$

and so ϕ_a is continuous at the point s.

We now consider the case $0 < a < 1$. In this case we may write

$$\phi_a(r) = a^r = (1/a)^{-r} = \phi_{1/a}(-r).$$

We have written ϕ_a as the composition of two functions: a linear function $r \mapsto -r$ and the exponential function $\phi_{1/a}$ with base $1/a > 1$. Both of these functions are continuous, and so the composition ϕ_a with $0 < a < 1$ is also continuous by Corollary 6.33. □

6.5.2 The exponential function for real-valued exponents

We will now extend the definition of the exponential function to arbitrary real-valued exponents. In other words, for arbitrary base $a \in \mathbb{R}$, we wish to define a continuous function

$$f_a : \mathbb{R} \to \mathbb{R}, \quad f(x) = a^x.$$

We will do this by constructing a function for which, for every sequence of rational numbers (r_n) that converges to a real number x, the sequence $(a^{r_n})_n$ converges to a^x.

Proposition 6.49: Sequences of rational exponents

Let $a \in \mathbb{R}^+$ be any nonnegative real number.

(i) If (r_n) is a sequence of rational numbers that converges to any real number, then the sequence of rational numbers $(a^{r_n})_n$ also converges.

(ii) If (r_n) and (s_n) are two rational sequences that both converge to the same real number, then

$$\lim_{n\to\infty}(a^{r_n}) = \lim_{n\to\infty}(a^{s_n}).$$

Proof.

(i) Because the rational sequence (r_n) converges, it is also bounded, so there exist

rational numbers $p, q \in \mathbb{Q}$ for which

$$p \leq r_n \leq q \text{ for all } n \in \mathbb{N}.$$

Define $M = \max\{a^p, a^q\}$ and choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the function

$$\phi_a \colon \mathbb{Q} \to \mathbb{R}, r \mapsto a^r$$

is monotonic, $a^{r_n} \leq M$ for all $n \in \mathbb{N}$. Because ϕ_a is continuous at the point 0 by Proposition 6.48, there exists $\delta \in \mathbb{R}^+$ for which

$$\left|a^h - 1\right| < \frac{\epsilon}{M} \text{ for all } h \in \mathbb{Q} \cap (-\delta, \delta).$$

Next, because the sequence (r_n) converges, it must be a Cauchy sequence by Proposition 4.31, and so there exists $N \in \mathbb{N}$ for which

$$|r_m - r_n| < \delta \text{ for all } m, n \geq N,$$

which in turn implies that

$$\begin{aligned} |a^{r_m} - a^{r_n}| = \left|a^{r_n}(a^{r_m - r_n} - 1)\right| &\overset{\text{(a)}}{=} a^{r_n}\left|a^{r_m - r_n} - 1\right| \\ &\overset{\text{(b)}}{<} M\frac{\epsilon}{M} = \epsilon, \end{aligned}$$

where (a) holds because a is positive, and (b) uses the two bounds developed earlier in the proof. The result, i.e. $|a^{r_m} - a^{r_n}| < \epsilon$ for all $m, n \geq N$, means that $(a^{r_n})_n$ is also a Cauchy sequence and is thus convergent.

(ii) The sequences (r_n) and (s_n) converge to the same real number, and so

$$\lim_{n\to\infty}(r_n - s_n) = 0.$$

Because the function ϕ_a is continuous, Remark 6.29 allows us to change to order of limit and function evaluation to get

$$\lim_{n\to\infty} a^{r_n - s_n} = a^{\lim_{n\to\infty}(r_n - s_n)} = a^0 = 1. \tag{6.13}$$

This result in turn implies that

$$\begin{aligned} \lim_{n\to\infty} a^{r_n} &= \lim_{n\to\infty}\left(a^{r_n - s_n}a^{s_n}\right) \\ &\overset{\text{(a)}}{=} \lim_{n\to\infty} a^{r_n - s_n} \cdot \lim_{n\to\infty} a^{s_n} \\ &\overset{\text{(b)}}{=} \lim_{n\to\infty} a^{s_n}, \end{aligned}$$

where (a) holds by Proposition 4.17 (iii), and (b) by result of Equation 6.13. □

We are now prepared to define the real-valued exponential function.

Definition 6.50: The exponential function

For any $a \in \mathbb{R}^+$ and all $x \in \mathbb{R}$, we define the number a^x as

$$a^x = \lim_{n\to\infty} (a^{r_n}),$$

where (r_n) is an arbitrary sequence of real numbers that converges to x, and define the *exponential function* with *base* a as the function

$$\psi_a \colon \mathbb{R} \to \mathbb{R}, \quad \psi_a(x) = a^x.$$

The above definition is sensible because for all $x \in \mathbb{R}$, there exists at least one sequence of rational numbers (r_n) that converges to x, and by Proposition 6.49 the sequence $(a^{r_n})_n$ converges to the same limit for any rational sequence (r_n) that converges to x. More so, if x is a rational number, Definition 6.50 agrees with the definition of the number $\phi_a(x)$ from Definition 6.47.

Proposition 6.51: Properties of the exponential function

For all $a, b \in \mathbb{R}^+$ and all $x, y \in \mathbb{R}$ it holds that $a^x > 0$ and

(i) $a^x a^y = a^{x+y}$

(ii) $a^x b^x = (ab)^x$

(iii) $(a^x)^y = a^{xy}$

For all $a \in \mathbb{R}^+$, the exponential function ψ_a is continuous and obeys the following:

(iv) If $a > 1$, then ψ_a is strictly increasing.

(v) If $0 < a < 1$, then ψ_a is strictly decreasing.

(vi) If $a = 1$, then ψ_a is constant with value 1.

(vii) If $a \neq 1$, then $\psi_a(\mathbb{R}) = \mathbb{R}^+$, i.e. ψ_a's image is the set of positive real numbers.

Proof.

(i) Let (r_n) and (s_n) be arbitrary sequences of rational numbers that converge to x and y, respectively. By Proposition 4.17 (i), the sequence $(r_n + s_n)_n$ converges to $x + y$, and so

$$\begin{aligned} a^{x+y} &= a^{\lim_{n\to\infty}(r_n+s_n)} \overset{\text{(a)}}{=} \lim_{n\to\infty} (a^{r_n+s_n}) \\ &= \lim_{n\to\infty} (a^{r_n} a^{s_n}) \\ &= \lim_{n\to\infty} (a^{r_n}) \cdot \lim_{n\to\infty} (a^{s_n}) = a^x a^y, \end{aligned}$$

where in (a) we may change the order of limit and function evaluation because the exponential function is continuous for rational-valued exponents.

(ii) By a similar calculation,

$$\begin{aligned}(ab)^x &= (ab)^{\lim_{n\to\infty} r_n} = \lim_{n\to\infty} (ab)^{r_n} \\ &= \lim_{n\to\infty} (a^{r_n} b^{r_n}) = \lim_{n\to\infty} a^{r_n} \cdot \lim_{n\to\infty} b^{r_n} \\ &= a^x b^x.\end{aligned}$$

(iv) Assume $a > 1$, let $x, y \in \mathbb{R}$ be any real numbers for which $x < y$, and let (r_n) and (s_n) be two strictly increasing sequences of rational numbers for which (r_n) converges to x, (s_n) converges to y, and $x < s_1$. For all natural numbers $n \geq 2$ it then holds that

$$r_1 < r_n < x < s_1 < s_n < y.$$

Because the function $\phi_a = \psi_a|_{\mathbb{Q}}$ (i.e. the exponential function restricted to the set of rational numbers) is strictly increasing, the sequences $(a^{r_n})_n$ and $(a^{s_n})_n$ are also strictly increasing, and so for each $n \geq 2$ it holds that

$$0 < a^{r_1} < a^{r_n} < a^{s_1} < a^{s_n}.$$

From these inequalities it follows that

$$0 < a^{r_n} < \lim_{n\to\infty} a^{r_n} = a^x \leq a^{s_1} < \lim_{n\to\infty} a^{s_n} = a^y.$$

The result $a^x < a^y$ means that ψ_a is increasing (for all $a > 1$).

(v) Analogous to the proof of point (iv).

(vi) This point follows directly from the definition of the exponential function (and in turn the definitions of the exponential function with rational, integer,natural number exponents).

We now pause to show that the function ψ_a is continuous for all $a \in \mathbb{R}^+$. We first show that the function ψ_a is continuous at 0. Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because $\psi_a|_{\mathbb{Q}}$ is continuous at 0, there exists $N \in \mathbb{N}$ for which

$$\left|a^{1/N} - a^0\right| = \left|a^{1/N} - 1\right| < \epsilon \tag{6.14a}$$

$$\left|a^{-1/N} - a^0\right| = \left|a^{-1/N} - 1\right| < \epsilon. \tag{6.14b}$$

Combined with the fact that ψ_a is monotonic, Equations 6.14 mean that

$$|a^x - 1| \text{ for all } x \in (-1/N, 1/N),$$

and so ψ_a is continuous at 0.

We will now show that ψ_a is continuous at any point $u \in \mathbb{R}$. Again choose arbitrary $\epsilon \in \mathbb{R}^+$. Because ψ_a is continuous at 0, there exists $\delta \in \mathbb{R}^+$ for which

$$\left|a^h - 1\right| < \frac{\epsilon}{a^u} \text{ for all } h \in (-\delta, \delta),$$

which in turn implies that

$$\left|a^{u+h} - a^u\right| = a^u\left|a^h - 1\right| < a^u \frac{\epsilon}{a^u} = \epsilon.$$

In other words, for all $x \in (u-\delta, u+\delta)$ it holds that

$$|a^x - a^u| < \epsilon,$$

which means that ψ_a is continuous at u. We now complete the proof of the remaining points in Proposition 6.51.

(iii) First assume that $y \in \mathbb{Q}$, and choose a sequence of rational numbers (r_n) that converges to the real number $x \in \mathbb{R}$. In this case $(r_n y)_n$ is a sequence of rational numbers that converges to $xy \in \mathbb{R}$, and so

$$(a^x)^y = \left(\lim_{n\to\infty} a^{r_n}\right)^y = \lim_{n\to\infty}\left[(a^{r_n})^y\right] = \lim_{n\to\infty}(a^{r_n y}) = a^{xy}, \quad (y \in \mathbb{Q}) \quad (6.15)$$

where we may change the order of limit and function evaluation because ψ_a is continuous. Now let y be an arbitrary real number, and let (s_n) be a sequence of rational numbers that converges to y. In this case

$$(a^x)^y = \lim_{n\to\infty}(a^x)^{s_n} \overset{\text{(a)}}{=} \lim_{n\to\infty}(a^{xs_n}) = a^{\lim_{n\to\infty}(xs_n)} = a^{xy},$$

where (a) uses the equality $(a^x)^{s_n} = a^{xs_n}$, which follows from Equation 6.15 and that fact that s_n is a rational number.

(vii) We first make the computations

$$\lim_{n\to\infty} a^n = \infty \quad \text{and} \quad \lim_{n\to\infty} a^{-n} = 0, \quad (a > 1)$$

and

$$\lim_{n\to\infty} a^n = 0 \quad \text{and} \quad \lim_{n\to\infty} a^{-n} = \infty, \quad (a < 1).$$

The function ψ_a is monotonic and continuous and thus meets the assumptions of Corollary 6.39, which then implies that $f(\mathbb{R}) = \mathbb{R}^+$. □

6.5.3 The logarithmic function

Definition 6.52: Existence of the logarithm

For all real numbers $a \in \mathbb{R}^+ \setminus \{1\}$ and all nonnegative real numbers $A \in \mathbb{R}^+$, there exists a unique real number x for which

$$a^x = A.$$

We call this value the *base-a logarithm of A* and denote it by $x = \log_a A$. For all such $a \in \mathbb{R}^+ \setminus \{1\}$, the *base-a logarithmic function* is the function

$$\log_a \colon \mathbb{R}^+ \to \mathbb{R}, \quad x \mapsto \log_a x.$$

The value of x solving the equation $a^x = A$ is unique because the exponential function with

base $a \in \mathbb{R}^+ \setminus \{1\}$ is a continuous, strictly monotonic function—and thus also an injective function by Proposition 6.43—and because its image is the entire set $\mathbb{R}^+$.

By its definition, the logarithmic function with base $a \in \mathbb{R}^+ \setminus \{1\}$ is the inverse function of the exponential function with base a, meaning that

$$a^{\log_a(t)} = t \text{ for all } t \in \mathbb{R}^+ \quad \text{and} \quad \log_a(a^x) = x \text{ for all } x \in \mathbb{R}^+. \tag{6.16}$$

Like the exponential function, the logarithmic function with base a is also strictly increasing if $a > 1$ and strictly decreasing if $0 < a < 1$.

Proposition 6.53: Properties of the logarithmic function

For all $a \in \mathbb{R}^+ \setminus \{1\}$, all $x, y \in \mathbb{R}^+$, and all $\alpha \in \mathbb{R}^+$ the logarithmic function obeys

(i) $\log_a(xy) = \log_a x + \log_a y$

(ii) $\log_a(x/y) = \log_a x - \log_a y$

(iii) $\log_a(x^\alpha) = \alpha \log_a x$.

Proof.

(i) We directly verify this point with the computation

$$a^{\log_a x + \log_a y} \overset{\text{(a)}}{=} a^{\log_a x} a^{\log_a y} \overset{\text{(b)}}{=} xy \overset{\text{(c)}}{=} a^{\log_a(xy)},$$

where (a) follows from Proposition 6.51 (i) and (b) and (c) follow from Equation 6.16. Because the exponential function is injective, the resulting equality $a^{\log_a x + \log_a y} = a^{\log_a(xy)}$ implies $\log_a(xy) = \log_a x + \log_a y$.

(ii) Again using the properties of the logarithmic and exponential functions, we directly compute

$$\begin{aligned} a^{\log_a x - \log_a y} &= a^{\log_a x} a^{-\log_a y} = a^{\log_a x} \left(a^{\log_a y}\right)^{-1} \\ &= x/y = a^{\log_a(x/y)}, \end{aligned}$$

which implies $\log_a(x/y) = \log_a x - \log_a y$.

(iii) This point follows from the calculation

$$a^{\alpha \log_a x} = \left(a^{\log_a x}\right)^\alpha = x^\alpha = a^{\log_a(x^\alpha)},$$

which implies $\log_a(x^\alpha) = \alpha \log_a x$. □

We now show how to relate logarithmic functions with different bases.

Proposition 6.54: Change of base formula

For any two numbers $a, b \in \mathbb{R}^+ \setminus \{1\}$ and all $x \in \mathbb{R}^+$, the logarithmic function obeys

$$\log_b x = \frac{\log_a x}{\log_a b}. \tag{6.17}$$

Equation 6.17 is sometimes called the *change of base formula* for logarithms.

Proof. The change of base formula follows from the calculation

$$x \overset{\text{(a)}}{=} a^{\log_a x} \overset{\text{(b)}}{=} b^{\log_b x} \overset{\text{(c)}}{=} \left(a^{\log_a b}\right)^{\log_b x} \overset{\text{(d)}}{=} a^{(\log_a b)(\log_b x)},$$

where (a), (b), and (c) follow from Equation 6.16 and (d) from Proposition 6.51 (iii). The exponential function is injective, so the resulting equality $a^{\log_a x} = a^{(\log_a b)(\log_b x)}$ implies $\log_a x = (\log_a b)(\log_b x)$. □

The change of basis formula means that logarithmic functions with different bases are proportional to each other.

Remark 6.55: Common bases of the logarithm function

The most common bases for the logarithmic function are the numbers 10 and e (Euler's number). The logarithm with base e is called the *natural logarithm* and is denoted by

$$\ln = \log_{\mathrm{e}} \quad \text{(the natural logarithm)},$$

while the logarithm with base 10 is sometimes called *Briggs's logarithm*.

Sometimes a logarithmic function is written simply "log", and the logarithm's base is not written explicitly. In this case the base must be determined from context—the implicit base is usually 10 in the natural sciences and e in mathematics.

6.5.4 The power function

In the definition of the exponential function (Definition 6.50), the base is a constant and the exponent is the variable. Reversing the roles—choosing the variable to be constant and letting the base vary—produces the power function.

Definition 6.56: Power function

For all $\alpha \in \mathbb{R}$, the *power function with exponent* α is the function

$$\cdot^{\alpha} \colon \mathbb{R}^+ \to \mathbb{R}, \quad x \mapsto x^{\alpha}.$$

The power function can be written as the composition of a logarithmic, linear, and exponential function in the form

$$x^{\alpha} = \mathrm{e}^{\ln(x^{\alpha})} = \mathrm{e}^{\alpha \ln x}. \tag{6.18}$$

The power function is thus continuous for all $\alpha \in \mathbb{R}^+$ by Corollary 6.33, since the exponential, logarithmic, and linear functions are all continuous. The power function $x \mapsto x^{\alpha}$ is strictly increasing for all $\alpha > 0$ and strictly decreasing for all $\alpha < 0$. The power function's image is $\mathbb{R}^+$ for all $\alpha \neq 0$.

Example 6.57: Euler's number

Let us show that Euler's number obeys the identity

$$\lim_{x \to \infty} (1 + x)^{1/x} = \mathrm{e}.$$

To show this, for all $x \in (0, 1)$ let $n(x) \in \mathbb{N}$ be the integer part of the real number

$1/x$, i.e. the natural number for which

$$n(x) \leq 1/x < n(x) + 1.$$

For all $x \in (0, 1)$ we can then make the bound

$$1 + \frac{1}{n(x)+1} < 1 + x \leq 1 + \frac{1}{n(x)}, \tag{6.19}$$

which in turn implies

$$\begin{aligned} \left(1 + \frac{1}{n(x)+1}\right)^{n(x)} &\leq \left(1 + \frac{1}{n(x)+1}\right)^{1/x} \overset{\text{(a)}}{<} (1+x)^{1/x} \\ &\overset{\text{(b)}}{\leq} \left(1 + \frac{1}{n(x)}\right)^{1/x} \\ &< \left(1 + \frac{1}{n(x)}\right)^{n(x)+1}, \end{aligned} \tag{6.20}$$

where (a) and (b) use Equation 6.19. We then note that $\lim_{x \to 0^+} n(x) = \infty$, implying

$$\lim_{x \to 0^+} \left(1 + \frac{1}{n(x)+1}\right)^{n(x)} = \lim_{n \to \infty} \left(1 + \frac{1}{n+1}\right)^{n} \overset{\text{(a)}}{=} \mathrm{e}$$

and

$$\lim_{x \to 0^+} \left(1 + \frac{1}{n(x)}\right)^{n(x)+1} = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^{n+1} \overset{\text{(b)}}{=} \mathrm{e},$$

where (a) and (b) follow from the definition of Euler's number in Example 4.23. We then combine the bound in Equation 6.19 and the inequality in Equation 6.20 to conclude that

$$\lim_{x \to 0^+} (1+x)^{1/x} = \mathrm{e}.$$

In an analogous manner we can show that

$$\lim_{x \to 0^-} (1+x)^{1/x} = \mathrm{e}.$$

6.5.5 The hyperbolic functions

The hyperbolic functions play a fundamental role in hyperbolic geometry, which falls beyond the scope of this text. Here we give only the functions' definitions.

Definition 6.58: The hyperbolic functions

- The *hyperbolic sine* function $\sinh\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$\sinh x = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{2}.$$

The sinh function is continuous, odd, strictly increasing, and bijective.

- The *hyperbolic cosine* function $\cosh\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$\cosh x = \frac{\mathrm{e}^x + \mathrm{e}^{-x}}{2}.$$

  The cosh function is continuous and even, and its image is the interval $[1, \infty)$.

- The *hyperbolic tangent* function $\tanh\colon \mathbb{R} \to \mathbb{R}$ is defined as

$$\tanh x = \frac{\sinh x}{\cosh x}.$$

  The hyperbolic tangent function is continuous, odd, and injective, and its image is the interval $(-1, 1)$.

- The *hyperbolic cotangent* function $\coth\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}$ is defined as

$$\coth x = \frac{\cosh x}{\sinh x}.$$

  The hyperbolic cotangent function is continuous, odd, and injective, and its image is $(-\infty, -1) \cup (1, \infty)$.

Proposition 6.59: Properties of the hyperbolic functions

The hyperbolic functions obey the following Pythagorean identities for all $x \in \mathbb{R}$ for which the expressions are defined:

$$\begin{aligned} \cosh^2 x - \sinh^2 x &= 1, \\ 1 - \tanh^2 x &= \frac{1}{\cosh^2 x}, \\ 1 - \coth^2 x &= \frac{-1}{\sinh^2 x} \quad (x \neq 0). \end{aligned}$$

The hyperbolic functions also obey the following summation identities for all $x \in \mathbb{R}$ for which the expressions are defined:

$$\begin{aligned} \sinh(x+y) &= \sinh x \cosh y + \sinh y \cosh x, \\ \cosh(x+y) &= \cosh x \cosh y + \sinh x \sinh y, \\ \tanh(x+y) &= \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}, \\ \coth(x+y) &= \frac{1 + \coth x \coth y}{\coth x + \coth y}, \quad (x, y \neq 0; x + y \neq 0). \end{aligned}$$

Proof. The properties can be directly verified by straightforward but rather tedious computation using the definitions of the hyperbolic functions and the properties of the exponential function in Proposition 6.51. □

Definition 6.60: The inverse hyperbolic functions

The inverse hyperbolic functions are also called the *area functions* for their relationship to the areas of hyperbolic sectors. They are defined as follows:

- The *inverse hyperbolic sine* function $\operatorname{arsinh}\colon \mathbb{R} \to \mathbb{R}$ is the inverse of the hyperbolic sine, and is thus continuous, odd, strictly increasing, and bijective. The arsinh function can be expressed explicitly as

$$\operatorname{arsinh} x = \ln\left(x + \sqrt{x^2+1}\right) \text{ for all } x \in \mathbb{R}.$$

- The *inverse hyperbolic cosine* function $\operatorname{arcosh}\colon [1,\infty) \to [0,\infty)$ is the inverse of the restricted, injective function $\cosh|_{[0,\infty)}$. The arcosh function is continuous, injective, and strictly increasing, and its image is the interval $[0,\infty)$. The arcosh function can be expressed explicitly as

$$\operatorname{arcosh} x = \ln\left(x + \sqrt{x^2-1}\right) \text{ for all } x \in [0,\infty).$$

- The *inverse hyperbolic tangent* function $\operatorname{artanh}\colon (-1,1) \to \mathbb{R}$ is the inverse of the injective function tanh function. The artanh function is continuous, odd, strictly increasing, and bijective, and can be expressed explicitly as

$$\operatorname{artanh} x = \frac{1}{2}\ln\frac{1+x}{1-x} \text{ for all } x \in (-1,1).$$

- The *inverse hyperbolic cotangent* function $\operatorname{arcoth}\colon (-\infty,-1)\cup(1,\infty) \to \mathbb{R}\setminus\{0\}$ is the inverse of the injective function $\coth\colon \mathbb{R}\setminus\{0\} \to \mathbb{R}$ function. The arcoth function is continuous, odd, and injective, and its image is $\mathbb{R}\setminus\{0\}$. The arcoth function can be expressed explicitly as

$$\operatorname{arcoth} x = \frac{1}{2}\ln\frac{x+1}{x-1} \text{ for all } x \in (-\infty,-1)\cup(1,\infty).$$

Remark 6.61: Elementary functions

We have seen that the polynomials, trigonometric functions, inverse trigonometric functions, exponential function, and logarithmic function are all continuous. These functions may be combined through addition, subtraction, multiplication, division, and composition; such combinations are again continuous functions, and are called *elementary functions*. For example, the rational, hyperbolic, inverse hyperbolic, and power functions are all elementary functions.

7 The derivative

7.1 Foundational concepts

Definition 7.1: The derivative of a scalar function at a point

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and let a be an interior point of U. The function f is said to be *differentiable at the point* $a \in U$ if the following limit exists:

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} \in \mathbb{R}.$$

If the limit exists, it is called the *derivative of* f *at the point* a and denoted by

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

By defining $h = x - a$, we can introduce a second, common form of the derivative of f at a, which is

$$f'(a) = \frac{\mathrm{d}f}{\mathrm{d}x}(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}.$$

Definition 7.2: The derivative of a scalar function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \subset \mathbb{R}$. The function f is said to be *differentiable* if it is differentiable at every point in U. In this case we define a new function called the *derivative* of f, defined as

$$f' = \frac{\mathrm{d}f}{\mathrm{d}x}\colon U \to \mathbb{R}, \quad x \mapsto f'(x). \text{(the derivative of } f)$$

Remark 7.3: Geometric interpretation of the derivative

Let the scalar function $f\colon U \to \mathbb{R}$ be differentiable at the interior point $a \in U \subset \mathbb{R}$. We first note that the quotient

$$\frac{f(x) - f(a)}{x - a}$$

is the slope of the line in $\mathbb{R}^2$ passing through the points $(a, f(a))$ and $(x, f(x))$. As the point x approaches a, the line through the points $(a, f(a))$ and $(x, f(x))$ approaches the *tangent line* to the graph of f at the point $a, f(a)$, and the slope k of this tangent line is precisely the derivative

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

In other words, $f'(a)$ is the slope of the line tangent to f's graph at the point $a, f(a)$, which has the equation

$$y = f'(a) \cdot (x - a) + f(a).$$

Remark 7.4: Derivatives and approximations

Let the scalar function $f\colon U \to \mathbb{R}$ be differentiable at the interior point $a \in U \subset \mathbb{R}$. From Remark 7.3, recall that the equation of the line tangent to the graph of f at the point $a, f(a)$ is

$$y = f'(a) \cdot (x - a) + f(a).$$

This tangent line is the line that best approximates the graph of f at a, so we may view the value of the linear function

$$x \mapsto f'(a) \cdot (x - a) + f(a)$$

as a linear approximation of the function f near the point a. We may view the approximation's error as a function of the difference $h = x - a$, and denote this error by

$$\begin{aligned} \alpha(h) &= f(x) - \big[f(a) + f'(a) \cdot (x - a)\big] \\ &= f(a + h) - f(a) - f'(a) \cdot h. \end{aligned}$$

For small enough values of h, we may write the function $f(x)$ as

$$f(x) = f(a) + f'(a) \cdot h + \alpha(h),$$

where we note that

$$\begin{aligned} \lim_{h \to 0} \frac{\alpha(h)}{h} &= \lim_{h \to 0} \frac{f(a + h) - f(a) - f'(a) \cdot h}{h} \\ &= \lim_{h \to 0} \left[\frac{f(a + h) - f(a)}{h}\right] - f'(a) \\ &\overset{\text{(a)}}{=} f'(a) - f'(a) = 0, \end{aligned} \tag{7.1}$$

where (a) uses the definition of the derivative. In other words, the error of the linear approximation of f near a grows arbitrarily small as x approaches a.

In fact, Equation 7.1 characterizes the differentiability of a function and its derivative at a point. More specifically, let $f\colon U \to \mathbb{R}$ be an arbitrary function defined on the subset $U \subset \mathbb{R}$ and let a be an arbitrary interior point of U. If there exists a real number $v_a \in \mathbb{R}$ for which

$$\lim_{h \to 0} \frac{f(a + h) - f(a) + v_a \cdot h}{h} = 0,$$

then f is differentiable at a and $f'(a) = v_a$.

Proposition 7.5: Differentiability implies continuity

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the real subset $U \subset \mathbb{R}$, and let a be an interior point of U. If f is differentiable at a, then f is also continuous at a.

Proof. For any small-enough $h \in \mathbb{R}$ we may write $f(a + h)$ as

$$f(a + h) = f(a) + h \left[\frac{f(a + h) - f(a)}{h}\right].$$

Using the fact that f is differentiable at a, we then make the calculation

$$\begin{aligned}\lim_{h\to 0} f(a+h) &\overset{\text{(a)}}{=} \lim_{h\to 0} f(a) + \lim_{h\to 0} h\left[\frac{f(a+h)-f(a)}{h}\right] \\ &\overset{\text{(b)}}{=} \lim_{h\to 0} f(a) + \lim_{h\to 0} h \cdot \lim_{h\to 0}\left[\frac{f(a+h)-f(a)}{h}\right] \\ &\overset{\text{(c)}}{=} f(a) + 0 \cdot f'(a) = f(a),\end{aligned}$$

where (a) simply adds and subtracts the quantity $f(a)$, (b) uses the limit rules from Proposition 6.24 and (c) uses the definition of the derivative at a point. The result $\lim_{h\to 0} f(a+h) = f(a)$ means that the function value $f(a)$ equals f's limit at a, and so f is continuous by Proposition 6.28 (ii). □

Note that continuity does *not* in general imply differentiability, as we will illustrate in the following example.

Example 7.6: Derivatives of common functions

(1) The function $f(x) = |x|$ is continuous, but it is not differentiable at the point 0, since the limit

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

is undefined. Thus continuity does not in general imply differentiability.

(2) For any $c \in \mathbb{R}$, the constant function $f(x) = c$ is differentiable at every point $x \in \mathbb{R}$, and its derivative is

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

(3) For any $k, n \in \mathbb{R}$, the linear function $f(x) = kx + n$ is differentiable at every point $x \in \mathbb{R}$, and its derivative is

$$\begin{aligned}f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{\big[k(x+h)+n\big]-(kx+n)}{h} \\ &= \lim_{h\to 0} \frac{kh}{h} = \lim_{h\to 0} k = k.\end{aligned}$$

(4) For any natural number n the power function $f(x) = x^n$ is differentiable for

all $x \in \mathbb{R}$, and its derivative is

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \\ &= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \\ &\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{\left(x^n + nx^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \cdots + h^n\right) - x^n}{h} \\ &= \lim_{h\to 0} \left[nx^{n-1} + \binom{n}{2}x^{n-2}h + \cdots + h^{n-1}\right] \\ &= nx^{n-1}, \end{aligned}$$

where (a) uses the binomial expansion.

(5) The function $f(x) = \sin x$ is differentiable at all $x \in \mathbb{R}$, and its derivative is

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{\sin(x+h) - \sin x}{h} \\ &\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{2\cos\left(x+\frac{h}{2}\right)\sin\frac{h}{2}}{h} \\ &= \lim_{h\to 0} \cos\left(x+\frac{h}{2}\right) \lim_{h\to 0} \frac{\sin(h/2)}{(h/2)} \\ &\overset{\text{(b)}}{=} \lim_{h\to 0} \cos\left(x+\frac{h}{2}\right) = \cos(x), \end{aligned}$$

where (a) uses the difference-to-product formula for $\sin a - \sin b$ and (b) uses the limit $\lim_{x\to 0} \frac{\sin x}{x} = 1$ (which we have not yet proven and simply quote here).

(6) The function $f(x) = \cos x$ is differentiable at all $x \in \mathbb{R}$, and its derivative is

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{\cos(x+h) - \cos x}{h} \\ &\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{\sin\left(x+h+\frac{\pi}{2}\right) - \sin\left(x+\frac{\pi}{2}\right)}{h} \\ &\overset{\text{(b)}}{=} \frac{\mathrm{d}}{\mathrm{d}x}\left[\sin\left(x+\frac{\pi}{2}\right)\right] \overset{\text{(c)}}{=} \cos\left(x+\frac{\pi}{2}\right) \\ &\overset{\text{(d)}}{=} \sin\left(x+\frac{\pi}{2}+\frac{\pi}{2}\right) = -\sin x, \end{aligned}$$

where (a) and (d) use the identity $\cos x = \sin(x+\pi/2)$, (b) uses the definition of the derivative, and (c) uses the known derivative of the sine function.

(7) Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \to \mathbb{R}$. If f is differentiable, then for any open subset $V \subset U$, the restricted function $f|_V$ is also differentiable and obeys $(f|_V)' = f'|_V$.

Topic 7.1: Higher-order derivatives

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \subset \mathbb{R}$. If f is differentiable, then its derivative $f'\colon U \to \mathbb{R}$ is also a scalar function. In this context the function f' is called the *first derivative* of f, is conventionally denoted by

$$f' = f^{(1)} \qquad \text{(first derivative of } f\text{)}.$$

If f' is itself differentiable, then *its* derivative, i.e.

$$(f')' = f'' = f^{(2)}\colon U \to \mathbb{R},$$

is again a scalar function. In this context f'' is called the *second derivative* of f, and f is said to be *twice differentiable*.

Continuing in this manner, it is possible to define arbitrarily higher-order derivatives of the original function f. We first introduce the notation

$$f^{(0)} = f,$$

and then let $k \in \mathbb{N}$ be an arbitrary natural number. The function f is then said to be *k-times differentiable* if there exists a family of functions

$$f^{(j)}\colon U \to \mathbb{R}, \quad j = 1, 2, \ldots, k,$$

such that $f^{(j-1)}$ is differentiable and

$$\left(f^{(j-1)}\right)' = f^{(j)} \text{ for all } j = 1, 2, \ldots, k.$$

In this context $f^{(k)}$ is called the *k-th derivative* of f, and is denoted by

$$f^{(k)} = \frac{\mathrm{d}^k f}{\mathrm{d}x^k}.$$

If the function $f^{(k)}$ is also continuous, then the original function f is said to be *k-times continuously differentiable*.

Topic 7.2: Smooth functions

The set of all k-times continuously differentiable functions mapping from U to $\mathbb{R}$ is denoted by $C^k(U)$. Using the same notation, the set of all continuous functions functions $U \to \mathbb{R}$ is denoted by $C(U) = C^{(0)}(U)$.

If a function f is k-times differentiable for some $k \in \mathbb{N}$, then each of the functions $f, f', \ldots, f^{(k-1)}$ is differentiable, and thus also continuous by Proposition 7.5. It follows that

$$C^{(0)}(U) \supset C^{(1)}(U) \supset \cdots \supset C^{(k)}(U) \supset C^{(k+1)}(U) \supset \cdots$$

Functions in the set

$$C^{\infty}(U) = \bigcap_{k \in \mathbb{N}} C^k(U)$$

are called *smooth functions*.

Example 7.7: The power function is smooth

From Example 7.6, the power function $f(x) = x^n$ is differentiable for any $n \in \mathbb{N}$, and its derivative is again a power function, i.e.

$$f'(x) = nx^{n-1}.$$

It follows that f is smooth, and that

$$f^{(k)}(x) = n(n-1)\cdots(n-k+1)x^{n-k}$$

for all $k = 1, 2, \ldots, n$, while $f^{(k)} = 0$ for all $k > n$.

7.2 Differentiation rules

In practice, the derivatives of functions are rarely computed directly from the definition of the derivative, but instead using a number of far more convenient differentiation rules, which we will derive in this section.

Proposition 7.8: Derivatives of sums, products, and quotients

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the subsets $U, V \subset \mathbb{R}$, and let $a \in \mathbb{R}$ be an interior point of $U \cap V$. If both f and g are differentiable at the point a, then

(i) The function $f + g$ is differentiable at a and obeys

$$(f+g)'(a) = f'(a) + g'(a).$$

(ii) The function fg is differentiable at a and obeys

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

(iii) If $g(a) \neq 0$, then a is an interior point of the subset $U \cap g^{-1}(\mathbb{R} \setminus \{0\})$ and the function f/g is differentiable at a and obeys

$$(f/g)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2}.$$

Proof.

(i) This point follows from the definition of the derivative and the rules for computing limits of scalar functions:

$$\begin{aligned}(f+g)'(a) &= \lim_{h\to 0} \frac{(f+g)(a+h) - (f+g)(a)}{h} \\ &= \lim_{h\to 0} \frac{f(a+h) + g(a+h) - f(a) - g(a)}{h} \\ &= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} + \lim_{h\to 0} \frac{g(a+h) - g(a)}{h} \\ &= f'(a) + g'(a).\end{aligned}$$

(ii) The rule again follows from the definition of the derivative and the rules for computing limits of scalar functions:

$$\begin{aligned}
(fg)'(a) &= \lim_{h\to 0} \frac{(fg)(a+h) - (fg)(a)}{h} \\
&= \lim_{h\to 0} \frac{f(a+h)g(a+h) - f(a)g(a)}{h} \\
&\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{f(a+h)g(a+h) - f(a)g(a+h)}{h} \\
&\qquad + \lim_{h\to 0} \frac{f(a)g(a+h) - f(a)g(a)}{h} \\
&= \lim_{h\to 0} g(a+h) \cdot \lim_{h\to 0} \left(\frac{f(a+h) - f(a)}{h} \right) \\
&\qquad + \lim_{h\to 0} f(a) \cdot \lim_{h\to 0} \left(\frac{g(a+h) - g(a)}{h} \right) \\
&= f'(a) \lim_{h\to 0} g(a+h) + f(a)g'(a) \\
&\overset{\text{(b)}}{=} f'(a)g(a) + f(a)g'(a),
\end{aligned}$$

where (a) follows from adding and subtracting $f(a)g(a+h)$ from the numerator and (b) holds because g is differentiable, and thus continuous, at the point a, meaning its limit and function values at a are equal.

(iii) Because a is an interior point of $U \cap V$, by definition there exists $\delta \in \mathbb{R}^+$ such that $(a-\delta, a+\delta)$ is a subset of $U \cap V$. Because $g(a) \neq 0$ and g is continuous at a, there exists a small enough value of δ that the function g has no zeros on the interval $(a-\delta, a+\delta)$, and so

$$(a-\delta, a+\delta) \subset U \cap g^{-1}(\mathbb{R} \setminus \{0\}),$$

meaning that a is an interior point of the set $U \cap g^{-1}(\mathbb{R} \setminus \{0\})$.

We then apply the rules for computing the limits of scalar functions to get

$$\begin{aligned}
(f/g)'(a) &= \lim_{h\to 0} \frac{(f/g)(a+h) - (f/g)(a)}{h} \\
&= \lim_{h\to 0} \frac{\frac{f(a+h)}{g(a+h)} - \frac{f(a)}{g(a)}}{h} \\
&\overset{\text{(a)}}{=} \lim_{h\to 0} \frac{f(a+h)g(a) - f(a)g(a+h)}{hg(a+h)g(a)} \\
&\overset{\text{(b)}}{=} \lim_{h\to 0} \frac{f(a+h)g(a) - f(a)g(a)}{hg(a+h)g(a)} + \lim_{h\to 0} \frac{f(a)g(a) - f(a)g(a+h)}{hg(a+h)g(a)} \\
&= \frac{\lim_{h\to 0} \left(\frac{f(a+h)-f(a)}{h} \right) g(a)}{\lim_{h\to 0} g(a+h)g(a)} - \frac{f(a) \lim_{h\to 0} \left(\frac{g(a+h)-g(a)}{h} \right)}{\lim_{h\to 0} g(a+h)g(a)} \\
&\overset{\text{(c)}}{=} \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2},
\end{aligned}$$

where (a) follows from multiplying the numerator and denominator by the quantity $g(a)g(a+h)$, (b) from adding and subtracting $f(a)g(a)$ from the

numerator, and (c) holds from the assumption that g is differentiable, and thus continuous, at the point a. □

We now state the very important chain rule, which enables the straightforward computation of the derivatives of composite functions.

Proposition 7.9: The chain rule

Let $f\colon U \to \mathbb{R}$ and $g\colon V \to \mathbb{R}$ be two scalar functions defined on the subsets $U, V \subset \mathbb{R}$, and let $a \in \mathbb{R}$ be an interior point of V for which $g(a)$ is an interior point of U.

If g is differentiable at a and if f is differentiable at $g(a)$, then a is an interior point of the set $g^{-1}(U)$ and the function $f \circ g$ is differentiable at a and obeys

$$(f \circ g)'(a) = f'(g(a)) \cdot g'(a).$$

Proof. Because a is an interior point of V, there exists $\delta \in \mathbb{R}^+$ such that $(a-\delta, a+\delta)$ is a subset of V, and because $g(a)$ is an interior point of U, there exists $\epsilon \in \mathbb{R}^+$ such that $(g(a)-\epsilon, g(a)+\epsilon)$ is a subset of U. Because g is differentiable at a, it is also continuous at a by Proposition 7.5, there exists $\delta \in \mathbb{R}^+$ for which

$$g(x) \in \big(g(a)-\epsilon, g(a)+\epsilon\big) \text{ for all } x \in (a-\delta, a+\delta) \cap g^{-1}(U). \tag{7.2}$$

Equation 7.2 directly implies that

$$g\big((a-\delta, a+\delta)\big) \subset (g(a)-\epsilon, g(a)+\epsilon) \subset U,$$

which in turn implies that $(a-\delta, a+\delta) \subset g^{-1}(U)$, and so a is an interior point of the subset $g^{-1}(U)$.

We now define the function $Q\colon U \to \mathbb{R}$ as follows:

$$Q(t) = \begin{cases} \dfrac{f(t) - f\big(g(a)\big)}{t - g(a)}, & t \in U \setminus \{g(a)\}, \\ f'\big(g(a)\big), & t = g(a). \end{cases}$$

By the definition of the derivative, the function value $Q\big(g(a)\big)$ equals the limit of Q at the point $g(a)$. Because Q's limit and function value at $g(a)$ are equal, Q is continuous at $g(a)$. Because g is differentiable at a, it is also continuous at a, and so by Corollary 6.33, the composite function $Q \circ g\colon g^{-1}(U) \to \mathbb{R}$ is also continuous at a and thus obeys

$$\lim_{x \to a} Q\big(g(x)\big) = Q\big(g(a)\big) = f'\big(g(a)\big). \tag{7.3}$$

The chain rule then follows from the calculation

$$\begin{aligned} \lim_{x \to a} \frac{f\big(g(x)\big) - f\big(g(a)\big)}{x - a} &\overset{\text{(a)}}{=} \lim_{x \to a} \left[\frac{f\big(g(x)\big) - f\big(g(a)\big)}{g(x) - g(a)} \cdot \frac{g(x) - g(a)}{x - a} \right] \\ &\overset{\text{(b)}}{=} \lim_{x \to a} \left[Q\big(g(x)\big) \cdot \frac{g(x) - g(a)}{x - a} \right] \\ &= \lim_{x \to a} Q\big(g(x)\big) \cdot \lim_{x \to a} \frac{g(x) - g(a)}{x - a} \\ &\overset{\text{(c)}}{=} f'\big(g(a)\big) \cdot g'(a), \end{aligned}$$

where (a) follows from multiplying the numerator and denominator by $g(x) - g(a)$, (b) from the definition of Q, and (c) from Equation 7.3 and the definition of the derivative of g at the point a. □

Proposition 7.10: Differentiation of inverse functions

Let $f\colon U \to \mathbb{R}$ be a continuous injective function defined on the subset $U \subset \mathbb{R}$, and let a be an interior point of U. If f is differentiable at a and $f'(a) \neq 0$, then $f(a)$ is an interior point of $f(U)$, and the inverse function $f^{-1}\colon f(U) \to \mathbb{R}$ is differentiable at $f(a)$ and obeys

$$(f^{-1})'\big(f(a)\big) = \frac{1}{f'(a)}.$$

Proof. Without loss of generality we may assume that U is an interval. Since f is injective, it is also strictly monotonic by Proposition 6.43 (i); together with the fact that a is an interior point of U, f being strictly monotonic and continuous implies that $f(a)$ is an interior point of the image $f(U)$.

Since f is injective, its inverse is also injective, and so by Proposition 6.43 (i) and (ii), the inverse function f^{-1} is a continuous, strictly monotonic function. After defining the shorthand notation $b = f(a)$, the rule for differentiation of inverse functions then follows from the calculation

$$\begin{aligned}
(f^{-1})'\big(f(a)\big) &= (f^{-1})'(b) \\
&= \lim_{y\to b} \frac{f^{-1}(y) - f^{-1}(b)}{y - b} \\
&= \lim_{y\to b} \frac{f^{-1}(y) - f^{-1}(b)}{f\big(f^{-1}(y)\big) - f\big(f^{-1}(b)\big)} \\
&\overset{(a)}{=} \lim_{x\to a} \frac{x - a}{f(x) - f(a)} = \lim_{x\to a} \frac{1}{\frac{f(x)-f(a)}{x-a}} \\
&= \frac{1}{f'(a)},
\end{aligned}$$

where in (a) we have written $x = f^{-1}(y)$ and used the fact that f^{-1} is a continuous function and thus obeys $\lim_{y\to b} f^{-1}(y) = f^{-1}(b) = f^{-1}(f(a)) = a$. □

Using the above differentiation rules, we may compute the derivatives of many common functions, as shown in the following example.

Example 7.11: Using the differentiation rules

(1) (*Scaled functions*) Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$, let $a \in \mathbb{R}$ be an interior point of U, and let $c \in \mathbb{R}$ be an arbitrary real constant. If f is differentiable at a, then the scaled function cf is also differentiable at a and obeys

$$(cf)'(a) = cf'(a).$$

This follows from Proposition 7.8 (ii), the rule for the derivative of a product.

Namely, the constant function with value c is differentiable, and its derivative is zero at every point, so

$$(cf)'(a) = (c)' \cdot f(a) + cf'(a) = 0 + cf'(a) = cf'(a).$$

(2) (*Polynomials*) A real polynomial of degree $n \in \mathbb{N} \cup \{0\}$ is a scalar function $P\colon \mathbb{R} \to \mathbb{R}$ of the form

$$P(x) = a_n x^n + x_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_0, a_1, \ldots, a_n$ are real constants, and $a_n \neq 0$. A polynomial is thus a sum of scaled power functions, so by the rules for the derivative of a sum and product in Proposition 7.8, the polynomial P is a differentiable function and for all $x \in \mathbb{R}$ obeys

$$P'(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + 2a_2 x + a_1.$$

The derivative of a polynomial of degree $n \geq 1$ is thus a polynomial of degree $n-1$; the derivative of a polynomial of degree 0 is a constant function with value 0.

(3) (*Rational functions*) A rational function is the quotient of two polynomial functions, and so is itself differentiable; its derivative can be computed from the quotient rule in Proposition 7.8 (iii). Thus the derivative of a rational function is again a rational function.

(4) (*Power function with integer exponent*) From Example 7.6, the power function $f(x) = x^n$ is differentiable for any $n \in \mathbb{N}$ and obeys

$$f'(x) = xx^{n-1}.$$

The same formula holds when $n = 0$ and $x \neq 0$, and even at $x = 0$ the power function $x \mapsto x^0$ is differentiable with derivative equal to 0.

Now consider the function $x \mapsto x^{-n}$, $n \in \mathbb{N}$, defined on the subset $\mathbb{R} \subset \{0\}$. By the quotient rule in Proposition 7.8, the function $x \mapsto x^{-n}$ is differentiable and obeys

$$(x^{-n})' = \left(\frac{1}{x^n}\right)' = \frac{-nx^{n-1}}{x^{2n}} = (-n)x^{-n-1},$$

which is the same formula as for natural number exponents. Thus, for any integer exponent $p \in \mathbb{Z}$ and all $x \in \mathbb{R} \setminus \{0\}$ the power function $x \mapsto x^p$ is differentiable and obeys

$$(x^p)' = px^{p-1}.$$

(5) (*The root function*) Consider any $n \in \mathbb{N}$ for which $n \geq 2$. From item (4) just above, the power function $x \mapsto x^n$ is differentiable and its derivative is nonzero at all points $x \neq 0$ as long as $n \geq 2$. Using the inverse function derivative rule (Proposition 7.10), the root function $x \mapsto \sqrt[n]{x}$ is also differentiable and obeys

$$(\sqrt[n]{x})' = \frac{1}{n(\sqrt[n]{x})^{n-1}} = \frac{\sqrt[n]{x}}{nx}.$$

If n is even, the root function is defined on the interval $[0, \infty)$ and is differentiable for $x \in \mathbb{R}^+$, but not at $x = 0$, which is not even an interior point of its domain. If n is odd, the root function is defined on $\mathbb{R}$ and is differentiable on $\mathbb{R} \setminus \{0\}$, but not at $x = 0$.

(6) (*Trigonometric functions*) As already shown in Example 7.6, the sine and cosine function are differentiable and their derivatives are

$$(\sin x)' = \cos x \quad \text{and} \quad (\cos x)' = -\sin x.$$

Using the quotient rule, the tangent and cotangent functions are also differentiable and obey

$$(\tan x)' = \left(\frac{\sin x}{\cos x}\right)' = \frac{\cos x \cos x - \sin x(-\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x}$$
$$(\cot x)' = \left(\frac{\cos x}{\sin x}\right)' = \frac{-\sin x \sin x - \cos x \cos x}{\sin^2 x} = -\frac{1}{\sin^2 x}.$$

(7) (*Inverse trigonometric functions*) The inverse sine function arcsin is the inverse of the restriction of the sine function to the interval $[-\pi/2, \pi/2]$. The derivative of the sine function is nonzero for all points in the open interval $(-\pi/2, \pi/2)$, and so by the inverse function derivative rule the arcsin function is differentiable for all points in the interval $(-1, 1)$, and for all x in this interval obeys

$$\begin{aligned}(\arcsin x)' &= \frac{1}{\sin'(\arcsin x)} = \frac{1}{\cos(\arcsin x)} \\ &= \frac{1}{\sqrt{1 - \sin^2(\arcsin x)}} = \frac{1}{\sqrt{1 - x^2}}.\end{aligned}$$

The arcsin function is not differentiable at the points -1 or 1.

Similarly, inverse cosine function arccos is the inverse of the restriction of the cosine function to the interval $[0, \pi]$. The derivative of the cosine function is nonzero for all points in the open interval $(0, \pi)$, and so by the inverse function derivative rule the arccos function is differentiable for all points in the interval $(-1, 1)$ and for all x in this interval obeys

$$\begin{aligned}(\arccos x)' &= \frac{1}{\cos'(\arccos x)} = \frac{1}{-\sin(\arccos x)} \\ &= \frac{1}{\sqrt{1 - \cos^2(\arccos x)}} = \frac{-1}{\sqrt{1 - x^2}}.\end{aligned}$$

The arccos function is not differentiable at the points -1 or 1.

Similarly, again using the inverse function derivative rule, the inverse tangent and inverse cotangent functions are differentiable, with derivatives

$$(\arctan x)' = \frac{1}{\frac{1}{\cos^2(\arctan x)}} = \frac{1}{1 + \tan^2(\arctan x)} = \frac{1}{1 + x^2}.$$
$$(\operatorname{arccot} x)' = \frac{1}{\frac{-1}{\sin^2(\operatorname{arccot} x)}} = \frac{-1}{1 + \cot^2(\operatorname{arccot} x)} = \frac{-1}{1 + x^2}.$$

Both the derivatives of the arctan and arccot functions are defined on the entire real line.

Note that the derivatives of each of the inverse trigonometric functions are combinations of rational and root functions.

(8) (*The logarithmic function*) The logarithmic function $f(x) = \log_a x$ is differentiable for any base $a \in \mathbb{R}^+$. For any $x \in \mathbb{R}^+$, we can compute its derivative directly from the definition of the derivative using the calculation

$$\begin{aligned}
(\log_a x)' &= \lim_{h\to 0} \frac{\log_a(x+h) - \log_a x}{h} = \lim_{h\to 0} \frac{\log_a \frac{x+h}{x}}{h} \\
&\overset{\text{(a)}}{=} \lim_{h\to 0} \left[\frac{1}{x} \cdot \frac{x}{h} \cdot \log_a\left(1 + \frac{h}{x}\right)\right] \\
&= \frac{1}{x} \lim_{h\to 0} \left[\frac{x}{h} \cdot \log_a\left(1 + \frac{h}{x}\right)\right] \\
&\overset{\text{(b)}}{=} \frac{1}{x} \lim_{h\to 0} \log_a\left(1 + \frac{h}{x}\right)^{x/h} \\
&\overset{\text{(c)}}{=} \frac{1}{x} \log_a\left[\lim_{h\to 0}\left(1 + \frac{h}{x}\right)^{x/h}\right] \\
&\overset{\text{(d)}}{=} \frac{1}{x} \log_a\left[\lim_{u\to 0}(1+u)^{1/u}\right] \\
&\overset{\text{(e)}}{=} \frac{1}{x} \log_a \mathrm{e} = \frac{1}{x \ln a},
\end{aligned}$$

where in (a) we have multiplied above and below by x, in (b) we have used the logarithm property $\alpha \log_a x = \log_a(x^\alpha)$ for any $\alpha \in \mathbb{R}^+$ (see Proposition 6.53), in (c) we have switched the order of limit and function evaluation because the logarithm function is continuous, in (d) we have defined $u = h/x$, and in (e) we have used the definition of Euler's number from Example 4.23.

In the case when $a = \mathrm{e}$, the above result simplifies to

$$(\ln x)' = \frac{1}{x}.$$

(9) (*The exponential function*) For any base $a \in \mathbb{R}^+$, the exponential function $f(x) = a^x$ is differentiable by the inverse function derivative rule, since the exponential function is the inverse of the logarithmic function, whose derivative does not have any zeros. Using the inverse function derivative rule, the derivative of the exponential function $x \mapsto a^x$ is

$$(a^x)' = \frac{1}{\log_a'(a^x)} = \frac{1}{\frac{1}{a^x \ln a}} = a^x \ln a.$$

This results holds even when $a = 1$, in which case the exponential function is constant and $(a^x)' = 0$, since $\ln 1 = 0$. In the special case when $a = \mathrm{e}$, the above result simplifies to

$$(\mathrm{e}^x)' = \mathrm{e}^x.$$

The exponential function $x \mapsto \mathrm{e}^x$ thus equals its own derivative—this is one of the reasons for the importance of Euler's number in mathematics and science.

(10) (*The power function*) As in Equation 6.18, for any $\alpha \in \mathbb{R}$ we may write the power function $x \mapsto x^\alpha$ with domain $\mathbb{R}^+$ as a composition of a logarithmic, linear, and exponential function in the form

$$x^\alpha = \mathrm{e}^{\alpha \ln x}.$$

By the chain rule, the power function is thus differentiable with derivative

$$(x^\alpha)' = \left(\mathrm{e}^{\alpha \ln x}\right)' = \mathrm{e}^{\alpha \ln x} \cdot (\alpha \ln x)' = \mathrm{e}^{\alpha \ln x} \cdot \frac{\alpha}{x} = \alpha x^{\alpha-1},$$

which takes the same general form as the derivative of the power function for integer exponents in point (4).

7.2.1 Summary of differentiation rules

For convenience, here is a summary of the derivative rules covered in this section:

$$\begin{aligned}
(\alpha f)' &= \alpha f', \ \alpha \in \mathbb{R} && \text{(scaled function rule)} \\
(f+g)' &= f' + g' && \text{(sum rule)} \\
(fg)' &= f'g + fg' && \text{(product rule)} \\
(f/g)' &= \frac{f'g - fg'}{g^2} && \text{(quotient rule)} \\
(f \circ g)' &= (f' \circ g) \cdot g' && \text{(chain rule)} \\
(f^{-1})' &= \frac{1}{f' \circ f^{-1}} && \text{(inverse function rule).}
\end{aligned}$$

And here are some important functions and their derivatives:

Function	Derivative	Validity
x^α	$\alpha x^{\alpha-1}$	$a \in \mathbb{R}$, $x \in \mathbb{R}^+$
$\sin x$	$\cos x$	$x \in \mathbb{R}$
$\cos x$	$-\sin x$	$x \in \mathbb{R}$
$\tan x$	$1/\cos^2 x$	$x \in \mathbb{R} \setminus \{\frac{\pi}{2} + \pi k\}$
$\cot x$	$-1/\sin^2 x$	$x \in \mathbb{R} \setminus \{\pi k\}$
a^x	$a^x \ln a$	$a \in \mathbb{R}^+$, $x \in \mathbb{R}$
e^x	e^x	$x \in \mathbb{R}$
$\log_a x$	$1/(x \ln a)$	$a, x \in \mathbb{R}^+$
$\ln x$	$1/x$	$x \in \mathbb{R}^+$
$\arcsin x$	$1/\sqrt{1-x^2}$	$x \in (-1, 1)$
$\arccos x$	$-1/\sqrt{1-x^2}$	$x \in (-1, 1)$
$\arctan x$	$1/(1+x^2)$	$x \in \mathbb{R}$
$\operatorname{arccot} x$	$-1/(1+x^2)$	$x \in \mathbb{R}$

7.3 Applications of the derivative

7.3.1 Local extrema and stationary points

In this section we show how the derivative can be used to find the points at which a function attains its minimum and maximum values.

Definition 7.12: Extrema of a scalar function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$, and let a be an arbitrary point in U.

(i) The function f is said to have a *local maximum* at a if there exists $\delta \in \mathbb{R}^+$ for which
$$f(x) \leq f(a) \text{ for all } x \in (a+\delta, a-\delta) \cap U.$$

(ii) The function f is said to have a *local minimum* at a if there exists $\delta \in \mathbb{R}^+$ for which
$$f(x) \geq f(a) \text{ for all } x \in (a+\delta, a-\delta) \cap U.$$

(iii) The function f is said to have a *local extremum* at a if it has either a local maximum or a local minimum at a.

(iv) The function f is said to have a *strict local maximum* at a if there exists $\delta \in \mathbb{R}^+$ for which
$$f(x) < f(a) \text{ for all } x \in (a+\delta, a-\delta) \cap (U \setminus \{a\}).$$

(v) The function f is said to have a *strict local minimum* at a if there exists $\delta \in \mathbb{R}^+$ for which
$$f(x) > f(a) \text{ for all } x \in (a+\delta, a-\delta) \cap (U \setminus \{a\}).$$

(vi) The function f is said to have a *strict local extremum* at a if it has either a strict local maximum or a strict local minimum at a.

(vii) The function f is said to have a *global maximum* at a if
$$f(x) \leq f(a) \text{ for all } x \in U,$$
and a *global strict maximum* at a if
$$f(x) < f(a) \text{ for all } x \in U \setminus \{a\}.$$
In both cases we write $f(a) = \max(f)$.

(viii) The function f is said to have a *global minimum* at a if
$$f(x) \geq f(a) \text{ for all } x \in U,$$
and a *global strict minimum* at a if
$$f(x) > f(a) \text{ for all } x \in U \setminus \{a\}.$$
In both cases we write $f(a) = \min(f)$.

(ix) The function f is said to have an *global extremum* at a if it has either a global maximum or a global minimum at a, and a *global strict extremum* at a if it has either a strict global maximum or a strict global minimum at a.

Note that the definition of local extrema involves only the immediate surroundings of the extremum, while global extrema are defined with respect to all points in a function's domain. It follows that every global extremum is also a local extremum and that a function can

have at most one strict global maximum and at most one strict global minimum.

We will now show that differentiable scalar functions have derivatives equal to zero, and thus horizontal tangent lines, at their local extrema.

Definition 7.13: Stationary point

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$. An interior point $a \subset U$ is called a *stationary point* of the function f if f is differentiable at a and

$$f'(a) = 0.$$

Every local extremum of a scalar function must satisfy the following condition.

Proposition 7.14: Necessary condition for a local extremum

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the subset $U \subset \mathbb{R}$, let $a \in U$ be an interior point of U, and let f be differentiable at at a. If f has a local extremum at the point a, then a must be a stationary point of f.

Proof. We first define the function $Q\colon U \to \mathbb{R}$ with the formula

$$Q(x) = \begin{cases} \frac{f(x)-f(a)}{x-a}, & x \in U \setminus \{a\}, \\ f'(a), & x = a. \end{cases}$$

Because f is differentiable at a, the function Q must obey

$$\lim_{x\to a} Q(x) = \lim_{x\to a} \frac{f(x)-f(a)}{x-a} = f'(a) = Q(a),$$

which means that Q is continuous at a. We must show that $Q(a) = 0$ under the assumption that f has a local extremum at a; we will do this by showing that $Q(a)$ is neither less than nor greater than zero.

We will first show by contradiction that $Q(a)$ is not greater than zero. Assume $Q(a) > 0$. Because a is an interior point of U, there exists $\delta \in \mathbb{R}^+$ for which $(a-\delta, a+\delta) \subset U$ and, because Q is continuous at a and assuming $Q(a) > 0$, for which

$$Q(x) > 0 \text{ for all } x \in (a-\delta, a+\delta).$$

By the definition of Q, it then holds that

$$\frac{f(x)-f(a)}{x-a} > 0 \text{ for all } x \in (a-\delta, a+\delta) \setminus \{a\},$$

which in turn implies that

$$f(x) - f(a) < 0 \text{ for all } x \in (a-\delta, a) \text{ i.e. when } (x-a) < 0,$$
$$f(x) - f(a) > 0 \text{ for all } x \in (a, a+\delta) \text{ i.e. when } (x-a) > 0.$$

This result means that $f(x) - f(a)$ changes sign on the interval $(a-\delta, a+\delta)$; this contradicts the proposition's assumption that f has a local extremum at a, and so $Q(a)$ cannot be greater than zero.

We can use an analogous procedure to show that $Q(a)$ is also not less than zero. This implies $Q(a) = 0$, and so $f'(a) = 0$ is a necessary condition for f to have a local extremum at a. □

Note that the converse of Proposition 7.14 is not true—a scalar function may have a stationary point at which it does not have a local extremum, as we show in the following example.

Example 7.15: Extrema of common functions

(1) Consider the scalar function $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin\bigl(x^2\bigr)$. This function is the composition of a sine and quadratic function; these are both differentiable, so by the chain rule f is also differentiable with derivative equal to

$$f'(x) = 2x\cos\bigl(x^2\bigr).$$

The function f's stationary points, i.e. the zeros of f', are thus

$$x_k^{\pm} = \pm\sqrt{\frac{\pi}{2} + k\pi}, \quad k = 0, 1, 2, \ldots$$

A graph of $f(x)$ would show that f indeed has a local extremum at each of the points $x_k^{\pm}$.

(2) The scalar function $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$ has a stationary point, but not a local extremum, at $x = 0$, showing that a function need not, in general, have an extremum at a stationary point.

(3) It is also possible for a scalar function f to have an extremum at a point a that is not a stationary point of f, but *only if* the assumptions in Proposition 7.14 are not met. In practice, this might occur when a is not an interior point of f's domain, or when f is not differentiable at a. As an example, consider the function $f\colon [-1, 1] \to \mathbb{R}$ given by $f(x) = |x|$. This function does not have any stationary points, but nonetheless has a has a local minimum at $x = 0$ and local maxima at $x = -1$ and $x = 1$.

Theorem 7.16: Rolle's theorem

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let the function $f\colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). If $f(a) = f(b)$, there then exists a point $c \in (a, b)$ for which

$$f'(c) = 0.$$

Proof. The function f is continuous and defined on a closed, bounded subset, and is thus bounded by Theorem 6.36. Because f is bounded, there exist points $u, v \in [a, b]$ for which

$$f(u) = \sup(f) \quad \text{and} \quad f(v) = \inf(f);$$

we stress that f thus has extrema at the points u and v.

Generally speaking, the points u and v could either be interior points or endpoints of the interval $[a, b]$. If $u \in (a, b)$, then u is an interior point of $[a, b]$; the point u thus meets the criterion of Proposition 7.14 and so

$$f'(u) = 0.$$

In this case we may choose $c = u$, proving this case of the Rolle's theorem by construction. Similarly, if $v \in (a, b)$, then $f'(v) = 0$ and we may choose $c = v$.

The only remaining possibility is that u and v are endpoints of the interval $[a, b]$, i.e. that $\{u, v\} \subset \{a, b\}$. We can then make the calculation

$$\sup(f) = f(u) = f(a) \overset{\text{(a)}}{=} f(b) = f(v) = \inf(f),$$

where (a) is required by the theorem's assumption that $f(a) = f(b)$. In this case f is constant and $f'(c) = 0$ for any point $c \in (a, b)$. □

Corollary 7.17: Monotonic functions on closed intervals

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let the function $f\colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$, differentiable on the open interval (a, b), and let $f'(x) \neq 0$ for all $x \in (a, b)$. In this case f is a strictly monotonic function.

Proof. This corollary follows directly from Rolle's theorem using a simple proof by contradiction. Assume f is not strictly monotonic. In this case there would exist points $u, v \in [a, b]$ for which $u \neq v$ and $f(u) = f(v)$. Since the labels u and v are arbitrary, will assume without loss of generality that $u < v$; otherwise we could simply reverse the roles of u and v. By Rolle's theorem applied to the restricted function $f|_{[u,v]}$, there would then exist a point $c \in (u, v) \subset (a, b)$ for which

$$f'(c) = 0,$$

which contradicts the assumption that $f'(x) \neq 0$ for all $x \in (a, b)$. Thus f must be strictly monotonic. □

Lagrange's mean value theorem, which we prove below, is an important generalization of Rolle's theorem; Because being interesting in its own right, we will regularly use Lagrange's mean value theorem to prove other propositions later in this book.

Theorem 7.18: Lagrange's mean value theorem

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let the function $f\colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). In this case there exists a point $c \in (a, b)$ for which

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

Proof. We first define the function $h\colon [a, b] \to \mathbb{R}$ according to

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

By construction, the function h is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), and in addition obeys

$$\begin{aligned} h(a) &= f(a)(b - a) - a\big[f(b) - f(a)\big] = f(a)b - af(b) \\ h(b) &= f(b)(b - a) - b\big[f(b) - f(a)\big] = -f(b)a + bf(a) = h(a). \end{aligned}$$

Because $h(a) = h(b)$, the function h meets the assumptions of Rolle's theorem, and so by Rolle's theorem there exists a point $c \in (a, b)$ for which $h'(c) = 0$. Because

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

it follows from $h'(c) = 0$ that

$$0 = h'(c) = f'(c)(b - a) - \big[f(b) - f(a)\big],$$

which we then simply rearrange to get

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

Lagrange's theorem is sometimes called simply the *mean value theorem*—we will occasionally use this name when the context is clear. To interpret the mean value theorem, note that the quantity

$$\frac{f(b) - f(a)}{b - a}$$

is the slope of the line passing through the endpoints $\big(a, f(a)\big)$ and $\big(b, f(b)\big)$ of f's graph. In this light, the mean value theorem guarantees the existence of some point c in the interval (a, b) at which the slope $f'(c)$ of the tangent line passing through the point $\big(c, f(c)\big)$ equals the slope of the line through the interval endpoints.

We offer a generalization of Lagrange's mean value theorem in the following theorem.

Theorem 7.19: Cauchy's mean value theorem

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let the functions $f, g \colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). In this case there exists a point $c \in (a, b)$ for which

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

Note that Cauchy's mean value theorem reduces to Lagrange's mean value theorem if g is the identity function, i.e. if $g(x) = x$.

Proof. The proof is similar to the proof of Lagrange's mean value theorem: we first define the function $h \colon [a, b] \to \mathbb{R}$ according to

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

By construction, the function h is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), and in addition obeys

$$\begin{aligned} h(a) &= f(a)\big[g(b) - g(a)\big] - g(a)\big[f(b) - f(a)\big] \\ &= f(a)g(b) - g(a)f(b), \\ h(b) &= f(b)\big[g(b) - g(a)\big] - g(b)\big[f(b) - f(a)\big] \\ &= -f(b)g(a) + g(b)f(a) = h(a). \end{aligned}$$

Because $h(a) = h(b)$, the function h meets the assumptions of Rolle's theorem, and so by Rolle's theorem there exists a point $c \in (a, b)$ for which $h'(c) = 0$. Because

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

it follows from $h'(c) = 0$ that

$$0 = h'(c) = f'(c)\big[g(b) - g(a)\big] - g'(c)\big[f(b) - f(a)\big],$$

which we then simply rearrange to get

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$ □

We dedicate the remainder of this section to some very useful consequences of the mean value theorem related to the growth and extrema of scalar functions.

Proposition 7.20: Existence of derivative limit implies differentiability

Let $f\colon U \to \mathbb{R}$ be a continuous scalar function defined on the open subset $U \subset \mathbb{R}$, define the point $a \in U$, and assume that f is differentiable on the set $U \setminus \{a\}$ and that the limit of the derivative $f'\colon U \setminus \{a\} \to \mathbb{R}$ exists at a. In this case the function f is differentiable at a and is equal to

$$f'(a) = \lim_{x \to a} f'(x).$$

Proof. Because U is an open set, there exists $\delta \in \mathbb{R}^+$ for which $(a - \delta, a + \delta) \subset U$. For any $x \in (a - \delta, a + \delta) \setminus \{a\}$, Lagrange's theorem guarantees the existence of a point c_x in the open interval between x and a for which

$$f'(c_x) = \frac{f(x) - f(a)}{x - a} \implies f(x) - f(a) = f'(c_x) \cdot (x - a).$$

We can the directly verify the proposition with the calculation

$$\begin{aligned} f'(a) &\overset{\text{(a)}}{=} \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \\ &\overset{\text{(b)}}{=} \lim_{x \to a} \frac{f'(c_x)(x - a)}{x - a} \\ &= \lim_{x \to a} f'(c_x) \overset{\text{(c)}}{=} \lim_{x \to a} f'(x), \end{aligned}$$

where (a) follows from the definition of the derivative, (b) from the earlier Lagrange's theorem equality $f(x) - f(a) = f'(c_x) \cdot (x - a)$ and (c) from the fact that c_x lies in the interval *between* x and a, and so the limit $x \to a$ also results in $c_x \to a$. □

Proposition 7.21: The first derivative and function growth

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let the function $f\colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). In this case:

(i) The function f is increasing on the interval $[a, b]$ if, and only if,

$$f'(x) \geq 0 \text{ for all } x \in (a, b).$$

(ii) The function f is decreasing on the interval $[a, b]$ if, and only if,

$$f'(x) \leq 0 \text{ for all } x \in (a, b).$$

(iii) If $f'(x) > 0$ for all $x \in (a, b)$, then the function f is strictly increasing on the interval $[a, b]$.

(iv) If $f'(x) < 0$ for all $x \in (a, b)$, then the function f is strictly decreasing on the interval $[a, b]$.

(v) The function f is constant on the interval $[a, b]$ if, and only if,

$$f'(x) = 0 \text{ for all } x \in (a, b).$$

Proof.

(i) If f is increasing, then

$$\frac{f(y) - f(x)}{y - x} \geq 0 \text{ for any distinct } x, y \in (a, b),$$

which immediately implies that

$$f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x} \geq 0 \text{ for any distinct } x, y \in (a, b).$$

Conversely, assume that $f'(x) \geq 0$ for all $x \in (a, b)$, and consider any two points $x, y \in [a, b]$ for which $x < y$. By the mean value theorem applied to the restriction of the function f to the interval $[x, y]$, there exists $c \in (x, y)$ for which

$$f'(c) = \frac{f(y) - f(x)}{y - x}.$$

The assumption that f' is nonnegative implies that

$$f'(c) = \frac{f(y) - f(x)}{y - x} \geq 0,$$

which combined with the assumption that $x < y$ implies

$$f(y) - f(x) = f'(c)(y - x) \geq 0,$$

meaning that $f(x) \leq f(y)$ and f is increasing on $[a, b]$.

(ii) Analogous to the proof of (i).

(iii) Consider any two points $x, y \in [a, b]$ for which $x < y$. By the mean value theorem applied to the restriction of the function f to the interval $[x, y]$, there exists $c \in (x, y)$ for which $f(y) - f(x) = f'(c)(y - x)$. The assumption that $f'(c) > 0$ then implies that

$$f(y) - f(x) = f'(c)(y - x) > 0,$$

which combined with the assumption that $x < y$ implies

$$f(y) - f(x) = f'(c)(y - x) > 0,$$

meaning that $f(x) < f(y)$ and f is strictly increasing on $[a, b]$.

(iv) Analogous to the proof of (iii).

(v) ($\Rightarrow$) If f is constant on $[a, b]$ then $f'(x) = 0$ on (a, b) because the derivative of a constant function is zero.

($\Leftarrow$) Assume that $f'(x) = 0$ for all $x \in (a, b)$ and consider an arbitrary point $x \in (a, b]$. By the mean value theorem applied to the restriction of the function f to the interval $[a, x]$, there exists $c \in (a, x)$ for which

$$f(x) - f(a) = f'(c)(x - a).$$

The assumption that $f'(x) = 0$ for all $x \in (a, b)$ means that $f'(c) = 0$ and thus $f(x) = f(a)$. And since $f(x) = f(a)$ for any $x \in (a, b]$, the function f must be constant. □

Definition 7.22: Locally constant function

A scalar function $f\colon U \to \mathbb{R}$ defined on the real subset $U \subset \mathbb{R}$ is said to be *locally constant at the point* $a \in U$ if there exists $\delta \in \mathbb{R}^+$ for which

$$f(x) = f(a) \text{ for all } x \in (a - \delta, a + \delta) \cap U,$$

i.e. if there exists a neighborhood of a on which f is constant.

A function is *locally constant* if it is locally constant at every point in its domain.

Any locally constant function $f\colon U \to \mathbb{R}$ is necessarily constant on every interval V that lies inside the subset U.

Corollary 7.23: Condition for locally constant functions

As a direct consequence of Proposition 7.21 (v), a function $f\colon U \to \mathbb{R}$ defined on an open subset $U \subset \mathbb{R}$ is locally constant if, and only if, it is differentiable and its derivative is zero for all points in U.

We now present three conditions for testing if a scalar function has a local extremum at a given point in its domain.

Proposition 7.24: Sufficient condition for an extremum I

Let $f\colon U \to \mathbb{R}$ be a continuous function defined on the open subset $U \subset \mathbb{R}$, define the point $a \in U$, and let f be differentiable on the set $U \setminus \{a\}$. In this case:

(i) If there exists $\delta \in \mathbb{R}^+$ for which $f'(x) \leq 0$ for all $x \in (a-\delta, a) \cap U$ and $f'(x) \geq 0$ for all $x \in (a, a + \delta) \cap U$, then f has a local minimum at a.

(ii) If there exists $\delta \in \mathbb{R}^+$ for which $f'(x) \geq 0$ for all $x \in (a-\delta, a) \cap U$ and $f'(x) \leq 0$

for all $x \in (a, a+\delta) \cap U$, then f has a local maximum at a.

(iii) If there exists $\delta \in \mathbb{R}^+$ for which $f'(x) < 0$ for all $x \in (a-\delta, a) \cap U$ and $f'(x) > 0$ for all $x \in (a, a+\delta) \cap U$, then f has a strict local minimum at a.

(iv) If there exists $\delta \in \mathbb{R}^+$ for which $f'(x) > 0$ for all $x \in (a-\delta, a) \cap U$ and $f'(x) < 0$ for all $x \in (a, a+\delta) \cap U$, then f has a strict local maximum at a.

Proof. The conditions follow directly from Proposition 7.21, the definition of increasing and decreasing functions in Definition 6.3, and the definition of extrema in Definition 7.12. □

Proposition 7.25: Sufficient condition for an extremum II

Let $f\colon U \to \mathbb{R}$ be a twice continuously differentiable function defined on the open subset $U \subset \mathbb{R}$, and define the point $a \in U$. In this case:

(i) If $f'(a) = 0$ and $f''(a) > 0$, the function f has a strict local minimum at a.

(ii) If $f'(a) = 0$ and $f''(a) < 0$, the function f has a strict local maximum at a.

Proof.

(i) Because $f''(a) > 0$ and because the function f'' is continuous, there exists $\delta \in \mathbb{R}^+$ for which $(a-\delta, a+\delta) \subset U$ and

$$f''(x) > 0 \text{ for all } x \in (a-\delta, a+\delta).$$

By Proposition 7.21 (iii) the function f' is thus strictly increasing on the interval $(a-\delta, a+\delta)$; combined with the fact that $f'(a) = 0$, this in turn implies that

$$f'(x) < 0 \text{ for all } x \in (a-\delta, a) \quad \text{and} \quad f'(x) > 0 \text{ for all } x \in (a, a+\delta).$$

Thus, by Proposition 7.24 (iii), the function f has a strict local minimum at a.

(ii) Analogous to the proof of (i). □

Note that Proposition 7.25 does not provide any information about a function's extrema at a given point when the function's second derivative is zero at this point. In such cases one must turn to higher derivatives, as formalized in the next condition.

Proposition 7.26: Sufficient condition for an extremum III

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \subset \mathbb{R}$, let $n \in \mathbb{N}$ be a natural number for which $n \geq 2$, and define the point $a \in \mathbb{R}$. More so, assume that f is at least n-times continuously differentiable at a, that $f^{(n)}(a) \neq 0$, and finally that

$$f^{(k)}(a) = 0 \text{ for all } k \in \mathbb{N} \text{ for which } k < n.$$

In this case:

(i) If n is even and $f^{(n)} > 0$, then f has a strict local minimum at a.

(ii) If n is even and $f^{(n)} < 0$, then f has a strict local maximum at a.

(iii) If n is odd and $f^{(n)} > 0$, then f does not have a local extremum at a and is strictly increasing in some neighborhood $V \subset U$ of the point a.

(iv) If n is odd and $f^{(n)} < 0$, then f does not have a local extremum at a and is strictly decreasing in some neighborhood $V \subset U$ of the point a.

Proof. We will prove the proposition by induction on the number n. By Proposition 7.25, the proposition holds immediately when $n = 2$, which we take as the base case. Now assume $n \geq 3$. By the induction step, the proposition holds for the function f', since $(f')^{(k)}(a) = 0$ and $(f')^{(n-1)}(a) \neq 0$ for all $k < n - 1$.

(i) Assume that n is even and that $f^{(n)}(a) > 0$. In this case $n - 1$ is odd and $(f')^{n-1}(a) > 0$. By the induction step it follows that f' is strictly increasing in some neighborhood $V \subset U$ of a. In addition $f'(a) = 0$, so by Proposition 7.24 (iii) the function f has a strict local minimum at a.

(ii) Analogous to the proof of (i).

(iii) Now assume that n is odd and that $f^{(n)}(a) > 0$. Then $n - 1$ is even and

$$(f')^{(n-1)} > 0.$$

By the induction step, the function f' has a strict local minimum at a. In addition $f'(a) = 0$, so by Proposition 7.21 (iii) f is strictly increasing in a neighborhood $V \subset U$ of a, and so f does not have a local extremum at a.

(iv) Analogous to the proof of (iii). □

7.3.2 Convexity and concavity

Definition 7.27: Convexity and concavity

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on an interval $U \subset \mathbb{R}$.

(i) The function f is said to be *convex* if for any two points $x, y \in U$ for which $x < y$ and for all $t \in [0,1]$, the function obeys

$$f\big(ty + (1-t)x\big) \leq tf(y) + (1-t)f(x).$$

(ii) The function f is said to be *concave* if for any two points $x, y \in U$ for which $x < y$ and for all $t \in [0,1]$, the function obeys

$$f\big(ty + (1-t)x\big) \geq tf(y) + (1-t)f(x).$$

(iii) The function f is said to be *strictly convex* if for any two points $x, y \in U$ for which $x < y$ and for all $t \in [0,1]$, the function obeys

$$f\big(ty + (1-t)x\big) < tf(y) + (1-t)f(x).$$

(iv) The function f is said to be *strictly concave* if for any two points $x, y \in U$ for which $x < y$ and for all $t \in [0,1]$, the function obeys

$$f\big(ty + (1-t)x\big) > tf(y) + (1-t)f(x).$$

The quantities $ty + (1-t)x$ and $tf(y) + (1-t)f(x)$ are parameterizations of the x and y coordinates, respectively, of the line in $\mathbb{R}^2$ connecting the points $(x, f(x))$ and $(y, f(y))$ as the parameter t ranges from 0 to 1.

In other words, as t traverses the interval $[0,1]$, the point

$$\big(ty + (1-t)x, tf(y) + (1-t)f(x)\big) \in \mathbb{R}^2$$

traverses the line connecting $(x, f(x))$ and $(y, f(y))$. A convex function's graph lies below this line for any $x, y \in U$, and a concave function's graph lies above it.

As an aside, an expression of the form

$$ty + (1-t)x, \quad \text{where } t \in [0,1] \text{ and } x, y \in \mathbb{R},$$

is called the *convex combination* of the points x and y.

Proposition 7.28: Convexity and the first derivative

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let the scalar function $f\colon [a,b] \to \mathbb{R}$ be continuous on the closed interval $[a,b]$ and differentiable on the open interval (a,b). In this case:

(i) The function f is convex on $[a,b]$ if, and only if, its derivative f' is increasing on (a,b).

(ii) The function f is concave on $[a,b]$ if, and only if, its derivative f' is decreasing on (a,b).

(iii) If f' is strictly increasing on (a,b), then f is strictly convex on $[a,b]$.

(iv) If f' is strictly decreasing on (a,b), then f is strictly concave on $[a,b]$.

Proof.

(iii) Assume that f' is strictly increasing, let $x, y \in [a, b]$ be any two points for which $x < y$, and define $t \in (0, 1)$ and the quantity $z = ty + (1 - t)x$. By the mean value theorem, there exist points $c \in (x, z)$ and $d \in (z, y)$ for which

$$\begin{aligned} f(z) - f(x) &= f'(c)(z - x) \\ f(y) - f(z) &= f'(d)(y - z). \end{aligned}$$

We first make the auxiliary calculation

$$\begin{aligned} (1 - t)(z - x) &= z - x - tz + tx \\ &\overset{\text{(a)}}{=} \big[ty + (1 - t)x\big] - x - tz + tx \\ &= t(y - z), \end{aligned}$$

where (a) follows from the definition of z, which we then use in the calculation

$$\begin{aligned} (1 - t)\big[f(z) - f(x)\big] &\overset{\text{(a)}}{=} (1 - t)f'(c)(z - x) \\ &\overset{\text{(b)}}{<} (1 - t)f'(d)(z - x) \\ &\overset{\text{(c)}}{=} tf'(d)(y - z) \\ &\overset{\text{(d)}}{=} t\big[f(y) - f(z)\big], \end{aligned}$$

where (a) holds from the mean value theorem, (b) because f' is strictly increasing and so $f'(c) < f'(d)$, (c) from the intermediate result $(1-t)(z-x) = t(y-z)$, and (d) from the mean value theorem. We then rearrange and simplify this result to get

$$f(z) < tf(y) + (1 - t)f(x),$$

which, after substituting in $z = ty + (1 - t)x$, gives

$$f\big(ty + (1 - t)x\big) < tf(y) + (1 - t)f(x),$$

meaning that f is strictly convex.

(iv) Analogous to the proof of (iii).

(i) ($\Rightarrow$) Assume that f is convex, and choose arbitrary $x, y \in (a, b)$ for which $x < y$. For any $t \in (0, 1)$, it then holds that

$$\begin{aligned} f\big(x + t(y - x)\big) &= f\big(y + (1 - t)(x - y)\big) \\ &= f\big(ty + (1 - t)x\big) \\ &\overset{\text{(a)}}{\leq} tf(y) + (1 - t)f(x) \\ &= f(y) + (1 - t)\big[f(x) - f(y)\big] \\ &= f(x) + t\big[f(y) - f(x)\big], \end{aligned}$$

where (a) holds because f is convex. We then rearrange this result to get the inequality

$$\begin{aligned} \frac{f\big(x + t(y - x)\big) - f(x)}{t} &\leq f(y) - f(x) \\ &\leq -\frac{f\big(y + (1 - t)(x - y)\big) - f(y)}{1 - t}. \end{aligned} \tag{7.4}$$

We then compute the two limits

$$\begin{aligned}\lim_{t\to 0}\frac{f\big(x+t(y-x)\big)-f(x)}{t} &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\cdot(y-x)\\ &= f'(x)(y-x)\end{aligned}$$

and

$$\begin{aligned}\lim_{t\to 1}\frac{f\big(y+(1-t)(x-y)\big)-f(y)}{1-t} &= \lim_{s\to 0}\frac{f\big(y+s(x-y)\big)-f(y)}{s}\\ &= \lim_{h\to 0}\frac{f(y+h)-f(y)}{h}\cdot(x-y)\\ &= f'(y)(x-y).\end{aligned}$$

Using these limits, the inequality in Equation 7.4, in the limits $t\to 0$ and $t\to 1$, simplifies to

$$f'(x)(y-x)\leq f(y)-f(x)\leq f'(y)(y-x),$$

and so $f'(x)\leq f'(y)$ and f' is increasing on (a,b).

($\Leftarrow$) We prove the converse case of (i) analogously to the proof of (iii).

(ii) Analogous to the proof of (i). □

Corollary 7.29: Convexity and the second derivative

Let $a,b\in\mathbb{R}$ be two real numbers for which $a<b$ and let the scalar function $f\colon[a,b]\to\mathbb{R}$ be continuous on the closed interval $[a,b]$ and twice differentiable on the open interval (a,b). In this case:

(i) The function f is convex on $[a,b]$ if, and only if, $f''(x)\geq 0$ for all $x\in(a,b)$.

(ii) The function f is concave on $[a,b]$ if, and only if, $f''(x)\leq 0$ for all $x\in(a,b)$.

(iii) If $f''(x)>0$ for all $x\in(a,b)$, then the function f is strictly convex on $[a,b]$.

(iv) If $f''(x)<0$ for all $x\in(a,b)$, then the function f is strictly concave on $[a,b]$.

Proof. The corollary follows from the combination of (1) the relationship between convexity and the growth of the first derivative in Proposition 7.28 and (2) the relationship between function growth and the first derivative in Proposition 7.21. □

Definition 7.30: Inflection points

Let $f\colon U\to\mathbb{R}$ be a scalar function defined on a real subset $U\subset\mathbb{R}$, and define the point $a\in U$.

(i) The function f is said to have an *inflection point* (*from concavity to convexity*) at the point $a\in U$ if there exists $\delta\in\mathbb{R}^+$ for which $[a-\delta,a+\delta]\subset U$ and for which f is concave on $[a-\delta,a]$ and convex on $[a,a+\delta]$.

(ii) The function f is said to have an *inflection point* (*from convexity to concavity*)

at the point $a \in U$ if there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset U$ and for which f is convex on $[a-\delta, a]$ and concave on $[a, a+\delta]$.

(iii) The function f is said to have an *inflection point* at the point $a \in U$ if it has either an inflection point from from concavity to convexity or an inflection point from convexity to concavity.

(iv) The function f is said to have an *inflection point* (*from strict concavity to strict convexity*) at the point $a \in U$ if there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset U$ and for which f is strictly concave on $[a-\delta, a]$ and strictly convex on $[a, a+\delta]$.

(v) The function f is said to have an *inflection point* (*from strict convexity to strict concavity*) at the point $a \in U$ if there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset U$ and for which f is strictly convex on $[a-\delta, a]$ and strictly concave on $[a, a+\delta]$.

The inflection points of a scalar function must satisfy the following condition.

Proposition 7.31: Necessary condition for an inflection point

Let $f\colon U \to \mathbb{R}$ be a differentiable scalar function defined on the open subset $U \subset \mathbb{R}$. If f has an inflection point at the point $a \in U$ and the derivative f' is differentiable at a, then

$$(f')'(a) = 0.$$

Proof. Assume that f has an inflection point from concavity to convexity at a. By Definition 7.30 there then exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset U$ and for which f is concave on the interval $[a-\delta, a]$ and convex on the interval $[a, a+\delta]$.

By Proposition 7.28 the function f' is thus decreasing on $(a-\delta, a)$ and increasing on $(a, a+\delta)$ (note that the intervals are open). More so, the function f' is assumed to be differentiable at a, so f' is also continuous at a, which means f' is decreasing on $(a-\delta, a]$ and increasing on $[a, a+\delta)$ (the intervals are now half-closed, since f' is continuous at a). By the sufficient conditions for an extremum in Proposition 7.24, the function f' has a local minimum at a. Finally, by the necessary condition for an extremum in Proposition 7.14, a must be a stationary point of f' and so

$$(f')'(a) = 0$$

The proof is analogous if f has an inflection point from convexity to concavity at a. □

Proposition 7.32: Sufficient condition for an inflection point

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \subset \mathbb{R}$; and let $n \in \mathbb{N}$ be a natural number and $a \in U$ a point satisfying the following conditions:

- $n \geq 3$,
- the function f is at least n-times continuously differentiable,
- $f^{(k)}(a) = 0$ for all $k \in \mathbb{N}$ for which $2 \leq k < n$, and
- $f^{(n)}(a) \neq 0$.

In this case:

(i) If n is odd and $f^{(n)}(a) > 0$, then f has an inflection point from strict concavity to strict convexity at a.

(ii) If n is odd and $f^{(n)}(a) < 0$, then f has an inflection point from strict convexity to strict concavity at a.

(iii) If n is even and $f^{(n)}(a) > 0$, then f does not have an inflection point at a and there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \in U$ and the function f is strictly convex on $[a-\delta, a+\delta]$.

(iv) If n is even and $f^{(n)}(a) < 0$, then f does not have an inflection point at a and there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \in U$ and the function f i s strictly concave on $[a-\delta, a+\delta]$.

Proof.

(i) By a combination of Proposition 7.26 (iii) and Proposition 7.21 (iii), the function f'' is strictly increasing in some neighborhood $V \subset U$ of the point a. Because f'' is strictly increasing in the neighborhood V, and $f''(a) = 0$ by this proposition's assumptions, there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset V$ and

$$f''(x) < 0 \text{ for all } x \in (a-\delta, a)$$
$$f''(x) > 0 \text{ for all } x \in (a, a+\delta).$$

It then follows from Corollary 7.29 that f is strictly concave on $[a-\delta, a]$ and strictly convex on $[a, a+\delta]$, and so a is an inflection point from strict concavity to strict convexity.

(ii) Analogous to the proof of (i).

(iii) By Proposition 7.26 (iii), the function f' (note the use of f' and not f'', since n is now even) is strictly increasing on some neighborhood $V \subset U$ of the point a. Since V is a neighborhood of a, there exists $\delta \in \mathbb{R}^+$ for which $[a-\delta, a+\delta] \subset V$. By Proposition 7.28 (iii), the function f is then strictly convex on the entire interval $[a-\delta, a+\delta]$, and so f does not have an inflection point at a.

(iv) Analogous to the proof of (iii). □

7.3.3 L'Hôpital's rule

In this section we cover progressively more general forms of L'Hôpital's rule, which is a supremely useful technique for computing the limits of scalar functions.

Proposition 7.33: L'Hôpital's rule at interval endpoints

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let $f, g\colon (a, b) \to \mathbb{R}$ be two differentiable functions, where $g'(x) \neq 0$ for all $x \in (a, b)$. In this case:

(i) If the right-sided limit of the function f'/g' exists at a, and if

$$\lim_{x \to a^+} f(x) = \lim_{x \to a^+} g(x) = 0,$$

then the right-sided limit of the function f/g also exists at a and is equal to

$$\lim_{x\to a^+}\frac{f(x)}{g(x)} = \lim_{x\to a^+}\frac{f'(x)}{g'(x)}.$$

(ii) If the left-sided limit of the function f'/g' exists at b, and if

$$\lim_{x\to b^-} f(x) = \lim_{x\to b^-} g(x) = 0,$$

then the left-sided limit of the function f/g also exists at b and is equal to

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

Proof.

(i) We first the functions $\phi, \psi\colon [a,b) \to \mathbb{R}$ according to

$$\phi(x) = \begin{cases} 0, & x = a \\ f(x) & x \in (a,b) \end{cases} \quad \text{and} \quad \psi(x) = \begin{cases} 0, & x = a \\ g(x) & x \in (a,b). \end{cases}$$

Because the functions f and g are differentiable (and thus continuous) on (a,b), and because the right-sided limits of ϕ and ψ equal the function values at $x = a$, the functions ϕ and ψ are differentiable on the open interval (a,b) and continuous on the half-closed interval $[a,b)$.

Because the derivative $g' = (\psi|_{(a,b)})'$ has no zeros by the proposition's assumptions, the function ψ is strictly monotonic. And because $\psi(a) = 0$ and ψ is strictly monotonic, it follows that

$$g(x) = \psi(x) \neq \psi(a) = 0 \text{ for all } x \in (a,b).$$

By the Cauchy mean value theorem applied to the functions ϕ and ψ on the interval (a,x), for all $x \in (a,b)$ there exists a point $c_x \in (a,x)$ for which

$$\frac{f'(c_x)}{g'(c_x)} \overset{\text{(a)}}{=} \frac{\phi'(c_x)}{\psi'(c_x)} \overset{\text{(b)}}{=} \frac{\phi(x)-\phi(a)}{\psi(x)-\psi(a)} \overset{\text{(c)}}{=} \frac{\phi(x)}{\psi(x)} \overset{\text{(d)}}{=} \frac{f(x)}{g(x)},$$

where (a) holds because $f' = \phi'$ and $g' = \psi'$ on (a,b), (b) holds by the Cauchy mean value theorem, (c) holds because $\phi(a) = \psi(a) = 0$, and (d) holds because $f = \phi$ and $g = \psi$ on (a,b). Removing the intermediate steps, the result reads

$$\frac{f(x)}{g(x)} = \frac{f'(c_x)}{g'(c_x)}.$$

Since c_x lies inside the open interval (a,x) for all $x \in (a,b)$, the point c_x is closer to a than the point x is, and so

$$\lim_{x\to a^+}\frac{f(x)}{g(x)} = \lim_{x\to a^+}\frac{f(c_x)}{g(c_x)} = \lim_{x\to a^+}\frac{f'(x)}{g'(x)},$$

which proves point (i).

(ii) Analogous to the proof of (i). □

Proposition 7.34: L'Hôpital's rule in an interval

Let $U \subset \mathbb{R}$ be an open subset of $\mathbb{R}$ and define the point $a \in U$. Let $f, g \colon U \setminus \{a\} \to \mathbb{R}$ be differentiable functions for which $g'(x) \neq 0$ for all $x \in U \setminus \{a\}$

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

If the limit of the function f'/g' exists at a, then the limit of the function f/g also exists at a and is equal to

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

Proof. We apply Proposition 7.33 to the left-sided limit, and then to the right-sided limit, of the functions f/g and f'/g' at a. □

Proposition 7.35: L'Hôpital's rule at infinity

Let $f, g \colon U \to \mathbb{R}$ be two differentiable functions defined on the open interval $U \subset \mathbb{R}$, and assume $g'(x) \neq 0$ for all $x \in U$. We then separately consider the following cases:

(i) Assume that the interval U is unbounded above and that

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

If the limit of the function f'/g' exists at infinity, then the limit of the function f/g also exists at infinity and is equal to

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}.$$

(ii) Assume that the interval U is unbounded below and that

$$\lim_{x \to -\infty} f(x) = \lim_{x \to -\infty} g(x) = 0.$$

If the limit of the function f'/g' exists at negative infinity, then the limit of the function f/g also exists at negative infinity and is equal to

$$\lim_{x \to -\infty} \frac{f(x)}{g(x)} = \lim_{x \to -\infty} \frac{f'(x)}{g'(x)}.$$

Proof.

(i) We use Proposition 7.33 (i) to directly compute

$$\begin{aligned} \lim_{x \to \infty} \frac{f(x)}{g(x)} &\overset{\text{(a)}}{=} \lim_{t \to 0^+} \frac{f(1/t)}{g(1/t)} \\ &\overset{\text{(b)}}{=} \lim_{t \to 0^+} \frac{f'(1/t)(-1/t^2)}{g'(1/t)(-1/t^2)} \\ &= \lim_{t \to 0^+} \frac{f'(1/t)}{g'(1/t)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}, \end{aligned}$$

where (a) follows from defining the new variable $t = 1/x$, and (b) from the chain rule and L'Hôpital's rule for interval endpoints in Proposition 7.33 (i).

(ii) Analogous to the proof of (i). □

Proposition 7.36: L'Hôpital's rule for divergent functions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let $f, g\colon (a,b) \to \mathbb{R}$ be two differentiable functions, where $g'(x) \neq 0$ for all $x \in (a,b)$. In this case:

(i) If $\lim_{x\to a^+} |g(x)| = \infty$ and the right-sided limit of the function f'/g' exists at a, then the right-sided limit of the function f/g also exists at a and is equal to

$$\lim_{x\to a^+} \frac{f(x)}{g(x)} = \lim_{x\to a^+} \frac{f'(x)}{g'(x)}.$$

(ii) If $\lim_{x\to b^-} |g(x)| = \infty$ and the left-sided limit of the function f'/g' exists at b, then the left-sided limit of the function f/g also exists at b and is equal to

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

Proof.

(i) Because the derivative g' does not have any zeros, the function g is strictly monotonic. We first choose arbitrary $\epsilon \in \mathbb{R}^+$ and define the shorthand notation

$$L = \lim_{x\to a^+} \frac{f'(x)}{g'(x)}$$

By the definition of a function's limit, there exists $s \in (a,b)$ for which

$$\left|\frac{f'(x)}{g'(x)} - L\right| < \frac{\epsilon}{8} \text{ for all } x \in (a,s). \tag{7.5}$$

Because $\lim_{x\to a^+} |g(x)| = \infty$, there exists $\delta \in \mathbb{R}^+$ for which:

- $a + \delta < s$,
- the function g does not have zeros on the interval $(a, a+\delta)$, and
- the following bounds hold for all $x \in (a, a+\delta)$:

$$\left|\frac{f(s)}{g(x)}\right| < \frac{\epsilon}{2}, \quad |L|\left|\frac{g(s)}{g(x)}\right| < \frac{\epsilon}{4}, \quad \text{and} \quad \left|1 - \frac{g(s)}{g(x)}\right| < 2.$$

More so, by the Cauchy mean value theorem, for all $x \in (a, a+\delta)$ there exists a corresponding point $c_x \in (x,s)$ for which

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(s) - f(x)}{g(s) - g(x)},$$

which implies that

$$\left|\frac{f(s) - f(x)}{g(s) - g(x)} - L\right| = \left|\frac{f'(c_x)}{g'(c_x)} - L\right| \overset{\text{(a)}}{<} \frac{\epsilon}{8}.$$

where (a) follows from the bound in Equation 7.5. For all $x \in (a, a+\delta)$ we can then complete the proof with the following, rather elaborate, calculation:

$$\begin{aligned}
\left|\frac{f(x)}{g(x)} - L\right| &\overset{\text{(a)}}{=} \left|\left(\frac{f(s)-f(x)}{g(s)-g(x)}\right)\left(\frac{g(s)-g(x)}{-g(x)}\right) + \frac{f(s)}{g(x)} - L\right| \\
&\overset{\text{(b)}}{=} \left|\left(\frac{f(s)-f(x)}{g(s)-g(x)} - L + L\right)\left(1 - \frac{g(s)}{g(x)}\right) + \frac{f(s)}{g(x)} - L\right| \\
&\overset{\text{(c)}}{=} \left|\left(\frac{f(s)-f(x)}{g(s)-g(x)} - L\right)\left(1 - \frac{g(s)}{g(x)}\right) - L\frac{g(s)}{g(x)} + \frac{f(s)}{g(x)}\right| \\
&\overset{\text{(d)}}{\leq} \left|\frac{f(s)-f(x)}{g(s)-g(x)} - L\right|\left|1 - \frac{g(s)}{g(x)}\right| + |L|\left|\frac{g(s)}{g(x)}\right| + \left|\frac{f(s)}{g(x)}\right| \\
&\overset{\text{(e)}}{<} \frac{\epsilon}{8}\cdot 2 + \frac{\epsilon}{4} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where (a) follows from adding and subtracting the quantity $f(s)/g(x)$, (b) from adding and subtracting L, (c) from multiplying out and simplifying, (d) from the triangle inequality, and (e) from the various bounds developed earlier in the proof. With the intermediate steps removed, the result means that

$$\lim_{x\to a^+} \frac{f(x)}{g(x)} = L = \lim_{x\to a^+} \frac{f'(x)}{g'(x)},$$

proving point (i).

(ii) Analogous to the proof of (i). □

Proposition 7.37: L'Hôpital's rule for divergent functions II

Let U be an open subset of $\mathbb{R}$, define the point $a \in U$, and let $f, g\colon U\setminus\{a\} \to \mathbb{R}$ be two differentiable functions, for which $g'(x) \neq 0$ for all $x \in U\setminus\{a\}$ and $\lim_{x\to a} |g(x)| = \infty$. If the limit of the function f'/g' exists at a, then the limit of the function f/g also exists at a and is equal to

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

Proof. We apply Proposition 7.36 to the left-sided limit, and then to the right-sided limit, of the functions f/g and f'/g' at a. □

Proposition 7.38: L'Hôpital's rule for divergent functions at infinity

Let $f, g\colon U \to \mathbb{R}$ be two differentiable functions defined on the open interval $U \subset \mathbb{R}$, where $g'(x) \neq 0$ for all $x \in U$. We then separately consider the following cases:

(i) Assume that the interval U is unbounded above and that $\lim_{x\to\infty} |g(x)| = \infty$. If the limit of the function f'/g' exists at infinity, then the limit of the function f/g also exists at infinity and is equal to

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

(ii) Assume that the interval U is unbounded below and that $\lim_{x\to-\infty}|g(x)| = \infty$. If the limit of the function f'/g' exists at negative infinity, then the limit of the function f/g also exists at negative infinity and is equal to

$$\lim_{x\to-\infty}\frac{f(x)}{g(x)} = \lim_{x\to-\infty}\frac{f'(x)}{g'(x)}.$$

Proof.

(i) We use Proposition 7.36 (ii) to directly compute

$$\begin{aligned}\lim_{x\to\infty}\frac{f(x)}{g(x)} &\overset{\text{(a)}}{=} \lim_{t\to0^+}\frac{f(1/t)}{g(1/t)} \overset{\text{(b)}}{=} \frac{f'(1/t)(-1/t^2)}{g'(1/t)(-1/t^2)}\\ &= \lim_{t\to0^+}\frac{f'(1/t)}{g'(1/t)} = \lim_{x\to\infty}\frac{f'(x)}{g'(x)},\end{aligned}$$

where in (a) we have introduced the new variable $t = 1/x$ and (b) follows from Proposition 7.36 (ii) and the chain rule.

(ii) Analogous to the proof of (i). □

We conclude this section with some practical applications of L'Hôpital's rule.

Example 7.39: Using L'Hôpital's rule

(1) Consider the rational function $f\colon \mathbb{R}\setminus\{-1,1\}\to\mathbb{R}$ given by

$$f(x) = \frac{x^3-1}{x^2-1}.$$

The point $x = 1$ is a zero of both the numerator and denominator, so the expression for $f(x)$ can be factored and simplified as follows:

$$f(x) = \frac{x^3-1}{x^2-1} = \frac{(x-1)(x^2+x+1)}{(x-1)(x+1)} = \frac{x^2+x+1}{x+1}.$$

The simplified expression on the right-hand side *is* defined at $x = 1$, and its value is $3/2$. We could also have found this expression by applying L'Hôpital's rule to the limit of the function f at the point $x = 1$, i.e.

$$\lim_{x\to1}\frac{x^3-1}{x^2-1} = \lim_{x\to1}\frac{3x^2}{2x} = \frac{3}{2}.$$

(2) We can use L'Hôpital's rule to compute

$$\lim_{x\to0}\frac{\mathrm{e}^{ax}-1}{x} = \lim_{x\to0}\frac{a\mathrm{e}^{ax}}{1} = a \text{ for all } a\in\mathbb{R}.$$

(3) Let us use L'Hôpital's rule to compute the limit of the function $x\mapsto(\sin x)/x$ at the point $x = 0$:

$$\lim_{x\to0}\frac{\sin x}{x} = \lim_{x\to0}\frac{\cos x}{1} = 1.$$

(4) Consider the function $f\colon \mathbb{R}^+ \to \mathbb{R}$ given by $f(x) = x \ln x$. Using L'Hôpital's rule, the function's limit at zero is

$$\begin{aligned}\lim_{x\to 0^+} (x \ln x) &\overset{\text{(a)}}{=} \lim_{x\to 0^+} \frac{\ln x}{(1/x)} \overset{\text{(b)}}{=} \lim_{x\to 0^+} \frac{(1/x)}{(-1/x^2)} \\ &= \lim_{x\to 0^+} (-x) = 0,\end{aligned}$$

where (a) follows from writing x as $1/(1/x)$ and (b) from L'Hôpital's rule.

(5) For any natural number n, we can iteratively compute the following limit with multiple applications of L'Hôpital's rule:

$$\begin{aligned}\lim_{x\to\infty} \frac{x^n}{\mathrm{e}^x} &= \lim_{x\to\infty} \frac{x x^{n-1}}{\mathrm{e}^x} \\ &= \lim_{x\to\infty} \frac{n(n-1)x^{n-1}}{\mathrm{e}^x} = \cdots = \lim_{x\to\infty} \frac{n!}{\mathrm{e}^x} = 0.\end{aligned}$$

As x approaches infinity, the exponential function grows faster than the power function for any $n \in \mathbb{N}$, and so the denominator is larger than the numerator and the limit is zero for arbitrary n.

7.3.4 Analysis of scalar functions

The results of this chapter provide the tools to analyze an arbitrary elementary scalar function and at least approximately sketch the function's graph. A typical analysis process includes the following points:

(i) If the function's domain is not explicitly given, we define the domain as the set of points for which the function's formula is defined.

(ii) We compute the function's zeros.

(iii) We use the limit to analyze the function's behavior at the boundary points of its domain and at positive and negative infinity, if applicable.

(iv) We compute the function's derivative and find the function's stationary points, intervals of increasing and decreasing growth, and local and global extrema.

(v) We compute the function's second derivative and find the function's points of inflection and intervals of concavity and convexity.

(vi) We sketch the functions graph on the basis of the above points.

8 The indefinite integral

Definition 8.1: Primitive function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open interval $U \subset \mathbb{R}$. A function $F\colon U \to \mathbb{R}$ is called a *primitive function* of f if F is differentiable and

$$F' = f.$$

Proposition 8.2: Primitive functions are defined up to a locally constant function

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open interval $U \subset \mathbb{R}$.

(i) If $F, G\colon U \to \mathbb{R}$ are two primitive functions of f, then the function $F - G$ is locally constant.

(ii) If $F\colon U \to \mathbb{R}$ is a primitive function of f and $C\colon U \to \mathbb{R}$ is an arbitrary locally constant function, then the sum $F + C$ is also a primitive function of F.

Proof.

(i) This point follows directly from the calculation

$$(F - G)' = F' - G' \overset{\text{(a)}}{=} f - f = 0,$$

where (a) holds because F and G are both primitive functions of f. Since $(F - G)' = 0$, the function $F - G$ is locally constant by Corollary 7.23.

(ii) This point follows directly from the calculation

$$(F + C)' = F' + C' \overset{\text{(a)}}{=} f + 0 = f,$$

where in (a) $C' = 0$ because C is locally constant. □

Proposition 8.2 implies that if a function $f\colon U \to \mathbb{R}$ has at least one primitive function, it in fact has infinitely many primitive functions, which can be parametererized by the set of all locally constant functions defined on f's domain $U \subset \mathbb{R}$.

Topic 8.1: Indefinite integration

Let $f\colon U \to \mathbb{R}$ be a scalar function defined on the open subset $U \subset \mathbb{R}$. The set of all of f's primitive functions is called f's *indefinite integral* is denoted by

$$\int f(x)\,\mathrm{d}x.$$

In this context, the process of finding the function f's primitive function is called *integration*, and the function $f\colon x \mapsto f(x)$ is called the *integrand*.

Because primitive functions are defined only up to a locally constant function, the problem of finding the primitive functions of a given function $f\colon U \to \mathbb{R}$ reduces to finding only *one*

of f's primitive functions, say F; we can then write f's indefinite integral in the form

$$\int f(x)\,\mathrm{d}x = F + C,$$

where $C\colon U \to \mathbb{R}$ is an arbitrary locally constant function.

Remark 8.3: Notation

The notation used in indefinite integration calls for a few comments. The symbol $\int$ is called an *integral symbol* and the symbol $\mathrm{d}x$ is called a *differential*. The x in the differential $\mathrm{d}x$ explicitly states that the integration is carried out over the variable x. Why the integral symbol and differential look like they do will make more sense after we cover definite integration and the relationship between definite and indefinite integration in Chapter 9.

In passing, we note that an expression of the form $f(x)\,\mathrm{d}x$ is called a *differential 1-form*. For any differentiable function $f\colon U \to \mathbb{R}$ defined on the open subset $U \subset \mathbb{R}$, the differential one form

$$\mathrm{d}f = f'(x)\,\mathrm{d}x$$

is called the *differential* of the function f. A precise definition of differential forms and their calculus is more natural in higher dimensions; for the time being, we mention only that the product of the differential 1-form $\mathrm{d}f$ with an arbitrary function $g\colon U \to \mathbb{R}$ is again a differential given by

$$g\,\mathrm{d}f = \big[g(x)f'(x)\big]\,\mathrm{d}x.$$

Not every function has a set of primitive functions, but, as we will show later, every continuous scalar function defined on an open subset of $\mathbb{R}$ *does* have a primitive function. We dedicate the following sections largely to the computation of these primitive functions.

8.1 Elementary integration rules

One can fairly straightforwardly reverse engineer the set of elementary derivative rules in Section 7.2.1 to produce a corresponding set of elementary indefinite integrals. We first present these elementary integrals, and then cover more integration techniques in the following sections.

Example 8.4: Elementary indefinite integrals

(1) The indefinite integrals of power functions are given by

$$\int x^r\,\mathrm{d}x = \frac{x^{r+1}}{r+1} + C, \quad r \in \mathbb{R}\setminus\{-1\},$$
$$\int \frac{\mathrm{d}x}{x} = \ln|x| + C.$$

(2) The indefinite integrals of exponential functions are given by

$$\int a^x\,\mathrm{d}x = \frac{a^x}{\ln a} + C, \quad a \in \mathbb{R}^+\setminus\{1\},$$
$$\int \mathrm{e}^x\,\mathrm{d}x = \mathrm{e}^x + C.$$

(3) The indefinite integrals of the sine and cosine functions are

$$\int \sin x \, \mathrm{d}x = -\cos x + C \quad \text{and} \quad \int \cos x \, \mathrm{d}x = \sin x + C.$$

(4) The indefinite integrals of the secant and cosecant functions are

$$\int \frac{\mathrm{d}x}{\cos^2 x} = \tan x + C \quad \text{and} \quad \int \frac{\mathrm{d}x}{\sin^2 x} = -\cot x + C.$$

(5) The indefinite integrals of some selected rational functions are

$$\begin{aligned}
\int \frac{\mathrm{d}x}{\sqrt{1-x^2}} &= \arcsin x + C, \\
\int \frac{\mathrm{d}x}{1+x^2} &= \arctan x + C, \\
\int \frac{\mathrm{d}x}{\sqrt{x^2+a^2}} &= \ln\left(x + \sqrt{x^2+a^2}\right) + C, \quad a \in \mathbb{R}^+ \\
\int \frac{\mathrm{d}x}{\sqrt{x^2-a^2}} &= \ln\left(x + \sqrt{x^2-a^2}\right) + C, \quad a \in \mathbb{R}^+.
\end{aligned}$$

Each case requires some caution regarding the domain of the relevant functions: the functions on the right-hand side of each equality are the primitive functions of the corresponding integrands on the left-hand side only on open subsets consisting of those points on which the functions on the right-hand side are differentiable.

The integral of the function $x \mapsto 1/x$ perhaps deserves a special comment. We can directly compute via differentiation that $x \mapsto \ln|x| + C$ is indeed a primitive function of $x \mapsto 1/x$; for $x \in R^-$ this calculation reads

$$(\ln|x|)' \overset{\text{(a)}}{=} \big[\ln(-x)\big]' \overset{\text{(b)}}{=} \frac{1}{(-x)} \cdot (-x)' = \frac{1}{x} \quad x \in \mathbb{R}^-,$$

where (a) holds for $x \in \mathbb{R}^-$ and (b) follows from the chain rule. Similarly for, $x \in \mathbb{R}^+$,

$$(\ln|x|)' \overset{\text{(a)}}{=} (\ln x)' = \frac{1}{x} \cdot x' = \frac{1}{x}, \quad x \in \mathbb{R}^+.$$

where (a) holds for $x \in \mathbb{R}^+$. The function $x \mapsto \ln|x|$ is defined for $\mathbb{R} \setminus \{0\}$, and so

$$\int \frac{\mathrm{d}x}{x} = \ln|x| + C \text{ for all } x \in \mathbb{R} \setminus \{0\}.$$

Because the set $\mathbb{R} \setminus \{0\}$ is the union of the two disjoint intervals $\mathbb{R}^-$ and $\mathbb{R}^+$, the locally constant function $C\colon \mathbb{R} \setminus \{0\}$ is of the form

$$C(x) = \begin{cases} C_1 & x \in \mathbb{R}^- \\ C_2 & x \in \mathbb{R}^+, \end{cases}$$

where C_1 and C_2 are two arbitrary locally constant functions.

Proposition 8.5: The indefinite integral is linear

Let $f, g\colon U \to \mathbb{R}$ be two scalar functions defined on the open subset $U \subset \mathbb{R}$, and let $\alpha, \beta \in \mathbb{R}$ be two real constants. If both f and g have primitive functions, then the function $\alpha f + \beta g$ also has a primitive function and obeys

$$\int \big[\alpha f(x) + \beta g(x)\big]\,\mathrm{d}x = \alpha \int f(x)\,\mathrm{d}x + \beta \int g(x)\,\mathrm{d}x.$$

Proof. Let F and G be primitive functions of f and g, respectively. In this case

$$(\alpha F + \beta G)' \overset{\text{(a)}}{=} \alpha F' + \beta G' = \alpha f + \beta g,$$

where (a) follows from the linearity of the derivative. □

Example 8.6: Integral of a polynomial

By the linearity of the indefinite integral in Proposition 8.5, for all $n \in \mathbb{N}$ the indefinite integral of a real polynomial of the form $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is

$$\begin{aligned}\int P(x)\,\mathrm{d}x &= a_n \int x^n\,\mathrm{d}x + \cdots + a_1 \int x\,\mathrm{d}x + a_0 \int 1\,\mathrm{d}x \\ &= a_n \frac{x^{n+1}}{n+1} + a_{n-1}\frac{x^n}{n} + \cdots + a_1 \frac{x^2}{2} + a_0 x + C.\end{aligned}$$

The primitive function of a n-th degree polynomial is thus an $(n+1)$-th degree polynomial.

We now cover the very useful technique of integration by parts, which can be viewed loosely as a reverse-engineered form of the product rule for differentiation.

Proposition 8.7: Integration by parts

Let $f, g\colon U \to \mathbb{R}$ be two differentiable scalar functions defined on the open subset $U \subset \mathbb{R}$. If the function $f'g$ has a primitive function, then the function fg' also has a primitive function and obeys

$$\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x. \tag{8.1}$$

Proof. We will show by direct calculation that the derivative of the right-hand side of Equation 8.1 equals the integrand on the left-hand side:

$$\begin{aligned}\left[f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x\right]' &\overset{\text{(a)}}{=} \big[f(x)g(x)\big]' - \left[\int f'(x)g(x)\,\mathrm{d}x\right]' \\ &\overset{\text{(b)}}{=} \big[f(x)g(x)\big]' - f'(x)g(x) \\ &\overset{\text{(c)}}{=} f(x)g'(x) + f'(x)g(x) - f'(x)g(x) \\ &= f(x)g'(x),\end{aligned}$$

where (a) follows from the linearity of the derivative, (b) from the definition of the indefinite integral, and (c) from the product rule for differentiation. □

The rule for integration by parts is commonly written in terms of differential 1-forms in the form

$$\int f\,\mathrm{d}g = fg - \int g\,\mathrm{d}f.$$

In this context, when using integration by parts to compute primitive functions one must write the differential 1-form being integrated in the form $f\,\mathrm{d}g$. Intuition for productively choosing f and $\mathrm{d}g$ comes with experience, and we now offer a few examples.

Example 8.8: Using integration by parts

(i) Let us use integration by parts to compute the integral

$$\int x^r \ln x\,\mathrm{d}x, \quad r \in \mathbb{R} \setminus \{-1\}.$$

We choose $f(x) = \ln x$ and $g'(x) = x^r$ (so that $\mathrm{d}g = x^r\,\mathrm{d}x$), and first compute

$$\mathrm{d}f = f'(x)\,\mathrm{d}x = \frac{1}{x}\,\mathrm{d}x.$$

We then find a primitive function of $g'(x) = x^r$, for example

$$g(x) = \frac{x^{r+1}}{r+1},$$

and then apply the integration by parts formula to get

$$\begin{aligned}
\int x^r \ln x\,\mathrm{d}x = \int f\,\mathrm{d}g &= fg - \int g\,\mathrm{d}f \\
&= (\ln x)\cdot\frac{x^{r+1}}{r+1} - \int \frac{x^{r+1}}{r+1}\cdot\frac{1}{x}\,\mathrm{d}x \\
&= \frac{x^{r+1}\ln x}{r+1} - \int \frac{x^r}{r+1}\,\mathrm{d}x \\
&= \frac{x^{r+1}\ln x}{r+1} - \frac{x^{r+1}}{(r+1)^2} + C \\
&= \frac{x^{r+1}}{r+1}\left(\ln x - \frac{1}{r+1}\right) + C.
\end{aligned}$$

When $r = 0$, this result simplifies to

$$\int \ln x\,\mathrm{d}x = x(\ln x - 1) + C.$$

(ii) We will now solve the integral

$$\int x\sin x\,\mathrm{d}x.$$

We choose $f(x) = x$ and $g'(x) = \sin x$, so that $\mathrm{d}g = \sin x\,\mathrm{d}x$. In this case $\mathrm{d}f = \mathrm{d}x$, $g(x) = -\cos x$, and integration by parts produces

$$\begin{aligned}
\int x\sin x\,\mathrm{d}x = \int f\,\mathrm{d}g &= fg - \int g\,\mathrm{d}f \\
&= -x\cos x + \int \cos x\,\mathrm{d}x \\
&= -x\cos x + \sin x + C.
\end{aligned}$$

(iii) Finally, we will compute the indefinite integral of the function $x \mapsto x^2 \mathrm{e}^x$ using two successive applications of integration by parts, first choosing $f(x) = x^2$ and $\mathrm{d}g = \mathrm{e}^x \,\mathrm{d}x$, and in the second application choosing $f(x) = 2x$ and $\mathrm{d}g = \mathrm{e}^x \,\mathrm{d}x$. The computation reads

$$\begin{aligned}\int x^2 \mathrm{e}^x \,\mathrm{d}x &= x^2 \mathrm{e}^x - \int 2x \mathrm{e}^x \,\mathrm{d}x \\ &= x^2 \mathrm{e}^x - \left(2x \mathrm{e}^x - \int 2 \mathrm{e}^x \,\mathrm{d}x \right) \\ &= \mathrm{e}^x (x^2 - 2x + 2) + C.\end{aligned}$$

The following, also very useful, technique may be viewed loosely as a reverse-engineered form of the chain rule for differentiation.

Proposition 8.9: Integration by change of variables

Let $F \colon U \to \mathbb{R}$ be a primitive function of the scalar function $f \colon U \to \mathbb{R}$ defined on the open subset $U \subset \mathbb{R}$, let $s \colon V \to \mathbb{R}$ be a differentiable function defined on the open subset $V \subset \mathbb{R}$, and assume $s(V) \subset U$. In this case the function $(f \circ s) \cdot s'$ has a primitive function and obeys

$$\int f\big(s(x)\big) s'(x) \,\mathrm{d}x = F \circ s + C.$$

Proof. We use the chain rule to directly verify

$$(F \circ s + C)' = (F' \circ s) \cdot s' = (f \circ s) \cdot s'. \qquad \square$$

It is instructive to write the formula for integration by change of variables in the form

$$\begin{aligned}\int f\big(s(x)\big) s'(x) \,\mathrm{d}x &\overset{\text{(a)}}{=} \int f\big(s(x)\big) \,\mathrm{d}s \\ &\overset{\text{(b)}}{=} \int f(s) \,\mathrm{d}s \\ &= F(s) + C \\ &\overset{\text{(c)}}{=} F\big(s(x)\big) + C,\end{aligned}$$

where in (a) we have introduced the differential 1-form $\mathrm{d}s = s'(x) \,\mathrm{d}x$, in (b) we have dropped the variable x and integrated over the new variable s, and in (c) we have again made explicit that s is itself a function of x, which is justified by Proposition 8.9. The technique gets its name by changing the variable of integration from x to s; this integration technique is also commonly called *integration by substitution*.

Choosing an appropriate function s when using integration by substitution is often a matter of experience, and we give a few examples below.

Example 8.10: Using integration by change of variables

(1) Let us use change of variables to solve the integral

$$\int \frac{\mathrm{d}x}{x-a}, \quad a \in \mathbb{R}.$$

We can solve the integral by introducing the new variable $s = x - a$, in which case $\mathrm{d}s = \mathrm{d}x$ and Proposition 8.9 gives

$$\int \frac{\mathrm{d}x}{x-a} = \int \frac{\mathrm{d}s}{s} = \ln|s| + C = \ln|x-a| + C.$$

(2) Let us now solve the integral

$$\int \frac{\mathrm{d}x}{a^2+x^2} = \int \frac{\mathrm{d}x}{a^2\left[1+(x/a)^2\right]}, \quad a \in \mathbb{R}^+.$$

We will use the new variable $s = x/a$, in which case $\mathrm{d}s = (1/a)\,\mathrm{d}x$ and

$$\begin{aligned}\int \frac{\mathrm{d}x}{a^2+x^2} &= \frac{1}{a^2}\int \frac{\mathrm{d}x}{1+(x/a)^2} \\ &= \frac{1}{a}\int \frac{\mathrm{d}s}{1+s^2} = \frac{1}{a}\arctan(s) + C \\ &= \frac{1}{a}\arctan\left(\frac{x}{a}\right) + C.\end{aligned}$$

(3) Finally, we will solve the integral

$$\int \tan x\,\mathrm{d}x = \int \frac{\sin x}{\cos x}\,\mathrm{d}x.$$

We choose the new variable $t = \cos x$, in which case $\mathrm{d}t = -\sin x\,\mathrm{d}x$ and

$$\begin{aligned}\int \tan x\,\mathrm{d}x &= \int \frac{\sin x}{\cos x}\,\mathrm{d}x \\ &= \int \frac{-\,\mathrm{d}t}{t} = -\ln|t| + C \\ &= -\ln|\cos x| + C.\end{aligned}$$

The domain of the tangent function is a countably infinite union of open, pairwise disjoint integrals, so in this case the locally constant function C is determined by a countably infinite number of constants, one for each of the intervals in the tangent function's domain.

8.2 Integration of rational functions

This chapter covers techniques that make possible the integration of an arbitrary rational function—it turns out that the integral of an arbitrary rational function can always be written as a linear combination of rational functions and inverse trigonometric functions. We first consider two important cases, which we will use later.

Example 8.11: Two important rational integrals

We will make use of the following two integrals later in this section.

(i) For any two real numbers $a, A \in \mathbb{R}$ and any natural number $k \geq 2$ it holds that

$$\int \frac{A\,\mathrm{d}x}{(x-a)^k} = -\frac{A}{(k-1)(x-a)^{k-1}} + C,$$

and, when $k = 1$, it holds that

$$\int \frac{A\,\mathrm{d}x}{x-a} = A \ln|x-a| + C.$$

(ii) Let $M, N \in \mathbb{R}$ be two arbitrary real numbers, and let $p, q \in \mathbb{R}$ be two real numbers such that the function

$$x \mapsto x^2 + px + q$$

is an arbitrary second-degree polynomial without any real zeros; in this case the polynomial's discriminant is necessarily negative, i.e. $D = p^2 - 4q < 0$.

Let us then solve the indefinite integral

$$\int \frac{Mx+N}{x^2+px+q}\,\mathrm{d}x.$$

We begin by writing the denominator in the form

$$\begin{aligned} x^2 + px + q &= \left(x + \frac{p}{2}\right)^2 - \frac{p^2}{4} + q \\ &= \left(x + \frac{p}{2}\right)^2 + \frac{-p^2+4q}{4} \\ &\overset{\text{(a)}}{=} t^2 + \mu^2, \end{aligned}$$

where in (a) we have defined the quantities

$$t = x + \frac{p}{2} \quad \text{and} \quad \mu = \sqrt{\frac{-p^2+4q}{4}} = \sqrt{\frac{-D}{4}}.$$

In terms of the new variable t and the constant μ, the integral becomes

$$\begin{aligned} \int \frac{Mx+N}{x^2+px+q}\,\mathrm{d}x &= \int \frac{Mt - (Mp)/2 + N}{t^2+\mu^2}\,\mathrm{d}t \\ &= M\int \frac{t\,\mathrm{d}t}{t^2+\mu^2} + \left(N - \frac{Mp}{2}\right)\int \frac{\mathrm{d}t}{t^2+\mu^2}. \end{aligned}$$

Borrowing the result of Example 8.10 (2), the second integral comes out to

$$\int \frac{\mathrm{d}t}{t^2+\mu^2} = \frac{1}{\mu}\arctan\left(\frac{t}{\mu}\right) + C,$$

while we can solve the first integral by substitution using the new variable $s^2 = t^2 + \mu^2 \implies \mathrm{d}s = 2t\,\mathrm{d}t$. The calculation reads

$$\int \frac{\mathrm{d}t}{t^2+\mu^2} = \frac{1}{2}\int \frac{\mathrm{d}s}{s} = \frac{1}{2}\ln|s| + C = \frac{1}{2}\ln\left(t^2+\mu^2\right) + C.$$

The complete result is then

$$\int \frac{Mx+N}{x^2+px+q}\,\mathrm{d}x = \frac{M}{2}\ln\left(x^2+px+q\right) + \frac{2N-Mp}{\sqrt{-p^2+4q}}\arctan\left(\frac{2x+p}{\sqrt{-p^2+4q}}\right)+C$$

We now show how to manipulate a general rational function into a form suitable for integrating. In general, a scalar rational function is a quotient of two real polynomials P and Q and takes the general form

$$x \mapsto \frac{P(x)}{Q(x)}.$$

In the following discussion we assume without loss of generality that:

- the degree of the denominator polynomial Q is at least 1 (if Q were of degree 0, i.e. simply a constant function, then P/Q would be a real polynomial, which we have already shown how to integrate in Example 8.6), and
- the leading coefficient of Q is 1 (if it were not, and were equal to some other number, say $a \in \mathbb{R}\setminus\{1\}$, we could simply factor out a from each of Q's terms; in the resulting polynomial $(P/a)/\widetilde{Q}$ the new denominator $\widetilde{Q} = Q/a$ would have leading coefficient equal to one).

Topic 8.2: Simplifying rational functions I

We aim to rearrange P/Q into a form conducive to integration; a good first step is to simplify P/Q with polynomial division, which in general produces two real polynomials S and R for which

$$P = SQ + R, \quad \deg(R) < \deg(Q).$$

The integral of a general rational function can thus be written in the form

$$\int \frac{P(x)}{Q(x)}\,\mathrm{d}x = \int S(x)\,\mathrm{d}x + \int \frac{R(x)}{Q(x)}\,\mathrm{d}x, \quad \deg(R) < \deg(Q).$$

The polynomial S may be integrated as in Example 8.6); it remains to integrate the rational function R/Q, where we are now sure that $\deg(R) < \deg(Q)$.

Topic 8.3: Simplifying rational functions II

It is possible to further simplify the expression R/Q by finding the zeros of the polynomial Q and then writing Q as a product of irreducible terms of degree 1 or 2. Such a decomposition of Q is always possible by the *fundamental theorem of algebra*, which states that every complex polynomial of positive degree has at least one complex zero. Because non-real zeros of a real polynomial always occur in complex conjugate pairs, we may then write any real polynomial Q as a product of the form

$$Q(x) = (x-x_1)^{\alpha_1}\cdots(x-x_n)^{\alpha_n}\cdot(x^2+p_1x+q_1)^{\beta_1}\cdots(x^2+p_mx+q_m)^{\beta_m},$$

where:

- $x_1, \ldots, x_n$ are unique real zeros of Q;
- the terms $(x^2 + p_1 x + q_1), \cdots, (x^2 + p_m x + q_m)$ are unique real polynomials of degree 2 without real zeros;
- $n, m \in \mathbb{N} \cup \{0\}$; and
- $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m \in \mathbb{N}$.

For each natural number $i \leq n$, the exponent α_i is called the *multiplicity* of the corresponding zero x_i. Finally, we note that because the polynomial $x^2 + p_j x + q_j$ does not have real zeros, it has a negative discriminant, i.e.

$$D_j = p_j^2 - 4q_j < 0 \text{ for all } j \leq m.$$

The remaining procedure for integrating the rational function R/Q depends on the decomposition of the polynomial Q. In certain cases it is easiest to compute the integral of R/Q using a partial fraction decomposition. In more general cases the integral of R/Q can be computed using an ansatz.

8.2.1 Partial fraction decomposition

In this section we will consider how to solve the integration problem

$$\int \frac{P(x)}{Q(x)}\,\mathrm{d}x = \int S(x)\,\mathrm{d}x + \int \frac{R(x)}{Q(x)}\,\mathrm{d}x,$$

where S, R, and Q are polynomials and $\deg(R) < \deg(Q)$. For review, the denominator Q can be written in the general form

$$Q(x) = (x - x_1)^{\alpha_1} \cdots (x - x_n)^{\alpha_n} \cdot (x^2 + p_1 x + q_1)^{\beta_1} \cdots (x^2 + p_m x + q_m)^{\beta_m},$$

where:

- $x_1, \ldots, x_n$ are unique real zeros of Q;
- the terms $(x^2 + p_1 x + q_1), \cdots, (x^2 + p_m x + q_m)$ are unique real polynomials of degree 2 without real zeros;
- $n, m \in \mathbb{N} \cup \{0\}$; and
- $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m \in \mathbb{N}$.

Topic 8.4: Partial fraction decomposition when $m = 0$

We first consider the case when $m = 0$, in which case Q can be factored into the product of first-degree polynomials with only real zeros in the form

$$Q(x) = (x - x_1)^{\alpha_1} \cdot (x - x_n)^{\alpha_n}.$$

In this case the rational function R/Q can be written as a sum of partial fractions:

$$\begin{aligned}\frac{R(x)}{Q(x)} = &\frac{A_{11}}{(x-x_1)} + \frac{A_{12}}{(x-x_1)^2} + \cdots + \frac{A_{1\alpha_1}}{(x-x_1)^{\alpha_1}} \\ &+ \frac{A_{21}}{(x-x_2)} + \frac{A_{22}}{(x-x_2)^2} + \cdots + \frac{A_{2\alpha_2}}{(x-x_2)^{\alpha_2}} \\ &+ \quad \cdots \\ &+ \frac{A_{n1}}{(x-x_n)} + \frac{A_{n2}}{(x-x_n)^2} + \cdots + \frac{A_{n\alpha_n}}{(x-x_n)^{\alpha_n}},\end{aligned}$$

where $A_{11}, \ldots, A_{1\alpha_1}, A_{21}, \ldots, A_{n\alpha_n} \in \mathbb{R}$ are uniquely determined real constants. These constants can by computed by forming a common denominator on the right-hand side of the above equation for $R(x)/Q(x)$, and then equating the numerators of the left-hand side and right-hand side; this process produces a system of linear equations that uniquely determines the constants.

With the constants $A_{11}, \ldots, A_{n\alpha_n}$ known, it follows that

$$\int \frac{R(x)}{Q(x)}\,\mathrm{d}x = \int \frac{A_{11}}{x-x_1} + \cdots + \int \frac{A_{n\alpha_n}}{(x-x_n)^{\alpha_n}}\,\mathrm{d}x,$$

where the integrals on the right-hand side can be computes as in Example 8.11 (i).

Topic 8.5: Partial fraction decomposition when $m > 0$ and $\beta_j = 1$

We now assume that $m > 0$ and that each of Q's complex zeros has multiplicity 1, i.e. $\beta_1 = \beta_2 = \cdots = \beta_m = 1$. In this case Q takes the general form

$$Q(x) = (x-x_1)^{\alpha_1} \cdots (x-x_n)^{\alpha_n} \cdot (x^2 + p_1 x + q_1) \cdots (x^2 + p_m x + q_m),$$

and the function R/Q can be written as a partial fraction decomposition of the form

$$\begin{aligned}\frac{R(x)}{Q(x)} = &\frac{A_{11}}{(x-x_1)} + \frac{A_{12}}{(x-x_1)^2} + \cdots + \frac{A_{1\alpha_1}}{(x-x_1)^{\alpha_1}} \\ &+ \frac{A_{21}}{(x-x_2)} + \frac{A_{22}}{(x-x_2)^2} + \cdots + \frac{A_{2\alpha_2}}{(x-x_2)^{\alpha_2}} \\ &+ \quad \cdots \\ &+ \frac{A_{n1}}{(x-x_n)} + \frac{A_{n2}}{(x-x_n)^2} + \cdots + \frac{A_{n\alpha_n}}{(x-x_n)^{\alpha_n}} \\ &+ \frac{M_1 x + N_1}{x^2 + p_1 x + q_1} + \cdots + \frac{M_m x + N_m}{x^2 + p_m x + q_m},\end{aligned}$$

where $A_{11}, \ldots, A_{n\alpha_n}$ and $M_1, \ldots, M_m, N_1, \ldots, N_m \in \mathbb{R}$ are uniquely determined real constants. We compute these constants analogously to the case when $m = 0$, i.e. by forming a common denominator on the right-hand side and then equating the numerators of the left-hand side and right-hand side. With the constants known, the

integral of R/Q can be written in the form

$$\int \frac{R(x)}{Q(x)}\,\mathrm{d}x = \int \frac{A_{11}}{x-x_1} + \cdots + \int \frac{A_{n\alpha_n}}{(x-x_n)^{\alpha_n}}\,\mathrm{d}x + \int \frac{M_1 x + N_1}{x^2 + p_1 x + q_1}\,\mathrm{d}x + \cdots + \int \frac{M_m x + N_m}{x^2 + p_m x + q_m}\,\mathrm{d}x,$$

where each of the integrals may be computed as in Example 8.11.

Example 8.12: Rational integration with partial fraction decomposition

Let us solve the following integral of a rational function:

$$\int \frac{4x+1}{(x+1)^2(x-1)}\,\mathrm{d}x.$$

The degree of the numerator is already less than the degree of the denominator, so an initial polynomial division step is unnecessary. The denominator is already written as a product of irreducible terms, so we may proceed directly to partial fraction decomposition; a decomposition of the integrand reads

$$\begin{aligned}\frac{4x+1}{(x+1)^2(x-1)} &= \frac{A_{11}}{x+1} + \frac{A_{12}}{(x+1)^2} + \frac{A_{21}}{x-2} \\ &= \frac{A_{11}(x+1)(x-2) + A_{12}(x-2) + A_{21}(x+1)^2}{(x+1)^2(x-2)} \\ &= \frac{x^2(A_{11}+A_{21})}{(x+1)^2(x-2)} + \frac{x(-A_{11}+A_{12}+2A_{21})}{(x+1)^2(x-2)} \\ &\qquad + \frac{(-2A_{11}-2A_{12}+A_{21})}{(x+1)^2(x-2)}.\end{aligned}$$

Equating the numerators on the left-hand side and right-hand side produces the system of linear equations

$$\begin{aligned} A_{11} + A_{21} &= 0, \\ -A_{11} + A_{12} + 2A_{21} &= 4, \\ -2A_{11} - 2A_{12} + A_{21} &= 1, \end{aligned}$$

which (without derivation) has the solution $A_{11} = -1$, $A_{12} = 1$, and $A_{21} = 1$. The partial decomposition and subsequent solution of the original integral is then

$$\begin{aligned}\int \frac{4x+1}{(x+1)^2(x-1)}\,\mathrm{d}x &= -\int \frac{\mathrm{d}x}{x+1} + \int \frac{\mathrm{d}x}{(x+1)^2} + \int \frac{\mathrm{d}x}{x-2} \\ &= -\ln|x+1| - \frac{1}{x+1} + \ln|x-2| + C.\end{aligned}$$

When $m > 0$ and at least one of Q's complex zeros has multiplicity greater than 1, i.e. $\beta_j > 1$ for at least one $j = 1, \ldots, m$, it is preferable to proceed with integration using a generic ansatz instead of using partial fraction decomposition.

8.2.2 Using an ansatz to integrate rational functions

In this section we consider the integration problem

$$\int \frac{P(x)}{Q(x)}\,\mathrm{d}x = \int S(x)\,\mathrm{d}x + \int \frac{R(x)}{Q(x)}\,\mathrm{d}x,$$

where S, R, and Q are polynomials, $\deg(R) < \deg(Q)$, and Q can be factored into the form

$$Q(x) = (x - x_1)^{\alpha_1} \cdots (x - x_n)^{\alpha_n} \cdot (x^2 + p_1 x + q_1)^{\beta_1} \cdots (x^2 + p_m x + q_m)^{\beta_m}.$$

In general (without formal proof), the integral of any scalar rational function is a linear combination of rational functions, the natural logarithmic function ln, and the inverse tangent function arctan. As a result, the integral of an arbitrary rational function can technically be written using an ansatz consisting of rational functions, ln, arctan, and unknown constants. This ansatz reads

$$\begin{aligned}\int \frac{R(x)}{Q(x)}\,\mathrm{d}x = {} & \frac{A_s x^s + A_{s-1}x^{s-1} + \cdots + A_1 x + A_0}{(x - x_1)^{\alpha_1 - 1} \cdots (x - x_n)^{\alpha_n - 1}(x^2 + p_1 x + q_1)^{\beta_1 - 1} \cdots (x^2 + p_m x + q_m)^{\beta_m - 1}} \\ & + B_1 \ln|x - x_1| + \cdots + B_n \ln|x - x_n| \\ & + U_1 \ln\!\left(x^2 + p_1 x + q_1\right) + \cdots + U_m \ln\!\left(x^2 + p_m x + q_m\right) \\ & + V_1 \arctan\left(\frac{2x + p_1}{\sqrt{-p_1^2 + 4q_1}}\right) + \cdots + V_m \arctan\left(\frac{2x + p_m}{\sqrt{-p_m^2 + 4q^m}}\right) + C,\end{aligned}$$

where $A_0, A_1, \ldots, A_s, B_1, \ldots, B_n, U_1, \ldots U_m, V_1, \ldots, V_m \in \mathbb{R}$ are real constants, and

$$s = (\alpha_1 - 1) + \cdots + (\alpha_n - 1) + 2(\beta_1 - 1) + \cdots + 2(\beta_m - 1) - 1.$$

The value of s means that the degree of the numerator of the ansatz's rational function is at least one less than the degree of denominator. We compute the constants by differentiating the ansatz's right-hand side, forming a common denominator for the entire resulting derivative, equating the resulting numerator to R, and finally solving the resulting system of linear equations for the unknown constants.

Example 8.13: Integrating rational functions using an ansatz

(1) Let us compute the indefinite integral

$$\int \frac{\mathrm{d}x}{(1 + x^2)^2}.$$

Because the irreducible quadratic term $(1 + x^2)$ has multiplicity, we will use an ansatz instead of integration by parts; the ansatz for this case reads

$$\int \frac{\mathrm{d}x}{(1 + x^2)^2} = \frac{A_1 x + A_0}{1 + x^2} + U_1 \ln\!\left(1 + x^2\right) + V_1 \arctan\left(\frac{2x}{\sqrt{4}}\right) + C. \qquad (8.2)$$

We begin by differentiating the Equation 8.2 to get

$$\begin{aligned}\frac{1}{(1+x^2)^2} &= \frac{A_1(1+x^2)-2x(A_1x+A_0)}{(1+x^2)^2}+\frac{2U_1x}{1+x^2}+\frac{V_1}{1+x^2}\\ &= \frac{A_1(1+x^2)-2x(A_1x+A_0)+2U_1x(1+x^2)+V_1(1+x^2)}{(1+x^2)^2}\\ &= \frac{x^3(2U_1)+x_2(-A_1+V_1)+x(-2A_0+2U_1)+(A_1+V_1)}{(1+x^2)^2}.\end{aligned}$$

We then equate the numerators on the left-hand and right-hand sides to get the system of equations

$$\begin{aligned}2U_1 &= 0\\ -A_1+V_1 &= 0\\ -2A_0+2U_1 &= 0\\ A_1+V_1 &= 1;\end{aligned}$$

without derivation, the solution to this system of equations is

$$A_0=0,\quad A_1=1/2,\quad U_1=0,\quad V_1=1/2.$$

We then substitute these constants into Equation 8.2 to get the solution

$$\int\frac{\mathrm{d}x}{(1+x^2)^2}=\frac{x}{2(1+x^2)}+\frac{1}{2}\arctan x+C. \tag{8.3}$$

(2) Let us once again use an ansatz to compute the indefinite integral

$$\int\frac{2x^3+7x^2+8x+4}{(x^2+2x+2)^2}\,\mathrm{d}x.$$

In this case the appropriate ansatz is

$$\begin{aligned}\int\frac{2x^3+7x^2+8x+4}{(x^2+2x+2)^2}\,\mathrm{d}x = \frac{Ax+B}{x^2+2x+2}&+E\ln\left(x^2+2x+2\right)\\ &+F\arctan\left(\frac{2x+2}{2}\right)+C;\end{aligned} \tag{8.4}$$

we then differentiate Equation 8.4 to get

$$\begin{aligned}\frac{2x^3+7x^2+8x+4}{(x^2+2x+2)^2} &= \frac{A(x^2+2x+2)-(Ax+B)(2x+2)}{(x^2+2x+2)^2}\\ &\quad+\frac{E(2x+2)}{x^2+2x+2}+\frac{F}{1+(x+1)^2}\\ &= \frac{A(x^2+2x+2)-(Ax+B)(2x+2)}{(x^2+2x+2)^2}\\ &\quad+\frac{\left[E(2x+2)+F\right](x^2+2x+2)}{(x^2+2x+2)^2}\\ &= \frac{x^3(2E)+x^2(-A+6E+F)+x(-2B+8E+2F)}{(x^2+2x+2)^2}\\ &\quad+\frac{(2A-2B+4E+2F)}{(x^2+2x+2)^2}.\end{aligned}$$

We then equate the numerators on the left-hand and right-hand sides to get the system of equations

$$\begin{aligned} 2E &= 2 \\ -A + 6E + F &= 7 \\ -2B + 8E + 2F &= 8 \\ 2A - 2B + 4E + 2F &= 4, \end{aligned}$$

which (without derivation) has the solution

$$A = 0, \quad B = 1, \quad E = 1, \quad F = 1.$$

We then substitute these constants into Equation 8.4 to get the solution

$$\int \frac{2x^3 + 7x^2 + 8x + 4}{(x^2 + 2x + 2)^2}\,\mathrm{d}x = \frac{1}{x^2 + 2x + 2} + \ln\left(x^2 + 2x + 2\right) + \arctan(x + 1) + C.$$

8.3 Integrals of selected irrational functions

We will now consider a few special cases of indefinite integrals of irrational and trigonometric functions. We will consider only rational functions whose denominator is the square root of a second-degree polynomial—such functions arise regularly in integration of the equations of motion of physical systems with one degree of freedom.

Example 8.14: Two special cases of irrational functions

We begin by computing the indefinite integrals of three special irrational functions, which we will then use later in the section.

(1) Let us compute the indefinite integral

$$\int \frac{\mathrm{d}x}{\sqrt{x^2 + px + q}}, \tag{8.5}$$

where $p, q \in \mathbb{R}$ are arbitrary real numbers. Let us denote the discriminant of the quadratic polynomial $x^2 + px + q$ by

$$D = p^2 - 4q.$$

The integrand in Equation 8.5 is defined for those values of the variable x for which the polynomial $x^2 + px + q$ is positive—thus the integrand's domain is $\mathbb{R}$ if $D < 0$ and otherwise a union of two disjoint, unbounded, open intervals.

Before computing the integral, we first write the polynomial in the form

$$x^2 = px + q = \left(x + \frac{p}{2}\right)^2 + \frac{-p^2 + 4q}{4} = t^2 + \nu,$$

where we have introduced the new variable t and constant ν given by

$$t = x + \frac{p}{2} \quad \text{and} \quad \nu = \frac{-p^2 + 4q}{4} = -\frac{D}{4}.$$

We then separately consider the following two subcases:

(a) If $D \neq 0$, we can use integration by substitution with the new variable t to get

$$\begin{aligned}\int \frac{\mathrm{d}x}{\sqrt{x^2+px+q}} &= \int \frac{\mathrm{d}t}{\sqrt{t^2+\nu}} \overset{\text{(a)}}{=} \ln\left|t+\sqrt{t^2+\nu}\right| + C \\ &= \ln\left|x+\frac{p}{2}+\sqrt{x^2+px+q}\right| + C,\end{aligned}$$

where in (a) we have used the elementary integral from Example 8.4 (5).

(b) If $D = 0$, then $\nu = 0$, the quantity under the square root is a perfect square, and we can compute

$$\begin{aligned}\int \frac{\mathrm{d}x}{\sqrt{x^2+px+q}} &= \int \frac{\mathrm{d}t}{\sqrt{t^2}} = \int \frac{\mathrm{d}t}{|t|} = \operatorname{sgn}(t)\ln|t| + C \\ &= \operatorname{sgn}\left(x+\frac{p}{2}\right)\ln\left|x+\frac{p}{2}\right| + C.\end{aligned}$$

(2) Next, let us compute the integral

$$\int \frac{\mathrm{d}x}{\sqrt{-x^2+px+q}}, \tag{8.6}$$

where $p, q \in \mathbb{R}$ are again arbitrary real numbers. Let us denote the discriminant of the quadratic polynomial $-x^2+px+q$ by

$$D = p^2 + 4q.$$

The integrand in Equation 8.6 is defined for those values of the variable x for which the polynomial $-x^2+px+q$ is positive, and so the integrand's domain is empty if $D \leq 0$. Thus we will assume that $D > 0$, in which case the integrand's domain is a bounded, open interval.

We first write the polynomial $-x^2+px+q$ in the form

$$\begin{aligned}-x^2+px+q &= -\left(x-\frac{p}{2}\right)^2 + \frac{p^2+4q}{4} \\ &= -(\nu t)^2 + \nu^2 \\ &= \nu^2(1-t^2),\end{aligned}$$

where we have introduced the new variable t and constant η given by

$$t = \frac{x-p/2}{\nu} \quad \text{and} \quad \nu = \sqrt{\frac{p^2+4q}{4}} = \sqrt{\frac{D}{4}}.$$

We may then use integration by substitution to compute

$$\begin{aligned}\int \frac{\mathrm{d}x}{\sqrt{-x^2+px+q}} &= \int \frac{\eta\,\mathrm{d}t}{\sqrt{\eta^2(1-t^2)}} = \int \frac{\mathrm{d}t}{\sqrt{1-t^2}} \overset{\text{(a)}}{=} \arcsin t + C \\ &= \arcsin\left(\frac{2x-p}{\sqrt{p^2+4q}}\right) + C,\end{aligned}$$

where in (a) we have gain used an integral from Example 8.4 (5).

Points (1)a and (2) in Example 8.14 allow us to compute the integral

$$\int \frac{\mathrm{d}x}{\sqrt{ax^2+bx+c}},$$

where $a, b, c \in \mathbb{R}$ are arbitrary real numbers. Namely:

- If $a > 0$, we factor out a from the square root and use point (1)a.
- If $a < 0$, we factor out $|a|$ from the square root and use point (2).
- If $a = 0$ and $b \neq 0$, we may compute the integral by substitution, using the new variable $t = bx + c$.
- If $a = b = 0$ and $c \neq 0$, the integrand is simply a constant function, and

$$\int \frac{1}{\sqrt{c}}\,\mathrm{d}x = \frac{x}{\sqrt{c}} + C.$$

- If $a = b = c = 0$, then the integrand has an empty domain, and so the integral is not well-defined.

8.3.1 Using an ansatz to integrate irrational functions

Let P be an arbitrary polynomial of degree $n \geq 1$ and let $a, b, c \in \mathbb{R}$ be any three real numbers for which $a \neq 0$. In this case we may use an ansatz compute the indefinite integral

$$\int \frac{P(x)\,\mathrm{d}x}{\sqrt{ax^2+bx+c}}.$$

The approriate the ansatz is

$$\int \frac{P(x)\,\mathrm{d}x}{\sqrt{ax^2+bx+c}} = T(x)\sqrt{ax^2+bx+c} + K\int \frac{\mathrm{d}x}{\sqrt{ax^2+bx+c}}, \tag{8.7}$$

where T is a real polynomial of degree at most $n-1$ and $K \in \mathbb{R}$ is an arbitrary real constant. We compute the coefficients of the polynomial T and the value of the constant K by first differentiating Equation 8.7, arranging the right-hand side to use the common demoninator $\sqrt{ax^2+bx+c}$, equating the resulting numerator of the right-hand to $P(x)$ and then solving the resulting system of linear equations. We can compute the remaining integral on the right-hand side of Equation 8.7 using the results of Example 8.14.

Example 8.15: Integrating an irrational function using an ansatz

Let us use an ansatz to compute the indefinite integral

$$\int \sqrt{\alpha^2 - x^2}\,\mathrm{d}x,$$

where $\alpha \in \mathbb{R} \setminus \{0\}$ is an arbitrary nonzero constant. We first write the integral in the form

$$\int \sqrt{\alpha^2 - x^2}\,\mathrm{d}x = \int \frac{\alpha^2 - x^2}{\sqrt{\alpha^2 - x^2}}\,\mathrm{d}x,$$

which allows to then introduce the ansatz

$$\int \frac{\alpha^2 - x^2}{\sqrt{\alpha^2 - x^2}}\,\mathrm{d}x = (Ax+B)\sqrt{\alpha^2 - x^2} + K\int \frac{\mathrm{d}x}{\sqrt{\alpha^2 - x^2}}. \tag{8.8}$$

We will first compute the values of the constants A, B, and K; we begin by differentiating Equation 8.8 to get

$$\begin{aligned}\frac{\alpha^2 - x^2}{\sqrt{\alpha^2 - x^2}} &= A\sqrt{\alpha^2 - x^2} + \frac{(Ax + B)(-2x)}{2\sqrt{\alpha^2 - x^2}} + \frac{K}{\sqrt{\alpha^2 - x^2}} \\ &= \frac{A(\alpha^2 - x^2) - (Ax + B)x + K}{\sqrt{\alpha^2 - x^2}} \\ &= \frac{x^2(-2A) + x(-B) + (A\alpha^2 + K)}{\sqrt{\alpha^2 - x^2}}.\end{aligned}$$

We then equate coefficients on the left-hand side and right-hand side's numerators to get the system of equations

$$\begin{aligned}-2A &= -1 \\ -B &= 0 \\ A\alpha^2 + K &= \alpha^2,\end{aligned}$$

which (without derivation), has the solution

$$A = \frac{1}{2}, \quad B = 0, \quad K = \frac{\alpha^2}{2}.$$

We can then substitute these constant back into Equation 8.8 and complete the remaining integral to get the result

$$\begin{aligned}\int \frac{\alpha^2 - x^2}{\sqrt{\alpha^2 - x^2}}\,\mathrm{d}x &= \frac{x}{2}\sqrt{\alpha^2 - x^2} + \frac{\alpha^2}{2}\int \frac{\mathrm{d}x}{\sqrt{\alpha^2 - x^2}} \\ &= \frac{x}{2}\sqrt{\alpha^2 - x^2} + \frac{\alpha^2}{2}\arcsin\left(\frac{x}{\alpha}\right) + C.\end{aligned}$$

8.4 Change of variables for integrating trigonometric functions

We conclude this chapter with a brief section on integrating trigonometric functions.

Topic 8.6: The universal trigonometric substitution

Let f be any scalar function produced by any combination of the addition, multiplication, and division of constant functions, the sine function, and the cosine function of the scalar variable x. Integrals of such functions can be solved with change of variables using the so-called *universal trigonometric substitution*

$$t = \tan\left(\frac{x}{2}\right), \quad \mathrm{d}x = \frac{2\,\mathrm{d}t}{1 + t^2}.$$

In this case it is straightforward to verify using half-angle formulae and the definition of the tangent function that

$$\sin x = \frac{2t}{1 + t^2} \quad \text{and} \quad \cos x = \frac{1 - t^2}{1 + t^2},$$

and so changing any integral of f over x to the new variable t results in an integral of a rational function of the variable t, which can in principle always be solved using

the methods of Section 8.2. Note that universal trigonometric substitution is also called the *tangent half-angle substitution*.

We illustrate this substitution technique in the following example.

Example 8.16: A simple example of the universal trigonometric substitution

We will use the change of variables $t = \tan(x/2)$ to solve the indefinite integral

$$\int \frac{\mathrm{d}x}{\sin x} = \int \frac{1+t^2}{2t} \cdot \frac{2}{1+t^2}\,\mathrm{d}t = \int \frac{\mathrm{d}t}{t} = \ln|t| + C = \ln\left|\tan\left(\frac{x}{2}\right)\right| + C.$$

While, in principle, the universal trigonometric substitution can be used to solve any trigonometric integral, in certain special cases it is much simpler to use more specialized integration techniques. We conclude this section with a technique for computing products of sine and cosine functions.

Topic 8.7: Integrating products of the sine and cosine functions

Let f be any scalar function produced by any combination of the sums and products of constant functions, sine functions, and cosine functions of the scalar variable x. Because of the linearity of the indefinite integral, the integral of such functions reduces to solving integrals of the form

$$F_{m,n}(x) = \int (\sin^m x) \cdot (\cos^n x)\,\mathrm{d}x, \quad m, n \in \mathbb{N} \cup \{0\}.$$

Such integrals can be computed recursively using induction on the sum $m+n$. As a base case we take the integrals $F_{0,0}$, $F_{1,0}$, and $F_{0,1}$, which are simply the integrals of the constant function, sine function, and cosine function, respectively. For $m+n \geq 2$, we then consider the following cases:

(i) If m is odd, we solve the integral $F_{m,n}$ using the change of variables $t = \cos x$.

(ii) If n is odd, we solve the integral $F_{m,n}$ using the change of variables $s = \sin x$.

(iii) If both m and n are even, we first rearrange the integral $F_{m,n}$ using the power reduction formulae

$$\sin^2 x = \frac{1-\cos(2x)}{2} \quad \text{and} \quad \cos^2 x = \frac{1+\cos(2x)}{2}.$$

We then introduce the new variable $u = 2x$, which reduces the integral to a linear combination of integrals of the form

$$F_{i,j}, \quad i+j \leq \frac{n+m}{2},$$

which we can solve by the induction step using the cases mentioned so far.

We illustrate this technique in the following example.

Example 8.17: Integrating products of sine and cosine functions

(1) We begin by computing the following integral

$$\int \sin^{2k+1} x \, \mathrm{d}x, \quad k \in \mathbb{N} \cup \{0\}.$$

Since the power of the sine function is odd, we can solve the integral with the change of variables $t = \cos x$; the calculation reads

$$\begin{aligned} \int \sin^{2k+1} x \, \mathrm{d}x &= \int \left(\sin^2 x\right)^k \sin x \, \mathrm{d}x = \int (1 - \cos^2 x)^k \sin x \, \mathrm{d}x \\ &\overset{\text{(a)}}{=} -\int (1 - t^2)^k \, \mathrm{d}t, \end{aligned}$$

where in (a) we have reduced the problem to the integral of a polynomial, which we could then solved as in Example 8.6.

We can analogously compute the following integral for all $k \in \mathbb{N} \cup \{0\}$, now using the substitution $s = \sin x$:

$$\begin{aligned} \int \cos^{2k+1} x \, \mathrm{d}x &= \int \left(\cos^2 x\right)^k \cos x \, \mathrm{d}x = \int (1 - \sin^2 x)^k \cos x \, \mathrm{d}x \\ &= \int (1 - s^2)^k \, \mathrm{d}s, \end{aligned}$$

again reducing the problem to the integral of a polynomial function.

(2) We now consider even powers of sine and cosine and will compute the integrals

$$\int \sin^{2k} x \, \mathrm{d}x \quad \text{and} \quad \int \cos^{2k} x \, \mathrm{d}x, \quad k \in \mathbb{N}.$$

In this case we will recursively reduce the power of the sine and cosine using power reduction formulae as in point (iii) of Topic 8.7; the calculation reads

$$\int \sin^{2k} x \, \mathrm{d}x = \int \left[\frac{1 - \cos(2x)}{2}\right]^k \mathrm{d}x = \frac{1}{2^k} \sum_{j=0}^{k} (-1)^j \binom{j}{k} \int \cos^j(2x) \, \mathrm{d}x,$$

$$\int \cos^{2k} x \, \mathrm{d}x = \int \left[\frac{1 + \cos(2x)}{2}\right]^k \mathrm{d}x = \frac{1}{2^k} \sum_{j=0}^{k} \binom{j}{k} \int \cos^j(2x) \, \mathrm{d}x.$$

In both cases we have reduced the original integrals to a linear combination of integrals in which the maximum power of the cosine function is reduced to half of its original value. We could then, in principle, solve the integral of any even-powered sine or cosine function by repeating the process recursively until reaching only integrals of the cosine function to the first power.

As a concrete example, let us consider the case $k = 4$, which we solve as follows:

$$\begin{aligned}
\int \cos^4 x \, \mathrm{d}x &\overset{\text{(a)}}{=} \int \left[\frac{1+\cos(2x)}{2}\right]^2 \mathrm{d}x = \frac{1}{4}\int \left[1 + 2\cos(2x) + \cos^2(2x)\right] \mathrm{d}x \\
&= \frac{1}{4}\int 1 \, \mathrm{d}x + \frac{1}{4}\int 2\cos(2x) \, \mathrm{d}x + \frac{1}{4}\cos^2(2x) \, \mathrm{d}x \\
&\overset{\text{(b)}}{=} \frac{x}{4} + \frac{1}{4}\cdot 2 \cdot \frac{\sin(2x)}{2} + \frac{1}{4}\left[\frac{1+\cos(4x)}{2}\right] \mathrm{d}x \\
&= \frac{x}{4} + \frac{1}{4}\sin(2x) + \frac{1}{8}\int 1 \, \mathrm{d}x + \frac{1}{8}\int \cos(4x) \, \mathrm{d}x \\
&= \frac{x}{4} + \frac{1}{4}\sin(2x) + \frac{x}{8} + \frac{1}{8}\frac{\sin(4x)}{4} + C \\
&= \frac{3x}{8} + \frac{\sin(2x)}{4} + \frac{\sin(4x)}{32} + C,
\end{aligned}$$

where both (a) and (b) use a power reduction formula to halve the power of the cosine function.

(3) We compute the following integral using a change of variables:

$$\begin{aligned}
\int \sin^3 x \cos^3 x \, \mathrm{d}x &= \int (\sin^3 x)(1 - \sin^2 x)\cos x \, \mathrm{d}x \\
&= \int s^3(1 - s^2) \, \mathrm{d}x = \frac{s^4}{4} - \frac{s^6}{6} + C \\
&= \frac{\sin^4 x}{4} - \frac{\sin^6 x}{6} + C.
\end{aligned}$$

(4) Let us conclude by computing the integral

$$\int \sin^2 x \cos^2 x \, \mathrm{d}x.$$

In this case both the power of the sine and cosine functions are even, so we will simplify the integral with power reduction formulae. The calculation reads

$$\begin{aligned}
\int \sin^2 x \cos^2 x \, \mathrm{d}x &= \int \left[\frac{1-\cos(2x)}{2}\right]\left[\frac{1+\cos(2x)}{2}\right] \mathrm{d}x \\
&= \frac{1}{4}\int \left[1 - \cos^2(2x)\right] \mathrm{d}x \\
&= \frac{1}{4}\int \left[1 - \frac{1+\cos(4x)}{2}\right] \mathrm{d}x \\
&= \frac{1}{8}\int 1 \, \mathrm{d}x - \frac{1}{8}\int \cos(4x) \, \mathrm{d}x \\
&= \frac{x}{8} - \frac{\sin(4x)}{32} + C.
\end{aligned}$$

9 The definite integral

The definite integral is closely related to the problem of finding the area under the graph of a scalar function. Indefinite and definite integration are related by the fundamental theorem of calculus, which we will cover later in this chapter.

9.1 The Darboux integral

Definition 9.1: Partition of a closed interval

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$. A *partition* of the closed interval $[a, b]$ is any *finite* subset $\mathcal{D} \subset [a, b]$ for which $a, b \in \mathcal{D}$.

We now introduce some partition-related notation that we will use throughout the chapter.

Topic 9.1: Some partition-related notation

For any $n \in \mathbb{N}$, we will write partitions containing $n + 1$ points in the forms

$$\mathcal{D} = \{x_0, x_1, \ldots, x_{n-1}, x_n\} = \{x_k\}_{k=0}^{n},$$

where we will assume the $n + 1$ points in the partition are ordered by size, i.e.

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

A partition $\mathcal{D} = \{x_k\}_{k=0}^{n}$ divides an interval $[a, b]$ into n subintervals

$$[x_{k-1}, x_k], \quad k = 1, 2, \ldots, n;$$

we will denote the length of the partition's k-th subinterval by

$$\Delta x_k = x_k - x_{k-1}, \quad k = 1, 2, \ldots, n.$$

We will use $\Delta\mathcal{D}$ to denote the maximum subinterval length in the partition, i.e.

$$\Delta\mathcal{D} = \max \Delta x_k = \max \big\{\Delta x_k \,;\, k \in \{1, 2, \ldots, n\}\big\}.$$

We will sometimes use the symbol n without qualification in this chapter when defining partitions—in such cases n implicitly refers to an arbitrary natural number.

Topic 9.2: Suprema and infima on partition subintervals

Let $f\colon [a, b] \to \mathbb{R}$ be an arbitrary bounded function and let $\mathcal{D} = \{x_k\}_{k=0}^{n}$ be an arbitrary partition of the interval $[a, b]$. For all $k = 1, 2, \ldots, n$, we will denote the supremum and infimum of f on the partition's k-th subinterval by

$$\begin{aligned} \mathrm{m}_k &= \mathrm{m}_k(f, \mathcal{D}) = \inf\big(f([x_{k-1}, x_k])\big), \\ \mathrm{M}_k &= \mathrm{M}_k(f, \mathcal{D}) = \sup\big(f([x_{k-1}, x_k])\big). \end{aligned}$$

For all $k = 1, 2, \ldots, n$, these subinterval suprema and infima obey the bounds

$$\inf(f) \leq \mathrm{m}_k(f, \mathcal{D}) \leq \mathrm{M}_k(f, \mathcal{D}) \leq \sup(f).$$

Topic 9.3: Approximating the area under a function's curve

For a given function $f\colon [a,b] \to \mathbb{R}$, a partition $\mathcal{D}$ of the interval $[a,b]$ defines a segmentation of the area between f's graph and the horizontal axis into vertical slices. In this case—assuming the function is bounded—the area of the k-th slice may be approximated by either

(i) the area of the rectangle of width Δx_k bounded by the horizontal lines $[x_{k-1}, x_k] \times \{0\}$ and $[x_{k-1}, x_k] \times \{\mathrm{m}_k\}$, or

(ii) the area of the rectangle of width Δx_k bounded by the horizontal lines $[x_{k-1}, x_k] \times \{0\}$ and $[x_{k-1}, x_k] \times \{\mathrm{M}_k\}$.

By adding up the areas of each partition-induced slice between f's graph and the horizontal axis, we may approximate the entire area between f's graph and the horizontal axis.

Each of the above two approximation methods for a slice's area produces leads to a corresponding approximations for the area between f's graph and the horizontal axis. Loosely, these two approximations are called lower and upper Darboux sums; we define these concepts precisely below.

Definition 9.2: Lower and upper Darboux sum

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let $f\colon [a,b] \to \mathbb{R}$ be an arbitrary bounded function, and let $\mathcal{D} = \{x_k\}_{k=0}^n$ be any partition of the interval $[a,b]$.

In this case the *lower Darboux sum* of the function f *with respect to the partition* $\mathcal{D}$ is the real number

$$\mathrm{s}(f,\mathcal{D}) = \sum_{k=1}^{n} \mathrm{m}_k(f,\mathcal{D})\Delta x_k \in \mathbb{R},$$

while the *upper Darboux sum* of the function f with respect to the partition $\mathcal{D}$ is

$$\mathrm{S}(f,\mathcal{D}) = \sum_{k=1}^{n} \mathrm{M}_k(f,\mathcal{D})\Delta x_k \in \mathbb{R}.$$

The definition of the Darboux sums and the numbers m_k and M_k immediately implies the following useful bound, which we will refer back to later in this section.

Proposition 9.3: A simple bound on lower and upper Darboux sums

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be an arbitrary bounded function. In this case the following inequality holds for any partition $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a,b]$:

$$\inf(f) \cdot (b-a) \leq \mathrm{s}(f,\mathcal{D}) \leq \mathrm{S}(f,\mathcal{D}) \leq \sup(f) \cdot (b-a).$$

Proof. The function f is bounded on $[a,b]$, so it has a supremum and infimum on

both $[a, b]$ and on each subinterval of $[a, b]$. We can then directly compute

$$\begin{aligned}
\mathrm{s}(f, \mathcal{D}) &= \sum_{k=1}^{n} \mathrm{m}_k(f, \mathcal{D}) \Delta x_k = \sum_{k=1}^{n} \inf\big(f([x_{k-1}, x_k])\big) \Delta x_k \overset{\text{(a)}}{\geq} \sum_{k=1}^{n} \inf(f) \Delta x_k \\
&= \inf(f) \sum_{k=1}^{n} \Delta x_k = \inf(f) \cdot (b - a), \\
\mathrm{S}(f, \mathcal{D}) &= \sum_{k=1}^{n} \mathrm{M}_k(f, \mathcal{D}) \Delta x_k = \sum_{k=1}^{n} \sup\big(f([x_{k-1}, x_k])\big) \Delta x_k \overset{\text{(b)}}{\leq} \sum_{k=1}^{n} \sup(f) \Delta x_k \\
&= \sup(f) \sum_{k=1}^{n} \Delta x_k = \sup(f) \cdot (b - a),
\end{aligned}$$

where (a) and (b) follow from the inequalities $\inf(f) \leq \inf\big(f([x_{k-1}, x_k])\big)$ and $\sup(f) \geq \sup\big(f([x_{k-1}, x_k])\big)$ for all $k = 1, 2, \ldots, n$.

The final bound, i.e. $\mathrm{s}(f, \mathcal{D}) \leq \mathrm{S}(f, \mathcal{D})$, follows immediately from the fact that $\mathrm{m}_k \leq \mathrm{M}_k$ for all $k = 1, 2, \ldots, n$. □

Proposition 9.4: Darboux sums on different partitions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let $f\colon [a, b] \to \mathbb{R}$ be an arbitrary bounded function. In this case

(i) Any two partitions $\mathcal{D}$ and $\mathcal{D}'$ of the interval $[a, b]$ obey the inequality

$$\mathrm{s}(f, \mathcal{D}) \leq \mathrm{S}(f, \mathcal{D}').$$

In other words, a function's lower Darboux sum will always be less than or equal to any upper Darboux sum, regardless of the partition used for each sum.

(ii) If $\mathcal{D}$ and $\mathcal{D}'$ are two partitions of the interval $[a, b]$ for which $\mathcal{D} \subset \mathcal{D}'$, then

$$\mathrm{s}(f, \mathcal{D}) \leq \mathrm{s}(f, \mathcal{D}') \leq \mathrm{S}(f, \mathcal{D}') \leq \mathrm{S}(f, \mathcal{D}).$$

Proof.

(ii) If $\mathcal{D}$ and $\mathcal{D}'$ are equal, the bound follows immediately from Proposition 9.3.

Now assume that the partition $\mathcal{D}'$ contains only one element more than $\mathcal{D}$, and let x' denote the single element in the set difference $\mathcal{D}' \setminus \mathcal{D}$. We then denote $\mathcal{D} = \{x_k\}_{k=0}^{n}$, let i be the index in the set $\{1, 2, \ldots, n\}$ for which $x' \in (x_{i-1}, x_i)$, and define the quantities

$$\begin{aligned}
\mathrm{m}_k &= \mathrm{m}_k(f, \mathcal{D}), \quad k = 1, 2, \ldots, n, \\
\mathrm{m}' &= \inf\big(f([x_{i-1}, x'])\big), \\
\mathrm{m}'' &= \inf\big(f([x', x_i])\big).
\end{aligned}$$

The point x' divides the i-th subinterval of the partition $\mathcal{D}$ into two parts, while the remaining subintervals in $\mathcal{D}$ are identical to the corresponding subintervals

in $\mathcal{D}'$. Because $x' \in (x_{i-1}, x_i)$, the interval $[x_{i-1}, x_i]$ is larger than either $[x_{i-1}, x']$ or $[x', x_i]$, and so

$$\mathrm{m}_i \leq \mathrm{m}' \quad \text{and} \quad \mathrm{m}_i \leq \mathrm{m}''.$$

These two inequalities in turn imply that

$$\begin{aligned} \mathrm{s}(f, \mathcal{D}') - \mathrm{s}(f, \mathcal{D}) &= \mathrm{m}' \cdot (x' - x_{i-1}) + \mathrm{m}'' \cdot (x_i - x') - \mathrm{m}_i \cdot (x_i - x_{i-1}) \\ &\geq \mathrm{m}_i \cdot (x' - x_{i-1}) + \mathrm{m}_i \cdot (x_i - x') - \mathrm{m}_i \cdot (x_i - x_{i-1}) \\ &= \mathrm{m}_i \cdot (x' - x_{i-1} + x_i - x' - x_i + x_{i-1}) \\ &= 0, \end{aligned}$$

which we then rearrange to get $\mathrm{s}(f, \mathcal{D}) \leq \mathrm{s}(f, \mathcal{D}')$. The proof of the inequality $\mathrm{S}(f, \mathcal{D}') \leq \mathrm{S}(f, \mathcal{D})$ when $\mathcal{D}$ and $\mathcal{D}'$ differ by only one element is analogous.

We now consider the case when the set difference $\mathcal{D}' \setminus \mathcal{D}$ has more than one element. In this case, for some $p \in \mathbb{N}$, there exists a chain of partitions

$$\mathcal{D} = \mathcal{D}_0 \subset \mathcal{D}_1 \subset \cdots \subset \mathcal{D}_p = \mathcal{D}'$$

of the interval $[a, b]$ for which each partition $\mathcal{D}_j$ will have exactly one element more than $\mathcal{D}_{j-1}$ for all $j = 1, 2, \ldots, p$. We have already shown that the proposition holds for every neighboring partition pair $(\mathcal{D}_{j-1}, \mathcal{D}_j)$ (which differ by one element), and so the proposition must also hold for the pair $(\mathcal{D}, \mathcal{D}')$ by the transitive nature of the subset and inequality relations.

(i) We first define the union $\mathcal{D}'' = \mathcal{D} \cup \mathcal{D}'$. By construction, $\mathcal{D}''$ is a partition of the interval $[a, b]$ for which $\mathcal{D} \subset \mathcal{D}''$ and $\mathcal{D}' \subset \mathcal{D}''$, and so we can apply point (ii) to directly show that

$$\mathrm{s}(f, \mathcal{D}) \overset{\text{(a)}}{\leq} \mathrm{s}(f, \mathcal{D}'') \overset{\text{(b)}}{\leq} \mathrm{S}(f, \mathcal{D}'') \overset{\text{(c)}}{\leq} \mathrm{S}(f, \mathcal{D}'),$$

where (a) and (c) follow from point (ii) and (b) from Proposition 9.3. □

Prop 9.4 (ii) states that for any bounded function f and any partition $\mathcal{D}'$ of f's domain containing more points than some other partition $\mathcal{D}$,

- any lower sum using $\mathcal{D}'$ will be larger than any lower sum using $\mathcal{D}$, and
- any upper sum using $\mathcal{D}'$ will be smaller than any upper sum using $\mathcal{D}$.

Intuitively, this observation suggests that as more and more points are added to a partition, the corresponding lower and upper sums grow closer and closer together. It is reasonable to ask if a function's lower and upper sums might "converge" to a fixed value as the partition on which they are computed contains more and more points. In certain cases this intuition turns out to be (loosely) correct—we will formalize the idea in the following two definitions. Before doing so, we first note the following:

- Proposition 9.4 (i) implies that every upper Darboux sum of a bounded function f is an upper bound of the set of f's lower Darboux sums. This means that the set of f's lower Darboux sums is bounded, and thus has a supremum.

- Similarly, Proposition 9.4 (i) implies that every lower Darboux sum of the function f is a lower bound of the set of f's upper Darboux sums. This means that the set of f's upper Darboux sums is bounded, and thus has a infimum.

These two observations provide the context for the following definition.

Definition 9.5: Upper and lower Darboux integral

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be a bounded function. In this case:

(i) The function f's *lower Darboux integral* is defined as the supremum of the set of f's lower Darboux sums and is denoted by

$$\mathrm{s}(f) = \underline{\int_a^b} f(x)\,\mathrm{d}x = \sup\big\{\mathrm{s}(f,\mathcal{D})\,;\,\mathcal{D} \text{ is a partition of } [a,b]\big\}.$$

(ii) The function f's *upper Darboux integral* is defined as the infimum of the set of f's upper Darboux sums and is denoted by

$$\mathrm{S}(f) = \overline{\int_a^b} f(x)\,\mathrm{d}x = \inf\big\{\mathrm{S}(f,\mathcal{D})\,;\,\mathcal{D} \text{ is a partition of } [a,b]\big\}.$$

We pause momentarily to note the useful inequality

$$\mathrm{s}(f,\mathcal{D}) \leq \mathrm{s}(f) \leq \mathrm{S}(f) \leq \mathrm{S}(f,\mathcal{D}), \tag{9.1}$$

which holds for any bounded function $f\colon [a,b] \to \mathbb{R}$, where $a < b$ and $\mathcal{D}$ is any partition of the interval $[a,b]$. Equation 9.1 follows directly from Proposition 9.3 and the definition of the upper and lower Darboux integrals, and we will refer back to it when proving a few propositions later in the chapter.

Definition 9.6: Darboux integrability

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be a scalar function. The function f is said to be *Darboux integrable* if it is bounded and its upper and lower Darboux integrals are equal, i.e. $\mathrm{s}(f) = \mathrm{S}(f)$. In this case we write

$$\int_a^b f(x)\,\mathrm{d}x = \mathrm{s}(f) = \mathrm{S}(f) \in \mathbb{R},$$

and call the resulting real number the *Darboux integral* of f.

Remark 9.7: Technicality: integrals over intervals of zero length

Definition 9.6 assumes that the scalar function f is defined on a closed, bounded interval of finite length. But it is straightforward to generalize Definition 9.6 to functions defined at a single point. Namely, for any $a \in \mathbb{R}$ we define functions of the form $f\colon [a,a] \to \mathbb{R}$ to be Darboux integrable with Darboux integral equal to zero, i.e.

$$\int_a^a f(x)\,\mathrm{d}x = 0.$$

Remark 9.8: Some terminology related to Darboux integration

(1) Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be any Darboux integrable function. The function f's Darboux integral

$$\int_a^b f(x)\,\mathrm{d}x$$

is also called the *definite integral of f from a to b* or the *definite integral of f on the interval $[a,b]$*; the numbers a and b are called the *limits* or *bounds* of integration.

We may also define the definite integral of f from b to a, which is equal to

$$\int_b^a f(x)\,\mathrm{d}x = -\int_a^b f(x)\,\mathrm{d}x.$$

In other words, reversing limits of integration changes the sign of the integral.

(2) We now introduce a shorthand notation for integration over a closed interval contained within an open subset. Let $f\colon U \to \mathbb{R}$ be a scalar function defined on an arbitrary open subset $U \subset \mathbb{R}$, let $a, b \in \mathbb{R}$ be two numbers for which $a < b$, and let $[a,b]$ be a nonempty, closed, bounded interval for which $[a,b] \subset U$.

In this case, if the restricted function $f|_{[a,b]}$ is Darboux integrable, we define the shorthand notation

$$\int_a^b \big(f|_{[a,b]}\big)(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x.$$

Going forward, we will refer to Darboux integrable functions as simply *integrable* functions unless explicitly noted otherwise. We denote the set of all integrable scalar functions defined on the nonempty, closed, bounded interval $[a,b]$ by

$$\mathcal{R}\big([a,b]\big).$$

As mentioned in this chapter's introduction, the definite integral of an integrable function is directly related to the problem of finding the area between the function's graph and the horizontal axis. Having defined Darboux integrability, we are now ready to formalize this notation of area.

Remark 9.9: Area between a function's graph and the horizontal axis

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be any integrable function. We first consider two special cases:

(i) If $f(x) \geq 0$ for all $x \in [a,b]$, then we define the area of the planar region bounded by the graph of f and the horizontal line $[a,b] \times \{0\}$ to

$$\int_a^b f(x)\,\mathrm{d}x.$$

(ii) If $f(x) \leq 0$ for all $x \in [a,b]$, then the definite integral of f over $[a,b]$ is also less than or equal to zero. Since negative area doesn't make sense, in this case

we define the area between the graph of f and the horizontal line $[a,b] \times \{0\}$ to be equal to

$$-\int_a^b f(x)\,\mathrm{d}x.$$

More generally, when $f(x)$ is neither nonnegative nor nonpositive for all $x \in [a,b]$, we must consider the sign of f when defining the area between f's graph and the horizontal axis. In such cases we split the interval $[a,b]$ onto subintervals on which f is either nonnegative or nonpositive, apply the above two cases to each subinterval, and sum the resulting areas to get the area between f and the horizontal axis.

Example 9.10: Darboux integrability

(1) Any constant function $f\colon \mathbb{R} \to \mathbb{R}$ with value $\alpha \in \mathbb{R}$ is integrable on every nonempty, closed, and bounded interval of the real line, and for any two $a, b \in \mathbb{R}$ has an integral equal to

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b \alpha\,\mathrm{d}x = \alpha(b-a).$$

To verify this, we first note that the equality holds trivially when $a = b$ by Remark 9.7, simply reducing to the equality $0 = 0$.

Let us then assume that $a < b$ and let $\mathcal{D} = \{x_k\}_{k=0}^n$ be an arbitrary partition of the interval $[a,b]$. Because f is constant,

$$\mathrm{m}_k(f,D) = \mathrm{M}_k(f,D) = \alpha \text{ for all } k = 1,2,\ldots,n.$$

This implies that f's upper and lower Darboux sums are both equal to

$$\mathrm{s}(f,\mathcal{D}) = \mathrm{S}(f,\mathcal{D}) = \alpha(b-a),$$

which in turn implies that

$$\mathrm{s}(f) = \mathrm{S}(f) = \alpha(b-a).$$

Finally, if $a > b$, the statement follows from the calculation

$$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x = -\alpha(a-b) = \alpha(b-a).$$

As a special case when $\alpha = 1$, this example implies that

$$\int_a^b 1\,\mathrm{d}x = \int_a^b \mathrm{d}x = b-a \text{ for all } a,b \in \mathbb{R}.$$

(2) Consider the function $f\colon [0,1] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 1 & x \in [0,1] \cap \mathbb{Q}, \\ 0 & x \in [0,1] \setminus \mathbb{Q}. \end{cases}$$

This function's value is equal to one at rational numbers in the interval $[0,1]$ and equal to zero at real numbers the real interval $[0,1]$.

For an arbitrary partition $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[0,1]$, it holds that

$$\mathrm{m}_k(f,\mathcal{D}) = 0 \quad \text{and} \quad \mathrm{M}_k(f,\mathcal{D}) = 1 \text{ for all } k = 1,2,\ldots,n,$$

which in turn implies that

$$\mathrm{s}(f,\mathcal{D}) = 0 \quad \text{and} \quad \mathrm{S}(f,\mathcal{D}) = 1.$$

Thus $\mathrm{s}(f) = 0$ while $\mathrm{S}(f) = 1$, and so the function f is not Darboux integrable.

In passing, although it falls beyond the scope of this course, we note that f is *Lebesgue integrable*. Loosely, the Lebesgue integral of f can be interpreted as the probability that a real number chosen at random from the interval $[0,1]$ is rational. This probability, and thus the Lebesgue integral of f, is equal to zero, since there are only countably many rational numbers (but uncountably many real numbers) in the interval $[0,1]$.

Proposition 9.11: Criterion for integrability

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$. A scalar function $f\colon [a,b] \to \mathbb{R}$ is integrable if, and only if, it is bounded and for all $\epsilon \in \mathbb{R}^+$ there exists a partition $\mathcal{D}$ of the interval $[a,b]$ for which

$$\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon.$$

Proof. ($\Leftarrow$) Assume that f bounded and satisfies the condition $\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon$. We first recall from Equation 9.1 that the function f's upper and lower Darboux integrals and sums are ordered as follows:

$$\mathrm{s}(f,\mathcal{D}) \leq \mathrm{s}(f) \leq \mathrm{S}(f) \leq \mathrm{S}(f,\mathcal{D}).$$

This inequality, combined with the assumption that $\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon$ implies

$$\mathrm{S}(f) - \mathrm{s}(f) < \epsilon.$$

And because $\epsilon \in \mathbb{R}^+$ is arbitrary, the inequality $\mathrm{S}(f) - \mathrm{s}(f) < \epsilon$ implies that $\mathrm{s}(f) = \mathrm{S}(f)$, i.e. f is integrable.

($\Rightarrow$) Assume f is integrable and choose arbitrary $\epsilon \in \mathbb{R}^+$. Because f's lower integral $\mathrm{s}(f)$ is the supremum of f's lower Darboux sums, there exists a partition $\mathcal{D}'$ of the interval $[a,b]$ for which

$$\mathrm{s}(f) - \mathrm{s}(f,\mathcal{D}') < \frac{\epsilon}{2}.$$

Similarly, because f's upper integral $\mathrm{S}(f)$ is the infimum of f's upper Darboux sums, there exists a partition $\mathcal{D}''$ of the interval $[a,b]$ for which

$$\mathrm{S}(f,\mathcal{D}') - \mathrm{S}(f) < \frac{\epsilon}{2}.$$

Now let $\mathcal{D} = \mathcal{D}' \cup \mathcal{D}''$. Because $\mathcal{D}' \subset \mathcal{D}$ and $\mathcal{D}'' \subset \mathcal{D}$, Proposition 9.4 (ii) then implies

$$\begin{aligned} \mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) &\leq \mathrm{S}(f,\mathcal{D}'') - \mathrm{s}(f,\mathcal{D}') \\ &\overset{\text{(a)}}{=} \mathrm{S}(f,\mathcal{D}'') - \mathrm{S}(f) + \mathrm{s}(f) - \mathrm{s}(f,\mathcal{D}') \\ &\overset{\text{(b)}}{=} \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

where (a) holds because f is integrable and so $\mathrm{s}(f) = \mathrm{S}(f)$, and (b) holds by the bounds established earlier in the proof. □

Proposition 9.12: Continuity implies integrability

Every continuous scalar function $f\colon [a,b] \to \mathbb{R}$ is integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

Proof. If $a = b$, the proposition follows trivially from Remark 9.7, so we will assume, without loss of generality, that $a < b$. We will prove the proposition by showing that f meets the conditions of Proposition 9.11.

First choose arbitrary $\epsilon \in \mathbb{R}^+$. Because f is continuous and defined on a closed, bounded interval, f is both bounded and uniformly continuous by Theorem 6.36 and Proposition 6.46. There thus exists $\delta \in \mathbb{R}^+$ for which

$$|f(y) - f(x)| < \frac{\epsilon}{b-a} \text{ for all } x, y \in [a,b] \text{ for which } |y - x| < \delta.$$

Now let $D = \{x_k\}_{k=0}^n$ be a partition of the interval $[a,b]$ for which $\max \Delta x_k < \delta$. The function f is continuous, so it assumes a maximum and minimum value (i.e. attains is supremum and infimum) on every subinterval of the partition $\mathcal{D}$ by Theorem 6.37. Thus, for every natural number $k \leq n$, there exist points $u_k, v_k \in [x_{k-1}, x_k]$ for which

$$f(u_k) = \mathrm{M}_k(f,\mathcal{D}) \quad \text{and} \quad f(v_k) = \mathrm{m}_k(f,\mathcal{D}).$$

We then combine the inequalities $|u_k - v_k| < \max \Delta x_k$ and $\max \Delta x_k < \delta$ to get $|u_k - v_k| < \delta$, which in turn implies that

$$\mathrm{M}_k(f,\mathcal{D}) - \mathrm{m}_k(f,\mathcal{D}) = f(u_k) - f(v_k) \overset{\text{(a)}}{<} \frac{\epsilon}{b-a} \text{ for all } k \leq n,$$

where (a) follows from f's uniform continuity. We then complete the proof with the calculation

$$\begin{aligned} \mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) &= \sum_{k=1}^n \mathrm{M}_k(f,\mathcal{D})\Delta x_k - \sum_{k=1}^n \mathrm{m}_k(f,\mathcal{D})\Delta x_k \\ &= \sum_{k=1}^n \big[\mathrm{M}_k(f,\mathcal{D}) - \mathrm{m}_k(f,\mathcal{D})\big]\Delta x_k \\ &< \sum_{k=1}^n \frac{\epsilon}{b-a}\Delta x_k = \frac{\epsilon}{b-a}\sum_{k=1}^n \Delta x_k \\ &= \frac{\epsilon}{b-a} \cdot (b-a) = \epsilon. \end{aligned}$$

We have shown that f is bounded and $\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon$ for arbitrary $\epsilon \in \mathbb{R}^+$, so f is integrable by Proposition 9.11. □

Definition 9.13: Piecewise continuity

A scalar function defined on a closed, bounded interval is said to be *piecewise-continuous* if it is continuous at all points in its domain, with the possible exception of finitely many points.

Having defined piecewise continuity, we can now give a generalization of Proposition 9.12.

Proposition 9.14: Piecewise continuity implies integrability

Every bounded, piecewise-continuous scalar function $f\colon [a,b] \to \mathbb{R}$ is integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

Proof. If f is continuous in addition to being piecewise-continuous, the proposition follows immediately from Proposition 9.12, so we will assume that f is piecewise-continuous, but not continuous.

Because f is piecewise-continuous, it is continuous at all points in the interval $[a,b]$ except for at finitely many points, which we will denote by

$$z_1, z_2, \ldots, z_p \in [a,b], \quad p \in \mathbb{N}.$$

We also note that because f is piecewise-continuous but not continuous, it cannot be constant, and so $\inf(f) < \sup(f)$. We then choose arbitrary $\epsilon \in \mathbb{R}^+$ and define

$$\alpha = \frac{\epsilon}{8p\big[\sup(f) - \inf(f)\big]}.$$

Since f is piecewise-continuous, it is continuous on the closed, bounded subset

$$A = [a,b] \setminus \bigcup_{i=1}^{p} (z_i - \alpha, z_i + \alpha) \subset \mathbb{R}.$$

More so, since A is closed and bounded, f is uniformly continuous on A by Proposition 6.46, and so there exists $\delta \in \mathbb{R}^+$ for which $\delta < \alpha$ and

$$|f(y) - f(x)| < \frac{\epsilon}{2(b-a)} \text{ for all } x, y \in A \text{ for which } |y - x| < \delta.$$

Next, let $\mathcal{D} = \{x_k\}_{k=0}^n$ be a partition of the interval $[a,b]$ for which $\max \Delta x_k < \delta$, and define

$$J = \{k \in \mathbb{N};\ k \leq n,\ [x_{k-1}, x_k] \subset A\},$$
$$J' = \{k \in \mathbb{N};\ k \leq n,\ [x_{k-1}, x_k] \not\subset A\}.$$

On the partition subintervals for which $k \in J$, we will use a similar procedure to that used in the proof of Proposition 9.12. Namely, for all $k \in J$ the function f is continuous on the interval $[x_{k-1}, x_k]$, so there exist points $u_k, v_k \in [x_{k-1}, x_k]$ for which

$$f(u_k) = \mathrm{M}_k(f, \mathcal{D}) \quad \text{and} \quad f(v_k) = \mathrm{m}_k(f, \mathcal{D}).$$

Because $\Delta x_k < \delta$, it follows that $|u_k - v_k| < \delta$ and so

$$\mathrm{M}_k(f, \mathcal{D}) - \mathrm{m}_k(f, \mathcal{D}) = f(u_k) - f(v_k) \overset{\text{(a)}}{<} \frac{\epsilon}{2(b-a)} \text{ for all } k \in J, \tag{9.2}$$

where (a) follows from f's uniform continuity on A; we will use this bound shortly.

Next, for the partition subintervals for which $k \in J'$, we make the bound

$$\sum_{k \in J'} \Delta x_k \overset{\text{(a)}}{<} p \cdot (2\alpha + 2\delta) \overset{\text{(b)}}{<} p \cdot 4\alpha, \tag{9.3}$$

where (a) holds because $\max \Delta x_k < \delta$ and (b) holds under the assumption that $\delta < \alpha$. The function f's integrability then follows from the calculation

$$\begin{aligned}
\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) &= \sum_{k=1}^{n} \big[\mathrm{M}_k(f, \mathcal{D}) - \mathrm{m}_k(f, \mathcal{D})\big] \Delta x_k \\
&= \sum_{k \in J} \big[\mathrm{M}_k(f, \mathcal{D}) - \mathrm{m}_k(f, \mathcal{D})\big] \Delta x_k \\
&\qquad + \sum_{k \in J'} \big[\mathrm{M}_k(f, \mathcal{D}) - \mathrm{m}_k(f, \mathcal{D})\big] \Delta x_k \\
&\overset{\text{(a)}}{<} \sum_{k \in J} \frac{\epsilon}{2(b-a)} \Delta x_k + \sum_{k \in J'} \big[\sup(f) - \inf(f)\big] \Delta x_k \\
&\overset{\text{(b)}}{<} \frac{\epsilon}{2(b-a)} (b-a) + \big[\sup(f) - \inf(f)\big] \cdot p \cdot 4\alpha = \epsilon,
\end{aligned}$$

where (a) uses Equation 9.2 and (b) uses Equation 9.3. The result $\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \epsilon$ means that f is integrable by Proposition 9.11. □

Proposition 9.15: Monotonic functions are integrable

Every monotonic scalar function $f\colon [a, b] \to \mathbb{R}$ is integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

Proof. Like in the proof of Proposition 9.14, we will assume without loss of generality that $a < b$. Assume f is increasing and—since we have already shown in Example 9.10 that constant functions are integrable—that f is not constant. We then choose arbitrary $\epsilon \in \mathbb{R}^+$, define

$$\delta = \frac{\epsilon}{f(b) - f(a)},$$

and let $\mathcal{D} = \{x_k\}_{k=0}^{n}$ be a partition of the interval $[a, b]$ for which $\max \Delta x_k < \delta$. Finally, we note that, because f is increasing,

$$\mathrm{m}_k(f, \mathcal{D}) = f(x_{k-1}) \quad \text{and} \quad \mathrm{M}_k(f, \mathcal{D}) = f(x_k) \text{ for all } k \leq n.$$

The proposition then follows from the calculation

$$\begin{aligned}
\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) &= \sum_{k=1}^{n} \big[\mathrm{M}_k(f,\mathcal{D}) - \mathrm{m}_k(f,\mathcal{D})\big]\Delta x_k \\
&= \sum_{k=1}^{n} \big[f(x_k) - f(x_{k-1})\big]\Delta x_k \\
&< \sum_{k=1}^{n} \big[f(x_k) - f(x_{k-1})\big]\delta \\
&= \Big(\big[f(x_1) - f(x_0)\big] + \cdots + \big[f(x_n) - f(x_{n-1})\big]\Big)\delta \\
&= \big[f(x_n) - f(x_0)\big]\delta \\
&= \big[f(b) - f(a)\big] \cdot \frac{\epsilon}{f(b) - f(a)} = \epsilon.
\end{aligned}$$

The result $\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon$ means that f is integrable by Proposition 9.11. The proof is analogous if f is decreasing instead of increasing. □

9.2 The Riemann integral

Definition 9.16: Riemann sum

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let $f\colon [a,b] \to \mathbb{R}$ be an arbitrary scalar function. For an arbitrary partition $\mathcal{D} = \{x_k\}_{k=0}^{n}$ of the interval $[a,b]$ and for any set of points

$$\xi_k \in [x_{k-1}, x_k], \quad k = 1, 2, \ldots, n,$$

the *Riemann sum* of the function f (with respect to the partition $\mathcal{D}$ and points $\{\xi_k\}_{k=0}^{n}$) is defined as

$$\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=1}^{n}) = \sum_{k=1}^{n} f(\xi_k)\Delta x_k.$$

The points $\{\xi_k\}_{k=1}^{n}$ are sometimes called *tagging points* and the partition $\mathcal{D}$ together with the points $\{\xi_k\}$ is called a *tagged partition*.

For a given partition, a function's Riemann sum will generally fall between the function's lower and upper Darboux sums on the same partition. More formally, if the function $f\colon [a,b] \to \mathbb{R}$ is bounded, then for any partition $\mathcal{D} = \{x_k\}_{k=0}^{n}$ of the interval $[a,b]$ and for any set of tagging points $\xi_k \in [x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$, the function f's Riemann sum and upper and lower Darboux sums obey the inequality

$$\mathrm{s}(f,\mathcal{D}) \leq \mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=0}^{n}) \leq \mathrm{S}(f,\mathcal{D}). \tag{9.4}$$

This inequality follows from $\mathrm{s}(f,\mathcal{D})$ and $\mathrm{S}(f,\mathcal{D})$ being defined in terms of f's infimum and supremum on each partition subinterval, respectively, while $\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=0}^{n})$ is defined in terms of f's value at an, in general, arbitrary point in each subinterval.

Definition 9.17: Riemann integrability and Riemann integral

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$. A scalar function $f\colon [a,b] \to \mathbb{R}$ is said to be *Riemann integrable* if there exists a real number $I \in \mathbb{R}$ for which for all $\epsilon \in \mathbb{R}^+$ there exists corresponding $\delta \in \mathbb{R}^+$ for which

$$|\mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n) - I| < \epsilon$$

for all partitions $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a,b]$ for which $\max \Delta x_k < \delta$ and for any choice of tagging points $\xi_k \in [x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$.

In this case we define

$$I = \lim_{\max \Delta x_k \to 0} \mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n),$$

and call the number I the *Riemann integral* of the function f.

It turns out that many conditions for Riemann integrability are completely analogous to the conditions for Darboux integrability already developed in Section 9.1. We list some of these conditions below.

Corollary 9.18: Conditions for Riemann integrability

(i) Every continuous scalar function $f\colon [a,b] \to \mathbb{R}$ is Riemann integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

(ii) Every bounded, piecewise-continuous scalar function $f\colon [a,b] \to \mathbb{R}$ is Riemann integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

(iii) Every monotonic scalar function $f\colon [a,b] \to \mathbb{R}$ is Riemann integrable for any two real numbers $a, b \in \mathbb{R}$ for which $a \leq b$.

Proof.

(i) First choose arbitrary $\epsilon \in \mathbb{R}^+$. As in the proof of Proposition 9.12, because f is continuous and defined on a closed, bounded interval, f is both bounded and uniformly continuous by Theorem 6.36 and Proposition 6.46. There thus exists $\delta \in \mathbb{R}^+$ for which

$$|f(y) - f(x)| < \frac{\epsilon}{b-a} \text{ for all } x, y \in [a,b] \text{ for which } |y - x| < \delta.$$

Let $D = \{x_k\}_{k=0}^n$ be a partition of the interval $[a,b]$ for which $\max \Delta x_k < \delta$. In this case, as we have already shown in Proposition 9.12,

$$\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \epsilon.$$

Since f is integrable, its upper and lower Darboux integrals are equal, and we define

$$I_{\mathrm{D}} = \mathrm{s}(f) = \mathrm{S}(f).$$

From Equation 9.1 and Equation 9.4 we then recall the inequalities

$$\mathrm{s}(f, \mathcal{D}) \leq \mathrm{s}(f) \leq \mathrm{S}(f) \leq \mathrm{S}(f, \mathcal{D}),$$
$$\mathrm{s}(f, \mathcal{D}) \leq \mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=0}^n) \leq \mathrm{S}(f, \mathcal{D}),$$

which hold for any partition $\mathcal{D}$ of the interval $[a,b]$ and any set of tagging points $\{\xi_k\}_{k=0}^n$. Together with the earlier result $\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon$, these inequalities imply that

$$|\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=0}^n) - I_{\mathrm{D}}| < \epsilon$$

for any partition $D = \{x_k\}_{k=0}^n$ for which $\max \Delta x_k < \delta$ and any set of tagging points $\{\xi_k\}_{k=0}^n$. Thus f is Riemann integrable on $[a,b]$.

(ii) Analogous to the proof of (i), but this time using the proof of Proposition 9.14.

(iii) Analogous to the proof of (i), but this time using the proof of Proposition 9.15. □

In fact, as we show in the following proposition, Riemann integrability is in general equivalent to Darboux integrability (this is why we have used the term "integrability" without further qualification in this chapter).

Proposition 9.19: Riemann and Darboux integrability are equivalent

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$. A function $f\colon [a,b] \to \mathbb{R}$ is Riemann integrable if, and only if, it is Darboux integrable, and in this case the Riemann and Darboux integrals of f are equal.

Proof. ($\Rightarrow$) We will first show that Riemann integrability implies Darboux integrability. Assume f is Riemann integrable and let I denote its Riemann integral. The proof is a bit lengthy, and we will complete it in two stages.

(i) We will first show, by contradiction, that f is bounded. Assume that f is unbounded above. Since f is Riemann integrable, there exists $\delta_1 \in \mathbb{R}^+$ such that, for all partitions $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a,b]$ for which $\max \Delta x_k < \delta_1$, and for any choice of tagging points $\xi_k \in [x_{k-1}, x_k], k = 1,2,\ldots,n$, it holds that

$$|\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=1}^n) - I| < 1. \tag{9.5}$$

Next, we choose a large-enough natural number $q \in \mathbb{N}$ such that

$$\frac{b-a}{q} < \delta_1,$$

define the points $\{v_j\}$ according to

$$v_j = a + j \cdot \frac{(b-a)}{q}, \quad j = 0,1,\ldots,q,$$

and define the partition $\mathcal{D} = \{v_j\}_{j=0}^q$. Note that, by construction,

$$\Delta v_j = \frac{b-a}{q} \text{ for all } j = 1,2,\ldots,q, \text{ implying } \max \Delta v_j < \delta_1. \tag{9.6}$$

Since we have assumed that f is unbounded above, there exists a point $v \in [a,b]$ for which

$$f(v) \geq f(v_j) + \frac{2q}{b-a} \text{ for all } j = 1,2,\ldots,q. \tag{9.7}$$

We then choose a natural number $l \leq q$ for which $v \in [v_{l-1}, v_l]$, and define $\zeta_l = v$ and $\zeta_j = v_j$ for all $j \leq q$ for which $j \neq l$. We can then make the bound

$$\begin{aligned} \mathrm{R}(f, \mathcal{D}, \{\zeta_j\}_{j=1}^q) - \mathrm{R}(f, \mathcal{D}, \{v_j\}_{j=1}^q) &= \big[f(v) - f(v_l)\big]\Delta v_l \\ &\overset{\text{(a)}}{\geq} \frac{2q}{b-a} \cdot \frac{b-a}{q} \\ &= 2, \end{aligned}$$

where (a) follows from Equations 9.6 and 9.7. On the other hand, because f is Riemann integrable, we may also make the bound

$$\begin{aligned} &\Big|\mathrm{R}(f, \mathcal{D}, \{\zeta_j\}_{j=1}^q) - \mathrm{R}(f, \mathcal{D}, \{v_j\}_{j=1}^q)\Big| \\ &\quad \leq \Big|\mathrm{R}(f, \mathcal{D}, \{\zeta_j\}_{j=1}^q) - I\Big| + \Big|I - \mathrm{R}(f, \mathcal{D}, \{v_j\}_{j=1}^q)\Big| \\ &\quad < 1 + 1 = 2, \end{aligned}$$

where (a) follows from Equation 9.5. The two bounds contradict each other, which means that f is not unbounded above. We can prove analogously that f is not unbounded below, and so f is bounded.

(ii) We will now show, using Proposition 9.11, that a Riemann integrable (and thus, by point (i), necessarily bounded) function f is Darboux integrable. Fix arbitrary $\epsilon \in \mathbb{R}^+$. Because f is Riemann integrable, there exists $\delta \in \mathbb{R}^+$ such that, for all partitions $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a, b]$ for which $\max \Delta x_k < \delta$, and for any choice of tagging points $\xi_k \in [x_{k-1}, x_k], k = 1, 2, \ldots, n$, it holds that

$$|\mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n) - I| < \frac{\epsilon}{4}.$$

Let $\mathcal{D} = \{x_k\}_{k=0}^n$ be any partition of the interval $[a, b]$ for which $\max \Delta x_k < \delta$. Because the function f is bounded, for all $k = 1, 2, \ldots, n$, there exists a point $\xi_k \in [x_{k-1}, x_k]$ for which

$$\mathrm{M}_k(f, \mathcal{D}) - f(\xi_k) < \frac{\epsilon}{4(b-a)}; \tag{9.8}$$

similarly, there exists a point $\eta_k \in [x_{k-1}, x_k]$ for which

$$f(\eta_k) - \mathrm{m}_k(f, \mathcal{D}) < \frac{\epsilon}{4(b-a)}. \tag{9.9}$$

We then continue by computing the bound

$$\begin{aligned} 0 &\overset{\text{(a)}}{\leq} \mathrm{S}(f, \mathcal{D}) - \mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n) \\ &= \sum_{k=1}^n \big[\mathrm{M}_k(f, \mathcal{D}) - f(\xi_k)\big]\Delta x_k \\ &\overset{\text{(b)}}{<} \sum_{k=1}^n \frac{\epsilon}{4(b-a)}\Delta x_k = \frac{\epsilon}{4}, \end{aligned}$$

where (a) holds by Equation 9.4 and (b) by Equation 9.8. Analogously,

$$\begin{aligned} 0 &\leq \mathrm{R}(f, \mathcal{D}, \{\eta_k\}_{k=1}^n) - \mathrm{s}(f, \mathcal{D}) \\ &= \sum_{k=1}^{n} \big[f(\eta_k) - \mathrm{m}_k(f, \mathcal{D})\big] \Delta x_k \\ &\overset{(a)}{<} \sum_{k=1}^{n} \frac{\epsilon}{4(b-a)} \Delta x_k = \frac{\epsilon}{4}, \end{aligned}$$

where (a) holds by Equation 9.9

We then combine the results so far to get

$$\begin{aligned} \mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) \leq\ & |\mathrm{S}(f, \mathcal{D}) - \mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n)| \\ & + |\mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n) - I| \\ & + |I - \mathrm{R}(f, \mathcal{D}, \{\eta_k\}_{k=1}^n)| \\ & + |\mathrm{R}(f, \mathcal{D}, \{\eta_k\}_{k=1}^n) - \mathrm{s}(f, \mathcal{D})| \\ <\ & \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon. \end{aligned}$$

Since $\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \epsilon$, the function f is Darboux integrable by Proposition 9.11.

($\Leftarrow$) We will now show that Darboux integrability implies Riemann integrability. Assume f is Darboux integrable and let I denote its Darboux integral. Since f is Darboux integrable it is also bounded, so there exists constant $K \in \mathbb{R}^+$ for which

$$\sup(|f|) \leq K.$$

Next, choose arbitrary $\epsilon \in \mathbb{R}^+$. Because f is Darboux integrable, by Proposition 9.11 there exists a partition $\mathcal{D}_0$ of the interval $[a, b]$ for which

$$\mathrm{S}(f, \mathcal{D}_0) - \mathrm{s}(f, \mathcal{D}_0) < \frac{\epsilon}{2}.$$

Let $N \in \mathbb{N}$ denote the number of points in the partition $\mathcal{D}_0$, and define

$$\delta = \frac{\epsilon}{8NK}.$$

Next, let $\mathcal{D} = \{x_k\}_{k=0}^n$ be any partition of the interval $[a, b]$ for which $\max \Delta x_k < \delta$, and define an additional partition $\mathcal{D}' = \mathcal{D} \cup \mathcal{D}_0$, which we will write in the form $\mathcal{D}' = \{u_i\}_{i=0}^p$ for some $p \in \mathbb{N}$. Because $\mathcal{D}_0 \subset \mathcal{D}'$, Proposition 9.4 (ii) then implies that

$$\mathrm{S}(f, \mathcal{D}_0) - \mathrm{s}(f, \mathcal{D}_0) < \frac{\epsilon}{2} \implies \mathrm{S}(f, \mathcal{D}') - \mathrm{s}(f, \mathcal{D}') < \frac{\epsilon}{2}.$$

Next, for any natural number $k \leq n$, we define the set J_k as

$$J_k = \big\{i \in \mathbb{N};\ i \leq p \text{ and } [u_{i-1}, u_i] \subset [x_{k-1}, x_k]\big\}.$$

Note that for all $k \leq n$ the set J_k must have at least one element, since $\mathcal{D} \subset \mathcal{D}'$. If J_k has only one element for a given k, then the open interval (x_{k-1}, x_k) does not

contain any elements in $\mathcal{D}_0$, in which case the closed subinterval $[x_{k-1}, x_k]$ is also a subinterval in the partition $\mathcal{D}'$ and

$$\mathrm{M}_k(f, \mathcal{D})\Delta x_k - \sum_{i \in J_k} \mathrm{M}_i(f, \mathcal{D}')\Delta u_i = 0 \quad (\text{if } |J_k| = 1).$$

Alternatively, if J_k has at least two elements for a given k, then the open interval (x_{k-1}, x_k) contains points in the partition $\mathcal{D}_0$. In this case we make the bound

$$\begin{aligned}
\mathrm{M}_k(f, \mathcal{D})\Delta x_k - \sum_{i \in J_k} \mathrm{M}_i(f, \mathcal{D}')\Delta u_i &\leq K\Delta x_k - \sum_{i \in J_k} (-K)\Delta u_i \\
&= K\left(\Delta x_k + \sum_{i \in J_k} \Delta u_i\right) \\
&= K\left(\Delta x_k + \Delta x_k\right) \\
&= K \cdot 2\Delta x_k < K \cdot 2\delta \quad (\text{if } |J_k| \geq 2).
\end{aligned}$$

Next, we note that there can never be more than N indices k for which the set J_k has at least two elements, since the partition $\mathcal{D}_0$ has only N elements. With this observation in mind, we make the bound

$$\begin{aligned}
0 &\leq \mathrm{S}(f, \mathcal{D}) - \mathrm{S}(f, \mathcal{D}') \\
&= \sum_{k=1}^{n} \mathrm{M}_k(f, \mathcal{D})\Delta x_k - \sum_{i=1}^{p} \mathrm{M}_i(f, \mathcal{D}')\Delta u_i \\
&= \sum_{k=1}^{n} \mathrm{M}_k(f, \mathcal{D})\Delta x_k - \sum_{k=1}^{n} \sum_{i \in J_k} \mathrm{M}_i(f, \mathcal{D}')\Delta u_i \\
&= \sum_{k=1}^{n} \left[\mathrm{M}_k(f, \mathcal{D})\Delta x_k - \sum_{i \in J_k} \mathrm{M}_i(f, \mathcal{D}')\Delta u_i\right] \\
&< N \cdot K \cdot 2\delta = 2NK \cdot \frac{\epsilon}{8NK} = \frac{\epsilon}{4}.
\end{aligned}$$

Analogously, this time considering f's lower Darboux sums, we can show that

$$0 \leq \mathrm{s}(f, \mathcal{D}') - \mathrm{s}(f, \mathcal{D}) < \frac{\epsilon}{4}.$$

We then use these results to make the bound

$$\begin{aligned}
\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) &= \mathrm{S}(f, \mathcal{D}) - \mathrm{S}(f, \mathcal{D}') \\
&\quad + \mathrm{S}(f, \mathcal{D}') - \mathrm{s}(f, \mathcal{D}') \\
&\quad + \mathrm{s}(f, \mathcal{D}') - \mathrm{s}(f, \mathcal{D}) \\
&< \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.
\end{aligned}$$

Finally, recalling that $I = \mathrm{s}(f) = \mathrm{S}(f)$ denotes f's Darboux integral and referring back to Equations 9.1 and 9.4, we combine the three relationships

$$\begin{aligned}
&\mathrm{s}(f, \mathcal{D}) \leq I \leq \mathrm{S}(f, \mathcal{D}), \\
&\mathrm{s}(f, \mathcal{D}) \leq \mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^{n}) \leq \mathrm{S}(f, \mathcal{D}), \\
&\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \epsilon
\end{aligned}$$

to conclude that

$$|\mathrm{R}(f, \mathcal{D}, \{\xi_k\}_{k=1}^n) - I| < \epsilon,$$

for any set of tagging points $\xi_k \in [x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. This completes the proof that Darboux integrability implies Riemann integrability, and thus that Darboux and Riemann integrability are equivalent. □

9.3 Properties of the definite integral

Having defined the definite integral and shown that all continuous functions, bounded piecewise-continuous functions, and monotonic functions are integrable, we will now cover some important properties of integrable functions. In particular, in this section we will state and prove the fundamental theorem of calculus, which relates definite and indefinite integration.

Proposition 9.20: Integrability of composite functions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let $f\colon [a, b] \to \mathbb{R}$ be an integrable function, and let $\phi\colon V \to \mathbb{R}$ be a continuous function defined on a closed, bounded subset $V \subset \mathbb{R}$ for which $f([a, b]) \subset V$. In this case the composite function $\phi \circ f\colon [a, b] \to \mathbb{R}$ is integrable.

Proof. The function ϕ is continuous and defined on a closed, bounded subset, so it is bounded and uniformly continuous by Theorem 6.36 and Proposition 6.46. Because ϕ is bounded, the composite function $\phi \circ f\colon [a, b] \to \mathbb{R}$ must also be bounded.

Using the same reasoning as in the proof of Proposition 9.12, we will assume without loss of generality that $a < b$. We then choose arbitrary $\epsilon \in \mathbb{R}^+$ and define

$$\mu = \frac{\epsilon}{b - a + 2\sup(|\phi|)}.$$

Because ϕ is uniformly continuous, there exists $\delta \in \mathbb{R}^+$ for which $\delta \leq \mu$ and,

$$|\phi(v) - \phi(u)| < \mu \text{ for all } u, v \in V \text{ for which } |v - u| < \delta.$$

Because f is integrable, by Prop 9.11 there exists a partition $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a, b]$ for which

$$\mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \delta^2.$$

Next, for all natural numbers $k \leq n$ we define the quantities

$$\begin{aligned} \mathrm{m}_k &= \mathrm{m}_k(f, \mathcal{D}) & \mathrm{M}_k &= \mathrm{M}_k(f, \mathcal{D}) \\ \mathrm{m}'_k &= \mathrm{m}'_k(\phi \circ f, \mathcal{D}) & \mathrm{M}'_k &= \mathrm{M}'_k(\phi \circ f, \mathcal{D}). \end{aligned}$$

Additionally, we define the sets A and B as

$$\begin{aligned} A &= \{k \in \mathbb{N};\ k \leq n, \mathrm{M}_k - \mathrm{m}_k < \delta\}, \\ B &= \{k \in \mathbb{N};\ k \leq n, \mathrm{M}_k - \mathrm{m}_k \geq \delta\}. \end{aligned}$$

If $k \in A$, then $\mathrm{M}_k - \mathrm{m}_k < \delta$ and so, by the definition of m_k and M_k,

$$|f(y) - f(x)| < \delta \text{ for all } x, y \in [x_{k-1}, x_k].$$

In this case, by ϕ's uniform continuity,

$$\big|\phi\big(f(y)\big) - \phi\big(f(x)\big)\big| < \mu \text{ for all } x, y \in [x_{k-1}, x_k],$$

which in turn implies that $\mathrm{M}'_k - \mathrm{m}'_k \leq \mu$ for $k \in A$.

Meanwhile, the sum of the lengths of all subintervals with indices $k \in B$ obeys

$$\begin{aligned}
\delta \sum_{k\in B} \Delta x_k = \sum_{k\in B} \delta \Delta x_k &\overset{\text{(a)}}{\leq} \sum_{k\in B} (\mathrm{M}_k - \mathrm{m}_k)\Delta x_k \\
&\overset{\text{(b)}}{\leq} \sum_{k=1}^{n} (\mathrm{M}_k - \mathrm{m}_k)\Delta x_k \\
&= \mathrm{S}(f, \mathcal{D}) - \mathrm{s}(f, \mathcal{D}) < \delta^2,
\end{aligned}$$

where (a) follows from the definition of the set B and (b) holds because the set B can contain at most, but no more than, all indices k in the partition $\mathcal{D}$. Removing the intermediate steps, the above result reads

$$\delta \sum_{k\in B} \Delta x_k \leq \sum_{k\in B} (\mathrm{M}_k - \mathrm{m}_k)\Delta x_k < \delta^2,$$

which after dividing through by δ, implies that

$$\sum_{k\in B} \Delta x_k < \delta \overset{\text{(a)}}{\leq} \mu, \tag{9.10}$$

where (a) follows from the definition of the set B. Additionally, for $k \in B$, we can liberally make the bound

$$\mathrm{M}'_k - \mathrm{m}'_k \leq 2\sup(|\phi|), \quad k \in B;$$

a simple graph of a hypothetical bounded and continuous function ϕ should convince you of this bound. We then complete the proof with the calculation

$$\begin{aligned}
\mathrm{S}(\phi \circ f, \mathcal{D}) - \mathrm{s}(\phi \circ f, \mathcal{D}) &= \sum_{k=1}^{n} (\mathrm{M}'_k - \mathrm{m}'_k)\Delta x_k \\
&= \sum_{k\in A} (\mathrm{M}'_k - m'_k)\Delta x_k + \sum_{k\in B} (\mathrm{M}'_k - \mathrm{m}'_k)\Delta x_k \\
&\leq \sum_{k\in A} \mu \Delta x_k + \sum_{k\in B} 2\sup(|\phi|)\Delta x_k \\
&\leq \mu \sum_{k\in A} \Delta x_k + 2\sup(|\phi|) \sum_{k\in B} \Delta x_k \\
&\overset{\text{(a)}}{\leq} \mu(b-a) + 2\sup(|\phi|)\mu \\
&= \frac{\epsilon \cdot (b-a)}{b-a+2\sup(|\phi|)} + \frac{\epsilon \cdot 2\sup(|\phi|)}{b-a+2\sup(|\phi|)} \\
&= \epsilon,
\end{aligned}$$

where the second term in (a) makes use of the bound in Equation 9.10. The function $\phi \circ f$ is thus integrable by Proposition 9.11. □

Proposition 9.21: Combinations of integrable functions

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let $f, g\colon [a,b] \to \mathbb{R}$ be two integrable functions, and let $\alpha \in \mathbb{R}$ be an arbitrary constant. In this case:

(i) The function $f+g$ is integrable and obeys

$$\int_a^b \big[f(x)+g(x)\big]\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$$

(ii) The function αf is integrable and obeys

$$\int_a^b (\alpha f)(x)\,\mathrm{d}x = \alpha \int_a^b f(x)\,\mathrm{d}x.$$

(iii) The function fg is integrable.

(iv) The function $|f|$ is integrable and satisfies the inequality

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)|\,\mathrm{d}x.$$

(v) If $f(x) \leq g(x)$ for all $x \in [a,b]$, then

$$\int_a^b f(x)\,\mathrm{d}x \leq \int_a^b g(x)\,\mathrm{d}x.$$

(vi) If $M \in \mathbb{R}$ is a real constant for which $|f(x)| \leq M$ for all $x \in [a,b]$, then

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq M \cdot (b-a).$$

Proof. As in Proposition 9.12, we will assume without loss of generality that $a < b$.

(i) We first note that for an arbitrary partition $\mathcal{D}$ of the interval $[a,b]$, the infimum of f plus the infimum of g on each partition subinterval is always less than or or equal to the infimum of $f+g$ on the same subinterval, implying

$$\mathrm{s}(f,\mathcal{D}) + \mathrm{s}(g,\mathcal{D}) \leq \mathrm{s}(f+g,\mathcal{D}).$$

For any two partitions $\mathcal{D}'$ and $\mathcal{D}''$ of the interval $[a,b]$, it then holds that

$$\begin{aligned}\mathrm{s}(f,\mathcal{D}') + \mathrm{s}(g,\mathcal{D}'') &\overset{\text{(a)}}{\leq} \mathrm{s}(f,\mathcal{D}'\cup\mathcal{D}'') + \mathrm{s}(g,\mathcal{D}'\cup\mathcal{D}'')\\ &\leq \mathrm{s}(f+g,\mathcal{D}'\cup\mathcal{D}''),\end{aligned}$$

where (a) holds by Proposition 9.4. Since the partitions $\mathcal{D}'$ and $\mathcal{D}''$ are arbitrary, it follows that

$$\mathrm{s}(f) + \mathrm{s}(g) \leq \mathrm{s}(f+g).$$

We can use an analogous procedure to show that $\mathrm{S}(f+g) \leq \mathrm{S}(f) + \mathrm{S}(g)$, so

$$\mathrm{s}(f) + \mathrm{s}(g) \leq \mathrm{s}(f+g) \leq \mathrm{S}(f+g) \leq \mathrm{S}(f) + \mathrm{S}(g)$$

Since f and g are both integrable, it must hold that $\mathrm{s}(f) + \mathrm{s}(g) = \mathrm{S}(f) + \mathrm{S}(g)$, which, combined with the above inequality, gives

$$\mathrm{s}(f) + \mathrm{s}(g) = \mathrm{s}(f+g) = \mathrm{S}(f+g) = \mathrm{S}(f) + \mathrm{S}(g),$$

and so $f + g$ is integrable and equal to the sum of the integrals of f and g.

(ii) Let the function $\phi\colon \mathbb{R} \to \mathbb{R}$ be given by $\phi(t) = \alpha t$, so that $\alpha f = \phi \circ f$. Because ϕ is continuous by construction and because f is integrable, the function $\alpha f = \phi \circ f$ is also integrable by Proposition 9.20.

It remains to show that the integral of αf equals the integral of f times α. Let $\mathcal{D}$ be an arbitrary partition of the interval $[a, b]$. We begin by assuming that $\alpha \leq 0$, in which case

$$\mathrm{s}(\alpha f, \mathcal{D}) = \alpha \cdot \mathrm{s}(f, \mathcal{D})$$

for any partition $\mathcal{D}$ of the interval $[a, b]$; since $\mathcal{D}$ is arbitrary, this implies that

$$\mathrm{s}(\alpha f) = \alpha \cdot \mathrm{s}(f).$$

We can show analogously that

$$\mathrm{S}(\alpha f) = \alpha \cdot \mathrm{S}(f).$$

Since f is integrable, $\mathrm{S}(f) = \mathrm{s}(f)$, which in turn implies that $\alpha \cdot \mathrm{S}(f) = \alpha \cdot \mathrm{s}(f)$ and so

$$\mathrm{s}(\alpha f) = \alpha \cdot \mathrm{s}(f) = \alpha \cdot \mathrm{S}(f) = \mathrm{S}(\alpha f),$$

which is the desired proof of equality when $\alpha \geq 0$.

We now consider the case when $\alpha < 0$. In this case

$$\mathrm{s}(\alpha f, \mathcal{D}) = \alpha \cdot \mathrm{S}(f, \mathcal{D}),$$

since a corresponding inequality holds for the infima and suprema of $\alpha \cdot f$ and (αf) on every subinterval of the partition $\mathcal{D}$. In this case

$$\begin{aligned}
\mathrm{s}(\alpha f) &= \sup\left\{\mathrm{s}(\alpha f, \mathcal{D})\,;\, \mathcal{D} \text{ is a partition of } [a,b]\right\} \\
&= \sup\left\{\alpha \mathrm{S}(f, \mathcal{D})\,;\, \mathcal{D} \text{ is a partition of } [a,b]\right\} \\
&= \alpha \cdot \inf\left\{\mathrm{S}(f, \mathcal{D})\,;\, \mathcal{D} \text{ is a partition of } [a,b]\right\} \\
&= \alpha \cdot \mathrm{S}(f).
\end{aligned}$$

Similarly, we can show that $S(\alpha f) = \alpha \cdot \mathrm{s}(f)$, and, following the same procedure as when $\alpha \geq 0$, that

$$\mathrm{s}(\alpha f) = \alpha \cdot \mathrm{S}(f) = \alpha \cdot \mathrm{s}(f) = \mathrm{S}(\alpha f),$$

which completes the proof of equality for all $\alpha \in \mathbb{R}$.

(iii) Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be the continuous function given by $\psi(t) = t^2$. Because the functions f and g are integrable, by points (i) and (ii) the functions $f + g$ and $f - g$ are also integrable, and so by Proposition 9.20 the functions

$$(f+g)^2 = \psi \circ (f+g) \quad \text{and} \quad (f-g)^2 = \psi \circ (f-g)$$

are also integrable. Again by points (i) and (ii), it then follows that the function

$$fg = \frac{1}{4}\left[(f+g)^2 - (f-g)^2\right]$$

is also integrable.

(v) The assumption $f(x) \leq g(x)$ for all $x \in [a,b]$ means that

$$\mathrm{s}(f, \mathcal{D}) \leq \mathrm{s}(g, \mathcal{D})$$

for every partition $\mathcal{D}$ of the interval $[a,b]$, which in turn implies that

$$\mathrm{s}(f) \leq \mathrm{s}(g).$$

Combined with the fact that $\mathrm{s}(f) = \mathrm{S}(f)$ and $\mathrm{s}(g) = \mathrm{S}(g)$ (because f and g are both integrable) the inequality $\mathrm{s}(f) \leq \mathrm{s}(g)$ implies that

$$\int_a^b f(x)\,\mathrm{d}x = \mathrm{S}(f) = \mathrm{s}(f) \leq \mathrm{s}(g) = \mathrm{S}(g) = \int_a^b g(x)\,\mathrm{d}x.$$

(iv) Let the function $\theta\colon \mathbb{R} \to \mathbb{R}$ be given by $\theta(t) = |t|$. Because θ is continuous and f is integrable, the function $|f| = \theta \circ f$ is also integrable by Proposition 9.20.

It remains to prove the inequality in point (iv). To do this, note that

$$-|f(x)| \leq f(x) \leq |f(x)| \text{ for all } x \in [a,b],$$

which be combine with points (ii) and (v) to get

$$-\int_a^b |f(x)|\,\mathrm{d}x = \int_a^b \left(-|f(x)|\right)\mathrm{d}x \leq \int_a^b f(x)\,\mathrm{d}x \leq \int_a^b |f(x)|\,\mathrm{d}x.$$

Written slightly more compactly, the inequality

$$-\int_a^b |f(x)|\,\mathrm{d}x \leq \int_a^b f(x)\,\mathrm{d}x \leq \int_a^b |f(x)|\,\mathrm{d}x,$$

together with the definition of the absolute value, implies that

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)|\,\mathrm{d}x.$$

(vi) This point follows directly from the calculation

$$\begin{aligned}\left|\int_a^b f(x)\,\mathrm{d}x\right| &\overset{\text{(a)}}{\leq} \int_a^b |f(x)|\,\mathrm{d}x \overset{\text{(b)}}{\leq} \int_a^b M\,\mathrm{d}x \\ &= M(b-a).\end{aligned}$$

where (a) follows from point (iv) and (b) from point (v) and the assumption that $|f(x)| \leq M$ for all $x \in [a,b]$. □

Proposition 9.22: Splitting an integration interval in two parts

Let $a, b, c \in \mathbb{R}$ be thee real numbers for which $a \leq b \leq c$. A function $f\colon [a,c] \to \mathbb{R}$ is integrable on the interval $[a,c]$ if, and only if, it is integrable on the intervals $[a,b]$ and $[b,c]$, and in this case it holds that

$$\int_a^c f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_b^c f(x)\,\mathrm{d}x.$$

Proof. For the same reason as in Proposition 9.12, we will assume that $a < b < c$.

($\Rightarrow$) Assume that f is integrable on $[a,c]$. This means that f must be bounded on $[a,c]$, and so f is also bounded on $[a,b]$ and $[b,c]$. By the criterion for integrability in Proposition 9.11, for all $\epsilon \in \mathbb{R}^+$ there exists a partition $\mathcal{D}$ of the interval $[a,c]$ for which

$$\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon.$$

We now define the partitions $\mathcal{D}_1$ and $\mathcal{D}_2$ according to

$$\begin{aligned}\mathcal{D}_1 &= (\mathcal{D} \cup \{b\}) \cap [a,b]\\ \mathcal{D}_2 &= (\mathcal{D} \cup \{b\}) \cap [b,c].\end{aligned}$$

Thus $\mathcal{D}_1$ is a partition of $[a,b]$ and $\mathcal{D}_2$ is a partition of $[b,c]$; by construction, the partitions $\mathcal{D}$, $\mathcal{D}_1$, and $\mathcal{D}_2$ obey

$$\begin{aligned}\mathrm{s}(f,\mathcal{D}\cup\{b\}) &= \mathrm{s}(f|_{[a,b]},\mathcal{D}_1) + \mathrm{s}(f|_{[b,c]},\mathcal{D}_2)\\ \mathrm{S}(f,\mathcal{D}\cup\{b\}) &= \mathrm{S}(f|_{[a,b]},\mathcal{D}_1) + \mathrm{S}(f|_{[b,c]},\mathcal{D}_2).\end{aligned}$$

We can then make the bound

$$\begin{aligned}&\big[\mathrm{S}(f|_{[a,b]},\mathcal{D}_1) - \mathrm{s}(f|_{[a,b]},\mathcal{D}_1)\big] + \big[\mathrm{S}(f|_{[b,c]},\mathcal{D}_2) - \mathrm{s}(f|_{[b,c]},\mathcal{D}_2)\big]\\ &\qquad = \mathrm{S}(f,\mathcal{D}\cup\{b\}) - \mathrm{s}(f,\mathcal{D}\cup\{b\})\\ &\qquad \overset{\text{(a)}}{\leq} \mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon,\end{aligned}$$

where (a) holds by Proposition 9.4. This in turn implies that

$$\begin{aligned}\mathrm{S}(f|_{[a,b]},\mathcal{D}_1) - \mathrm{s}(f|_{[a,b]},\mathcal{D}_1) &< \epsilon\\ \mathrm{S}(f|_{[b,c]},\mathcal{D}_2) - \mathrm{s}(f|_{[b,c]},\mathcal{D}_2) &< \epsilon,\end{aligned}$$

and so the restricted functions $f|_{[a,b]}$ and $f|_{[b,c]}$ are both integrable by Proposition 9.11.

($\Leftarrow$) Assume the restricted functions $f|_{[a,b]}$ and $f|_{[b,c]}$ are both integrable. Thus both restrictions are bounded, and so the function f is itself bounded. For any partition $\mathcal{D}'$ of the interval $[a,b]$ and for any partition $\mathcal{D}''$ of the interval $[b,c]$ it holds that

$$\begin{aligned}\mathrm{s}(f|_{[a,b]},\mathcal{D}') + \mathrm{s}(f|_{[b,c]},\mathcal{D}'') &= \mathrm{s}(f,\mathcal{D}'\cup\mathcal{D}'')\\ &\leq \mathrm{S}(f,\mathcal{D}'\cup\mathcal{D}'')\\ &= \mathrm{S}(f|_{[a,b]},\mathcal{D}') + \mathrm{S}(f|_{[b,c]},\mathcal{D}''),\end{aligned}$$

which, since the partitions $\mathcal{D}'$ and $\mathcal{D}''$ are arbitrary, in turn implies that

$$\mathrm{s}(f_{[a,b]}) + \mathrm{s}(f_{[b,c]}) \leq \mathrm{s}(f) \leq \mathrm{S}(f) \leq \mathrm{S}(f_{[a,b]}) + \mathrm{S}(f_{[b,c]}).$$

The functions $f|_{[a,b]}$ and $f|_{[b,c]}$ are both integrable, so

$$\mathrm{s}(f_{[a,b]}) + \mathrm{s}(f_{[b,c]}) = \mathrm{S}(f_{[a,b]}) + \mathrm{S}(f_{[b,c]}),$$

which immediately implies that

$$\mathrm{s}(f_{[a,b]}) + \mathrm{s}(f_{[b,c]}) = \mathrm{s}(f) = \mathrm{S}(f) = \mathrm{S}(f_{[a,b]}) + \mathrm{S}(f_{[b,c]}),$$

and so f is integrable. □

Proposition 9.22 immediately implies the following:

Corollary 9.23: Splitting an integration interval into subintervals

For any integrable function $f\colon U \to \mathbb{R}$ defined on a closed, bounded, nonempty subset $U \subset \mathbb{R}$:

(i) The function f is integrable on every closed, bounded, nonempty subinterval of the interval U.

(ii) For any $a, b, c \in U$ it holds that

$$\int_a^c f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_b^c f(x)\,\mathrm{d}x.$$

Corollary 9.23 can then be applied recursively to split an integration interval into an arbitrary number of subintervals.

Proposition 9.24: The definite integral is unaffected by changing a finite set of function values

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$ and let $f, g\colon [a,b] \to \mathbb{R}$ be two functions for which $f(x) = g(x)$ at every point $x \in [a,b]$, with the possible exception of *finitely* many points. If f is integrable, then g is also integrable and obeys

$$\int_a^b g(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x.$$

Proof. Without loss of generality, we will assume that $a < b$. Because the value of the function $f - g$ is nonzero only at a finite number of points, the function $f - g$ must be bounded and piecewise-continuous, and thus integrable by Proposition 9.14; by Proposition 9.21 (iv) the function $|f - g|$ is also integrable. The function $|f - g|$ is nonzero at only a finite number of points, which means that

$$\mathrm{s}(|f - g|, \mathcal{D}) = 0 \tag{9.11}$$

for any partition $\mathcal{D}$ of the interval $[a,b]$, in turn implying that

$$\left|\int_a^b \big[f(x) - g(x)\big]\,\mathrm{d}x\right| \overset{\text{(a)}}{\leq} \int_a^b |f(x) - g(x)|\,\mathrm{d}x \overset{\text{(b)}}{=} 0,$$

where (a) follows from Proposition 9.21 (iv) and (b) from Equation 9.11. But the absolute value of a real number can be less than or equal to zero only if the number

itself is zero, and so

$$\int_a^b \big[f(x) - g(x)\big]\,\mathrm{d}x = 0.$$

The function g is thus integrable by Proposition 9.21, since g is the difference of the integrable functions f and $f - g$. We can then complete the proof with the calculation

$$\int_a^b g(x)\,\mathrm{d}x \overset{\text{(a)}}{=} \int_a^b f(x)\,\mathrm{d}x - \int_a^b \big[f(x) - g(x)\big]\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x - 0,$$

where (a) holds by Proposition 9.21. □

Proposition 9.24 means that an integrable function's integral is unchanged by changing a finite number of the function's values. This fact will allow us to somewhat generalize the notion of integrability and the definite integral, but we will first need to introduce the concept of the extension of a function to an interval.

Definition 9.25: Extension of a function

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let P be a finite subset of the interval $[a, b]$, and let $f\colon [a, b] \setminus P \to \mathbb{R}$ be an arbitrary function. We then define the *extension* of f to the interval $[a, b]$ as the function $\hat{f}\colon [a, b] \to \mathbb{R}$ given by

$$\hat{f}(x) = \begin{cases} f(x) & x \in [a, b] \setminus P \\ 0 & x \in P. \end{cases}$$

In other words, $\hat{f}$ equals f wherever f is defined, and zero elsewhere.

We now extend the definition of integrability to functions that may be undefined at isolated points in their domain.

Definition 9.26: Integrability on an interval

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let P be a finite, nonempty subset of the interval $[a, b]$, and let $f\colon [a, b] \setminus P \to \mathbb{R}$ be any function. The function f is said to be *integrable on the interval* $[a, b]$ if its extension $\hat{f}\colon [a, b] \to \mathbb{R}$ is integrable. In this case f and $\hat{f}$'s definite integrals from a to b are equal, and we write

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b \hat{f}(x)\,\mathrm{d}x.$$

Example 9.27: Integrability on subintervals

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let P be a finite subset of the closed, bounded interval $[a, b]$. In this case every bounded, continuous function $f\colon [a, b] \setminus P \to \mathbb{R}$ is integrable, since its extension $\hat{f}$ to the interval P is bounded and piecewise-continuous.

Proposition 9.28: Integrals involving the zero function

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$ and let $f\colon [a,b] \to \mathbb{R}$ be a continuous function for which $f(x) \geq 0$ for all $x \in [a,b]$. In this case if

$$\int_a^b f(x)\,\mathrm{d}x = 0,$$

then $f(x) = 0$ for all $x \in [a,b]$.

Proof. We will prove the proposition by contradiction. Let x be an arbitrary point in the interval $[a,b]$, and assume that $f(x) \neq 0$. Because f is nonnegative at all points in its domain and $f(x) \neq 0$, it follows that

$$f(x) > 0 \text{ for all } x \in [a,b].$$

Because f is continuous, f's values on any sufficiently small neighborhood of an arbitrary point $x \in [a,b]$ are arbitrarily close to $f(x)$. As a result, there exist points $u, v \in [a,b]$ for which $u < v$

$$f(t) > \frac{f(x)}{2} \text{ for all } t \in [u,v].$$

Now let the set $\mathcal{D} = \{a, u, v, b\}$ be a partition of the interval $[a,b]$. Because f is bounded below by the number $f(x)/2$ on the subinterval $[u,v]$ of the partition $\mathcal{D}$, we can make the bound

$$\mathrm{s}(f) \geq \mathrm{s}(f,\mathcal{D}) \geq \frac{f(x)}{2} \cdot (v-u) \overset{\text{(a)}}{>} 0,$$

where (a) holds from the assumption that $f(x) > 0$ for all $x \in [a,b]$. The result $\mathrm{s}(f) > 0$ contradicts the proposition's assumption that $\int_a^b f(x)\,\mathrm{d}x = 0$, which requires that $\mathrm{s}(f) = 0$. We thus conclude that $f(x) = 0$ for all $x \in [a,b]$. □

Definition 9.29: Average value of a scalar function

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, and let $f\colon [a,b] \to \mathbb{R}$ be an integrable function. The *average value* of the function f on the interval $[a,b]$ is then equal to

$$\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x.$$

Proposition 9.30: Continuous functions take on their average value

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$. For any continuous function $f\colon [a,b] \to \mathbb{R}$ there exists a point $\xi \in [a,b]$ for which

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x.$$

In other words, every continuous function defined on a closed interval assumes its average value on this interval.

Proof. Because f is continuous and defined on a closed, bounded interval, it is bounded by Theorem 6.36. We then note that

$$\inf(f) \leq f(x) \leq \sup(f) \text{ for all } x \in [a, b];$$

combined with Proposition 9.21 (v), this inequality implies that

$$\inf(f) \cdot (b-a) = \int_a^b \inf(f)\,\mathrm{d}x \leq \int_a^b f(x)\,\mathrm{d}x \leq \int_a^b \sup(f)\,\mathrm{d}x = \sup(f) \cdot (b-a).$$

We then divide through by the quantity $(b-a)$ to get

$$\inf(f) \leq \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x \leq \sup(f).$$

By Proposition 6.38, a continuous function defined on a closed interval attains all values between its supremum and infimum; since $\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x$ lies between f's supremum and infimum, there exists a point $\xi \in [a, b]$ for which

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x. \qquad \square$$

Theorem 9.31: First fundamental theorem of calculus

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let $f\colon [a, b] \to \mathbb{R}$ be an integrable function, and let c be an arbitrary point in the interval $[a, b]$. In this case the function $F\colon [a, b] \to \mathbb{R}$ given by

$$F(x) = \int_c^x f(t)\,\mathrm{d}t, \quad x \in [a, b],$$

is continuous. More so, if f is continuous at any point $u \in (a, b)$, then F is differentiable at u and obeys

$$F'(u) = f(u).$$

The first fundamental theorem of calculus thus guarantees the existence of a primitive function F for a continuous function f.

Proof. We first note that the function f is integrable on every closed, nonempty subinterval of the interval $[a, b]$ by Corollary 9.23, and so the function F is well-defined for any $c \in [a, b]$.

We will first show that F is uniformly continuous, then show that $F'(u) = f(u)$.

(i) Because f is integrable, it is necessarily bounded by Proposition 9.11. For any

two $x, y \in [a, b]$, we then make the calculation

$$\begin{aligned} |F(y) - F(x)| &\overset{\text{(a)}}{=} \left| \int_c^y f(t)\,\mathrm{d}t - \int_c^x f(t)\,\mathrm{d}t \right| \\ &\overset{\text{(b)}}{=} \left| \int_x^c f(t)\,\mathrm{d}t + \int_c^y f(t)\,\mathrm{d}t \right| \\ &\overset{\text{(c)}}{=} \left| \int_x^y f(t)\,\mathrm{d}t \right| \\ &\overset{\text{(d)}}{\leq} \sup(f) \cdot |y - x|, \end{aligned} \tag{9.12}$$

where (a) follows from the definition of the function F, (b) from swapping the limits of integration in the second integral, (c) from Proposition 9.22, and (d) from Proposition 9.21 (vi) and the fact that $f(t) \leq \sup(f)$ for all $t \in [x, y]$. Now choose arbitrary $\epsilon > 0$ and define

$$\delta = \frac{\epsilon}{\sup(f)}.$$

In this case, for all $\xi, \eta \in [a, b]$ for which $|\xi - \eta| < \delta$ it holds that

$$|F(\eta) - F(\xi)| \overset{\text{(a)}}{\leq} \sup(f) \cdot |\eta - \xi| < \sup(f) \cdot \delta = \sup(f) \cdot \frac{\epsilon}{\sup(f)} = \epsilon,$$

where (a) follows from the result of Equation 9.12, and so F is uniformly continuous (and thus also continuous) on $[a, b]$.

(ii) We will now show that $F'(u) = f(u)$ for all $u \in (a, b)$ at which f is continuous. Let f be continuous at the point $u \in (a, b)$, which means that for all $\epsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ for which

$$|f(t) - f(u)| < \frac{\epsilon}{2} \text{ for all } t \in [a, b] \text{ for which } |t - u| < \delta.$$

For all $x \in [a, b]$ for which $0 < |x - u| < \delta$ we can then make the bound

$$\begin{aligned} \left| \frac{F(x) - F(u)}{x - u} - f(u) \right| &\overset{\text{(a)}}{=} \left| \frac{1}{x - u} \left(\int_c^x f(t)\,\mathrm{d}t - \int_c^u f(t)\,\mathrm{d}t \right) - f(u) \right| \\ &\overset{\text{(b)}}{=} \left| \frac{1}{x - u} \left(\int_c^x f(t)\,\mathrm{d}t + \int_u^c f(t)\,\mathrm{d}t \right) - f(u) \right| \\ &\overset{\text{(c)}}{=} \left| \frac{1}{x - u} \int_u^x f(t)\,\mathrm{d}t - \frac{f(u)}{x - u} \int_u^x \mathrm{d}t \right| \\ &\overset{\text{(d)}}{=} \left| \frac{1}{x - u} \left(\int_u^x f(t)\,\mathrm{d}t - \int_u^x f(u)\,\mathrm{d}t \right) \right| \\ &\overset{\text{(e)}}{=} \frac{1}{|x - u|} \left| \int_u^x \left[f(t) - f(u) \right] \mathrm{d}t \right| \\ &\overset{\text{(f)}}{\leq} \frac{1}{|x - u|} \cdot \frac{\epsilon}{2} \cdot |x - u| < \epsilon, \end{aligned}$$

where (a) follows from the definition of the function F, (b) from swapping the limits of integration in the second integral, (c) from Proposition 9.22 and the

fact that $\int_u^x \mathrm{d}t = x-u$, (d) from factoring out the common term $1/(x-u)$, and (e) and (f) from Proposition 9.21 and the earlier bound $|f(t)-f(u)| < \epsilon/2$. The above calculation shows that

$$f(u) = \lim_{x\to u} \frac{F(x)-F(u)}{x-u},$$

i.e. that $F'(u) = f(u)$. □

Theorem 9.32: The second fundamental theorem of calculus

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \le b$, let $f\colon [a,b] \to \mathbb{R}$ be an integrable function, and let the function $F\colon [a,b] \to \mathbb{R}$ be continuous on the closed interval $[a,b]$, differentiable on the open interval (a,b), and let

$$F'(x) = f(x) \text{ for all } x \in (a,b).$$

In this case

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

The second fundamental theorem of calculus thus makes possible the computation of definite integrals from indefinite integrals.

Proof. As in Proposition 9.12, we will assume without loss of generality that $a < b$. First choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the function f is integrable, by Proposition 9.11 there exists a partition $\mathcal{D} = \{x_k\}_{k=0}^n$ of the interval $[a,b]$ for which

$$\mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon.$$

For all $k = 1, 2, \ldots, n$, the mean value theorem applied to the function F on the partition subinterval $[x_{k-1}, x_k]$ guarantees the existence of a point $\xi_k \in (x_{k-1}, x_k)$ for which

$$F(x_k) - F(x_{k-1}) = F'(\xi_k)\cdot(x_k - x_{k-1}) = f(\xi_k)\Delta x_k.$$

Using this equality, the Riemann sum of f with respect to the partition $\mathcal{D}$ and the set of tagging points $\{\xi_k\}_{k=1}^n$ is equal to

$$\begin{aligned}\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=1}^n) &= \sum_{k=1}^n f(\xi_k)\Delta x_k \\ &= \sum_{k=1}^n \big[F(x_k) - F(x_{k-1})\big] \\ &= F(x_n) - F(x_0) = F(b) - F(a).\end{aligned}$$

Recalling Equations 9.1 and 9.4, we momentarily pause to note the inequalities

$$\mathrm{s}(f,\mathcal{D}) \le \int_a^b f(x)\,\mathrm{d}x \le \mathrm{S}(f,\mathcal{D}), \quad \text{and} \quad \mathrm{s}(f,\mathcal{D}) \le \mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=1}^n) \le \mathrm{S}(f,\mathcal{D})$$

which we then use to make the calculation

$$\begin{aligned}\left|F(b) - F(a) - \int_a^b f(x)\,\mathrm{d}x\right| &= \left|\mathrm{R}(f,\mathcal{D},\{\xi_k\}_{k=1}^n) - \int_a^b f(x)\,\mathrm{d}x\right| \\ &\le \mathrm{S}(f,\mathcal{D}) - \mathrm{s}(f,\mathcal{D}) < \epsilon.\end{aligned}$$

The inequality holds for arbitrary $\epsilon \in \mathbb{R}^+$, which implies that

$$F(b) - F(a) - \int_a^b f(x)\,\mathrm{d}x = 0,$$

which we then trivially rearrange to get the desired result

$$F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x. \qquad \square$$

Oftentimes, the expression $F(b) - F(a)$ in the second fundamental theorem of calculus is written in the shorthand form

$$F(b) - F(a) = F(x)\big|_{x=a}^{x=b} = F(x)\big|_a^b.$$

We will now use the second fundamental theorem of calculus to apply some of the indefinite integration rules from Chapter 8 to the computation of definite integrals.

Proposition 9.33: Integration by parts for computation of definite integrals

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, let $f, g\colon [a,b] \to \mathbb{R}$ be two integrable functions and let the functions $F, G\colon [a,b] \to \mathbb{R}$ be continuous on the closed interval $[a,b]$, differentiable on the open interval (a,b), and obey $F'(x) = f(x)$ and $G'(x) = g(x)$ for all $x \in (a,b)$. In this case

$$\int_a^b F(x)g(x)\,\mathrm{d}x = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)\,\mathrm{d}x.$$

Proof. We first define the function $H = F \cdot G$. As the product of continuous and differentiable functions, H is itself continuous on $[a,b]$ and differentiable on (a,b). More so, for all $x \in (a,b)$ it holds that

$$H'(x) \overset{\text{(a)}}{=} F'(x)G(x) + F(x)G'(x) = (fG + Fg)(x),$$

where (a) uses the product rule for differentiation. Because the functions f, g, F, and G are all integrable, the function $fG + Fg$ is also integrable. We can then apply the second fundamental theorem of calculus to get

$$\begin{aligned}\int_a^b \big[f(x)G(x) + F(x)g(x)\big]\,\mathrm{d}x &= \int_a^b H'(x)\,\mathrm{d}x \\ &= H(b) - H(a) \\ &= F(b)G(b) - F(a)G(a). \qquad \square\end{aligned}$$

Proposition 9.34: Change of variables for computation of definite integrals

Let $f\colon U \to \mathbb{R}$ be a function defined on the real subset $U \subset \mathbb{R}$, let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$, and let $s\colon [a,b] \to \mathbb{R}$ be a continuous function for which $s([a,b]) \subset U$. More so, let $F\colon U \to \mathbb{R}$ be a continuous function for which the following properties hold for all $x \in (a,b)$:

- $s(x)$ is an interior point of U,

- F is differentiable at the point $s(x)$, and
- $F'\big(s(x)\big) = f\big(s(x)\big)$.

Finally, assume that s is differentiable on the interval (a, b) and that the function $(f \circ s) \cdot s' \colon (a, b) \to \mathbb{R}$ is integrable on the interval $[a, b]$.

In this case,

$$\int_a^b f\big(s(x)\big)s'(x)\,\mathrm{d}x = F\big(s(b)\big) - F\big(s(a)\big).$$

Proof. The functions F and s are both continuous, so the composite function $F \circ s$ is itself continuous on $[a, b]$. Similarly, by the chain rule, $F \circ s$ is differentiable on (a, b) and for all $x \in (a, b)$, its derivative is equal to

$$(F \circ s)'(x) = F'\big(s(x)\big) \cdot s'(x) = f\big(s(x)\big) \cdot s'(x).$$

Next, let $\phi \colon [a, b] \to \mathbb{R}$ be the extension of the function $(f \circ s) \cdot s' \colon (a, b) \to \mathbb{R}$ to the interval $[a, b]$, defined so that $\phi(a) = \phi(b) = 0$. The function $(f \circ s) \cdot s' \colon (a, b) \to \mathbb{R}$ is integrable by the proposition's assumptions, so by Definition 9.26 the extension ϕ is also integrable. We then apply the second fundamental theorem of calculus to get

$$\int_a^b f\big(s(x)\big)s'(x)\,\mathrm{d}x = \int_a^b \phi(x)\,\mathrm{d}x = (F \circ s)(b) - (F \circ s)(a). \qquad \square$$

Admittedly, Proposition 9.34 is written in a rather general form. Under stricter restrictions, the proposition simplifies considerably.

Corollary 9.35: Simplified change of variables for definite integration

Let $v, w \in \mathbb{R}$ be two real numbers for which $v \leq w$, and let $f \colon [v, w] \to \mathbb{R}$ be a continuous function. Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$ and let the function $s \colon [a, b] \to \mathbb{R}$ be continuous on the closed interval $[a, b]$, continuously differentiable on the open interval (a, b), and let $s\big((a, b)\big) \subset (v, w)$. If the function $(f \circ s) \cdot s' \colon (a, b) \to \mathbb{R}$ is bounded, then

$$\int_a^b f\big(s(x)\big)s'(x)\,\mathrm{d}x = \int_{s(a)}^{s(b)} f(t)\,\mathrm{d}t.$$

Proof. We first define the function $F \colon [v, w] \to \mathbb{R}$ according to

$$F(u) = \int_{s(a)}^{u} f(t)\,\mathrm{d}t, \quad u \in [v, w].$$

By the first fundamental theorem of calculus, the function F is continuous on $[v, w]$, differentiable on (v, w), and for all $u \in (v, w)$ obeys

$$F'(u) = f(u).$$

Because s is continuous and $s\big((a, b)\big) \subset (v, w)$ by the proposition's assumptions, it follows from Corollary 6.40 that $s([a, b]) \subset [v, w]$. Again by the proposition's assumptions, the function $(f \circ s) \cdot s' \colon (a, b) \to \mathbb{R}$ is continuous and bounded, and so

is integrable on the interval $[a, b]$. We then apply Proposition 9.34 and the second fundamental theorem of calculus to get

$$\int_a^b f\big(s(x)\big)s'(x)\,\mathrm{d}x = F\big(s(b)\big) - F\big(s(a)\big) = \int_{s(a)}^{s(b)} f(t)\,\mathrm{d}t. \qquad \square$$

We will now illustrate the above integration rules with a number of practical examples.

Example 9.36: Computing definite integrals

(1) (*Integrating the power function*) For any natural number $n \in \mathbb{N}$ and real numbers $a, b \in \mathbb{R}$, we can use the second fundamental theorem of calculus to compute

$$\int_a^b x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1}\bigg|_a^b = \frac{b^{n+1} - a^{n+1}}{n+1}.$$

(2) (*Integrating the sine function*) The definite integral of the sine function over the interval spanned by any two real numbers $a, b \in \mathbb{R}$ is

$$\int_a^b \sin x\,\mathrm{d}x = -\cos x\big|_a^b = -\cos b + \cos a.$$

As a concrete example, the area between the sine function's graph and the horizontal axis on the interval from 0 to π is

$$\int_0^\pi \sin x\,\mathrm{d}x = \cos 0 - \cos \pi = -(-1) + 1 = 2.$$

On the other hand, the integral of the sine function from 0 to 2π is

$$\int_0^{2\pi} \sin x\,\mathrm{d}x = \cos 0 - \cos 2\pi = -(-1) - 1 = 0.$$

Even though the area between the sine function's graph and the horizontal axis on the interval from π to 2π is also equal to 2, this area contributes negatively to the above definite integral because the sine function is negative on the interval x in $(\pi, 2\pi)$ (perhaps see the discussion in Remark 9.9 for context).

(3) (*Computing the area of a circle*) Consider the semicircle bounded by horizontal axis and the graph of the function

$$f(x) = \sqrt{r^2 - x^2}, \quad x \in [-r, r].$$

The area S_c of a circle of radius r is twice the area, say S_sc, of the above semicircle. We will compute the integral with the change of variables

$$x = r\sin t, \quad \mathrm{d}x = r\cos t\,\mathrm{d}t.$$

In this case the interval of integration changes to $t \in [-\pi/2, \pi/2]$, and the area

of the circle is

$$\begin{aligned} S_{\mathrm{c}} = 2S_{\mathrm{sc}} &= 2\int_{-r}^{r} \sqrt{r^2 - x^2}\,\mathrm{d}x \\ &= 2\int_{-\pi/2}^{\pi/2} \sqrt{r^2 - r^2\sin^2 t}\; r\cos t\,\mathrm{d}t \\ &= 2r^2\int_{-\pi/2}^{\pi/2} \sqrt{1-\sin^2 t}\;\cos t\,\mathrm{d}t = 2r^2\int_{-\pi/2}^{\pi/2}\cos^2 t\,\mathrm{d}t \\ &= 2r^2\int_{-\pi/2}^{\pi/2} \frac{1+\cos 2t}{2}\,\mathrm{d}t \\ &= r^2\left(t + \frac{\sin 2t}{2}\right)_{-\pi/2}^{\pi/2} = \pi r^2. \end{aligned}$$

As expected from elementary geometry, the area of a circle of radius r is πr^2.

9.4 Numerical integration

The second fundamental theorem of calculus is a powerful tool for the computation of definite integrals, but is useful only when one can analytically compute the primitive function of the integrand. When a primitive function is not known or its computation is infeasible, it is possible to approximate definite integrals to arbitrary precision using integration. Numerical integration falls within the scope of the more general field of numerical methods, and we will only cover two simple techniques here.

Topic 9.4: Building up to the trapezoidal rule

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let $f\colon [a,b] \to \mathbb{R}$ be an integrable function, and let $n \in \mathbb{N}$ be an arbitrary natural number. We first divide the interval $[a,b]$ into n subintervals determined by the evenly-spaced boundary points

$$x_k = a + k\frac{b-a}{n}, \quad k = 0, 1, \ldots, n.$$

On each of these n subintervals, we approximate the region between the horizontal line $[x_{k-1}, x_k] \times \{0\}$ and f's graph with the region between $[x_{k-1}, x_k] \times \{0\}$ and the line connecting the points $\big(x_{k-1}, f(x_{k-1})\big)$ and $\big(x_k, f(x_k)\big)$. As long as $f(x_{k-1})$ and $f(x_k)$ have the same sign, this approximation region is a trapezoid, hence the name "trapezoidal rule".

Letting $y_k = f(x_k)$ for $k = 0, 1, \ldots, n$, we then sum the area of each of the trapezoids to get the approximation

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{n}\left(\frac{y_n - y_0}{2} + y_1 + y_2 + \cdots + y_{n-1}\right).$$

Of course, any approximation is of little use without an estimate of its error; we formulate the trapezoidal rule's error precisely in the following proposition.

Proposition 9.37: The trapezoidal rule

Let $f\colon U \to \mathbb{R}$ be a twice continuously differentiable function defined on the open interval $U \subset \mathbb{R}$, let $a, b \in U$ be two points for which $a < b$, fix arbitrary $n \in \mathbb{N}$, and define

$$y_k = f\left(a + k\frac{b-a}{n}\right), \quad k = 0, 1, \ldots, n.$$

In this case (without proof) there exists a point $c \in [a, b]$ for which

$$\int_a^b f(x)\,\mathrm{d}x = \left(\frac{y_n + y_0}{2} + y_1 + y_2 + \cdots + y_{n-1}\right) - \frac{(b-a)^3}{12n^2} f''(c).$$

Because f is twice continuously differentiable, the function $|f''|$, being continuous, is bounded above by some constant on the interval $[a, b]$, and so the above proposition gives an estimate of the error of the trapezoidal rule; this error is inversely proportional to the square of the number of subintervals n.

In most cases the area approximation using the trapezoidal rule is better than the corresponding Riemann sum approximation for a given number of subintervals n.

Simpson's rule is a generalization of the trapezoidal rule that uses parabolas instead of line segments to approximate the area between a function's graph and the horizontal axis on each subinterval. We state Simpson's rule and its corresponding error bound in the following proposition.

Proposition 9.38: Simpson's rule

Let $f\colon U \to \mathbb{R}$ be a four-times continuously differentiable function defined on the open interval $U \subset \mathbb{R}$, let $a, b \in U$ be two points for which $a < b$, fix arbitrary $n \in \mathbb{N}$, and define

$$y_k = f\left(a + k\frac{b-a}{2n}\right), \quad k = 0, 1, \ldots, 2n.$$

In this case (without proof) there exists a point $c \in [a, b]$ for which

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{6n}\left(y_0 + 4\sum_{k=1}^{n} y_{2k-1} + 2\sum_{k=1}^{n-1} y_{2k} + y_{2n}\right) - \frac{(b-a)^5}{2880n^4} f^{(4)}(c).$$

Because f is twice continuously differentiable, the function $|f''|$, being continuous, is bounded above by some constant on the interval $[a, b]$, and so the above proposition gives an estimate of the error of the trapezoidal rule; this error is inversely proportional to the fourth power of the number of subintervals n.

9.5 Improper integrals

We have thus far computed definite integrals only of bounded functions defined on closed, bounded intervals. In this section we will expand the definition of the definite integral to apply to unbounded functions or to functions defined on unbounded domains. This type of generalized definite integral is called an *improper integral*; we will see that its existence depends on the existence of one or more limits. In cases where an improper integral exists, it is said to *converge*; when an improper integral does not exist, it is said to *diverge*.

9.5.1 Improper integrals on bounded intervals

For the entirety of this section, let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$.

Topic 9.5: Improper integrals on bounded intervals

(I) The improper integral of a continuous function $f\colon (a, b] \to \mathbb{R}$ exists if the limit

$$\lim_{\epsilon \to 0^+} \int_{a+\epsilon}^{b} f(x)\,\mathrm{d}x$$

also exists. In this case the limit is called the improper integral of f in the limits from a to b, and we use the shorthand notation

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\epsilon \to 0^+} \int_{a+\epsilon}^{b} f(x)\,\mathrm{d}x.$$

(II) The improper integral of a continuous function $f\colon [a, b) \to \mathbb{R}$ exists if the limit

$$\lim_{\epsilon \to 0^+} \int_{a}^{b-\epsilon} f(x)\,\mathrm{d}x$$

also exists. In this case the limit is called the improper integral of f in the limits from a to b, and we use the shorthand notation

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\epsilon \to 0^+} \int_{a}^{b-\epsilon} f(x)\,\mathrm{d}x.$$

(III) Let P be a nonempty, finite subset of the interval $[a, b]$, let $f\colon [a, b] \setminus P \to \mathbb{R}$ be a continuous function, and let $r_0, r_1, \ldots, r_n \in \mathbb{R}$ be real numbers for which

$$a = r_0 < r_1 < \cdots < r_{n-1} < r_n = b$$

and either

(a) $P \cap [r_{k-1}, r_k] = \{r_{k-1}\}$ or

(b) $P \cap [r_{k-1}, r_k] = \{r_k\}$ for all $k = 1, 2, \ldots, n$.

In this case the function f's improper integral exists if each of the type (I) or type (II) improper integrals

$$\int_{r_{k-1}}^{r_k} f(x)\,\mathrm{d}x, \quad k = 1, 2, \ldots$$

also exist. In this case, f's improper integral of f in the limits from a to b is

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{k=1}^{n} \int_{r_{k-1}}^{r_k} f(x)\,\mathrm{d}x.$$

Note that there are many possible choices of the numbers $r_0, r_1, \ldots, r_n$, but the resulting value of the improper integral is independent of the choice.

If the function f mentioned in the cases (I), (II), and (III) above is bounded in addition to being continuous, then f is integrable by Example 9.27, and f's improper integral in the limits from a to b converges to f's definite integral from a to b.

Proposition 9.39: Applying the second fundamental theorem of calculus to improper integrals

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let P be a nonempty, finite subset of the interval $[a, b]$, and let $f\colon [a, b] \setminus P \to \mathbb{R}$ be a continuous function. If there exists a continuous function $F\colon [a, b] \to \mathbb{R}$ that is differentiable on the subset $(a, b) \setminus P$ and obeys

$$F'(x) = f(x) \text{ for all } x \in (a, b) \setminus P,$$

then the improper integral of f exists and obeys

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Proof. We divide the proof into the following three cases:

(i) If $P = \{a\}$, the proposition follows from the calculation

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \lim_{\epsilon \to 0^+} \int_{a+\epsilon}^b f(x)\,\mathrm{d}x \\ &\overset{\text{(a)}}{=} \lim_{\epsilon \to 0^+} \big[F(b) - F(a+\epsilon)\big] \\ &\overset{\text{(b)}}{=} F(b) - F(a),\end{aligned}$$

where (a) follows from the second fundamental theorem of calculus and (b) from the continuity of F at a.

(ii) If $P = \{b\}$, we make the analogous calculation

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \lim_{\epsilon \to 0^+} \int_a^{b-\epsilon} f(x)\,\mathrm{d}x \\ &= \lim_{\epsilon \to 0^+} \big[F(b-\epsilon) - F(a)\big] \\ &= F(b) - F(a).\end{aligned}$$

(iii) Now let P be an arbitrary nonempty, finite subset of the interval $[a, b]$. Let $r_0, r_1, \ldots, r_n \in \mathbb{R}$ be real numbers for which

$$a = r_0 < r_1 < \cdots < r_{n-1} < r_n = b$$

and either

(a) $P \cap [r_{k-1}, r_k] = \{r_{k-1}\}$ or

(b) $P \cap [r_{k-1}, r_k] = \{r_1\}$ for all $k = 1, 2, \ldots, n$.

In this case points (i) and (ii) imply that for each $k = 1, 2, \ldots, n$ the improper integral of the function f exists on the interval $[r_{k-1}, r_k]$ and is equal to

$$\int_{r_{k-1}}^{r_k} f(x)\,\mathrm{d}x = F(r_k) - F(r_{k-1}).$$

We then divide the integration integral as in Corollary 9.23 to get

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{k=1}^{n}\int_{r_{k-1}}^{r_k} f(x)\,\mathrm{d}x \\ &= \sum_{k=1}^{n}\big[F(r_k)-F(r_{k-1})\big] \\ &= F(b)-F(a),\end{aligned}$$

which completes the proof. □

Example 9.40: Improper integrals on bounded intervals

(1) The improper integral of the function $x \mapsto 1/\sqrt{x}$ in the limits from 0 to 1 converges by the calculation

$$\int_0^1 \frac{\mathrm{d}x}{\sqrt{x}} = \lim_{\epsilon\to 0^+}\int_\epsilon^1 \frac{\mathrm{d}x}{\sqrt{x}} = \lim_{\epsilon\to 0^+} 2\sqrt{x}\Big|_\epsilon^1 = \lim_{\epsilon\to 0^+}(2-\epsilon) = 2.$$

(2) The improper integral of the function $x \mapsto 1/x$ in the limits from 0 to 1 diverges by the calculation

$$\int_0^1 \frac{\mathrm{d}x}{x} = \lim_{\epsilon\to 0^+}\int_\epsilon^1 \frac{\mathrm{d}x}{x} = \lim_{\epsilon\to 0^+} \ln|x|\Big|_\epsilon^1 = \lim_{\epsilon\to 0^+}(\ln 1 - \ln\epsilon) = \infty.$$

Notice that both of the above integrands take the form $f(x) = 1/x^\alpha$ for some $\alpha \in \mathbb{R}$; we consider a more general form of this type of integrand in the following proposition.

Proposition 9.41: Convergence criterion for improper integrals of power-function denominators on bounded domains

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let $g\colon [a,b] \to \mathbb{R}$ be a continuous function, and define the constant $\alpha \in \mathbb{R}$. In this case the improper integral

$$\int_a^b \frac{g(x)}{(x-a)^\alpha}\,\mathrm{d}x$$

(i) converges if $\alpha < 1$ and

(ii) diverges if $\alpha \geq 1$ and $g(a) \neq 0$.

Proof.

(i) Let $n \in \mathbb{N}$ be a natural number for which $n(1-\alpha) > 1$ and define the new integration variable

$$t = \sqrt[n]{x-a},$$

in which case $x = a + t^n$ and $\mathrm{d}x = nt^{n-1}$. We then make the calculation

$$\begin{aligned}\int_a^b \frac{g(x)}{(x-a)^\alpha}\,\mathrm{d}x &\overset{\text{(a)}}{=} \lim_{\epsilon\to 0^+}\int_{a+\epsilon}^b \frac{g(x)}{(x-a)^\alpha}\,\mathrm{d}x \\ &\overset{\text{(b)}}{=} \lim_{\epsilon\to 0^+}\int_{\sqrt[n]{\epsilon}}^{\sqrt[n]{b-a}} ng(a+t^n)t^{n(1-\alpha)-1}\,\mathrm{d}t \\ &\overset{\text{(c)}}{=} \int_0^{\sqrt[n]{b-a}} ng(a+t^n)t^{n(1-\alpha)-1}\,\mathrm{d}t,\end{aligned}$$

where (a) holds by the definition of an improper integral, (b) from the change of variables $x = a + t^n$, and in (c) we have switched the order of limit and integral evaluation because the function $t \mapsto ng(a+t^n)t^{n(1-\alpha)-1}$ is continuous on the interval $[0, \sqrt[n]{b-a}]$. And since the function $t \mapsto ng(a+t^n)t^{n(1-\alpha)-1}$ is continuous and defined on a closed interval, it must also be integrable, and so the above improper integral reduces to the corresponding definite integral.

(ii) Assume $g(a) > 0$. The function g is continuous, so there exists a point $c \in (a, b]$ for which $c - a < 1$ and

$$g(x) > g(a)/2 \text{ for all } x \in [a, c].$$

It suffices to show that the improper integral of the function $x \mapsto g(x)/(x-a)^\alpha$ diverges in the limits from a to c. To show this integral indeed diverges, for any small-enough $\epsilon \in \mathbb{R}^+$ we first make the calculation

$$\begin{aligned}\int_{a+\epsilon}^c \frac{g(x)}{(x-a)^\alpha}\,\mathrm{d}x &\geq \int_{a+\epsilon}^c \frac{g(a)}{2(x-a)} \\ &= \frac{g(a)}{2}\big[\ln(c-a) - \ln\epsilon\big].\end{aligned}$$

The improper integral evidently diverges because the limit

$$\lim_{\epsilon\to 0^+}\big[\ln(c-a) - \ln\epsilon\big]$$

also diverges. The proof is analogous when $g(a) < 0$. □

9.5.2 Improper integrals on unbounded intervals

Topic 9.6: Improper integrals on unbounded intervals

(IV) Let $b \in \mathbb{R}$ be an arbitrary real number and let $f\colon (-\infty, b] \to \mathbb{R}$ be a continuous function. The improper integral of the function f exists if the limit

$$\lim_{a\to-\infty}\int_a^b f(x)\,\mathrm{d}x$$

also exists. In this case the limit is called the improper integral of f in the limits from $-\infty$ to b, and we use the shorthand notation

$$\int_{-\infty}^{b} f(x)\,\mathrm{d}x = \lim_{a\to-\infty}\int_{a}^{b} f(x)\,\mathrm{d}x.$$

(V) Let $a \in \mathbb{R}$ be an arbitrary real number and let $f\colon [a,\infty) \to \mathbb{R}$ be a continuous function. The improper integral of the function f exists if the limit

$$\lim_{b\to\infty}\int_{a}^{b} f(x)\,\mathrm{d}x$$

also exists. In this case the limit is called the improper integral of f in the limits from a to ∞, and we use the shorthand notation

$$\int_{a}^{\infty} f(x)\,\mathrm{d}x = \lim_{b\to\infty}\int_{a}^{b} f(x)\,\mathrm{d}x.$$

(VI) Let $b \in \mathbb{R}$ be an arbitrary real number, let P be a nonempty, finite subset of the interval $(-\infty, b]$, let $f\colon (-\infty,b]\setminus P \to \mathbb{R}$ be a continuous function, and let $v \in \mathbb{R}$ be any real number less than the smallest value in P. The improper integral of the function f exists if the type (IV) and type (III) improper integrals

$$\int_{-\infty}^{v} f(x)\,\mathrm{d}x \quad \text{and} \quad \int_{v}^{b} f(x)\,\mathrm{d}x$$

also exist. In this case the improper integral of f is given by

$$\int_{-\infty}^{b} f(x)\,\mathrm{d}x = \int_{-\infty}^{v} f(x)\,\mathrm{d}x + \int_{v}^{b} f(x)\,\mathrm{d}x.$$

Note that the result is independent of the choice of v.

(VII) Let $a \in \mathbb{R}$ be an arbitrary real number, let P be a nonempty, finite subset of the interval $[a,\infty)$, let $f\colon [a,\infty)\setminus P \to \mathbb{R}$ be a continuous function, and let $w \in \mathbb{R}$ be any real number larger than the largest value in P. The improper integral of the function f exists if the type (III) and type (IV) improper integrals

$$\int_{a}^{w} f(x)\,\mathrm{d}x \quad \text{and} \quad \int_{w}^{\infty} f(x)\,\mathrm{d}x$$

also exist. In this case the improper integral of f is given by

$$\int_{a}^{\infty} f(x)\,\mathrm{d}x = \int_{a}^{w} f(x)\,\mathrm{d}x + \int_{w}^{\infty} f(x)\,\mathrm{d}x$$

Note that the result is independent of the choice of w.

(VIII) Let P be a finite subset of the real numbers, let $f\colon \mathbb{R}\setminus P \to \mathbb{R}$ be a continuous function, and let $u \in \mathbb{R}\setminus P$ be an arbitrary real number. In this case the improper integral of the function f exists if the type (IV)/(VI) or (V)/(VII) improper integrals

$$\int_{-\infty}^{u} f(x)\,\mathrm{d}x \quad \text{and} \quad \int_{u}^{\infty} f(x)\,\mathrm{d}x$$

also exist. In this case the improper integral of f is given by

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{u} f(x)\,\mathrm{d}x + \int_{u}^{\infty} f(x)\,\mathrm{d}x.$$

Note that the result is independent of the choice of $u \in \mathbb{R} \setminus P$.

Example 9.42: Improper integrals on unbounded domains

(1) Let us first solve the following integral:

$$\begin{aligned}\int_1^{\infty} \frac{\mathrm{d}x}{x} &= \lim_{b\to\infty} \int_1^b \frac{\mathrm{d}x}{x} \\ &= \lim_{b\to\infty} \ln|x|\Big|_1^b \\ &= \lim_{b\to\infty} (\ln b - \ln 1) = \infty.\end{aligned}$$

The limit diverges, and so the improper integral of the function $x \mapsto 1/x$ in the limits from 1 to ∞ also diverges.

(2) We now consider the exponential function $x \mapsto \mathrm{e}^x$ on the interval $(-\infty, 0]$. This improper integral *does* exist, as shown by the calculation

$$\begin{aligned}\int_{-\infty}^{0} \mathrm{e}^x\,\mathrm{d}x &= \lim_{a\to-\infty} \int_a^0 \mathrm{e}^x\,\mathrm{d}x \\ &= \lim_{a\to-\infty} \mathrm{e}^x\Big|_a^0 = \lim_{a\to-\infty} \left(\mathrm{e}^0 - \mathrm{e}^a\right) = 1.\end{aligned}$$

(3) We conclude the example with one more convergent improper integral:

$$\begin{aligned}\int_1^{\infty} \frac{\mathrm{d}x}{x^3} &= \lim_{b\to\infty} \int_1^b \frac{\mathrm{d}x}{x^3} = \lim_{b\to\infty} \frac{1}{-2x^2}\Big|_1^b \\ &= \lim_{b\to\infty} \left(-\frac{1}{2b^2} + \frac{1}{2}\right) = \frac{1}{2}.\end{aligned}$$

We conclude the chapter with two convergence criteria for improper integrals on unbounded domains.

Proposition 9.43: Convergence criterion for improper integrals on unbounded domains

Let $a \in \mathbb{R}$ be an arbitrary real number and let $f\colon [a, \infty) \to \mathbb{R}$ be a continuous function. The improper integral

$$\int_a^{\infty} f(x)\,\mathrm{d}x$$

converges if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists a real number $b \in \mathbb{R}$ for which

$b > a$ and

$$\left| \int_u^v f(x)\,\mathrm{d}x \right| < \epsilon \text{ for all } u, v \in \mathbb{R} \text{ for which } b < u < v.$$

Proof. We first define the function $F\colon [a, \infty) \to \mathbb{R}$ according to

$$F(t) = \int_a^t f(x)\,\mathrm{d}x.$$

($\Rightarrow$) Assume the improper integral converges, which means that the limit

$$L = \lim_{t\to\infty} F(t)$$

exists. For all $\epsilon \in \mathbb{R}^+$ there thus exists a large-enough real number $b > a$ such that

$$|F(t) - L| < \frac{\epsilon}{2} \text{ for all } t > b.$$

For any real number $u, v \in \mathbb{R}$ for which $b < u < v$ we then make the calculation

$$\begin{aligned} \left| \int_u^v f(x)\,\mathrm{d}x \right| &= |F(v) - F(u)| = |F(V) - L + L - F(u)| \\ &\leq |F(v) - L| + |L - F(u)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

($\Leftarrow$) Assume the proposition's convergence criterion holds for the function f. For arbitrary $\epsilon \in \mathbb{R}^+$ there thus exists a sufficiently large real number $b > a$ such that

$$|F(v) - F(u)| = \left| \int_u^v f(x)\,\mathrm{d}x \right| < \frac{\epsilon}{2}$$

for any two real numbers u and v for which $b < u < v$. The result $|F(v) - F(u)| < \epsilon$ means that the sequence $\big(F(a+n)\big)_{n\in\mathbb{N}}$ is a Cauchy sequence, and is thus convergent by Proposition 4.31. Let Q denote the limit of this sequence and choose a large-enough natural number $n \in \mathbb{N}$ such that $a + n > b$ and

$$|F(a+n) - Q| < \frac{\epsilon}{2}.$$

For every real number t for which $t > b$, we can make the bound

$$\begin{aligned} |F(t) - Q| &= |F(t) - F(a+n) + F(a+n) - Q| \\ &\leq |F(t) - F(a+n)| + |F(a+n) - Q| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

The result $|F(t) - Q| < \epsilon$ for arbitrary $\epsilon \in \mathbb{R}^+$ means that

$$Q = \lim_{t\to\infty} F(t),$$

and so the improper integral converges. □

Proposition 9.44: Convergence criterion for improper integrals of power-function denominators on unbounded domains

Let $a \in \mathbb{R}^+$ be any positive real number, let $g\colon [a,\infty) \to \mathbb{R}$ be a bounded, continuous function, and let $\alpha \in \mathbb{R}$ be a continuous constant. In this case the improper integral

$$\int_a^\infty \frac{g(x)}{x^\alpha}\,\mathrm{d}x$$

(i) converges if $\alpha > 1$ and

(ii) diverges if $\alpha \leq 1$ and $\inf\{|g(x)|\,;\, x \in [c,\infty)\} > 0$ for some $c \in [a,\infty)$.

Proof. We prove the two parts in turn:

(i) For any real numbers $u, v \in \mathbb{R}$ for which $a < u < v$ we first make the calculation

$$\begin{aligned}\left|\int_u^v \frac{g(x)}{x^\alpha}\,\mathrm{d}x\right| &\leq \sup(|g|)\int_u^v \frac{\mathrm{d}x}{x^\alpha} \\ &= \sup(|g|)\frac{x^{1-\alpha}}{1-\alpha}\bigg|_u^v, \\ &= \frac{\sup(|g|)}{1-\alpha}\left(v^{1-\alpha} - u^{1-\alpha}\right) \\ &\overset{\text{(a)}}{\leq} \frac{\sup(|g|)}{\alpha-1}u^{1-\alpha}.\end{aligned}$$

where (a) follows from the fact that u and v are both positive. We note that

$$\lim_{u\to\infty} u^{1-\alpha} = 0 \quad (\text{if } \alpha > 1),$$

then complete the proof with the calculation

$$\lim_{u\to\infty}\left|\int_u^v \frac{g(x)}{x^\alpha}\,\mathrm{d}x\right| \leq \frac{\sup(|g|)}{\alpha-1}\cdot\lim_{u\to\infty} u^{1-\alpha} = 0,$$

which, by Proposition 9.43, means that the improper integral of g/x^α converges in the limits from a to ∞.

(ii) We will assume without loss of generality that $a = c$ and $a, c > 1$, and define

$$W = \inf\{|g(x)|\,;\, x \in [a,\infty)\}.$$

Because $W > 0$ and the function g is continuous, either

$$g([a,\infty)) \subset [W,\infty) \quad \text{or} \quad g([a,\infty)) \subset (-\infty,-W].$$

For any real number $b > a$ we can then make the bound

$$\int_a^b \frac{g(x)}{x^\alpha}\,\mathrm{d}x \overset{\text{(a)}}{\geq} \int_a^b \frac{g(x)}{x}\,\mathrm{d}x \geq W\int_a^b \frac{\mathrm{d}x}{x} = W(\ln b - \ln a),$$

where (a) holds because $\alpha \leq 1$ and $a > 0$. We then note that

$$\lim_{b\to\infty}(\ln b - \ln a) = \infty,$$

and so the associated improper integral diverges in the limits from a to ∞. The proof is analogous when $g([a,\infty)) \subset (-\infty,-W]$. □

9.6 The integral test for series convergence

We conclude this chapter with the *integral test*, in which definite integration is used to test the convergence of series of real numbers as studied in Chapter 5.

Proposition 9.45: The integral test

Let $f\colon [1,\infty) \to \mathbb{R}$ be a decreasing, continuous function for which

$$f(x) \geq 0 \text{ for all } x \in [1,\infty).$$

In this case the improper integral

$$\int_1^\infty f(x)\,\mathrm{d}x$$

converges if, and only if, the series of real numbers

$$\sum \big[f(k)\big]_{k=1}^{\infty}$$

also converges.

Proof. We will first prove an auxiliary inequality, the prove the proposition itself. Choose arbitrary $n \in \mathbb{N}$ and define the set

$$\mathcal{D}_n = \{1, 2, \ldots, n+1\} \subset \mathbb{R}.$$

The set $\mathcal{D}_n$ is a partition of the interval $[1, n+1]$ into n subintervals, and we will write it in the form $\mathcal{D}_n = \{k+1\}_{k=0}^{n}$. Because the function f is decreasing, its upper and lower Darboux sums on the partition $\mathcal{D}_n$ are

$$\mathrm{s}(f,\mathcal{D}_n) = \sum_{k=1}^{n} f(k+1) \quad \text{and} \quad \mathrm{S}(f,\mathcal{D}_n) = \sum_{k=1}^{n} f(k+1),$$

which we then combine with Equation 9.1 to get the desired inequality

$$\sum_{k=1}^{n} f(k+1) \leq \int_1^{n+1} f(x)\,\mathrm{d}x \leq \sum_{k=1}^{n} f(k). \tag{9.13}$$

We will now prove the proposition itself.

($\Rightarrow$) Assume that the improper integral $\int_1^\infty f(x)\,\mathrm{d}x$ converges to the number $I \in \mathbb{R}$. For all $n \in \mathbb{N}$, Equation 9.13 then implies the inequality

$$\sum_{j=1}^{n+1} f(j) = f(1) + \sum_{k=1}^{n} f(k+1) \leq f(1) + \int_1^{n+1} f(x)\,\mathrm{d}x \leq f(1) + I. \tag{9.14}$$

Equation 9.14 means that

$$\sum_{k=1}^{n} f(k+1) \leq I \text{ for all } n \in \mathbb{N},$$

and so the increasing sequence of partial sums of the series of nonnegative numbers $\sum \big[f(k)\big]_{k=1}^{\infty}$ is bounded above by I, and is thus convergent by the monotone convergence theorem (Proposition 4.15).

($\Leftarrow$) assume the series of real numbers $\sum \big[f(k)\big]_{k=1}^{\infty}$ converges to the sum $A \in \mathbb{R}$. Equation 9.13 implies that

$$\int_1^{n+1} f(x)\,\mathrm{d}x \leq \sum_{k=1}^{n} f(k) \leq A,$$

and so the increasing function $I\colon [1,\infty) \to \mathbb{R}$ given by

$$I(b) = \int_1^b f(x)\,\mathrm{d}x$$

is bounded above by the number A. Because I is bounded above it must have a supremum; because I is increasing this supremum must also be I's limit as b approaches infinity, and so the improper integral

$$\int_1^{\infty} f(x)\,\mathrm{d}x$$

converges. □

A classic application of the integral test is proving the divergence of the harmonic series, which we show in the following example.

Example 9.46: Using the integral test

Consider the series of real numbers

$$\sum \left(\frac{1}{k^{\alpha}}\right)_{k=1}^{\infty}, \quad \alpha \in \mathbb{R}, \tag{9.15}$$

and let the function $f\colon \mathbb{R}^{+} \to \mathbb{R}$ be given by

$$f(x) = \frac{1}{x^{\alpha}}.$$

By the integral test, the series in Equation 9.15 converges if, and only if, the improper integral of the function f on the interval from 1 to ∞ converges, and by Proposition 9.44 this improper integral converges if $\alpha > 1$, and diverges otherwise.

In particular, letting $\alpha = 1$, this result means that the harmonic series

$$\sum \left(\frac{1}{k}\right)_{k=1}^{\infty}$$

diverges, as we have already shown in Example 5.7.

10 Taylor series and power series

In this section we will explore some elementary properties of power series, focusing in particular on the construction of Taylor series.

10.1 Taylor's formula

Loosely, Taylor's formula approximates a differentiable function around a point in its domain using a polynomial weighted by the function's derivatives at that point.

As a simple motivating example using only the first derivative, let $f\colon U \to \mathbb{R}$ be a differentiable function defined on the open subset $U \subset \mathbb{R}$ and define the point $a \in U$. As mentioned in Remark 7.4, the tangent line to the graph of the function f at the point a is the line that best approximates f's graph at the point $\big(a, f(a)\big)$. We may thus use f's derivative to approximate f's function values around a using the *linear*, or *first-order*, approximation

$$f(x) \approx f(a) + f'(a)(x-a) \qquad \text{(if } x \text{ is "near" } a\text{)}.$$

We will have more to say on the validity of the approximation and how to be sure x is "near enough" to a in the coming pages.

Example 10.1: Computing a cube root with a linear approximation

Let us use a linear approximation to compute the cube root of the number 1003. We first note that—intentionally—the number 1003 is relatively close to 1000, which has a cube root of 10.

To use a first-order approximation to compute $\sqrt[3]{1003}$, let the function $f\colon \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \sqrt[3]{x},$$

and so its derivative is

$$f'(x) = \frac{1}{3}x^{-2/3}.$$

We will then form a linear approximation of f about the point 1000. The function value and derivative value at 1000 are

$$f(1000) = 10 \quad \text{and} \quad f'(1000) = \frac{1}{300},$$

and so the linear approximation of f about a reads

$$f(x) \approx f(1000) + f'(1000)(x-1000) = 10 + \frac{(x-1000)}{300}.$$

We then this approximation to compute the estimate

$$\sqrt[3]{1003} \approx 10 + \frac{(1003-1000)}{300} = 10.01.$$

The true value of $\sqrt[3]{1003}$ to six decimal places turns out to be 10.009990, so the approximation appears to be quite "good".

Importantly, the approximation method used in the Example 10.1 offered no self-contained estimate of the approximation's error—we had to resort to external means (in this case, a

simple scientific calculator)—to compute the true value of $\sqrt[3]{1000}$.

Quite, generally, an approximation method is of little use without a self-contained estimate of its error. Fortunately, Taylor's approximation formula *does* include an estimate of its error. We will build up to this formula in the coming pages, and begin with an auxiliary result showing how to express a real-valued polynomial in terms of its derivatives.

Proposition 10.2: Expressing a polynomial in terms of its derivatives

Let $P\colon \mathbb{R} \to \mathbb{R}$ be a scalar polynomial of degree at most $n \in \mathbb{N} \cup \{0\}$ and let $a \in \mathbb{R}$ be a real constant. In this case, for all $x \in \mathbb{R}$, the polynomial P can be expressed as

$$P(x) = P(a) + P'(a)(x-a) + \frac{P''(a)}{2}(x-a)^2 + \cdots + \frac{P^{(n)}(a)}{n!}(x-a)^n.$$

Proof. We first write the polynomial P in the form

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_0, a_1, \ldots, a_n \in \mathbb{R}$ are real constants. We then define the new variable $h = x - a$ (and so $x = h + a$) and expand the polynomial P in powers of h to get

$$P(a+h) = a_n(a+h)^n + a_{n-1}(a+h)^{n-1} + \cdots + a_1(a+h) + a_0.$$

This expression can alternatively be written in the general form

$$P(a+h) = b_n h^n + b_{n-1}h^{n-1} + \cdots + b_1 h + b_0,$$

where the new coefficients $b_0, b_1, \ldots, b_n$ depend on both the point a and the original coefficients of the polynomial P. We could compute the explicit form of the $\{b_k\}$ coefficients using the binomial formula, but we will instead compute them using differentiation. Namely, we repeatedly differentiate the expression for $P(a+h)$ with respect to h to get

$$\begin{aligned}
\big(P(a+h)\big)' &= nb_n h^{n-1} + (n-1)b_{n-1}h^{n-2} + \cdots 2b_2 h + b_1 \\
\big(P(a+h)\big)'' &= n(n-1)b_n h^{n-2} + (n-1)(n-2)b_{n-1}h^{n-2} + \cdots + 2b_2 \\
&\vdots \\
\big(P(a+h)\big)^{(n)} &= n(n-1)\cdots\big[n-(n-1)\big]b_n.
\end{aligned}$$

We then simply evaluate each derivative at $h = 0$ to get

$$P^{(k)}(a) = k!b_k \implies b_k = \frac{P^{(k)}(a)}{k!}, \quad k = 0, 1, \ldots, n.$$

Having solved for the coefficients $\{b_k\}$, we can immediately write $P(x)$ in the form

$$\begin{aligned}
P(x) = P(a+h) &= b_0 + b_1 h + \cdots + b_{n-1}h^{n-1} + b_n h^n \\
&= P(a) + P'(a)(x-a) + \cdots + \frac{P^{(n-1)}(a)}{(n-1)!} + \frac{P^{(n)}(a)}{n!}(x-a)^n,
\end{aligned}$$

which completes the proof by construction. □

Proposition 10.2 implies that a scalar polynomial P of degree at most $n \in \mathbb{N}$ is uniquely determined by the set of numbers

$$\{P(a), P'(a), \ldots, P^{(n)}(a)\}$$

for a given point $a \in \mathbb{R}$. Thus, for any smooth function $f\colon U \to \mathbb{R}$, Proposition 10.2 means it is possible to find a unique polynomial whose derivatives up to order n at a chosen point a in the function's domain agree exactly with f's derivatives at a. More precisely:

Corollary 10.3: Polynomial approximation of a function at a point

Define the integer $n \in \mathbb{N} \cup \{0\}$, let $U \subset \mathbb{R}$ be an open subset of $\mathbb{R}$, let $f\colon U \to \mathbb{R}$ be an n-times differentiable function, and define the point $a \in U$. In this case there exists exactly one real polynomial $P\colon \mathbb{R} \to \mathbb{R}$ of degree at most n for which

$$P^{(k)}(a) = f^{(k)}(a), \quad k = 0, 1, \ldots, n.$$

This polynomial is given by the formula

$$P(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

Proof. The corollary follows directly from Proposition 10.2. □

The polynomial P in Corollary 10.3 is a special case of a so-called Taylor polynomial, which we define more precisely below.

Definition 10.4: Taylor polynomial and Taylor expansion

Define the integer $n \in \mathbb{N} \cup \{0\}$, let $U \subset \mathbb{R}$ be an open subset of $\mathbb{R}$, let $f\colon U \to \mathbb{R}$ be an n-times differentiable function, and define the point $a \in U$. In this case the *n-th order Taylor polynomial* of the function f *expanded about the point* a is defined as

$$T_n f(x; a) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

This polynomial is sometimes also called the n-th order *Taylor expansion* of the function f about the point a, and is used to approximate f's values near a.

We will now compute the Taylor polynomials of some common functions.

Example 10.5: Taylor polynomials of common functions

(1) (*A polynomial*) Let use compute the Taylor expansion of the polynomial

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

about the point 2, i.e. in powers of $(x-2)$. Currently, the polynomial is expanded in powers of the variable x about the point 0; to switch to powers of

$(x-2)$, we first compute the three derivatives

$$\begin{aligned} f'(x) &= 3x^2 - 4x - 4 & f'(2) &= 0 \\ f''(x) &= 6x - 4 & f''(2) &= 8 \\ f'''(x) &= 6 & f'''(2) &= 6. \end{aligned}$$

We then note that $f(2) = 0$ and apply the polynomial expansion formula in Proposition 10.2 to write $f(x)$ in the form

$$\begin{aligned} f(x) &= 0 + 0(x-2) + \frac{8}{2}(x-2)^2 + \frac{6}{3!}(x-2)^3 \\ &= 4(x-2)^2 + (x-2)^3. \end{aligned}$$

The first few Taylor polynomial of the function f are

$$\begin{aligned} T_0 f(x;2) &= 0 \\ T_1 f(x;2) &= 0 \\ T_2 f(x;2) &= 4(x-2)^2 \\ T_3 f(x;2) &= 4(x-2)^2 + (x-2)^3. \end{aligned}$$

Because f is a polynomial of degree 3, its higher-order Taylor polynomials are equal to the function f itself, i.e.

$$f(x) = T_n f(x;2) \text{ for all natural numbers } n \geq 3.$$

(2) (*Exponential function*) Let us compute the Taylor expansion of the function

$$f\colon x \mapsto \mathrm{e}^x$$

about the point $a = 0$. We first note the simple relationship $f^{(k)}(x) = \mathrm{e}^x$, so

$$f^{(k)}(0) = 1 \text{ for all integers } k \geq 0.$$

For all $n \in \mathbb{N} \cup \{0\}$, the Taylor expansion of the exponential function is thus

$$T_n f(x;0) = \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{x^n}{n!}.$$

(3) (*Sine function*) Let us compute Taylor polynomial of the sine function

$$f\colon x \mapsto \sin x$$

about the point $a = 0$. The sine function's first four derivatives are

$$\begin{aligned} f'(x) &= \cos x \\ f''(x) &= -\sin x \\ f'''(x) &= -\cos x \\ f''''(x) &= \sin x. \end{aligned}$$

Notice that $f'''' = f$. In fact, f's higher derivatives repeat with period 4, and so for all integers $p \geq 0$ the sine function's derivatives at $x = 0$ are

$$\begin{aligned} f^{(4p)}(0) &= 0 \\ f^{(4p+1)}(0) &= 1 \\ f^{(4p+2)}(0) &= 0 \\ f^{(4p+3)}(0) &= -1. \end{aligned}$$

For every integer $m \geq 0$, the m-th order Taylor polynomial of the sine function expanded about the point 0 is thus

$$\begin{aligned} T_{2m+1}f(x;0) &= 0 + \frac{1}{1}x + \frac{0}{2}x^2 - \frac{1}{3!}x^3 + \frac{0}{4!}x^4 + \cdots + \frac{(-1)^m x^{2m+1}}{(2m+1)!} \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + \frac{(-1)^m x^{2m+1}}{(2m+1)!} \\ &= \sum_{k=0}^{m} \frac{(-1)^k x^{2k+1}}{(2k+1)!}. \end{aligned}$$

Note that this Taylor expansion contains only terms with odd powers, which is in fact a consequence of the sine function being odd.

(4) (*Cosine function*) We conclude this example with the Taylor polynomial of the cosine function $f\colon x \mapsto \cos x$ about the point $a = 0$. Following an analogous procedure to that used above for the sine function, for all integers $p \geq 0$ the cosine function's derivatives at $x = 0$ are

$$\begin{aligned} f^{(4p)}(0) &= 1 \\ f^{(4p+1)}(0) &= 0 \\ f^{(4p+2)}(0) &= -1 \\ f^{(4p+3)}(0) &= 0 \end{aligned}$$

For every integer $m \geq 0$, the m-th order Taylor polynomial of the cosine function expanded about the point 0 is thus

$$\begin{aligned} T_{2m}f(x;0) &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + \frac{(-1)^m x^{2m}}{(2m)!} \\ &= \sum_{k=0}^{m} \frac{(-1)^k x^{2k}}{(2k)!}. \end{aligned}$$

This Taylor expansion contains only terms with even powers, which is in fact a consequence of the cosine function being even.

Definition 10.6: Taylor approximation and remainder

Define the integer $n \in \mathbb{N} \cup \{0\}$, let $f\colon U\colon \mathbb{R}$ be an arbitrary $(n+1)$-times differentiable function defined on the open subset $U \subset \mathbb{R}$, and let $a \in U$ be a constant. For each

$x \in U \setminus \{a\}$ the *n-th order Taylor approximation* of the function f at the point a is

$$f(x) \approx T_n f(x;a) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

The *error* or *remainder* of the n-th order approximation, denoted by R_n, is the difference between the true function value and the Taylor approximation T_n, i.e.

$$R_n(x) = f(x) - T_n f(x;a).$$

In terms of this remainder, for every $x \in U$ the function value $f(x)$ can be written exactly in the form

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x).$$

We will dedicate the next few pages to deriving (and thus constructively proving) Taylor's formula, which gives a bound on the remainder in Taylor approximations.

Topic 10.1: Deriving Taylor's formula

Define the integer $n \in \mathbb{N} \cup \{0\}$, let $f : U : \mathbb{R}$ be an arbitrary $(n+1)$-times differentiable function defined on the open subset $U \subset \mathbb{R}$, and let $a \in U$ be a constant. Fix an arbitrary point $x \in U \setminus \{a\}$ and natural number $p \in \mathbb{N}$ and define the function $F : U \to \mathbb{R}$ according to

$$F(t) = f(x) - f(t) - f'(t)(x-t) - \cdots - \frac{f^{(n)}(t)(x-t)^n}{n!} - R_n(x)\left(\frac{x-t}{x-a}\right)^p.$$

Assume that the chosen point x is close enough to the expansion point a that the entire interval between x and a lies inside the set U, so that the function F of the variable t is defined at all points $t \in [x, a]$.

We first note that, by construction, the function F is continuous on the closed interval $[a, x]$ and differentiable on the open interval (a, x). At the interval endpoints, the function F obeys

$$\begin{aligned} F(a) &= f(x) - f(a) - f'(a)(x-a) - \cdots - \frac{f^{(n)}(a)(x-a)^n}{n!} - R_n(x)\left(\frac{x-a}{x-a}\right)^p \\ &= f(x) - f(x) - 0 - \cdots - 0 - 0 = 0, \\ F(x) &= f(x) - f(x) - f'(x)(x-x) - \cdots - \frac{f^{(n)}(x)(x-x)^n}{n!} - R_n(x)\left(\frac{x-x}{x-a}\right)^p \\ &= f(x) - f(x) - 0 - \cdots - 0 - 0 = 0. \end{aligned}$$

Thus $F(a) = F(x) = 0$, and so F meets the criteria for Rolle's theorem. By Rolle's theorem there then exists a point ξ in the open interval (a, x) for which $F'(\xi) = 0$.

Before computing the derivative F', we first use the product rule to make the auxiliary calculation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{f^{(k)}(t)(x-t)^k}{k!}\right] = \frac{f^{(k+1)}(t)(x-t)^k}{k!} - \frac{f^{(k)}(t)(x-t)^{k-1}}{(k-1)!}. \tag{10.1}$$

We then differentiate F with respect to t to get

$$\begin{aligned}
F'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(t)(x-t)^k}{k!} - R_n(x)\left(\frac{x-t}{x-a}\right)^p\right] \\
&\overset{\text{(a)}}{=} 0 - \sum_{k=0}^{n} \frac{f^{(k+1)}(t)(x-t)^k}{k!} + \sum_{k=1}^{n} \frac{f^{(k)}(t)(x-t)^{k-1}}{(k-1)!} + \frac{pR_n(x)}{x-a}\left(\frac{x-t}{x-a}\right)^{p-1} \\
&\overset{\text{(b)}}{=} -\sum_{k=0}^{n} \frac{f^{(k+1)}(t)(x-t)^k}{k!} + \sum_{l=0}^{n-1} \frac{f^{(l+1)}(t)(x-t)^l}{l!} + \frac{pR_n(x)}{x-a}\left(\frac{x-t}{x-a}\right)^{p-1} \\
&= -\frac{f^{n+1}(t)(x-t)^n}{n!} + \frac{pR_n(x)}{x-a}\left(\frac{x-t}{x-a}\right)^{p-1},
\end{aligned}$$

where (a) follows from Equation 10.1 and (b) from introducing the new summation index $l = k - 1$. We then evaluate the above equality at the Rolle's theorem point $\xi \in (a, x)$ at which $F'(\xi) = 0$ to get

$$F'(\xi) = 0 \implies \frac{f^{n+1}(\xi)(x-\xi)^n}{n!} = \frac{pR_n(x)}{x-a}\left(\frac{x-\xi}{x-a}\right)^{p-1},$$

then rearrange and solve for the remainder to get

$$R_n(x) = \frac{(x-\xi)^{n-p+1} f^{(n+1)}(\xi)}{p \cdot n!}(x-a)^p.$$

Different choices of the number $p \in \mathbb{N}$ lead to different expressions for the remainder, but in any case the remainder depends on the $(n+1)$-th derivative of the function f. Resultantly, an upper bound on the magnitude of the function $f^{(n+1)}$ will also place an upper bound on the magnitude of the remainder $R_n(x)$; finding such an upper bound is the motivation for Taylor's formula, which we are close to deriving.

Next, we will parameterize the line between the points x and a in terms of the parameter $\theta \in (0, 1)$, and write the number $\xi \in (a, x)$ in the form

$$\xi = a + \theta(x-a), \quad \theta \in (0,1).$$

The parameter θ indicates where the point ξ lies on the line between a and x; note that $\theta = 0$ corresponds to $\xi = a$ and $\theta = 1$ to $\xi = x$. It then follows that

$$x - \xi = x - a - \theta(x-a) = (x-a)(1-\theta),$$

in terms of which the expression for the remainder $R_n(x)$ reads

$$R_n(x) = \frac{(1-\theta)^{n-p+1} f^{(n+1)}\big(a + \theta(x-a)\big)}{p \cdot n!} \cdot (x-a)^{n+1}.$$

Choosing $p = n + 1$ in this expression gives the *Lagrange form* of the remainder, which reads

$$R_n(x) = \frac{f^{(n+1)}\big(a + \theta(x-a)\big)}{(n+1)!} \cdot (x-a)^{n+1}$$

for some value of the parameter $\theta \in (0,1)$. Meanwhile, choosing $p = 1$ gives the *Cauchy form* of the remainder, which reads

$$R_n(x) = \frac{(1-\theta)^n f^{(n+1)}\big(a+\theta(x-a)\big)}{n!}(x-a)^{n+1},$$

again for some value of the parameter $\theta \in (0,1)$.

We stress that, for a given function $f\colon U \to \mathbb{R}$, order $n \in \mathbb{N}$, and expansion point $a \in U$, the value of θ will differ for different choices of the number $p \in \mathbb{N}$; similarly, the value of θ also depends on the value of x at which the remainder $R_n(x)$ is evaluated. The above derivation proves by construction the following forms of Taylor's formula:

Theorem 10.7: Taylor's formula

Define the integer $n \in \mathbb{N} \cup \{0\}$, let $U \subset \mathbb{R}$ be an open interval, let $f\colon U \to \mathbb{R}$ be an arbitrary $(n+1)$-times differentiable function, and let $a \in U$ be a constant. In this case for every $x \in U \setminus \{a\}$ there exists $\theta \in (0,1)$ for which

(i) (*Lagrange form of the remainder*)

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(n+1)}\big(a+\theta(x-a)\big)}{(n+1)!}(x-a)^{n+1}$$

(ii) (*Cauchy form of the remainder*)

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{(1-\theta)^n f^{(n+1)}\big(a+\theta(x-a)\big)}{n!}(x-a)^{n+1}.$$

10.2 Taylor series

In both the Lagrange and Cauchy forms of the Taylor series remainder in Theorem 10.7, the bound on the approximation's error decreases rapidly with increasing n, i.e. as more and more terms are added to the approximating Taylor polynomial. This observation raises the question of what happens in the limit as n approaches infinity. As we will see in this section, under certain conditions the resulting *Taylor series* converges *exactly* to the function being approximated, potentially over the function's entire domain.

As we have seen in Definition 10.6, for a given $n \in \mathbb{N} \cup \{0\}$ and sufficiently differentiable function $f\colon U \to \mathbb{R}$, the function value $f(x)$ equals the sum of the n-th order Taylor approximation $T_n f(x;a)$ and the associated remainder $R_n(x)$, i.e.

$$f(x) = T_n f(x;a) + R_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + R_n(x).$$

If the function f is not a polynomial, then in general the remainder R_n is nonzero. It is possible, however, that for a given $x \in U$ the remainder value $R_n(x)$ converges to zero as n approaches infinity. If $\lim_{n\to\infty} R_n(x) = 0$, then

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

The real number on the right hand side is the sum of f's Taylor series, which we now define.

Definition 10.8: Taylor series

Let $f\colon U \to \mathbb{R}$ be a smooth function defined on the open subset $U \subset \mathbb{R}$, and define the point $a \in U$. The *Taylor series* of the function f *expanded about the point* a is the infinite series

$$\sum \left[\frac{f^{(k)}(a)}{k!}(x-a)^k \right]_{k=0}^{\infty}.$$

Note that a function's Taylor series expanded about a given point need not converge to a function's value at that given point. The series can converge to some other value or not converge at all—a Taylor series converges to the corresponding function value *only* when the remainder R_n vanishes at that point.

We also stress the subtle but important distinction between the Taylor series

$$\sum \left[\frac{f^{(k)}(a)}{k!}(x-a)^k \right]_{k=0}^{\infty},$$

which is an infinite series, and the *sum* of the series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k,$$

which is a real number (assuming the series converges).

Example 10.9: Taylor series of the exponential function

Let us compute the Taylor series of the exponential function about the point $a = 0$. From Example 10.5, the exponential function's Taylor polynomials about $a = 0$ are

$$T_n f(x;0) = \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{x^n}{n!},$$

and so the corresponding Taylor series is

$$\sum \left(\frac{x_k}{k!} \right)_{k=0}^{\infty}.$$

We now aim to find for which points $x \in \mathbb{R}$ the exponential Taylor series converges to the exponential function. First, noting that $f^{(n+1)}(x) = \mathrm{e}^x$, the Lagrange form of the series's remainder is

$$R_n = \frac{\mathrm{e}^{\theta x}}{(n+1)!}x^{n+1} \text{ for some } \theta \in (0,1).$$

Because $\theta \in (0,1)$, we can bound the expression $\mathrm{e}^{\theta x}$ above by the number e^x. We will now show that the sequence $(|x|^n/n!)_n$ converges to zero for all $x \in \mathbb{R}$. First fix $N \in \mathbb{N}$ for which $N \geq |x|$. For every natural number $n \geq N$ it then holds that

$$\frac{|x|^{n+1}}{(n+1)!} = \frac{|x|^n}{n!} \cdot \frac{|x|}{n+1} < \frac{|x|^n}{n!} \cdot \frac{|x|}{N} \overset{\text{(a)}}{<} \frac{|x|^n}{n!},$$

where (a) holds because $N \geq |x|$. The sequence of nonnegative (and thus bounded below by zero) real numbers $(|x|^n/n!)_n$ is thus strictly decreasing from the N-th term forward, and thus converges by Proposition 4.15. More so, we can directly compute the sequence's limit with the calculation

$$\begin{aligned}\lim_{n\to\infty} \frac{|x|^n}{n!} &\overset{\text{(a)}}{=} \lim_{n\to\infty} \frac{|x|^{n+1}}{(n+1)!} = \lim_{n\to\infty} \frac{|x|^n}{n!} \cdot \lim_{n\to\infty} \frac{|x|}{n+1} \\ &= \lim_{n\to\infty} \frac{|x|^n}{n!} \cdot 0 = 0,\end{aligned}$$

where (a) holds because offsetting the index n by one does not changes the limit's value. The sequence $(|x|^n/n!)_n$ converges to zero, and so

$$\lim_{n\to\infty} R_n(x) = 0 \text{ for all } x \in \mathbb{R},$$

which in turn implies that the exponential Taylor series converges to the exponential for all $x \in \mathbb{R}$, i.e.

$$\mathrm{e}^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

In particular, setting $x = 1$ gives the interesting equality.

$$\mathrm{e} = \sum_{k=0}^{\infty} \frac{1}{k!} = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots.$$

Example 10.10: Taylor series of the sine function

Let us compute the Taylor series of the sine function expanded about the point $a = 0$. From Example 10.5, the sine function's Taylor polynomials about $a = 0$ are

$$\begin{aligned}T_{2n+1}f(x;0) &= \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + \frac{(-1)^n x^{2n+1}}{(2n+1)!},\end{aligned}$$

and the corresponding Taylor series is

$$\sum \left[\frac{(-1)^k x^{2k+1}}{(2k+1)!}\right]_{k=0}^{\infty}.$$

We now aim to find the associated remainder. We first note that the derivatives of the sine function are—up to a sign—either sine or cosine functions, and are thus bounded by 1 in absolute value. For all $n \in \mathbb{N} \cup \{0\}$ we can thus make the bound

$$|R_n(x)| = \left|\frac{f^{(n+1)}(\theta x)}{(n+1)!} x^{n+1}\right| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

We showed in Example 10.9 that the sequence $(|x|^n/n!)_n$ converges to zero, and so

$$\lim_{n\to\infty} R_n(x) = 0.$$

The sine series thus converges to the sine function for all $x \in \mathbb{R}$, i.e.

$$\sin x = \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \cdots .$$

Example 10.11: Taylor series of the cosine function

From Example 10.5, the Taylor polynomials of the cosine function are

$$\begin{aligned} T_{2n} f(x; 0) &= \sum_{k=0}^{n} \frac{(-1)^k x^{2k}}{(2k)!} \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + \frac{(-1)^n x^{2n}}{(2n)!}. \end{aligned}$$

The corresponding Taylor series is

$$\sum \left[\frac{(-1)^k x^{2k}}{(2k)!} \right]_{k=0}^{\infty}.$$

Following an analogous procedure to that used for the remainder of the sine series, the remainder of the cosine series also converges to zero:

$$\lim_{n \to \infty} R_n(x) = 0.$$

The cosine series thus converges to the cosine function for every $x \in \mathbb{R}$, i.e.

$$\cos x = \sum_{k=0}^{n} \frac{(-1)^k x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \pm \cdots .$$

Example 10.12: Taylor series of the logarithmic function

Let us compute the Taylor series of the logarithmic function

$$f \colon x \mapsto \ln(x+1)$$

about the point $a = 0$. Note that the function f has a vertical asymptote at the point $x = -1$, so we should not expect that the function's Taylor series will converge to the function for all $x \in \mathbb{R}$.

We first compute the function's first few derivatives:

$$\begin{aligned} f'(x) &= \frac{1}{x+1} = (x+1)^{-1} \\ f''(x) &= -1!(x+1)^{-2} \\ f'''(x) &= 2!(x+1)^{-3} \\ f''''(x) &= -3!(x+1)^{-4}. \end{aligned}$$

We can then deduce that for every $k \in \mathbb{N}$ the function f's k-th derivative is

$$f^{(k)}(x) = (-1)^{k-1}(k-1)!(x+1)^{-k}.$$

The function value at the expansion point is $f(0) = 0$ and for all $k \in \mathbb{N}$ the function's k-th order derivative value at $a = 0$ is

$$f^{(k)}(0) = (-1)^{k-1}(k-1)!.$$

The function f's Taylor series about $a = 0$ is thus

$$\sum \left[\frac{(-1)^{k-1}x^k}{k} \right]_{k=1}^{\infty}.$$

For all $n \in \mathbb{N} \cup \{0\}$, the Lagrange form of the series's remainder is

$$\begin{aligned} R_n(x) &= \frac{f^{(n+1)}(\theta x)}{(n+1)!} = \frac{(-1)^n n!(1+\theta x)^{-n-1}}{(n+1)!} \\ &= \frac{(-1)^n}{n+1}\left(\frac{x}{1+\theta x}\right)^{n+1}, \quad \theta \in (0,1). \end{aligned}$$

If $x \in [0,1]$ then $1 + \theta x \geq 1$, which leads to the bound

$$|R_n(x)| = \left| \frac{(-1)^n}{n+1}\left(\frac{x}{1+\theta x}\right)^{n+1} \right| \leq \frac{x^{n+1}}{n+1}.$$

The sequence $(x^n/n)_n$ converges to 0 for $x \in [0,1]$, and so

$$\lim_{n\to\infty} R_n(x) = 0, \quad x \in [0,1].$$

When $x \in (-1,0)$, it is more useful to consider the series's Cauchy remainder, which for all $n \in \mathbb{N} \cup \{0\}$ reads

$$\begin{aligned} R_n(x) &= \frac{(1-\theta)^n f^{(n+1)}(\theta x)}{n!}x^{n+1} \\ &= \frac{(1-\theta)^n(-1)^n n!(1+\theta x)^{-n-1}}{n!}x^{n+1} \\ &= \frac{(-1)^n x}{1+\theta x} \cdot \left[\frac{(1-\theta)x}{1_\theta x}\right]^n, \quad \theta \in (0,1). \end{aligned}$$

Assuming $x \in (-1,0)$, it follows that $1 + \theta x > 1 - \theta$ and so

$$\left|\frac{(1-\theta)x}{1+\theta x}\right| < \left|\frac{(1-\theta)x}{1-\theta}\right| = |x|,$$

which in turn leads to the bound

$$|R_n(x)| = \left| \frac{(-1)^n x}{1+\theta x} \cdot \left[\frac{(1-\theta)x}{1+\theta x}\right]^n \right| < \frac{|x|}{1-|x|} \cdot |x|^n.$$

The right-hand side is geometric sequence with initial term $a = |x|/(1-|x|)$ and ratio $q = |x| \in (0,1)$, and thus converges to zero by Example 4.20, so

$$\lim_{n\to\infty} R_n(x) = 0, \quad x \in (-1,0).$$

We have shown that the logarithmic series converges to the logarithmic function for all $x \in (-1, 1]$ and that

$$\ln(x+1) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1} x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots .$$

When evaluated at $x = -1$, the logarithmic Taylor series reduces to the negative of the divergent harmonic series, and thus diverges. Similarly, the logarithmic Taylor series diverges for $|x| > 1$, since in this case its terms increase without bound in absolute value with increasing n.

Remark 10.13: Computing logarithms of larger numbers

With the help of a small trick, one can use the logarithmic Taylor series to compute the logarithms of numbers larger than 1 in absolute value. We first recall from Example 10.12 that for arbitrary $x \in (-1, 1)$ it holds that

$$\begin{aligned} \ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots , \\ \ln(1-x) &= -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots , \end{aligned}$$

and that both series converge. We then subtract the two series to get

$$\begin{aligned} \ln(1+x) - \ln(1-x) &= \ln\left(\frac{1+x}{1-x}\right) \\ &= 2x + \frac{2x^3}{3} + \frac{2x^5}{5} + \frac{2x^7}{7} + \cdots \\ &= 2x\left(1 + \frac{x^2}{3} + \frac{x^4}{5} + \frac{x^6}{7} + \cdots\right). \end{aligned}$$

The result, i.e.

$$\ln\left(\frac{1+x}{1-x}\right) = 2x\left(1 + \frac{x^2}{3} + \frac{x^4}{5} + \frac{x^6}{7} + \cdots\right),$$

makes it possible to compute the logarithm of an arbitrary positive number, since the function $x \mapsto (1+x)/(1-x)$ takes on every value in $\mathbb{R}^+$ for $x \in (-1, 1)$.

Example 10.14: The binomial series

We will derive the binomial series from the Taylor series of the power function.

Let $\alpha \in \mathbb{R}$ be an arbitrary real number and let f be the scalar function $f\colon x \mapsto (1+x)^\alpha$. We will expand the power function f into a Taylor series about the point $a = 0$. We

first compute the derivatives of the function f, which are

$$\begin{aligned} f'(x) &= \alpha(1+x)^{\alpha-1} \\ f''(x) &= \alpha(\alpha-1)(1+x)^{\alpha-2} \\ &\vdots \\ f^{(k)}(x) &= \alpha(\alpha-1)\cdots(\alpha-k+1)(1+x)^{\alpha-k}, \quad k \in \mathbb{N}. \end{aligned}$$

For all $k \in \mathbb{N}$, the derivative values at zero are

$$f^{(k)}(0) = \alpha(\alpha-1)\cdots(\alpha-k+1).$$

For shorthand, we now introduce the *generalized binomial symbol*

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}, \quad \alpha \in \mathbb{R}, \quad k \in \mathbb{N} \cup \{0\},$$

in terms of which the Taylor series of the power function f expanded about the point $a = 0$, also called the *binomial series*, reads

$$\sum \left[\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k \right]_{k=0}^{\infty} = \sum \left[\binom{\alpha}{k} x^k \right]_{k=0}^{\infty}.$$

We will divide the analysis of its convergence into two separate cases:

(i) If $\alpha \in \mathbb{N} \cup \{0\}$, i.e. if α is a nonnegative integer, then f is a polynomial, and we will write $\alpha = m$ to make clear that the power is an integer. In this case the generalized binomial symbols reduces to the familiar binomial symbol

$$\begin{aligned} \binom{m}{k} &= \frac{m(m-1)\cdot(m-k+1)}{k!} \\ &= \frac{m!}{k!(m-k)!}, \quad k = 0, 1, \ldots, m. \end{aligned}$$

Meanwhile, for $k > m$ we define $\binom{m}{k} = 0$. In this case the binomial series has only a finite number of nonzero terms and thus converges for all $x \in \mathbb{R}$. For all $x \in \mathbb{R}$ we can then write the power function $x \mapsto (1+x)^m$ in the familiar form

$$(1+x)^m = \sum_{k=0}^{m} \binom{m}{k} x^k.$$

(ii) Now assume $\alpha \in \mathbb{R}$ is an arbitrary real number. We aim to find a bound for the remainder of the binomial series. If $x \in [0, 1)$, it is convenient to use the Lagrange form of the remainder, which for all $n \in \mathbb{N} \cup \{0\}$ and some $\theta \in (0, 1)$ reads

$$\begin{aligned} R_n(x) &= \frac{f^{(n+1)}(\theta x)}{(n+1)!} x^{n+1} \\ &= \binom{\alpha}{n+1} (1+\theta x)^{\alpha-(n+1)} x^{n+1} \\ &= (1+\theta x)^{\alpha} \cdot \frac{\alpha(\alpha-1)\cdots(\alpha-n)}{1\cdot 2\cdots(n+1)} \left(\frac{x}{1+\theta x} \right)^{n+1}. \end{aligned}$$

We can bound the expression $(1+\theta x)^\alpha$ above by the constant $\max\{1, 2^\alpha\}$. Because $x \in [0,1)$ and $\theta \in (0,1)$, it follows that $x/(1+\theta x) \leq x$, which leads to the bound

$$\begin{aligned}|R_n(x)| &\leq \max\{1, 2^\alpha\} \cdot \left|\frac{\alpha}{1}\right| \cdot \left|\frac{\alpha-1}{2}\right| \cdots \left|\frac{\alpha-n}{n+1}\right| x^{n+1} \\ &= \max\{1, 2^\alpha\} \cdot \left|\frac{\alpha}{1}x\right| \cdot \left|\frac{\alpha-1}{2}x\right| \cdots \left|\frac{\alpha-n}{n+1}x\right|\end{aligned}$$

We then note that

$$\lim_{n\to\infty} \left|\frac{\alpha-n}{n+1}x\right| = \lim_{n\to\infty} \left|\frac{(\alpha/n)-1}{1+(1/n)}x\right| = |x| < 1,$$

which means that for large enough n we can bound the sequence $(|R_n(x)|)_n$ above by the convergent geometric sequence, which means that the remainder must also converge, i.e.

$$\lim_{n\to\infty} R_n(x) = 0, \quad x \in [0,1).$$

For $x \in (-1,0)$, we will use the Cauchy form of the binomial series's remainder, which for all $n \in \mathbb{N} \cup \{0\}$ and some $\theta \in (0,1)$ reads

$$\begin{aligned}R_n(x) &= (1-\theta)^n \frac{\alpha(\alpha-1)\cdots(\alpha-n)}{n!} \cdot (1+\theta x)^{\alpha-(n+1)} x^{n+1} \\ &= (1+\theta x)^{\alpha-1} x \left(\alpha \cdot \frac{\alpha-1}{1} \cdot \frac{\alpha-2}{2} \cdots \frac{\alpha-n}{n}\right) \cdot \left[\frac{(1-\theta)x}{1+\theta x}\right]^n.\end{aligned}$$

Similarly to the process used in the analysis of the logarithmic series, because $x \in (-1,0)$ we can make the bound

$$\left|\frac{(1-\theta)x}{1+\theta x}\right| < \left|\frac{(1-\theta)x}{1-\theta}\right| = |x|.$$

More so, we can bound the expression $\left|(1+\theta x)^{\alpha-1}x\right|$ above by the constant value $\max\{1, 2^{\alpha-1}\}$, which leads to the bound

$$|R_n(x)| \leq \max\{1, 2^{\alpha-1}\} \cdot |\alpha| \cdot \left|\frac{\alpha-1}{1}x\right| \cdot \left|\frac{\alpha-2}{2}x\right| \cdots \left|\frac{\alpha-n}{n}x\right|..$$

We then note that

$$\lim_{n\to\infty} \left|\frac{\alpha-n}{n}x\right| = |x| < 1,$$

which means that for large enough n we can again bound the sequence $(|R_n(x)|)_n$ above by the convergent geometric sequence, which means that the remainder must also converge for $x \in (-1,0)$, i.e.

$$\lim_{n\to\infty} R_n(x) = 0, \quad x \in (-1,0).$$

We have shown that the binomial series converges to the power function $f\colon x \mapsto (1+x)^\alpha$ for all $x \in (-1,1)$, i.e.

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k, \quad x \in (-1,1).$$

10.3 Power series

In this section we will generalize Taylor series to power series, which allow for complex-valued variables.

10.3.1 Convergence of power series

Definition 10.15: Power series and their convergence

A *power series expanded about the point* $a \in \mathbb{C}$ is a series of the form

$$\sum \left[a_n(z-a)^n\right]_{n=0}^{\infty},$$

where $a_0, a_1, \ldots \in \mathbb{C}$ are complex constants and $z \in \mathbb{C}$ is a complex-valued variable.

Setting the value of the complex variable z to an arbitrary complex number produces a series of complex numbers of the form studied in Chapter 5. If the resulting series is absolutely or conditionally convergent, we say that the power series *converges absolutely* or *converges conditionally*, respectively, *at the point* z. Otherwise, the power series is said to *diverge at* z.

The *region of convergence* of a power series is the set of all complex numbers at which the power series converges. Generally the region of convergence of a power series is a circular region centered at the point $a \in \mathbb{C}$ about which the series is expanded.

Remark 10.16: Translation of power series

Most often, we will analyze power series expanded about the point $a = 0$, i.e. power series of the form

$$\sum \left(a_n z^n\right)_{n=0}^{\infty}.$$

Note, however, that we can derive an arbitrary power series

$$\sum \left[a_n(z-a)^n\right]_{n=0}^{\infty}$$

from the zero-centered series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ from a simple translation by a in the complex plane. As a consequence of this observation, all results that we will prove for complex series expanded about $a = 0$ can be straightforwardly transfered to power series expanded about an arbitrary point $a \in \mathbb{C}$.

Proposition 10.17: Convergence of power series

If a power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ converges at the point $z_0 \in \mathbb{C}$, then the series also converges at every point $z \in \mathbb{C}$ for which $|z| < |z_0|$.

Proof. We will assume that $z_0 \neq 0$, because in this case the condition $|z| < |z_0|$ is meaningless. The power series $\sum \left(a_n z_0^n\right)_{n=0}^{\infty}$ converges, so by Corollary 5.6 its individual terms converge to zero, which in turns means that the sequence $(a_n z_0^n)_n$ is bounded. There thus exists a real number $M > 0$ for which

$$|a_n z_0^n| \leq M \text{ for all } n \in \mathbb{N} \cup \{0\}.$$

Now let $z \in \mathbb{C}$ be any complex number for which $|z| < |z_0|$. We then define the

nonnegative real number

$$q = \left|\frac{z}{z_0}\right| < 1 \iff |z| < |z_0|,$$

and note that the series $\sum \left(Mq^n\right)_{n=0}^{\infty}$ is a convergent geometric series because $q < 1$. We then conclude the proof with the bound

$$|a_n z^n| = \left|a_n z_0^n \left(\frac{z}{z_0}\right)^n\right| = |a_n z_0^n| a^n \leq Mq^n,$$

which means the convergent series $\sum \left(Mq^n\right)_{n=0}^{\infty}$ majorizes the series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$, which thus converges absolutely by the comparison test in Proposition 5.13. □

Proposition 10.17 implies that if a power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ converges at some point $z_0 \in \mathbb{C}$, then it converges absolutely at all points in the interior of the disk of radius $|z_0|$ centered at the origin at the complex plane. This observation raises the question of what is the largest-radius disk for which a given power series converges, which leads to the concept of a series's radius of convergence.

Definition 10.18: Radius of convergence

Let $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ be an arbitrary power series and define the set

$$B = \left\{|z|;\, z \in \mathbb{C} \text{ and } \sum \left(a_n z^n\right)_{n=0}^{\infty} \text{ converges at } z\right\} \subset \mathbb{R}.$$

The *radius of convergence* of the series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$, denoted by R, is then defined as

$$R = \begin{cases} \sup B, & B \text{ is bounded above} \\ \infty, & B \text{ is not bounded above.} \end{cases}$$

Note that B is always nonempty, since it always contains at least the number 0. If $R = 0$, the power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ converges only at the point $z = 0$. If $R = \infty$, the radius of convergence is obviously not a real number. Rather, we take $R = \infty$ to mean the radius of convergence is larger than every real number, in which case the series converges absolutely in the entire complex plane.

The radius of convergence is characterized by the following properties, which follow directly from its definition in Definition 10.18.

Corollary 10.19: Characterizing the radius of convergence

The radius of convergence of the power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ is the real number $R \in [0, \infty) \cup \{\infty\}$ for which:

(i) The series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ converges absolutely for all $z \in \mathbb{C}$ for which $|z| < R$.

(ii) The series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ diverges for all $z \in \mathbb{C}$ for which $|z| > R$.

It follows that a power series with radius of convergence R converges absolutely inside the complex disk with radius R, diverges outside the complex disk of radius R, and may either

converge or diverge on the disk's border.

Although we have stated Proposition 10.17 and Corollary 10.19 in terms of a series expanded about the point $a = 0 \in \mathbb{C}$, we can straightforwardly generalize these statements to power series expanded about an arbitrary point $a \in \mathbb{C}$ using a simple translation by a in the complex plane. Namely, the radii of convergence of the power series

$$\sum \left[a_n(z-a)^n\right]_{n=0}^{\infty} \quad \text{and} \quad \sum \left(a_n z^n\right)_{n=0}^{\infty}$$

are identical and equal to, say, R. The series $\sum \left[a_n z^n\right]_{n=0}^{\infty}$ converges absolutely for all z inside the complex disk $\mathcal{K}(0, R)$, while the translated series $\sum \left[a_n(z-a)^n\right]_{n=0}^{\infty}$ converges absolutely for all z inside the complex disk $\mathcal{K}(a, R)$.

Definition 10.20: Limit inferior and limit superior

Let (r_n) be an arbitrary sequence of real numbers and let E be the set of all cluster points of the sequence (r_n). In this case the *limit superior* of the sequence (r_n), denoted by $\limsup_{n\to\infty}(r_n)$, is defined as

$$\limsup_{n\to\infty}(r_n) = \begin{cases} \infty, & (r_n) \text{ unbounded above} \\ \sup E, & (r_n) \text{ bounded above and } E \neq \emptyset \\ -\infty, & (r_n) \text{ unbounded above.} \end{cases}$$

The number $\limsup_{n\to\infty}(r_n)$ can be loosely interpreted as the sequence's largest cluster point. The limit superior of a convergent real sequence is by definition equal to the sequence's limit.

We may also define the *limit inferior* of an arbitrary real sequence (r_n) as

$$\liminf_{n\to\infty}(r_n) = -\limsup_{n\to\infty}(-r_n).$$

Loosely, a real sequence's limit inferior is the sequence's smallest cluster point.

Remark 10.21: Some abuse of notation

We will use the following very convenient abuses of notation when working with inferior and superior limits:

- $-(\infty) = -\infty$,
- $-(-\infty) = \infty$,
- $x/\infty = 0$ for all $x \in \mathbb{R}^+$, and
- $x/0 = \infty$ for all $x \in \mathbb{R}^+$.

These are all abuses of notation because they treat ∞ and $-\infty$ as a a real number, when really ∞ and $-\infty$ are a shorthand notation used to indicate that a sequence or series diverges without bound, as mentioned in Remarks 4.22 and 5.3

Using the limit superior, we can now generalize many of series convergence tests from Chapter 5 to power series.

Proposition 10.22: Radius of convergence using the root test

The radius of convergence R of an arbitrary power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ is equal to

$$R = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}}.$$

Proof. We first define the quantity

$$\rho = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}},$$

then make the calculation

$$\limsup_{n\to\infty} \sqrt[n]{|a_n z^n|} = \limsup_{n\to\infty} \left(|z| \sqrt[n]{|a_n|}\right) = |z| \limsup_{n\to\infty} \sqrt[n]{|a_n|} = \frac{|z|}{\rho}.$$

If $|z| < \rho$, the largest cluster point of the sequence $\left(\sqrt[n]{|a_n z^n|}\right)_n$ is strictly less than 1, meaning there exists a natural number $N \in \mathbb{N}$ and real number $q \in \mathbb{R}$ for which

$$\sqrt[n]{|a_n z^n|} \leq q < 1 \text{ for all } n \geq N.$$

We can then apply the root test for series convergence (Proposition 5.17) to conclude that the series $\sum (a_n z^n)_{n=0}^{\infty}$ converges absolutely for all $|z| < \rho$.

If $|z| > \rho$, then either the largest cluster point of the sequence $\left(\sqrt[n]{|a_n z^n|}\right)_n$ is strictly greater than 1 or the sequence $\left(\sqrt[n]{|a_n z^n|}\right)_n$ is unbounded below. In either case there exists a natural number $M \in \mathbb{N}$ for which $\sqrt[n]{|a_n z^n|} \geq 1$ for all $n \geq M$, and so the series $\sum (a_n z^n)_{n=0}^{\infty}$ diverges by the root test.

We have shown that the series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ converges for $|z| < \rho$ and diverges for $|z| > \rho$, and so ρ is the series's radius of convergence. □

Proposition 10.23: Radius of convergence using the ratio test

The radius of convergence R of an arbitrary power series $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ is equal to

$$R = \frac{1}{\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|}.$$

Proof. The proof is essentially analogous to the proof of Proposition 10.22, with the ratio test replacing the root test. For the sake of completeness, we first define

$$\rho = \frac{1}{\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|},$$

and then make the calculation

$$\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n} z\right| = \limsup_{n\to\infty} |z| \left|\frac{a_{n+1}}{a_n}\right| = |z| \limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = \frac{|z|}{\rho}.$$

If $|z| < \rho$, the largest cluster point of the sequence $\left(\left|\frac{a_{n+1}}{a_n} z\right|\right)_n$ is strictly less than 1, meaning there exists a natural number $N \in \mathbb{N}$ and real number $q \in \mathbb{R}$ for which

$$\left|\frac{a_{n+1}}{a_n} z\right| \leq q < 1 \text{ for all } n \geq N.$$

We can then apply the ratio test for series convergence (Proposition 5.18) to conclude that the series $\sum(a_n z^n)_{n=0}^{\infty}$ converges absolutely for all $|z| < \rho$.

If $|z| > \rho$, then either the largest cluster point of the sequence $(|(a_{n+1}z)/(a_n)|)_n$ is strictly greater than 1 or the sequence $(|(a_{n+1}z)/(a_n)|)_n$ is unbounded below. In either case there exists a natural number $M \in \mathbb{N}$ for which $(|(a_{n+1}z)/(a_n)|) \geq 1$ for all $n \geq M$, and so the series $\sum(a_n z^n)_{n=0}^{\infty}$ diverges by the ratio test. □

Under certain conditions it is possible multiply power series in a well-defined manner. Series multiplication is, loosely, a generalization of polynomial multiplication; we give a more precise formulation below.

Definition 10.24: Multiplication of series and the Cauchy product

Let $\sum\left(a_n z^n\right)_{n=0}^{\infty}$ and $\sum\left(b_n z^n\right)_{n=0}^{\infty}$ be two power series. The product of these two series is the series

$$\sum\left(c_n z^n\right)_{n=0}^{\infty} = \left[\sum\left(a_n z^n\right)_{n=0}^{\infty}\right] \cdot \left[\sum\left(b_n z^n\right)_{n=0}^{\infty}\right],$$

whose coefficients are given by

$$c_n = a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0 = \sum_{k=0}^{n} a_k b_{n-k}.$$

More generally, let $\sum\left(u_n\right)_{n=0}^{\infty}$ and $\sum\left(v_n\right)_{n=0}^{\infty}$ be two series of complex numbers. The *Cauchy product* of these two series is the series

$$\sum\left(w_n\right)_{n=0}^{\infty} = \left[\sum\left(u_n\right)_{n=0}^{\infty}\right] \cdot \left[\sum\left(v_n\right)_{n=0}^{\infty}\right],$$

whose terms are given by

$$w_n = \sum_{k=0}^{n} u_k v_{n-k}.$$

Note that, by construction, the Cauchy product of two series is commutative. In general, a Cauchy product of two series does not converge; the following proposition gives a condition when a Cauchy product does converge.

Proposition 10.25: Convergence of the Cauchy product

Let $\sum\left(u_n\right)_{n=0}^{\infty}$ and $\sum\left(v_n\right)_{n=0}^{\infty}$ be two convergent series of complex numbers. If at least one of the series is also absolutely convergent, then the Cauchy product of the

series converges, and its sum is equal to

$$\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}u_k v_{n-k}\right)=\left(\sum_{k=0}^{\infty}u_n\right)\cdot\left(\sum_{k=0}^{\infty}v_n\right).$$

Proof. Assume the series $\sum\left(u_n\right)_{n=0}^{\infty}$ is absolutely convergent. For conciseness, we first define the shorthand sums

$$U=\sum_{n=0}^{\infty}u_n,\qquad V=\sum_{n=0}^{\infty}v_n,$$

and also the following shorthand partial sums for all $n\in\mathbb{N}$:

$$\begin{aligned}w_n&=\sum_{k=0}^{n}u_k v_{n-k}\\U_n&=u_0+u_1+\cdots+u_n,\\V_n&=v_0+v_1+\cdots+v_n,\\W_n&=w_0+w_1+\cdots+w_n.\end{aligned}$$

To complete the proof's preparation, we note that for all $n,m\in\mathbb{N}$ it holds that

$$\begin{aligned}W_n&=u_0V_n+u_1V_{n-1}+\cdots+u_nV_0\\W_{n+m}&=u_0V_{n+m}+\cdots u_mV_n+u_{m+1}V_{n-1}+\cdots+u_{m+n}V_0.\end{aligned}$$

We aim to show that the sequence of partial sums (W_n) converges to the number UV. To do so, first choose arbitrary $\epsilon\in\mathbb{R}^+$. Because the series $\sum\left(v_n\right)_{n=0}^{\infty}$ converges, it sequence of partials sums converges and is thus bounded, so there exists a sufficiently large real number $B\in\mathbb{R}^+$ for which

$$|V_k|\leq B\text{ for all }k\in\mathbb{N}\cup\{0\}.$$

More so, there exists a sufficiently large natural number $n\in\mathbb{N}$ for which

$$|V_{n+p}-V|<\epsilon\text{ for all }p\in\mathbb{N}\cup\{0\};$$

for each such p let $\eta_p=V_{n+p}-V$. For each $m\in\mathbb{N}$ we can then write

$$u_0V_{n+m}+\cdots+u_mV_n=(u_0+\cdots+u_m)V+u_0\eta_m+\cdots+u_m\eta_0.$$

Because each η_p is less than ϵ in absolute value, we can make the bound

$$|u_0\eta_m+\cdots+u_m\eta_0|<\big(|u_0|+|u_1|+\cdots+|u_m|\big)\epsilon\leq A\epsilon,$$

where A is the sum of the convergent series

$$A=|u_0|+|u_1|+|u_2|+\cdots=\sum_{k=0}^{\infty}|u_k|\geq 0.$$

Continuing, there exists a sufficiently large natural number m_0 such that

$$|u_{m+1}|+\cdots+|u_{m+l}|<\epsilon\text{ for all }m\geq m_0\text{ and }l\in\mathbb{N}.$$

We can then complete the proof with the calculation

$$\begin{aligned}
|W_{n+m} - UV| &= |u_0 V_{n+m} + \cdots u_m V_n \\
&\quad u_{m+1} V_{n-1} + \cdots + u_{m+n} V_0 - UV| \\
&= |(u_0 + \cdots + u_m)V + u_0\eta_m + \cdots + u_m\eta_0 \\
&\quad u_{m+1} V_{n-1} + \cdots + u_{m+n} V_0 - UV| \\
&= |(U_m - U)V + u_0\eta_m + \cdots + u_m\eta_0 \\
&\quad + u_{m+1} V_{n-1} + \cdots + u_{m+n} V_0| \\
&\overset{\text{(a)}}{\leq} |(U_m - U)V| + |u_0\eta_m + \cdots + u_m\eta_0| \\
&\quad + |u_{m+1} V_{n-1}| + \cdots + |u_{m+n} V_0| \\
&\leq \epsilon|V| + A\epsilon + \big(|u_{m+1}| + \cdots + |u_{m+n}|\big)B \\
&\leq \big(|V| + A + B\big)\epsilon,
\end{aligned}$$

where (a) follows from the triangle inequality. Since $\epsilon \in \mathbb{R}^+$ was arbitrary and the quantity $|V| + A + B$ is just a constant, the end result, i.e.

$$|W_{n+m} - UV| \leq \big(|V| + A + B\big)\epsilon$$

means that, for large enough m, the partial sum W_{n+m} grows arbitrarily close to the number UV, which completes the proof. □

Corollary 10.26: Cauchy product of power series

Let $\sum \left(a_n z^n\right)_{n=0}^{\infty}$ and $\sum \left(b_n z^n\right)_{n=0}^{\infty}$ be two power series that both converge at the point $z \in \mathbb{C}$. If at least one of these series also converges absolutely at z, then the product of the series converges at z, and the sum of the product is equal to

$$\sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k b_{n-k} \right) z^n = \left(\sum_{n=0}^{\infty} a_n z^n \right) \cdot \left(\sum_{n=0}^{\infty} b_n z^n \right).$$

As a result, the radius of convergence of the product must be greater than or equal to the radius of convergence of both of the multiplied series.

10.3.2 The complex exponential function

In this section we will use power series to expand the exponential function to the complex numbers; in this sense, this section can be seen as a continuation of Section 6.5, in which we defined the real-valued exponential function.

If fact, power series make it possible to generalize the definition of many elementary functions to the complex numbers—we will see some in this section.

Definition 10.27: The complex exponential function

The complex exponential function $\exp\colon \mathbb{C} \to \mathbb{C}$ is defined as

$$\exp(z) = \mathrm{e}^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}, \quad z \in \mathbb{C}.$$

This expression is well-defined because, as shown in Example 10.9, the exponential series converges absolutely on the entire complex plane.

Definition 10.28: Te complex sine and cosine functions

The complex sine and cosine functions $\sin\colon \mathbb{C} \to \mathbb{C}$ and $\cos\colon \mathbb{C} \to \mathbb{C}$ are defined as

$$\sin z = \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k+1}}{(2k+1)!}, \quad z \in \mathbb{C}$$
$$\cos z = \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k}}{(2k)!}, \quad z \in \mathbb{C}.$$

Like the complex exponential function, the complex sine and cosine function are defined on the entire complex plane.

We now prove the famous Euler formula, which uses power series to relate the complex exponential function to the sine and cosine functions.

Proposition 10.29: Euler's formula

For any $x \in \mathbb{R}$ the complex exponential, sine, and cosine functions obey the equality

$$\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i} \sin x.$$

This relationship is usually called *Euler's formula*.

Proof. Euler's formula follows directly from the calculation

$$\begin{aligned}
\cos x + \mathrm{i} \sin x &\overset{\text{(a)}}{=} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} + \mathrm{i} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \\
&\overset{\text{(b)}}{=} \sum_{k=0}^{\infty} \frac{(\mathrm{i}x)^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{(\mathrm{i}x)^{2k+1}}{(2k+1)!} \\
&= \sum_{n=0}^{\infty} \frac{(\mathrm{i}x)^n}{n!} \overset{\text{(c)}}{=} \mathrm{e}^{\mathrm{i}x},
\end{aligned}$$

where (a) follows from the power series for the complex sine and cosine functions, (b) from the identities $(-1)^k = (\mathrm{i}^2)^k = \mathrm{i}^{2k}$ and $\mathrm{i} \cdot (-1)^k = \mathrm{i} \cdot (\mathrm{i}^2)^k = \mathrm{i}^{2k+1}$, and (c) from the power series definition of the complex exponential function. □

Proposition 10.30: Product of two complex exponents

For any $z, w \in \mathbb{C}$, the complex exponential function obeys the identity

$$e^{z+w} = e^z \cdot e^w.$$

Proof. The exponential series is absolutely convergent, so we may take the Cauchy product of two exponential series to get

$$e^z \cdot e^w = \left(\sum_{n=0}^{\infty} \frac{z^n}{n!}\right) \cdot \left(\sum_{n=0}^{\infty} \frac{w^n}{n!}\right) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \frac{z^k}{k!} \cdot \frac{w^{n-k}}{(n-k)!}\right). \tag{10.2}$$

Meanwhile, the quantity e^{z+w} can be written in the form

$$\begin{aligned} e^{z+w} &= \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} \left[\sum_{k=0}^{n} \binom{n}{k} z^k w^{n-k}\right] \\ &= \sum_{n=0}^{\infty} \left[\sum_{k=0}^{n} \frac{1}{n!} \cdot \frac{n!}{k!(n-k)!} z^k w^{n-k}\right] \\ &= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \frac{z^k}{k!} \cdot \frac{w^{n-k}}{(n-k)!}\right). \end{aligned} \tag{10.3}$$

The right-hand sides of Equations 10.2 and 10.3 are equal, and so $e^{z+w} = e^z \cdot e^w$. □

Remark 10.31: Geometric interpretation of the complex exponent

Let $z \in \mathbb{C}$ be an arbitrary complex number, which we will write in the form $z = x + iy$ for two real numbers $x, y \in \mathbb{R}$. The complex exponent e^z may then be written

$$e^z = e^{x+iy} \overset{(a)}{=} e^x \cdot e^{iy} \overset{(b)}{=} e^x(\cos y + i \sin y),$$

where (a) uses Euler's formula and (b) uses Proposition 10.30. This expression offers a geometric interpretation of the complex exponential function as a real exponential function $x \mapsto e^x$ modulated by the complex sinusoidal function $y \mapsto \cos y + i \sin y$; note that when $y = 0$, i.e. when z is purely real, the above expression reduces to the familiar $e^z = e^x$.

10.3.3 Uniform convergence

We have seen that a power series can converge (or absolutely converge) at certain points and diverge at others. But so far when studying convergence, we have only analyzed convergence at a given point. Uniform convergence considers the convergence of a sequence or series at multiple points simultaneously. In particular, uniform convergence can be defined not only for power series, but more generally for sequences and series of arbitrary scalar functions.

We begin this chapter with a few definitions.

Definition 10.32: Sequences and series of scalar functions

A *sequence of scalar functions* is a mapping $n \mapsto f_n$ that associates every natural number $n \in \mathbb{N}$ with a corresponding scalar function $f_n \colon \mathbb{R} \to \mathbb{R}$. Usually we consider only the case when all of the functions f_n are defined on the same subset $U \subset \mathbb{R}$. We will variously denote sequences of functions by

$$(f_1, f_2, f_3, \ldots) = (f_n)_{n=1}^{\infty} = (f_n)_{n\in\mathbb{N}} = (f_n)_n = (f_n \colon U \to \mathbb{R})_n = (f_n)$$

With any sequence of functions $(f_n)_n$ we can associate (1) a *series of functions*

$$\sum (f_n)_{n=1}^{\infty}$$

and (2) the corresponding sequence of the series's partial sums, given by

$$\left(\sum_{k=1}^{n} f_k \right)_n .$$

Definition 10.33: Convergence of function sequences

Let $(f_n \colon U \to \mathbb{R})$ be a sequence of scalar functions defined on the subset $U \subset \mathbb{R}$.

(i) The sequence (f_n) *converges at the point* $x \in U$ if the sequence of scalar function values $\big(f_n(x)\big)_n$ converges—this sequence of function values is a regular sequence of real numbers of the form studied in Chapter 4.

(ii) The sequence (f_n) *converges pointwise* to the function $f \colon U \to \mathbb{R}$ if the sequence (f_n) converges at all points $x \in U$ and obeys.

$$f(x) = \lim_{n\to\infty} f_n(x) \text{ for all } x \in U.$$

In this case we say that f is the *limit* of the sequence (f_n) and write

$$f = \lim_{n\to\infty} f_n.$$

(iii) The sequence (f_n) *converges uniformly* on the set U to the function $f \colon U \to \mathbb{R}$ if for all $\epsilon \in \mathbb{R}^+$ there exists a sufficiently large $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \epsilon \text{ for all } n \geq N \text{ and all } x \in U.$$

Intuitively, uniform convergence means that a sequence of functions converges "equally quickly" across all points in the function domains.

Definition 10.34: Convergence of function series

Let $(f_n \colon U \to \mathbb{R})_n$ be a sequence of scalar functions defined on the subset $U \subset \mathbb{R}$, and let $\sum (f_n)_{n=1}^{\infty}$ be the corresponding series of functions. In this case

(i) The series $\sum (f_n)_{n=1}^{\infty}$ *converges at the point* $x \in U$ if the series of function values $\sum \big(f_n(x)\big)_{n=1}^{\infty}$ converges—this series of function values is a regular series

of real numbers of the form studied in Chapter 5.

Analogously, the series $\sum(f_n)_{n=1}^{\infty}$ *converges absolutely at the point* $x \in U$ if the series of function values $\sum \big(f_n(x)\big)_{n=1}^{\infty}$ converges absolutely in the sense of Definition 5.8.

(ii) The series $\sum(f_n)_{n=1}^{\infty}$ *converges pointwise* to the function $s\colon U \to \mathbb{R}$ if the sequence of partial sums $\big(\sum_{k=1}^{n} f_k\big)_n$ converges pointwise to s. In this case we say that the function s is the *sum* of the series $\sum(f_n)_{n=1}^{\infty}$ and write

$$s = \sum_{n=1}^{\infty} f_n.$$

Equivalently, the series $\sum(f_n)_{n=1}^{\infty}$ converges pointwise to the function $s\colon U \to \mathbb{R}$ if

$$s(x) = \sum_{n=1}^{\infty} f_n(x) \text{ for all } x \in U.$$

(iii) The series $\sum(f_n)_{n=1}^{\infty}$ *converges uniformly* on the set U to the function $s\colon U \to \mathbb{R}$ if the sequence of partial sums $(\sum_{k=1}^{n} f_k)_n$ converges uniformly on the set U to the function s.

By definition, every uniformly convergent sequence also converges pointwise. The converse, i.e. that every pointwise convergent sequence converges uniformly, is not in general true, as shown in the following example.

Example 10.35: Pointwise convergence does not imply uniform convergence

As an example of a sequence of functions that converges pointwise but not uniformly, consider the sequence given by

$$f_n\colon [0,1] \to \mathbb{R}, \quad f_n(x) = x^n.$$

This sequence converges pointwise on the interval $[0,1]$ to the function $f\colon [0,1] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0 & x \in [0,1), \\ 1 & x = 1. \end{cases}$$

However, the sequence does not converge uniformly on $[0,1]$, since for $\epsilon \in (0,1)$ there does not exist any natural number N for which $|f_n - f(x)| < \epsilon$ for all $n \geq N$ and all $x \in [0,1]$. For example, every element of the sequence (f_n) differs from the limit function f by more than $\epsilon/10$ at all points sufficiently close to the interval endpoint 1.

Proposition 10.36: Cauchy criterion for uniform convergence

A sequence of functions $(f_n\colon U \to \mathbb{R})_n$ defined on the subset $U \subset \mathbb{R}$ converges uniformly on U if, and only if, for all $\epsilon \in \mathbb{R}^+$ there exists sufficiently large $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \epsilon \text{ for all } m, n \geq N \text{ and all } x \in U.$$

Proof. ($\Rightarrow$) Assume the sequence (f_n) converges uniformly on U to the function $f\colon U \to \mathbb{R}$, and choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the sequence (f_n) converges uniformly, there exists sufficiently large $N \in \mathbb{N}$ such that

$$|f(x) - f_n(x)| < \frac{\epsilon}{2} \text{ for all } n \geq N \text{ and all } x \in U.$$

For any $m, n \geq N$ we may then complete this part of the proof with the calculation

$$\begin{aligned} |f_n(x) - f_m(x)| &= |f_n(x) - f(x) + f(x) - f_m(x)| \\ &\overset{\text{(a)}}{\leq} |f_n(x) - f(x)| + |f(x) - f_m(x)| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

where (a) uses the triangle inequality

($\Leftarrow$) Under the proposition's assumptions, the sequence of numbers $\big(f_n(x)\big)_n$ is a Cauchy sequence for all $x \in U$. Being a Cauchy sequence, the sequence $\big(f_n(x)\big)_n$ is convergent for $x \in U$ and so the sequence of functions (f_n) converges pointwise to some function $f\colon U \to \mathbb{R}$. Now choose arbitrary $\epsilon \in \mathbb{R}^+$. By the proposition's assumptions there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \frac{\epsilon}{2} \text{ for all } m, n \geq N \text{ and all } x \in U.$$

For all $x \in U$ and all $n \geq N$ it then holds that

$$|f_n(x) - f(x)| \overset{\text{(a)}}{=} \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \frac{\epsilon}{2} < \epsilon,$$

where (a) holds because the sequence (f_m) converges pointwise to f. □

Proposition 10.37: Weierstrass M-test for uniform convergence

Let $\sum(f_n\colon U \to \mathbb{R})_{n=1}^{\infty}$ be a series of functions defined on the subset $U \subset \mathbb{R}$ and let $\sum(M_n)_{n=1}^{\infty}$ be a convergent series of nonnegative numbers for which

$$|f_n(x)| \leq M_n \text{ for all } n \in \mathbb{N} \text{ and all } x \in U.$$

In this case the series $\sum(f_n)_{n=1}^{\infty}$ converges uniformly on the set U and converges absolutely at every point in U.

Proof. Let (s_n) denote the sequence of partial sums of the series $\sum(f_n)_{n=1}^{\infty}$, and choose arbitrary $\epsilon \in \mathbb{R}^+$. The series $\sum(M_n)_{n=1}^{\infty}$ converges, so by the Cauchy criterion for the convergence of number series (Proposition 5.5) there exists $N \in \mathbb{N}$ for which

$$\sum_{k=n}^{\infty} M_k < \epsilon \text{ for all } n, m \in \mathbb{N} \text{ for which } m \geq n \geq N.$$

For all $m, n \in \mathbb{N}$ for which $m > n \geq N$ and all $x \in U$ we then compute

$$|s_m(x) - s_n(x)| = \left|\sum_{k=n+1}^{m} f_k(x)\right| \overset{\text{(a)}}{\leq} \sum_{k=n+1}^{m} |f_k(x)| \leq \sum_{k=n+1}^{m} M_k \overset{\text{(b)}}{\leq} \sum_{k=n}^{m} M_k < \epsilon,$$

where (a) holds by the triangle inequality and (b) under the M-test's assumption that M_k is nonnegative for all $k \in \mathbb{N}$. □

Proposition 10.38: A uniformly convergent sequence of continuous functions converges to a continuous function

If $(f_n \colon U \to \mathbb{R})_n$ is a sequence of continuous functions that converges uniformly on the subset $U \subset \mathbb{R}$ to the function $f \colon U \to \mathbb{R}$, then the function f is also continuous.

Proof. We will show that the function f is continuous at every point in U. Let $x \in U$ be an arbitrary point in U and choose arbitrary $\epsilon \in \mathbb{R}^+$. The sequence (f_n) converges uniformly to f, so there exists $n \in \mathbb{N}$ for which

$$|f(y) - f_n(y)| < \frac{\epsilon}{3} \text{ for all } y \in U.$$

And because the function f_n is continuous at x, there exists $\delta \in \mathbb{R}^+$ for which

$$|f_n(x) - f_n(y)| < \frac{\epsilon}{3} \text{ for all } y \in U \text{ for which } |x - y| < \delta.$$

For all $y \in (x - \delta, x + \delta) \cap U$ we then make the calculation

$$\begin{aligned} |f(y) - f(x)| &= |f(y) - f_n(y) + f_n(y) - f_n(x) + f_n(x) - f(x)| \\ &\overset{\text{(a)}}{\leq} |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon, \end{aligned}$$

where (a) holds by the triangle inequality. The inequality $|f(y) - f(x)| < \epsilon$ for all $y \in (x - \delta, x + \delta) \cap U$ means that the function f is continuous at x; because the point $x \in U$ was arbitrary, f is continuous on U. □

Corollary 10.39: A uniformly convergent series of continuous functions converges to a continuous function

If $\sum(f_n \colon U \to \mathbb{R})_{n=1}^{\infty}$ is a series of continuous functions that converges uniformly on the subset $U \subset \mathbb{R}$, then the function's sum $\sum_{n=1}^{\infty} f_n \colon U \to \mathbb{R}$ is also a continuous function.

Proof. The corollary follows directly from applying Proposition 10.38 to the sequence of partial sums $\left(\sum_{k=1}^{n} f_k\right)_n$. □

Proposition 10.40: A uniformly convergent sequence of integrable functions converges to an integrable function

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$ and let $(f_n \colon [a, b] \to \mathbb{R})_n$ be a sequence of integrable functions that converges uniformly to the function $f \colon [a, b] \to \mathbb{R}$ on the nonempty, closed, bounded interval $[a, b]$. In this case the function f is also

integrable and obeys

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x.$$

Proof. Each of the functions f_n is integrable and thus bounded, and so its uniformly convergent limit f is also bounded. For arbitrary $n \in \mathbb{N}$ define the number ϵ_n as

$$\epsilon_n = \sup\big\{|f_n(x) - f(x)|\,;\, x \in [a,b]\big\}.$$

Since the functions f_n converge uniformly and thus also pointwise, for all $x \in [a,b]$ it holds that

$$f_n(x) - \epsilon_n \leq f(x) \leq f_n(x) + \epsilon_n, \tag{10.4}$$

which in turn leads to the calculation

$$\begin{aligned}\int_a^b f_n(x)\,\mathrm{d}x - \epsilon_n(b-a) = \int_a^b \big[f_n(x) - \epsilon_n\big]\,\mathrm{d}x &\overset{\text{(a)}}{\leq} \mathrm{s}(f) \\ \leq \mathrm{S}(f) &\overset{\text{(b)}}{\leq} \int_a^b \big[f_n(x) + \epsilon_n\big]\,\mathrm{d}x \\ = \int_a^b f_n(x)\,\mathrm{d}x + \epsilon_n(b-a),\end{aligned} \tag{10.5}$$

where (a) and (b) both follow from Equation 10.4 and the definition of f's upper and lower Darboux integrals $\mathrm{s}(f)$ and $\mathrm{S}(f)$. Equation 10.5 in turn implies the bound

$$0 \leq \mathrm{S}(f) - \mathrm{s}(f) \leq 2\epsilon_n(a-b). \tag{10.6}$$

We then note that $\lim_{n\to\infty} \epsilon_n = 0$ because the sequence (f_n) converges uniformly to the function f, then apply the limit $\lim_{n\to\infty} \epsilon_n = 0$ to Equation 10.6 to get

$$0 \leq \mathrm{S}(f) - \mathrm{s}(f) \leq 0 \implies \mathrm{s}(f) = \mathrm{S}(f).$$

The result $\mathrm{s}(f) = \mathrm{S}(f)$ means that the function f is integrable. More so, Equation 10.5 implies the bound

$$\left|\int_a^b f_n(x)\,\mathrm{d}x - \int_a^b f(x)\,\mathrm{d}x\right| \leq \epsilon_n(b-a),$$

which in the limit $n \to \infty$ means that

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x. \qquad \square$$

Remark 10.41: On switching the order of limit and integral evaluation

Proposition 10.40 means that the uniformly convergent limit of a sequence of integrable functions is again an integrable function. More so, for functions satisfying the conditions of Proposition 10.40, it is possible to switch the order of limit and

integral evaluation, i.e.

$$\int_a^b \left[\lim_{n\to\infty} f_n(x)\right] \mathrm{d}x = \lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x. \tag{10.7}$$

We can verify the interchange of integration and limit evaluation in Equation 10.7 is valid directly from the calculation

$$\int_a^b \left[\lim_{n\to\infty} f_n(x)\right] \mathrm{d}x \overset{\text{(a)}}{=} \int_a^b f(x)\,\mathrm{d}x \overset{\text{(b)}}{=} \lim_{n\to\infty} \int_a^b f_n(x)\,\mathrm{d}x,$$

where (a) follows from the uniform convergence of (f_n) to f and (b) from Proposition 10.40.

Corollary 10.42: A uniformly convergent series of integrable functions converges to an integrable function

Let $a, b \in \mathbb{R}$ be two real numbers for which $a \leq b$ and let $\sum \left(f_n \colon [a,b] \to \mathbb{R}\right)_{n=1}^{\infty}$ be a series of integrable functions that converges uniformly on the nonempty, closed, bounded interval $[a,b]$. In this case the series's sum $\sum_{n=1}^{\infty} f_n \colon [a,b] \to \mathbb{R}$ is also integrable and obeys

$$\int_a^b \left[\sum_{n=1}^{\infty} f_n(x)\right] \mathrm{d}x = \sum_{n=1}^{\infty} \int_a^b f_n(x)\,\mathrm{d}x.$$

Proof. The corollary follows directly from applying Proposition 10.40 to the sequence of partial sums $\left(\sum_{k=1}^{n} f_k\right)_n$. □

Example 10.43: Uniform convergence to the absolute value function

Let $(f_n \colon \mathbb{R} \to \mathbb{R})_n$ be the sequence of functions given by

$$f_n \colon x \mapsto \sqrt{x^2 + 1/n} \text{ for all } n \in \mathbb{N} \text{ and all } x \in \mathbb{R}.$$

This sequence of functions converges uniformly to the absolute value function

$$f \colon x \mapsto |x|.$$

As motivation for the next proposition, we note that although the functions f_n are differentiable for all $n \in \mathbb{N}$ and converge uniformly to f, the function f itself is not differentiable at the point 0. It turns out that the sequence of derivatives (f_n') converges pointwise, but *not* uniformly, to the sign function $\operatorname{sgn} \colon \mathbb{R} \to \mathbb{R}$; this is to be intuitively expected, since the limit function sgn is not continuous.

The following proposition gives a condition under which the order of differentiation and limit evaluation can be switched for sequences of functions.

Proposition 10.44: Differentiation of uniformly convergent sequences

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let $\big(f_n \colon (a,b) \to \mathbb{R}\big)_n$ be a sequence of differentiable functions defined on the nonempty, open, bounded interval (a,b) that converges at some point $x_0 \in (a,b)$, and let the sequence of derivatives $(f_n')_n$ converge uniformly on the interval (a,b) to some function $g \colon (a,b) \to \mathbb{R}$. In this case the sequence of functions (f_n) converges uniformly on (a,b) to a differentiable function $f \colon (a,b) \to \mathbb{R}$ for which $f' = g$.

Proof. Choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the sequence (f_n) converges at the point x_0, by the Cauchy criterion for sequence convergence there exists $N_1 \in \mathbb{N}$ for which

$$|f_n(x_0) - f_m(x_0)| < \frac{\epsilon}{2} \text{ for all } m, n \geq N_1, \tag{10.8}$$

and because the sequence $(f')_n$ converges uniformly, by the Cauchy criterion for uniform function sequence convergence there exists $N_2 \in \mathbb{N}$ for which

$$\big|f_n'(t) - f_m'(t)\big| < \frac{\epsilon}{2(b-a)} \text{ for all } m, n \geq N_2 \text{ and all } t \in (a,b).$$

Let $N = \max\{N_1, N_2\}$. For any two points $x, t \in (a,b)$ for which $x \neq t$ and any natural numbers $m, n \geq N$, we can then apply the mean value theorem to the function $f_n - f_m$ on the interval between x and t to guarantee the existence of a point ξ in the interval between x and t for which

$$\begin{aligned} \big|\big[f_n(x) - f_m(x)\big] - \big[f_n(t) - f_m(t)\big]\big| &\overset{\text{(a)}}{=} \big|f_n'(\xi) - f_m'(\xi)\big| \cdot |x - t| \\ &< \frac{\epsilon}{2(b-a)} \cdot |x - t| \overset{\text{(b)}}{<} \frac{\epsilon}{2}, \end{aligned} \tag{10.9}$$

where (a) holds by the mean value theorem applied to the function $f_n - f_m$ and (b) because $|x - t| < (b - a)$, since $x, t \in (a,b)$. We then evaluate the resulting inequality for $t = x_0$ to get the bound

$$\begin{aligned} |f_n(x) - f_m(x)| &= \big|\big[f_n(x) - f_m(x)\big] - \big[f_n(x_0) - f_m(x_0)\big] + \big[f_n(x_0) - f_m(x_0)\big]\big| \\ &\overset{\text{(a)}}{\leq} \big|\big[f_n(x) - f_m(x)\big] - \big[f_n(x_0) - f_m(x_0)\big]\big| + |f_n(x_0) - f_m(x_0)| \\ &\overset{\text{(b)}}{<} \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

for all $x \in (a,b)$ and for all natural numbers $m, n \geq N$; point (a) follows from the triangle inequality and (b) from Equations 10.8 and 10.9. By the Cauchy criterion for uniform convergence of function sequences, the result $|f_n(x) - f_m(x)| < \epsilon$ means that the sequence (f_n) converges uniformly on the interval (a,b) to some function $f \colon (a,b) \to \mathbb{R}$.

It remains to the show that the function f is differentiable and that $f' = g$. Let x be an arbitrary point in the open interval (a,b). For all $t \in (a,b) \setminus \{x\}$ we then define

$$\phi_n(t) = \frac{f_n(t) - f_n(x)}{t - x} \quad \text{and} \quad \phi(t) = \frac{f(t) - f(x)}{t - x}.$$

For all $t \in (a,b) \setminus \{x\}$ and all natural numbers $n, m \geq N$, we then apply the bound in Equation 10.9 to the function $\phi_n - \phi_m$ to get

$$|\phi_n(t) - \phi_m(t)| < \frac{\epsilon}{2(b-a)}.$$

By the Cauchy criterion for uniform convergence, the sequence (ϕ_n) converges uniformly to the function ϕ on the subset $(a,b) \setminus \{x\}$. There thus exists a sufficiently large natural number $n \geq N$ for which

$$|\phi(t) - \phi_n(t)| < \frac{\epsilon}{3} \text{ for all } t \in (a,b) \setminus \{x\}.$$

More so, because the sequence of derivatives $(f_n')_n$ converges uniformly, and thus also pointwise, to the function g, we may choose n to be large enough that

$$\left|f_n'(x) - g(x)\right| < \frac{\epsilon}{3}.$$

Next, we combine the definition of the derivative and the definition of the function ϕ_n as a difference quotient for the function f_n to get the relationship

$$\lim_{t \to x} \phi_n(t) = f_n'(x),$$

so there exists a sufficiently small open neighborhood $V \subset \mathbb{R}$ of the point x for which

$$\left|\phi_n(t) - f_n'(x)\right| < \frac{\epsilon}{3} \text{ for all } t \in \big(V \cap (a,b)\big) \setminus \{x\}.$$

For all $t \in \big(V \cap (a,b) \setminus \{x\}\big)$, we then combine the previous three bounds to get

$$\begin{aligned} |\phi(t) - g(x)| &= \left|\phi(t) - \phi_n(t) + \phi_n(t) - f_n'(x) + f_n'(x) - g(x)\right| \\ &\leq |\phi(t) - \phi_n(t)| + \left|\phi_n(t) - f_n'(x)\right| + \left|f_n'(x) - g(x)\right| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon, \end{aligned}$$

which means that $\lim_{t \to x} \phi(t) = g(x)$. Since x is an arbitrary point in (a,b), we can then complete the proof with the calculation

$$f'(x) \overset{\text{(a)}}{=} \lim_{t \to x} \frac{f(t) - f(x)}{t - x} \overset{\text{(b)}}{=} \lim_{t \to x} \phi(t) = g(x),$$

where (a) follows from the definition of the derivative and (b) from the definition of the function ϕ. □

Remark 10.45: On switching the order of limit and derivative evaluation

It is possible to switch the order of limit and derivative evaluation for functions satisfying the conditions of Proposition 10.44. More precisely, if the sequence of differentiable functions $\big(f_n \colon (a,b) \to \mathbb{R}\big)_n$ converges pointwise on the interval (a,b) and the sequence of function derivatives $(f_n')_n$ converges uniformly on the interval

(a, b) to some function $g\colon (a, b) \to \mathbb{R}$, then

$$\left(\lim_{n\to\infty} f_n\right)' = \lim_{n\to\infty} f_n'.$$

This equality follows directly from the calculation

$$\left(\lim_{n\to\infty} f_n\right)' \overset{\text{(a)}}{=} f' \overset{\text{(b)}}{=} g \overset{\text{(c)}}{=} \lim_{n\to\infty} f_n',$$

where (a) follows from the convergence of (f_n) to f and (b) and (c) from Proposition 10.44.

Corollary 10.46: Differentiation of uniformly convergent series

Let $a, b \in \mathbb{R}$ be two real numbers for which $a < b$, let $\sum \left(f_n\colon (a, b) \to \mathbb{R}\right)_{n=1}^{\infty}$ be a series of differentiable functions that converges at some point $x_0 \in (a, b)$, and let the series of derivatives $\sum(f_n')_{n=1}^{\infty}$ converge uniformly on the interval (a, b). In this case the series $\sum(f_n)_{n=1}^{\infty}$ also converges uniformly on (a, b) and the series's sum is a differentiable function for which

$$\left(\sum_{n=1}^{\infty} f_n\right)' = \sum_{n=1}^{\infty} f_n'.$$

Proof. The corollary follows from applying Proposition 10.44 to the sequences of partial sums $\left(\sum_{k=1}^{n} f_k\right)_n$ and $\left(\sum_{k=1}^{n} f_k'\right)_n$. □

10.3.4 Uniform convergence of real power series

We will now apply the results of this section to real power series. A real power series $\sum(a_n x^n)_{n=0}^{\infty}$ with radius of convergence R converges pointwise, but in general not uniformly, on the interval $(-R, R)$.

Proposition 10.47: Uniform convergence of real power series

Any real power series $\sum(a_n x^n)_{n=0}^{\infty}$ with radius of convergence R converges uniformly on the interval $[-r, r]$ for any real number $r \in [0, R)$.

Proof. Let $x_0 \in \mathbb{R}$ be a real number for which $r < x_0 < R$, and define the real number $q = r/x_0$, which by construction satisfies $q \in [0, 1)$. Because $x_0 < R$, the series $\sum(a_n x^n)_{n=0}^{\infty}$ converges absolutely at x_0, and so the sequence of absolute value terms $(|a_n x_0^n|)_n$ converges and is bounded. This in turn means there exists an upper bound $M \in \mathbb{R}^+$ for which

$$|a_n x_0^n| \leq M \text{ for all } n \in \mathbb{N} \cup \{0\}.$$

More so, by the definition of x_0 and q, it holds that

$$0 \leq \left|\frac{x}{x_0}\right| \leq \frac{r}{x_0} = q < 1 \text{ for all } x \in [-r, r].$$

The previous two bounds make it possible to majorize the power series $\sum(a_n x^n)_{n=0}^{\infty}$ with a convergent geometric series on the interval $[-r, r]$, i.e.

$$|a_n x^n| = \left|a_n x_0^n \left(\frac{x}{x_0}\right)^n\right| = |a_n x_0^n| \cdot \left|\left(\frac{x}{x_0}\right)^n\right| \leq Mq^n.$$

The power series $\sum(a_n x^n)_{n=0}^{\infty}$ thus converges uniformly on the interval $[-r, r]$ by the Weierstrass M-test. □

Proposition 10.48: Differentiation of real power series

Let $\sum(a_n x^n)_{n=0}^{\infty}$ be a real power series with radius of convergence of R and let the function $s\colon (-R, R) \to \mathbb{R}$ be the series's sum on the interval $(-R, R)$, i.e.

$$s\colon x \mapsto \sum_{n=0}^{\infty} a^n x^n, \quad x \in (-R, R).$$

In this case

(i) The power series $\sum(na_n x^{n-1})_{n=1}^{\infty}$ also has radius of convergence R.

(ii) The function s is differentiable on the interval $(-R, R)$, and for all $x \in (-R, R)$, the derivative s' is given by

$$s'(x) = \sum_{n=1}^{\infty} na_n x^{n-1}.$$

Proof.

(i) This point follows the calculation

$$\begin{aligned}\limsup_{n\to\infty} \sqrt[n]{|na_n|} &= \limsup_{n\to\infty} \left(\sqrt[n]{n}\sqrt[n]{|a_n|}\right) \\ &= \lim_{n\to\infty} \sqrt[n]{n} \cdot \limsup_{n\to\infty} \sqrt[n]{|a_n|} \\ &= 1 \cdot \limsup_{n\to\infty} \sqrt[n]{|a_n|} \overset{\text{(a)}}{=} \frac{1}{R},\end{aligned}$$

where (a) holds by Proposition 10.22. Again by Proposition 10.22, the power series with coefficients na^n, i.e. $\sum(na_n x^{n-1})_{n=1}^{\infty}$, has radius of convergence R.

(ii) By point (i), the series of derivatives $\sum(na_n x^{n-1})_{n=1}^{\infty}$ has radius of convergence R, and so by Proposition 10.47, converges uniformly on any closed interval $[-r, r]$ for which $r < R$. This point then follows directly from Corollary 10.46. □

Corollary 10.49: Sums of real power series are smooth

Let $\sum(a_n x^n)_{n=0}^{\infty}$ be a real power series with radius of convergence of R and let the function $s\colon (-R,R) \to \mathbb{R}$ be the series's sum on the interval $(-R,R)$, i.e.

$$s\colon x \mapsto \sum_{n=0}^{\infty} a^n x^n, \quad x \in (-R,R).$$

In this case s is a smooth function.

Proof. Differentiating a power series again produces a power series, so this corollary follows from repeated application of Proposition 10.48 to the function s. □

Example 10.50: Differentiating real power series

(1) If we differentiate the exponential series, whose sum is

$$\mathrm{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

term by term using Proposition 10.48, we get the expected result

$$\begin{aligned}(\mathrm{e}^x)' &= \left(\sum_{n=0}^{\infty} \frac{x^n}{n!}\right)' = \sum_{n=1}^{\infty} n\frac{x^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} \\ &\overset{\text{(a)}}{=} \sum_{k=0}^{\infty} \frac{x^k}{k!} = \mathrm{e}^x,\end{aligned}$$

where in (a) we have defined the new index of summation $k = n-1$.

(2) The logarithmic series has radius of convergence equal to 1, and thus by Proposition 10.47 converges on every closed, bounded subinterval of the interval $(-1,1)$, but not on the interval $(-1,1)$ itself. The sum of the logarithmic series is unbounded, but every finite partial sum of the logarithmic series is bounded on $(-1,1)$.

If we differentiate the logarithmic series, whose sum is

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots,$$

term by term, we get the expected result

$$\begin{aligned}\big[\ln(1+x)\big]' &= x' - \left(\frac{x^2}{2}\right)' + \left(\frac{x^3}{3}\right)' - \left(\frac{x^4}{4}\right)' + \cdots \\ &= 1 - x + x^2 - x^3 + \cdots \\ &= \frac{1}{1+x}, \quad x \in (-1,1).\end{aligned}$$

(3) If we differentiate the sine series, whose sum is

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \cdots,$$

term by term, we get the expected result

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \pm \cdots .$$

(4) Finally, let us consider the binomial series. For any $\alpha \in \mathbb{R} \setminus \{0\}$ and all $x \in (-1, 1)$, the series's sum is

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n.$$

If we differentiate the series's sum term by term using Proposition 10.48, the result is

$$\big[(1+x)^\alpha\big]' = \alpha(1+x)^{\alpha-1} = \sum_{n=1}^{\infty} n\binom{\alpha}{n} x^{n-1},$$

We can then divide through by α and rearrange to get

$$\begin{aligned}
(1+x)^{\alpha-1} &= \sum_{n=1}^{\infty} \frac{n}{\alpha}\binom{\alpha}{n} x^{n-1} \\
&= \sum_{n=1}^{\infty} \frac{n}{\alpha} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^{n-1} \\
&= \sum_{n=1}^{\infty} \frac{(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{(n-1)!} x^{n-1} \\
&\overset{\text{(a)}}{=} \sum_{k=0}^{\infty} \frac{(\alpha-1)(\alpha-2)\cdots\big[(\alpha-1)-k+1\big]}{k!} x^k \\
&= \sum_{k=0}^{\infty} \binom{\alpha-1}{k} x^k,
\end{aligned} \tag{10.10}$$

where (a) follows from defining the new summation index $k = n-1$. To belabor the point, we then take the example full circle with calculation

$$\begin{aligned}
\big[(1+x)^\alpha\big]' &= \sum_{n=1}^{\infty} n\binom{\alpha}{n} x^{n-1} \\
&= \sum_{n=1}^{\infty} n \cdot \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^{n-1} \\
&= \alpha \sum_{n=1}^{\infty} \frac{(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{(n-1)!} x^{n-1} \\
&\overset{\text{(a)}}{=} \alpha \sum_{k=0}^{\infty} \frac{(\alpha-1)(\alpha-2)\cdots\big[(\alpha-1)-k+1\big]}{k!} x^k \\
&= \alpha \sum_{k=0}^{\infty} \binom{\alpha-1}{k} x^k \overset{\text{(b)}}{=} \alpha(1+x)^{\alpha-1},
\end{aligned}$$

where (a) follows from defining the new summation index $k = n-1$ and (b) from Equation 10.10.

11 Functions of multiple variables

This chapter is a brief introduction to functions of multiple variables. We will see that many concepts from the theory of scalar functions can be straightforwardly extended to multivariable functions. We will briefly cover higher-dimensional Euclidean space, then cover continuity, differentiability, Taylor series, and the extrema of real-valued multivariable functions.

11.1 The Euclidean space $\mathbb{R}^n$

When the symbol n is used without qualification in this section, it implicitly refers to any natural number $n \in \mathbb{N}$.

Topic 11.1: The set $\mathbb{R}^n$ and n-dimensional Euclidean space

The set $\mathbb{R}^n$ is the n-times Cartesian product of the set of real numbers $\mathbb{R}$ with itself, i.e.

$$\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \cdots \mathbb{R} = \{(v_1, v_2, \ldots, v_n);\, v_1, v_1, \ldots, v_n \in \mathbb{R}\}$$

As for the three-dimensional Euclidean space $\mathbb{R}^3$ covered in Chapter 3, we can generalize the set $\mathbb{R}^n$ to a vector space by equipping $\mathbb{R}^n$ with the operations of vector addition and scalar multiplication. To this end, for any two n-tuples $(v_1, v_2, \ldots, v_n), (w_1, w_2, \ldots, w_n) \in \mathbb{R}^n$ and for any real number $\alpha \in \mathbb{R}$ we define

(i) The *vector sum* of $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_n)$ as

$$(v_1, \ldots, v_n) + (w_1, \ldots, w_n) = (v_1 + w_1, \ldots, v_n + w_n) \in \mathbb{R}^n$$

(ii) The *scalar multiple* of the n-tuple $(v_1, \ldots, v_n)$ and the real number $\alpha \in \mathbb{R}$ as

$$\alpha(v_1, \ldots, v_n) = (\alpha v_1, \ldots, \alpha v_n) \in \mathbb{R}^n.$$

The set $\mathbb{R}^n$ equipped with the above two operations of vector addition and scalar multiplication is called the *n-dimensional Euclidean vector space* $\mathbb{R}^n$. In this context, elements of $\mathbb{R}^n$ are also called *vectors* or *points* in $\mathbb{R}^n$, and real numbers are called *scalars*. We will denote the components of an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^n$ by

$$v_1, v_2, \ldots, v_n,$$

and write this vector as

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n.$$

The vector space $\mathbb{R}^n$ is called "Euclidean" because it obeys the intuitive Euclidean geometry familiar from everyday life; non-Euclidean geometries fall beyond the scope of this course.

11.1.1 The scalar product on $\mathbb{R}^n$

In this section we define the scalar product operation on $\mathbb{R}^n$ and introduce some useful geometrical concepts made possible by the scalar product.

Definition 11.1: Scalar product

The *scalar product* of any two vectors $\boldsymbol{v}, \boldsymbol{w}, \in \mathbb{R}^n$ is the function

$$\cdot : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$$

defined as

$$\boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n \in \mathbb{R}.$$

The scalar product is also called the *dot product* and *inner product*; the name "inner product" is more commonly used in the theory of general vector spaces.

The scalar product can be used to define the length of a vector in $\mathbb{R}^n$.

Definition 11.2: Length of a vector

The *length* of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{v}\|$ and defined as

$$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

The length of a vector in Euclidean space is also called the vector's *magnitude*. In more abstract vector spaces, the length of a vector generalizes to a quantity called the vector's *norm*.

In Euclidean space the length of a vector can be denoted by either $|\boldsymbol{v}|$ or $\|\boldsymbol{v}\|$; we will use the notation $\|\boldsymbol{v}\|$ to avoid ambiguity with the absolute value of a scalar.

The scalar product in $\mathbb{R}^n$ has analogous properties to the scalar product in $\mathbb{R}^3$ covered in Section 3.2.1; we briefly review some of these below.

Proposition 11.3: Some properties of the scalar product

(i) (*The triangle inequality*) For any two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$,

$$\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\|.$$

The proof is analogous to the proof of the three-dimensional case in Proposition 3.8, so we omit it here.

(ii) (*The Cauchy-Schwartz inequality*) For any two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$,

$$|\boldsymbol{v} \cdot \boldsymbol{w}| \leq \|\boldsymbol{v}\| \|\boldsymbol{w}\|, \quad \boldsymbol{v}, \boldsymbol{w}, \in \mathbb{R}^n.$$

The proof is analogous to the proof of the three-dimensional case in Proposition 3.7, so we omit it here.

(iii) For any vector $\boldsymbol{v} \in \mathbb{R}^n$ and for every index $i = 1, 2, \ldots, n$:

$$|v_i| \leq \|\boldsymbol{v}\| \leq \sqrt{n} \max\{|v_1|, |v_2|, \ldots, |v_n|\}.$$

This inequality follows directly from the definition of the norm on $\mathbb{R}^n$ in Definition 11.2.

Definition 11.4: Perpendicular vectors

Two vectors in $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ are said to be *perpendicular* if

$$\boldsymbol{v} \cdot \boldsymbol{w} = 0.$$

In this case we write $\boldsymbol{v} \perp \boldsymbol{w}$. In more abstract vector spaces, perpendicularity generalizes to the concept of *orthogonality*, and two perpendicular vectors would be said to be *orthogonal*.

Definition 11.5: Distance between vectors

The *distance* between two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ is a function

$$d(\,\cdot\,,\,\cdot\,)\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$$

defined as

$$d(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\| = \sqrt{(v_1 - w_1)^2 + \cdots + (v_n - w_n)^2}.$$

In more abstract vector spaces, the distance between two vectors generalizes to a concept called the *metric* on the vector space.

11.1.2 Intervals and neighborhoods

Having established a well-defined concept of distance between vectors, we may now define basic topological concepts on $\mathbb{R}^n$, for example the notion of neighborhoods and open and closed sets. This section closely parallels the one-dimensional treatment of neighborhoods open and closed sets in Section 2.2.6.

Definition 11.6: Open ball

A *ball* is the mathematical name for the solid region bounded by a spherical surface. For any vector $\boldsymbol{a} \in \mathbb{R}^n$ and positive number $R \in \mathbb{R}^+$, the *open ball* centered at $\boldsymbol{a}$ with radius R, denoted by $\mathcal{K}(\boldsymbol{a}, R)$, is the subset

$$\mathcal{K}(\boldsymbol{a}, R) = \{\boldsymbol{v} \in \mathbb{R}^n;\, d(\boldsymbol{a}, \boldsymbol{v}) = \|\boldsymbol{a} - \boldsymbol{v}\| < R\} \subset \mathbb{R}^n.$$

In other words, the open ball $\mathcal{K}(\boldsymbol{a}, R)$ is the set of all points $\mathbb{R}^n$ whose distance from $\boldsymbol{a}$ is less than R.

Definition 11.7: Neighborhoods; interior and exterior points

Let $A \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$ and let $\boldsymbol{a} \in \mathbb{R}^n$ be an arbitrary vector.

(i) The subset A is a *neighborhood* of the point $\boldsymbol{a}$ if there exists $\delta \in \mathbb{R}^+$ for which

$$\mathcal{K}(\boldsymbol{a}, \delta) \subset A.$$

(ii) The point $\boldsymbol{a}$ is an *interior point* of the subset A if A is a neighborhood of $\boldsymbol{a}$.

(iii) The point $\boldsymbol{a}$ is an *exterior point* of the subset A if A's complement $A^{\mathsf{c}} = \mathbb{R}^n \setminus A$ is a neighborhood of $\boldsymbol{a}$.

(iv) The point $\boldsymbol{a}$ is a *boundary point* of the subset A if it is neither an interior point nor an exterior point of A.

Definition 11.8: Open, closed, and bounded subsets

Let $A \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$.

(i) The subset A is said to be *open* if A is a neighborhood of every point in A.

(ii) The subset A is said to be *closed* if its complement $A^{\mathsf{c}} = \mathbb{R}^n \setminus A$ is open.

(iii) The subset A is said to be *bounded* if there exists $M \in \mathbb{R}^+$ for which

$$A \subset \mathcal{K}(0, M),$$

i.e. if there exists a ball of finite radius M entirely containing A.

Definition 11.9: Cluster point of a subset of $\mathbb{R}^n$

Let $A \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$ and let $\boldsymbol{a} \in \mathbb{R}^n$ be an arbitrary vector. The point $\boldsymbol{a}$ is called a *cluster point* of A if

$$A \cap \mathcal{K}(\boldsymbol{a}, \delta) \setminus \{\boldsymbol{a}\} \neq \emptyset \text{ for all } \delta \in \mathbb{R}^+,$$

i.e. if every open ball centered at $\boldsymbol{a}$ always contains points in A other than $\boldsymbol{a}$.

11.1.3 Sequences in $\mathbb{R}^n$

A *sequence* in the vector space $\mathbb{R}^n$ is a function mapping the natural numbers to elements of $\mathbb{R}^n$, i.e. a function of the form

$$\mathbb{N} \to \mathbb{R}^n, \quad k \mapsto \boldsymbol{a}_k,$$

where $k \in \mathbb{N}$ and $\boldsymbol{a}_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$. We will denote sequences in $\mathbb{R}^n$ by

$$(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots) = (\boldsymbol{a}_k)_{k=1}^{\infty} = (\boldsymbol{a}_k)_k = (\boldsymbol{a}_k).$$

We will also refer to sequences in $\mathbb{R}^n$ as *vector sequences*. We will now define cluster points and limits of sequences in $\mathbb{R}^n$; these are defined analogously to the cluster points and limits of sequences of real numbers from Definitions 4.6 and 4.11.

Definition 11.10: Cluster point of a sequence in $\mathbb{R}^n$

The point $\boldsymbol{a} \in \mathbb{R}^n$ is called a *cluster point* of the sequence $(\boldsymbol{a}_k)$ if every neighborhood of $\boldsymbol{a}$ contains infinitely many elements of the sequence $(\boldsymbol{a}_k)$, i.e. if the set

$$\{k \in \mathbb{N};\, \boldsymbol{a}_k \in A\}$$

contains infinitely many elements for every neighborhood A of $\boldsymbol{a}$.

Definition 11.11: Limit of a sequence in $\mathbb{R}^n$

The point $\boldsymbol{a} \in \mathbb{R}^n$ is called the *limit* of the sequence $(\boldsymbol{a}_k)$ if every neighborhood A of the point $\boldsymbol{a}$ contains all elements of the sequence $(\boldsymbol{a}_k)$ with the possible exception of finitely many elements, i.e. if the set $\{k \in \mathbb{N};\ \boldsymbol{a}_k \neq A\}$ is finite for every neighborhood A of $\boldsymbol{a}$. In this case we write

$$\boldsymbol{a} = \lim_{k\to\infty} \boldsymbol{a}_k.$$

A sequence in $\mathbb{R}^n$ is said to be *convergent* if it has a limit; otherwise the sequence is said to be *divergent*.

Remark 11.12: Components of vector sequences

If $(\boldsymbol{a}_k)$ is a sequence of vectors in $\mathbb{R}^n$, then every element of the sequence can be written in the component form

$$\boldsymbol{a}_k = (a_{k,1}, a_{k,2}, \ldots, a_{k,n}).$$

For every $i = 1, 2, \ldots, n$ the i-th component of the sequence's elements forms a sequence of real numbers $(a_{k,i})_k$. Without proof, a sequence $(\boldsymbol{a}_k)$ in $\mathbb{R}^n$ converges if, and only if, all n sequences of its components converge, and in this case

$$\lim_{k\to\infty} \boldsymbol{a}_k = \left(\lim_{k\to\infty} a_{k,1}, \lim_{k\to\infty} a_{k,2}, \ldots, \lim_{k\to\infty} a_{k,n} \right).$$

The properties of cluster points and limits of sequences of real numbers developed in Chapter 4 apply in a generalized form to sequences in $\mathbb{R}^n$. We survey some of the more important properties below.

Proposition 11.13: Properties of sequences in $\mathbb{R}^n$

(i) If a sequence in $\mathbb{R}^n$ has a limit, it has exactly one cluster point. This is the analog of Proposition 4.13.

(ii) A *subsequence* of a sequence $(\boldsymbol{a}_k)$ in $\mathbb{R}^n$ is a sequence $(\boldsymbol{a}_{k_l})_l$, where $(k_l)_l$ is an arbitrary strictly increasing sequence of natural numbers. This is the analog of Definition 4.24.

(iii) Every cluster point of a subsequence of a given sequence is also a cluster point of the parent sequence. This is the analog of Proposition 4.26 (i).

(iv) Every subsequence of a convergent sequence is also convergent. This is the analog of Proposition 4.26 (ii).

(v) The point $\boldsymbol{a} \in \mathbb{R}^n$ is a cluster point of the sequence $(\boldsymbol{a}_k)$ if, and only if, there exists a subsequence of the sequence $(\boldsymbol{a}_k)$ that converges to $\boldsymbol{a}$. This is the analog of Proposition 4.27.

11.1.4 Scalar-valued and vector-valued functions of multiple variables

Definition 11.14: Scalar- and vector-valued multivariable function

Let $U \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$. A *scalar-valued function of n variables* is a mapping $f\colon U \to \mathbb{R}$ that associates vectors in U with real numbers in $\mathbb{R}$, i.e.

$$f\colon (x_1, x_2, \ldots, x_n) \in U \mapsto f(x_1, x_2, \ldots, x_n) \in \mathbb{R}.$$

Let $m, n \in \mathbb{N}$ be any two real numbers and let $V \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$. *A vector-valued function of n variables* is a mapping $f\colon V \to \mathbb{R}^m$ that associates vectors in V with vectors in $\mathbb{R}^m$, i.e.

$$\boldsymbol{g}\colon (x_1, x_2, \ldots, x_n) \in V \mapsto \boldsymbol{g}(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m.$$

We will also write the function value $\boldsymbol{g}(x_1, x_2, \ldots, x_n)$ in the form

$$\boldsymbol{g}(x_1, \ldots, x_n) = \big(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)\big) \in \mathbb{R}^m.$$

In this sense the vector-valued function $\boldsymbol{g}$ can be thought of as having components $g_1, g_2, \ldots, g_m\colon V \to \mathbb{R}$, which are all scalar-valued functions of n variables mapping from V to $\mathbb{R}$. We then write $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)$. The i-th scalar-valued function g_i determines the i-th component of the vector $\boldsymbol{g}(x_1, \ldots, x_n) \in \mathbb{R}^m$.

Example 11.15: A scalar-valued function of two variables

Let $n = 2$, in which case we conventionally denote the two variable components x_1 and x_2 by x and y. Consider the two-variable scalar function given by

$$f(x, y) = \frac{1}{x^2 + y^2}.$$

The function's domain is the subset

$$U = \big\{(x, y) \in \mathbb{R}^2;\, (x, y) \neq (0, 0)\big\} = \mathbb{R}^2 \setminus \{(0, 0)\},$$

and the function's image is the set of positive real numbers. The function's graph is the three-dimensional subset

$$\operatorname{graph}(f) = \left\{ \left(x, y, \frac{1}{x^2 + y^2} \right) ;\, (x, y) \in U \right\} \subset \mathbb{R}^2 \times \mathbb{R} = \mathbb{R}^3.$$

Geometrically, we can interpret f's graph as a surface in $\mathbb{R}^3$. For a given real number $x_0 \in \mathbb{R}$, the intersection of f's graph with the plane $\{x_0\} \times \mathbb{R} \times \mathbb{R}$, i.e. the plane in $\mathbb{R}^3$ parallel to the yz plane and passing through x_0, is a curve corresponding to the graph of the single-variable function $y \mapsto f(x_0, y)$. Similarly, for any real number $y_0 \in \mathbb{R}$, the intersection of f's graph with the place $\mathbb{R} \times \{y_0\} \times \mathbb{R}$ is a curve corresponding to the graph of the single-variable function $x \mapsto f(x, y_0)$. The union of these two sets of curves for all $x_0, y_0 \in \mathbb{R}$ is the graph of f.

Remark 11.16: Contour lines

Contour lines are a useful tool for visualizing the graphs of scalar functions of two variables, such as the function f in the Example 11.15. For a given function $f\colon \mathbb{R}^2 \to \mathbb{R}$, contour lines are curves in the plane $\mathbb{R}^2$ given by the solution sets $\{(x,y) \in \mathbb{R}^2;\, f(x,y) = z_0\}$ for a given constant $z_0 \in \mathbb{R}$.

If a function's graph represents the surface of some terrain in three-dimensional space (for example in the context of cartography), the function's contour line at z_0 is the set of all points at the height z_0.

Topic 11.2: Combinations of multivariable functions

Let $n, m \in \mathbb{N}$ be any two natural numbers, let $U, V, Z \subset \mathbb{R}^n$ be three arbitrary subsets of $\mathbb{R}^n$, let $\boldsymbol{g}\colon U \to \mathbb{R}^m$ and $\boldsymbol{h}\colon V \to \mathbb{R}^m$ be two vector-valued functions of n variables, and let $\alpha\colon Z \to \mathbb{R}$ be a scalar-valued function of n variables. In this case

(i) The *sum* of the functions $\boldsymbol{g}$ and $\boldsymbol{h}$ is the function $\boldsymbol{g}+\boldsymbol{h}\colon U \cap V \to \mathbb{R}^m$ given by

$$(\boldsymbol{g}+\boldsymbol{h})(\boldsymbol{v}) = \boldsymbol{g}(\boldsymbol{v}) + \boldsymbol{h}(\boldsymbol{v}), \quad \boldsymbol{v} \in U \cap V.$$

This definition also holds for scalar-valued functions; one simply takes $m = 1$, in which case $\boldsymbol{g}$ and $\boldsymbol{h}$ reduce to scalar-valued functions of n variables.

(ii) The *product* of the vector-valued function $\boldsymbol{g}$ with the scalar-valued function α is the vector-valued function $\alpha\boldsymbol{g}\colon Z \cap U \to \mathbb{R}^m$ given by

$$(\alpha\boldsymbol{g})(\boldsymbol{v}) = \alpha(\boldsymbol{v})\boldsymbol{g}(\boldsymbol{v}), \quad \boldsymbol{v} \in U \cap Z.$$

(iii) The *pointwise scalar product* of the functions $\boldsymbol{g}$ and $\boldsymbol{h}$ is the function $\boldsymbol{g}\cdot\boldsymbol{h}\colon U \cap V \to \mathbb{R}$ given by

$$(\boldsymbol{g}\cdot\boldsymbol{h})(\boldsymbol{v}) = \boldsymbol{g}(\boldsymbol{v}) \cdot \boldsymbol{h}(\boldsymbol{v}), \quad \boldsymbol{v} \in U \cap Z.$$

It is also possible to extend function composition to multivariable functions. Namely, let $p \in \mathbb{N}$ be an arbitrary natural number, let $W \subset \mathbb{R}^m$ be any subset of $\mathbb{R}^m$, and let $\boldsymbol{f}\colon W \to \mathbb{R}^p$ be a vector-valued function of m variables. In this case the *composition* of the functions $\boldsymbol{g}$ and $\boldsymbol{f}$ is the function $\boldsymbol{f}\circ\boldsymbol{g}\colon \boldsymbol{g}^{-1}(W) \to \mathbb{R}^p$ given by

$$(\boldsymbol{f}\circ\boldsymbol{g})(\boldsymbol{v}) = \boldsymbol{f}\big(\boldsymbol{g}(\boldsymbol{v})\big), \quad \boldsymbol{v} \in \boldsymbol{g}^{-1}(W).$$

11.1.5 Limits and continuity of multivariable functions

Definition 11.17: Limit of a multivariable function

Let $n, m \in \mathbb{N}$ be any two natural numbers, let $\boldsymbol{g}\colon U \to \mathbb{R}^m$ be a vector-valued function of n variables defined on the subset $U \subset \mathbb{R}^n$, and let $\boldsymbol{a}$ be a cluster point of U. In this case the vector $\boldsymbol{L} \in \mathbb{R}^m$ is called the *limit* of the function $\boldsymbol{g}$ at the point $\boldsymbol{a}$ if for all $\epsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ for which

$$\boldsymbol{g}(\boldsymbol{v}) \in \mathcal{K}(\boldsymbol{L}, \epsilon) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a}, \delta) \cap U \setminus \{\boldsymbol{a}\}.$$

In this case we write

$$\boldsymbol{L} = \lim_{\boldsymbol{v} \to \boldsymbol{a}} \boldsymbol{g}(\boldsymbol{v}).$$

Note that the open ball $\mathcal{K}(\boldsymbol{L}, \epsilon)$ is simply an n-dimensional generalization of the open interval $(L - \epsilon, L + \epsilon)$ used in the definition of the limit of a scalar function in Definition 6.16.

Remark 11.18: The limit must hold component-wise

Let $n, m \in \mathbb{N}$ be two arbitrary natural numbers, let $\boldsymbol{g} \colon U \to \mathbb{R}^m$ be a vector-valued function of n variables defined on the subset $U \subset \mathbb{R}^n$, and let $\boldsymbol{a}$ be a cluster point of U. In this case $\boldsymbol{L} \in \mathbb{R}^m$ is the limit of the function $\boldsymbol{g}$ at the point $\boldsymbol{a}$ if, and only if,

$$L_i = \lim_{\boldsymbol{v} \to \boldsymbol{a}} g_i(\boldsymbol{v}) \text{ for all } i = 1, 2, \ldots, m,$$

where L_i are the components of the vector $\boldsymbol{L}$ and g_i are the component functions of the function $\boldsymbol{g}$.

The properties of limits of scalar functions developed in Section 6.1 apply in a generalized form to limits of vector-valued functions. These properties are proved in an analogous manner to the one-dimensional counterparts, so we will not prove them here.

Definition 11.19: Continuity of multivariable functions

Let $n, m \in \mathbb{N}$ be two arbitrary natural numbers, let $\boldsymbol{g} \colon U \to \mathbb{R}^m$ be a vector-valued function of n variables defined on the subset $U \subset \mathbb{R}^n$, and define the point $\boldsymbol{a} \in U$. In this case the function $\boldsymbol{g}$ is said to be *continuous at* $\boldsymbol{a}$ if for all $\epsilon \in \mathbb{R}^+$ there exists $\delta \in \mathbb{R}^+$ for which

$$\boldsymbol{g}(\boldsymbol{v}) \in \mathcal{K}\big(\boldsymbol{g}(\boldsymbol{a}), \epsilon\big) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a}, \delta) \cap U.$$

A multivariable function is said to be *continuous* if it is continuous at all points in its domain.

The set of all continuous functions $\boldsymbol{g} \colon U \to \mathbb{R}^m$ is denoted by $C(U, \mathbb{R}^m)$ or $C^0(U, \mathbb{R}^m)$. In this notation, $C(U) = C^0(U)$ are shorthand for $C(U, \mathbb{R})$, the set of all continuous scalar-valued functions mapping from U to $\mathbb{R}$.

Loosely, Definition 11.19 states that the function $\boldsymbol{g}$ is continuous at $\boldsymbol{a}$ if the function values $\boldsymbol{g}(\boldsymbol{v})$ come arbitrarily close to $\boldsymbol{g}(\boldsymbol{a})$ whenever $\boldsymbol{v}$ is arbitrarily close to $\boldsymbol{a}$. Note that the open balls $\mathcal{K}\big(\boldsymbol{g}(\boldsymbol{a}), \epsilon\big)$ and $\mathcal{K}(\boldsymbol{a}, \delta)$ are just the n-dimensional generalizations of the open intervals $\big(f(a) - \epsilon, f(a) + \epsilon\big)$ and $(a - \delta, a + \delta)$ used in the definition of the continuity of a scalar function in Definition 6.26. Like in Definition 6.26, the actual function value $\boldsymbol{g}(\boldsymbol{a})$ has no influence on the limit of $\boldsymbol{g}$ at $\boldsymbol{a}$.

Remark 11.20: Continuity must hold component-wise

Let $n, m \in \mathbb{N}$ be any two natural numbers, let $\boldsymbol{g} \colon U \to \mathbb{R}^m$ be a vector-valued function of n variables defined on the subset $U \subset \mathbb{R}^n$, and define the point $\boldsymbol{a} \in U$.

The function $\boldsymbol{g}$ is continuous at $\boldsymbol{a}$ if, and only if, each of its component functions

$$g_1, g_2, \ldots, g_m \colon U \to \mathbb{R}$$

are also continuous at $\boldsymbol{a}$.

Properties analogous to those developed for continuous scalar functions in Section 6.2 apply to continuous vector-valued functions. We briefly state two of these below.

Proposition 11.21: Properties of continuous multivariable functions

Let $n, m \in \mathbb{N}$ be any two natural numbers, let $\boldsymbol{g} \colon U \to \mathbb{R}^m$ be a vector-valued function of n variables defined on the subset $U \subset \mathbb{R}^n$, and define the point $\boldsymbol{a} \in U$. In this case

(i) If $\boldsymbol{a}$ is a cluster point of U, the multivariable function $\boldsymbol{g}$ is continuous at $\boldsymbol{a}$ if, and only if, the function value at $\boldsymbol{a}$ equals the function's limit at $\boldsymbol{a}$, i.e. if

$$\boldsymbol{g}(\boldsymbol{a}) = \lim_{\boldsymbol{v} \to \boldsymbol{a}} \boldsymbol{g}(\boldsymbol{v}).$$

This is the analog of Proposition 6.28 (ii).

(ii) The multivariable function $\boldsymbol{g}$ is continuous at $\boldsymbol{a}$ if, and only if, for every vector sequence $(\boldsymbol{a}_k)$ in U that converges to $\boldsymbol{a}$, the corresponding sequence of function values $\big(\boldsymbol{g}(\boldsymbol{a}_k)\big)_k$ converges to the function value $\boldsymbol{g}(\boldsymbol{a})$. In this case the order of limit and function evaluation is interchangeable, i.e.

$$\lim_{k \to \infty} \boldsymbol{g}(\boldsymbol{a}_k) = \boldsymbol{g}\left(\lim_{k \to \infty} \boldsymbol{a}_k \right).$$

This is the analog of Proposition 6.28 and Remark 6.29.

The following properties are the higher-dimensional generalizations of Corollaries 6.32 and 6.33 in Section 6.2.

Topic 11.3: Combinations of continuous multivariable functions

Let $n, m \in \mathbb{N}$ be arbitrary natural numbers, let $U, V, Z \subset \mathbb{R}^n$ be three arbitrary subsets of $\mathbb{R}^n$, let $\boldsymbol{g} \colon U \to \mathbb{R}^m$ be a vector-valued function of n variables let $\boldsymbol{h} \colon V \to \mathbb{R}^m$ be a vector-valued function of n variables, and let $\alpha \colon Z \to \mathbb{R}$ be a scalar-valued function of n variables. In this case

(i) If the functions $\boldsymbol{g}$ and $\boldsymbol{h}$ are continuous, then their sum $\boldsymbol{g} + \boldsymbol{h}$ is also continuous.

(ii) If the functions $\boldsymbol{g}$ and α are continuous, then their product $\alpha \boldsymbol{g}$ is also continuous.

(iii) If the functions $\boldsymbol{g}$ and $\boldsymbol{h}$ are continuous, then their pointwise scalar product $\boldsymbol{g} \cdot \boldsymbol{h}$ is also continuous.

(iv) Let $p \in \mathbb{N}$ be an arbitrary natural number, let $W \subset \mathbb{R}^m$ be an arbitrary subset of $\mathbb{R}^m$, and let $\boldsymbol{f} \colon W \to \mathbb{R}^p$ be a vector-valued function of m variables. If both $\boldsymbol{f}$ and $\boldsymbol{g}$ continuous, then their composition $\boldsymbol{f} \circ \boldsymbol{g}$ is also continuous.

Example 11.22: Projection functions

For all $i = 1, 2, \ldots, n$, a *projection* is a scalar-valued function $\mathrm{pr}_i \colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\mathrm{pr}_i(x_1, x_2, \ldots, x_n) = x_i.$$

A projection function thus returns a single (scalar) component of its input vector. Projections are continuous functions. To show this, we first recall from Proposition 11.3 (iii) the inequality

$$|u_i| \leq \|\boldsymbol{u}\| \text{ for all } \boldsymbol{u} \in \mathbb{R}^n \text{ and all } i = 1, 2, \ldots, n.$$

For any two vectors $\boldsymbol{v}, \boldsymbol{a} \in \mathbb{R}^n$, we then apply this inequality to the vector $\boldsymbol{v} - \boldsymbol{a}$ to get

$$|v_i - a_i| \leq \|\boldsymbol{v} - \boldsymbol{a}\|.$$

This inequality implies that $|v_i - a_i| < \delta$ if $\|\boldsymbol{v} - \boldsymbol{a}\| < \delta$; setting $\delta = \epsilon$ in the definition of continuity then implies that projections are continuous functions.

Example 11.23: Multivariable polynomials

All polynomial scalar-valued functions of n variables are continuous functions, since polynomials are the sums and products of projection functions. Since projection functions are continuous and the sums and products of continuous multivariable functions are again continuous, n-variable scalar-valued polynomials are continuous. For example, the following two polynomials are continuous:

$$f(x, y, z) = xyz^3,$$
$$g(x, y, z) = xy + 3yz^3 + 12yz^4 - xz^5.$$

11.1.6 Path-connectedness and compactness

In this section we introduce the important topological concepts of path-connectedness and compactness. Unlike higher-dimensional neighborhoods, openness, limits, and continuity, which we encountered in previous chapters, path-connectedness and compactness are new topics in this book.

Definition 11.24: Paths and path-connectedness

(i) Let $A \subset \mathbb{R}^n$ be a subset of $\mathbb{R}^n$ and define the points $\boldsymbol{v}, \boldsymbol{w} \in A$. A *path in the subset* A from the point $\boldsymbol{v}$ to the point $\boldsymbol{w}$ is a continuous vector-valued function

$$\boldsymbol{\gamma} \colon [0, 1] \to \mathbb{R}^n$$

for which $\boldsymbol{\gamma}(0) = \boldsymbol{v}$, $\boldsymbol{\gamma}(1) = \boldsymbol{w}$, and $\boldsymbol{\gamma}\big([0, 1]\big) \subset A$. A path can be interpreted is a continuous "walk" through the set A from the *initial point* $\boldsymbol{v}$ to the *terminal point* $\boldsymbol{w}$ as the function argument runs from 0 to 1.

Note that the term "path" refers to the function $\boldsymbol{\gamma}$, and *not* to function's image $\boldsymbol{\gamma}\big([0, 1]\big)$, i.e. the n-dimensional curve in A consisting of points along the path.

(ii) A subset $A \subset \mathbb{R}^n$ is *path-connected* if for any two points $\boldsymbol{v}, \boldsymbol{w} \in A$ there exists a path in A from $\boldsymbol{v}$ to $\boldsymbol{w}$.

Example 11.25: Real intervals are path-connected

All intervals on the real line are path-connected. In fact, intervals are the *only* path-connected subsets of the real numbers. By the definition of paths and continuity, for any subset of $\mathbb{R}$ other than an interval (which would necessarily contains "gaps" in the real numbers), it would be impossible to construct a path connecting any two points separated by a gap.

The following proposition generalizes Theorem 6.35.

Proposition 11.26: Zeros of continuous functions on path-connected subsets

Let $f\colon U \to \mathbb{R}$ be a continuous scalar-valued function defined on a path-connected subset $U \subset \mathbb{R}^n$. If $f(\boldsymbol{v}) \cdot f(\boldsymbol{w}) < 0$ (i.e. if $f(\boldsymbol{v})$ and $f(\boldsymbol{w})$ have opposite signs) for any two points $\boldsymbol{v}, \boldsymbol{w} \in U$, then there exists a point $\boldsymbol{\xi} \in U$ for which $f(\boldsymbol{\xi}) = 0$.

Proof. The subset U is path-connected, so there exists a path $\boldsymbol{\gamma}$ in U from the point $\boldsymbol{v}$ to the point $\boldsymbol{w}$. Both f and $\boldsymbol{\gamma}$ are continuous, so the composite function $f \circ \boldsymbol{\gamma}\colon [0,1] \to \mathbb{R}$ is also continuous. By the proposition's assumptions, the function $f \circ \boldsymbol{\gamma}$ obeys

$$(f \circ \boldsymbol{\gamma})(0) \cdot (f \circ \boldsymbol{\gamma})(1) = f(\boldsymbol{v}) f(\boldsymbol{v}) < 0.$$

And since $f \circ \boldsymbol{\gamma}$ is a continuous, scalar function defined on a closed interval that takes on oppositely-signed values at the interval endpoints, by Theorem 6.35 there exists a point $c \in [0,1]$ for which $f\big(\boldsymbol{\gamma}(c)\big) = 0$. The point $\boldsymbol{\xi} = \boldsymbol{\gamma}(c) \in \mathbb{R}^n$ thus obeys $f(\boldsymbol{\xi}) = 0$, completing the proof. □

Proposition 11.27: The image of a continuous function on a path-connected domain is path-connected

Let $m, n \in \mathbb{N}$ be two natural numbers and let $\boldsymbol{g}\colon U \to \mathbb{R}^m$ be a continuous vector-valued function defined on a path-connected subset $U \subset \mathbb{R}^n$. In this case the function's image $\boldsymbol{g}(U)$ is also path-connected.

Proof. Consider two arbitrary points $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{g}(U)$ and let $\boldsymbol{v}, \boldsymbol{w} \in U$ denote two points in U for which $\boldsymbol{a} = \boldsymbol{g}(\boldsymbol{v})$ and $\boldsymbol{b} = \boldsymbol{g}(\boldsymbol{w})$, respectively. The subset U is path-connected, so there exists a path $\boldsymbol{\gamma}$ in U from $\boldsymbol{v}$ to $\boldsymbol{w}$. The functions $\boldsymbol{g}$ and $\boldsymbol{\gamma}$ are continuous, so the composite function $\boldsymbol{g} \circ \boldsymbol{\gamma}\colon [0,1] \to \boldsymbol{g}(U) \subset \mathbb{R}^m$ is also continuous.

Because $\boldsymbol{g} \circ \boldsymbol{\gamma}$ is continuous and necessarily only takes on values in $\boldsymbol{g}(U)$, $\boldsymbol{g} \circ \boldsymbol{\gamma}$ is a path in the set $\boldsymbol{g}(U)$ from the point $\boldsymbol{a}$ to the point $\boldsymbol{b}$; since the points $\boldsymbol{a}$ and $\boldsymbol{b}$ were arbitrary, there thus exists a path in $\boldsymbol{g}(U)$ between any two points in $\boldsymbol{g}(U)$, so the subset $\boldsymbol{g}(U)$ is path-connected. □

Definition 11.28: Compactness

A subset $A \subset \mathbb{R}^n$ is said to be *compact* if every sequence of points in A has at least one cluster point in A.

In preparation for the coming proposition, we will first extend the Bolzano-Weierstrass theorem (Theorem 4.9) from the real numbers $\mathbb{R}$ to the n-dimensional Euclidean space $\mathbb{R}^n$.

Lemma 11.29: Bounded sequences have a cluster point

Every bounded sequence in $\mathbb{R}^n$ has at least one cluster point.

Proof. Let $(\boldsymbol{a}_k)$ be an arbitrary sequence of vectors in $\mathbb{R}^n$. We have already proved the $n = 1$ case in Theorem 4.9, and can in fact prove the n-dimensional case simply by applying the $n = 1$ case component-wise to every component of $(\boldsymbol{a}_k)$. Namely, n such applications would produce a subsequence of $(\boldsymbol{a}_k)$ that converges in each of its components, and is thus itself convergent by Remark 11.12. There thus exists a convergent subsequence of $(\boldsymbol{a}_k)$, and so by Proposition 11.13 (v) this subsequence's limit is a cluster point of $(\boldsymbol{a}_k)$. □

We may now state the following important characterization of compact sets.

Proposition 11.30: Compact subsets are closed and bounded

A subset $A \subset \mathbb{R}^n$ is compact if, and only if, it is closed and bounded.

Proof. ($\Rightarrow$) Let $A \subset \mathbb{R}^n$ be a compact subset. We will first show, by contradiction, that A is bounded. If A were not bounded, then for all $k \in \mathbb{N}$ there would exist a point $\boldsymbol{a}_k \in A$ for which $\|\boldsymbol{a}_k\| > k$. Any thus-defined sequence $(\boldsymbol{a}_k)$ would not have any cluster points, which contradicts the assumption that A is compact.

We will now show, again by contradiction, that A is closed. Assume that A is not closed, in which case $A^{\mathsf{c}} = \mathbb{R}^n \setminus A$ is not open, meaning there would exist a point $\boldsymbol{b} \in \mathbb{R}^n \setminus A$ that is not an interior point of $\mathbb{R}^n \setminus A$. Thus $\mathbb{R}^n \setminus A$ is not a neighborhood of $\boldsymbol{b}$, which implies that

$$\mathcal{K}(\boldsymbol{b}, 1/k) \cap A \neq \emptyset \text{ for all } k \in \mathbb{N},$$

and so there would exist a corresponding point $\boldsymbol{b}_k \in \mathbb{R}^n$ for which

$$\boldsymbol{b}_k \in \mathcal{K}(\boldsymbol{b}, 1/k) \cap A.$$

Any thus-constructed sequence $(\boldsymbol{b}_k)$ would consist only of points in A, but would converge to $\boldsymbol{b}$, which is not in A. Thus the sequence $(\boldsymbol{b}_k)$ of points in A would not have a cluster point in A, which contradicts the assumption that A is compact. It follows that A being compact implies that A is closed and bounded.

($\Leftarrow$) Assume that $A \subset \mathbb{R}^n$ is closed and bounded, and let $(\boldsymbol{a}_k)$ be an arbitrary sequence of vectors in A. Because A is bounded, the sequence $(\boldsymbol{a}_k)$ is necessarily also bounded, and thus $(\boldsymbol{a}_k)$ has at least one cluster point $\boldsymbol{a} \in \mathbb{R}^n$ by Lemma 11.29. Because A is closed, the cluster point $\boldsymbol{a}$ must lie in A, and since the sequence $(\boldsymbol{a}_k)$ was arbitrary, every sequence of vectors in A has at least one cluster point in A, and so A is compact. □

Example 11.31: Closed, bounded intervals are compact

An interval on the real line is compact if, and only if, it is closed and bounded; Neither open, nonempty intervals nor unbounded intervals are compact.

Proposition 11.32: The image of a continuous function on a compact domain is compact

Let $m, n \in \mathbb{N}$ be two natural numbers and let $\boldsymbol{g}\colon U \to \mathbb{R}^m$ be a continuous vector-valued function defined on a compact subset $U \subset \mathbb{R}^n$. In this case the function's image $\boldsymbol{g}(U)$ is also compact.

Proof. We aim to show that that any sequence $(\boldsymbol{b}_k)$ of points in the subset $\boldsymbol{g}(U) \subset \mathbb{R}^m$ has at least one cluster point in $\boldsymbol{g}(U)$. To show this, let $(\boldsymbol{a}_k)$ be a sequence of points in U for which

$$\boldsymbol{b}_k = \boldsymbol{g}(\boldsymbol{a}_k) \text{ for all } k \in \mathbb{N}.$$

The set U is compact, so the sequence $(\boldsymbol{a}_k)$ has a cluster point $\boldsymbol{a} \in U$; by Proposition 11.13 (v) there then exists a subsequence $(\boldsymbol{a}_{k_l})_l$ of the sequence $(\boldsymbol{a}_k)$ that converges to $\boldsymbol{a}$. Because the function $\boldsymbol{g}$ is continuous, by Proposition 11.21 the subsequence $(\boldsymbol{b}_{k_l})_l = \big(\boldsymbol{g}(\boldsymbol{a}_{k_l})\big)_l$ converges to the point $\boldsymbol{g}(\boldsymbol{a}) \in \boldsymbol{g}(U)$, and so the point $\boldsymbol{g}(\boldsymbol{a})$ is a cluster point of the sequence $(\boldsymbol{b}_k)$. □

11.2 Partial and total differentiability

This section introduces some generalizations of the scalar derivative from Chapter 7 that extend the differentiation operation to multivariable functions.

11.2.1 Partial derivatives

Topic 11.4: Motivation for the partial derivative

Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of n variables defined on the open subset $U \subset \mathbb{R}^n$, let $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ be an interior point of U, and let $k \in 1, 2, \ldots, n$ denote any index from 1 to n.

If we take all of the function f's n variables, with the exception of the k-th variable x_k, to be constant, the result is a single-variable function of the form

$$x_k \mapsto f(a_1, \ldots, a_{k-1}, x_k, a_{k+1}, \ldots, a_n).$$

The graph of this single-variable function is a curve along the graph of the original n-variable function f that progresses in the direction of the k-th coordinate and passes through the point $\big(\boldsymbol{a}, f(\boldsymbol{a})\big) \in \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$; the derivative of this single-variable function, if it exists, is the slope of the tangent line to this curve in the direction of the k-th coordinate.

Definition 11.33: Partial derivative and gradient a scalar-valued function

Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of $n \in \mathbb{N}$ variables defined on the open subset $U \subset \mathbb{R}^n$, let $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ be an interior point of U, and let

$k \in 1, 2, \ldots, n$ be an arbitrary index value.

(i) The *partial derivative of the function f with respect to the variable x_k at the point $\boldsymbol{a}$* is denoted by $\frac{\partial f}{\partial x_k}(\boldsymbol{a})$ and defined as

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = \lim_{h \to 0} \left[\frac{f(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n)}{h} - \frac{f(a_1, \ldots, a_{k-1}, a_k, a_{k+1}, \ldots, a_n)}{h} \right],$$

assuming the limit exists. If the limit does exist, the function f is said to be *partially differentiable with respect to the variable x_k at the point $\boldsymbol{a}$*.

Note that when $n = 1$ the partial derivative reduces to the familiar derivative of a single-variable scalar function from Definition 7.1.

(ii) The function f is said to be *partially differentiable at the point $\boldsymbol{a}$* if it is partially differentiable at $\boldsymbol{a}$ with respect to all of its variables $x_1, x_2, \ldots, x_n$.

(iii) If the function f is partially differentiable at $\boldsymbol{a}$, then the *gradient of the function f at the point $\boldsymbol{a}$* is defined as the vector

$$(\boldsymbol{\nabla} f)(\boldsymbol{a}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \frac{\partial f}{\partial x_2}(\boldsymbol{a}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{a}) \right] \in \mathbb{R}^n.$$

(iv) The function f is said to be *partially differentiable with respect to the variable x_k* if it is partially differentiable with respect to x_k at every point in its domain. In this case the *partial derivative of f with respect to the variable x_k* is defined as the scalar-valued function

$$\frac{\partial f}{\partial x_k} \colon U \to \mathbb{R}, \quad \boldsymbol{x} \mapsto \frac{\partial f}{\partial x_k}(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in U.$$

(v) The function f is said to be *partially differentiable* if it is partially differentiable at every point in its domain. In this case the *gradient* of the function f is defined as the vector-valued function

$$(\boldsymbol{\nabla} f) \colon U \to \mathbb{R}^n, \quad \boldsymbol{x} \mapsto (\boldsymbol{\nabla} f)(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in U.$$

Definition 11.34: Continuous differentiability of a scalar-valued function

A scalar-valued function $f \colon U \to \mathbb{R}$ of $n \in \mathbb{N}$ variables defined on the open subset $U \subset \mathbb{R}^n$ is said to be *continuously-differentiable* if f is continuous and partially differentiable and if all of f's partial derivatives $\frac{\partial f}{\partial x_k} \colon U \to \mathbb{R}, j = 1, 2, \ldots, n$ are themselves continuous functions.

We denote the set of all continuously-differentiable scalar-valued functions $f \colon U \to \mathbb{R}$ by $C^1(U)$.

We now define an alternative and more compact notation for the partial derivative in terms of unit vectors, with which you will gain more experience in a course on linear algebra.

Remark 11.35: Alternative notation for the partial derivative

Let $\hat{\mathbf{e}}_k$ denote the vector of unit norm whose k-th component is 1 and whose j-th component is zero for all $j \in \{1, 2, \ldots, n\} \setminus \{k\}$; some example unit vectors are

$$\begin{aligned}\hat{\mathbf{e}}_1 &= (1, \underbrace{0, 0, \ldots, 0}_{n-1 \text{ terms}}) \\ \hat{\mathbf{e}}_k &= (\underbrace{0, \ldots, 0}_{k-1 \text{ terms}}, 1, \underbrace{0, \ldots, 0}_{n-k \text{ terms}}) \\ \hat{\mathbf{e}}_n &= (\underbrace{0, 0, \ldots, 0}_{n-1 \text{ terms}}, 1).\end{aligned}$$

Using unit vector notation, we can write the partial derivative of a scalar-valued function $f : U \to \mathbb{R}^n$ at some point $\boldsymbol{a} \in U$ in the more compact form

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = \lim_{h \to 0} \frac{f(\boldsymbol{a} + h\hat{\mathbf{e}}_k) - f(\boldsymbol{a})}{h}.$$

Finally, here are some more notations regularly used for the partial derivative:

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = \frac{\partial}{\partial x_k}(f)(\boldsymbol{a}) = f_{x_k}(\boldsymbol{a}) = (D_k f)(\boldsymbol{a}) = (D_{\hat{\mathbf{e}}_k} f)(\boldsymbol{a}).$$

We have defined the partial derivative of a multivariable scalar function as the derivative of the single-variable function obtained by restricting the multivariable function to vary only along a given coordinate axis $\hat{\mathbf{e}}_k$. More generally, is quite reasonable to consider restricting a multivariable function to vary along a line in the direction of an arbitrary vector $\boldsymbol{u} \in \mathbb{R}^n$. This idea motivates the concept of the *directional derivative*, which we define below.

Definition 11.36: Directional derivatives

Let $f : U \to \mathbb{R}$ be a scalar-valued function of $n \in \mathbb{N}$ variables defined on the open subset $U \subset \mathbb{R}^n$, let $\boldsymbol{a} \in U$ be an interior point of U, and let $\boldsymbol{u} \in \mathbb{R}^n$ be an arbitrary vector. In this case the *directional derivative of the function f in the direction of $\boldsymbol{u}$ at the point $\boldsymbol{a}$* is denoted by $(D_{\boldsymbol{u}} f)(\boldsymbol{a})$ and defined as

$$(D_{\boldsymbol{u}} f)(\boldsymbol{a}) = \lim_{h \to 0} \frac{f(\boldsymbol{a} + h\boldsymbol{u}) - f(\boldsymbol{a})}{h} \in \mathbb{R},$$

assuming the limit exists.

Having introduced the differentiability for scalar-valued functions, we now extend some of these concepts to vector-valued functions.

Definition 11.37: Partial differentiability of vector-valued functions

Let $m, n \in \mathbb{N}$ be two arbitrary natural numbers, let $U \subset \mathbb{R}^n$ be an open subset of $\mathbb{R}^n$, and let $\boldsymbol{g} : U \to \mathbb{R}^m$ be a vector-valued function with scalar components $\boldsymbol{g} = (g_1, \ldots, g_m)$. In this case:

(i) The vector-valued function $\boldsymbol{g}$ is said to be *partially differentiable* if each of its scalar components $g_1, \ldots, g_m$ is partially differentiable. In this case we define

the *partial derivative of* $\boldsymbol{g}$ *with respect to the variable* x_k as the vector-valued function

$$\frac{\partial \boldsymbol{g}}{\partial x_k} = \left(\frac{\partial g_1}{\partial x_k}, \ldots, \frac{\partial g_m}{\partial x_k} \right) : U \to \mathbb{R}^m, \quad \boldsymbol{x} \mapsto \left[\frac{\partial g_1}{\partial x_k}(\boldsymbol{x}), \ldots, \frac{\partial g_m}{\partial x_k}(\boldsymbol{x}) \right].$$

(ii) The vector-valued function $\boldsymbol{g}$ is said to be *continuously-differentiable* if each of its scalar components $g_1, \ldots, g_m$ is continuously-differentiable.

(iii) We denote the set of all continuously-differentiable vector-valued functions $f : U \to \mathbb{R}^m$ by $C^1(U, \mathbb{R}^m)$.

Example 11.38: Partial derivatives of a polynomial

Let f be the scalar-valued polynomial function of two variables given by

$$f(x, y) = x^2 + 2xy^3 + 4x.$$

The function f's partial derivatives are then

$$\frac{\partial f}{\partial x}(x, y) = 2x + 2xy^3 + 4 \quad \text{and} \quad \frac{\partial f}{\partial y}(x, y) = 6xy^2.$$

11.2.2 The tangent hyperplane

Topic 11.5: Three-dimensional motivation for the tangent hyperplane

Let $U \subset R^2$ be an open subset of $\mathbb{R}^2$, let $f : U \to \mathbb{R}$ be a continuously-differentiable function of two variables, let $(x_0, y_0) \in U$ be an arbitrary point in U, and define the point $z_0 = f(x_0, y_0)$. Using f's two partial derivatives, it is possible to compute the equation of the tangent plane to the graph of f at any point (x_0, y_0, z_0) on f's graph.

Consider taking the intersection of f's graph with the planes $\mathbb{R} \times \{y_0\} \times \mathbb{R}$ and $\{x_0\} \times \mathbb{R} \times \mathbb{R}$, i.e. the three-dimensional planes of constant y value $y = y_0$ and constant x value $x = x_0$, respectively. The intersection of f's graph with these two planes produces two curves whose slopes at the point (x_0, y_0, z_0) are given by the partial derivatives $f_x(x_0, y_0)$ and $f_y(x_0, y_0)$, respectively.

The tangent vectors to these two curves at the point (x_0, y_0, z_0) are thus

$$\boldsymbol{s}_1 = \left(1, 0, \frac{\partial f}{\partial x}(x_0, y_0) \right) \quad \text{and} \quad \boldsymbol{s}_2 = \left(0, 1, \frac{\partial f}{\partial y}(x_0, y_0) \right).$$

The normal to the tangent plane at (x_0, y_0, z_0) must be perpendicular to both of these two tangent vectors, and we can construct such a normal vector using the cross product

$$\boldsymbol{s}_2 \times \boldsymbol{s}_1 = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, -1 \right).$$

Having found the normal to the tangent plane, the point normal form (see Definition 3.9 for review) of the equation of the plane tangent to f's graph at the point

(x_0, y_0, z_0) is

$$\frac{\partial f}{\partial x}(x_0, y_0) \cdot (x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0) \cdot (y - y_0) - (z - z_0) = 0.$$

The line normal to the graph of f at the point (x_0, y_0, z_0) is given by the equation

$$\frac{x - x_0}{f_x(x_0, y_0)} = \frac{y - y_0}{f_y(x_0, y_0)} = \frac{z - z_0}{-1};$$

this equation is of course defined only as long as $f_x(x_0, y_0) \neq 0$ and $f_y(x_0, y_0) \neq 0$.

Example 11.39: Tangent plane to the graph of a two-variable function

Let us determine the equations of the three-dimensional tangent plane and normal line to the graph of the two-variable function

$$f\colon (x, y) \mapsto x^2 + 2xy^3 + 4x$$

at the point $(x_0, y_0) = (1, 0)$. We have already computed the function's partial derivatives in Example 11.38, these are given by

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) &= 2x + 2y^3 + 4 \implies f_x(1, 0) = 6, \\ \frac{\partial f}{\partial y}(x, y) &= 6xy^2 \qquad\quad \implies f_y(1, 0) = 0. \end{aligned}$$

Noting that $z_0 = f(1, 0) = 5$, the point-normal equation of the tangent plane to f's graph at the point $(1, 0, 5)$ is then

$$6(x - 1) + 0(y - 0) - (z - 5) = 0 \implies 6x - z = 1,$$

while the equation of the normal to this plane at $(1, 0, 5)$ is

$$\frac{x - 1}{6} = \frac{z - 5}{-1}; \quad y = 0.$$

A tangent hyperplane is the higher-dimensional generalization of the tangent plane to the graph of a two-variable function, and is defined for functions of n variables.

Topic 11.6: Tangent hyperplanes

Define $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be an open subset, let $f\colon U \to \mathbb{R}$ be a continuously-differentiable function of n variables, let $\boldsymbol{a} \in U$ be an arbitrary point in U, and define the point $z = f(\boldsymbol{a}) \in \mathbb{R}$. Finally, for every $k = 1, 2, \ldots, n$, let

$$\boldsymbol{s}_k = \left(\underbrace{0, \ldots, 0}_{k-1 \text{ terms}}, 1, \underbrace{0, \ldots, 0}_{n-k \text{ terms}}, \frac{\partial f}{\partial x_k}(\boldsymbol{a}) \right) = \left(\hat{\mathbf{e}}_k, \frac{\partial f}{\partial x_k}(\boldsymbol{a}) \right) \in \mathbb{R}^{n+1}$$

denote the tangent to the graph of f in the direction of the k-th coordinate axis at the point $\big(\boldsymbol{a}, f(\boldsymbol{a})\big) \in \mathbb{R}^{n+1}$.

In this case the *normal vector to the tangent hyperplane* of the graph of f at the point $\big(\boldsymbol{a}, f(\boldsymbol{a})\big)$ is the vector

$$\left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \dots, \frac{\partial f}{\partial x_n}(\boldsymbol{a}), -1\right) = \big((\boldsymbol{\nabla} f)(\boldsymbol{a}), -1\big) \in \mathbb{R}^{n+1};$$

this vector is by construction perpendicular to each of the tangent vectors $\boldsymbol{s}_k$, since

$$\left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \dots, \frac{\partial f}{\partial x_n}(\boldsymbol{a}), -1\right) \cdot \left(\hat{\mathbf{e}}_k, \frac{\partial f}{\partial x_k}(\boldsymbol{a})\right) = \frac{\partial f}{\partial x_k}(\boldsymbol{a}) - \frac{\partial f}{\partial x_k}(\boldsymbol{a}) = 0.$$

The *tangent hyperplane* to the graph of f at the point $\big(\boldsymbol{a}, f(\boldsymbol{a})\big)$ is the hyperplane in $\mathbb{R}^{n+1}$ consisting of all vectors $\boldsymbol{r} = (x_1, \dots, x_n, z) \in \mathbb{R}^{n+1}$ for which

$$\big((\boldsymbol{\nabla} f)(\boldsymbol{a}), -1\big) \cdot \Big[\boldsymbol{r} - \big(\boldsymbol{a}, f(\boldsymbol{a})\big)\Big] = 0,$$

i.e. vectors passing through the point $\big(\boldsymbol{a}, f(\boldsymbol{a})\big)$ and perpendicular to the normal to the tangent hyperplane. The equation of the tangent hyperplane in point-normal form is

$$\frac{\partial f}{\partial x_1}(\boldsymbol{a}) \cdot (x_1 - a_1) + \dots + \frac{\partial f}{\partial x_n}(\boldsymbol{a}) \cdot (x_n - a_n) - \big(z - f(\boldsymbol{a})\big) = 0,$$

where $a_1, \dots, a_n$ are the components of the vector $\boldsymbol{a}$.

11.2.3 Total differentiability

Definition 11.40: Total differentiability

Define $m, n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$, and let $\boldsymbol{a} \in U$ be an interior point of U.

(i) A scalar-valued function $f\colon U \to \mathbb{R}$ is said to be *totally differentiable at the point* $\boldsymbol{a}$ if there exists a vector $\boldsymbol{v} \in \mathbb{R}^n$ for which

$$\lim_{\boldsymbol{w} \to \boldsymbol{0}} \frac{f(\boldsymbol{a} + \boldsymbol{w}) - f(\boldsymbol{a}) - \boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{w}\|} = 0.$$

(ii) A vector-valued function $\boldsymbol{g}\colon U \to \mathbb{R}^m$ with components $\boldsymbol{g} = (g_1, \dots, g_m)$ is said to be *totally differentiable at* a if each of its components $g_1, \dots, g_m$ is totally differentiable at a.

(iii) A scalar-valued function $f\colon V \to \mathbb{R}$ defined on an open subset $V \subset \mathbb{R}^n$ is said to be *totally differentiable* if it is totally differentiable at every point in V.

(iv) A vector-valued function $\boldsymbol{g}\colon U \to \mathbb{R}^m$ with components $\boldsymbol{g} = (g_1, \dots, g_m)$ is said to be *totally differentiable* if each of its components $g_1, \dots, g_m$ is totally differentiable.

Total differentiability is a stricter condition than both continuity and partial differentiability; we formulate this statement more precisely in the following proposition.

Proposition 11.41: Implications of total differentiability

Define $n \in \mathbb{N}$, let $U \subset \mathbb{R}^n$ be an arbitrary subset of $\mathbb{R}^n$, let $a \in U$ be an interior point of U, and let the function $f\colon U \to \mathbb{R}$ be totally differentiable at $\boldsymbol{a}$. In this case

(i) The function f is continuous at $\boldsymbol{a}$.

(ii) The function f is partially differentiable at $\boldsymbol{a}$.

(iii) For each vector $\boldsymbol{u} \in \mathbb{R}^n$ the directional derivative of f in the direction of $\boldsymbol{u}$ exists at $\boldsymbol{a}$ and is equal to

$$(D_{\boldsymbol{u}} f)(\boldsymbol{a}) = (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{u}.$$

(iv) There exists exactly one vector $\boldsymbol{v} \in \mathbb{R}^n$ for which

$$\lim_{\boldsymbol{w} \to \boldsymbol{0}} \frac{f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - \boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{w}\|} = 0,$$

and this vector is the gradient $\boldsymbol{v} = (\boldsymbol{\nabla} f)(\boldsymbol{a})$.

Proof. Since f is totally differentiable at $\boldsymbol{a}$, there exists a vector $\boldsymbol{v} \in \mathbb{R}^n$ for which

$$\lim_{\boldsymbol{w} \to \boldsymbol{0}} \frac{f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - \boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{w}\|} = 0.$$

For shorthand, we denote the above fraction's numerator by

$$\alpha(\boldsymbol{w}) = f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - \boldsymbol{v} \cdot \boldsymbol{w}.$$

We show that f is continuous at $\boldsymbol{a}$ with the calculation

$$\begin{aligned}\lim_{\boldsymbol{w} \to \boldsymbol{0}} \big[f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a})\big] &\overset{\text{(a)}}{=} \lim_{\boldsymbol{w} \to \boldsymbol{0}} \big[\alpha(\boldsymbol{w}) + \boldsymbol{v} \cdot \boldsymbol{w}\big] \\ &= \lim_{\boldsymbol{w} \to \boldsymbol{0}} \left[\|\boldsymbol{w}\| \frac{\alpha(\boldsymbol{w})}{\|\boldsymbol{w}\|} + \boldsymbol{v} \cdot \boldsymbol{w}\right] \overset{\text{(b)}}{=} 0,\end{aligned}$$

where (a) follows from the definition of $\alpha(w)$ and (b) because f is totally differentiable, so $\lim_{\boldsymbol{w} \to \boldsymbol{0}} \alpha(\boldsymbol{w})/\|\boldsymbol{w}\| = 0$. We have shown that

$$\lim_{\boldsymbol{w} \to \boldsymbol{0}} \big[f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a})\big] = 0 \implies \lim_{\boldsymbol{x} \to \boldsymbol{a}} f(\boldsymbol{x}) = f(\boldsymbol{a}).$$

Thus f's function value and limit at $\boldsymbol{a}$ are equal, and so f is continuous by Proposition 11.21 (i).

We will now prove points (ii)-(iv). Define $\boldsymbol{u} \in \mathbb{R}^n$ and $h \in \mathbb{R}$. We will first show that

$$\lim_{h \to 0} \frac{\alpha(h\boldsymbol{u})}{h} = 0.$$

When $\boldsymbol{u} = \boldsymbol{0}$, this follows directly from the fact that $\alpha(\boldsymbol{0}) = 0$. When $\boldsymbol{u} \neq \boldsymbol{0}$, the above equality follows from the calculation

$$\frac{1}{\|\boldsymbol{u}\|} \lim_{h \to 0} \frac{\alpha(h\boldsymbol{u})}{|h|} = \lim_{h \to 0} \frac{\alpha(h\boldsymbol{u})}{\|h\boldsymbol{u}\|} \overset{\text{(a)}}{=} \lim_{\boldsymbol{w} \to \boldsymbol{0}} \frac{\alpha(\boldsymbol{w})}{\|\boldsymbol{w}\|} \overset{\text{(b)}}{=} 0.$$

where (a) follows from defining $\boldsymbol{w} = h\boldsymbol{u}$ and (b) holds because f is totally differentiable at $\boldsymbol{a}$, and so $\lim_{h\to 0}(\alpha(h\boldsymbol{u}))/h = 0$ when $\boldsymbol{u} \neq \boldsymbol{0}$.

We now compute the directional derivative

$$\begin{aligned}(D_{\boldsymbol{u}}f)(\boldsymbol{a}) &= \lim_{h\to 0}\frac{f(\boldsymbol{a}+h\boldsymbol{u}) - f(\boldsymbol{a})}{h} \\ &= \lim_{h\to 0}\frac{\alpha(h\boldsymbol{u}) + \boldsymbol{v}\cdot(h\boldsymbol{u})}{h} \\ &= \lim_{h\to 0}\frac{\alpha(h\boldsymbol{u})}{h} + \boldsymbol{v}\cdot\boldsymbol{u} \overset{\text{(a)}}{=} \boldsymbol{v}\cdot\boldsymbol{u}.\end{aligned}$$

where (a) follows from the earlier result $\lim_{h\to 0}\frac{\alpha(h\boldsymbol{u})}{h} = 0$. The result $(D_{\boldsymbol{u}}f)(\boldsymbol{a}) = \boldsymbol{v}\cdot\boldsymbol{u}$ means the directional derivative of f exists at $\boldsymbol{a}$ for any vector $\boldsymbol{u} \in \mathbb{R}^n$, which implies that all of f's partial derivatives (which are just directional derivatives in the direction of unit vectors) exist at $\boldsymbol{a}$, and so f is partially differentiable at $\boldsymbol{a}$, proving point (ii).

For each of the unit vectors $\hat{\mathbf{e}}_k, k = 1, 2, \ldots, n$ we then make the calculation

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = (\mathcal{D}_{\hat{\mathbf{e}}_k})(\boldsymbol{a}) = \boldsymbol{v}\cdot\hat{\mathbf{e}}_k = v_k,$$

Thus the k-th component v_k of the vector $\boldsymbol{v}$ equals f's partial derivative at $\boldsymbol{a}$ with respect to x_k, which in turn implies that

$$\boldsymbol{v} = (v_1, \ldots, v_n) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{a})\right) = (\boldsymbol{\nabla} f)(\boldsymbol{a}).$$

This result shows that $\boldsymbol{v}$ is uniquely equal to $(\boldsymbol{\nabla} f)(\boldsymbol{a})$, proving point (iv), and that

$$(D_{\boldsymbol{u}}f)(\boldsymbol{a}) = (\boldsymbol{\nabla} f)(\boldsymbol{a})\cdot\boldsymbol{u},$$

proving point (iii) and completing the proof. □

Proposition 11.42: Continuously-differentiable functions are totally differentiable on open subsets

Every continuously-differentiable function $f\colon U \to \mathbb{R}$ defined on an open subset $U \subset \mathbb{R}^n$ is totally differentiable.

Proof. Define the arbitrary point $\boldsymbol{a} \in U$ and choose arbitrary $\epsilon \in \mathbb{R}^+$. All of f's partial derivatives are continuous, so for all $k = 1, 2, \ldots, n$ there exists $\delta \in \mathbb{R}^+$ for which $\mathcal{K}(\boldsymbol{a}, \delta) \subset U$ and

$$\left|\frac{\partial f}{\partial x_k}(\boldsymbol{v}) - \frac{\partial f}{\partial x_k}(\boldsymbol{a})\right| < \frac{\epsilon}{n} \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a}, \delta). \tag{11.1}$$

Next, let $\boldsymbol{w}$ denote any vector in $\mathcal{K}(0,\delta) \subset \mathbb{R}^n$ and define the vectors

$$\begin{aligned}
\boldsymbol{b}_0 &= \mathbf{0} \\
\boldsymbol{b}_1 &= (w_1, 0, \ldots, 0) = w_1 \hat{\mathbf{e}}_1 \\
\boldsymbol{b}_2 &= (w_1, w_2, 0, \ldots, 0) = w_1 \hat{\mathbf{e}}_1 + w_2 \hat{\mathbf{e}}_2 \\
&\vdots \\
\boldsymbol{b}_n &= (w_1, w_2, \ldots, w_n) = \boldsymbol{w},
\end{aligned}$$

where $w_1, w_2, \ldots, w_n$ are the components of the vector $\boldsymbol{w}$. Because $\boldsymbol{w} \in \mathcal{K}(0,\delta)$,

$$\|\boldsymbol{b}_k\| < \|\boldsymbol{w}\| < \delta \text{ for all } k = 1, 2, \ldots, n,$$

and so all points of the form $\boldsymbol{a} + \boldsymbol{b}_k$ lie in the ball $\mathcal{K}(\boldsymbol{a},\delta)$.

More so, the ball $\mathcal{K}(\boldsymbol{a},\delta)$ also contains the image of the path constructed by using straight line segments to connect the points $\boldsymbol{a} = \boldsymbol{a} + \boldsymbol{b}_0$, $\boldsymbol{a} + \boldsymbol{b}_1$, $\boldsymbol{a} + \boldsymbol{b}_2$, and so on for all $k = 1, 2, \ldots, n$ until reaching the terminal point $\boldsymbol{a} + \boldsymbol{b}_n = \boldsymbol{w}$. By the construction of this path and the definition of the points $\{\boldsymbol{b}_k\}$, the restriction of the function f to the line segment between the points $\boldsymbol{a} + \boldsymbol{b}_{k-1}$ and $\boldsymbol{a} + \boldsymbol{b}_{k+1}$ depends only the variable x_k as it takes on values from a_k to $a_k + w_k$. Because f is continuously-differentiable, this restriction is necessarily a differentiable function of x_k, and so we can apply the mean value theorem to the restricted function on the interval $[a_k, a_k + w_k]$, which guarantees the existence of a point $c_k \in (0,1)$ for which

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) = \frac{f(\boldsymbol{a} + \boldsymbol{b}_k) - f(\boldsymbol{a} + \boldsymbol{b}_{k-1})}{(a_k + w_k) - a_k},$$

or, after rearranging,

$$f(\boldsymbol{a} + \boldsymbol{b}_k) - f(\boldsymbol{a} + \boldsymbol{b}_{k-1}) = w_k \frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k). \tag{11.2}$$

We then pause to note the auxiliary identity

$$f(\boldsymbol{a} + \boldsymbol{w}) - f(\boldsymbol{a}) = \sum_{k=1}^{n} \big[f(\boldsymbol{a} + \boldsymbol{b}_k) - f(\boldsymbol{a} + \boldsymbol{b}_{k-1}) \big], \tag{11.3}$$

which we then use in the calculation

$$\begin{aligned}
&f(\boldsymbol{a} + \boldsymbol{w}) - f(\boldsymbol{a}) - (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{w} \\
&\qquad \overset{\text{(a)}}{=} f(\boldsymbol{a} + \boldsymbol{w}) - f(\boldsymbol{a}) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\boldsymbol{a}) w_k \\
&\qquad \overset{\text{(b)}}{=} \sum_{k=1}^{n} \big[f(\boldsymbol{a} + \boldsymbol{b}_k) - f(\boldsymbol{a} + \boldsymbol{b}_{k-1}) \big] - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\boldsymbol{a}) w_k \\
&\qquad \overset{\text{(c)}}{=} \sum_{k=1}^{n} w_k \frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\boldsymbol{a}) w_k \\
&\qquad = \sum_{k=1}^{n} w_k \left[\frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) - \frac{\partial f}{\partial x_k}(\boldsymbol{a}) \right],
\end{aligned} \tag{11.4}$$

where (a) follows from the definition of the gradient and dot product, (b) from Equation 11.3, and (c) from Equation 11.2. We then take the absolute value of the result of Equation 11.4 and continue the calculation to get

$$\begin{aligned}
&\left|f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{w}\right| \\
&\qquad = \left|\sum_{k=1}^{n} w_k \left[\frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) - \frac{\partial f}{\partial x_k}(\boldsymbol{a})\right]\right| \\
&\qquad \overset{\text{(a)}}{\leq} \sum_{k=1}^{n} |w_k| \left|\frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) - \frac{\partial f}{\partial x_k}(\boldsymbol{a})\right| \\
&\qquad \overset{\text{(b)}}{\leq} \sum_{k=1}^{n} \|\boldsymbol{w}\| \left|\frac{\partial f}{\partial x_k}(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k) - \frac{\partial f}{\partial x_k}(\boldsymbol{a})\right| \\
&\qquad \overset{\text{(c)}}{<} \sum_{k=1}^{n} \|\boldsymbol{w}\| \frac{\epsilon}{n} = \|\boldsymbol{w}\| \epsilon,
\end{aligned} \tag{11.5}$$

where (a) follows from the triangle inequality, (b) from the bound $|w_k| \leq \|\boldsymbol{w}\|$ from Proposition 11.3 (iii), and (c) from Equation 11.1 and the fact that the vector $(\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k)$ lies in $\mathcal{K}(\boldsymbol{a}, \delta)$, since any point of the form $\boldsymbol{a} + \boldsymbol{b}_k$ lies in $\mathcal{K}(\boldsymbol{a}, \delta)$, and $\boldsymbol{a} + \boldsymbol{b}_{k-1} + c_k w_k \hat{\mathbf{e}}_k$ lies on the line segment between $\boldsymbol{a} + \boldsymbol{b}_{k-1}$ and $\boldsymbol{a} + \boldsymbol{b}_{k+1}$. Rearranging the result of Equation 11.5 produces

$$\frac{|f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} < \epsilon;$$

since $\epsilon \in \mathbb{R}^+$ and the point $\boldsymbol{a} \in U$ were arbitrary, it follows that

$$\lim_{\boldsymbol{w} \to \mathbf{0}} \frac{|f(\boldsymbol{a}+\boldsymbol{w}) - f(\boldsymbol{a}) - (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{w}|}{\|\boldsymbol{w}\|} = 0 \text{ for all } \boldsymbol{a} \in U,$$

and so f is totally differentiable. □

Example 11.43: Interpreting the gradient

Let $f : U \to \mathbb{R}$ be a continuously-differentiable function defined on the open subset $U \subset \mathbb{R}^n$, let $\boldsymbol{a} \in U$ be a point in U for which $(\boldsymbol{\nabla} f)(\boldsymbol{a}) \neq \mathbf{0}$, and let $\hat{\mathbf{u}} \in \mathbb{R}^n$ be an arbitrary unit-norm vector, i.e. a vector for which $\|\hat{\mathbf{u}}\| = 1$. By Proposition 11.42, f is totally differentiable on U, which allows us to compute the directional derivative of f in the direction of $\hat{\mathbf{u}}$ as follows:

$$\begin{aligned}
(D_{\hat{\mathbf{u}}} f)(\boldsymbol{a}) &\overset{\text{(a)}}{=} (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \hat{\mathbf{u}} = \|\hat{\mathbf{u}}\| \|(\boldsymbol{\nabla} f)(\boldsymbol{a})\| \cos\phi \\
&\overset{\text{(b)}}{=} \|(\boldsymbol{\nabla} f)(\boldsymbol{a})\| \cos\phi,
\end{aligned}$$

where (a) holds by Proposition 11.41 (iii), (b) because $\hat{\mathbf{u}}$ is of unit norm, and ϕ denotes the angle between the vectors $\hat{\mathbf{u}}$ and $(\boldsymbol{\nabla} f)(\boldsymbol{a})$. The result $(D_{\hat{\mathbf{u}}} f)(\boldsymbol{a}) = \|(\boldsymbol{\nabla} f)(\boldsymbol{a})\| \cos\phi$ means that f's directional derivative at $\boldsymbol{a}$ is maximized when $\cos\phi = 1$. But $\cos\phi = 1$ when $\hat{\mathbf{u}}$ points in the same direction as the gradient of f at $\boldsymbol{a}$, which means that the gradient $(\boldsymbol{\nabla} f)(\boldsymbol{a})$ points in the direction in which the function f is most rapidly increasing at the point $\boldsymbol{a}$.

11.2.4 The chain rule

Partial derivatives are computed analogously to the derivatives of scalar functions—all variables except one are set to constants, and the resulting partial derivative is analogous to the derivative of a scalar function. Thus the rules for partial differentiation of sums, products, and quotients of multivariable functions are analogous to their scalar function counterparts in Section 7.2.

However, the chain rule for computing the derivative of a composition of multivariable functions requires special attention, since this rule is more involved than its scalar analog; we will dedicate this section to the statement and proof of the multivariable chain rule.

Proposition 11.44: The chain rule for multivariable functions

Define $n, m \in \mathbb{N}$ and the subsets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, let the vector-valued function $\boldsymbol{g}\colon U \to \mathbb{R}^m$ be totally differentiable at the interior point $\boldsymbol{a} \subset U$, let $\boldsymbol{g}(\boldsymbol{a})$ be an interior point of V, and let the scalar-valued function $f\colon V \to \mathbb{R}$ be totally differentiable at $\boldsymbol{g}(\boldsymbol{a})$. In this case $\boldsymbol{a}$ is an interior point of $\boldsymbol{g}^{-1}(V)$, the composite function $f \circ \boldsymbol{g}\colon \boldsymbol{g}^{-1}(V) \to \mathbb{R}$ is totally differentiable at $\boldsymbol{a}$, and

$$\frac{\partial(f \circ \boldsymbol{g})}{\partial x_k}(\boldsymbol{a}) = (\boldsymbol{\nabla} f)\big(\boldsymbol{g}(\boldsymbol{a})\big) \cdot \frac{\partial \boldsymbol{g}}{\partial x_k}(\boldsymbol{a}) \text{ for all } k = 1, 2, \ldots, n.$$

Proof. Because $\boldsymbol{a}$ is an interior point of U, because $\boldsymbol{g}$ is totally differentiable and thus continuous at $\boldsymbol{a}$, and because V is a neighborhood of $\boldsymbol{g}(\boldsymbol{a})$, the point $\boldsymbol{a}$ is an interior point of $\boldsymbol{g}^{-1}(V)$.

Before completing the remainder of the proof, we first establish some notation: We will denote the variables of the function $\boldsymbol{g}$ by $x_1, \ldots, x_n$ and the variables of f by $u_1, \ldots, u_m$. Let $\boldsymbol{b} = \boldsymbol{g}(\boldsymbol{a})$, let $\boldsymbol{g}$ have scalar-valued components $\boldsymbol{g} = (g_1, \ldots, g_m)$, and let $\boldsymbol{v} \in \mathbb{R}^n$ be the vector with components

$$v_k = (\boldsymbol{\nabla} f)(\boldsymbol{b}) \cdot \frac{\partial \boldsymbol{g}}{\partial x_k}(\boldsymbol{a}).$$

For any vector $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ it then holds that

$$\begin{aligned}
\boldsymbol{v} \cdot \boldsymbol{w} &\overset{\text{(a)}}{=} \sum_{k=1}^{n} \left[(\boldsymbol{\nabla} f)(\boldsymbol{b}) \cdot \frac{\partial \boldsymbol{g}}{\partial x_k}(\boldsymbol{a}) \right] w_k \\
&\overset{\text{(b)}}{=} \sum_{k=1}^{n} \sum_{i=1}^{m} \frac{\partial f}{\partial u_i}(\boldsymbol{b}) \frac{\partial g_i}{\partial x_k}(\boldsymbol{a}) w_k \\
&= \sum_{i=1}^{m} \frac{\partial f}{\partial u_i}(\boldsymbol{b}) \left[\sum_{k=1}^{n} \frac{\partial g_i}{\partial x_k}(\boldsymbol{a}) w_k \right] \\
&\overset{\text{(c)}}{=} \sum_{i=1}^{m} \frac{\partial f}{\partial u_i}(\boldsymbol{b}) \big[(\boldsymbol{\nabla} g_i)(\boldsymbol{a}) \cdot \boldsymbol{w} \big] \\
&\overset{\text{(d)}}{=} (\boldsymbol{\nabla} f)(\boldsymbol{b}) \cdot \boldsymbol{\eta}(\boldsymbol{w}),
\end{aligned}$$

where (a) follows from the definition of the dot product, (b) and (c) from the definition of the dot product and gradient, and in (d) we have defined the vector

$$\boldsymbol{\eta}(\boldsymbol{w}) = \big[(\boldsymbol{\nabla} g_1)(\boldsymbol{a}) \cdot \boldsymbol{w}, \ldots, (\boldsymbol{\nabla} g_m)(\boldsymbol{a}) \cdot \boldsymbol{w} \big].$$

We then fix an upper bound $M \in \mathbb{R}^+$ for which

$$|(\boldsymbol{\nabla} f)(\boldsymbol{b})| \leq M \quad \text{and} \quad |(\boldsymbol{\nabla} g_i)(\boldsymbol{a})| \leq M \text{ for all } i = 1, \ldots, m; \tag{11.6}$$

we can be sure such an upper bound exists because f and $\boldsymbol{g}$'s total derivatives exist at the points $\boldsymbol{b}$ and $\boldsymbol{a}$, respectively.

Next, choose arbitrary $\epsilon \in \mathbb{R}^+$ and choose sufficiently small $\nu \in \mathbb{R}^+$ for which $\nu < 1$ and

$$\nu\sqrt{m}(1 + 2M) < \epsilon. \tag{11.7}$$

Because f is totally differentiable at $\boldsymbol{b}$, there exists $\mu \in \mathbb{R}^+$ for which $\mathcal{K}(\boldsymbol{b}, \mu) \subset V$ and

$$|f(\boldsymbol{b} + \boldsymbol{r}) - f(\boldsymbol{b}) - (\boldsymbol{\nabla} f)(\boldsymbol{b}) \cdot \boldsymbol{r}| \overset{\text{(a)}}{=} |\alpha(\boldsymbol{r})| \leq \nu\|\boldsymbol{r}\| \text{ for all } \boldsymbol{r} \in \mathcal{K}(\boldsymbol{0}, \mu), \tag{11.8}$$

where in (a) we have introduced the shorthand notation

$$\alpha(\boldsymbol{r}) = f(\boldsymbol{b} + \boldsymbol{r}) - f(\boldsymbol{b}) - (\boldsymbol{\nabla} f)(\boldsymbol{b}) \cdot \boldsymbol{r}.$$

Because $\boldsymbol{g}$ is totally differentiable, and thus continuous, at $\boldsymbol{a}$, there exists $\delta \in \mathbb{R}^+$ for which

$$\mathcal{K}(\boldsymbol{a}, \delta) \subset U \quad \text{and} \quad \boldsymbol{g}\big(\mathcal{K}(\boldsymbol{a}, \delta)\big) \subset \mathcal{K}(\boldsymbol{b}, \mu) \subset \mathcal{K}\big(\boldsymbol{g}(\boldsymbol{a}), \mu\big).$$

Additionally, again because $\boldsymbol{g}$ is totally differentiable at $\boldsymbol{a}$, it holds that

$$|\beta_i(\boldsymbol{w})| = |g_i(\boldsymbol{a} + \boldsymbol{w}) - g_i(\boldsymbol{w}) - (\boldsymbol{\nabla} g_i)(\boldsymbol{a}) \cdot \boldsymbol{w}| \leq \nu\|\boldsymbol{w}\| \text{ for all } \boldsymbol{w} \in \mathcal{K}(\boldsymbol{0}, \delta),$$

where we have defined the quantities

$$\begin{aligned} \beta_i(\boldsymbol{w}) &= g_i(\boldsymbol{a} + \boldsymbol{w}) - g_i(\boldsymbol{w}) - (\boldsymbol{\nabla} g_i)(\boldsymbol{a}) \cdot \boldsymbol{w}, \quad i = 1, \ldots, m \\ \boldsymbol{\beta}(\boldsymbol{w}) &= \big(\beta_1(\boldsymbol{w}), \ldots, \beta_m(\boldsymbol{w})\big) \in \mathbb{R}^m. \end{aligned}$$

The inequality $|\beta_i(\boldsymbol{w})| \leq \nu\|\boldsymbol{w}\|$ for all $i = 1, \ldots, m$ in turn implies that

$$\|\boldsymbol{\beta}(\boldsymbol{w})\| \leq \sqrt{m}\nu\|\boldsymbol{w}\|. \tag{11.9}$$

Next, we choose arbitrary $\boldsymbol{w} \in \mathcal{K}(\boldsymbol{0}, \delta) \subset \mathbb{R}^n$ for which $\boldsymbol{w} \neq \boldsymbol{0}$ and define

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{w}) = \boldsymbol{g}(\boldsymbol{a} + \boldsymbol{w}) - \boldsymbol{g}(\boldsymbol{a}).$$

The function $\boldsymbol{g}$ is continuous at $\boldsymbol{a}$ and thus $\boldsymbol{g}\big(\mathcal{K}(\boldsymbol{a}, \delta)\big) \subset \mathcal{K}(\boldsymbol{b}, \mu) = \mathcal{K}\big(\boldsymbol{g}(\boldsymbol{a}), \mu\big)$, which in turn implies that

$$\boldsymbol{r} \in \mathcal{K}(\boldsymbol{0}, \mu).$$

More so, the components of the vector $\boldsymbol{r}$ obey

$$\begin{aligned} |r_i| &= |g_i(\boldsymbol{a} + \boldsymbol{w}) - g_i(\boldsymbol{a})| = |\beta_i(\boldsymbol{w}) + (\boldsymbol{\nabla} g_i)(\boldsymbol{a}) \cdot \boldsymbol{w}| \\ &\overset{\text{(a)}}{\leq} |\beta_i(\boldsymbol{w})| + |(\boldsymbol{\nabla} g_i)(\boldsymbol{a}) \cdot \boldsymbol{w}| \overset{\text{(b)}}{\leq} |\beta_i(\boldsymbol{w})| + \|(\boldsymbol{\nabla} g_i)(\boldsymbol{a})\|\|\boldsymbol{w}\| \\ &\leq \nu\|\boldsymbol{w}\| + M\|\boldsymbol{w}\| \\ &\overset{\text{(c)}}{<} (1 + M)\|\boldsymbol{w}\|, \end{aligned}$$

where (a) and (b) follow from the triangle and Cauchy-Schwartz inequalities, respectively, and (c) from the requirement that $\mu < 1$. The result $|r_i| < (1+M)\|\boldsymbol{w}\|$ in turn implies that

$$\|\boldsymbol{r}\| < \sqrt{m}(1+M)\|\boldsymbol{w}\|. \tag{11.10}$$

We then continue with the calculation

$$\begin{aligned} f\big(\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{w})\big) - f\big(\boldsymbol{g}(\boldsymbol{a})\big) - \boldsymbol{v}\cdot\boldsymbol{w} &= f(\boldsymbol{b}+\boldsymbol{r}) - f(\boldsymbol{b}) - \boldsymbol{v}\cdot\boldsymbol{w} \\ &= \alpha(\boldsymbol{r}) + (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\boldsymbol{r} - (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\boldsymbol{\eta}(\boldsymbol{w}) \\ &= \alpha(\boldsymbol{r}) + (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\big(\boldsymbol{r} - \boldsymbol{\eta}(\boldsymbol{w})\big) \\ &= \alpha(\boldsymbol{r}) + (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\boldsymbol{\beta}(\boldsymbol{w}), \end{aligned}$$

which we then use to make the bound

$$\begin{aligned} \frac{\big|f\big(\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{w})\big) - f\big(\boldsymbol{g}(\boldsymbol{a})\big) - \boldsymbol{v}\cdot\boldsymbol{w}\big|}{\|\boldsymbol{w}\|} &= \frac{|\alpha(\boldsymbol{r}) + (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\boldsymbol{\beta}(\boldsymbol{w})|}{\|\boldsymbol{w}\|} \\ &\overset{\text{(a)}}{\leq} \frac{|\alpha(\boldsymbol{r})|}{\|\boldsymbol{w}\|} + |(\boldsymbol{\nabla} f)(\boldsymbol{b})|\frac{\|\boldsymbol{\beta}(\boldsymbol{w})\|}{\|\boldsymbol{w}\|} \\ &\overset{\text{(b)}}{\leq} \frac{\nu\|\boldsymbol{r}\|}{\|\boldsymbol{w}\|} + |(\boldsymbol{\nabla} f)(\boldsymbol{b})|\frac{\|\boldsymbol{\beta}(\boldsymbol{w})\|}{\|\boldsymbol{w}\|} \\ &\overset{\text{(c)}}{<} \nu\sqrt{m}(1+M) + |(\boldsymbol{\nabla} f)(\boldsymbol{b})|\frac{\|\boldsymbol{\beta}(\boldsymbol{w})\|}{\|\boldsymbol{w}\|} \\ &\overset{\text{(d)}}{\leq} \nu\sqrt{m}(1+M) + \nu M\sqrt{m} \\ &= \nu\sqrt{m}(1+2M) \overset{\text{(e)}}{<} \epsilon, \end{aligned}$$

where (a) follows from the triangle inequality, (b) from Equation 11.8, (c) from Equation 11.10, (d) from Equations 11.6 and 11.9, and (e) from Equation 11.7. The result of this calculation is

$$\frac{\big|f\big(\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{w})\big) - f\big(\boldsymbol{g}(\boldsymbol{a})\big) - \boldsymbol{v}\cdot\boldsymbol{w}\big|}{\|\boldsymbol{w}\|} < \epsilon \text{ for arbitrary } \epsilon \in \mathbb{R}^+,$$

which implies that the function $f\circ\boldsymbol{g}$ is totally differentiable at $\boldsymbol{a}$, and that its partial derivatives are equal to the components of the vector $\boldsymbol{v}$, i.e.

$$\frac{\partial(f\circ\boldsymbol{g})}{\partial x_k}(\boldsymbol{a}) = v_k = (\boldsymbol{\nabla} f)(\boldsymbol{b})\cdot\frac{\partial\boldsymbol{g}}{\partial x_k}(\boldsymbol{a}). \qquad \square$$

Example 11.45: Using the multivariable chain rule

(1) Consider the vector-valued function $\boldsymbol{g}\colon R^2 \to \mathbb{R}^2$ given by

$$\boldsymbol{g}(r,\phi) = (r\cos\phi, r\sin\phi).$$

This function encodes the change of variables from polar to Cartesian coordinates, since we could also write it in the form

$$\boldsymbol{g} = (g_1, g_2) = (x, y),$$

where the component functions x and y are given by

$$x(r,\phi) = r\cos\phi \quad \text{and} \quad y(r,\phi) = r\sin\phi.$$

For any two-variable function $f\colon \mathbb{R}^2 \to \mathbb{R}$ of the Cartesian coordinates x and y, we will aim to find f's partial derivatives with respect to the polar coordinates r and ϕ. To this end, we apply the chain rule to get

$$\begin{aligned}
\frac{\partial(f\circ g)}{\partial r}(r,\phi) &= \frac{\partial f}{\partial x}(r\cos\phi, r\sin\phi)\frac{\partial x}{\partial r}(r,\phi) + \frac{\partial f}{\partial y}(r\cos\phi, r\sin\phi)\frac{\partial y}{\partial r}(r,\phi) \\
&= \frac{\partial f}{\partial x}(r\cos\phi, r\sin\phi)\cos\phi + \frac{\partial f}{\partial y}(r\cos\phi, r\sin\phi)\sin\phi, \\
\frac{\partial(f\circ g)}{\partial \phi}(r,\phi) &= \frac{\partial f}{\partial x}(r\cos\phi, r\sin\phi)\frac{\partial x}{\partial \phi}(r,\phi) + \frac{\partial f}{\partial y}(r\cos\phi, r\sin\phi)\frac{\partial y}{\partial \phi}(r,\phi) \\
&= \frac{\partial f}{\partial x}(r\cos\phi, r\sin\phi)(-r\sin\phi) + \frac{\partial f}{\partial y}(r\cos\phi, r\sin\phi)r\cos\phi.
\end{aligned}$$

As a concrete example, let us find the partial derivatives of the function

$$f(x,y) = x^2 + y^2,$$

for which the composite function $f\circ g$ reads

$$(f\circ g)(r,\phi) = f(r\cos\phi, r\sin\phi) = r^2\cos^2\phi + r^2\sin^2\phi = r^2.$$

Thus in polar coordinates the function f depends only on the variable r, so we expect that f's partial derivative with respect to ϕ will be zero. We first note that $f_x = 2x$ and $f_y = 2y$, then apply the chain rule to get the expected result

$$\begin{aligned}
\frac{\partial(f\circ g)}{\partial r}(r,\phi) &= 2r\cos^2\phi + 2r\sin^2\phi = 2r, \\
\frac{\partial(f\circ g)}{\partial \phi}(r,\phi) &= -2r^2\cos\phi\sin\phi + 2r^2\sin\phi\cos\phi = 0.
\end{aligned}$$

(2) Let $\boldsymbol{\gamma}\colon J \to \mathbb{R}^n$ be a continuously-differentiable vector-valued function of one variable defined on the open interval $J \subset \mathbb{R}$ and defined as

$$\boldsymbol{\gamma}(t) = \big(\gamma_1(t), \ldots, \gamma_n(t)\big) \text{ for all } t \in J.$$

The function $\boldsymbol{\gamma}$ represents a parameterized curve in $\mathbb{R}^n$, and its derivative at any point $t_0 \in J$, i.e.

$$\frac{\mathrm{d}\boldsymbol{\gamma}}{\mathrm{d}t} = \dot{\boldsymbol{\gamma}}(t_0) = \big(\dot\gamma_1(t_0), \ldots, \dot\gamma_n(t_0)\big) \in \mathbb{R}^n,$$

is a vector tangent to the curve at the point $\boldsymbol{\gamma}(t_0)$ that represents the velocity of the parameterization.

Next, let $f\colon U \to \mathbb{R}$ be any continuously-differentiable function defined on the open subset $U \subset \mathbb{R}^n$ and defined the point $\boldsymbol{a} = \boldsymbol{\gamma}(t_0) \in U$. We then apply the chain rule to compute the derivative

$$\frac{\mathrm{d}(f\circ\boldsymbol{\gamma})}{\mathrm{d}t}(t_0) = (\boldsymbol{\nabla} f)(\boldsymbol{a})\cdot\dot{\boldsymbol{\gamma}}(t_0) \overset{\text{(a)}}{=} \big(D_{\dot{\boldsymbol{\gamma}}(t_0)} f\big)(\boldsymbol{a}),$$

where (a) follows from Proposition 11.41 (iii). If the parameterized curve $\boldsymbol{\gamma}$ lies inside a set of points on which the function f is constant, i.e. inside a so-called *level hypersurface* of f, then the function $f \circ \boldsymbol{\gamma}$ is also constant and obeys

$$\frac{\mathrm{d}(f \circ \boldsymbol{\gamma})}{\mathrm{d}t}(t_0) = (\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \dot{\boldsymbol{\gamma}}(t_0) = 0,$$

meaning that the gradient of f at the point $\boldsymbol{a}$ is orthogonal to the level hypersurface of f passing through the point $\boldsymbol{a}$.

11.2.5 Higher-order partial derivatives

Definition 11.46: Higher-order partial derivatives

Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of n variables defined on the open subset $U \subset \mathbb{R}^n$. If f is partially differentiable with respect to the variable x_i, then the partial derivative $f_{x_i}\colon U \to R$ is itself a function of n variables, and if f_{x_i} is partially differentiable with respect to the variable x_k, then the *second partial derivative of f with respect to the variables x_i and x_k* is defined as

$$\frac{\partial^2 f}{\partial x_k \partial x_i} = \frac{\partial}{\partial x_k}\left[\frac{\partial}{\partial x_i}(f)\right] = \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_k}$$

We will also write second partial derivatives in the more compact forms

$$\frac{\partial^2 f}{\partial x_k \partial x_i} = (f_{x_i})_{x_k} = f_{x_i x_k} = D_k D_i f = D_{ki} f.$$

We define higher-order derivatives similarly: the k-th partial derivative of the function f with respect to the variables $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ is defined as

$$\frac{\partial^k f}{\partial x_{i_k} \cdots \partial x_{i_2} \partial x_{i_1}} = \frac{\partial}{\partial x_{i_k}}\left(\cdots\left(\frac{\partial}{\partial x_{i_2}}\left(\frac{\partial}{\partial x_{i_1}}(f)\right)\right)\cdots\right),$$

assuming the appropriate partial derivatives exist. If the function f is k-times differentiable with respect to the same variable x_i, we write the corresponding higher-order partial derivative in the shorthand form

$$\frac{\partial^k f}{\partial x_i \cdots \partial x_i \partial x_i} = \frac{\partial^k f}{\partial x_i^k} = D_i^k f.$$

Definition 11.47: Continuous higher-order partial differentiability

(i) Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of the n variables $x_1, x_2, \ldots, x_n$ defined on the open subset $U \subset \mathbb{R}^n$ and let $k \in \mathbb{N} \cup \{0\}$ be an arbitrary positive integer. The function f is said to be *k-times continuously-differentiable* if it is continuous and, for all natural numbers $p \leq k$ and all possible choices of the

indices $i_1, i_2, \ldots, i_p \in \{1, 2, \ldots, n\}$, the partial derivatives of the form

$$\frac{\partial^p f}{\partial x_{i_p} \cdots \partial x_{i_2} \partial x_{i_1}}$$

exist at every point in U and are themselves continuous functions of n variables mapping from U to $\mathbb{R}$.

(ii) We denote the set of all k-times continuously-differentiable scalar-valued functions $f\colon U \to \mathbb{R}$ by $C^k(U)$.

(iii) Let $m, n \in \mathbb{N}$ be arbitrary natural numbers, let $\boldsymbol{g}\colon U \to \mathbb{R}^m$ be a vector-valued function with scalar-valued components $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)$ defined on the open subset $U \subset \mathbb{R}^n$, and let $k \in \mathbb{N} \cup \{0\}$ be an arbitrary nonnegative integer. The function $\boldsymbol{g}$ is said to be *k-times continuously-differentiable* if each of its components $g_1, g_2, \ldots, g_m$ is k-times continuously-differentiable.

(iv) We denote the set of all k-times continuously-differentiable vector-valued functions $\boldsymbol{g}\colon U \to \mathbb{R}^m$ by $C^k(U, \mathbb{R}^m)$.

(v) Functions inside the set intersections

$$C^\infty(U) = \bigcap_{k=0}^{\infty} c^k(U) \quad \text{and} \quad C^\infty(U, \mathbb{R}^m) = \bigcap_{k=0}^{\infty} c^k(U, \mathbb{R}^m)$$

are called *smooth functions*.

Example 11.48: Second partial derivatives of a polynomial

Let f be the scalar-valued polynomial function of two variables given by

$$f(x, y) = x^2 + 2xy^3 + 4x.$$

The function f's second partial derivatives are

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2}(x, y) &= 2 \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &= 6y^2 \\
\frac{\partial^2 f}{\partial y \partial x}(x, y) &= 6y^2 \\
\frac{\partial^2 f}{\partial y^2}(x, y) &= 12xy.
\end{aligned}$$

Note that f's two mixed partial derivatives in Example 11.48 are equal—in this particular case this equality of partial derivatives is not a coincidence, but mixed partial derivatives are not equal in general. In this section we will show that mixed higher-order partial derivatives are equal whenever the relevant partial derivatives exist and are continuous. In preparation, we first prove the following lemma.

Lemma 11.49: An auxiliary result for proving equality of mixed partial derivatives

Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of n variables defined on the open subset $U \subset \mathbb{R}^n$, let $i, j \in \{1, 2, \ldots, n\}$ be two indices for which $i \neq j$, assume the partial derivatives $D_i f$ and D_{ji} exist at all points in U, and define the point $\boldsymbol{a} \in U$. More so, let $h, k \in \mathbb{R} \setminus \{0\}$ be two nonzero real numbers for which

$$\boldsymbol{a} + t\hat{\mathbf{e}}_i + s\hat{\mathbf{e}}_j \in U \text{ for all } t \in [0, h] \text{ and all } s \in [0, k].$$

In this case there exists a point $t_0 \in (0, h)$ and a point $s_0 \in (0, k)$ for which

$$f(\boldsymbol{a} + h\hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - f(\boldsymbol{a} + h\hat{\mathbf{e}}_i) - f(\boldsymbol{a} + k\hat{\mathbf{e}}_j) + f(\boldsymbol{a}) \\ = hk(D_{ji}f)(\boldsymbol{a} + t_0\hat{\mathbf{e}}_i + s_0\hat{\mathbf{e}}_j).$$

Proof. We first define the function $u\colon [0, h] \to \mathbb{R}$ given by

$$u(t) = f(\boldsymbol{a} + t\hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - f(\boldsymbol{a} + t\hat{\mathbf{e}}_i), \quad t \in [0, h].$$

We then apply the mean value theorem to u on the interval $[0, h]$, which guarantees the existence of a point $t_0 \in (0, h)$ for which

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t_0) = \frac{u(h) - u(0)}{h - 0},$$

or, after rearranging,

$$\begin{aligned} u(h) - u(0) &= h \cdot \frac{\mathrm{d}u}{\mathrm{d}t}(t_0) \\ &\overset{\text{(a)}}{=} h \cdot \big[(D_i f)(\boldsymbol{a} + t_0\hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - (D_i f)(\boldsymbol{a} + t_0\hat{\mathbf{e}}_i)\big], \end{aligned} \tag{11.11}$$

where (a) follows from the chain rule.

We then define the function $v\colon [0, k] \to \mathbb{R}$ given by

$$v(s) = (D_i f)(\boldsymbol{a} + t_0\hat{\mathbf{e}}_i + s\hat{\mathbf{e}}_j), \quad s \in [0, k].$$

Like for the function u, we apply the mean value theorem to v on the interval $[0, k]$, which guarantees the existences of a point $s_0 \in (0, k)$ for which

$$\frac{\mathrm{d}v}{\mathrm{d}s}(s_0) = \frac{v(k) - v(0)}{k - 0},$$

or, after rearranging,

$$v(k) - v(0) = k \cdot \frac{\mathrm{d}v}{\mathrm{d}s}(s_0) = k \cdot (D_j D_i f)(\boldsymbol{a} + t_0\hat{\mathbf{e}}_i + s_0\hat{\mathbf{e}}_j). \tag{11.12}$$

On the one hand, by the definition of u, it then holds that

$$u(h) - u(0) = f(\boldsymbol{a} + h\hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - f(\boldsymbol{a} + h\hat{\mathbf{e}}_i) - f(\boldsymbol{a} + k\hat{\mathbf{e}}_j) + f(\boldsymbol{a}),$$

while on the other hand the equalities earlier in this proof imply that

$$\begin{aligned} u(h) - u(0) &\overset{\text{(a)}}{=} h \cdot \big[(D_i f)(\boldsymbol{a} + t_0 \hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - (D_i f)(\boldsymbol{a} + t_0 \hat{\mathbf{e}}_i)\big] \\ &\overset{\text{(b)}}{=} h \cdot \big[v(k) - v(0)\big] \\ &\overset{\text{(c)}}{=} hk \cdot (D_{ji} f)(\boldsymbol{a} + t_0 \hat{\mathbf{e}}_i + s_0 \hat{\mathbf{e}}_j), \end{aligned}$$

where (a) follows from Equation 11.11, (b) from the definition of $v(s)$, and (c) from Equation 11.12. We then complete the proof by equating the above two expressions for $u(h) - u(0)$. □

Proposition 11.50: Condition for changing the order of partial differentiation

Let $f\colon U \to \mathbb{R}$ be a scalar-valued function of n variables defined on the open subset $U \subset \mathbb{R}^n$, and define the indices $i, j \in \{1, 2, \ldots, n\}$. More so, assume the partial derivatives $D_i f$, $D_j f$ and $D_{ji} f$ exist at all points in U and that the function $D_{ji} f\colon U \to \mathbb{R}$ is continuous at some point $\boldsymbol{a} \in U$. In this case the second partial derivative $D_{ij} f$ exists at $\boldsymbol{a}$ and obeys

$$(D_{ij} f)(\boldsymbol{a}) = (D_{ji} f)(\boldsymbol{a}).$$

Proof. If $i = j$ the proposition is trivial, so we will assume that $i \neq j$. First choose arbitrary $\epsilon \in \mathbb{R}^+$. Because the function $D_{ji} f$ is continuous at $\boldsymbol{a}$, there exists $\delta \in \mathbb{R}^+$ for which $\mathcal{K}(\boldsymbol{a}, \delta) \subset U$ and

$$|(D_{ij} f)(\boldsymbol{v}) - (D_{ji} f)(\boldsymbol{a})| < \epsilon \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a}, \delta).$$

By Lemma 11.49, for any two nonzero real numbers $h, k \in \mathbb{R} \setminus \{0\}$ for which $h^2 + k^2 < \delta^2$, it holds that

$$\left| \frac{f(\boldsymbol{a} + h\hat{\mathbf{e}}_i + k\hat{\mathbf{e}}_j) - f(\boldsymbol{a} + h\hat{\mathbf{e}}_i)}{hk} - \frac{f(\boldsymbol{a} + k\hat{\mathbf{e}}_j) - f(\boldsymbol{a})}{hk} - (D_{ji} f)(\boldsymbol{a}) \right| < \epsilon.$$

We now consider the limit of the left-hand side of the above equality as $k \to 0$ at a fixed value of h. By recognizing the expressions for partial derivatives and remembering that the partial derivative $D_j f$ is assumed to exist at $\boldsymbol{a}$, this limit reduces to

$$\left| \frac{(D_j f)(\boldsymbol{a} + h\hat{\mathbf{e}}_i) - (D_j f)(\boldsymbol{a})}{h} - (D_{ji} f)(\boldsymbol{a}) \right| \leq \epsilon.$$

This resulting inequality in turn means that the function $D_j f$ is partially differentiable with respect to x_i at the point $\boldsymbol{a}$ and obeys

$$(D_i D_j f)(\boldsymbol{a}) = \lim_{h \to 0} \frac{(D_j f)(\boldsymbol{a} + h\hat{\mathbf{e}}_i) - (D_j f)(\boldsymbol{a})}{h} = (D_{ji} f)(\boldsymbol{a}),$$

which completes the proof that $(D_{ij} f)(\boldsymbol{a}) = (D_{ji} f)(\boldsymbol{a})$. □

Proposition 11.50 is very useful in practice, since it means that the mixed partial derivatives $D_{ij} f$ and $D_{ji} f$ are equal for any function $f \in C^2(U)$. More so, the proposition can be

naturally extended to apply to the mixed partial derivatives of arbitrary order of any smooth function.

11.3 Taylor's formula for multivariable functions

In this section we will extend Taylor expansions to functions of multiple variables.

Topic 11.7: Deriving Taylor's formula for multivariable functions

Define $n \in \mathbb{N}$, $m \in \mathbb{N} \cup \{0\}$, the open subset $U \subset \mathbb{R}^n$, the point $\boldsymbol{a} = (a_1, \ldots, a_n) \in U$, and let $f \colon U \to \mathbb{R}$ be an $(m+1)$-times continuously-differentiable function of n variables. The set U is open, so there exists $R \in \mathbb{R}^+$ for which $\mathcal{K}(\boldsymbol{a}, R) \subset U$. Define the point $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{K}(\boldsymbol{a}, R) \setminus \{\boldsymbol{a}\} \subset \mathbb{R}^n$, and let

$$\boldsymbol{w} = (w_1, \ldots, w_n) = (x_1 - a_1, \ldots, x_n - a_n).$$

Note that, by construction, the entire line segment between $\boldsymbol{a}$ and $\boldsymbol{a} + \boldsymbol{w}$ lies inside the open ball $\mathcal{K}(\boldsymbol{a}, R)$. We will parameterize the line containing the points $\boldsymbol{a}$ and $\boldsymbol{a} + \boldsymbol{w}$ with the vector-valued function $\boldsymbol{\gamma} \colon \mathbb{R} \to \mathbb{R}^n$ given by

$$\boldsymbol{\gamma}(t) = \boldsymbol{a} + t\boldsymbol{w}.$$

By construction, the function $\boldsymbol{\gamma}$ is smooth and obeys $\boldsymbol{\gamma}(0) = \boldsymbol{a}$ and $\boldsymbol{\gamma}(1) = \boldsymbol{a} + \boldsymbol{w}$, the set $\boldsymbol{\gamma}([0,1]) \subset \mathbb{R}^n$ is the line segment between $\boldsymbol{a}$ and $\boldsymbol{a} + \boldsymbol{w}$, and $\dot{\boldsymbol{\gamma}}(t) = \boldsymbol{w}$ for all $t \in \mathbb{R}$. Next, we define the scalar function $g \colon \boldsymbol{\gamma}^{-1}(U) \to \mathbb{R}$ as the composition

$$g = f \circ \boldsymbol{\gamma}, \quad g(t) = f\big(\boldsymbol{\gamma}(t)\big).$$

Because f is continuous, the set $\boldsymbol{\gamma}^{-1}(U) \subset \mathbb{R}$ is open, and in addition

$$[0,1] \subset \boldsymbol{\gamma}^{-1}(U).$$

Because the function f is $(m+1)$-times continuously-differentiable and because $\boldsymbol{\gamma}$ is smooth, by the chain rule the function g is itself $(m+1)$-times continuously-differentiable, and in addition obeys

$$g(0) = f(\boldsymbol{a}) \quad \text{and} \quad g(1) = f(\boldsymbol{a} + \boldsymbol{w}).$$

By Taylor's formula for scalar functions applied to g when expanded about 0, there exists a point $\theta \in (0,1)$ for which

$$\begin{aligned} f(\boldsymbol{a} + \boldsymbol{w}) = g(1) &= \sum_{k=0}^{m} \frac{g^{(k)}(0)}{k!}(1-0)^k + \frac{g^{(m+1)}(\theta)}{(m+1)!}(1-0)^{m+1} \\ &= \sum_{k=0}^{m} \frac{g^{(k)}(0)}{k!} + \frac{g^{(m+1)}(\theta)}{(m+1)!}. \end{aligned} \tag{11.13}$$

We can then find the corresponding Taylor's formula for f by computing g's derivatives using the chain rule, and then expressing these derivatives in terms of the partial derivatives of f. To do this, we first write $\boldsymbol{\gamma}$ by components as

$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n) \quad \text{and} \quad \dot{\gamma}_j(t) = w_j, \quad t \in \mathbb{R},\ j = 1, 2, \ldots, n.$$

We then apply the chain rule, which for all $t \in g^{-1}(U)$ produces

$$\begin{aligned} g'(t) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\big(\boldsymbol{\gamma}(t)\big)\dot{\gamma}_i(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\boldsymbol{a}+t\boldsymbol{w})w_i \\ &= \sum_{i=1}^{n} (D_i f)(\boldsymbol{a}+t\boldsymbol{w})w_i. \end{aligned}$$

We should remember that the function $(D_i f)(\boldsymbol{a}+t\boldsymbol{w})$ is a composition of function $\boldsymbol{\gamma}$ and the function $D_i f$ (similarly to how g is a composition of $\boldsymbol{\gamma}$ and f); with this in mind we can then apply the chain rule once more to get

$$\begin{aligned} g''(t) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \big(D_j(D_i f)\big)(\boldsymbol{a}+t\boldsymbol{w})w_j \right] w_i \\ &= \sum_{i,j=1}^{n} (D_{ji} f)(\boldsymbol{a}+t\boldsymbol{w})w_j w_i. \end{aligned}$$

More generally, the k-th derivative of the $(m+1)$-times continuously-differentiable function g is

$$g^{(k)}(t) = \sum_{i_1,\ldots,i_k=1}^{n} (D_{i_1\ldots i_k} f)(\boldsymbol{a}+t\boldsymbol{w})w_{i_1}\cdots w_{i_k}, \quad k = 1,\ldots,m+1.$$

We then substitute these derivatives of g into the earlier Taylor's formula expression for the scalar function g in Equation 11.13 to get

$$\begin{aligned} f(\boldsymbol{a}+\boldsymbol{w}) = \sum_{k=0}^{m} \frac{1}{k!} &\left[\sum_{i_1,\ldots,i_k=1}^{n} (D_{i_1\ldots i_k} f)(\boldsymbol{a}) w_{i_1}\cdots w_{i_k} \right] \\ &+ \frac{1}{(m+1)!} \sum_{i_1,\ldots,i_{m+1}=1}^{n} (D_{i_1\ldots i_{m+1}=1} f)(\boldsymbol{a}+\theta\boldsymbol{w}) w_{i_1}\cdots w_{i_{m+1}} \end{aligned}$$

for some $\theta \in (0,1)$. This expression is the m-th order Taylor's formula for the function f expanded about the point $\boldsymbol{a}$ with the remainder in Lagrange form.

We now continue the above derivation with some simplifications to mixed partial derivatives.

Topic 11.8: Simplifying mixed partial derivatives

The function f is $(m+1)$-times continuously-differentiable, so by Proposition 11.50 the order of partial differentiation is immaterial. As a result, we can arrange the partial derivatives so that we first differentiate f with respect to the first variable (as many times as required), then with respect to the second variable, and so on. As a concrete example, consider the partial derivative

$$D_{i_1\ldots i_k} f = \frac{\partial^k f}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_k}},$$

and suppose r_1 of the indices $i_1, \dots, i_k$ are equal to 1, r_2 of the indices are equal to 2, and in general r_j of the indices are equal to j for $j = 1, 2, \dots, n$. We can then write

$$D_{i_1 \dots i_k} f = \frac{\partial^k f}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_k}} = \frac{\partial^k f}{\partial x_1^{r_1} \partial x_2^{r_2} \cdots \partial x_n^{r_n}}.$$

Different sequences of the indices $i_1, \dots, i_k$ can lead to identical sequences of the indices $r_1, \dots, r_n$. It turns out—this follows from basic combinatorics—that the number of different i sequences that lead to the same r sequence is exactly

$$\frac{k!}{r_1! r_2! \cdots r_n!}. \tag{11.14}$$

As an aside, if each of the indices $i_1, \dots, i_k$ are different, then the denominator is exactly 1, in which case, as might be expected, the number of i sequences leading to the same r sequence (which in this case consists only of 1s) equals $k!$, the number of permutations of i indices.

Using Equation 11.14, we can then write the derivatives in Taylor's formula for the function f as

$$\sum_{i_1, \dots, i_k = 1}^{n} (D_{i_1 \dots i_k} f)(\boldsymbol{a}) w_{i_1} \cdots w_{i_k} = \sum_{\substack{0 \le r_1, \dots, r_n \le k \\ r_1 + \cdots + r_n = k}} \frac{k!}{r_1! \cdots r_n!} \frac{\partial^k f}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}}(\boldsymbol{a}) w_1^{r_1} \cdots w_n^{r_n},$$

while the full expression for Taylor's formula becomes

$$f(\boldsymbol{a} + \boldsymbol{w}) = \sum_{k=0}^{m} \left[\sum_{\substack{0 \le r_1, \dots, r_n \le k \\ r_1 + \cdots + r_n = k}} \frac{\partial^k f}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}}(\boldsymbol{a}) \frac{w_1^{r_1}}{r_1!} \cdots \frac{w_n^{r_n}}{r_n!} \right] + \sum_{\substack{0 \le r_1, \dots, r_n \le m+1 \\ r_1 + \cdots + r_n = m+1}} \frac{\partial^k f}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}}(\boldsymbol{a} + \theta \boldsymbol{w}) \frac{w_1^{r_1}}{r_1!} \cdots \frac{w_n^{r_n}}{r_n!}$$

for some $\theta \in (0, 1)$.

Remark 11.51: Simplifications of Taylor's formula

The general form of Taylor's formula given above for a function of n variables is admittedly rather complicated. Here we show three special cases often used in practice; these are considerably simpler.

(1) (*Low-order expansions*) The first-order Taylor polynomial of a function $f \colon \mathbb{R}^n \to \mathbb{R}$ of n variables $x_1, \dots, x_n$ expanded about the point $\boldsymbol{a} \in \mathbb{R}^n$ is

$$f(\boldsymbol{a}) + \left[\frac{\partial f}{\partial x_1}(\boldsymbol{a})(x_1 - a_1) + \cdots + \frac{\partial f}{\partial x_n}(\boldsymbol{a})(x_n - a_n) \right].$$

The corresponding second order Taylor polynomial is

$$f(\boldsymbol{a}) + \left[\frac{\partial f}{\partial x_1}(\boldsymbol{a})(x_1 - a_1) + \cdots + \frac{\partial f}{\partial x_n}(\boldsymbol{a})(x_n - a_n)\right] + \left[\frac{1}{2}\sum_{1\le i,j\le n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{a})(x_i - a_i)(x_j - a_j)\right].$$

(2) (*Function of two variables*) For a two-variable function $f\colon \mathbb{R}^2 \to \mathbb{R}$ we denote the two variables by x and y. If we expand the function f about the point $(x_0, y_0) \in \mathbb{R}^2$, the resulting second-order Taylor polynomial is

$$f(x_0,y_0) + \left[\frac{\partial f}{\partial x}(x_0,y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y - y_0)\right] + \frac{1}{2}\left[\frac{\partial^2 f}{\partial x^2}(x_0,y_0)(x - x_0)^2 + \frac{\partial^n f}{\partial y^n}(x_0,y_0)(y - y_0)^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x_0,y_0)(x - x_0)(y - y_0)\right].$$

The general m-th order Taylor polynomial of the two-variable function f can be written using the binomial formula in the form

$$\sum_{k=0}^{m} \frac{1}{k!}\left[\sum_{r=0}^{k} \binom{k}{r} \frac{\partial^k f}{\partial x^r \partial y^{k-r}}(x_0,y_0)(x - x_0)^r (y - y_0)^{k-r}\right].$$

Example 11.52: Taylor expansion of a multivariable polynomial

Let us compute the third-order Taylor polynomial of the two-variable function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = x^3 + y^2 + xy$$

when expanded about the point $(1,2)$. The function's partial derivatives are

$$\begin{aligned} f_x &= 3x^2 + y & f_x(1,2) &= 5 \\ f_y &= 2y + x & f_y(1,2) &= 5 \\ f_{xx} &= 6x & f_{xx}(1,2) &= 6 \\ f_{xy} &= f_{yx} = 1 & f_{yx}(1,2) &= 1 \\ f_{yy} &= 2 & f_{yy}(1,2) &= 2 \\ f_{xxx} &= 6 & f_{xxx}(1,2) &= 6; \end{aligned}$$

the remaining third-order derivatives are all equal to zero. The first-order Taylor polynomial expanded about $(1,2)$ is then

$$\begin{aligned} T_1 f\big(x,y;(1,2)\big) &= f(1,2) + f_x(1,2)\cdot(x-1) + f_y(1,2)(y-2) \\ &= 7 + 5(x-1) + 5(y-2). \end{aligned}$$

The second-order Taylor polynomial about $(1,2)$ is

$$\begin{aligned} T_2 f\big(x,y;(1,2)\big) &= T_1 f\big(x,y;(1,2)\big) + \frac{1}{2}\big[f_{xx}(1,2)(x-1)^2 + f_{yy}(1,2)(y-2)^2 \\ &\qquad + 2f_{xy}(1,2)(x-1)(y-2)\big] \\ &= 7 + 5(x-1) + 5(y-2) + 3(x-1)^2 + (y-2)^2 + (x-1)(y-2). \end{aligned}$$

Similarly, the third-order Taylor polynomial about $(1,2)$ is

$$\begin{aligned} T_3 f\big(x,y;(1,2)\big) &= 7 + 5(x-1) + 5(y-2) \\ &\qquad + 3(x-1)^2 + (y-2)^2 + (x-1)(y-2) + 6\frac{(x-1)^3}{3!}. \end{aligned}$$

As might be expected and is straightforwardly verified, this expression is exactly equal to the original function $f(x,y)$, which is itself a third-order polynomial.

11.4 Extrema of multivariable scalar-valued functions

In this section we will give a necessary condition for existence of extrema of multivariable scalar-valued functions and a sufficient condition for the existence of extrema of a two-variable function. These conditions are generalizations of those developed in Section 7.3 for scalar functions. Besides minima and maxima, multivariable functions also have critical points called *saddle points*, which do not have a scalar analog. Throughout the section, we will again assume that $n \in \mathbb{N}$ is an arbitrary natural number.

Definition 11.53: Extrema of multivariable functions

Let $f\colon U \to \mathbb{R}$ be an n-variable scalar-valued function defined on the subset $U \subset \mathbb{R}^n$, and define the point $\boldsymbol{a} \in U$.

(i) The function f has a *local maximum* at $\boldsymbol{a}$ if there exists $\delta \in \mathbb{R}^+$ for which

$$f(\boldsymbol{v}) \leq f(\boldsymbol{a}) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a},\delta) \cap U.$$

(ii) The function f has a *local minimum* at $\boldsymbol{a}$ if there exists $\delta \in \mathbb{R}^+$ for which

$$f(\boldsymbol{v}) \geq f(\boldsymbol{a}) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a},\delta) \cap U.$$

(iii) The function f has a *local extremum* at $\boldsymbol{a}$ if it has either a local maximum or a local minimum at $\boldsymbol{a}$.

(iv) The function f has a *strict local maximum* at $\boldsymbol{a}$ if there exists $\delta \in \mathbb{R}^+$ for which

$$f(\boldsymbol{v}) < f(\boldsymbol{a}) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a},\delta) \cap (U \setminus \{a\}).$$

(v) The function f has a *strict local minimum* at $\boldsymbol{a}$ if there exists $\delta \in \mathbb{R}^+$ for which

$$f(\boldsymbol{v}) > f(\boldsymbol{a}) \text{ for all } \boldsymbol{v} \in \mathcal{K}(\boldsymbol{a},\delta) \cap (U \setminus \{a\}).$$

(vi) The function f has a *strict local extremum* at $\boldsymbol{a}$ if it has either a strict local maximum or a strict local minimum at $\boldsymbol{a}$.

(vii) The function f has a *global maximum* at $\boldsymbol{a}$ if $f(\boldsymbol{v}) \leq f(\boldsymbol{a})$ for all $\boldsymbol{v} \in U$, and a *strict global maximum* at $\boldsymbol{a}$ if $f(\boldsymbol{v}) < f(\boldsymbol{a})$ for all $\boldsymbol{v} \in U \setminus \{\boldsymbol{a}\}$. In both cases we then write

$$\max(f) = f(\boldsymbol{a}).$$

(viii) The function f has a *global minimum* at $\boldsymbol{a}$ if $f(\boldsymbol{v}) \geq f(\boldsymbol{a})$ for all $\boldsymbol{v} \in U$, and a *strict global minimum* at $\boldsymbol{a}$ if $f(\boldsymbol{v}) > f(\boldsymbol{a})$ for all $\boldsymbol{v} \in U \setminus \{\boldsymbol{a}\}$. In both cases we then write

$$\min(f) = f(\boldsymbol{a}).$$

(ix) The function f has an *(strict) global extremum* at $\boldsymbol{a}$ if it has either a (strict) maximum or a (strict) minimum at $\boldsymbol{a}$.

Definition 11.54: Stationary point of a multivariable function

Let $f\colon U \to \mathbb{R}$ be an n-variable function defined on the subset $U \subset \mathbb{R}^n$. An interior point $\boldsymbol{a} \in U$ is called a *stationary point* of f if f is totally differentiable at $\boldsymbol{a}$ and

$$(\boldsymbol{\nabla} f)(\boldsymbol{a}) = \boldsymbol{0}.$$

Proposition 11.55: Necessary condition for the extremum of a multivariable function

Let $f\colon U \to \mathbb{R}$ be an n-variable function defined on the subset $U \subset \mathbb{R}^n$, and assume f is totally differentiable at some interior point $\boldsymbol{a} \in U$. In this case if f has a local extremum at $\boldsymbol{a}$, then $\boldsymbol{a}$ is necessarily a stationary point of f.

Proof. Let $\boldsymbol{u} \in \mathbb{R}^n$ be an arbitrary vector and let $\boldsymbol{\gamma}\colon \mathbb{R} \to \mathbb{R}^n$ be the smooth vector-valued function given by

$$\boldsymbol{\gamma}(t) = \boldsymbol{a} + t\boldsymbol{u}.$$

In this case $\boldsymbol{\gamma}(0) = \boldsymbol{a}$ and $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{u}$ and the single-variable function $f \circ \boldsymbol{\gamma}\colon \boldsymbol{\gamma}^{-1}(U) \to \mathbb{R}$ is differentiable at 0. Since the function f has a local extremum at $\boldsymbol{a}$, the function $f \circ \boldsymbol{\gamma}$ has an extremum at 0, and so, by the necessary condition for the extrema of scalar-valued functions in Proposition 7.14,

$$(f \circ \boldsymbol{\gamma})'(0) = 0. \tag{11.15}$$

We then make the calculation

$$(\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot \boldsymbol{u} \overset{\text{(a)}}{=} (\boldsymbol{\nabla} f)\big(\boldsymbol{\gamma}(0)\big) \cdot \dot{\boldsymbol{\gamma}}(0) \overset{\text{(b)}}{=} (f \circ \boldsymbol{\gamma})'(0) \overset{\text{(c)}}{=} 0,$$

where (a) follows from the definition of $\boldsymbol{\gamma}$, (b) from the chain rule, and (c) from Equation 11.15. Since the vector $\boldsymbol{u} \in \mathbb{R}^n$ was arbitrary and f is totally differentiable at $\boldsymbol{a}$, we can choose $\boldsymbol{u} = (\boldsymbol{\nabla} f)(\boldsymbol{a})$, which we then substitute into the previous equality to get

$$(\boldsymbol{\nabla} f)(\boldsymbol{a}) \cdot (\boldsymbol{\nabla} f)(\boldsymbol{a}) = 0 \overset{\text{(a)}}{\Longrightarrow} (\boldsymbol{\nabla} f)(\boldsymbol{a}) = \boldsymbol{0},$$

where (a) follows from the positive-definiteness of the scalar product. □

In Proposition 7.25 we showed that a first derivative equal to zero together with a nonzero second derivative was a sufficient condition for the extrema of a scalar function. Formulating a sufficient condition for extrema of multivariable functions is somewhat more involved—it involves a matrix of second derivative values called the *Hessian matrix*.

Since we have not covered matrices or linear algebra in this course, we will consider only the case of a two-variable function, which can be straightforwardly formulated without detailed knowledge of matrices. To this end we now define the determinant of the Hessian matrix of a two-variable function.

Remark 11.56: The Hessian matrix of a scalar-valued multivariable function

The *determinant of the Hessian matrix* of a scalar-valued function of two variables $f\colon U \to \mathbb{R}$ defined on an open subset $U \subset \mathbb{R}^2$ is the scalar-valued function $\det \mathbf{H}(f)\colon U \to \mathbb{R}$ given by

$$\det \mathbf{H}(f)(x,y) = \frac{\partial^2 f}{\partial x^2}(x,y)\frac{\partial^2 f}{\partial y^2}(x,y) - \left[\frac{\partial^2 f}{\partial x \partial y}(x,y)\right]^2.$$

We can now give a sufficient condition for the existence of extrema of a two-variable function.

Proposition 11.57: Sufficient condition for the extrema of a two-variable function

Let $f\colon U \to \mathbb{R}$ be a twice continuously-differentiable function of the two variables x and y defined on the open subset $U \subset \mathbb{R}^2$ and let $(x_0, y_0) \in U$ be a stationary point of the function f. In this case

(i) If $\det \mathbf{H}(f)(x_0, y_0) > 0$ and $f_{xx}(x_0, y_0) < 0$, then f has a strict local maximum at the stationary point (x_0, y_0).

(ii) If $\det \mathbf{H}(f)(x_0, y_0) > 0$ and $f_{xx}(x_0, y_0) > 0$, then f has a strict local minimum at the stationary point (x_0, y_0).

(iii) If $\det \mathbf{H}(f)(x_0, y_0) < 0$, then f does not have a local extremum at the stationary point (x_0, y_0). Instead, f is said to have a *saddle point* at (x_0, y_0).

Proof. Because the point (x_0, y_0) is a stationary point of the function f, all of f's first partial derivatives at (x_0, y_0) are zero, and so for sufficiently small vectors $(h, k) \in \mathbb{R}^2$, f's Taylor expansion about the point (x_0, y_0) reads

$$\begin{aligned} f(x_0+h, y_0+k) = f(x_0, y_0) + \frac{1}{2}\Bigg[&\frac{\partial^2 f}{\partial x^2}(x_0+\theta h, y_0+\theta k)h^2 \\ &+ 2\frac{\partial^2 f}{\partial x \partial y}(x_0+\theta h, y_0+\theta k)hk \\ &+ \frac{\partial^2 f}{\partial y^2}(x_0+\theta h, y_0+\theta k)k^2\Bigg]. \end{aligned}$$

We first introduce the more compact notation

$$\alpha = \alpha(h,k) = \frac{\partial^2 f}{\partial x^2}(x_0 + \theta h, y_0 + \theta k)$$
$$\beta = \beta(h,k) = \frac{\partial^2 f}{\partial x \partial y}(x_0 + \theta h, y_0 + \theta k)$$
$$\gamma = \gamma(h,k) = \frac{\partial^2 f}{\partial y^2}(x_0 + \theta h, y_0 + \theta k),$$

in terms of which f's Taylor expansion about the point (x_0, y_0) reads

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + \frac{1}{2}\big[\alpha(h,k)h^2 + 2\beta(h,k)hk + \gamma(h,k)k^2\big].$$

In the context of this proof, we are interested in the sign of the quantity

$$B = B(h,k) = \alpha(h,k)h^2 + 2\beta(h,k)hk + \gamma(h,k)k^2;$$

if this quantity is positive (negative) for all sufficiently small vectors (h,k), the f has a strict local minimum (maximum) at the point (x_0, y_0).

By the proposition's assumptions, $\det \mathbf{H}(f)(x_0, y_0) \neq 0$. Because $f \in C^2(U)$, all of its second partial derivatives are continuous, and so the function $\det \mathbf{H}(f)$ is itself continuous; as a result, the function $\det \mathbf{H}(f)$ is nonzero and of the same sign as $\det \mathbf{H}(f)(x_0, y_0)$ in any sufficiently small neighborhood of the point (x_0, y_0).

To prove points (i) and (ii), we will assume that $\det \mathbf{H}(f)(x_0, y_0) > 0$. As a result $\alpha(h,k) \cdot \beta(h,k) > 0$ for sufficiently small (h,k). More so, because the function α is continuous, the sign of the function value $\alpha(h,k)$ is the same as the sign of $\alpha(0,0)$ for sufficiently small (h,k). If $k \neq 0$, we may define $t = h/k$ and write

$$B = \alpha h^2 + 2\beta hk + \gamma k^2 = k^2(\alpha t^2 + 2\beta t + \gamma).$$

The expression for B is a quadratic function of the variable t with a discriminant equal to

$$D = (2\beta)^2 - 4\alpha\gamma = 4(\beta^2 - \alpha\gamma) = -4 \det \mathbf{H}(f)(x_0 + h, y_0 + k).$$

For sufficiently small (h,k), the discriminant D is negative and the quantity B is of the same sign as α. Additionally, in the case when $k = 0$ and h is nonzero and sufficiently small, the quantity

$$\alpha h^2 + 2\beta hk + \gamma k^2 = \alpha h^2$$

is of the same sign as α. For sufficiently small small nonzero (h,k) the quantity

$$\alpha h^2 + 2\beta hk + \gamma k^2$$

is thus negative when α is negative and positive when α is positive. If follows that f has a strict local maximum at (x_0, y_0) in the case (i) and a strict local minimum at (x_0, y_0) in the case (ii).

To prove point (iii), we will assume $\det \mathbf{H}(f)(x_0, y_0) < 0$ and show that in this case there exist arbitrarily small vectors (h_1, k_1) and (h_2, k_2) for which

$$\begin{aligned} f(x_0 + h_1, y_0 + k_1) &> f(x_0, y_0) \\ f(x_0 + h_2, y_0 + k_2) &< f(x_0, y_0), \end{aligned}$$

which means that f does not have an extremum at (x_0, y_0). To begin, the inequality $\det \mathbf{H}(f)(x_0, y_0) < 0$ implies that the quadratic function

$$\alpha(0,0)t^2 + 2\beta(0,0)t + \gamma(0,0)$$

has a positive discriminant, which means there exist real numbers $t_1, t_2 \in \mathbb{R}$ for which

$$\begin{aligned} \alpha(0,0)t_1^2 + 2\beta(0,0)t_1 + \gamma(0,0) &> 0 \\ \alpha(0,0)t_2^2 + 2\beta(0,0)t_2 + \gamma(0,0) &< 0. \end{aligned}$$

More so, the continuity of the function f's second partial derivatives means that for sufficiently small (h, k), it holds that

$$\begin{aligned} \alpha(h,k)t_1^2 + 2\beta(h,k)t_1 + \gamma(h,k) &> 0 \\ \alpha(h,k)t_2^2 + 2\beta(h,k)t_2 + \gamma(h,k) &< 0. \end{aligned}$$

Let us now choose h_1 and k_1 for which $t_1 = h_1/k_1$, and note that the set of all such pairs (h_1, k_1) satisfying $t_1 = h_1/k_1$ forms a line in the plane. If h_1 and k_1 are sufficiently small and $k_1 \neq 0$, then

$$\alpha(h_1,k_1)t_1^2 + 2\beta(h_1,k_1)t_1 + \gamma(h_1,k_1) > 0,$$

which in turn implies that

$$\alpha(h_1,k_1)h_1^2 + 2\beta(h_1,k_1)h_1k_1 + \gamma(h_1,k_1)k_1^2 > 0.$$

We have thus found arbitrarily small vectors (h_1, k_1) for which

$$\begin{aligned} f(x_0 + h_1, y_0 + k_1) = f(x_0, y_0) &+ \frac{1}{2}\big[\alpha(h_1,k_1)h_1^2 \\ &+ 2\beta(h_1,k_1)h_1k_1 + \gamma(h_1,k_1)k_1^2\big] \\ &> f(x_0, y_0). \end{aligned}$$

Analogously, we can find arbitrarily small vectors (h_2, k_2) for which $t_2 = h_2/k_2$ and

$$f(x_0 + h_2, y_0 + k_2) < f(x_0, y_0).$$

Combining the results $f(x_0+h_1, y_0+k_1) > f(x_0, y_0)$ and $f(x_0+h_2, y_0+k_2) < f(x_0, y_0)$ implies that f does not have an extremum at the point (x_0, y_0). □

Example 11.58: Extrema of two-variable functions

(1) Let us find the extrema of the two-variable function

$$f(x,y) = x^3 + y^3 - 3xy.$$

We will first find the function f's stationary points—f's first partial derivatives are

$$f_x = 3x^2 - 3y \quad \text{and} \quad f_y = 3y^2 - 3x,$$

and so any of f's stationary points must satisfy the system of equations

$$\begin{aligned} f_x &= 3x^2 - 3y = 0, \\ f_y &= 3y^2 - 3x = 0. \end{aligned}$$

The first equation implies $x^2 = y$, while the second implies that $y^2 = x$; we square the second equation and substitute the result into the second equation to get

$$y^4 = y.$$

The equation $y^4 = y$ has two real solutions—$y = 0$ and $y = 1$, which combined with the equations $x^2 = y$ and $y^2 = x$ lead to the corresponding x values $x = 0$ and $x = 1$. The function f thus has two stationary points; these are

$$P_1(0,0) \quad \text{and} \quad P_2(1,1).$$

The function f's second partial derivatives are

$$f_{xx} = 6x, \qquad f_{xy} = f_{yx} = -3, \qquad f_{yy} = 6y,$$

and so the determinant of f's Hessian matrix is

$$\det \mathbf{H}(f)(x,y) = 6x \cdot 6y - (-3)^2 = 36xy - 9.$$

At f's stationary points, the Hessian matrix's determinant equals

$$\det \mathbf{H}(f)(0,0) = -9 < 0 \quad \text{and} \quad \det \mathbf{H}(f)(1,1) = 27 > 0,$$

and so f has a saddle point at P_1 and, because $f_{xx}(1,1) = 6 > 0$, a strict local minimum at P_2.

(2) Let us now consider the function given by

$$f(x,y) = 2x^2 + 3y^2 - 4x + 2 = 2(x-1)^2 + 3y^2.$$

The function f's first partial derivatives are

$$f_x = 4x - 4 \quad \text{and} \quad f_y = 6y,$$

and, without derivation, f has a single stationary point, $P_1(1,0)$. The function f's second partial derivatives are

$$f_{xx} = 4, \qquad f_{xy} = f_{yx} = 0, \qquad f_{yy} = 6,$$

and so the determinant of f's Hessian matrix is

$$\det \mathbf{H}(f)(x,y) = 4 \cdot 6 - 0 = 24 > 0.$$

Because $\det \mathbf{H}(f)(x,y) > 0$ for all $(x,y) \in \mathbb{R}^2$, the function f has a strict local minimum at its single stationary point $P_1(1,0)$.

(3) Finally, let us consider the function given by

$$f(x, y) = y^2 - x^2.$$

The function's first partial derivatives are

$$f_x = -2x \quad \text{and} \quad f_y = 2y,$$

which implies that f has a single stationary point, $P_1(0, 0)$. The function f's second partial derivatives are

$$f_{xx} = -2, \qquad f_{xy} = 0, \qquad f_{yy} = 2,$$

and so the determinant of f's Hessian matrix is

$$\det \mathbf{H}(f)(x, y) = (-2) \cdot 2 = 0 = -4 < 0.$$

Because $\det \mathbf{H}(f)(x, y) < 0$ for all $(x, y) \in \mathbb{R}^2$, the function f does not have an extremum (but instead has a saddle point) at its single stationary point $P_1(0, 0)$.