University of Chicago
“A Textbook for Advanced Calculus”

John Boller and Paul J. Sally, Jr.

Chapter 0

Number Systems and Cardinality

> *Dans la présente Note, on va essayer de préciser une terminologie propre à l'étude des ensembles abstraits. Cette étude a pour but principal d'étendre les propriétés des ensembles linéaires à des ensembles de plus en plus généraux, et par voie de conséquence, de disséquer ces propriétés et d'en rechercher pour chacune la véritable origine. Il en résulte que le fond des démonstrations est pour ainsi dire donné d'avance et que la difficulté consiste en grande partie a préciser dans quel cadre elles viendront se placer. Adopter une terminologie, c'est donc tracer d'avance toute la théorie. Au fur et à mesure du développement de cette théorie, la terminolgie a varié et variera encore. Mais il n'est peut-être pas inutile d'en proposer une, adaptée à la situation présente.*
>
> *– M. M. Fréchet,*
>
> *Extrait des Comptes rendus du Congrès des Sociétés savantes en 1924.*

0.1 The Integers

The set of natural numbers is the familiar collection $\mathbb{N} = \{1, 2, 3, \ldots, n, \ldots\}$. It would be possible to rigorously develop the properties of the natural numbers deriving from the Peano postulates. We choose not to do this here, but we refer the interested reader to [La].

We do wish to take a more formal approach towards another familiar set of numbers, namely the integers. The integers form the collection $\{0, 1, -1, 2, -2, \ldots\}$, which we study in elementary arithmetic. We denote the integers by the symbol $\mathbb{Z}$ (from the German word *Zahlen*). The operations in the integers are addition $(+)$ and multiplication $(\cdot)$, and here are the rules. We expect that the reader is well versed in the arithmetic of the integers, but we are stating these properties explicitly for two reasons. First, these properties are used in arithmetic from the earliest grades, but are seldom justified. Second, these properties will be used to describe other algebraic structures that we will meet later.

Rules of Arithmetic in $\mathbb{Z}$ 0.1.1

$$\left.\begin{array}{ll} \text{(A1)} & \text{If } a, b \in \mathbb{Z},\ \text{then } a + b \in \mathbb{Z}. \\ \text{(M1)} & \text{If } a, b \in \mathbb{Z},\ \text{then } a \cdot b \in \mathbb{Z}. \end{array}\right\} \text{Closure}$$

$$\left.\begin{array}{ll} \text{(A2)} & \text{If } a, b, c \in \mathbb{Z},\ \text{then } a + (b + c) = (a + b) + c. \\ \text{(M2)} & \text{If } a, b, c \in \mathbb{Z},\ \text{then } a \cdot (b \cdot c) = (a \cdot b) \cdot c. \end{array}\right\} \text{Associativity}$$

$$\left.\begin{array}{ll} \text{(A3)} & \text{If } a, b \in \mathbb{Z},\ \text{then } a + b = b + a. \\ \text{(M3)} & \text{If } a, b \in \mathbb{Z},\ \text{then } a \cdot b = b \cdot a. \end{array}\right\} \text{Commutativity}$$

$$\left.\begin{array}{ll}\mathsf{(A4)} & \exists\, 0 \in \mathbb{Z} \ni \forall a \in \mathbb{Z},\ a+0=0+a=a. \\ \mathsf{(M4)} & \exists\, 1 \in \mathbb{Z} \ni \forall a \in \mathbb{Z},\ a\cdot 1=1\cdot a=a. \end{array}\right\}\text{Identities}$$

$$\mathsf{(A5)} \quad \forall a \in \mathbb{Z},\ \exists -a \in \mathbb{Z} \ni a+(-a)=(-a)+a=0. \ \}\ \text{Additive inverses}$$

In general, elements in $\mathbb{Z}$ do not have multiplicative inverses in $\mathbb{Z}$. That is, given an element $a \in \mathbb{Z}$, we cannot necessarily find another element $b \in \mathbb{Z}$ such that $ab = 1$. However, some integers do have multiplicative inverses, namely 1 and -1.

The operations of addition and multiplication are tied together by the distributive law.

(D) If $a, b, c \in \mathbb{Z}$, then $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

Without the distributive law, there would be no connection between addition and multiplication. The richness of the structure is embodied in the interaction between the two operations.

Let's stop and investigate some of the implications of these 10 axioms.

Facts 0.1.2

1. Additive identities are unique.

   Proof. Suppose that 0 and $0'$ are additive identities. Then $0 = 0 + 0' = 0'$.

2. Multiplicative identities are unique.

   Proof. Exercise. (Hint: Use the same technique as above.)

3. Additive inverses are unique.

   Proof. Suppose that $a \in \mathbb{Z}$ and $a + a' = 0$. Then $-a + (a + a') = -a + 0 = -a$. On the other hand, by associativity ((A2)), we have $-a + (a + a') = ((-a) + a) + a' = 0 + a' = a'$. Thus, $a' = -a$.

4. (Cancellation for addition) If $a, b, c \in \mathbb{Z}$ and $a + b = a + c$, then $b = c$.

   Proof. If $a + b = a + c$, then $-a + (a + b) = -a + (a + c)$. By associativity ((A2)), $((-a) + a) + b = ((-a) + a) + c$, and hence $0 + b = 0 + c$, from which we conclude that $b = c$.

5. If $a \in \mathbb{Z}$, then $a \cdot 0 = 0$.

   Proof. We can write

$$\begin{aligned} a \cdot 0 &= a \cdot (0 + 0) \\ (a \cdot 0) + 0 &= a \cdot 0 + a \cdot 0 \end{aligned}$$

   by properties of the additive identity and the distributive law. Now cancel to get $a \cdot 0 = 0$.

   This is really quite something, and it emphasizes the role of the distributive law. What we have here is multiplication by the additive identity reproducing the additive identity. We have more interaction between multiplication and addition in the following statements.

6. If $a \in \mathbb{Z}$, then $(-1) \cdot a = -a$.

   Proof. We can write $a + (-1) \cdot a = 1 \cdot a + (-1) \cdot a = (1 + (-1)) \cdot a = 0 \cdot a = 0$. But additive inverses are unique, so $-a = (-1) \cdot a$.

   Notice that this really says something. That is, the left-hand expression, $(-1) \cdot a$, represents the additive inverse of the multiplicative identity multiplied by a. The right-hand side, $-a$, on the other hand, represents the additive inverse of a.

Notice that, when convenient, we drop the dot which signifies multiplication.

Exercise 0.1.3 If $a, b \in \mathbb{Z}$, then $(-a)b = a(-b) = -(ab)$.

Exercise 0.1.4 If $a, b \in \mathbb{Z}$, then $(-a)(-b) = ab$.

Now, what other properties do the integers have? In the integers, cancellation for multiplication doesn't follow from the first 10 axioms. Cancellation for multiplication should be familiar; many texts introduce it as an additional axiom for the integers in the following form.

(C) If $a, b, c \in \mathbb{Z}$ with $a \neq 0$ and $ab = ac$, then $b = c$.

Exercise 0.1.5 Why is $a = 0$ excluded?

However, we will see shortly that because the integers are also ordered, cancellation in the integers is a consequence of the order properties.

Exercise 0.1.6 Cancellation can be phrased in another way. Show that the statement "if $a, b \in \mathbb{Z}$ and $ab = 0$, then either $a = 0$ or $b = 0$" is equivalent to cancellation.

What else do we have for the integers? We have inequalities. The $<$ sign should be familiar to you. It is subject to the following *rules of order*.

(O1) If $a, b \in \mathbb{Z}$, then one and only one of the following holds: $a < b$, $a = b$, or $b < a$. (Trichotomy)

(O2) If $a, b, c \in \mathbb{Z}$ with $a < b$ and $b < c$, then $a < c$. (Transitivity)

(O3) If $a, b, c \in \mathbb{Z}$ and $a < b$, then $a + c < b + c$. (Addition)

(O4) If $a, b, c \in \mathbb{Z}$, $a < b$, and $0 < c$, then $ac < bc$. (Multiplication by positive elements)

We adopt the usual notation and terminology. That is, if $a < b$, we say that "a is less than b." If $a < b$ or $a = b$, we say that "a is less than or equal to b" and write $a \leq b$. If $a < b$ we may also write $b > a$ and say that "b is greater than a." The statement $b \geq a$ is now self-explanatory.

Here are some examples of recreational exercises and facts which go with the order axioms. For these statements and the following exercises, let $a, b, c \in \mathbb{Z}$.

Facts 0.1.7

1. $a > 0$ iff $-a < 0$.

   Proof. Suppose $a > 0$. Add $-a$ to both sides.

2. If $a > 0$ and $b > 0$, then $ab > 0$.

   Proof. Suppose $a > 0$. Then, since $b > 0$, $ab > 0 \cdot b = 0$.

3. If $a > 0$ and $b < 0$, then $ab < 0$.

   Proof. Suppose $a > 0$ and $b < 0$. Then $-b > 0$ and $a(-b) = -(ab) > 0$. So $ab < 0$.

4. If $a < 0$ and $b < 0$, then $ab > 0$.

   Proof. If $a < 0$ and $b < 0$, then $-a > 0$ and $-b > 0$. Hence $(-a)(-b) = ab > 0$.

5. If $a \neq 0$, then $a^2 > 0$.

   Proof. If a is greater then 0, use Fact 2. If a is less then 0, use Fact 4.

6. $1 > 0$.

   Proof. $1 = 1^2$.

7. If $a > b$ and $c < 0$, then $ac < bc$.

   Proof. If $a > b$, then $a - b > 0$. Since $-c > 0$, $(-c)(a - b) = -ac + bc > 0$. Hence, $bc > ac$.

8. If $a > b$, then $-a < -b$.

   Proof. Let $c = -1$.

Are you having fun yet? Good, try these exercises.

Exercise 0.1.8 Suppose that $0 < a$ and $0 < b$. Show that $a < b$ iff $a^2 < b^2$.

Exercise 0.1.9 Suppose that $a < 0$ and $b < 0$. Show that $a < b$ iff $b^2 < a^2$.

Exercise 0.1.10 Show that $2ab \leq a^2 + b^2$.

The set $\mathbb{N}$ of natural numbers is the set of positive elements in $\mathbb{Z}$, that is, the set of elements which are greater than 0. It is clear that $\mathbb{N}$ is closed under addition and multiplication. If we add trichotomy, these properties lead to an alternate characterization of order.

Exercise 0.1.11 Suppose now that we have only the first 10 axioms for $\mathbb{Z}$ as well as the cancellation property (C). Let P be a set of integers with the following properties.

1. If $a \in \mathbb{Z}$, then one and only one of the following holds: $a \in P$, $a = 0$, or $-a \in P$.

2. If $a, b \in P$, then $a + b \in P$ and $ab \in P$.

For $a, b \in \mathbb{Z}$, define $a < b$ if $b - a \in P$. Show that this relation satisfies (O1)–(O4). Moreover, if we have a relation that satisfies (O1)–(O4), and we define $P = \{a \in \mathbb{Z} \mid a > 0\}$, then show that P satisfies properties 1 and 2 above.

Exercise 0.1.12 Show that the cancellation property (C) can be proved using the axioms for addition and multiplication and the order axioms.

So far, the integers have five axioms for addition, four for multiplication, one for the distributive law, and four for order. There is one more axiom which plays a crucial role. It is called the *Well-Ordering Principle.* This Principle assures us that 1 is the smallest positive integer. This should not come as a surprise but we do need something to confirm this. In the rational numbers, which we construct in the next section, the first fourteen axioms are satisfied, but there is actually no smallest positive element. Thus, we need to introduce the Well-Ordering Principle as an axiom for $\mathbb{Z}$.

(0.1.13 WO) Well-Ordering Principle for $\mathbb{Z}$ If A is a nonempty subset of the positive integers, then A has a least element. That is, there exists an element $a_0 \in A$, such that for all $a \in A$, $a_0 \leq a$.

That does it! We now have the 15 properties, and they completely characterize the integers. (For a proof of this, see Project 2 in this chapter.) Most of the work with the Well-Ordering Principle will be done later. However, here are a couple of facts which follow immediately from the Well-Ordering Principle.

Facts 0.1.14

1. There are no integers between 0 and 1.

   Proof. Let $A = \{a \in \mathbb{Z} \mid 0 < a < 1\}$. If $A \neq \varnothing$, then it has a least element a_0 which is in A. So, $0 < a_0 < 1$, and, by property (O4), $0 < a_0^2 < a_0$. But then $a_0^2 \in A$ and a_0 is not the least element.

2. (Mathematical Induction) Let A be a set of positive integers such that $1 \in A$, and if $k \in A$, then $k+1 \in A$. Then A is the set of all positive integers.

   Proof. Suppose there exists a positive integer which is not in A, and let A' be the set of all such positive integers. Then A' is a nonempty subset of the positive integers, and hence has a least element c. Now $c > 1$ since $1 \in A$, and there is no integer between 0 and 1. So $c-1$ is an integer greater than 0. Since $c-1 < c$, it follows that $c-1 \in A$. And, so, $(c-1)+1 = c$ is also in A, which is a contradiction.

Exercise 0.1.15 If n and k are non-negative integers with $n \geq k$, we define the *binomial coefficient* $\binom{n}{k}$ by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $n! = n(n-1)\cdots 2 \cdot 1$, and we set $0! = 1$ (this will be explained later in the book when we discuss the Gamma function). Prove the *Binomial Theorem*: If $a, b \in \mathbb{Z}$ and n is a positive integer, then

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

(Use Mathematical Induction.)

Remark 0.1.16 Observe that the binomial coefficient $\binom{n}{k}$ represents the number of ways of choosing k objects from n objects where order does not matter. The binomial coefficient $\binom{n}{k}$ is the number of subsets of k elements in a set with n elements. Of course the binomial theorem implies that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$, the total number of subsets of a set with n elements.

Exercise 0.1.17 *i.* Prove by induction that if A and B are finite sets, A with n elements and B with m elements, then $A \times B$ has nm elements.

ii. Prove by induction the corresponding result for a collection of k finite sets, where $k > 2$.

0.2 Equivalence Relations and the Construction of $\mathbb{Q}$

Next we turn to the idea of a relation on a set. Here is the formal definition of a relation.

Definition 0.2.1 A *relation* on a set X is a subset R of $X \times X$.

For example, we can define a relation on $\mathbb{Z}$ by setting R equal to $\{(a,b) | a, b \in \mathbb{Z} \text{ and } a < b\}$.

Equivalence relations are a special type of relation, which we define below. They appear everywhere in mathematics, and we really mean that. What an equivalence relation does is take a set and partition it into subsets. Some equivalence relations appear to be very natural, some appear to be supernatural, and others appear to make no sense at all.

Definition 0.2.2 Let X be a set. An *equivalence relation* on X is a relation R on X such that

(ER1) For all $a \in X$, $(a,a) \in R$. (Reflexive)

(ER2) For $a, b \in X$, if $(a,b) \in R$, then $(b,a) \in R$. (Symmetric)

(ER3) For $a, b, c \in X$, if $(a,b), (b,c) \in R$, then $(a,c) \in R$. (Transitive)

The "twiddle" notation $(\sim)$ is often used in mathematics. Here we use it as follows: if $(a,b) \in R$, we write $a \sim b$. Then the definition of equivalence relation becomes

(ER1) For all $a \in X$, $a \sim a$. (Reflexive)

(ER2) For $a, b \in X$, if $a \sim b$ then $b \sim a$. (Symmetric)

(ER3) For $a, b, c \in X$, if $a \sim b$ and $b \sim c$, then $a \sim c$. (Transitive)

Again, speaking loosely, we can refer to $\sim$ as an equivalence relation on X.

Exercise 0.2.3 Let R be a relation on X that satisfies the following two conditions.

a. For all $a \in X$, $(a, a) \in R$.

b. For $a, b, c \in X$ if $(a, b), (b, c) \in R$, then $(c, a) \in R$.

Show that R is an equivalence relation.

Example 0.2.4 The most basic example of an equivalence relation is equality. That is, $a \sim b$ iff $a = b$. Prove this, but please don't write anything.

Example 0.2.5 If A and B are triangles in the plane, write $A \sim B$ if and only if A is similar to B.

Example 0.2.6 Let n be an integer greater than or equal to 2. If $a, b \in \mathbb{Z}$, we say that $a \sim b$ iff $a - b$ is a multiple of n, that is, n divides $a - b$.

This last example requires a little more elucidation. So, we present a brief discussion about divisibility in $\mathbb{Z}$.

Definition 0.2.7 Suppose that a and b are integers. We say that a *divides* b, written $a|b$, if there is an element $c \in \mathbb{Z}$ such that $b = ac$. When a divides b, the number a is called a *divisor* of b.

We need the following facts about divisibility.

Facts 0.2.8

1. If $a \in \mathbb{Z}$, then $a|a$.
2. If $a|b$ then $a| - b$.
3. If $a|b$ and $b|c$, then $a|c$.

These facts are easy to prove. For example, if $a|b$ and $b|c$, there are integers h and k such that $b = ha$ and $c = kb$. But then $c = (hk)a$, and since hk is an integer by axiom $(\mathsf{M1})$, that does it.

Exercise 0.2.9 Show that, if $a \in \mathbb{Z}$, then $a|0$.

Exercise 0.2.10 Show that, if a and b are integers such that $a|b$ and $b|a$, then $a = \pm b$.

Exercise 0.2.11 Show that, if $c|a$ and $c|b$, and $s, t \in \mathbb{Z}$, then $c|(sa + tb)$.

There is one other type of integer that should be familiar to the reader.

Definition 0.2.12 Let p be a positive integer greater than or equal to 2. We say that p is *prime* if the only positive divisors of p are 1 and p.

If n is a positive integer greater than 2 which is not prime, then n is called *composite*. So, if n is composite there exist integers a and b both greater than or equal to 2, such that $n = ab$.

Exercise 0.2.13 Let n be a positive integer greater than or equal to 2. Then there exists a prime p such that p divides n.

With this discussion of divisibility under our belt, we define the notion of congruence in the integers.

Let n be an integer greater than or equal to 2. For $a, b \in \mathbb{Z}$, we say that a is *congruent to b modulo n*, $a \equiv b \pmod{n}$, provided that $n \mid a - b$.

Exercise 0.2.14 For a fixed integer $n \geq 2$, we define $a \sim b$ if and only if $a \equiv b \pmod{n}$. Show that this is an equivalence relation on $\mathbb{Z}$.

Now we return to equivalence relations in general. The partitioning into subsets relative to an equivalence relation comes about as follows. If $a \in X$, we write $C(a) = \{b \in X \mid b \sim a\}$. $C(a)$ is called *the class of a* or *the equivalence class containing a*. Here are the properties of equivalence classes.

Theorem 0.2.15 (Properties of equivalence classes)

1. $a \in C(a)$.

   Proof. Reflexivity.

2. If $a \sim b$, then $C(a) = C(b)$.

   Proof. Transitivity.

3. If a is not equivalent b ($a \nsim b$), then $C(a) \cap C(b) = \varnothing$.

   Proof. If $c \in C(a) \cap C(b)$, then $c \sim a$ and $c \sim b$, so $a \sim b$. So $C(a) \cap C(b) \neq \varnothing$ iff $C(a) = C(b)$.

4. $\bigcup_{a \in X} C(a) = X$.

   Proof. Use (1) above.

This all means that an equivalence relation on a set X partitions X into a collection of pairwise disjoint subsets. Although this looks quite special, it's really not that impressive. For example, take a set X and break it up into pairwise disjoint nonempty subsets whose union is all of X. Then, for $a, b \in X$, define $a \sim b$ if a and b are in the same subset.

Exercise 0.2.16 Prove that this is an equivalence relation on X.

One more example of an equivalence relation will prove useful for future developments. This is a method for constructing the rational numbers $\mathbb{Q}$ from the integers $\mathbb{Z}$ using the properties discussed in the last section. We consider the set $F = \{(a, b) \mid a, b \in \mathbb{Z} \text{ and } b \neq 0\}$. We are thinking (for example) of the pair $(2, 3)$ as the fraction $\frac{2}{3}$. For $(a, b), (c, d) \in F$, we define $(a, b) \sim (c, d)$ if $ad = bc$. Thus, for instance, $(2, 3) \sim (8, 12) \sim (-6, -9)$.

Exercise 0.2.17 Show that $\sim$ is an equivalence relation on F.

The set of equivalence classes determined by this equivalence relation is called the *rational numbers* and is denoted by $\mathbb{Q}$. You should be extremely happy about this since it explains all that business about equivalent fractions that you encountered in elementary school. What a relief!

We have several things to do with this example. First, we have to add and multiply rational numbers, that is, add and multiply equivalence classes. The fundamental principle to be established here is that, when we add or multiply equivalence classes, we do it by selecting an element from each equivalence class and adding or multiplying these. We must be certain that the result is independent of the representatives that we choose in the equivalence classes. For simplicity, we denote the class of (a, b) by $\{(a, b)\}$ rather than $C((a, b))$.

For $\{(a,b)\}, \{(c,d)\} \in \mathbb{Q}$, we define addition and multiplication as follows.

$$\begin{aligned}\{(a,b)\} + \{(c,d)\} &= \{(ad+bc, bd)\}, and\\ \{(a,b)\} \cdot \{(c,d)\} &= \{(ac, bd)\}.\end{aligned}$$

What we must establish is the fact that if $(a,b) \sim (a',b')$ and $(c,d) \sim (c',d')$, then $(ad+bc, bd) \sim (a'd'+b'c', b'd')$ and $(ac,bd) \sim (a'c', b'd')$. All this requires is a little elementary algebra, but, for your sake, we'll actually do one and you can do the other. Of course, we do the easier of the two and leave the more complicated one for you. So, here goes: $(a,b) \sim (a',b')$ means that $ab' = a'b$, and $(c,d) \sim (c',d')$ means that $cd' = c'd$. Multiplying the first equality by cd', and then substituting $cd' = c'd$ on the right hand side of the resulting equation, we get the desired equality $acb'd' = a'c'bd$.

Exercise 0.2.18 You do addition. It's messy.

When we are defining some operation which combines equivalence classes, we often do this by choosing representatives from each class and then showing that it doesn't make any difference which representatives are chosen. We have a formal name for this. We say that the operation under consideration is *well-defined* if the result is independent of the representatives chosen in the equivalence classes. Throughout this book, we will encounter equivalence relations on a regular basis. You will be fortunate enough to have the opportunity to prove that these are actually equivalence relations.

What properties are satisfied by addition and multiplication as defined above? For example, what about the associativity of addition? We must prove that $(\{(a,b)\}+\{(c,d)\})+\{(e,f)\} = \{(a,b)\}+(\{(c,d)\}+\{(e,f)\})$. Well,

$$\begin{aligned}(\{(a,b)\} + \{(c,d)\}) + \{(e,f)\} &= \{(ad+bc, bd)\} + \{(e,f)\}\\ &= \{((ad+bc)f + (bd)e, (bd)f)\}.\end{aligned}$$

Now we use associativity and distributivity in $\mathbb{Z}$ to rearrange things in an appropriate fashion. This gives $\{(((ad)f+(bc)f)+(bd)e, (bd)f)\}$, and using the acrobatics of parentheses, we get $\{(a(df)+b(cf+de), b(df))\} = \{(a,b)\} + (\{(c,d)\} + \{(e,f)\})$. This is all rather simple, that is, to prove various properties of addition and multiplication in $\mathbb{Q}$, we reduce them to known properties from $\mathbb{Z}$.

Exercise 0.2.19

i. Prove the associative law for multiplication in $\mathbb{Q}$.

ii. Prove the commutative laws for addition and multiplication in $\mathbb{Q}$.

iii. Show that $\{(0,1)\}$ is an additive identity in $\mathbb{Q}$.

iv. Show that $\{(1,1)\}$ is a multiplicative identity in $\mathbb{Q}$.

v. Show that $\{(-a,b)\}$ is an additive inverse for $\{(a,b)\}$.

vi. Prove the distributive law for $\mathbb{Q}$.

Notice here that if $\{(a,b)\} \neq \{(0,1)\}$, that is, $a \neq 0$, then $\{(a,b)\} \cdot \{(b,a)\} = \{(1,1)\}$. Thus, in $\mathbb{Q}$, we have multiplicative inverses for nonzero elements.

Let's tidy this up a bit. First of all, we have no intention of going around writing rational numbers as equivalence classes of ordered pairs of integers. So let's decide once and for all to write the rational number $\{(a,b)\}$ as $\frac{a}{b}$. Most of the time this fraction will be reduced to lowest terms, but, if it is not reduced to lowest terms, it will certainly be in the same equivalence class as a fraction which is reduced to lowest terms. Note that it is always possible to choose a fraction from the equivalence class which is in lowest terms because of the well-ordering principle in the integers, applied to the numerators. With this, addition and multiplication of rational numbers have their usual definition:

$$\begin{aligned}\frac{a}{b}+\frac{c}{d} &= \frac{ad+bc}{bd},\\ \frac{a}{b}\cdot\frac{c}{d} &= \frac{ac}{bd}.\end{aligned}$$

Now consider the axioms for the integers (A1)–(A5), (M1)–(M4), and (D). All of these hold for the rational numbers, and there is an additional multiplicative property, multiplicative inverses.

(M5) If $a \neq 0$, then there is an element a^{-1} such that $aa^{-1} = a^{-1}a = 1$.

Remark 0.2.20 Note that the natural numbers $\mathbb{N}$ may be regarded as a subset of $\mathbb{Z}$, and in turn the integers $\mathbb{Z}$ may be regarded as a subset of $\mathbb{Q}$ by identifying the integer n with the equivalence class $\frac{n}{1}$.

The operations of addition and multiplication are sometimes called *binary operations* or *internal laws of composition*

Definition 0.2.21 Let R be a non-empty set. An *internal law of composition (ILC)* on R is a map $\circ : R \times R \to R$. If $a, b \in R$ then we usually write $\circ((a,b)) = a \circ b$

Of course the more properties that are satisfied by internal laws of composition, the better life gets.

Definition 0.2.22 A set with two internal laws of composition, $+$ and $\cdot$, that satisfy (A1)–(A5), (M1)–(M4), and (D) is called a *commutative ring with* 1. If, in addition, cancellation (C) holds for multiplication, the commutative ring with 1 is called an *integral domain.* If (M5) also holds, the structure is called a *field.*

Note that the word "commutative" in this definition refers not to the commutativity of addition but to the commutativity of multiplication. Thus, in our latest terminology, $\mathbb{Z}$ is a integral domain and $\mathbb{Q}$ is a field. What about cancellation for multiplication? This followed from order in $\mathbb{Z}$, but for $\mathbb{Q}$ (or any field for that matter) cancellation for multiplication holds automatically.

Exercise 0.2.23 Prove that the cancellation law (C) holds in any field.

Exercise 0.2.24 Let X be a nonempty set and $R = \wp(X)$, the power set of X. Show that R with symmetric difference as addition and intersection as multiplication is a commutative ring with 1. When is R a field? (See Appendix A for the relevant set-theoretic definitions.)

There is another definition which will prove useful in our discussions about these various algebraic structures.

Definition 0.2.25 Suppose that R is a commutative ring with 1. A subset R_0 of R is a *subring* if R_0 is a ring itself with the same operations of addition and multiplication as in R. We don't necessarily require that R_0 have a multiplicative identity and in this case we call R_0 simply a *commutative ring.*

The same idea can be used to define *subintegral domain.* Finally, if F is a field and F_0 is a subset of F, we say that F_0 is a *subfield* if it is a field with the same operations of addition and multiplication as in F.

Exercise 0.2.26

i. Let R be a ring and R_0 a non empty subset of R, show that R_0 is a subring if and only if for any $a, b \in R_0$ we have $a - b$ and ab in R_0.

ii. If F is a field and F_0 is non-empty subset of F, are the properties in (*i*) enough to ensure that F_0 is a subfield?

Just as we consider two sets to be the same when there exists a bijection between them, there is a notion of sameness for other mathematical structures defined by the existence of certain maps, called isomorphisms.

Definition 0.2.27 Let R and R' be two commuative rings with 1. We say that R and R' are *isomorphic* if there exists a bijection $\phi : R \to R'$ such that

a. $\phi(x+y) = \phi(x) + \phi(y)$ for all $x, y \in R$;

b. $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in R$.

We call the map ϕ an *isomorphism*.

What about order in $\mathbb{Q}$? It is simple to extend the order from $\mathbb{Z}$ to $\mathbb{Q}$. We do this using the notion of a set of positive elements. We say that $\frac{a}{b} \in \mathbb{Q}$ is positive if $ab > 0$ in $\mathbb{Z}$.

Exercise 0.2.28 Show that the above notion of positivity in $\mathbb{Q}$ satisfies the properties in Exercise A.5.11, or equivalently, the properties of order given in (O1)–(O4).

Definition 0.2.29 An integral domain or field in which there is an order relation satisfying (O1)–(O4) is called an *ordered integral domain* or *ordered field*, respectively. See Project 2 for more about this.

Definition 0.2.30 Suppose that R and R' are ordered integral domains. We say that R and R' are *order isomorphic* if there exists an isomorphism $\phi : R \to R'$ such that if $x, y \in R$ and $x < y$ then $\phi(x) < \phi(y)$ in R'. We call the map ϕ an *order isomorphism*.

So what is this all about? We have rules for the integers, and the same rules, along with (M5), are satisfied by the rational numbers. Actually, there are lots of structures other than the integers and the rational numbers which have operations of addition, multiplication, and, sometimes, an order relation. For example, the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, the algebraic numbers $\mathbb{A}$, the collection of $n \times n$ matrices $M_n(\mathbb{R})$, all satisfy some or all of these properties.

We want to give two more examples before we leave this section. First, let n be a positive integer greater than or equal to 2 and consider the equivalence relation given in Example A.6.6. What are the equivalence classes? For example, take $n = 5$. Then we have 5 classes. They are

$$\begin{aligned}
C(0) = \overline{0} &= \{0, 5, -5, 10, -10, \ldots\} \\
C(1) = \overline{1} &= \{1, 6, -4, 11, -9, \ldots\} \\
C(2) = \overline{2} &= \{2, 7, -3, 12, -8, \ldots\} \\
C(3) = \overline{3} &= \{3, 8, -2, 13, -7, \ldots\} \\
C(4) = \overline{4} &= \{4, 9, -1, 14, -6, \ldots\}.
\end{aligned}$$

Note that, in this example, we have simplified the notation of equivalence class by writing the equivalence class $C(a)$ by $\overline{a}$. Observe that $\overline{5} = \overline{0}$, $\overline{6} = \overline{1}$, etc. In general, for an arbitrary n, we will have n classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$. These are called *the equivalence classes modulo* n, or, for short, *mod* n. Moreover, for any integer a, we denote the equivalence class in which a lies by $\overline{a}$. Of course, it is always true that $\overline{a}$ is equal to one of the classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$. Let's define addition and multiplication mod n.

Definition 0.2.31 Denote the set of equivalence classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$ by $\mathbb{Z}_n$. For $\overline{a}, \overline{b} \in \mathbb{Z}_n$, define $\overline{a} + \overline{b} = \overline{a+b}$ and $\overline{a}\overline{b} = \overline{ab}$.

Exercise 0.2.32

i. Show that addition and multiplication in $\mathbb{Z}_n$ are well-defined.

ii. Show that, with these operations, $\mathbb{Z}_n$ is a commutative ring with 1.

iii. Show that $\mathbb{Z}_n$ cannot satisfy the order axioms no matter how $>$ is defined.

iv. Show that $\mathbb{Z}_2$ is a field but $\mathbb{Z}_4$ is not.

v. For p prime show that $\mathbb{Z}_p$ is a field.

The second example is the real numbers denoted by $\mathbb{R}$. A construction and complete discussion of the real numbers is given in the next chapter. We will see that the real numbers are an ordered field which contains $\mathbb{Q}$ and has one additional property called the least upper bound property.

0.3 Countability

Our discussion of number systems would not be complete without mentioning infinite sets. Indeed, most of the sets we deal with in analysis are infinite. Moreover, any discussion of continuity and change must involve infinite sets. Thus we are motivated to begin a formal discussion of what it means for a set to be infinite. (See Appendix A for a consideration of the more elementary notions of set theory.)

Definition 0.3.1 A set A is *finite* if A is empty or there exists $n \in \mathbb{N}$ such that there is a bijection $f : A \to \{1, 2, \ldots, n\}$, where $\{1, 2, \ldots, n\}$ is the set of all natural numbers less than or equal to n. In this case, we say A has n elements.

Exercise 0.3.2 If A is a finite set and B is a subset of A, show that B is a finite set. In addition show that if B is a proper subset then the number of elements in B is less then the number of elements in A.

There is a natural and useful property of finite sets, which in fact will turn out to be a characterization of them:

Theorem 0.3.3 If A is a finite set and B is a proper subset of A, then there is no bijection between B and A.

Proof. Suppose A has n elements and B has m elements with $m < n$. Then the Pigeonhole Principle (see Appendix A.3) tells us that, for any function from A to B, there is an element of B which is the image of two different elements of A.

Exercise 0.3.4 Show that the following are finite sets:

i. The English alphabet.

ii. The set of all possible twelve letter words made up of letters from the English alphabet.

iii. The set of all subsets of a given finite set.

This approach to things makes the definition of infinite sets quite simple:

Definition 0.3.5 An *infinite* set is a set that is not finite.

One of the most important characteristics of a set is its cardinality, which formalizes the notion of the size of the set. A thorough treatment of cardinality would take us too far afield, but we can say what it means for two sets to have the same cardinality.

Definition 0.3.6 The *cardinal number* of a finite set A is the number of elements in A, that is, the cardinal number of A is the natural number n if there is a bijection between A and $\{k \in \mathbb{N} \,|\, 1 \leq k \leq n\}$.

Definition 0.3.7 A set A has *cardinality* $\aleph_0$ (pronounced “aleph null” or “aleph naught”) if it can be put in one-to-one correspondence with $\mathbb{N}$, that is, there is a bijection between the set and $\mathbb{N}$.

In general, two sets have the same cardinality if they can be put in one-to-one correspondence with each other.

Example 0.3.8 The set $\mathbb{N}$ has cardinality $\aleph_0$ (this should not come as a surprise).

We will see later in this section that there are infinite sets with cardinality other than $\aleph_0$.

Example 0.3.9 The set $\mathbb{N} \cup \{0\}$ has cardinality $\aleph_0$ because the function $f : \mathbb{N} \cup \{0\} \to \mathbb{N}$ given by $f(n) = n + 1$ is a bijection.

Example 0.3.10 The set $\mathbb{Z}$ has cardinality $\aleph_0$ because the function $f : \mathbb{Z} \to \mathbb{N}$ given by

$$f(z) = \begin{cases} 2z + 2 & \text{if } z \geq 0 \\ -2z - 1 & \text{if } z < 0 \end{cases}$$

is a bijection.

There is a very useful theorem which asserts the existence of a one-to-one correspondence between two sets. This relieves us of the burden of constructing a bijection between two sets to show that they have the same cardinality.

Theorem 0.3.11 (Schröder-Bernstein) If A and B are sets, and there exist injections $f : A \to B$ and $g : B \to A$, then there exists a bijection between A and B.

Proof. First, we divide A into three disjoint subsets. For each $x \in A$, consider the list of elements

$$S_x = \{x, g^{-1}(x), f^{-1} \circ g^{-1}(x), g^{-1} \circ f^{-1} \circ g^{-1}(x), \dots\}.$$

The elements of this sequence are called *predecessors* of x. Notice that in S_x, we start with $x \in A$. Then $g^{-1}(x) \in B$ if $g^{-1}(x)$ exists (x may not be in the image of g). For each $x \in A$, exactly one of the three following possibilities occurs.

1. The list S_x is infinite.

2. The last term in the list is an element of A. That is, the last term is of the form $y = f^{-1} \circ g^{-1} \circ \cdots \circ g^{-1}(x)$, and $g^{-1}(y)$ does not exist (i.e. y is not in the image of g). In this case, we say that S_x *stops in* A.

3. The last term in the list is an element of B. That is, the last term is of the form $z = g^{-1} \circ f^{-1} \circ \cdots \circ g^{-1}(x)$ and $f^{-1}(z)$ does not exist (i.e. z is not in the image of f). In this case, we say that S_x *stops in* B.

Let the corresponding subsets of A be denoted by A_1, A_2, A_3. Similarly, define the corresponding subsets of B. That is

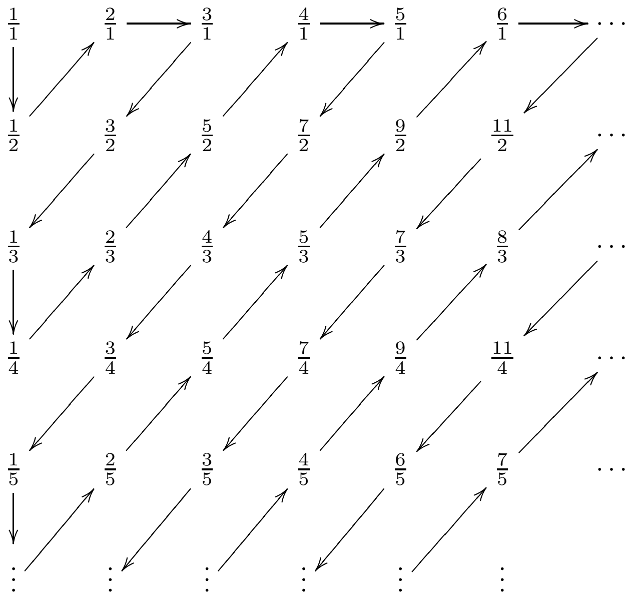
$$\begin{aligned} B_1 &= \{y \in B \mid y \text{ has infinitely many predecessors }\}, \\ B_2 &= \{y \in B \mid \text{the predecessors of } y \text{ stop in } A\}, \text{ and} \\ B_3 &= \{y \in B \mid \text{the predecessors of } y \text{ stop in } B\}. \end{aligned}$$

Now observe that $f : A_1 \to B_1$, $g : B_1 \to A_1$ are both bijections. Also, $g : B_2 \to A_2$ and $f : A_3 \to B_3$ are bijections. ∎

Exercise 0.3.12 Suppose A, B, and C are subsets of a set X such that $A \subseteq B \subseteq C$. Show that if A and C have the same cardinality, then A and B have the same cardinality.

Example 0.3.13 $\mathbb{Q}_+$ has cardinality $\aleph_0$ (recall that $\mathbb{Q}_+$ denotes the positive rational numbers). Here are three proofs:

1. This is a very common and very sloppy proof. However the underlying idea will stand us in good stead.



   To find a bijection between $\mathbb{N}$ and $\mathbb{Q}_+$, we write all the positive fractions in a grid, with all fractions with denominator 1 in the first row, all fractions with denominator 2 in the second row, all fractions with denominator 3 in the third row, etc. Now go through row by row and throw out all the fractions that aren't written in lowest terms. Then, starting at the upper left hand corner, trace a path through all the remaining numbers as above.

   We can count along the path we drew, assigning a natural number to each fraction. So $\frac{1}{1} \to 1$, $\frac{1}{2} \to 2$, $\frac{2}{1} \to 3$, $\frac{3}{1} \to 4$, $\frac{3}{2} \to 5$, etc. This is a bijection. Therefore, $\mathbb{Q}_+$ is countable. Although this is a very common proof, the bijection is not at all obvious. It is very difficult to see, for example, which rational number corresponds to $1,000,000$.

2. In this proof, we'll make use of the Schröder-Bernstein Theorem. It is easy to inject $\mathbb{N}$ into $\mathbb{Q}_+$; simply send n to n.

   For the injection from $\mathbb{Q}_+$ to $\mathbb{N}$, we consider $\mathbb{Q}_+$ to be the set of positive rational numbers written as fractions in base 10, and we let $\mathbb{N}$ be the natural numbers but written in base 11 with the numerals being 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, d. So, for example, the integer which is written 21 in base 10 is $1d$ in base 11, and $1222_{(10)} = d11_{(11)}$. (Incidentally, we are writing the bases in base 10.) Now define a function $f : \mathbb{Q}_+ \to \mathbb{N}$ by writing $\frac{a}{b} = \frac{a_n \ldots a_2 a_1}{b_m \ldots b_2 b_1}$, where a_i and b_i are the i^{th} digits of the numerator and denominator (and, of course, integers between 0 and 9). Then, set $f(\frac{a}{b}) = a_n \ldots a_2 a_1 d b_m \ldots b_2 b_1$. The fraction $\frac{a}{b}$ will always be written in lowest terms. For instance, if we take the fraction $2/3$ in base 10, then $f(2/3) = 2d3$ in base 11 which is the same as the integer 355 written in base 10. Each number which is the image of a fraction has one and only one d in it, so it is easy to see which fraction is represented by a given integer.

   According to Schröder-Bernstein, two injections make a bijection, so $\mathbb{Q}_+$ is countable.

3. Write each positive fraction in lowest terms and factor the numerator and denominator into primes, so that $\frac{p}{q} = \frac{p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}}{q_1^{\beta_1} q_2^{\beta_2} \cdots q_m^{\beta_m}}$, with $p_i \neq q_j$. If by chance p or q is 1, and can't be factored, write it as 1^1.

Then let $f : \mathbb{Q}_+ \to \mathbb{N}$ be defined by

$$f\left(\frac{p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}}{q_1^{\beta_1} q_2^{\beta_2} \cdots q_m^{\beta_m}}\right) = p_1^{2\alpha_1} p_2^{2\alpha_2} \cdots p_n^{2\alpha_n} q_1^{2\beta_1 - 1} q_2^{2\beta_2 - 1} \cdots q_m^{2\beta_m - 1}.$$

In particular, note that if $a \in \mathbb{Q}_+$ is an integer, then $f(a) = a^2$.

Exercise 0.3.14 For this exercise, consider the function f in the third proof above.

1. Verify that f is a bijection.
2. Suppose that $N = 10^k$ for some integer k. Find $\frac{p}{q} \in \mathbb{Q}$ such that $f\left(\frac{p}{q}\right) = N$.

Exercise 0.3.15 Use any one of the above three proofs to show that $\mathbb{Q}$ has cardinality $\aleph_0$.

Exercise 0.3.16 Show that the natural numbers are an infinite set.

Exercise 0.3.17 Show that any set that has the same cardinal number as $\mathbb{N}$ is an infinite set.

Note: A set is called *countable*, or sometimes *denumerable*, if it has cardinality $\aleph_0$ (that is, if it is in one-to-one correspondence with the natural numbers). The term countable is used in several ways. Many people use it to refer to infinite sets which are in one-to-one correspondence with $\mathbb{N}$, while others include finite sets when they say countable. This is not something to get disturbed about. Usually, when we refer to a countable set, we mean countably infinite (cardinality $\aleph_0$) . When we refer to a finite set, we will generally say "A is a finite set."

Exercise 0.3.18 Show that a subset of a countable set is countable or finite.

Exercise 0.3.19 Show that the set of all polynomial functions with integer coefficients is a countable set.

Theorem 0.3.20 If A is an infinite set, then A has a countable subset.

Proof. Take any infinite set A and choose an element a_1 in A. Let $A_1 = A \setminus \{a_1\}$. By the definition of infinite set, A_1 is infinite. So we choose a_2 in A_1 and define $A_2 = A \setminus \{a_1, a_2\}$. Since A is not finite, we can continue to choose elements. Thus, if we have chosen $a_1, \ldots, a_n$, we consider $A_n = A \setminus \{a_1, \ldots, a_n\}$. Since A is infinite, we can choose an element a_{n+1} in A_n. Continuing inductively, we obtain our desired countable subset. Note that this countable set may be all of A.

Remark 0.3.21 There is some discussion among mathematicians as to whether the preceding proof involves the Axiom of Choice. The Axiom of Choice in its fullest form will be discussed below. However, one can make the argument that it requires some sort of choice mechanism to pick an element from a non-empty set. The technique that we use in the proof of Theorem A.8.21 is sometimes referred to as "the countable Axiom of Choice."

We could pursue an alternate definition of an infinite set. In fact, we could define infinite sets first and then say that a finite set is a set that is not infinite. We use Theorem A.8.21 as motivation for the following.
Redefinition 0.3.22 A set is *infinite* if there is a bijection between the set and one of its proper subsets.

Redefinition 0.3.23 A *finite* set is a set that is not infinite.

To show the equivalence of the two definitions, recall that in Theorem A.8.3 we showed there is no bijection between a finite set and any of its proper subsets. This means that if a set is infinite by our new definition, it is not finite (hence, infinite) by the old definition too. Next, let's show that any set that is infinite by the old definition is bijective with one of its proper subsets.

Proof. Say A is an infinite set and $B \subseteq A$ is countable. Then we can write $B = \{b_1, b_2, \ldots, b_n, \ldots\}$. Now define $f : A \to A \setminus \{b_1\}$ as follows: for $a \in A \setminus B$, $f(a) = a$, and for $b_i \in B$, $f(b_i) = b_{i+1}$. Thus f is a bijection between A and $A \setminus \{b_1\}$. Therefore, our definitions are equivalent.

We now turn to operations involving infinite sets.

Facts 0.3.24

1. If A_1 and A_2 are countable sets, then $A_1 \cup A_2$ is a countable set.
2. If A_1, A_2, …, A_n are countable sets, then $\cup_{j=1}^{n} A_j$ is a countable set.
3. Let $\{A_j\}_{j\in\mathbb{N}}$ be a countable collection of countable sets. Then $\cup_{j\in\mathbb{N}} A_j$ is a countable set.

Proof. We prove 3 only. You can prove the other two (or deduce them from 3).

Write $A_j = \{a_{j,1}, a_{j,2}, \ldots, a_{j,n}, \ldots\}$. We use the diagonal process, as in Example A.8.13. Simply write

$$\begin{gathered}
A_1 : a_{1,1}, a_{1,2}, \ldots, a_{1,n}, \ldots \\
A_2 : a_{2,1}, a_{2,2}, \ldots, a_{2,n}, \ldots \\
\vdots \\
A_m : a_{m,1}, a_{m,2}, \ldots, a_{m,n}, \ldots \\
\vdots
\end{gathered}$$

Now count diagonally, ignoring repetitions.

Now let's take a look at Cartesian products. It is clear from the ideas presented above that if A_1 and A_2 are countable, then $A_1 \times A_2$ is countable.

Exercise 0.3.25

i. Show that if A_1, A_2, …, A_n are countable, then $A_1 \times A_2 \times \cdots \times A_n$ is countable.

ii. What can you say about the countable Cartesian product of countable sets?

Next we look at the power set $\wp(A)$ for any set A.

Theorem 0.3.26 If A is any set (including the empty set), there is no bijection between A and $\wp(A)$.

Proof. This is clear if A is the empty set. Suppose that there is a bijection between A and $\wp(A)$. If $a \in A$, let P_a be the subset of A associated with it. Now consider the set $B = \{a | a \notin P_a\}$. The set B must be associated to some element of A, which we creatively call b, so that $B = P_b$. Is b in B? For b to be in B, we must have that $b \notin P_b$. But $B = P_b$, so therefore b is not in B. But then $b \in P_b$, which means that b is in B. This is a contradiction. Therefore, there is no bijection between A and $\wp(A)$.

Definition 0.3.27 If A is a countable set, then the cardinality of $\wp(A)$ is denoted by $\mathbf{c}$.

Exercise 0.3.28 Show that the definition of the cardinal number $\mathbf{c}$ does not depend on the choice of the countable set A. That is if A and B are countable sets then there is a bijection between $\wp(A)$ and $\wp(B)$.

Remark 0.3.29 At this point, we observe that if A is a countable set, $A = \{a_1, a_2, \ldots, a_n, \ldots\}$, then $\wp(A)$ is in one-to-one correspondence with the set of all functions from A to the set $\{0, 1\}$. This correspondence is defined as follows: If B is a subset of A, then we define the map $f_B : A \to \{0, 1\}$ by $f_B(a_j) = 1$ if a_j is in B, 0 if a_j is not in B. Observe that f_B can be viewed as a binary expansion of a real number between 0 and 1.

Exercise 0.3.30 Suppose that A is a nonempty set. Show that $\wp(A)$ is in one-to-one correspondence with the set of all functions from A to $\{0, 1\}$.

Remark 0.3.31 Based on the reasoning in the previous exercise, if A is a finite set with n elements, the cardinality of $\wp(A)$ is 2^n. We extend this reasoning to countable sets to write $\mathbf{c} = 2^{\aleph_0}$.

One of the most important sets of numbers that we deal with in this book is the collection of real numbers $\mathbb{R}$. In the next chapter, we will go through the formal construction of the real numbers from the rational numbers. For the present discussion, we can just consider the set of real numbers to be the set of all infinite decimals with the convention that no decimal expansion can end in repeating 9s. There are two things to show about the reals. The first is the proof due to Cantor that the reals are uncountable, and the second is that the cardinality of the real numbers is in fact $\mathbf{c}$.

Theorem 0.3.32 The set of all real numbers between 0 and 1 is not countable.

Proof. We first note that the decimal expansion is unique with the exception of those that end in all nines. In this case, we always round up the digit which occurs before the sequence of nines. To prove that this set is not countable, we assume that it is, and list the real numbers between 0 and 1 vertically.

$$\begin{gathered} a_1 = 0.a_{1,1}a_{1,2}\ldots a_{1,n}\ldots \\ a_2 = 0.a_{2,1}a_{2,2}\ldots a_{2,n}\ldots \\ \vdots \\ a_m = 0.a_{m,1}a_{m,2}\ldots a_{m,n}\ldots \\ \vdots \end{gathered}$$

We now proceed using a process similar to the one used in the proof of Theorem A.8.27 to produce a real number between 0 and 1 which is not on our list. We construct a number $b = 0.b_1b_2\ldots b_n\ldots$ by proceeding diagonally down the list as follows: if $a_{1,1} = 1$, take $b_1 = 2$. If $a_{1,1} \neq 1$, take $b_1 = 1$. Next, if $a_{2,2} = 1$, take $b_2 = 2$. If $a_{2,2} \neq 1$, take $b_2 = 1$. Continuing this process, we see that the decimal $b = 0.b_1b_2\ldots b_n\ldots$ cannot be on our list, since it differs from each number we list in at least one digit. Consequently, the real numbers between 0 and 1 are not countable. ☠

Theorem 0.3.33 The cardinality of the real numbers between 0 and 1 is $\mathbf{c} = 2^{\aleph_0}$.

Proof. To write down an exact bijection between $\wp(\mathbb{N})$ and the real numbers between 0 and 1 requires some care. The standard way to do this is to write all real numbers between 0 and 1 in their binary expansion in such a way that no expansion terminates in all ones. In considering the corresponding subsets of $\mathbb{N}$, we first remove two specific subsets of $\wp(\mathbb{N})$. We remove the two collections $A_f = \{C \in \wp(\mathbb{N}) \mid C \text{ is finite}\}$ and $A_{cf} = \{D \in \wp(\mathbb{N}) \mid {}^cD \text{ is finite}\}$. The collection $\wp(\mathbb{N}) \setminus (A_f \cup A_{cf})$ is in one-to-one correspondence with all binary expansions which have an infinite number of ones but do not terminate in all ones. We get the required bijection by Remark A.8.30.

We can place A_f into one-to-one correspondence with the set of all finite binary expansions with 0 in the first place, and A_{cf} can be put into one-to-one correspondence with the set of all finite binary expansions with 1 in the first place. ☠

Exercise 0.3.34 Write down these last two bijections explicitly.

Exercise 0.3.35

i. Prove that the countable union of sets of cardinality $\mathbf{c}$ again has cardinality $\mathbf{c}$.

ii. Prove that the set of all real numbers has cardinality $\mathbf{c}$.

iii. Prove that the set of irrational numbers in $\mathbb{R}$ has cardinality $\mathbf{c}$.

How big do cardinal numbers get? For instance, the power set of $\mathbb{R}$ is "bigger than" $\mathbf{c}$. In fact, the power set of $\mathbb{R}$ can be identified with the set of all maps from $\mathbb{R}$ into $\{0, 1\}$ just as we did above for the power set of $\mathbb{N}$. Thus, the cardinality of $\wp(\mathbb{R})$ is $2^{\mathbf{c}}$.

The following theorem is interesting, and we include it because it illustrates the kinds of non-intuitive results that one encounters when dealing with infinite sets.

Theorem 0.3.36 There is a bijection between the unit interval and the unit square.

Proof. Let

$$I = [0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

and $I^2 = [0, 1] \times [0, 1]$. This seems like a great time to use Schröder-Bernstein. The function $f : I \to I^2$ defined by $f(x) = (x, 0)$ is an injection. Define the function $g : I^2 \to I$ by the rule $g((a_0.a_1a_2 \ldots a_n \ldots, b_0.b_1b_2 \ldots b_n \ldots)) = (0.a_0b_0a_1b_1a_2b_2 \ldots a_nb_n \ldots)$, where $a_0a_1a_2 \ldots a_n \ldots$ and $b_0.b_1b_2 \ldots b_n \ldots$ are decimal expansions of the coordinates of any point in I^2 (of course, the decimal expansion is prohibited from ending in all 9s). The function $g : I^2 \to I$ is an injection. Therefore, there is a bijection between I and I^2.

0.4 Axiom of Choice

Definition 0.4.1 A *partially ordered set* is a set X with a relation $\leq$ which is reflexive, transitive, and anti-symmetric (that means that if $a \leq b$ and $b \leq a$, then $a = b$). A *totally ordered set* is a partially ordered set with the additional property that, for any two elements $a, b \in X$, either $a \leq b$ or $b \leq a$. A *well-ordered set* is a totally ordered set in which any non-empty subset has a least element.

Example 0.4.2

1. $(\mathbb{N}, \leq)$ is a totally ordered set, as are $(\mathbb{Z}, \leq)$, $(\mathbb{Q}, \leq)$ and $(\mathbb{R}, \leq)$.
2. Let X be a set, and let $\wp(X)$ be the collection of all subsets of X. Then $(\wp(X), \subseteq)$ is a partially ordered set.

Definition 0.4.3 Let Y be a subset of a partially ordered set X. An *upper bound* for Y is an element $a \in X$ such that $y \leq a$ for all $y \in Y$. A *least upper bound* for Y is an element $b \in X$ such that b is an upper bound for Y and if a is an upper bound for Y, then $b \leq a$. The least upper bound is sometimes abbreviated lub, and is also denoted as sup (supremum). You can figure out what a *lower bound* and *greatest lower bound (*glb*)* are. The greatest lower bound is also denoted by inf (infimum).

Observe that a subset of a partially ordered set may not have an upper bound or a lower bound.

Exercise 0.4.4 If a subset Y of a partially ordered set X has an upper bound, determine whether or not Y must have a least upper bound. If Y has a least upper bound, determine whether or not this least upper bound is unique.

Definition 0.4.5 In a partially ordered set, an element b is *maximal* if $a \geq b$ implies $a = b$.

We turn now to one of the major topics of this chapter, the axiom of choice, and various logically equivalent statements. For many years, there has been considerable discussion among mathematicians about the use of the axiom of choice and the seemingly contradictory results that come along with it. We find it indispensable in obtaining a number of results in mathematics.

The Axiom of Choice 0.4.6 Given a collection $\mathcal{C}$ of sets which does not include the empty set, there exists a function $\phi : \mathcal{C} \to \cup_{C \in \mathcal{C}} C$ with the property that $\forall A \in \mathcal{C}$, $\phi(A) \in A$.

Another way of looking at this is as follows. Suppose $\{A_i\}_{i \in I}$ is a collection of non-empty sets indexed by an index set I. A *choice function* is then defined as a map $\phi : I \to \bigcup_{i \in I} A_i$ such that $\phi(i) \in A_i$. The axiom of choice can then be rephrased.

The Axiom of Choice 0.4.7 For every collection of nonempty sets there exists a choice function.

The axiom of choice is equivalent to a number of other very useful statements which are not at all obvious. Here they are, in no particular order.

Let X be a partially ordered set. The collection $\wp(X)$ can be partially ordered by inclusion, see A.9.2. This partial ordering on $\wp(X)$ is used in some of the statements below.

Hausdorff Maximality Principle 0.4.8 Every partially ordered set X contains a totally ordered subset that is maximal with respect to the ordering on $\wp(X)$.

Zorn's Lemma 0.4.9 If a non-empty partially ordered set has the property that every non-empty totally ordered subset has an upper bound, then the partially ordered set has a maximal element.

Well-Ordering Principle 0.4.10 Every set can be well-ordered.

The following lemma is slightly complicated, but it will allow us to prove the equivalence of the above statements with little trouble.

Lemma 0.4.11 Suppose that $(X, \leq)$ is a non-empty partially ordered set such that every non-empty totally ordered subset has a least upper bound. If $f : X \to X$ is such that $f(x) \geq x$ for all $x \in X$, then there is some $w \in X$ such that $f(w) = w$.

Proof. First we reduce to the case when X contains a least element, call it b. In fact, if X is nonempty choose any $b \in X$ and replace X by $X' = \{x \in X \mid x \geq b\}$. It is clear that X' is stable under f (that is $f(X') \subseteq X'$) and has the same properties as X. We call a subset Y of X "admissible" if

1. $b \in Y$
2. $f(Y) \subseteq Y$
3. Every lub of a totally ordered subset of Y belongs to Y.

X is certainly admissible, and the intersection of any family of admissible sets is admissible. Let W be the intersection of all admissible sets. The set $\{x | b \leq x\}$ is admissible, so if $y \in W$, then $b \leq y$.

We will now construct a totally ordered subset of W with the property that its least upper bound is a fixed point of f. Consider the set $P = \{x \in W | \text{ if } y \in W \text{ and } y < x \text{ then } f(y) \leq x\}$. Note that P is non-empty since $b \in P$. First we show that any element of P can be compared to any element of W and hence P is totally ordered.

Now fix an $x \in P$ and define $A_x = \{z \in W | z \leq x \text{ or } z \geq f(x)\}$. We would like to show that A_x is admissible.

1. Obviously, $b \in A_x$ since $b \leq x$.
2. Suppose $z \in A_x$. There are three possibilities. If $z < x$, $f(z) \leq x$ by the conditions of P, so $f(z) \in A_x$. If $z = x$, $f(z) = f(x) \geq f(x)$ so $f(z) \in A_x$. If $z \geq f(x)$, then $f(z) \geq z \geq f(x)$ so $f(z) \in A_x$.
3. Finally, let Y be a totally ordered non-empty subset of A_x, and let y_0 be the lub of Y in X. Then $y_0 \in W$, since W is admissible. If $z \leq x$ for all $z \in Y$ then $y_0 \leq x$ and hence $y_0 \in A_x$. Otherwise $z \geq f(x)$ for some $z \in Y$, which implies $y_0 \geq f(x)$, so $y_0 \in A_x$.

Thus, A_x is admissible.

Since A_x is an admissible subset of W, $A_x = W$. Put another way, if $x \in P$ and $z \in W$, then either $z \leq x$ or $z \geq f(x) \geq x$, and thus P is totally ordered. Therefore P has a least upper bound, call it x_0. Again $x_0 \in W$ and $f(x_0) \in W$ because W is admissible. We will now show $f(x_0) = x_0$. First we claim $x_0 \in P$. Indeed, if $y \in W$ and $y < x_0$, then there exists $x \in P$ with $y < x \leq x_0$, whence $f(y) \leq x \leq x_0$. Let $y \in W$ and suppose $y < f(x_0)$. As we saw above $A_{x_0} = W$, so we have $y \leq x_0$. If $y = x_0$, then $f(y) = f(x_0) \leq f(x_0)$. If $y < x_0$, then $f(y) \leq x_0 \leq f(x_0)$. In either case, we find $f(x_0) \in P$. Hence $f(x_0) \leq x_0 \leq f(x_0)$.

Whew!

Theorem 0.4.12 (1) The Axiom of Choice, (2) Hausdorff Maximality Principle, (3) Zorn's Lemma, and (4) Well-Ordering Principle are all equivalent.

Proof. We will show that (1) implies (2), which implies (3), which implies (4), which implies (1), and then we will be done.

(1) ⇒ (2)

Take a non-empty partially ordered set $(E, \leq)$. Make $\mathcal{E}$, the family of totally ordered subsets of E, into a partially ordered set under inclusion. We wish to show that $\mathcal{E}$ has a maximal element (i.e., an element which is not smaller than any other element). So we will assume the opposite and reach a contradiction by applying Lemma A.9.11. We must first check to see if the lemma is applicable: Suppose $\mathcal{F}$ is a totally ordered subset of $\mathcal{E}$. Then it has a least upper bound, namely $\cup_{F \in \mathcal{F}} F$. Now, for a given $e \in \mathcal{E}$, let $S_e = \{x \in \mathcal{E} | e \subseteq x, e \neq x\}$. Then S_e can never be the empty set, because that would mean that e is maximal. So we apply the axiom of choice by defining a function $f : \{S_e | e \in \mathcal{E}\} \to \mathcal{E}$ with the property that $f(S_e) \in S_e$. Now define $g : \mathcal{E} \to \mathcal{E}$ by $g(e) = f(S_e)$. This gives us that $e \subsetneq g(e)$ for all $e \in \mathcal{E}$, contradicting the lemma.

(2) ⇒ (3)

Again, consider a partially ordered set $(E, \leq)$. Now let x be the upper bound for E_0, a maximal totally ordered subset of E. Suppose that there is some $y \in E$ such that $y > x$. Then $E_0 \cup \{y\}$ is a totally ordered set containing E_0, contradicting our assumption of maximality.

Exercise 0.4.13 Now you finish the proof. Show that Zorn's Lemma implies the Well Ordering Principle, and that the Well Ordering Principle implies the Axiom of Choice.

0.5 Independent Projects

0.5.1 Basic Number Theory The following statements present a number of facts about elementary number theory. Your goal in this project is to prove them.

1. The division algorithm: if $a, b \in \mathbb{Z}$ and $b \neq 0$, then there is a unique pair $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < |b|$.

2. If M is a subset of $\mathbb{Z}$ which is closed under subtraction and contains a nonzero element, then $M = \{np \mid n \in \mathbb{Z}\}$, where p is the least positive element of M.

   Definition 0.5.1 Let $a, b \in \mathbb{Z}$. The *greatest common divisor* of a and b is the largest positive integer d such that $d \mid a$ and $d \mid b$. We often denote the greatest common divisor of a and b by (a, b). We say that a and b are *relatively prime* if $(a, b) = 1$.

3. If $a, b \in \mathbb{Z}$ and $d = (a, b)$, then there exist $s, t \in \mathbb{Z}$ such that $d = sa + tb$.

4. Euclid's lemma: If p is prime and $p|ab$, then $p|a$ or $p|b$.

5. If $(a, c) = 1$, and $c|ab$, then $c|b$.

6. If $(a, c) = 1$, $a|m$ and $c|m$, then $ac|m$.

7. If $a > 0$ then $(ab, ac) = a(b, c)$.

8. The integers $\mathbb{Z}$ have unique factorization, that is, if n is an integer greater than or equal to 2, then there exist unique distinct primes $p_1, p_2, \ldots, p_k$, with $p_1 < p_2 < \cdots < p_k$, and positive integer exponents $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$.

9. If n is a positive integer greater than or equal to 2 with unique factorization $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$, then the number of positive divisors of n is
$$(\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_k + 1).$$

   Definition 0.5.2 Let $a, b, c \in \mathbb{Z}$, with $c > 1$. We say that a is *congruent* to b *modulo* c if $c \mid (a - b)$, and we denote ethos statement by $a \equiv b \pmod{c}$.

10. If $a \equiv b \pmod{m}$, then $-a \equiv -b \pmod{m}$, $a + x \equiv b + x \pmod{m}$, and $ax \equiv bx \pmod{m}$ for every $x \in \mathbb{Z}$.

11. If $(c, m) = 1$ and $ca \equiv cb \pmod{m}$, then $a \equiv b \pmod{m}$.

12. If $(c, m) = 1$, then $cx \equiv b \pmod{m}$ has a unique solution x modulo m. In other words, the congruence is satisfied by some $x \in \mathbb{Z}$, and if it is also satisfied by some other $x' \in \mathbb{Z}$, then $x \equiv x' \pmod{m}$.

13. If p is prime and $c \not\equiv 0 \pmod{p}$, then $cx \equiv b \pmod{p}$ has a unique solution x modulo p.

14. If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then $a + c \equiv b + d \pmod{m}$ and $ac \equiv bd \pmod{m}$.

15. If $a, b, c \in \mathbb{Z}$ and $d = (a, b)$, then $ax + by = c$ has integer solutions x and y if and only if $d|c$.

    Definition 0.5.3 Let $a, b \in \mathbb{Z}$. The *least common multiple* of a and b is the smallest positive integer d such that $a \mid d$ and $b \mid d$. We often denote the least common multiple of a and b by $[a, b]$.

16. If $a, b \in \mathbb{Z}$, then $m[a, b] = [ma, mb]$ when $m > 0$.

17. If a and b are positive integers, then $(a, b)[a, b] = ab$.

18. If $ca \equiv cb \pmod{m}$ and $d = (c, m)$, then $a \equiv b \pmod{\frac{m}{d}}$.

19. If $m, a, b \in \mathbb{Z}$, $m > 1$, the congruence $ax \equiv b \pmod{m}$ is solvable if and only if $(a, m)|b$. There are exactly (a, m) solutions distinct modulo m.

20. If $a, b, s, t \in \mathbb{Z}$ are such that $sa + tb = 1$, then $(a, b) = 1$.

Definition 0.5.4 Let $m \in \mathbb{Z}$ be greater than 1, and let $P = \{i \in \mathbb{Z} \mid 1 \leq i \leq m-1, \text{and } (i, m) = 1\}$. A *reduced residue system* modulo m is a set Q of integers such that each of the integers in P is congruent modulo m to exactly one of the elements in Q.

21. The number of elements in a reduced residue system modulo m is independent of the representatives chosen.

22. If p is a prime and ϕ denotes Euler's ϕ function (where $\phi(a)$ is the number of integers between 0 and a, inclusive, that are relatively prime to a), then $\phi(p^n) = p^n - p^{n-1} = p^n(1 - \frac{1}{p})$.

23. The number of elements in a reduced residue system modulo m is $\phi(m)$.

24. If $a_1, \ldots, a_{\phi(m)}$ is a reduced residue system modulo m and $(\kappa, m) = 1$, then $\kappa a_1, \ldots, \kappa a_{\phi(m)}$ is a reduced residue system modulo m.

25. If m is a positive integer and $(\kappa, m) = 1$, then $\kappa^{\phi(m)} \equiv 1 \pmod{m}$.

26. If $d_1, \ldots, d_k$ are the positive divisors of n, then $\sum_{i=1}^{k} \phi(d_i) = n$.

0.5.2 Ordered Integral DomainsThis project is designed to show that any ordered integral domain contains a copy of the integers. Thus, in particular, any ordered field such as the rationals or real numbers contains a copy of the integers. Let R be an ordered integral domain.

Definition 0.5.5 An *inductive set* in R is a subset S of R such that

a. $1 \in S$, and

b. if $x \in S$, then $x + 1 \in S$.

Example 0.5.6

i. R is an inductive subset of R.

ii. $S = \{x \in R \mid x \geq 1\}$ is an inductive subset of R.

Now define N to be the intersection of all the inductive subsets of R. It is clear that N is an inductive subset of R. Of course, N is supposed to be the natural numbers. Since of all the axioms for a commutative ring with 1, as well as the order axioms, hold in R, we can use them freely in N. The following facts are easy to prove, so prove them.

Facts 0.5.7 1. Suppose that S is a non-empty subset of N such that $1 \in S$ and if $x \in S$ then $x + 1 \in S$, show that $S = N$.

2. Show that N is closed under addition.

3. Show that N is closed under multiplication. Hint: fix $x \in N$ and look at the set $M_x = \{y \in N \mid xy \in N\}$. Show that M_x is an inductive subset.

4. Show that the well ordering principle holds in N.

5. Show that all elements of N are positive.

   This is all fine, but where do we get the integers? Well, of course, we just tack on 0 and the negative natural numbers. Before nodding your head and shouting "Hooray!", you must show that this new set $Z = N \cup \{0\} \cup \{n \in R \mid -n \in N\}$ is closed under multiplication and addition.

6. Show that if $m, n \in N$ then $m - n \in Z$. In particular if $m \in N$ then $m - 1 \in Z$.

7. Show that Z is closed under addition.

8. Show that Z is closed under multiplication.

   So we have that Z is an ordered integral domain in which the positive elements are well ordered.

9. Show that Z and the integers, $\mathbb{Z}$, are order isomorphic. That is, there exists a bijection $\phi : Z \to \mathbb{Z}$ such that

   (a) $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in Z$,

   (b) $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in Z$, and

   (c) if $x < y$ in Z, then $\phi(x) < \phi(y)$ in $\mathbb{Z}$.

Chapter 1

The Real and Complex Numbers

> *Thus the System of Real Numbers—the definition of irrationals and the extension of the four species to the new numbers—is established. The method has the advantage of simplicity in detail. It is well for the student, after a first study of the method of Dedekind, to work it through in detail. He will then return to the former method with increased power and greater zest.*
> *The method of regular sequences is a middle-of-the-road method. It is an easy way to reach the mountain top. The traveller buys his ticket and takes the funicular. Many people prefer this mode of travel. But some like a stiff climb over rocks and across streams, and such an ascent has its advantages if the heart is good and the muscles are strong.*
> *– William Fogg Osgood – Functions of Real Variables*

In Chapter 0, we defined the integers and discussed their properties in some detail. We then constructed the rational numbers from the integers and observed that the rational numbers form an ordered field. It follows from Project 0.2 that any field which contains the integers must also contain the rationals as a subfield.

Exercise 1.0.1 Prove that any field that contains the integers contains the rationals as a subfield.

In this chapter, we do several things. First, we introduce the real numbers by adding the Least Upper Bound Property to the axioms for an ordered field. Second, despite Osgood, we construct the real numbers from the rational numbers by the method of Cauchy sequences. Third, we construct the complex numbers from the real numbers and prove a few useful theorems about complex numbers. Intermingled in all of this is a discussion of the fields of algebraic numbers and real algebraic numbers. As a project at the end of the chapter, we lead the reader through the construction of the real numbers via Dedekind cuts. In a second project, we study decimal expansions of real numbers.

1.1 The Least Upper Bound Property and the Real Numbers

Definition 1.1.1 Let F be an ordered field. Let A be a nonempty subset of F. We say that A is *bounded above* if there is an element $M \in F$ with the property that if $x \in A$, then $x \leq M$. We call M an *upper bound* for A. Similarly, we say that A is *bounded below* if there is an element $m \in F$ such that if $x \in A$, then $m \leq x$. We call m a *lower bound* for A. We say that A is *bounded* if A is bounded above and A is bounded below.

Examples 1.1.2 *i.* Consider the subset A of $\mathbb{Q}$:

$$A = \left\{ 1 + \frac{(-1)^n}{n} \;\middle|\; n \in \mathbb{N} \right\}.$$

Then A is bounded above by $\frac{3}{2}$ and bounded below by 0.

ii. Let $A = \{x \in \mathbb{Q} \mid 0 < x^3 < 27\}$. Then A is bounded below by 0 and bounded above by 3.

Exercise 1.1.3 Let a be a positive rational number. Let $A = \{x \in \mathbb{Q} \mid x^2 < a\}$. Show that A is bounded in $\mathbb{Q}$.

Definition 1.1.4 Let F be an ordered field, and let A be a nonempty subset of F which is bounded above. We say that $L \in F$ is a *least upper bound* for A if the following two conditions hold:

a. L is an upper bound for A;

b. if M is any upper bound for A, then $L \leq M$.

The definition of a *greatest lower bound* is similar with all of the inequalities reversed.

Exercise 1.1.5 Show the least upper bound of a set is unique if it exists.

Previously, we have discussed the real numbers in an informal way as a collection of decimal expansions, with the property that no expansion ends in all nines. Of course, this is not a formal definition of the real numbers, but it is common practice to work with the real numbers with this particular representation. We now give a formal definition of the real numbers which provides a working basis for proving theorems. Later in this chapter, starting with the rational numbers as an ordered field we will give a precise construction of the real numbers as an ordered field in which the least upper bound property holds.

Definition 1.1.6 An ordered field F has the *least upper bound property* if every nonempty subset A of F that is bounded above has a least upper bound.

Exercise 1.1.7 Show that any two ordered fields with the least upper bound property are order isomorphic.

We will see in Section 1.5 that an ordered field with the least upper bound property exists. Thus, it makes sense to make the following definition.

Definition 1.1.8 The *real numbers* are the (unique, up to order isomorphism) ordered field that satisfies the least upper bound property. We will denote this field by $\mathbb{R}$.

We say that the real numbers are an ordered field with the least upper bound property. In many texts, the real numbers are defined as a *complete ordered field*. This is actually a misuse of the word "complete," which is defined in terms of the convergence of Cauchy sequences. This will be discussed later in this chapter.

Exercise 1.1.9 Find the least upper bound in $\mathbb{R}$ of the sets in Exercise 1.1.3 and Example 1.1.2.

Definition 1.1.10 An ordered field F has the *greatest lower bound property* if every nonempty subset A of F that is bounded below has a greatest lower bound. That is, there exists an element ℓ of F such that:

a. ℓ is a lower bound for A;

b. if m is any lower bound for A, then $m \leq \ell$.

Exercise 1.1.11 Prove that an ordered field has the least upper bound property iff it has the greatest lower bound property.

If L is the least upper bound of a set A, we write $L = \text{lub } A$ or $L = \sup A$ (sup stands for *supremum*). If ℓ is the greatest lower bound of a set A, we write $\ell = \text{glb } A$ or $\ell = \inf A$ (inf stands for *infimum*).

Exercise 1.1.12 Let n be a positive integer that is not a perfect square. Let $A = \{x \in \mathbb{Q} \mid x^2 < n\}$. Show that A is bounded in $\mathbb{Q}$ but has neither a greatest lower bound nor a least upper bound in $\mathbb{Q}$. Conclude that $\sqrt{n}$ exists in $\mathbb{R}$, that is, there exists a real number a such that $a^2 = n$.

We have observed that the rational numbers are contained in $\mathbb{R}$. A real number is *irrational* if it is not in $\mathbb{Q}$.

Fact 1.1.13 We can conclude from Exercise 1.1.12 that if n is a positive integer that is not a perfect square, then $\sqrt{n}$ exists in $\mathbb{R}$ and is irrational.

Exercise 1.1.14 Suppose that A and B are bounded sets in $\mathbb{R}$. Prove or disprove the following:

i. $\text{lub}(A \cup B) = \max\{\text{lub}(A), \text{lub}(B)\}$.

ii. If $A + B = \{a + b \mid a \in A, b \in B\}$, then $\text{lub}(A + B) = \text{lub}(A) + \text{lub}(B)$.

iii. If the elements of A and B are positive and $A \cdot B = \{ab \mid a \in A, b \in B\}$, then $\text{lub}(A \cdot B) = \text{lub}(A)\text{lub}(B)$.

iv. Formulate the analogous problems for the greatest lower bound.

1.2 Consequences of the Least Upper Bound Property

We now present some facts that follow from the least upper bound property and the properties of the integers. The first is the *Archimedean property* of the real numbers.

Theorem 1.2.1 (Archimedean Property of $\mathbb{R}$) *If a and b are positive real numbers, then there exists a natural number n such that $na > b$.*

Proof. If $a > b$, take $n = 1$. If $a = b$, take $n = 2$. If $a < b$, consider the set $S = \{na \mid n \in \mathbb{N}\}$. The set $S \neq \varnothing$ since $a \in S$. Suppose S is bounded above by b. Let $L = \text{lub } S$. Then, since $a > 0$, there exists an element $n_0 a \in S$ such that $L - a < n_0 a$. But then $L < (n_0 + 1)a$, which is a contradiction. ☠

Corollary 1.2.2 *If ε is a positive real number, there exists a natural number n such that $\frac{1}{n} < \varepsilon$.*

Definition 1.2.3 Let F be an ordered field. From Chapter **??**, we know that $\mathbb{Z} \subset F$ and by Exercise 1.0.1 we know $\mathbb{Q} \subset F$. We say that F is an *Archimedean ordered field* if for every $x \in F$ there exists $N \in \mathbb{Z}$ such that $x < N$.

The fields $\mathbb{Q}$ and $\mathbb{R}$ are Archimedean ordered fields.

Exercise 1.2.4 Let F be an Archimedean ordered field. Show that F is order isomorphic to a subfield of $\mathbb{R}$.

Next, we show that every real number lies between two successive integers.

Theorem 1.2.5 *If a is a real number, then there exists an integer N such that $N - 1 \leq a < N$.*

Proof. Let $S = \{n \in \mathbb{Z} \mid n > a\}$. Then by the Archimedean property, $S \neq \varnothing$. S is bounded below by a, so by the well ordering principle S has a least element N. Then $N - 1 \notin S$, so $N - 1 \leq a < N$. ☠

We now show that there is a rational number between any two real numbers.

Theorem 1.2.6 If a and b are real numbers with $a < b$, there exists a rational number $r = \frac{p}{q}$ such that $a < r < b$.

Proof. From the Archimedean property of $\mathbb{R}$ (Corollary 1.2.2) there exists $q \in \mathbb{N}$ such that $\frac{1}{q} < b - a$. Now consider the real number qa. By Theorem 1.2.5, there exists an integer p such that $p - 1 \leq qa < p$. It follows that $\frac{p-1}{q} \leq a < \frac{p}{q}$. This implies that $\frac{p}{q} - \frac{1}{q} \leq a$, or $a < \frac{p}{q} \leq a + \frac{1}{q} < b$.

Definition 1.2.7 A subset A of $\mathbb{R}$ is said to be *dense* in $\mathbb{R}$ if for any pair of real numbers a and b with $a < b$, there is an $r \in A$ such that $a < r < b$.

Corollary 1.2.8 The rational numbers are dense in the real numbers.

How do the irrational numbers behave?

Exercise 1.2.9

i. Show that any irrational number multiplied by any nonzero rational number is irrational.

ii. Show that the product of two irrational numbers may be rational or irrational.

Next we show that there is an irrational number between any two real numbers.

Corollary 1.2.10 The irrational numbers are dense in $\mathbb{R}$.

Proof. Take $a, b \in \mathbb{R}$ such that $a < b$. We know that $\sqrt{2}$ is irrational and greater than 0. But then $\frac{a}{\sqrt{2}} < \frac{b}{\sqrt{2}}$. By Corollary 1.2.8, there exists a rational number p/q, with $p \neq 0$ such that $\frac{a}{\sqrt{2}} < \frac{p}{q} < \frac{b}{\sqrt{2}}$. Thus $a < \sqrt{2}p/q < b$, and $\sqrt{2}p/q$ is irrational.

The real numbers are the union of two disjoint sets, the rational numbers and the irrational numbers, and each of these sets is dense in $\mathbb{R}$. Note that density implies nothing about cardinality since the rationals are countable and the irrationals are not, as shown in Section A.8.

1.3 Rational Approximation

We have just shown that both the rational numbers and the irrational numbers are dense in the real numbers. But, really, how dense are they? It is reasonable to think that proximity for rational numbers can be measured in terms of the size of the denominator. To illustrate this, we ask the question, "How close do two rational numbers have to be in order to be the same rational number?" This is not a trick question – it is designed to illustrate the principle mentioned above. Thus, if $a/b, c/d \in \mathbb{Q}$ and $|a/b - c/d| < 1/bd$, then $a/b = c/d$.

This idea can be encapsulated in the following theorem. Throughout this section, we shall assume that the denominator of a rational number is a positive integer and that the numerator and denominator are relatively prime.

Theorem 1.3.1 If a/b is a fixed rational number and p/q is a rational number such that $0 < |p/q - a/b| < 1/mb$ for some positive integer m, then $q > m$.

Proof. Simplify the subtraction of fractions and multiply both sides by bq.

We now present several facts on rational approximation. For the rest of this section, we assume that the reader is familiar with the results contained in Project 0.5.1 at the end of Chapter 0.

Exercise 1.3.2 Let a and b be relatively prime integers. Show that the equation $ax + by = 1$ has infinitely many solutions (q, p) with q and p relatively prime integers.

Theorem 1.3.3 Let $\alpha = a/b$ with a and b relatively prime and $b \neq 1$. Then there exist infinitely many $p/q \in \mathbb{Q}$ such that $|a/b - p/q| < 1/q$.

Proof. Let $(q, -p)$ be a solution to the equation $ax + by = 1$. Then $q \neq 0$ since $b \neq 1$. We may assume $q > 0$. We then have $|a/b - p/q| = 1/bq < 1/q$.

Remark 1.3.4 If $b = 1$ then the same result holds with $<$ replaced by $\leq$.

The next theorem characterizes rational numbers in terms of rational approximation. We first need the following exercise.

Exercise 1.3.5 Let α be a real number, and let η and t be positive real numbers. Show that there exists only a finite number of rational numbers p/q with $q < \eta$ that satisfy $|\alpha - p/q| < 1/q^t$.

Theorem 1.3.6 Let $\alpha = a/b \in \mathbb{Q}$. Then there are only finitely many p/q so that $|\alpha - p/q| \leq 1/q^2$.

Proof. Suppose there are infinitely many p/q satisfying the inequality. Then by the exercise above, q gets arbitrarily large. Thus there exists a p/q with $q > b$ such that $|a/b - p/q| < 1/q^2$. This implies that $|aq - bp| < b/q < 1$, which is a contradiction.

We next consider rational approximation of irrational numbers. The question is: if α is irrational, are there any rational numbers p/q satisfying the inequality $|\alpha - p/q| < 1/q^2$? The affirmative answer follows from a theorem of Dirichlet on rational approximation of any real number.

Theorem 1.3.7 (Dirichlet) Let α be a real number and n a positive integer. Then there is a rational number p/q with $0 < q \leq n$ satisfying the inequality

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{(n+1)q}.$$

Proof. If $n = 1$, then $p/q = [\alpha]$ or $p/q = [\alpha + 1]$ satisfies $|\alpha - p/q| \leq 1/2$. (Recall that $[\alpha]$ is the greatest integer less than or equal to α; for further details, see Appendix A.) Suppose that $n \geq 2$. Consider the $n + 2$ numbers

$$0, \alpha - [\alpha], 2\alpha - [2\alpha], \ldots, n\alpha - [n\alpha], 1$$

in the interval $[0, 1]$. Assume that the numbers in our list are distinct, which is the case if α is irrational. By the pigeonhole principle, two of the numbers differ in absolute value by at most $1/(n+1)$. If one of the numbers is 0 and the other is $i\alpha - [i\alpha]$, then $i \leq n$, $|i\alpha - [i\alpha]| \leq 1/(n+1)$, and

$$\left|\alpha - \frac{[i\alpha]}{i}\right| \leq \frac{1}{(n+1)i}.$$

After $[i\alpha]/i$ is reduced to lowest terms p/q, the rational number p/q satisfies the required inequality. Similarly, if the two numbers are $j\alpha - [j\alpha]$ and 1, then $j \leq n$ and reducing $([j\alpha] + 1)/j$ to lowest terms p/q, we have p/q satisfies the required inequality. Finally, if the two numbers are $i\alpha - [i\alpha]$ and $j\alpha - [j\alpha]$, where $i < j$, then

$$|j\alpha - [j\alpha] - (i\alpha - [i\alpha])| = |(j - i)\alpha + ([j\alpha] - [i\alpha])| \leq \frac{1}{n+1}.$$

Then

$$\left|\alpha - \frac{[j\alpha] - [i\alpha]}{j - i}\right| \leq \frac{1}{(n+1)(j-i)}.$$

Thus, after $([j\alpha] - [i\alpha])/(j - i)$ is reduced to lowest terms p/q, the rational number p/q satisfies the inequality because $j - i < n$.

In the event that the $n+2$ numbers are not distinct, then α itself is a rational number with denominator at most n. For this case, either there exists $1 \le i \le n$ so that

$$\alpha = \frac{[i\alpha]}{i}$$

or there exist $1 \le i < j \le n$ so that

$$\alpha = \frac{[j\alpha] - [i\alpha]}{j - i}$$

. Thus, if the numbers are not distinct the required inequality is trivially satisfied by α itself.

Corollary 1.3.8 Given any real number α there is a rational number p/q such that $|\alpha - p/q| < 1/q^2$.

Proof. This follows immediately from the theorem.

Now comes the good news (or bad news depending on how you look at it).

Theorem 1.3.9 If α is irrational, then there are infinitely many rational numbers p/q such that $|\alpha - p/q| < 1/q^2$.

Proof. Suppose there are only a finite number of rational numbers p_1/q_1, p_2/q_2, $\ldots, p_k/q_k$ satisfying the inequality. Then, there is a positive integer n such that $|\alpha - p_i/q_i| > 1/(n+1)q_i$ for $i = 1, 2, \ldots, k$. This contradicts Theorem 1.3.7 which asserts the existence of a rational number p/q satisfying $q \le n$ and $|\alpha - p/q| < 1/(n+1)q < 1/q^2$.

So, there you have it, a real number α is rational if and only if there exists only a finite number of rational numbers p/q such that $|\alpha - p/q| \le 1/q^2$. And a real number α is irrational if and only if there exists an infinite number of rational numbers p/q such that $|\alpha - p/q| \le 1/q^2$.

1.4 Intervals

At this stage we single out certain subsets of $\mathbb{R}$ which are called intervals.

Definition 1.4.1 A subset of $\mathbb{R}$ is an *interval* if it falls into one of the following categories.

a. For $a, b \in \mathbb{R}$ with $a < b$, the *open interval* (a, b) is defined by $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$.

b. For $a, b \in \mathbb{R}$ with $a \le b$, the *closed interval* $[a, b]$, is defined by $[a, b] = \{x \in \mathbb{R} \mid a \le x \le b\}$.

c. For $a, b \in \mathbb{R}$ with $a < b$, the *half open interval* $[a, b)$, is defined by $[a, b) = \{x \in \mathbb{R} \mid a \le x < b\}$.

d. For $a, b \in \mathbb{R}$ with $a < b$, the *half open interval* $(a, b]$, is defined by $(a, b] = \{x \in \mathbb{R} \mid a < x \le b\}$.

e. For $a \in \mathbb{R}$, the *infinite open interval* (a, ∞), is defined by $(a, \infty) = \{x \in \mathbb{R} \mid a < x\}$.

f. For $b \in \mathbb{R}$, the *infinite open interval* $(-\infty, b)$, is defined by $(-\infty, b) = \{x \in \mathbb{R} \mid x < b\}$.

g. For $a \in \mathbb{R}$, the *infinite closed interval* $[a, \infty)$, is defined by $[a, \infty) = \{x \in \mathbb{R} \mid a \le x\}$.

h. For $b \in \mathbb{R}$, the *infinite closed interval* $(-\infty, b]$, is defined by $(-\infty, b] = \{x \in \mathbb{R} \mid x \le b\}$.

i. $\mathbb{R} = (-\infty, \infty)$.

Definition 1.4.2 If $x \in \mathbb{R}$ a *neighborhood* of x is an open interval containing x. For many instances, it is useful to use symmetric neighborhoods. That is, if $x \in \mathbb{R}$ a *symmetric neighborhood* of x is an interval of the form $(x - \varepsilon, x + \varepsilon)$, where $\varepsilon > 0$.

These intervals, and their counterparts in other spaces, are used extensively throughout analysis.

Exercise 1.4.3 Suppose that I is a subset of $\mathbb{R}$, show that I is an interval if and only if for all $a, b \in I$, with $a \leq b$, the closed interval $[a, b]$ is contained in I.

The notion of interval is valid in any ordered field, and we will occasionally find this useful. We end with this section with a theorem about intervals in $\mathbb{R}$, which is called the Nested Intervals Theorem.

Theorem 1.4.4 (Nested Intervals Theorem) Let $([a_n, b_n])_{n \in \mathbb{N}}$ be a nested sequence of closed bounded intervals in $\mathbb{R}$. That is, for any n we have
$[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$, or equivalently, $a_n \leq a_{n+1} \leq b_{n+1} \leq b_n$ for all n. Then $\bigcap_{n \in \mathbb{N}} [a_n, b_n] \neq \varnothing$.

Proof. Let $A = \{a_n \mid n \in \mathbb{N}\}$. Then A is bounded above by b_1. If $a = \text{lub} A$, then $a \in \bigcap_{n \in \mathbb{N}} [a_n, b_n]$.

The nested intervals property is actually not exclusive to the real numbers. In fact, it is really a theorem about a sequence of nested compact sets in a metric space. This result will be proved in the Chapter **??**. There is often some confusion about the relationship between the nested interval theorem in $\mathbb{R}$ and the least upper bound property. Although our proof in $\mathbb{R}$ involves the least upper bound property, it can be done in alternate ways.

Exercise 1.4.5 *i.* Give an example of a nested sequence of closed (but not necessarily bounded) intervals such that the intersection is empty.

ii. Give an example of a nested sequence of bounded (but not necessarily closed) intervals such that the intersection is empty.

1.5 The Construction of the Real Numbers

We are now ready to proceed with the construction of the real numbers from the rational numbers using the fact that the rational numbers are the ordered field constructed from $\mathbb{Z}$ in Chapter 0. We have already defined $\mathbb{R}$ as an ordered field in which the least upper bound property holds. We now proceed to build such a field starting from $\mathbb{Q}$.

Recall that the *absolute value* on $\mathbb{Q}$ is defined as follows:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0. \end{cases}$$

Also recall that the absolute value on $\mathbb{Q}$ satisfies the following three properties.

1. For any $a \in \mathbb{Q}$, $|a| \geq 0$, and $|a| = 0$ if and only if $a = 0$.

2. For any $a, b \in \mathbb{Q}$, $|ab| = |a||b|$.

3. For any $a, b \in \mathbb{Q}$, $|a + b| \leq |a| + |b|$ (triangle inequality).

Exercise 1.5.1 Show that, for any $a, b \in \mathbb{Q}$, we have $||a| - |b|| \leq |a - b|$.

Definition 1.5.2 A *sequence* of rational numbers is a function $f : \mathbb{N} \to \mathbb{Q}$.

We often denote such a sequence by $(f(1), f(2), \ldots, f(k), \ldots)$, or $(a_1, a_2, \ldots, a_k, \ldots)$, where $a_k = f(k)$ for each $k \in \mathbb{N}$. Most frequently, we will use $(a_k)_{k \in \mathbb{N}}$ for this same sequence.

Definition 1.5.3 A sequence $(a_k)_{k\in\mathbb{N}}$ of rational numbers is a *Cauchy sequence* in $\mathbb{Q}$ if, given any rational number $r > 0$, there exists an integer N such that if $n, m \geq N$, then $|a_n - a_m| < r$.

Definition 1.5.4 A sequence $(a_k)_{k\in\mathbb{N}}$ *converges in* $\mathbb{Q}$ to $a \in \mathbb{Q}$ if, given any rational number $r > 0$, there exists an integer N such that, if $n \geq N$, then $|a_n - a| < r$. The rational number a is calld the *limit* of the sequence $(a_k)_{k\in\mathbb{N}}$. Sometimes, we just say that the sequence $(a_k)_{k\in\mathbb{N}}$ converges in $\mathbb{Q}$ without mentioning the limit a.

Exercise 1.5.5 If a sequence $(a_k)_{k\in\mathbb{N}}$ converges in $\mathbb{Q}$, show that $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{Q}$.

Exercise 1.5.6 Show that the limit a of a convergent sequence is unique.

Definition 1.5.7 Let $(a_k)_{k\in\mathbb{N}}$ be a sequence of rational numbers. We say that $(a_k)_{k\in\mathbb{N}}$ is a *bounded sequence* if the set $\{a_k \mid k \in \mathbb{N}\}$ is a bounded set in $\mathbb{Q}$.

Lemma 1.5.8 Let $(a_k)_{k\in\mathbb{N}}$ be a Cauchy sequence of rational numbers. Then $(a_k)_{k\in\mathbb{N}}$ is a bounded sequence.

Proof. Let $(a_k)_{k\in\mathbb{N}}$ be a Cauchy sequence of rational numbers. Pick $N \in \mathbb{N}$ such that $|a_n - a_m| < 1$ for $n, m \geq N$. Then $|a_n - a_N| < 1$ for all $n \geq N$, so that $|a_n| < 1 + |a_N|$ for all $n \geq N$. Let M be the max of $|a_1|, |a_2|, \ldots, |a_{N-1}|, 1 + |a_N|$. Then $\{|a_k| \mid k \in \mathbb{N}\}$ is bounded above by M.

Let $\mathcal{C}$ denote the set of all Cauchy sequences of rational numbers. We define addition and multiplication of Cauchy sequences term-wise, that is, $(a_n)_{n\in\mathbb{N}} + (b_n)_{n\in\mathbb{N}} = (a_n + b_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}(b_n)_{n\in\mathbb{N}} = (a_n b_n)_{n\in\mathbb{N}}$.

Exercise 1.5.9 Show that the sum of two Cauchy sequences in $\mathbb{Q}$ is a Cauchy sequence in $\mathbb{Q}$.

Theorem 1.5.10 The product of two Cauchy sequences in $\mathbb{Q}$ is a Cauchy sequence in $\mathbb{Q}$.

Proof. Let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be Cauchy sequences in $\mathbb{Q}$. By the lemma above, these sequences must be bounded. Let A and B be upper bounds for the sequences $(|a_k|)_{k\in\mathbb{N}}$ and $(|b_k|)_{k\in\mathbb{N}}$, respectively.

Let $r > 0$. Since $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, we can choose N_1 such that if $n, m > N_1$, then $|a_n - a_m| < \frac{r}{2B}$. Since $(b_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, we can choose N_2 such that if $n, m > N_2$, then $|b_n - b_m| < \frac{r}{2A}$. Let $N = \max\{N_1, N_2\}$. If $n, m > N$, then

$$\begin{aligned}
|a_n b_n - a_m b_m| &= |a_n b_n - a_n b_m + a_n b_m - a_m b_m| \\
&\leq |a_n||b_n - b_m| + |b_m||a_n - a_m| \\
&\leq A|b_n - b_m| + B|a_n - a_m| \\
&< r.
\end{aligned}$$

Exercise 1.5.11 Show that, with addition and multiplication defined as above, $\mathcal{C}$ is a commutative ring with 1.

Let $\mathcal{I}$ be the set of sequences $(a_k)_{k\in\mathbb{N}}$ in $\mathcal{C}$ with the property that, given any rational $r > 0$, there exists an integer N such that if $n \geq N$, then $|a_n| < r$. The set $\mathcal{I}$ consists of Cauchy sequences that converge to 0.

Lemma 1.5.12 Suppose $(a_k)_{k\in\mathbb{N}} \in \mathcal{C} \setminus \mathcal{I}$, then there exists a positive rational number r and an integer N such that $|a_n| \geq r$ for all $n \geq N$.

Proof. Suppose $(a_k)_{k \in \mathbb{N}} \notin \mathcal{I}$. Then there exists a rational number $r > 0$ such that $|a_k| \geq 2r$ infinitely often. Pick $N \in \mathbb{N}$ such that $|a_n - a_m| < r$ for $n, m \geq N$. This implies that

$$|a_n| > |a_m| - r \textbf{ for } n, m \geq N\textbf{.}$$

Fix an $m \geq N$ for which $|a_m| \geq 2r$. Then for all $n \geq N$, we have $|a_n| > r$.

We can rephrase the statement of this lemma. We say that a property of rational numbers holds *eventually* for the terms of a sequence $(a_k)_{k \in \mathbb{N}}$ if there exists some $N \in \mathbb{N}$ such that the property holds for a_n whenever $n \geq N$. So the lemma above says that for a Cauchy sequence $(a_k)_{k \in \mathbb{N}}$ that is not in $\mathcal{I}$, there exists a positive rational number r such that $|a_k|$ is eventually greater than or equal to r.

Exercise 1.5.13 Show that if a Cauchy sequence $(a_k)_{k \in \mathbb{N}}$ does not converge to 0, all the terms of the sequence eventually have the same sign.

Definition 1.5.14 Let $(a_k)_{k \in \mathbb{N}}$, and $(b_k)_{k \in \mathbb{N}}$ be Cauchy sequences in $\mathbb{Q}$. We say that $(a_k)_{k \in \mathbb{N}}$ is *equivalent* to $(b_k)_{k \in \mathbb{N}}$, denoted by $(a_k)_{k \in \mathbb{N}} \sim (b_k)_{k \in \mathbb{N}}$, if $(c_k)_{k \in \mathbb{N}} = (a_k - b_k)_{k \in \mathbb{N}}$ is in $\mathcal{I}$.

Exercise 1.5.15 Show that $\sim$ defines an equivalence relation on $\mathcal{C}$.

Denote by $\mathbf{R}$ the set of equivalence classes in $\mathcal{C}$. We claim that, with appropriate definitions of addition and multiplication (already indicated above) and order (to be defined below), $\mathbf{R}$ is an ordered field satisfying the least upper bound property.

If $(a_k)_{k \in \mathbb{N}}$ is a Cauchy sequence, denote its equivalence class by $[a_k]$. As one might expect, the sum and product of equivalence classes are defined as follows: $[a_k] + [b_k] = [a_k + b_k]$ and $[a_k][b_k] = [a_k b_k]$.

Exercise 1.5.16 Show that addition and multiplication are well defined on $\mathbf{R}$.

Exercise 1.5.17 Show that $\mathbf{R}$ is a commutative ring with 1, with $\mathcal{I}$ as the additive identity and $[a_k]$ such that $a_k = 1$ for all k as the multiplicative identity. This follows easily from Exercise 1.5.11.

Theorem 1.5.18 $\mathbf{R}$ is a field.

Proof. We need only show that multiplicative inverses exist for non-zero elements. So assume that $[a_k] \neq \mathcal{I}$. Then, as we saw in Lemma 1.5.12, a_k is eventually bounded below in absolute value. Hence, we can pick $M \in \mathbb{N}$ and $c > 0$ such that $|a_k| > c$ for all $k \geq M$. Define a sequence $(b_k)_{k \in \mathbb{N}}$ as follows: $b_k = 1$ for $k \leq M$, and $b_k = 1/a_k$ for $k > M$. Observe that for n, m large enough,

$$\left| \frac{1}{a_n} - \frac{1}{a_m} \right| = \frac{|a_n - a_m|}{|a_n a_m|} \leq \frac{1}{c^2} |a_n - a_m|.$$

Hence, $(b_k)_{k \in \mathbb{N}}$ is a Cauchy sequence. It is clear by construction that $[b_k]$ is the multiplicative inverse of $[a_k]$.

The next step is to define order on $\mathbf{R}$. Let $[a_k]$ and $[b_k]$ represent distinct elements of $\mathbf{R}$. Then $[c_k] = [a_k - b_k]$ is not equal to $\mathcal{I}$. Hence there exists $N \in \mathbb{N}$ such that all the terms of c_k have the same sign for $k \geq N$. Thus, either $a_k < b_k$ for all $k \geq N$ or $b_k < a_k$ for $k \geq N$. We use this fact to define an order on $\mathbf{R}$.

Definition 1.5.19 Let $a = [a_k], b = [b_k]$ be distinct elements of $\mathbf{R}$. We define $a < b$ if $a_k < b_k$ eventually and $b < a$ if $b_k < a_k$ eventually.

Exercise 1.5.20 Show that the order relation on $\mathbf{R}$ defined above is well-defined and makes $\mathbf{R}$ an ordered field.

To finish this off, we must show that $\mathbf{R}$ is an Archimedean ordered field that satisfies the least upper bound property. We will have then reached the Osgood's mountain top so we can dismount the funicular and ski happily down the slope.

Define a map $i : \mathbb{Q} \to \mathbf{R}$ by sending $r \in \mathbb{Q}$ to the equivalence class of the constant sequence $(r, r, \dots)$. It is evident that this map is injective and order-preserving, so we may consider $\mathbb{Q} \subset \mathbf{R}$ as ordered fields.

Theorem 1.5.21 The field $\mathbf{R}$ is an Archimedean ordered field.

Proof. Suppose $a \in \mathbf{R}$ and $a > 0$. Let $(a_k)_{k\in\mathbb{N}}$ represents a. As noted above, the Cauchy sequence $(a_k)_{k\in\mathbb{N}}$ is bounded above by some integer N, that is, $a_k < N$ for all sufficiently large k. It follows that a is less than the integer $(N, N, \dots)$ in $\mathbf{R}$ (under the inclusion $\mathbb{Q} \subset \mathbf{R}$).

Theorem 1.5.22 The least upper bound property holds in $\mathbf{R}$.

Proof. Let A be a nonempty subset of $\mathbf{R}$ that is bounded above by, say, m. Then, by the Archimedean property, we can find $M \in \mathbb{Z}$ with $m \leq M$. Let a be in A and let n be an integer with $n < a$. For $p \in \mathbb{N}$, set $S_p = \{k2^{-p} \mid k \in \mathbb{Z} \text{ and } n \leq k2^{-p} \leq M\} \cup \{m\}$. Note that $S_p \neq \varnothing$ and is finite. Now let $a_p = \min\{x \mid x \in S_p \text{ and } x \text{ is an upper bound for A}\}$.

Note that if $p < q$, then

$$a_p - 2^{-p} < a_q \leq a_p,$$

since, for example, $a_p - 2^{-p}$ is not an upper bound for A, while a_q is an upper bound. But this implies that

$$|a_p - a_q| \leq 2^{-p} \text{ for all } p < q$$

from which it follows that $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. Let $L = [a_k]$.

We claim that L is a least upper bound for A. Suppose $x \in A$ and $x > L$. Choose p such that $2^{-p} < (x - L)$ (using the Archimedean property). Since $a_p - 2^{-p} < a_q$ for $p < q$ and (a_p) is a decreasing Cauchy sequence, it follows that $a_p - 2^{-p} \leq L \leq a_p$. In particular if we add $2^{-p} < x - L$ and $a_p - 2^{-p} \leq L$ we obtain $a_p < x$ which is a contradiction. Therefore L is an upper bound for A.

Suppose that H is an upper bound for A and $H < L$. Choose p such that $2^{-p} < L - H$. Take $x \in A$ such that $a_p - 2^{-p} < x$. Then $a_p - 2^{-p} < H$. Adding, we get $a_p < L$. But, as noted above, $L \leq a_p$ for all $p \in \mathbb{N}$, so this is a contradiction.

In this section, we have constructed an ordered field $\mathbf{R}$ in which the least upper bound property holds. Following our discussion in Section 1.1, this means that $\mathbf{R}$ must in fact, up to order isomorphism, be the real numbers $\mathbb{R}$. From now on, we will refer to the real numbers exclusively as $\mathbb{R}$.

1.6 Convergence in $\mathbb{R}$

We define the absolute value on $\mathbb{R}$ in exactly the same manner as on $\mathbb{Q}$.

Definition 1.6.1 Suppose $x \in \mathbb{R}$. The *absolute value* of x is defined by

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

The following are the essential properties of the absolute value.

Theorem 1.6.2 Properties of absolute value on $\mathbb{R}$

1. For any $x \in \mathbb{R}$, $|x| \geq 0$, and $|x| = 0$ iff $x = 0$.

2. For any $x, y \in \mathbb{R}$, $|xy| = |x||y|$.

3. For any $x, y \in \mathbb{R}$, $|x+y| \leq |x| + |y|$ (triangle inequality).

Exercise 1.6.3 Prove the properties of the absolute value.

With absolute value defined, we can talk about Cauchy and convergent sequences in $\mathbb{R}$. Of course, we should first define what a sequence in $\mathbb{R}$ is.

Definition 1.6.4 A *sequence* of real numbers is a function $f : \mathbb{N} \to \mathbb{R}$.

As with rational numbers, we will usually denote a sequence of real numbers by $(a_k)_{k\in\mathbb{N}}$, where $a_k = f(k)$.

Definition 1.6.5 A sequence $(a_k)_{k\in\mathbb{N}}$ of real numbers is *convergent* if there exists an element $a \in \mathbb{R}$ such that given any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $k \geq N_\varepsilon$ implies that $|a_k - a| < \varepsilon$. We say that $(a_k)_{k\in\mathbb{N}}$ *converges to* a, and a is called the *limit* of the sequence $(a_k)_{k\in\mathbb{N}}$. Symbolically, we write

$$\lim_{k\to\infty} a_k = a.$$

We will often say that a sequence of real numbers is convergent without specific reference to the limit a. Note that N_ε depends on the choice of ε.

Exercise 1.6.6 Show that the limit a of a convergent sequence is unique.

Definition 1.6.7 Given a sequence $(a_k)_{k\in\mathbb{N}}$, a *subsequence* is a sequence of the form $(a_{k_j})_{j\in\mathbb{N}}$, where $(k_j)_{j\in\mathbb{N}}$ is a strictly monotonic increasing sequence of natural numbers.

In other words, a subsequence of a sequence contains some, but not necessarily all, terms of the original sequence in their original order.

Exercise 1.6.8 Let $(a_k)_{k\in\mathbb{N}}$ be a convergent sequence. Show that any subsequence converges to the same limit.

Definition 1.6.9 A sequence $(a_k)_{k\in\mathbb{N}}$ of real numbers is *bounded* if the set $\{a_k \mid k \in \mathbb{N}\}$ is bounded.

Exercise 1.6.10 Show that any convergent sequence of real numbers is bounded.

Exercise 1.6.11 Find a bounded sequence of real numbers that is not convergent.

Definition 1.6.12 A sequence $(a_k)_{k\in\mathbb{N}}$ of real numbers is *monotonic increasing* if $a_k \leq a_{k+1}$ for all $k \in \mathbb{N}$. A sequence $(a_k)_{k\in\mathbb{N}}$ of real numbers is *strictly monotonic increasing* if $a_k < a_{k+1}$ for all $k \in \mathbb{N}$. Similar definitions hold for monotonic decreasing and strictly monotonic decreasing sequences with the inequalities reversed. A sequence is called *monotonic* if it is monotonic increasing or monotonic decreasing.

The following lemmas fundamental in discussing convergence in $\mathbb{R}$.

Lemma 1.6.13 Let $(a_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}$. Then $(a_k)_{k\in\mathbb{N}}$ has a monotonic subsequence.

Proof. Suppose $(a_k)_{k\in\mathbb{N}}$ does not have a monotonic increasing subsequence. Then, there exists $n_1 \in \mathbb{N}$ such that $a_{n_1} > a_k$ for all $k > n_1$. Again, since $(a_k)_{k>n_1}$ does not have a monotonic increasing subsequence, there exists $n_2 > n_1$ such that $a_{n_2} > a_k$ for all $k > n_2$. Moreover $a_{n_1} > a_{n_2}$. Continuing in this way, we obtain a strictly monotonic decreasing subsequence $(a_{n_1}, a_{n_2}, \ldots)$.

Lemma 1.6.14 Every bounded monotonic sequence in $\mathbb{R}$ converges to an element in $\mathbb{R}$.

Proof. Without loss of generality, suppose $(a_k)_{k\in\mathbb{N}}$ is monotonic increasing and bounded. Let a be the least upper bound of the set $\{a_1, a_2, \ldots\}$. For all $\varepsilon > 0$, there exists an N such that $a - \varepsilon < a_N \leq a$. Since $(a_k)_{k\in\mathbb{N}}$ is increasing, if $k > N$, we have $a \geq a_k \geq a_N > a - \varepsilon$. So $\lim_{k\to\infty} a_k = a$.

Lemma 1.6.15 Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.

Exercise 1.6.16 Prove Lemma 1.6.15.

Even if a bounded sequence of real numbers is not itself convergent, we can still say something about its long-term behavior.

Definition 1.6.17 Let $(a_k)_{k\in\mathbb{N}}$ be a bounded sequence of real numbers. For each $n \in \mathbb{N}$, define $b_n = \sup\{a_k \mid k \geq n\}$, and $c_n = \inf\{a_k \mid k \geq n\}$. We define the *limit supremum* of the sequence $(a_k)_{k\in\mathbb{N}}$ to be $\lim_{n\to\infty} b_n$, and we denote this by $\limsup_{k\to\infty} a_k$. We define the *limit infimum* of $(a_k)_{k\in\mathbb{N}}$ similarly: $\liminf_{k\to\infty} a_k = \lim_{n\to\infty} c_n$.

Exercise 1.6.18 Show that if $(a_k)_{k\in\mathbb{N}}$ is a bounded sequence, then $\limsup_{k\to\infty} a_k$ and $\liminf_{k\to\infty} a_k$ exist.

Exercise 1.6.19 Show that if $(a_k)_{k\in\mathbb{N}}$ is convergent, then $\limsup_{k\to\infty} a_k = \liminf_{k\to\infty} a_k = \lim_{k\to\infty} a_k$.

Exercise 1.6.20 Show that if $\limsup_{k\to\infty} a_k = \liminf_{k\to\infty} a_k$, then $(a_k)_{k\in\mathbb{N}}$ is convergent, and $\lim_{k\to\infty} a_k = \limsup_{k\to\infty} a_k = \liminf_{k\to\infty} a_k$.

Exercise 1.6.21 For the following sequences, compute lim sup and lim inf.

i. $a_k = 1 + (-1)^k$.

ii. $b_k = (-1)^k + \frac{1}{k}$.

iii. $c_k = \cos k$.

As we did for sequences of rational numbers, we can define Cauchy sequences of real numbers. However, unlike the situation in the rational numbers, we will see that every Cauchy sequence of real numbers converges in $\mathbb{R}$.

Definition 1.6.22 (See Definition 1.5.3) A sequence $(a_k)_{k\in\mathbb{N}}$ in $\mathbb{R}$ is a *Cauchy sequence* if, given any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $n, m \geq N_\varepsilon$ implies $|a_m - a_n| < \varepsilon$.

Exercise 1.6.23

i. Prove that every Cauchy sequence in $\mathbb{R}$ is bounded.

ii. If $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$, show that for any $\varepsilon > 0$ there exists a subsequence $(a_{k_j})_{j\in\mathbb{N}}$ such that $|a_{k_j} - a_{k_{j+1}}| < \varepsilon/2^{j+1}$ for $j \in \mathbb{N}$.

Theorem 1.6.24 (Cauchy Criterion) A sequence $(a_k)_{k\in\mathbb{N}}$ of real numbers is convergent if and only if it is a Cauchy sequence.

Proof. We already did half of this in $\mathbb{Q}$, (see Exercise 1.5.5) but we will do it again. First, we prove that if $(a_k)_{k\in\mathbb{N}}$ is convergent, then it is Cauchy. Suppose $\lim_{k\to\infty} a_k = a$. Then, since the sequence converges, given $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $|a_n - a| < \frac{\varepsilon}{2}$ for all $n \geq N_\varepsilon$. Thus, if $n, m \geq N_\varepsilon$, we have

$$|a_n - a_m| \leq |a_n - a| + |a_m - a| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

and so $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence.

Suppose now that $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. Then, by Exercise 1.6.23, $(a_k)_{k\in\mathbb{N}}$ is a bounded sequence, and hence by Lemma 1.6.15 has a convergent subsequence. Call the limit of this subsequence a. Then, since $(a_k)_{k\in\mathbb{N}}$ is Cauchy, it is clear that $\lim_{k\to\infty} a_k = a$.

Exercise 1.6.25 Show that if $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are Cauchy sequences in $\mathbb{R}$, then $(a_n+b_n)_{n\in\mathbb{N}}$ and $(a_n\cdot b_n)_{n\in\mathbb{N}}$ are Cauchy sequences in $\mathbb{R}$.

Definition 1.6.26 Let S be a subset of $\mathbb{R}$. Then $x\in\mathbb{R}$ is an *accumulation point* of S if, for all $\varepsilon>0$, we have $((x-\varepsilon,x+\varepsilon)\setminus\{x\})\cap S\neq\varnothing$.

Remark 1.6.27 Thus, x is an accumulation point of S if every interval around x contains points of S other than x. Of course, x does not have to be an element of S in order to be an accumulation point of S.

Exercise 1.6.28 Find the accumulation points of the following sets in $\mathbb{R}$.

i. $S=(0,1)$;

ii. $S=\{(-1)^n+\frac{1}{n}\mid n\in\mathbb{N}\}$;

iii. $S=\mathbb{Q}$;

iv. $S=\mathbb{Z}$;

v. S is the set of rational numbers whose denominators are prime.

Lemma 1.6.29 Let S be a subset of $\mathbb{R}$. Then every neighborhood of an accumulation point of S contains infinitely many points of S.

Proof. Let x be an accumulation point of S. Given $\varepsilon>0$, there is a point $x_1\in(x-\varepsilon,x+\varepsilon)\cap S$ such that $x_1\neq x$. Let $\varepsilon_1=|x-x_1|$. Then, there is a point $x_2\in(x-\varepsilon_1,x+\varepsilon_1)\cap S$ such that $x_2\neq x$. Iterating this procedure, we get an infinite set of elements in S that is contained in $(x-\varepsilon,x+\varepsilon)$.

Exercise 1.6.30 You prove the converse.

Now here is a Big Time Theorem.

Theorem 1.6.31 (Bolzano–Weierstrass) Let S be a bounded, infinite subset of $\mathbb{R}$. Then S has an accumulation point in $\mathbb{R}$.

Proof. Pick an infinite sequence $(a_k)_{k\in\mathbb{N}}$ of distinct elements of S. Then, by Lemma 1.6.15, $(a_k)_{k\in\mathbb{N}}$ has a convergent subsequence,$(a_{k_j})_{j\in\mathbb{N}}$. If $\lim_{j\to\infty}a_{k_j}=b$, then b is an accumulation point of S.

Exercise 1.6.32

i. Find an infinite subset of $\mathbb{R}$ which does not have an accumulation point.

ii. Find a bounded subset of $\mathbb{R}$ which does not have an accumulation point.

Definition 1.6.33 Let S be a subset of $\mathbb{R}$. We say that S is an *open set* in $\mathbb{R}$ if, for each point $x\in S$, there is an $\varepsilon>0$ (depending on x) such that $(x-\varepsilon,x+\varepsilon)\subset S$.

Definition 1.6.34 Let $S\subset\mathbb{R}$. We say S is a *closed set* in $\mathbb{R}$ if the complement of S is an open set in $\mathbb{R}$.

Note that the empty set and $\mathbb{R}$ are both open and closed subsets of $\mathbb{R}$.

Exercise 1.6.35

i. Show that $\varnothing$ and $\mathbb{R}$ are the only subsets of $\mathbb{R}$ that are both open and closed in $\mathbb{R}$.

ii. Show that every nonempty open set in $\mathbb{R}$ can be written as a countable union of pairwise disjoint open intervals.

iii. Show that an arbitrary union of open sets in $\mathbb{R}$ is open in $\mathbb{R}$.

iv. Show that a finite intersection of open sets in $\mathbb{R}$ is open in $\mathbb{R}$.

v. Show, by example, that an infinite intersection of open sets is not necessarily open.

vi. Show that an arbitrary intersection of closed sets in $\mathbb{R}$ is a closed set in $\mathbb{R}$.

vii. Show that a finite union of closed sets in $\mathbb{R}$ is a closed set in $\mathbb{R}$.

viii. Show, by example, that an infinite union of closed sets in $\mathbb{R}$ is not necessarily a closed set in $\mathbb{R}$.

Exercise 1.6.36 Show that a subset of $\mathbb{R}$ is closed iff it contains all its accumulation points.

Exercise 1.6.37 We define the Cantor set to be a subset of the closed interval $[0,1]$. First, remove the open interval $(1/3, 2/3)$ from $[0,1]$. Next, remove the open intervals $(1/9, 2/9)$ and $(7/9, 8/9)$. At each step, remove middle third of the remaining closed intervals. Repeating this process a countable number of times, we are left with a subset of the closed interval $[0,1]$ called the *Cantor set.* Show that:

i. the Cantor set is closed;

ii. the Cantor set consist of all numbers in the closed $[0,1]$ whose ternary expansion consists of only 0s and 2s and may end in infinitely many 2s;

iii. the Cantor set is uncountable;

iv. every point of the Cantor set is an accumulation point of the Cantor set;

v. the complement of the Cantor set in $[0,1]$ is a dense subset of $[0,1]$.

The next theorem, the Heine-Borel theorem for $\mathbb{R}$, is the second of the two basic topological theorems for the real numbers, the first of which is the Bolzano-Weierstrass theorem. We shall see versions of these theorems again in Chapter 3.

Theorem 1.6.38 (Heine-Borel) Let S be a closed and bounded subset of $\mathbb{R}$. Given a collection $\{U_i\}_{i \in I}$ of open sets such that $S \subset \bigcup_{i \in I} U_i$, there exists a finite subcollection $U_1, \ldots, U_n$ of $\{U_i\}_{i \in I}$ such that $S \subset U_1 \cup \ldots \cup U_n$.

Proof. Suppose that S is a nonempty, closed, bounded subset of $\mathbb{R}$. If $a = \text{glb}(S)$ and $b = \text{lub}(S)$, then, since S is closed, a and b are in S, and $S \subset [a,b]$. Let $\{U_i\}_{i \in I}$ be a collection of open sets such that $S \subset \bigcup_{i \in I} U_i$. By adjoining the complement of S (if necessary), we obtain a collection $\mathcal{U}$ of open sets whose union contains $[a,b]$.

Now let $B = \{x \in [a,b] \mid [a,x] \text{ is covered by a finite number of open sets in } \mathcal{U}\ \}$. Then B is nonempty since $a \in B$, and B is bounded above by b. Let $c = \text{lub}(B)$.

Exercise 1.6.39 Prove that $c \in B$. (Hint: Prove that B is closed.)

Suppose $c < b$. Let $\mathcal{U}'$ be a finite subcollection of $\mathcal{U}$ that covers $[a,c]$. Because any element of $\mathcal{U}'$ that contains c is open, there exists y such that $c < y < b$ and $[c,y]$ is in the same open set that contains c. Thus $[a,y]$ is covered by $\mathcal{U}'$. This is a contradiction, and hence b must equal c. Thus $[a,b]$ is covered by a finite number of open sets from $\mathcal{U}$, and by throwing away the complement of S (if necessary), S is covered by a finite number of open sets from the original collection.

The ideas contained in the Heine-Borel theorem are important enough to deserve their own definitions. We introduce them here in the context of $\mathbb{R}$, but we will see them again in Chapter 3 when we talk about metric spaces.

Definition 1.6.40 Let A be a subset of $\mathbb{R}$. An *open covering of* A is a collection of open sets $\{U_i\}_{i\in I}$ such that $A \subset \bigcup_{i\in I} U_i$.

Definition 1.6.41 Let A be a subset of $\mathbb{R}$. We say that A is a *compact set* if every open covering of A has a finite subcovering. That is, if $\{U_i\}_{i\in I}$ is an open covering of A, there is a finite subcollection $U_1, U_2, \ldots, U_n$ of the collection $\{U_i\}_{i\in I}$ such that $A \subset U_1 \cup U_2 \cup \cdots \cup U_n$

This notion of compactness is more subtle than it appears. It does not say that in order for a set to be compact, it must have a finite open covering. Note that this is true of any subset of $\mathbb{R}$, since we can take the single set $U_1 = \mathbb{R}$ to be a covering. The stress in the definition is that *every* open covering must have a finite subcovering. This is a big idea in analysis. It allows us to reduce certain arguments about infinite sets to arguments about finite sets.

We can rephrase the Heine-Borel theorem to say that closed and bounded subsets of $\mathbb{R}$ are compact. In fact, the converse of this statement is also true, namely, that all compact subsets of $\mathbb{R}$ are closed and bounded.

Exercise 1.6.42 Show that a compact subset of $\mathbb{R}$ is both closed and bounded.

Exercise 1.6.43 Let $S = (a, b)$. Give an example of an open covering of S that does not have a finite subcovering.

Exercise 1.6.44 Let $S = \mathbb{Z}$. Give an example of an open covering of S that does not have a finite subcovering.

There is another closely related notion regarding subsets of $\mathbb{R}$.

Definition 1.6.45 A subset A of $\mathbb{R}$ is *sequentially compact* if every infinite sequence in A has a subsequence that converges to an element of A.

The following theorem can be proved easily using the Bolzano-Weierstrass and Heine-Borel theorems in $\mathbb{R}$.

Exercise 1.6.46 A subset of $\mathbb{R}$ is compact if and only if it is sequentially compact.

We will see in Chapter 3 that the same theorem is true in metric spaces. In the next section, we give an indication of how this works in $\mathbb{C}$.

1.7 The complex numbers $\mathbb{C}$

To start this section, we give a somewhat inexact definition of complex numbers. This notion is often used as the definition of the complex numbers, but it does contain some ambiguity which we will rectify immediately.

Definition 1.7.1 (Rural Definition) The set of *complex numbers*, $\mathbb{C}$, is the collection of expressions of the form $z = a + bi$ where $a, b \in \mathbb{R}$, and i is a symbol which satisfies $i^2 = -1$. If $z = a + bi$ and $w = c + di$ are in $\mathbb{C}$, then we define $z + w = (a + c) + (b + d)i$, and $zw = (ac - bd) + (bc + ad)i$.

Actually, one can go a long way with this definition if the symbol i with the property that $i^2 = -1$ doesn't cause insomnia. In fact, though, once you assert that $i^2 = -1$, you must accept the fact that $(-i)^2 = -1$, and hence there is some ambiguity in the choice of which of i and $-i$ is "the" square root of -1. This difficulty is avoided in the following construction.

We consider the Cartesian product $\mathbb{R} \times \mathbb{R}$ with addition and multiplication defined by

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d),\\ (a, b)(c, d) &= (ac - bd, bc + ad).\end{aligned}$$

Exercise 1.7.2 Show that $\mathbb{R} \times \mathbb{R}$ with addition and multiplication as defined above is a field, with $(0,0)$ as the additive identity, $(1,0)$ as the multiplicative identity, $-(a,b) = (-a,-b)$, and

$$(a,b)^{-1} = (a/(a^2+b^2), -b/(a^2+b^2)) \quad \text{if } (a,b) \neq (0,0).$$

Definition 1.7.3 *The field of complex numbers* is the set $\mathbb{C} = \mathbb{R} \times \mathbb{R}$ with the operations of addition and multiplication defined above.

Note that $\mathbb{R}$ is isomorphic to the subfield of $\mathbb{C}$ given by $\{(a,0) \mid a \in \mathbb{R}\}$. If we set $i = (0,1)$, then $i^2 = (-1,0)$. Finally, to fix things up real nice, we write $(a,b) = (a,0) + (b,0)(0,1)$, or, returning to our original rural definition, $(a,b) = a + bi$.

The first observation to make is that $\mathbb{C}$ cannot be made into an ordered field. That is, it cannot satisfy the order axioms given in Section A.5. This is immediate because in any ordered field, if $a \neq 0$ then $a^2 > 0$. This would imply that $i^2 = -1 > 0$, but $1^2 = 1 > 0$ and this is a contradiction.

Exercise 1.7.4 Show that the field of complex numbers is not isomorphic to the field of real numbers.

Definition 1.7.5 If $z = a + bi$ with $a, b \in \mathbb{R}$, we call a the *real part* of z and b the *imaginary part* of z. We write $a = \text{Re } z$ and $b = \text{Im } z$. The complex number z is called *pure imaginary* if $a = \text{Re } z = 0$.

Definition 1.7.6 If $z = a + bi$ with $a, b \in \mathbb{R}$, the *complex conjugate* of z, denoted $\overline{z}$, is the complex number $\overline{z} = a - bi$.

Exercise 1.7.7 Prove the following statements about complex conjugates:

i. $z \in \mathbb{R}$ iff $z = \overline{z}$;

ii. $\overline{z+w} = \overline{z} + \overline{w}$;

iii. $\overline{zw} = \overline{zw}$;

iv. $z\overline{z} \in \mathbb{R}$.

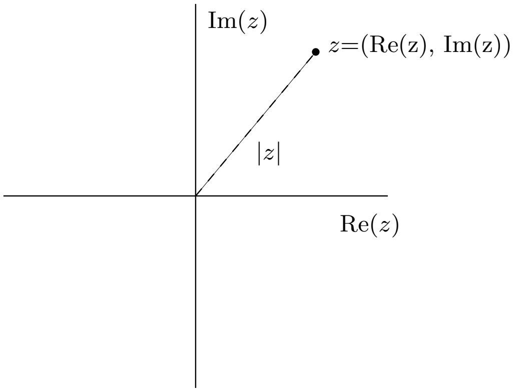
Definition 1.7.8 If $z = a + bi$ with $a, b \in \mathbb{R}$, the *absolute value* of z is

$$|z| = (z\overline{z})^{\frac{1}{2}} = (a^2 + b^2)^{\frac{1}{2}},$$

where, of course, we mean the nonnegative square root in $\mathbb{R}$.

If z and w are complex numbers, then $|z|$ and $|w|$ are real numbers, and hence it makes sense to say that $|z| < |w|$. However, it makes no sense to say that $z < w$ since $\mathbb{C}$ is not an ordered field.

Exercise 1.7.9 Show that if we identify $z = a + bi$ with the point $(a,b) \in \mathbb{R}^2$, then the absolute value of z is equal to the distance to the point (a,b) from $(0,0)$.



Exercise 1.7.10 Show that the absolute value on $\mathbb{C}$ satisfies all the properties of the absolute value on $\mathbb{R}$.

1. For any $z \in \mathbb{C}$, we have $|z| \geq 0$, and $|z| = 0$ iff $z = 0$.
2. For any $z, w \in \mathbb{C}$, we have $|zw| = |z||w|$.
3. For any $z, w \in \mathbb{C}$, we have $|z + w| \leq |z| + |w|$ (triangle inequality).

Exercise 1.7.11 Why is the triangle inequality so named?

Definition 1.7.12 If $z = x + iy \in \mathbb{C}$, $z \neq 0$, and $r = |z|$, then the *polar form* of z is $z = r(\cos\theta + i\sin\theta)$ where θ is the unique solution to the equations

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta, \end{aligned}$$

in the interval $[0, 2\pi)$. The angle θ is called *the principal branch of the argument of* z and is denoted $\mathrm{Arg}(z)$. For z as above, we often write $z = re^{i\theta}$, where $e^{i\theta}$ is defined to be $\cos\theta + i\sin\theta$. (In fact, $\cos\theta + i\sin\theta$ is the value of the complex exponential function $f(z) = e^z$, defined by the power series $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$, when $z = i\theta$.)

Exercise 1.7.13 Show that there is a unique θ with $0 \leq \theta < 2\pi$ simultaneously satisfying $\cos\theta = \left(\frac{a}{a^2+b^2}\right)^{\frac{1}{2}}$, $\sin\theta = \left(\frac{b}{a^2+b^2}\right)^{\frac{1}{2}}$ for a pair of real numbers a and b not both zero.

Exercise 1.7.14 Prove by induction that $(e^{i\theta})^n = e^{in\theta}$ for $n \in \mathbb{N}$.

Exercise 1.7.15 Suppose that $n \in \mathbb{N}$. Prove that, if $z = e^{\frac{2k\pi i}{n}}$, for $k \in \mathbb{Z}$ and $0 \leq k \leq n-1$, then $z^n = 1$. Such a z is called an *n-th root of unity.* Note that these n roots of unity are all distinct.

Remark 1.7.16 The n-th roots of unity form a cyclic group of order n under multiplication. An n-th root of unity is *primitive* if it is a generator of this group. In fact, the primitive n-th roots of unity are those of the form $e^{2\pi ik/n}$ where k and n are relatively prime. (See Project 2.6.1.)

Proposition 1.7.17 If $n > 1$, the sum of the n distinct n-th roots of unity is 0.

Proof. For any $z \in \mathbb{C}$,

$$(1 - z^n) = (1 - z)(1 + z + z^2 + \cdots + z^{n-1}).$$

Now let $z = e^{\frac{2\pi i}{n}}$.

Exercise 1.7.18 Suppose z is a nonzero complex number, and write $z = re^{i\theta}$. Show that z has exactly n distinct complex n-th roots given by $r^{1/n}e^{i(2\pi k+\theta)/n}$ for $0 \leq k \leq n-1$.

1.8 Convergence in $\mathbb{C}$

Now that we have an absolute value on $\mathbb{C}$, we can define the notion of Cauchy sequence and convergent sequence in $\mathbb{C}$.

Definition 1.8.1 A sequence $(z_k)_{k\in\mathbb{N}}$ of complex numbers is *convergent* if there exists an element $z \in \mathbb{C}$ such that the sequence satisfies the following property: given any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $k \geq N_\varepsilon$ implies that $|z_k - z| < \varepsilon$. We say that $(z_k)_{k\in\mathbb{N}}$ *converges to* z, and z is called the *limit* of the sequence $(z_k)_{k\in\mathbb{N}}$. Symbolically, we write

$$\lim_{k\to\infty} z_k = z.$$

We will often say that a sequence of complex numbers is convergent without specific reference to the limit z. Note that in the definition, ε is a positive *real* number (this is already implied by the use of the inequality symbol, but we repeat it here for emphasis). Note too that N_ε depends on ε. As usual, the limit of a convergent sequence is unique.

Definition 1.8.2 Let r be a positive real number, and let $z_0 \in \mathbb{C}$. The *open ball* of radius r with center at z_0 is the set

$$B_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| < r\}. \tag{1.1}$$

The *closed ball* of radius r with center z_0 is the set

$$\overline{B}_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| \leq r\}. \tag{1.2}$$

The open balls and closed balls in $\mathbb{C}$ are the analogs of open and closed intervals in $\mathbb{R}$. We can define open and closed sets in $\mathbb{C}$ in a fashion similar to the definitions in $\mathbb{R}$.

Definition 1.8.3 Let S be a subset of $\mathbb{C}$. We say that S is an *open set* in $\mathbb{C}$ if, for each point $z \in S$, there is an $\varepsilon > 0$ (depending on z) such that $B_\varepsilon(z) \subset S$.

Definition 1.8.4 Let $S \subset \mathbb{C}$. We say that S is a *closed set* in $\mathbb{C}$ if the complement of S is an open set in $\mathbb{C}$.

Exercise 1.8.5

i. Show that the empty set and $\mathbb{C}$ are both open and closed subsets of $\mathbb{C}$.

ii. Show that no other subsets of $\mathbb{C}$ besides $\varnothing$ and $\mathbb{C}$ are both open and closed in $\mathbb{C}$.

iii. Show that an arbitrary union of open sets in $\mathbb{C}$ is an open set in $\mathbb{C}$.

iv. Show that a finite intersection of open sets in $\mathbb{C}$ is an open set in $\mathbb{C}$.

v. Show, by example, that an infinite intersection of open sets in $\mathbb{C}$ need not be an open set in $\mathbb{C}$.

vi. Show that an arbitrary intersection of closed sets in $\mathbb{C}$ is a closed set in $\mathbb{C}$.

vii. Show that a finite union of closed sets in $\mathbb{C}$ is a closed set in $\mathbb{C}$.

viii. Show, by example, that an infinite union of closed sets in $\mathbb{C}$ is not necessarily a closed set in $\mathbb{C}$.

Exercise 1.8.6

i. Let (a, b) and (c, d) be open intervals in $\mathbb{R}$. Show that the *open rectangle* $(a, b) \times (c, d) = \{x + iy \in \mathbb{C} \mid x \in (a, b), y \in (c, d)\}$ is an open set in $\mathbb{C}$.

ii. Let $[a, b]$ and $[c, d]$ be closed intervals in $\mathbb{R}$. Show that the *closed rectangle* $[a, b] \times [c, d] = \{x + iy \in \mathbb{C} \mid x \in [a, b], y \in [c, d]\}$ is a closed set in $\mathbb{C}$.

iii. Let $z \in \mathbb{C}$, and let S be an open set containing z. Show that there exists an open rectangle $R = (a, b) \times (c, d)$ such that $z \in R$ and $R \subset S$.

Exercise 1.8.7 Consider the collection of open balls $\{B_r(z)\}$ in $\mathbb{C}$ where $r \in \mathbb{Q}$ and $\mathrm{Re}(z), \mathrm{Im}(z) \in \mathbb{Q}$. Show that any open set in $\mathbb{C}$ can be written as a finite or countable union from this collection of sets.

Exercise 1.8.8

Show, by example, that there are open sets in $\mathbb{C}$ that cannot be written as the countable union of pairwise disjoint open balls.

Definition 1.8.9 Let $A \subset \mathbb{C}$. The set A is *bounded* if there exists $r > 0$ such that $A \subset B_r(0)$.

Exercise 1.8.10 Define the notion of a bounded sequence in $\mathbb{C}$.

Definition 1.8.11 (See Definition 1.6.22) A sequence $(z_k)_{k\in\mathbb{N}}$ in $\mathbb{C}$ is a *Cauchy sequence* if, given any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $n, m \geq N$ implies $|z_m - z_n| < \varepsilon$.

Exercise 1.8.12 Prove that every Cauchy sequence in $\mathbb{C}$ is bounded.

Theorem 1.8.13 (Cauchy Criterion) A sequence $(z_k)_{k\in\mathbb{N}}$ of complex numbers is convergent if and only if it is a Cauchy sequence.

Proof. The first half of the proof is identical to the proof of Theorem 1.6.24.

Suppose now that $(z_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}$. Let $z_k = a_k + b_k i$, where $a_k, b_k \in \mathbb{R}$. Then $|z_m - z_n|^2 = (a_m - a_n)^2 + (b_m - b_n)^2$. Hence, $|a_m - a_n| \leq |z_m - z_n|$, and since $(z_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, it follows that $(a_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. Similarly, $(b_k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. If $\lim_{k\to\infty} a_k = a$ and $\lim_{k\to\infty} b_k = b$, then $\lim_{k\to\infty} z_k = z$ where $z = a + bi$.

Exercise 1.8.14 Show that every bounded sequence in $\mathbb{C}$ has a convergent subsequence.

Definition 1.8.15 Let S be a subset of $\mathbb{C}$. Then z is an *accumulation point* of S if, for all $\varepsilon > 0$ we have $(B_\varepsilon(z) \setminus \{z\}) \cap S \neq \varnothing$.

Remark 1.8.16 Thus, z is an accumulation point of S if every open ball around z contains points of S other than z. Of course, z does not have to be an element of S in order to be an accumulation point of S.

Exercise 1.8.17 Find the accumulation points of the following sets:

i. $S = \{z \in \mathbb{C} \mid |z| = 1\}$ (this is the *unit circle* in $\mathbb{C}$);

ii. $S = \{z \in \mathbb{C} \mid \text{Re } z > \text{Im } z\}$;

iii. $S = \{a + bi \mid a, b \in \mathbb{Q}\}$;

iv. $S = \{a + bi \mid a, b \in \mathbb{Z}\}$;

v. $S = \{\frac{1}{n} + \frac{1}{m}i \mid n, m \in \mathbb{N}\}$.

Exercise 1.8.18

i. Let S be a subset of $\mathbb{C}$. Show that every open set containing an accumulation of S contains infinitely many points of S.

ii. (Bolzano–Weierstrass Theorem for $\mathbb{C}$) Prove that any bounded infinite set in $\mathbb{C}$ has an accumulation point in $\mathbb{C}$.

Definition 1.8.19 Let S be a subset of $\mathbb{C}$. An *open covering of* S is a collection of open sets $\{U_i\}_{i\in I}$ such that $S \subset \bigcup_{i\in I} U_i$.

Definition 1.8.20 Let S be a subset of $\mathbb{C}$. We say that S is *compact* if every open covering of S has a finite subcovering. That is, if $\{U_i\}_{i\in I}$ is an open covering of S, there is a finite subcollection $U_1, U_2, \ldots, U_n$ of the collection $\{U_i\}_{i\in I}$ such that $S \subseteq U_1 \cup U_2 \cup \cdots \cup U_n$.

Theorem 1.8.21 (Heine-Borel) If S is a closed and bounded subset of $\mathbb{C}$, then S is compact.

Proof. For the purposes of this proof, we recall that $\mathbb{C} = \{(x, y) \mid x, y \in \mathbb{R}\}$ as in Definition 1.7.3. We prove it for $S = [a, b] \times [c, d]$, where $a, b, c, d \in \mathbb{R}$ and $a < b$ and $c < d$, and leave the general case as an exercise.

Take a point $x_0 \in [a, b]$ and consider the set $\{x_0\} \times [c, d]$. We take an open set $N \subseteq \mathbb{C}$ containing $\{x_0\} \times [c, d]$. We claim that there exists an open interval I containing x_0 such that $I \times [c, d] \subseteq N$. We see this as follows. For each point $(x_0, y) \in \{x_0\} \times [c, d]$, choose $r_y > 0$ such that the open square $(x_0 - r_y, x_0 + r_y) \times (y - r_y, y + r_y) \subseteq N$. (See Exercise 1.8.6.*iii.*) By intersecting these squares with $\{x_0\} \times \mathbb{R}$, we get a collection of open intervals of the form $(y - r_y, y + r_y)$ that cover $[c, d]$. By the Heine-Borel theorem in $\mathbb{R}$, there exists a finite subcollection of these open intervals that covers the interval $[c, d]$. Hence the corresponding collection of open squares also covers $\{x_0\} \times [c, d]$. Let r be the minimum of the r_y from this finite collection. Then $I = (x_0 - r, x_0 + r)$ is our desired interval.

Now let $\{U_j\}_{j \in J}$ be an open covering of S. For each $x \in [a, b]$, the collection $\{U_j\}_{j \in J}$ covers $\{x\} \times [c, d]$. As we did above, we choose a finite subcollection $U_1, \ldots, U_n$ that covers $\{x\} \times [c, d]$. The open set $N_x = U_1 \cup \cdots \cup U_n$ contains a set of the form $I_x \times [c, d]$ by the preceding discussion, where I_x is an open interval containing x. The collection $\{I_x\}_{x \in [a,b]}$ covers $[a, b]$, and hence by the Heine-Borel theorem for $\mathbb{R}$, there exists a finite subcollection $I_{x_1}, \ldots, I_{x_m}$ that covers $[a, b]$. We take our finite subcollection of the original open cover $\{U_j\}_{j \in J}$ to be $\{U \mid \text{for some } x_i\text{, the set } U \text{ is one of the elements in the union that defines } N_{x_i}\}$.

Exercise 1.8.22 Prove the general case of the Heine-Borel theorem in $\mathbb{C}$. (Hint: Take a closed bounded set in $\mathbb{C}$ and put it inside the product of two closed bounded intervals. Then use the result from the proof above.)

Exercise 1.8.23 Show that a subset of $\mathbb{C}$ is closed iff it contains all its accumulation points.

Exercise 1.8.24 Define the notion of sequentially compact for a subset of $\mathbb{C}$, and show that a subset of $\mathbb{C}$ is sequentially compact if and only if it is closed and bounded. (See Definition 1.6.45.)

1.9 Infinite Series

We assume that the reader has had at least an elementary introduction to infinite series and their convergence properties. In fact, the theory of infinite series actually reduces to the convergence of sequences, which we have covered thoroughly in this chapter. An infinite series is expressed as a sum of an infinite number of elements from some place where addition makes sense. These elements could be numbers, functions or what have you, so we begin with one-sided series of numbers.

We take an infinite series to be an expression of the form $\sum_{n=1}^{\infty} a_n$, where the elements a_n come from a number system in which addition makes sense. So that we don't wander around aimlessly, let's fix our number system to be the complex numbers, that is $a_n \in \mathbb{C}$, with the possibility of restricting to the real numbers or even the rational numbers. In the definition, we have chosen to use the natural numbers as the index set, but in considering infinite series we could start the summation with any integer n_0 and write $\sum_{n=n_0}^{\infty} a_n$. Later, we will also consider two-sided series where the index set is the entire set of integers and we write $\sum_{-\infty}^{\infty} a_n$. If these expressions are going to have any meaning at all, we must look at the partial sums.

Definition 1.9.1 If $\sum_{n=1}^{\infty} a_n$ is an infinite series of complex numbers, the *N-th partial sum* of the series is $S_N = \sum_{n=1}^{N} a_n$.

Examples 1.9.2

i. Let $a_n = 1$ for all n. Then $S_N = N$.

ii. Let $a_n = 1/n$. Then $S_N = 1 + 1/2 + \cdots + 1/N$.

iii. Let $a_n = 1/2^n$. Then $S_N = 1 - 1/2^N$.

iv. Let $a_n = (-1)^{n+1}$. In this case, $S_N = 1$ if N is odd and 0 if N is even.

v. Fix θ, with $0 < \theta < 2\pi$, and let $a_n = e^{in\theta}/n$. Then $S_N = \sum_{n=1}^{N} e^{in\theta}/n$, which is the best we can do without more information about θ.

vi. Let $a_n = \sin n\pi/n^2$. In this case, $S_N = \sum_{n=1}^{N} \sin(n\pi)/n^2$.

Definition 1.9.3 Let $\sum_{n=1}^{\infty} a_n$ be an infinite series of complex numbers. If $N \in \mathbb{N}$, we let $S_N = \sum_{n=1}^{N} a_n$. The sequence $(S_N)_{N\in\mathbb{N}}$ is called *the sequence of partial sums.* We say that the series $\sum_{n=1}^{\infty} a_n$ *converges* if the sequence of partial sums $(S_N)_{N\in\mathbb{N}}$ converges, and we call the *sum* of a convergent series $\sum_{n=1}^{\infty} a_n$ the number S to which the sequence $(S_N)_{N\in\mathbb{N}}$ converges. If the sequence $(S_N)_{N\in\mathbb{N}}$ does not converge we say that $\sum_{n=1}^{\infty} a_n$ *diverges.*

Of course, since we are working in $\mathbb{C}$, the series converges if and only if the sequence $(S_N)_{N\in\mathbb{N}}$ is a Cauchy sequence. That is, given $\varepsilon > 0$, there is a $N_\varepsilon \in \mathbb{N}$ such that for $n, m > N_\varepsilon$ (assuming $n > m$), then $|\sum_{k=m+1}^{n} a_n| < \varepsilon$.

Exercise 1.9.4 Determine which of the series in Example 1.9.2 converge.

We are faced with two problems. The first is, “How do we tell if a series converges?” The second is, “If a series does converge, how do we find the explicit sum?” There is extensive literature about these two questions, but the fact is that the second question presents many more difficulties than the first.

The most helpful series in all of this discussion is a geometric series.

Definition 1.9.5 Let z be a complex number. The *geometric series* defined by z is $\sum_{n=0}^{\infty} z^n$.

Exercise 1.9.6

i. If $N \in \mathbb{N}$ and $z \neq 1$, show that $S_N = \sum_{n=0}^{N} z^n = \frac{1-z^{N+1}}{1-z}$.

ii. If $|z| < 1$, show that $\lim_{n\to\infty} z^n = 0$.

iii. If $|z| > 1$, show that $\lim_{n\to\infty} z^n$ does not exist.

Theorem 1.9.7 Consider the geometric series defined by a complex number z. If $|z| < 1$, then the series converges. If $|z| > 1$, then the series diverges.

Exercise 1.9.8

i. Prove the theorem using the exercise above.

ii. What can you say about a geometric series for which $|z| = 1$?

Theorem 1.9.9 Suppose that a series $\sum_{n=1}^{\infty} a_n$ converges. Show that $\lim_{n\to\infty} a_n = 0$.

Exercise 1.9.10 Prove this.

The more useful phrasing of this theorem is often the contrapositive; namely, if the terms of a series do not go to zero, then the series must diverge.

Note, however, that the property that $\lim_{n\to\infty} a_n = 0$ does not ensure that the series $\sum_{n=1}^{\infty} a_n$ converges. The most useful example is given above where $a_n = 1/n$. In this case, $S_1 = 1$, $S_4 > 2$, it is easy to check that $S_{2^n} > n$ for $n \in \mathbb{N}$, and hence the series $\sum_{n=1}^{\infty} 1/n$ diverges.

Exercise 1.9.11 The series $S = \sum_{n=1}^{\infty} 1/n$ is often called the *harmonic series.* We have just proved that this series diverges. Show that, by suitably eliminating an infinite number of terms, the remaining sub-series can be made to converge to any positive real number.

Definition 1.9.12 A series $\sum_{n=1}^{\infty} a_n$ of complex numbers *converges absolutely* if the series $\sum_{n=1}^{\infty} |a_n|$ converges.

Proposition 1.9.13 If $\sum_{n=1}^{\infty} a_n$ converges absolutely, then $\sum_{n=1}^{\infty} a_n$ converges.

Proof. This follows from the fact that $|\sum_{k=m+1}^{n} a_k| \leq \sum_{k=m+1}^{n} |a_k|$, by the Triangle Inequality.

The converse to Proposition 1.9.13 is false, and is shown by the example $\sum_{n=1}^{\infty}(-1)^{n+1}/n$. This series converges since $|\sum_{k=m+1}^{n}(-1)^{k+1}/k| < 1/m$. However as we have seen above the series does not converge absolutely.

There are various tests to determine if a series converges. These include the comparison test, the ratio test, and the root test.

The comparison test is often very useful, but its use depends on knowing ahead of time a series which converges.

Theorem 1.9.14 (Comparison Test) Suppose that $a_n > 0$ for every $n \in \mathbb{N}$ and that $\sum_{n=1}^{\infty} a_n$ converges. If $b_n \in \mathbb{C}$ satisfies $|b_n| \leq a_n$ for all n, then the series $\sum_{n=1}^{\infty} b_n$ converges absolutely and hence converges.

Proof. For each $N \in \mathbb{N}$, let $S_N = \sum_{n=1}^{N} a_n$, and let $S = \lim_{N\to\infty} S_N$. Let $T_N = \sum_{n=1}^{N}|b_n|$. Then for every $N \in \mathbb{N}$, $T_N \leq S_N$, and hence $T_N \leq S$. Thus, T_N is a monotonic bounded sequence of real numbers, which must converge.

Exercise 1.9.15

i. If the series $\sum_{n=1}^{\infty} a_n$ converges to s and c is any constant, show that the series $\sum_{n=1}^{\infty} ca_n$ converges to cs.

ii. Suppose that $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ are infinite series. Suppose that $a_n > 0$ and $b_n > 0$ for all $n \in \mathbb{N}$ and that $\lim_{n\to\infty} a_n/b_n = c > 0$. Show that $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} b_n$ converges.

Exercise 1.9.16

i. If $p \in \mathbb{R}$ and $p < 1$, show that $\sum_{n=1}^{\infty} 1/n^p$ diverges.

ii. If $p \in \mathbb{R}$ and $p > 1$, show that $\sum_{n=1}^{\infty} 1/n^p$ converges. Hint: Use the fact from elementary calculus that

$$\sum_{n=2}^{N} \frac{1}{n^p} \leq \int_1^N \frac{1}{x^p}\, dx = \frac{1}{p-1}\left(1 - \frac{1}{N^{p-1}}\right).$$

The most useful series for comparison is the geometric series defined by a real number r, with $0 < r < 1$.

Theorem 1.9.17 (Ratio Test) Suppose that $\sum_{n=1}^{\infty} a_n$ is a series of non-zero complex numbers. If $r = \lim_{n\to\infty} |a_{n+1}/a_n|$ exists, then the series converges absolutely if $r < 1$, and the series diverges if $r > 1$.

Proof. Suppose $\lim_{n\to\infty} |a_{n+1}/a_n| = r < 1$. If ρ satisfies $r < \rho < 1$, then there exists $N \in \mathbb{N}$ such that $|a_{n+1}|/|a_n| < \rho$ for all $n \geq N$. Consequently, $|a_n| \leq |a_N|\rho^{n-N}$ for all $n \geq N$. The result follows from the Comparison Test.

The second half of the theorem follows from Theorem 1.9.9.

Exercise 1.9.18 Give examples to show that if $r = 1$ in the statement of the Ratio Test, anything may happen.

Our final test for convergence is called the root test. This can be quite effective when the comparison test and ratio test fail.

Theorem 1.9.19 (Root Test) Suppose that $\sum_{n=1}^{\infty} a_n$ is a series of complex numbers. Let $r = \limsup_{n\to\infty} |a_n|^{1/n}$. If $r < 1$, then the series converges absolutely. If $r > 1$, then the series diverges.

Proof. Suppose that $\limsup_{n\to\infty} |a_n|^{1/n} = r < 1$. Pick ρ such that that $r < \rho < 1$. Then, there exists $N \in \mathbb{N}$ such that $|a_n| \le \rho^n$ for all $n \ge N$. The convergence of the series now follows from the comparison test. The second half of the theorem is left as an exercise.

Exercise 1.9.20 Give examples to show that if $r = 1$ in the statement of the Root Test, anything may happen.

Exercise 1.9.21 Suppose that the ratio test applies to a series. That is, $\lim_{n\to\infty} |a_{n+1}|/|a_n| = r$. Show that $\limsup_{n\to\infty} |a_n|^{1/n} = r$.

Definition 1.9.22 Let z_0 be a fixed complex number. A *complex power series* around z_0 is a series of the form $\sum_{n=0}^{\infty} a_n(z - z_0)^n$, where the coefficients a_n are in $\mathbb{C}$ for all $n \in \mathbb{N}$. When this series converges, it converges to a function of the complex variable z.

Exercise 1.9.23 Show that if the series converges absolutely for a complex number z then it also converges for a any complex number w such that $|w - z_0| \le |z - z_0|$, that is the series converges on the disk $\{w \in \mathbb{C} \mid |w - z_0| \le |z - z_0|\}$.

From this exercise, it follows that a complex power series around z_0 that converges absolutely at any point other then z_0 will have a disk of convergence of the form $\{z \in \mathbb{C} \mid |z - z_0| < r\}$. The supremum of all such r is called the *radius of convergence* of the power series.

To determine the radius of convergence for a complex power series we use the convergence tests developed above, in particular the root test.

Theorem 1.9.24 Suppose that $\limsup_{n\to\infty} |a_n|^{1/n} = r$. If $r > 0$, then the power series $\sum_{n=0}^{\infty} a_n(z - z_0)^n$ has a radius of convergence $1/r$. If the number $r = 0$, the we say that the radius of convergence is infinity, and if the lim sup does not exist because $|a_n|^{\frac{1}{n}}$ is unbounded ($r = \infty$), we say that the radius of convergence is 0.

Examples 1.9.25

i. Consider the series $\sum_{n=0}^{\infty} n(z-z_0)^n$. Then $\lim_{n\to\infty} n^{1/n} = 1$, and the power series converges absolutely for $|z - z_0| < 1$, that is, the radius of convergence is 1.

ii. Consider the series $\sum_{n=1}^{\infty} n^n(z - z_0)^n$. Then $\lim_{n\to\infty}(n^n)^{1/n} = \infty$, so the radius of convergence is 0 and the series converges only for $z = z_0$.

Exercise 1.9.26 Determine the radius of convergence of the following power series:

i.
$$\sum_{n=1}^{\infty} \frac{z^n}{n!};$$

ii.
$$\sum_{n=2}^{\infty} \frac{z^n}{\ln(n)};$$

iii.
$$\sum_{n=1}^{\infty} \frac{n^n}{n!} z^n.$$

1.10 Algebraic notions in $\mathbb{R}$ and $\mathbb{C}$

In this chapter, we have constructed the real numbers and showed that they are the unique ordered field, up to order isomorphism, that satisfies the least upper bound property. We have also constructed the complex numbers, a field containing a subfield that is isomorphic to the real numbers. Moreover, we have seen that the complex numbers cannot be ordered. While the question of when two fields are isomorphic is fundamental, we now pursue the more refined question of studying the ways in which a field is isomorphic to itself.

Let us start simply. Consider the field of rational numbers and a function $f : \mathbb{Q} \to \mathbb{Q}$ such that $f(x+y) = f(x) + f(y)$. What can we say about f? We have $f(x+0) = f(x) + f(0) = f(x)$, so $f(0) = 0$. Next we have, for $n \in \mathbb{N}$,

$$f(n) = f(\underbrace{1+1+\cdots+1}_{n \text{ times}}) = \underbrace{f(1)+f(1)+\cdots+f(1)}_{n \text{ times}} = nf(1).$$

A similar argument shows that $f(m/n) = (m/n)f(1)$ for any positive rational number m/n. Also, for any positive rational number r, we have $0 = f(0) = f(r + (-r)) = f(r) + f(-r)$, so that $f(-r) = -f(r)$. Thus, $f(r) = rf(1)$ for all $r \in \mathbb{Q}$. Now suppose that we also want $f(xy) = f(x)f(y)$ for $x, y \in \mathbb{Q}$, that is, we want f to be an isomorphism of fields. Then, $f(1) = f(1 \cdot 1) = f(1)f(1)$, and if f is to be injective, we must have $f(1) = 1$. Thus, the function f can be none other than the identity function on $\mathbb{Q}$ (see Definition A.7.16).

What we have just done in the preceding paragraph is to show that the only field isomorphism from $\mathbb{Q}$ to itself is the identity. In general, an isomorphism from a field to itself is called an automorphism, a notion which we make precise with the following definition.

Definition 1.10.1 Let F be a field. An *automorphism of* F is a bijection, $\phi : F \to F$, such that

a. $\phi(x+y) = \phi(x) + \phi(y)$ for all $x, y \in F$,

b. $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in F$.

We denote the set of automorphisms of a field F by $\mathrm{Aut}(F)$.

Exercise 1.10.2 Show that the set $\mathrm{Aut}(F)$ of automorphisms of a field F has the structure of a group under function composition (see Project 2.1).

Let's see what happens for $\mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ be an automorphism. By the exact same reasoning as for $\mathbb{Q}$, we see that $f(1) = 1$, and in fact, $f(r) = r$ for all $r \in \mathbb{Q}$. The next thing to note here is that if $a \in \mathbb{R}$ and $a \neq 0$, then $a^2 > 0$ and $f(a^2) = (f(a))^2$, so $f(a^2) > 0$. Since all positive real numbers have unique positive square roots, we can conclude that if $c > 0$, then $f(c) > 0$. Thus, if $a < b$, then $f(a) < f(b)$ since $b - a > 0$. Now take any real number c. If $c \in \mathbb{Q}$, then $f(c) = c$. If $c \notin \mathbb{Q}$ and $f(c) \neq c$, then there are two possibilities. If $c < f(c)$, choose a rational number r such that $c < r < f(c)$. Then $f(c) < f(r) = r$, which is a contradiction. If $f(c) < c$, we run into the same problem. So we conclude that $f(c) = c$ for all $c \in \mathbb{R}$.

Theorem 1.10.3 The groups $\mathrm{Aut}(\mathbb{Q})$ and $\mathrm{Aut}(\mathbb{R})$ consist only of the identity.

Exercise 1.10.4 Find a field F such that $\mathrm{Aut}(F) \neq \{1\}$.

Exercise 1.10.5 Find nontrivial elements of $\mathrm{Aut}(\mathbb{C})$.

Exercise 1.10.6

i. Let F be a field and let ϕ be an element of $\mathrm{Aut}(F)$. Define $H_\phi = \{x \in F \mid \phi(x) = x\}$. Show that H_ϕ is a subfield of F.

ii. Suppose that F is a field and that $\mathbb{Q}$ is a subfield of F. If $\phi \in \mathrm{Aut}(F)$, show that $\mathbb{Q}$ is a subfield of H_ϕ.

Exercise 1.10.7

i. Find $\mathrm{Aut}(\mathbb{Z}_p)$ where p is a prime and $\mathbb{Z}_p$ is the finite field with p elements.

ii. Let $F = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$. Show that F is a field and find $\mathrm{Aut}(F)$. This is the beginning of the subject called Galois theory, in which one of the goals is to determine $\mathrm{Aut}(F)$ when F is a so called "algebraic extension" of $\mathbb{Q}$.

More generally, an algebraic extension of $\mathbb{Q}$ is a subfield of $\mathbb{C}$ all of whose elements are roots of polynomials with coefficients in $\mathbb{Q}$ (see Definition A.7.31).

If R is a commutative ring with 1, we write $R[x]$ for the collection of polynomials in the variable x with coefficients in R. We can add and multiply polynomials in the usual manner, and this makes $R[x]$ into a commutative ring with 1.

Exercise 1.10.8 Show that $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, and $\mathbb{C}[z]$ are integral domains. Determine the elements in each of these domains which have multiplicative inverses.

Definition 1.10.9 Let F be a field. We say that F is *algebraically closed* if every nonconstant polynomial in $F[x]$ has a root in F. That is, F is algebraically closed if, for every nonconstant $p(x) \in F[x]$, there is an element $r \in F$ such that $p(r) = 0$.

The most important example of an algebraically closed field is supplied by the Fundamental Theorem of Algebra, which states that the field of complex numbers is algebraically closed. There is a semi-infinite number of proofs of this theorem. We will present one of these in Project 3.10.4 using the properties of continuous functions developed in Chapter 3.

Exercise 1.10.10 Let F be a field and suppose that $p(x) \in F[x]$. Show that r is a root of $p(x)$ if and only if $(x - r)$ is a factor of $p(x)$. That is, we can write $p(x) = (x - r)q(x)$ for some $q(x) \in F[x]$. (Hint: Consider the division algorithm for polynomials.)

Definition 1.10.11 Let $\mathbb{A}$ be the collection of all roots of polynomials in $\mathbb{Z}[x]$. $\mathbb{A}$ is called *the set of algebraic numbers* in $\mathbb{C}$. The set $\mathbb{A}_{\mathbb{R}} = \mathbb{A} \cap \mathbb{R}$ is called *the set of real algebraic numbers.* A real number which is not a real algebraic number is called *transcendental.*

Example 1.10.12 Among the more famous algebraic numbers are i and $-i$. For real algebraic numbers, the most famous one is probably $\sqrt{2}$. The most famous transcendental numbers are π and e.

Exercise 1.10.13 Show that $\mathbb{A}$ and $\mathbb{A}_{\mathbb{R}}$ are fields.

Exercise 1.10.14 Show that the field $\mathbb{A}$ of algebraic numbers is countable.

Remark 1.10.15 It follows from the exercise above that the field $\mathbb{A}_{\mathbb{R}}$ of real algebraic numbers is countable and hence the set of transcendental numbers is uncountable.

Exercise 1.10.16 Find nontrivial elements of $\mathrm{Aut}(\mathbb{A}_{\mathbb{R}})$.

Exercise 1.10.17 Find nontrivial elements of $\mathrm{Aut}(\mathbb{A})$ that are not on your list from the previous problem.

1.11 Independent Projects

1.11.1 Another construction of $\mathbb{R}$

Definition 1.11.1 A subset α of $\mathbb{Q}$ is said to be a *cut* (or a *Dedekind cut*) if it satisfies the following:

a. the set $\alpha \neq \varnothing$ and $\alpha \neq \mathbb{Q}$;

b. if $r \in \alpha$ and $s \in \mathbb{Q}$ satisfies $s < r$, then $s \in \alpha$;

c. if $r \in \alpha$, then there exists $s \in \mathbb{Q}$ with $s > r$ and $s \in \alpha$.

Let R denote the collection of all cuts.

Definition 1.11.2 For $\alpha, \beta \in R$, we define $\alpha + \beta = \{r + s \mid r \in \alpha \text{ and } s \in \beta\}$. Let $\mathbf{0} = \{r \in \mathbb{Q} \mid r < 0\}$.

Exercise 1.11.3 If α and β are cuts, show that $\alpha + \beta$ is a cut, and also show that $\mathbf{0}$ is a cut.

Exercise 1.11.4 Show that, with this addition, $(R,+)$ is an abelian group with $\mathbf{0}$ as the identity element.

We now define an order on R.

Definition 1.11.5 If $\alpha, \beta \in R$, we say that $\alpha < \beta$ if α is a proper subset of β.

Exercise 1.11.6 Show that the relation $<$ satisfies the following properties:

1. if $\alpha, \beta \in R$, then one and only one of the following holds: $\alpha < \beta$, $\alpha = \beta$, or $\beta < \alpha$ (Trichotomy);

2. if $\alpha, \beta, \gamma \in R$ with $\alpha < \beta$ and $\beta < \gamma$, then $\alpha < \gamma$ (Transitivity);

3. if $\alpha, \beta, \gamma \in R$ with $\alpha < \beta$, then $\alpha + \gamma < \beta + \gamma$ (Additivity).

It is now possible to define the notions of bounded above, bounded below, bounded, upper bound, least upper bound, lower bound, and greatest lower bound, in Rm just as we did earlier in this Chapter.

Exercise 1.11.7 Show that the least upper bound property holds in R, that is, if A is a nonempty subset of R which is bounded above, then A has a least upper bound in R.

Next, we must define multiplication in R.

Definition 1.11.8 If $\alpha, \beta \in R$ with $\alpha, \beta > 0$, then

$$\alpha\beta = \{p \in \mathbb{Q} \mid \text{there are positive elements } r \in \alpha \text{ and } s \in \beta \text{ such that } p \leq rs\}.$$

The next step is multiplication by $\mathbf{0}$, which is exactly as it should be, namely, for any $\alpha \in R$, we define $\alpha\mathbf{0} = \mathbf{0}$.

If $\alpha < 0$ or $\beta < 0$, or both, replace any negative element by its additive inverse and use the multiplication of positive elements to define multiplication accordingly. For example, if $\alpha < 0$ and $\beta > 0$, $\alpha\beta = -[(-\alpha)(\beta)]$.

Exercise 1.11.9 Show that R with addition, multiplication, and order as defined above is an ordered field. (Hint: think carefully about how to define the multiplicative inverse of a nonzero cut.)

Exercise 1.11.10 Put it all together and show that R is an Archimedean ordered field in which the least upper bound property holds.

1.11.2 Decimal Expansions of Real numbers

In Chapter **??**, we used a decimal representation of the real numbers to show that the real numbers between 0 and 1 form an uncountable set. In this project, we actually prove that every real number between 0 and 1 has a unique decimal expansion that does not terminate in all 9s. In addition, we discuss the fact that rational numbers have decimal expansions of three different types. The first is those rational numbers whose denominators are are divisors of a power of 10, the second is those whose denominators are relatively prime to 10, and the third is an intermediate case.

Since we know every real number lies between two consecutive integers (see Theorem 1.2.1), we start with a real number x such that $0 < x < 1$. Let $D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Assume first that x is irrational. The construction proceeds as follows. Let a_1 be the largest element of D which is less then $10x$. Then $0 < x - a_1/10 < 1/10$. Let a_2 be the largest integer in D less then $100x - 10a_1$. Proceeding as before we get $0 < x - a_1/10 - a_2/10^2 < 1/10^2$. Continuing this process, we obtain a monotonic increasing sequence $S_n = a_1/10 + a_2/10^2 + \cdots + a_n/10^n$, where $a_j \in D$ and $0 < x - S_n < 1/10^n$. So we conclude that S_n converges to x, and we get $x = a_1/10 + a_2/10^2 + \cdots + a_n/10^n + \ldots = \sum_{n=1}^{\infty} \frac{a_n}{10^n}$. We call $0.a_1a_2\ldots$ the *decimal expansion of x*.

Exercise 1.11.11 Let x be a irrational number between 0 and 1. Show that the decimal expansion x is unique.

We now turn to rational numbers between 0 and 1. We can apply the above procedure to rational numbers but with the possibility of equality in any of the inequalities above.

Suppose that x has a *terminating decimal expansion* that is, there exists N so that $a_n = 0$ for all $n > N$ and $a_N \neq 0$. Then we can write $x = a_1/10 + a_2/10^2 + \cdots + a_N/10^N$.

Exercise 1.11.12 *i.* Show that if r is a rational number in $(0, 1)$, then the decimal expansion of r terminates if and only if the denominator of r (when r is reduced to lowest terms) has the form 2^a5^b, where a and b are non-negative integers.

ii. With r as above show that the last non-zero digit of r is in the m-th place where $m = \max\{a, b\}$.

Note that rational numbers with terminating decimal expansions are the only real numbers between 0 and 1 for which equality can occur in initial procedure.

Next, consider a rational number $r = p/q$ in $(0, 1)$ for which q is relatively prime to 10. From Exercise 0.5.1.23, q divides $10^{\phi(q)} - 1$. Let n be the smallest natural number such that q divides $10^n - 1$. Then $(p/q)(10^n - 1)$ is an integer, which we denote by m. That is,

$$m = \frac{p}{q}(10^n - 1) \text{ or } \frac{p}{q} = \frac{m}{10^n - 1}.$$

Using results about geometric series from Section 1.9, we can now write

$$\frac{p}{q} = \frac{m}{10^n - 1} = \frac{m}{10^n}(1 - 10^{-n})^{-1} = \frac{m}{10^n}(1 + 10^{-n} + 10^{-2n} + \ldots) = m/10^n + m/10^{2n} + \ldots.$$

As $0 < p/q < 1$ we have $m < 10^n$. Thus the right hand side of the equation above gives us a periodic decimal expansion of p/q whose period has length at most n.

Exercise 1.11.13 Prove that the period is exactly n.

Exercise 1.11.14 Let p/q be a rational number between 0 and 1. If q and 10 are relatively prime, show that p/q has a unique periodic decimal expansion with the length of the period equal to the order of 10 mod q, that is, the smallest power of 10 that is congruent to 1 mod q.

We now present the remaining case as an exercise.

Exercise 1.11.15 Let p/q be a rational number in $(0, 1)$ with $q = 2^a 5^b r$, where r is relatively prime to 10. Let $k = \max\{a, b\}$, and let n be the smallest positive integer such that r divides $10^n - 1$. Show that after k digits, the decimal expansion of p/q is periodic of length n.

We ask finally whether decimal expansions are unique. The answer is contained in the following exercise.

Exercise 1.11.16

i. Consider the decimal $0.9999\ldots = \sum_{n=1}^{\infty} \frac{9}{10^n}$. Show that this geometric series converges to 1. That is, $0.9999\ldots = 1$.

ii. Show that every number that has a decimal expansion that ends in repeating nines can be written as a terminating decimal.

iii. Show that the exceptional situation in part *ii* is the only non-uniqueness that can occur.

Chapter 2

Linear Algebra

The presentation of the material will perhaps incur the reproach of being too dogmatic in its methods. To this possible objection, we would like to make two answers. Firstly, that what the student may learn here is not designed to help him with problems he has already met but with those he will have to cope with in the future; it is therefore impossible to motivate the definitions and theorems by applications of which the reader does not know the existence as yet. Secondly, that one of the important pedagogical problems which a teacher of beginners in mathematics has to solve is to impart to his students the technique of rigorous mathematical reasoning; this is an exercise in rectitude of thought, of which it would be futile to disguise the austerity.
– Claude Chevalley, Fundamental Concepts of Algebra

This is a book on analysis. However, almost all of the structures that we deal with in analysis have an underlying algebraic component, and an understanding of this algebraic component makes it a lot easier to discuss the analysis. Our approach is more user friendly than Chevalley's. We will find that analysis is mostly about inequalities, while algebra is mostly about equalities. The fundamental algebraic ideas we discuss in this chapter concern vector spaces and linear algebra. Other algebraic structures that play a role in analysis include groups, rings, and fields. Some ideas about these were discussed in Chapters 0 and 1. More can be found in the Projects at the end of this chapter.

2.1 Fundamentals of Linear Algebra

The algebraic structures that are really fundamental for analysis are vector spaces (sometimes called linear spaces). Recall that a field has been defined in Definition A.6.20.

Definition 2.1.1 Let F be a field. A *vector space* over F is a triple $(V, +, \cdot)$ where $(V, +)$ satisfies the axioms (A1)-(A5) of Chapter **??** and $\cdot$ is a map from $F \times V$ to V satisfying the following properties:

a. if $\alpha \in F$ and $\mathbf{v} \in V$, then $\alpha \cdot \mathbf{v} \in V$;

b. if $\alpha \in F$ and $\mathbf{v}_1, \mathbf{v}_2 \in V$, then $\alpha \cdot (\mathbf{v}_1 + \mathbf{v}_2) = (\alpha \cdot \mathbf{v}_1) + (\alpha \cdot \mathbf{v}_2)$;

c. if $\alpha, \beta \in F$ and $\mathbf{v} \in V$, then $(\alpha + \beta) \cdot \mathbf{v} = (\alpha \cdot \mathbf{v}) + (\beta \cdot \mathbf{v})$;

d. if $\alpha, \beta \in F$ and $\mathbf{v} \in V$, then $(\alpha\beta) \cdot \mathbf{v} = \alpha \cdot (\beta \cdot \mathbf{v})$;

e. if 1 is the multiplicative identity in F and $\mathbf{v} \in V$, then $1 \cdot \mathbf{v} = \mathbf{v}$.

The function $(\alpha, \mathbf{v}) \mapsto \alpha \cdot \mathbf{v}$ is called *scalar multiplication* and the elements of F are called *scalars*. We frequently suppress the dot $(\cdot)$.

For completeness, we restate axioms (A1)-(A5).

(A1) If $\mathbf{v}_1, \mathbf{v}_2 \in V$, then $\mathbf{v}_1 + \mathbf{v}_2 \in V$.

(A2) If $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V$, then $\mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3) = (\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3$.

(A3) If $\mathbf{v}_1, \mathbf{v}_2 \in V$, then $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$.

(A4) There exists $\mathbf{0} \in V$, such that for all $\mathbf{v} \in V$, $\mathbf{v} + \mathbf{0} = \mathbf{0} + \mathbf{v} = \mathbf{v}$.

(A5) For every $\mathbf{v} \in V$, there exists $-\mathbf{v} \in V$ such that $\mathbf{v} + (-\mathbf{v}) = -\mathbf{v} + \mathbf{v} = \mathbf{0}$.

Any structure satisfying (A1) – (A5) above is called an *abelian group* (See Project 2.1). Hence, a vector space is an abelian group with scalar multiplication.

Exercise 2.1.2

i. If 0 is the additive identity in F and $\mathbf{0}$ is the additive identity in V, show that $0 \cdot \mathbf{v} = \mathbf{0}$ for any $\mathbf{v} \in V$. Note that this statement actually says something. It says that the additive identity in F has a property for scalar multiplication similar to multiplication by 0 in commutative rings. However, in this case, multiplication by the 0 in F gives $\mathbf{0}$ in V.

ii. Show that condition e. does not follow from the other axioms.

Examples 2.1.3

1. If F is a field, $V = \{\mathbf{0}\}$ is a vector space over F.

2. If F is a field, then F is a vector space over F with scalar multiplication being ordinary multiplication in F.

3. Let F be field and let $F^n = \{(x_1, x_2, \ldots, x_n) \mid x_j \in F, j = 1, 2, \ldots, n\}$. Addition in F^n is defined coordinatewise, that is, if $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, then $\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$. If $\alpha \in F$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in F^n$, we set $\alpha \cdot \mathbf{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n)$. Then F^n is a vector space over F.

4. The ring of polynomial functions in one variable, $F[x]$, forms a vector space over F.

5. Let X be a nonempty set, let F be a field, and let $V = \mathcal{F}(X, F)$ be the set of all functions from X to F. For $f, g \in V$, define $(f + g)(x) = f(x) + g(x)$, and for $\alpha \in F$ define $(\alpha f)(x) = \alpha f(x)$. Then V is a vector space over F.

6. The real numbers $\mathbb{R}$ form a vector space over $\mathbb{Q}$.

Exercise 2.1.4 Check that the above examples satisfy the axioms for a vector space.

Remark 2.1.5 What is a vector? That's easy to answer. A vector is an element of a vector space. Lots of people describe a vector as a quantity having magnitude and direction. This is not particularly useful in most contexts (see, for example, 5 above) . However, in this chapter, when we do the geometric interpretation of vectors in n-dimensional Euclidean space, it will be helpful to think of vectors this way.

Exercise 2.1.6 Consider the collection $\mathbb{R}[x]$ of all polynomial functions with coefficients in $\mathbb{R}$. Show that $\mathbb{R}[x]$ is a vector space over $\mathbb{R}$ and also over $\mathbb{Q}$.

There are three fundamental notions which need to be discussed right at the beginning of our treatment of vector spaces. These are linear combinations, linear independence, and linear dependence. These notions are the heart and soul of elementary vector space theory.

Definition 2.1.7 Let V be a vector space over a field F, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in V$. Then a vector of the form

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m$$

where $\alpha_1, \alpha_2, \ldots, \alpha_m \in F$ is called a *linear combination* of $\mathbf{v}_1, \mathbf{v}_1, \ldots, \mathbf{v}_m$.

Remark 2.1.8 Note that the definition of linear combination refers to a sum of a *finite* number of vectors from the vector space V.

Definition 2.1.9 Let V be a vector space over a field F and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ be nonzero vectors in V. We say that the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is a *linearly independent* set if, for any scalars $\alpha_1, \ldots, \alpha_m \in F$ such that

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m = \mathbf{0},$$

we have $\alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$.

Example 2.1.10 A set containing a single nonzero vector forms a linearly independent set.

Exercise 2.1.11 In the vector space F^n over a field F, set $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)$, and generally, $\mathbf{e}_j = (0, 0, \ldots, 1, \ldots, 0)$, where the 1 is in the jth coordinate and 0 is in the other coordinates. Show that, for any k, $1 \leq k \leq n$, the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k\}$ is a linearly independent set.

Exercise 2.1.12 Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a linearly independent set in a vector space V. Show that any nonempty subset of this set is linearly independent.

Remark 2.1.13 An infinite set of vectors is said to be *linearly independent* if each finite subset is linearly independent.

Exercise 2.1.14 Show that the set $\{1, x, x^2, \ldots, x^n, \ldots\}$ is a linearly independent set in $\mathbb{Q}[x]$.

Definition 2.1.15 Let V be a vector space over a field F, and let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ be vectors in V. The set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is a *linearly dependent* set if there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_m \in F$, not all zero, such that

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m = \mathbf{0}.$$

In other words, a finite set of vectors is linearly dependent if it is not linearly independent.

Remark 2.1.16 An infinite set of vectors is said to be *linearly dependent* if there exists a finite subset that is linearly dependent.

Exercise 2.1.17

i. Let $V = F^n$ and let $\mathbf{v}$ be any vector in V. Show that the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n, \mathbf{v}\}$ is a linearly dependent set in V.

ii. Show that, if $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a set of vectors in a vector space V and one of these vectors is the zero vector, then the set is a linearly dependent set.

iii. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be vectors in a vector space V. Show that the set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly dependent set iff one of these vectors is a scalar multiple of the other.

Lemma 2.1.18 Suppose that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a linearly dependent set of nonzero vectors in a vector space V over a field F. Then there exists k, $1 < k \leq m$, such that $\mathbf{v}_k$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$.

Proof. Note that $\{\mathbf{v}_1\}$ is a linearly independent set because $\mathbf{v}_1$ is a nonzero vector. Let k be the largest integer such that $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$ is a linearly independent set. Observe that $k \leq m$ because $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a linearly dependent set. Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a linearly dependent set, so there exist scalars $\alpha_1, \ldots, \alpha_k$ such that

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k = \mathbf{0},$$

and $\alpha_k \neq 0$. Then

$$v_k = -\frac{\alpha_1}{\alpha_k}\mathbf{v}_1 - \frac{\alpha_2}{\alpha_k}\mathbf{v}_2 - \cdots - \frac{\alpha_{k-1}}{\alpha_k}\mathbf{v}_{k-1},$$

so the integer k gives the desired result. ∎

There are two additional ideas that are closely related. One is that of a spanning set, and the second is that of a basis.

Definition 2.1.19 Let V be a vector space over a field F, and let S be a set of vectors in V. The set S is a *spanning set* for V if every element of V can be written as a (finite!) linear combination of the vectors in S. We say that the set S *spans* V.

Example 2.1.20 Let F be a field and let $V = F^n$. The set $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ as defined in Exercise 2.1.11 spans V. In particular, if $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, then $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$.

Exercise 2.1.21 Let $V = \mathbb{R}^3$ over the field $F = \mathbb{R}$. Show that the set $S = \{\mathbf{v}_1 = (1,2,3), \mathbf{v}_2 = (4,5,6), \mathbf{v}_3 = (7,8,9)\}$ does not span V by finding a vector $\mathbf{v} \in V$ that cannot be written as a linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$.

Example 2.1.22 Let $V = \mathbb{Q}[x]$ be the vector space of polynomials defined over the field $F = \mathbb{Q}$. The collection of monomoials $S = \{1, x, x^2, \ldots, x^n, \ldots\}$ is a spanning set for V because every element in $\mathbb{Q}[x]$ can be written as a finite linear combination of these monomials. Note that no proper subset of S has this property. In fact, no finite set of vectors spans V.

Exercise 2.1.23 Show that the vector space $\mathbb{Q}[x]$ is a countable set.

Exercise 2.1.24 Let V be a vector space with a spanning set S. Suppose T is a set of vectors containing S. Show that T also spans V.

The second of these big ideas is that of a basis, which combines the aspects of linear independent sets of vectors and spanning sets of vectors.

Definition 2.1.25 A spanning set S is a *basis* for V if S is linearly independent.

Example 2.1.26 Let F be a field and let $V = F^n$. The spanning set $S = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is a basis for V. This is called *the standard basis*, or the *canonical basis* for V.

Example 2.1.27 Let $V = \mathbb{Q}[x]$ over the field $F = \mathbb{Q}$. The spanning set $S = \{1, x, x^2, \ldots, x_n, \ldots\}$ is a basis for V.

Exercise 2.1.28 Let V be a vector space over the field F, and let S be a basis of V. Show that every vector $\mathbf{v} \in V$ can be written *uniquely* as a linear combination of vectors in S.

Exercise 2.1.29 Show that no proper subset of a linearly independent set can be a spanning set.

What we now prove is that if a vector space has a basis with a finite number of elements, then all bases have the same number of elements. This will allow us to define the dimension of a vector space as the cardinality of a basis. The following lemma plays a crucial role.

Lemma 2.1.30 (Exchange Lemma) Suppose that V is a vector space over a field F and that $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is a spanning set for V. If the set $T = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a linearly independent set in V, then $n \leq m$.

Proof. The idea of the proof is to replace the elements of S one-by-one with elements of T in such a way that at each stage, we still have a spanning set for V. If $n > m$ then we will have a proper subset of T that spans V, which is impossible by Exercise 2.1.29.

First, consider the collection $S_1 = \{\mathbf{v}_1\} \cup \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$. Since S spans V, the set S_1 also spans V by Exercise 2.1.24. Either $\mathbf{v}_1 = u_{j_1}$ for some j_1, or S_1 contains S as a proper subset, and hence is linearly dependent by Exercise 2.1.29. In the latter case, according to Lemma 2.1.18 there exists an element $\mathbf{u}_{j_1} \in S_1$ which is a linear combination of $\mathbf{v}_1, \mathbf{u}_1, \ldots, \mathbf{u}_{j_1-1}$. In either case, we define the set $S_1' = \{\mathbf{v}_1, u_1, u_2, \ldots, u_{j_1-1}, u_{j_1+1}, \ldots, u_m\}$. This set still spans V because S_1 did, and u_{j_1} was itself a linear combination of the elements of S_1'.

Let us iterate this procedure. Define

$$S_2 = \{\mathbf{v}_2\} \cup \{\mathbf{v}_1, \mathbf{u}_1, \ldots, \mathbf{u}_{j_1-1}, \mathbf{u}_{j_1+1}, \ldots, \mathbf{u}_m\}.$$

By the same reasoning, S_2 is a spanning set. So, proceeding as above, we can find j_2 such that the set

$$S_2' = \{\mathbf{v}_2, \mathbf{v}_1\} \cup (\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\} \setminus \{u_{j_1}, u_{j_2}\})$$

is a spanning set. We can continue by putting $\mathbf{v}_3, \mathbf{v}_4, \ldots$ at the beginning of our list, and each time we do that, we can eliminate an element of U that remains on our list. Our procedure (that is, using Lemma 2.1.18) will never eliminate one of the $\mathbf{v}$'s since they are linearly independent. So at each stage, we eliminate one of the $\mathbf{u}$'s and are left with a spanning set. If at some point, all the $\mathbf{u}$'s are gone and some $\mathbf{v}$'s are left, then a proper subset of the $\mathbf{v}$'s would be a spanning set. This contradicts the linear independence of the $\mathbf{v}$'s.

Exercise 2.1.31 Suppose that V is a vector space over a field F and that $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is a spanning set for V. If T is *any* linearly independent subset of V, then T has at most m elements.

Now, we assume that our vector space V over F has a finite subset that spans. Such a vector space is called *finite dimensional*. The next corollary proves the existence of a basis for a finite dimensional vector space.

Lemma 2.1.32 Let V be a nonzero finite dimensional vector space over a field F. Then V has a finite basis.

Proof. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a spanning set consisting of nonzero vectors. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is linearly dependent, then by Lemma 2.1.18 there exists an integer k such that $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}\}$ is a linearly independent set and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is linearly dependent. Eliminating $\mathbf{v}_k$ from the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$, we still have a spanning set. Continue this process (a finite number of times). This yields a linearly independent set that spans V, that is, a basis.

Exercise 2.1.33 Let V be a nonzero finite dimensional vector space over a field F. Show that any basis of V has a finite number of elements.

Now we get to the heart of the matter.

Theorem 2.1.34 If V is a finite dimensional vector space over a field F, then any two bases of V have the same number of elements.

Proof. Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ are bases for V. Then by the Exchange Lemma, since the first set spans and the second set is linearly independent, we must have $m \leq n$. Similarly, since the second set spans and the first set is linearly independent, we must have $n \leq m$.

Now we can talk about an n-dimensional vector space over a field F.

Definition 2.1.35 Suppose V is a vector space containing a spanning set of n linearly independent vectors, $n \geq 1$. The *dimension* of V, denoted $\dim V$, is equal to n, that is $\dim V = n$. If $V = \{\mathbf{0}\}$, we set $\dim V = 0$.

Theorem 2.1.36 Suppose that V is an n-dimensional vector space over a field F and that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a linearly independent set in V. Then $m \leq n$ and there exist vectors $\mathbf{v}_{m+1}, \ldots, \mathbf{v}_n$ such that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m, \mathbf{v}_{m+1}, \ldots, \mathbf{v}_n\}$ is a basis for V.

Proof. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a basis for V. Applying the procedure in the proof of the Exchange Lemma m times, we get a basis $\{\mathbf{v}_m, \mathbf{v}_{m-1}, \ldots, \mathbf{v}_1, \mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_{n-m}}\}$.

Remark 2.1.37 Informally, the above theorem says that we can take any linearly independent set, and "extend" it to a basis.

Definition 2.1.38 If V is a vector space over a field F, then a nonempty subset $W \subseteq V$ is a *subspace* if it is closed under addition and scalar multiplication. That is, if $\mathbf{v}, \mathbf{w} \in W$, then $\mathbf{v} + \mathbf{w} \in W$, and if $\mathbf{v} \in W$ and $\alpha \in F$, then $\alpha\mathbf{v} \in W$.

Exercise 2.1.39

i. Let V be a vector space over a field F, show that $\{\mathbf{0}\}$ and V are subspaces of V.

ii. When is it true that the only subspaces of V are $\{\mathbf{0}\}$ and V?

Definition 2.1.40 Let V be a vector space over a field F. Let S be any nonempty subset of V. We define the *span of* S to be the set $\mathrm{Span}(S)$ of all linear combinations of elements of S. That is,

$$\mathrm{Span}(S) = \{\mathbf{v} \in V \mid \mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m \text{ for some } \alpha_1, \alpha_2, \ldots, \alpha_m \in F, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in S\}.$$

Additionally, we define $\mathrm{Span}(\varnothing)$ to be the set $\{\mathbf{0}\}$.

Exercise 2.1.41 Let S be any subset of V. Show that $\mathrm{Span}(S)$ is a subspace of V.

Examples 2.1.42

i. Let $V = \mathbb{Q}[x]$, and let W be the collection of all polynomials in $\mathbb{Q}[x]$ whose degree is less than or equal to a fixed non-negative integer n. Then W is a subspace of $\mathbb{Q}[x]$

ii. Let $V = F^n$, and for a fixed $m \leq n$, let $W = \{\mathbf{v} \in V \mid \mathbf{v} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \cdots + \alpha_m\mathbf{e}_m, \alpha_j \in F\}$. Then W is a subspace of V.

Exercise 2.1.43 Find the dimension of W in the last two examples.

Theorem 2.1.44 Let V be a finite dimensional vector space over a field F. Suppose that W is a subspace of V. Then $\dim W \leq \dim V$.

Proof. Suppose that $\dim V = n$. Consider the set S of all positive integers that are the cardinalities of linearly independent sets in W. The set S is bounded above by n by Theorem 2.1.36. Let h be the largest element of S. Let $B = \{\mathbf{w}_1, \ldots, \mathbf{w}_h\}$ be a linearly independent set in W of cardinality h. Then B must be a basis for W. Otherwise, there would be an element $\mathbf{w} \in W$ not in the span of B. Then h would not be maximal since $B \cup \{\mathbf{w}\}$ would be linearly independent. So, $\dim W = h \leq n$. ∎

Most of the vector spaces which arise in analysis are not finite dimensional and thus are called *infinite dimensional*. We will often be dealing with vector spaces of functions whose domain and range are finite dimensional vector spaces over $\mathbb{R}$ (or over the complex numbers $\mathbb{C}$), but the spaces of functions themselves will ordinarily be infinite dimensional spaces. To prove the existence of a basis for an infinite dimensional space, Zorn's Lemma or some other equivalent statement is needed.

Exercise 2.1.45 Suppose X is a set.

i. If $V = \mathcal{F}(X, F)$, the set of functions from X to F, show that V is finite dimensional if and only if X is finite.

ii. If X is a finite set, find an explicit basis for $V = \mathcal{F}(X, F)$.

iii. Fix a subset $A \subseteq X$ and define $W(A) = \{f \in V \mid f(a) = 0 \, \forall a \in A\}$. Show that $W(A)$ is a subspace of V.

iv. Can you find an infinite set X and a field F where you can write an explicit basis for $V = \mathcal{F}(X, F)$?

Theorem 2.1.46 Let V be a nonzero vector space over a field F. Then V has a basis. That is, there exists a linearly independent subset B of V such that each element of V is a finite linear combination of elements of B.

Proof. Let $\mathcal{X}$ be the collection of linearly independent sets in V. This collection can be partially ordered by inclusion. We apply Zorn's Lemma to $\mathcal{X}$ Given any totally ordered subset of $\mathcal{X}$, the union of the elements in this subset provides a maximal element in $\mathcal{X}$. The conclusion of Zorn's Lemma says that this maximal element gives a basis for V. ∎

Example 2.1.47 This last theorem means that there exists a basis $B = \{\mathbf{v}_i \mid i \in I\}$ for $\mathbb{R}$ considered as a vector space over $\mathbb{Q}$. In particular, every real number can be written as a finite linear combination of elements of this basis with coefficients taken from $\mathbb{Q}$. The basis is not countable. It is called a *Hamel basis* for $\mathbb{R}$ over $\mathbb{Q}$.

2.2 Linear Transformations

One of the most important topics in the subject of linear algebra is the study of maps from one vector space over a field F to another vector space over F that preserve addition and scalar multiplication. Such maps are called *linear transformations* and they play a vital role throughout the remainder of this text.

Definition 2.2.1 Let V, W be vector spaces over a field F. A function $T : V \to W$ is called a *linear transformation*, *linear map*, or *linear operator* if

a. $T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2)$ for all $\mathbf{v}_1, \mathbf{v}_2 \in V$,

b. $T(\alpha\mathbf{v}) = \alpha T(\mathbf{v})$ for all $\alpha \in F$ and $\mathbf{v} \in V$.

Exercise 2.2.2 Let V and W be vector spaces over a field F and $T : V \to W$ a linear transformation. Show that $T(\mathbf{0}) = \mathbf{0}$ and $T(-\mathbf{v}) = -T(\mathbf{v})$ for all $\mathbf{v} \in V$.

Definition 2.2.3 Let V and W be vector spaces over a field F and $T : V \to W$ a linear transformation. The *image* of T is the set $T(V) = \{\mathbf{w} \in W \mid \mathbf{w} = T(\mathbf{v}) \text{ for some } \mathbf{v} \in V\}$.

Exercise 2.2.4 *i.* Show that $T(V)$ is a subspace of W.

ii. If T is an injection, show that $T^{-1} : T(V) \to V$ is a linear operator.

Example 2.2.5 Consider $V = F$ as a vector space over F and fix $a \in F$. Define $T_a(x) = ax$ for $x \in V$. Then T_a is a linear transformation on F.

Exercise 2.2.6 Consider $\mathbb{R}$ as a vector space over itself and fix $a, b \in \mathbb{R}$. Define $T_{a,b}(x) = ax + b$. Show that $T_{a,b}$ is a linear transformation if and only if $b = 0$.

Example 2.2.7 Let $V = F^n$ and $W = F$ considered as vector spaces over F. For each $i \in \{1, 2, \ldots, n\}$, let $P_i : V \to W$ be the map given by $P_i(x_1, x_2, \ldots, x_i, \ldots, x_n) = x_i$. This map is a linear transformation called the *ith coordinate projection.*

Exercise 2.2.8 Let V be a vector space over a field F, and let W be a subspace of V. Show that $T : W \to V$ defined by $T(\mathbf{w}) = \mathbf{w}$ is a linear transformation.

Exercise 2.2.9 Let V be a vector space over a field F, and let $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis for V. Fix $m \leq n$. Show that the function $T : V \to V$ defined by

$$T(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m + \cdots + \alpha_n\mathbf{v}_n) = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m$$

is a linear transformation.

Example 2.2.10 Let $V = \mathbb{R}[x]$ considered as a vector space over $\mathbb{R}$. Define $D : \mathbb{R}[x] \to \mathbb{R}[x]$ by $[D(p)](x) = p'(x)$, that is, the derivative of the polynomial p. It follows from the properties of the derivative that D is a linear transformation.

Exercise 2.2.11 Show that $D : \mathbb{R}[x] \to \mathbb{R}[x]$ as defined in the previous example is surjective but not injective.

If $T : V \to W$ is a linear transformation, how are the dimensions of V, W, and $T(V)$ related? Our first result along these lines says that the dimension of $T(V)$ is less than or equal to that of V. Informally, we might say that a linear transformation cannot increase the dimension of a vector space.

Theorem 2.2.12 Suppose that V and W are vector spaces over a field F and that $T : V \to W$ is a linear transformation. If V is finite dimensional with $\dim V = n$, then $\dim T(V) \leq n$.

Proof. It suffices to show that every subset of $T(V)$ consisting of $n+1$ elements is linearly dependent. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n+1}$ be vectors in $T(V)$. Pick $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n+1} \in V$ such that $T(\mathbf{v}_j) = \mathbf{w}_j$ for $j = 1, \ldots, n+1$. Since $\dim V = n$, the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n+1}\}$ is linearly dependent, so there exist scalars $\alpha_1, \ldots, \alpha_{n+1}$, not all zero, such that $\alpha_1\mathbf{v_1} + \cdots + \alpha_{n+1}\mathbf{v}_{n+1} = \mathbf{0}$. It follows that $\alpha_1\mathbf{w}_1 + \cdots + \alpha_{n+1}\mathbf{w}_{n+1} = T(\alpha_1\mathbf{v}_1 + \cdots + \alpha_{n+1}\mathbf{v}_{n+1}) = T(\mathbf{0}) = \mathbf{0}$. Hence the set $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n+1}\}$ is linearly dependent.

Definition 2.2.13 Let V and W be vector spaces over a field F and $T : V \to W$ a linear transformation. The transformation T is called a *linear isomorphism* if T is a bijection. In this case V and W are said to be *linearly isomorphic.*

Corollary 2.2.14 Suppose that V and W are finite dimensional vector spaces over a field F. If V and W are linearly isomorphic, then $\dim V = \dim W$. Moreover, if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V, then $\{T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)\}$ is a basis for W.

Exercise 2.2.15 Suppose that V and W are finite dimensional vector spaces over a field F such that $\dim V = \dim W$. Show that V and W are linearly isomorphic.

In general, if T is a linear transformation from V to W, T is neither injective nor surjective. This leads to an important idea, the kernel of a linear transformation.

Definition 2.2.16 Let V and W be vector spaces over a field F, and $T : V \to W$ a linear transformation. The *kernel of* T is defined by

$$\ker T = \{\mathbf{v} \in V \mid T\mathbf{v} = \mathbf{0}\}.$$

Exercise 2.2.17 *i.* Show that $\ker T$ is a subspace of V.

ii. Show that T is injective if and only if $\ker T = \{\mathbf{0}\}$.

iii. Let $D : \mathbb{R}[x] \to \mathbb{R}[x]$ be defined as above, that is, $[D(p)](x) = p'(x)$. Find $\ker D$.

iv. Let $P_i : F^n \to F$ be the ith coordinate projection as defined in Example 2.2.7. Find $\ker P_i$.

The notions of kernel and image allow us to give a more precise answer to the question we asked earlier about the relation between the dimensions of V, W, and $T(V)$.

Theorem 2.2.18 Suppose that V and W are vector spaces over a field F and that $\dim V$ is finite. Let $T : V \to W$ be a linear transformation. Then $\dim V = \dim \ker T + \dim T(V)$.

Proof. Since $\ker T$ is a subspace of V, we know that $k = \dim \ker T \leq n$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be a basis for $\ker T$. We can extend this to a basis $\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$ for V by Theorem 2.1.36. We claim that $\{T(\mathbf{v}_{k+1}), \ldots, T(\mathbf{v}_n)\}$ is a basis for $T(V)$. The equation in the statement of the theorem is now obvious once we verify the claim.

Let $\mathbf{w} \in T(V)$. Then there exists $\mathbf{v} \in V$ such that $T(\mathbf{v}) = \mathbf{w}$. Since $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a basis for V, there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_n \in F$ such that $\mathbf{v} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$. Then $\mathbf{w} = T(\mathbf{v}) = T(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n) = \alpha_1 T(\mathbf{v}_1) + \alpha_2 T(\mathbf{v}_2) + \cdots + \alpha_n T(\mathbf{v}_n) = \alpha_{k+1} T(\mathbf{v}_{k+1}) + \alpha_{k+2} T(\mathbf{v}_{k+2}) + \cdots + \alpha_n T(\mathbf{v}_n)$. Hence the set $\{T(\mathbf{v}_{k+1}), T(\mathbf{v}_{k+2}), \ldots, T(\mathbf{v}_n)\}$ spans $T(V)$.

To show that this set of vectors in linearly independent, suppose that $\alpha_{k+1}T(\mathbf{v}_{k+1}) + \alpha_{k+2}T(\mathbf{v}_{k+2}) + \cdots + \alpha_n T(\mathbf{v}_n) = 0$. Then $T(\alpha_{k+1}\mathbf{v}_{k+1} + \alpha_{k+2}\mathbf{v}_{k+2} + \cdots + \alpha_n\mathbf{v}_n) = 0$, hence $\alpha_{k+1}\mathbf{v}_{k+1} + \alpha_{k+2}\mathbf{v}_{k+2} + \cdots + \alpha_n\mathbf{v}_n \in \ker T$. Hence, there exist scalars $\beta_1, \beta_2, \ldots, \beta_k \in F$ such that $\alpha_{k+1}\mathbf{v}_{k+1} + \alpha_{k+2}\mathbf{v}_{k+2} + \cdots + \alpha_n\mathbf{v}_n = \beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \cdots + \beta_k\mathbf{v}_k$, or $\beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \cdots + \beta_k\mathbf{v}_k - \alpha_{k+1}\mathbf{v}_{k+1} - \alpha_{k+2}\mathbf{v}_{k+2} - \cdots - \alpha_n\mathbf{v}_n = 0$. Since $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a basis for V, it is a linearly independent set, all of the coefficients, including $\alpha_{k+1}, \alpha_{k+2}, \ldots, \alpha_n$, must be zero. Hence the set $\{T(\mathbf{v}_{k+1}), T(\mathbf{v}_{k+2}), \ldots, T(\mathbf{v}_n)\}$ is linearly independent, and thus a basis for $T(V)$.

Remark 2.2.19 Other authors refer to the preceding theorem as the Rank-Nullity Theorem. This is because the *rank* of T is defined to be $\dim T(V)$ and the *nullity* of T is defined to be $\dim \ker T$.

Exercise 2.2.20 Let V and W be finite dimensional vector spaces over a field F with $\dim V = \dim W$. Let $T : V \to W$ be a linear transformation. Show that the following are equivalent.

1. T is bijective.

2. T is surjective.

3. T is injective.

Definition 2.2.21 Let V and W be vector spaces over a field F. If T, T_1, and T_2 are linear transformations from V to W, we define

a. $(T_1 + T_2)(\mathbf{v}) = T_1(\mathbf{v}) + T_2(\mathbf{v})$, for $\mathbf{v} \in V$, and

b. $(\alpha T)(\mathbf{v}) = \alpha T(\mathbf{v})$, for $\alpha \in F$.

Theorem 2.2.22 Let V and W be vector spaces over a field F. Let $\mathscr{L}(V, W)$ denote the set of all linear transformations from V to W. Then, with the above operations, $\mathscr{L}(V, W)$ is a vector space over F.

Proof. Clear.

Exercise 2.2.23 Show that if $\dim V = n$ and $\dim W = m$, then $\mathscr{L}(V, W)$ is a finite dimensional vector space with $\dim \mathscr{L}(V, W) = nm$.

The proof of this exercise is facilitated by the use of bases in V and W. This will lead us to the notion of matrices representing linear transformations in the next section.

Finally, we consider the composition of linear transformations.

Exercise 2.2.24 Let U, V, and W be vector spaces over a field F. Let $S, S_1, S_2 \in \mathscr{L}(U, V)$ and $T, T_1, T_2 \in \mathscr{L}(V, W)$.

i. Show that $T \circ S \in \mathscr{L}(U, W)$.

ii. Show that $T \circ (S_1 + S_2) = (T \circ S_1) + (T \circ S_2)$.

iii. Show that $(T_1 + T_2) \circ S = T_1 \circ S + T_2 \circ S$.

2.3 Linear Transformations and Matrices

Let V be finite dimensional vector space over a field F, and suppose that $\dim V = n$. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for V. If $\mathbf{v} \in V$ then we can write $\mathbf{v} = \sum_{k=1}^{n} \beta_k \mathbf{v}_k$ for some scalars $\beta_1, \beta_2, \ldots, \beta_n \in F$. These scalars are called the *coefficients* of $\mathbf{v}$ relative to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Once we have chosen a basis, it is common to represent the vector $\mathbf{v}$ by writing these coefficients in a column:

$$\mathbf{v} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}.$$

We call this expression with n scalars a $n \times 1$ *column vector*, where n refers to the number of rows, and 1 refers to the number of columns, i.e., a single one. We sometimes abbreviate this notation by writing $\mathbf{v} = (\beta_j)$.

Now take finite dimensional vector spaces V and W over a field F and $T \in \mathscr{L}(V, W)$. Suppose that $\dim V = n$, $\dim W = m$, and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ are bases for V and W, respectively. For $1 \leq k \leq n$, we can write

$$T(\mathbf{v}_k) = \sum_{j=1}^{m} a_{jk} \mathbf{w}_j,$$

where each $a_{jk} \in F$. That is, the particular scalar a_{jk} is the coefficient of $\mathbf{w}_j$ when writing the vector $T(\mathbf{v}_k)$ in terms of the basis $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. These scalars are sufficient to characterize T, since any vector in V can be written as a unique linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and any vector in W can be written as a unique linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_m$. We encode this information about T in a rectangular array called a *matrix*. That is, we write

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Observe that the matrix A has m rows and n columns (rows are horizontal and columns are vertical). This matrix is called an *$m \times n$ (m by n) matrix over F*. Of course, the coefficients in the matrix depend on the choice of bases in V and W.

Now, for any $\mathbf{v} \in V$, we may write $\mathbf{v}$ with respect to the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. As above, we find scalars $\beta_1, \beta_2, \ldots, \beta_n \in F$ such that $\mathbf{v} = \sum_{k=1}^n \beta_k \mathbf{v}_k$. Then we write $T(\mathbf{v})$ with respect to the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ as follows:

$$T(\mathbf{v}) = \sum_{k=1}^n \beta_k \left(\sum_{j=1}^m a_{jk} \mathbf{w}_j \right) = \sum_{j=1}^m \left(\sum_{k=1}^n \beta_k a_{jk} \right) \mathbf{w}_j.$$

Expressing this same computation with the notation introduced above, we write

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n a_{1k}\beta_k \\ \sum_{k=1}^n a_{2k}\beta_k \\ \vdots \\ \sum_{k=1}^n a_{mk}\beta_k \end{pmatrix}.$$

Note that the resulting $m \times 1$ column vector is the expression for $T(\mathbf{v})$ in terms of its coefficients with respect to the basis $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$. We also refer to this column vector as the product of the matrix $A = (a_{ij})$ by the vector $\mathbf{v} = (\beta_j)$.

Incidentally, this might be a good time to formalize the definition of matrix.

Definition 2.3.1 Let R be a commutative ring with 1. Let a_{ij} be elements of R, where $1 \leq i \leq m$ and $1 \leq j \leq n$. An *$m \times n$ matrix over R* is a rectangular array given by

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

Exercise 2.3.2 Let V and W be finite dimensional vector spaces over F of dimensions n and m respectively. Let $M_{mn}(F)$ be the collection of $m \times n$ matrices over F. We use the notation $A = (a_{ij})$ for elements of $M_{mn}(F)$. If $A = (a_{ij})$, $B = (b_{ij})$, we define $A + B = (a_{ij} + b_{ij})$ and for $\alpha \in F$ we define $\alpha A = (\alpha a_{ij})$.

i. Show that $M_{mn}(F)$ is a vector space over F.

ii. Find a basis for $M_{mn}(F)$.

iii. By fixing bases for V and W give an explicit linear isomorphism between $\mathscr{L}(V, W)$ and $M_{mn}(F)$.

In the previous section, we saw that if T_1 and T_2 are in $L(V, W)$, then $T_1 + T_2$ is in $L(V, W)$, and in the previous exercise, we saw how to add their respective matrices. Similarly, we saw that if $\alpha \in F$ and $T \in L(V, W)$, then αT is in $L(V, W)$, and in the previous exercise, we saw how to multiply the matrix for T by the scalar α.

We turn now to the question of how to write the matrix for a linear transformation that is the composition of two linear transformations. Let U, V, and W be finite dimensional vector spaces over a field F. Let $S \in L(U, V)$, and $T \in L(V, W)$. An exercise in the previous section showed that $T \circ S$ is in $L(U, W)$. Let $\{\mathbf{u}_1, \ldots, \mathbf{u}_l\}$ be a basis for U. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for V. Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis for W. How are the matrices for S, T, and $T \circ S$ related? Let $A = (a_{jk})$ be the $n \times l$ matrix for S, and let $B = (b_{ij})$ be the $m \times n$ matrix for T. (Note that we have intentionally used j as an index that runs from 1 to n in both cases.) What is the matrix for $T \circ S$? We express $(T \circ S)(\mathbf{u}_k)$ in terms of the basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ as follows:

$$\begin{aligned}(T \circ S)(\mathbf{u}_k) &= T(S(\mathbf{u}_k))\\ &= T\left(\sum_{j=1}^{n} a_{jk}\mathbf{v}_j\right)\\ &= \sum_{j=1}^{n} a_{jk}T(\mathbf{v}_j)\\ &= \sum_{j=1}^{n} a_{jk}\sum_{i=1}^{m} b_{ij}\mathbf{w}_i\\ &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n} b_{ij}a_{jk}\right)\mathbf{w}_i.\end{aligned}$$

If $C = (c_{ik})$ is the matrix for $T \circ S$, with respect to the two bases $\{\mathbf{u}_1, \ldots, \mathbf{u}_l\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, then the preceding computation shows that $c_{ik} = \sum_{j=1}^{n} b_{ij}a_{jk}$. This inspires us to define matrix multiplication formally.

Definition 2.3.3 Let $A = (a_{jk})$ be an $n \times l$ matrix, and let $B = (b_{ij})$ be an $m \times n$ matrix. Then the *product* BA is the $m \times l$ matrix $C = (c_{ik})$, where $c_{ik} = \sum_{j=1}^{n} b_{ij}a_{jk}$.

Remark 2.3.4 Let $A = (a_{jk})$ be an $n \times l$ matrix, and let $B = (b_{ij})$ be an $m \times n$ matrix. The definition above shows us how to multiply the matrix B by the matrix A. In should be noted that given this setup, the product AB is not even defined unless $l = m$. Even if $l = m$, the matrices BA and AB will be different sizes unless $l = m = n$. These are severe hurdles to any thoughts of commutativity in doing matrix multiplication, but this comes back to the more fundamental idea that the composition of functions is, in general, not commutative.

The astute reader will notice the similarity between the way we defined the multiplication of two matrices, and the way we defined the multiplication of a matrix by a vector. In fact, the latter is a special case of the former.

Exercise 2.3.5 Let $A = (a_{ij})$ be an $m \times n$ matrix, and let $\mathbf{v} = (\beta_j)$ be an $n \times 1$ column vector. If we think of $A\mathbf{v}$ as a product of matrices, what linear transformations do $\mathbf{v}$ and $A\mathbf{v}$ represent?

The above development is of particular importance in the case when $V = W$. In this case, we use the notation $\mathscr{L}(V)$ for $\mathscr{L}(V, V)$. If $\dim V = n$, then $\dim \mathscr{L}(V) = n^2$, and each element of $\mathscr{L}(V)$ can be represented by an $n \times n$ matrix relative to a single chosen basis of V. Along with the operations of addition and scalar multiplication in $\mathscr{L}(V)$, we have composition of linear transformations. Suppose that $S, T \in \mathscr{L}(V)$. Then $S \circ T$ is defined in the usual way by $S \circ T(\mathbf{v}) = S(T(\mathbf{v}))$.

Exercise 2.3.6

i. If $S, T \in \mathscr{L}(V)$, show that $S \circ T \in \mathscr{L}(V)$.

ii. If $R, S, T \in \mathscr{L}(V)$, then $R \circ (S \circ T) = (R \circ S) \circ T$ (this actually follows from the associativity of composition of functions discussed in the Appendix).

iii. If $R, S, T \in \mathscr{L}(V)$, show that $R \circ (S + T) = (R \circ S) + (R \circ T)$ and $(R + S) \circ T = (R \circ T) + (S \circ T)$.

iv. Let $I \in \mathscr{L}(V)$ be defined by $I(\mathbf{v}) = \mathbf{v}$ for $\mathbf{v} \in V$. Show that $T \circ I = I \circ T = T$ for all $T \in \mathscr{L}(V)$.

v. Show that if $\dim V \geq 2$, then $\mathscr{L}(V)$ is not commutative with respect to $\circ$. That is, there exist $S, T \in \mathscr{L}(V)$ such that $S \circ T \neq T \circ S$.

A little vocabulary is in order here. In Chapter **??**, we used the terms commutative ring with 1, integral domain, and field. As pointed out there, the word "commutative" referred to the operation of multiplication. Some of the most important algebraic structures that occur in analysis are called *algebras*.

Definition 2.3.7 Let F be a field. An *algebra* over F is a set A such that A is a vector space over F and A has an internal law of composition $\circ$ satisfying the associative law, and left and right distributivity. That is, for $a, b, c \in A$, we have

$$\begin{aligned} a \circ (b \circ c) &= (a \circ b) \circ c, \\ a \circ (b + c) &= (a \circ b) + (a \circ c) \\ \text{and } (a + b) \circ c &= (a \circ c) + (b \circ c). \end{aligned}$$

For scalar multiplication we have for $\alpha \in F$, $(\alpha \cdot a) \circ b = \alpha \cdot (a \circ b) = a \circ (\alpha \cdot b)$.

An algebra A is an *algebra with identity* if there is an element $\mathbf{1} \in A$ so that $a \circ \mathbf{1} = \mathbf{1} \circ a = a$ for all $a \in A$. The algebra A is a *commutative algebra* if $a \circ b = b \circ a$ for all $a, b \in A$.

Example 2.3.8 If V is a nonzero vector space over a field F, then $\mathscr{L}(V)$ is an algebra with identity which is commutative if and only if $\dim V = 1$.

Exercise 2.3.9 Let $\mathbb{R}[x]$ be the vector space of polynomial functions in one variable over $\mathbb{R}$. Define multiplication of polynomials in the usual way. Show that $\mathbb{R}[x]$ is a commutative algebra with identity.

Exercise 2.3.10 Let $A = M_{nn}(F)$. Show that multiplication of $n \times n$ matrices over a field F is associative and that

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

is an identity for multiplication of matrices. Show further that, for $n \times n$ matrices, multiplication is left and right distributive over addition and the appropriate properties hold for scalar multiplication.

Conclusion: $M_{nn}(F)$ is an algebra with identity over F which is commutative if and only if $n = 1$.

For simplicity we will write $M_n(F)$ for $M_{nn}(F)$.

An invertible element of $M_n(F)$ is one that has a multiplicative inverse. The collection of these invertible elements plays a special role, which we will investigate in several ways. First, we will finish this section by discussing how to decompose an invertible matrix into "elementary matrices." Secondly, in the next section, we will investigate the determinant a matrix, which gives a precise condition under which a matrix is invertible. Thirdly, in Project 2.6.1, we will discuss the algebraic structure of $GL_n(F)$, the collection of invertible matrices. This will be of particular importance when we get to the Change of Variables Theorem in Chapter 5.

Definition 2.3.11 An *elementary transformation* in $GL_n(F)$ is a linear transformation of one of the following three forms:

1. multiplication of a coordinate by a nonzero constant k:

$$T(x_1, x_2, \ldots, x_{a-1}, x_a, x_{a+1}, \ldots, x_n) = (x_1, x_2, \ldots, x_{a-1}, kx_a, x_{a+1}, \ldots, x_n);$$

2. interchange of two coordinates:

$$T(x_1, x_2, \ldots, x_{a-1}, x_a, x_{a+1}, \ldots, x_{b-1}, x_b, x_{b+1}, \ldots x_n) = (x_1, x_2, \ldots, x_{a-1}, x_b, x_{a+1}, \ldots, x_{b-1}, x_a, x_{b+1}, \ldots, x_n);$$

3. replacement of a coordinate by the sum of itself and another:

$$T(x_1, x_2, \ldots, x_{a-1}, x_a, x_{a+1}, \ldots, x_b, \ldots, x_n) = (x_1, x_2, \ldots, x_{a-1}, x_a + x_b, x_{a+1}, \ldots, x_b, \ldots, x_n).$$

An *elementary matrix* is the matrix of an elementary transformation with respect to the standard basis.

Exercise 2.3.12 *i.* Find the elementary matrix for each elementary transformation.

ii. Show that any matrix in $GL_n(F)$ can be written as a product of elementary matrices.

2.4 Determinants

Exercise 2.4.1 Suppose that $T \in \mathscr{L}(V)$ is a bijection. Then T has an inverse, $T^{-1} : V \to V$. Show that $T^{-1} \in \mathscr{L}(V)$. (See Exercise 2.2.4.)

If $T \in \mathscr{L}(V)$, and T is invertible, then of course $T \circ T^{-1} = T^{-1} \circ T = I$, where I is the identity map. The problem that confronts us is the following: if $\dim V = n$, and $A = (a_{ij})$ is the matrix of T relative to a given basis of V, how do we find the matrix of T^{-1} relative to the same basis? We seek a matrix denoted by A^{-1} such that $A \cdot A^{-1} = A^{-1} \cdot A = I$.

Well, this shouldn't present a great problem. All we do is write the matrix for A^{-1} as (x_{ij}). This leads to n^2 linear equations in n^2 unknowns:

$$\begin{array}{rcl}
a_{11}x_{11} + a_{12}x_{21} + \cdots + a_{1n}x_{n1} & = & 1 \\
a_{21}x_{11} + a_{22}x_{21} + \cdots + a_{2n}x_{n1} & = & 0 \\
& \vdots & \\
a_{n1}x_{11} + a_{n2}x_{21} + \cdots + a_{nn}x_{n1} & = & 0 \\
a_{11}x_{12} + a_{12}x_{22} + \cdots + a_{1n}x_{n2} & = & 0 \\
& \vdots & \\
a_{n1}x_{1n} + a_{n2}x_{2n} + \cdots + a_{nn}x_{nn} & = & 1.
\end{array}$$

This looks somewhat tedious, so maybe at this stage, we should just tell you the answer and consider it further in a project at the end of the chapter. But, that would not be true to the nature of this book. So we are led to the quest for determinants, one of the great discoveries in mathematics.

To begin a discussion of determinants, we must first consider the collection S_n of all bijections from the set $[n] = \{1, 2, \ldots, n\}$ to itself. These bijections are called *permutations of* n *elements*.

Example 2.4.2 Let us consider two examples, namely S_2 and S_3. We can represent the elements of S_2 by arrays in which the top row lists domain and the bottom lists the corresponding elements in the image. There are two permutations of two elements, namely

$$I = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad r = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Note that these arrays should be thought of not as matrices, but simply as a way to represent permutations as functions. Similarly, we can write the six elements of S_3 as

$$I = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad r = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad r^2 = r \circ r = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$$

$$f_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad f_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \quad f_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

Exercise 2.4.3

i. Use the fundamental counting principle (see Theorem A.4.6) to show that the number of elements in S_n is $n!$.

ii. Show that the composition of two elements of S_n is also an element of S_n.

iii. Show that the elements of S_n satisfy the associative law under composition of functions.

iv. Define $I \in S_n$ by $I(x) = x$ for all $x \in [n]$. Show that I is an identity for S_n under composition.

v. If $\sigma \in S_n$, define σ^{-1} as one does for any bijection. That is, $\sigma(x) = y$ iff $\sigma^{-1}(y) = x$. Show that $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = I$.

The collection S_n with the internal law of composition $\circ$ satisfying the above properties is an example of an algebraic structure called a *group*. The general theory of groups is discussed in the Projects at the end of the chapter. The group S_n is called *the symmetric group on n objects*.

Definition 2.4.4 Let σ be an element of S_n. We define the *sign* of σ by

$$\operatorname{sgn}(\sigma) = \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\sigma(j) - \sigma(i)}{j - i}.$$

Exercise 2.4.5 Show that if $\sigma \in S_n$, then $\operatorname{sgn}(\sigma) = \pm 1$.

Exercise 2.4.6 Let σ be any function from $[n]$ to $[n]$. We can define $\operatorname{sgn}(\sigma)$ as above. Show that $\operatorname{sgn}(\sigma) = 0$ if and only if σ is not a bijection.

Proposition 2.4.7 For $\sigma, \tau \in S_n$, we have $\operatorname{sgn}(\sigma \circ \tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau)$

Proof. We have

$$\begin{aligned}
\operatorname{sgn}(\sigma \circ \tau) &= \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{j - i} \\
&= \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} \cdot \frac{\tau(j) - \tau(i)}{j - i} \\
&= \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\sigma(\tau(j)) - \sigma(\tau(i))}{\tau(j) - \tau(i)} \cdot \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\tau(j) - \tau(i)}{j - i} \\
&= \prod_{\substack{\{k,l\} \subset [n], \\ k \neq l}} \frac{\sigma(l) - \sigma(k)}{l - k} \cdot \prod_{\substack{\{i,j\} \subset [n], \\ i \neq j}} \frac{\tau(j) - \tau(i)}{j - i} \\
&= \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau)
\end{aligned}$$

Exercise 2.4.8 Find the signs of the permutations in S_2 and S_3.

Definition 2.4.9

a. The permutations σ in S_n for which $\operatorname{sgn}(\sigma) = 1$ are called *even permutations*. Those permutations σ for which $\operatorname{sgn}(\sigma) = -1$ are called *odd permutations*. The collection of even permutations is denoted by A_n.

b. A permutation which interchanges two distinct elements and leaves the remaining elements fixed is called a *transposition.* The transposition which sends i to j, j to i, and leaves everything else fixed is written (ij).

Exercise 2.4.10

i. Show that a transposition is an odd permutation.

ii. Show that every element of S_n can be decomposed as a product of transpositions.

iii. Show that $\sigma \in S_n$ is an even permutation if and only if σ can be decomposed into an even number of transpositions. Also show that $\sigma \in S_n$ is an odd permutation if and only if σ can be decomposed into an odd number of transpositions. The number of transpositions is not unique but the parity is always the same.

iv. Show that A_n is a group. (See Project 2.1. A_n is called the *Alternating Group* on n objects).

v. Show that the number of elements in A_n is $n!/2$.

vi. Show that $\text{sgn}(\sigma)$ can be defined as simply the sign of the integer $\prod_{1\leq i<j\leq n}(\sigma(j)-\sigma(i))$.

vii. Show that $A_2 = \{I\}$ and $A_3 = \{I, r, r^2\}$.

viii. Decompose each element in S_3 as a product of transpositions.

ix. Write explicitly as arrays the elements of A_4 and A_5.

We are now prepared to define the determinant of an $n \times n$ matrix.

Definition 2.4.11 Let $A = (a_{ij})$ be an $n \times n$ matrix over a field F. The *determinant of* A, denoted by $\det A$, is defined as

$$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)}$$

Example 2.4.12 Consider a 2×2 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Then $\det A = a_{11}a_{22} - a_{12}a_{21}$.

Exercise 2.4.13 Write out the expression for the determinant of a 3×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

It should have $3! = 6$ terms.

We have two tasks ahead. The first is to illustrate the role of the determinant in computing the inverse of an $n \times n$ matrix. The second is to find some reasonable way to compute the determinant of a matrix.

Definition 2.4.14 Suppose $A = (a_{ij})$ is an $n \times n$ matrix over a field F. Then *the transpose of* A, denoted ${}^tA$, is the matrix obtained by reflecting A around the main diagonal, that is, the collection of elements $a_{11}, a_{22}, \ldots, a_{nn}$. Thus,

$${}^tA = (a_{ji}).$$

Hence, if

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \text{ then } \quad {}^tA = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix},$$

and, if

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \text{ then } \quad {}^tA = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}.$$

Exercise 2.4.15 Let A be an $n \times n$ matrix over a field F. Show that $\det(A) = \det({}^tA)$.

Exercise 2.4.16 Let $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. Suppose that $\det A \neq 0$. Show that A^{-1} exists and find it.

Lemma 2.4.17 If $A = (a_{ij})$ is an $n \times n$ matrix over a field F such that, for some m, k, with $m \neq k$, the m-th row is equal to the k-th row, then $\det A = 0$.

Proof. For any $\sigma \in S_n$ let $\tilde{\sigma} = \sigma \circ (km)$, where (km) is the transposition defined above. Our assumption implies that

$$a_{1,\sigma(1)}a_{2,\sigma(2)} \cdots a_{n,\sigma(n)} = a_{1,\tilde{\sigma}(1)}a_{2,\tilde{\sigma}(2)} \cdots a_{n,\tilde{\sigma}(n)}$$

On the other hand, $\operatorname{sgn}(\tilde{\sigma}) = \operatorname{sgn}(\sigma)\operatorname{sgn}((km)) = \operatorname{sgn}(\sigma) \cdot (-1) = -\operatorname{sgn}(\sigma)$. This shows that $\det A = 0$. ∎

Exercise 2.4.18 Show that, if $A = (a_{ij})$ is an $n \times n$ matrix over a field F such that, for some m, k, with $m \neq k$, the m-th column is equal to the k-th column, then $\det A = 0$.

The following exercise will prove useful in our discussion of the properties of determinants.

Exercise 2.4.19 Suppose that $A = (a_{ij})$ and $B = (b_{ij})$ are $n \times n$ matrices over a field F. Suppose further that $\sigma : [n] \to [n]$ is not a bijection. Show that

$$\sum_{\rho \in S_n} \operatorname{sgn}(\rho) a_{1,\sigma(1)} b_{\sigma(1),\rho(1)} a_{2,\sigma(2)} b_{\sigma(2),\rho(2)} \cdots a_{n,\sigma(n)} b_{\sigma(n),\rho(n)} = 0.$$

Definition 2.4.20 Let $A = (a_{ij})$ be an $n \times n$ matrix over a field F. Let A_{ij} be the $(n-1) \times (n-1)$ matrix obtained by deleting the i-th row and the j-th column of A. The (i, j) *cofactor* of A is the element C_{ij} of F defined by $C_{ij} = (-1)^{i+j} \det A_{ij}$

Theorem 2.4.21 Let $A = (a_{ij})$ be an $n \times n$ matrix over a field F. Then, for any fixed k with $1 \leq k \leq n$,

$$\det A = a_{k1}C_{k1} + a_{k2}C_{k2} + \cdots + a_{kn}C_{kn}.$$

This is called *the expansion of the determinant of A with respect to the k-th row.*

Exercise 2.4.22 Let A be a 3×3 matrix over a field F. Show that the expansion of $\det A$ with respect to any row yields the same answer you obtained in the exercise above.

Proof of the theorem. By definition, $\det A$ is the sum of products of the form

$$\operatorname{sgn}(\sigma) a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)} \qquad (*)$$

where σ runs through the elements of S_n. We claim that the sum of all expressions of the form $(*)$ for which $\sigma(k) = j$ is equal to $a_{kj}C_{kj}$. If we show this for every j with $1 \leq j \leq n$, then, summing over all j, we get the desired result.

We have

$$\sum_{\substack{\sigma\in S_n\\ \sigma(k)=j}} \operatorname{sgn}(\sigma)a_{1,\sigma(1)}a_{2,\sigma(2)}\cdots a_{n,\sigma(n)} = a_{kj}\sum_{\substack{\sigma\in S_n\\ \sigma(k)=j}} \operatorname{sgn}(\sigma)a_{1,\sigma(1)}a_{2,\sigma(2)}\cdots \widehat{a_{k,\sigma(k)}}\cdots a_{n,\sigma(n)},$$

where $\widehat{\cdot}$ indicates that the factor is removed from the product. Thus, we need to check that

$$\sum_{\substack{\sigma\in S_n\\ \sigma(k)=j}} \operatorname{sgn}(\sigma)a_{1,\sigma(1)}\cdots \widehat{a_{k,\sigma(k)}}\cdots a_{n,\sigma(n)} = (-1)^{j+k}\det(A_{kj}). \tag{\#}$$

To compute $\det(A_{kj})$, we must first re-index the rows and columns such that the indices go from 1 to $n-1$. For this, define $\phi : \{1,2,\ldots,n-1\} \to \{1,2,\ldots,\hat{k},\ldots,n\}$, by

$$\phi(j) = \begin{cases} j & \text{for } 1\leq j\leq k-1,\\ j+1 & \text{for } k\leq j\leq n-1.\end{cases}$$

Similarly, with k replaced by j, define a bijection $\psi : \{1,2,\ldots,n-1\} \to \{1,2,\ldots\hat{j},\ldots,n\}$. Let $\sigma\in S_n$ be such that $\sigma(k)=j$. Then the map $\sigma\circ\phi : \{1,2,\ldots,n-1\}\to\{1,2,\ldots n\}$ does not contain j in its image. The map $\psi^{-1}\circ\sigma\circ\phi : \{1,2,\ldots,n-1\}\to\{1,2,\ldots,n-1\}$ is well defined. In fact, the map $\{\sigma\in S_n \mid \sigma(k)=j\}\to S_{n-1}$ given by $\sigma\mapsto\psi^{-1}\circ\sigma\circ\phi$ is a bijection.

Now, recalling the definition of $\det A_{kj}$, we see that the proof of (#) follows immediately from

$$\operatorname{sgn}(\sigma) = (-1)^{j+k}\operatorname{sgn}(\psi^{-1}\circ\sigma\circ\phi). \tag{\#\#}$$

Note that ϕ and ψ are strictly increasing maps so that $\operatorname{sgn}(\psi^{-1}\circ\sigma\circ\phi)$ coincides with the sign $\prod_{\substack{1\leq i<l\leq n\\ i,l\neq k}}(\sigma(l)-\sigma(i))$. Canceling the product on both sides of (##), we are left with showing that

$$\operatorname{sgn}\left(\prod_{\substack{1\leq i<l\leq n\\ i=k \text{ or } l=k}}(\sigma(l)-\sigma(i))\right) = (-1)^{j+k}.$$

Recalling that $\sigma(k)=j$, the last product is

$$\prod_{i=1}^{k-1}(j-\sigma(i))\cdot\prod_{l=k+1}^{n}(\sigma(l)-j) = (-1)^{k-1}\prod_{l\neq k}(\sigma(l)-j).$$

Moreover, $\operatorname{sgn}\left(\prod_{l\neq k}(\sigma(l)-j)\right)$ is clearly $(-1)^{j-1}$. Altogether, we obtain $(-1)^{k-1}\cdot(-1)^{j-1} = (-1)^{j+k}$ as desired.

Exercise 2.4.23 Suppose that $A=(a_{ij})$ is a $n\times n$ matrix over a field F. Use the fact that $\det A = \det{}^tA$ to give an expansion of the determinant with respect to the k-th column.

We can now assert a theorem about inverses.

Theorem 2.4.24 If A is an $n\times n$ matrix over a field F and $\det A\neq 0$,then A has an inverse. If $\det A\neq 0$, the matrix of A^{-1} is the transpose of the cofactor matrix multiplied by the inverse of the determinant of A. That is,

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} C_{11} & C_{21} & \cdots & C_{n1}\\ C_{12} & C_{22} & \cdots & C_{n2}\\ \vdots & \vdots & \vdots & \vdots\\ C_{1n} & C_{2n} & \cdots & C_{nn}\end{pmatrix}$$

Proof. Let $C = (C_{ij})$ and consider the product ${}^tCA$. Look at the diagonal elements in this product. The j-th diagonal element is $a_{1j}C_{1j} + \cdots + a_{nj}C_{nj} = \det A$. For the off-diagonal elements, we take $\{k, m\} \subset [n]$, $k \neq m$ and consider the (m,k)-th entry of ${}^tCA$. We get $a_{1k}C_{1m} + a_{2k}C_{2m} + \cdots + a_{nk}C_{nm}$. This represents expansion of the determinant of a matrix A' that is equal to A with the exception that the m-th column has been replaced by the k-th column. By the Exercise 2.4.18, this determinant is 0. Thus, if $\det A \neq 0$, then ${}^tC/\det(A)$ is the left inverse of A.

Exercise 2.4.25 Show that ${}^tC/\det(A)$ is also a right inverse for A.

We now wish to prove that an $n \times n$ matrix A over a field F has a multiplicative inverse if and only if $\det A \neq 0$. One half of this fact was proved above. That is, if $\det A \neq 0$, then A has an inverse. The other half depends on the following important theorem.

Theorem 2.4.26 If $A = (a_{ij})$ and $B = (b_{ij})$ are $n \times n$ matrices over a field F, then

$$\det(AB) = \det(A) \cdot \det(B).$$

Proof. We first expand

$$\det(A) \cdot \det(B) = \sum_{\tau,\sigma \in S_n} \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau) a_{1,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)} b_{1,\tau(1)} b_{2,\tau(2)} \cdots b_{n,\tau(n)}.$$

We re-index the product of the b_{jk}'s as follows. It is clear that $b_{1,\tau(1)} \cdots b_{n,\tau(n)} = b_{\sigma(1),\tau(\sigma(1))} \cdots b_{\sigma(n),\tau(\sigma(n))}$. For fixed σ in S_n, we see that, as τ runs through S_n, so does $\tau \circ \sigma$. Moreover, $\operatorname{sgn}(\sigma)\operatorname{sgn}(\tau) = \operatorname{sgn}(\tau \circ \sigma)$. Hence, by letting $\rho = \tau \circ \sigma$, we have

$$\det A \cdot \det B = \sum_{\sigma \in S_n} \sum_{\rho \in S_n} \operatorname{sgn}(\rho) a_{1,\sigma(1)} b_{\sigma(1),\rho(1)} a_{2,\sigma(2)} b_{\sigma(2),\rho(2)} \cdots a_{n,\sigma(n)} b_{\sigma(n),\rho(n)}$$

By Exercise 2.4.19 this last sum will not change if we allow σ to run over all maps $\sigma : [n] \to [n]$.

For a fixed ρ, we have

$$\sum_{\sigma} a_{1,\sigma(1)} b_{\sigma(1),\rho(1)} a_{2,\sigma(2)} b_{\sigma(2),\rho(2)} \cdots a_{n,\sigma(n)} b_{\sigma(n),\rho(n)}.$$

Now we let $C = AB$ and consider $\det(C)$. Let $C = (c_{ij})$, then we have

$$\det(C) = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) c_{1,\rho(1)} c_{2,\rho(2)} \cdots c_{n,\rho(n)}.$$

Now, from the definition of C, we know that $c_{j,\rho(j)} = \sum_{k_j=1}^{n} a_{jk_j} b_{k_j,\rho(j)}$. This gives

$$\prod_{j=1}^{n} c_{j,\rho(j)} = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_n=1}^{n} a_{1k_1} b_{k_1,\rho(1)} a_{2k_2} b_{k_2,\rho(2)} \cdots a_{nk_n} b_{k_n,\rho(n)}$$

For each term $(k_1, k_2, \ldots, k_n)$ in the sum, we can define a map $\sigma : [n] \to [n]$ by $1 \mapsto k_1, 2 \mapsto k_2, \ldots, n \mapsto k_n$. Notice that there are exactly n^n such maps. Hence, all maps $\sigma : [n] \to [n]$ arise in this way. So we can index the sum by σ as we let σ run over all maps. In this way, we get

$$\begin{aligned} \det(C) &= \sum_{\rho \in S_n} \sum_{\sigma} \operatorname{sgn}(\rho) a_{1,\sigma(1)} b_{\sigma(1),\rho(1)} a_{2,\sigma(2)} b_{\sigma(2),\rho(2)} \cdots a_{n,\sigma(n)} b_{\sigma(n),\rho(n)} \\ &= \sum_{\rho \in S_n} \sum_{\sigma \in S_n} \operatorname{sgn}(\rho) a_{1,\sigma(1)} b_{\sigma(1),\rho(1)} a_{2,\sigma(2)} b_{\sigma(2),\rho(2)} \cdots a_{n,\sigma(n)} b_{\sigma(n),\rho(n)} \\ &= \det A \cdot \det B. \end{aligned}$$

Theorem 2.4.27 Let A be an $n \times n$ matrix over a field F. Show that A has a multiplicative inverse if and only if $\det A \neq 0$.

Proof. Exercise.

The next exercise illustrates several of the important properties of determinants of $n \times n$ matrices.

Exercise 2.4.28 Find the determinants of each of the three types of elementary matrices.

Exercise 2.4.29 Suppose that A is an $n \times n$ matrix over a field F.

i. If we multiply a row or column of A by a scalar c, find the determinant of the resulting matrix.

ii. Show that if we interchange two rows or two columns of A, then the determinant is $-\det A$.

iii. Show that if we add a scalar multiple of any row to any other row, or if we add a scalar multiple of any column to any other column, then the determinant remains unchanged.

Exercise 2.4.30 Let A be a $n \times n$ matrix over a field F.

i. If $\det A \neq 0$, show that the columns of A are linearly independent and hence form a basis for F^n.

ii. Do the same for the rows.

iii. If the columns of A are linearly independent, show that $\det A \neq 0$. (Hint: consider the image of the linear transformation defined by A.)

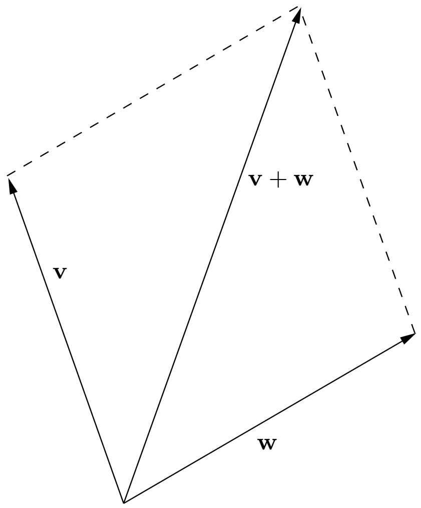
iv. Do the same for the rows.

2.5 Geometric Linear Algebra

We now wish to investigate the geometry of finite dimensional vector spaces over $\mathbb{R}$.

A point in $\mathbb{R}^n$ is represented by an n-tuple of elements of $\mathbb{R}$, written $\mathbf{p} = (p_1, \ldots, p_n)$, with each $p_i \in \mathbb{R}$. Geometrically, these are thought of as points in space, where the p_i's give the coordinates of the point $\mathbf{p}$. At the same time, we may consider n-tuples of real numbers as n-dimensional vectors giving the data of a direction and a magnitude, without specifying a base point from which this vector emanates. Thinking this way, we see that such vectors are elements of a vector space, $\mathbb{E}^n$, where elements can be written as $\mathbf{v} = (v_1, \ldots, v_n)$, with each $v_i \in \mathbb{R}$. We will consistently distinguish between the "points" of $\mathbb{R}^n$, and "vectors" in $\mathbb{E}^n$, since geometrically they are quite different. Observe that we are choosing the vectors $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_n$ from Exercise 2.1.11 as a basis for the vector space $\mathbb{E}^n$ and further that $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ may be written as the linear combination $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + \cdots + v_n\mathbf{e}_n$. Moreover, the coordinates of a point $\mathbf{p}$ are determined by a collection of mutually perpendicular coordinate axes and a distinguished point called the origin, namely, $(0, 0, \ldots, 0)$. The idea here is that vectors are free to wander around in space, and points have to stay where they are.

If we want to add two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{E}^n$, we represent $\mathbf{v} + \mathbf{w}$ as the diagonal of a parallelogram as pictured below.



and if $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$, then $\mathbf{v} + \mathbf{w} = (v_1 + w_1, \ldots, v_n + w_n)$. Continuing this idea, we specify that a direction $\mathbf{v}$ based at a point $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ is given by an element of $\mathbb{R}^n \times \mathbb{E}^n$. We also have a geometric operation on $\mathbb{R}^n \times \mathbb{E}^n$ which is to drag the point $\mathbf{p}$ along $\mathbf{v}$ to get a new point $\mathbf{q} = (q_1, q_2, \ldots, q_n)$. This geometric operation is algebraically encoded in the formula:

$$\begin{gathered}\mathbb{R}^n \times \mathbb{E}^n \to \mathbb{R}^n \\ (\mathbf{p}, \mathbf{v}) \mapsto \mathbf{p} + \mathbf{v} = \mathbf{q} \\ ((p_1, \ldots, p_n), (v_1, \ldots, v_n)) \mapsto (p_1 + v_1, \ldots, p_n + v_n) = (q_1, \cdots, q_n).\end{gathered}$$

With this in mind, we have the statements

$$\text{vector} + \text{vector} = \text{vector},$$

$$\text{point} + \text{vector} = \text{point},$$

so naturally,

$$\text{point} - \text{point} = \text{vector!}$$

To make sense of this formally, given an ordered pair of points $(\mathbf{p}, \mathbf{q}) \in \mathbb{R}^n \times \mathbb{R}^n$, there is a unique vector $\mathbf{v} \in \mathbb{E}^n$ such that $\mathbf{p} + \mathbf{v} = \mathbf{q}$. This $\mathbf{v}$ represents $\mathbf{q} - \mathbf{p}$. Of course, algebraically it is given by nothing more than $\mathbf{v} = (q_1 - p_1, \ldots, q_n - p_n)$.

We are now ready to define some geometric objects in $\mathbb{R}^n$. To describe a *line*, we need a point $\mathbf{p}_0 \in \mathbb{R}^n$ and a (nonzero) direction $\mathbf{v} \in \mathbb{E}^n$.

Definition 2.5.1 Let $\mathbf{p}_0$ be a point in $\mathbb{R}^n$ and $\mathbf{v}$ be a direction in $\mathbb{E}^n$. The *line ℓ through $\mathbf{p}_0$ in the direction* $\mathbf{v}$ is given by

$$\ell = \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} = \mathbf{p}_0 + t\mathbf{v},\ t \in \mathbb{R}\}.$$

Notice that we are using the formalism of "vector plus point equals point."

Definition 2.5.2 Suppose that $\mathbf{p}_0 \in \mathbb{R}^n$ and $\mathbf{v}$, $\mathbf{w}$ are linearly independent vectors in $\mathbb{E}^n$. The *plane through $\mathbf{p}_0$ spanned by $\mathbf{v}$ and $\mathbf{w}$* is

$$\mathscr{P} = \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} = \mathbf{p}_0 + t\mathbf{v} + s\mathbf{w},\ t, s \in \mathbb{R}.\}$$

More generally, we can use these ideas for other subsets of $\mathbb{R}^n$.

Definition 2.5.3 If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent vectors in $\mathbb{E}^n$, then we define the *k-dimensional affine subspace through $\mathbf{p}_0 \in \mathbb{R}^n$ spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$* as

$$\mathbf{H} = \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} = \mathbf{p}_0 + t_1\mathbf{v}_1 + \ldots + t_k\mathbf{v}_k, \text{ where } t_j \in \mathbb{R}, 1 \leq j \leq k\}.$$

Note: The collection of vectors $\{t_1\mathbf{v}_1 + \ldots + t_k\mathbf{v}_k, t_j \in \mathbb{R}\}$ is actually a subspace of $\mathbb{E}^n$. Thus, a k-dimensional affine subspace is constructed by taking a k-dimensional subspace of $\mathbb{E}^n$ and adding it to a point of $\mathbb{R}^n$. When $k = n - 1$, $\mathbf{H}$ is called a *hyperplane* in $\mathbb{R}^n$.

Definition 2.5.4 If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly independent vectors in $\mathbb{E}^n$, and $\mathbf{p}_0 \in \mathbb{R}^n$, we define the *$k$-dimensional parallelepiped with vertex $\mathbf{p}_0$ spanned by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$* as

$$\mathbf{P} = \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} = \mathbf{p}_0 + t_1\mathbf{v}_1 + \ldots + t_k\mathbf{v}_k, \text{ with } 0 \leq t_j \leq 1\}.$$

Note that if $k = n = 2$ then $\mathbf{P}$ is just a standard parallelogram in $\mathbb{R}^2$.

Much of the geometry that appears in this section will arise in a more general context in later chapters. We introduce only enough here to make the reader feel comfortable in $\mathbb{R}^n$. The rich interplay between $\mathbb{R}^n$ and the vector space $\mathbb{E}^n$ is what makes life interesting.

Definition 2.5.5 Let V be a vector space over a field F. A *bilinear form* $\langle \cdot, \cdot \rangle$ on V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to F$$

which satisfies linearity in both variables. That is, for all $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{w}, \mathbf{w}_1, \mathbf{w}_2 \in V$, and all $\alpha \in F$,

$$\begin{aligned}
\langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{w} \rangle &= \langle \mathbf{v}_1, \mathbf{w} \rangle + \langle \mathbf{v}_2, \mathbf{w} \rangle \\
\langle \alpha\mathbf{v}, \mathbf{w} \rangle &= \alpha \langle \mathbf{v}, \mathbf{w} \rangle \\
\langle \mathbf{v}, \mathbf{w}_1 + \mathbf{w}_2 \rangle &= \langle \mathbf{v}, \mathbf{w}_1 \rangle + \langle \mathbf{v}, \mathbf{w}_2 \rangle \\
\langle \mathbf{v}, \alpha\mathbf{w} \rangle &= \alpha \langle \mathbf{v}, \mathbf{w} \rangle.
\end{aligned}$$

The form $\langle \cdot, \cdot \rangle$ is said to be *symmetric* if $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$ for all $\mathbf{v}, \mathbf{w} \in V$.

Definition 2.5.6 Let V be a vector space over $\mathbb{R}$. The bilinear form $\langle \cdot, \cdot \rangle$ is said to be *positive definite* if $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$ for all $\mathbf{v} \in V$, and $\langle \mathbf{v}, \mathbf{v} \rangle = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

Bilinear forms and their companion Hermitian forms (over $\mathbb{C}$) will appear regularly throughout the book. For now, we assume $F = \mathbb{R}$. The main example of a positive definite symmetric bilinear form on $\mathbb{E}^n$ is the scalar product or dot product.

Definition 2.5.7 Suppose that $\mathbf{v} = (v_1, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$ are vectors in $\mathbb{E}^n$. The *scalar product* of $\mathbf{v}$ and $\mathbf{w}$ is $\langle \mathbf{v}, \mathbf{w} \rangle = v_1w_1 + \ldots + v_nw_n$. The scalar product is sometimes called the *dot product* and is denoted by $\mathbf{v} \cdot \mathbf{w}$. We will try our best to be consistent and use $\langle \, , \, \rangle$.

Exercise 2.5.8 Prove that the scalar product is a positive definite symmetric bilinear form on $\mathbb{E}^n$.

Exercise 2.5.9 Let $V = \mathbb{E}^2$. Show that the map $f : \mathbb{E}^2 \times \mathbb{E}^2 \to \mathbb{R}$ given by $f((v_1, v_2), (w_1, w_2)) = v_1w_1$ is a symmetric bilinear form that is not positive definite.

There are two concepts that arise immediately with the existence of a positive definite symmetric bilinear form. The first is the length or norm of a vector and the second is orthogonality.

Definition 2.5.10 If $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{E}^n$, then the *length* or *norm* of $\mathbf{v}$ is defined by

$$||\mathbf{v}|| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = (v_1^2 + \cdots + v_n^2)^{1/2}.$$

Exercise 2.5.11 Prove the following properties of the norm. If $\mathbf{v}, \mathbf{w} \in \mathbb{E}^n$, then:

i. $||\mathbf{v}|| \geq 0$;

ii. $||\mathbf{v}|| = 0$ iff $\mathbf{v} = \mathbf{0}$;

iii. $||\alpha\mathbf{v}|| = |\alpha|\,||\mathbf{v}||, \quad \alpha \in \mathbb{R}$;

iv. $||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}||$;

v. $\|\mathbf{v} + \mathbf{w}\|^2 + \|\mathbf{v} - \mathbf{w}\|^2 = 2\left(\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2\right)$.

There is one more fundamental inequality relating the scalar product of two vectors to their norms. We will use the famous quadratic formula from high school algebra to prove it.

Theorem 2.5.12 (Cauchy-Schwarz Inequality) Let $\mathbf{v}, \mathbf{w} \in \mathbb{E}^n$. Then $|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\|\,\|\mathbf{w}\|$.

Proof. Let λ be a real number. Then

$$\begin{aligned} 0 \leq \langle \mathbf{v} - \lambda\mathbf{w}, \mathbf{v} - \lambda\mathbf{w} \rangle &= \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{v}, \lambda\mathbf{w} \rangle - \langle \lambda\mathbf{w}, \mathbf{v} \rangle + \langle \lambda\mathbf{w}, \lambda\mathbf{w} \rangle \\ &= ||\mathbf{v}||^2 - 2\lambda\langle \mathbf{v}, \mathbf{w} \rangle + \lambda^2||\mathbf{w}||^2. \end{aligned}$$

This is a quadratic polynomial in λ which is always greater than or equal to 0. For this inequality to hold, the discriminant of this quadratic must be nonpositive. That is, we must have, in the usual notation, $b^2 - 4ac \leq 0$. With $a = \|w\|^2$, $b = -2\langle v, w \rangle$, and $c = \|v\|^2$, we get our desired inequality immediately.

Exercise 2.5.13 Prove that equality holds in the Cauchy-Schwarz Inequality iff one of the vectors is a scalar multiple of the other.

The definition of the norm leads to the usual definition of Euclidean distance between two points in $\mathbb{R}^n$. Thus, if $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^n$ then $d(\mathbf{p}_1, \mathbf{p}_2) = ||\mathbf{p}_1 - \mathbf{p}_2||$. The general study of distance is carried out in Chapter 3 where we discuss metric spaces.

Since we have a positive definite symmetric bilinear form, the concept of orthogonality (or perpendicularity) in $\mathbb{E}^n$ can be formalized as follows.

Definition 2.5.14 Let $\mathbf{v}, \mathbf{w} \in \mathbb{E}^n$. Then $\mathbf{v}$ and $\mathbf{w}$ are said to be *orthogonal* (or *perpendicular*) if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$. A set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ of vectors in $\mathbb{E}^n$ is said to be *mutually orthogonal* or *pairwise orthogonal* if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for all pairs i, j with $i \neq j$.

Exercise 2.5.15

i. Show that the vector $\mathbf{0}$ in $\mathbb{E}^n$ is orthogonal to every vector in $\mathbb{E}^n$.

ii. Show that the vectors in the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ are pairwise orthogonal, that is, $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$ if $i \neq j$, and further that $\langle \mathbf{e}_i, \mathbf{e}_i \rangle = 1$.

iii. If $\mathbf{v}$ is a nonzero vector in $\mathbb{E}^n$, show that the collection $W = \{\mathbf{w} \in \mathbb{E}^n \mid \langle \mathbf{w}, \mathbf{v} \rangle = 0\}$ is an $(n-1)$-dimensional subspace of $\mathbb{E}^n$.

iv. If $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are pairwise orthogonal non-zero vectors in $\mathbb{E}^n$, show that they form a linearly independent set in $\mathbb{E}^n$.

Now, we wish to consider the angle between two non-zero vectors $\mathbf{v}, \mathbf{w} \in \mathbb{E}^n$. If the vectors are linearly dependent, that is, $\mathbf{w} = \lambda\mathbf{v}$ for some nonzero scalar $\lambda \in \mathbb{R}$, then the angle between them is 0° if $\lambda > 0$ and 180° if $\lambda < 0$. If $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, we look at the plane through the origin spanned by $\mathbf{v}$ and $\mathbf{w}$. In this case, there are two angles associated with $\mathbf{v}$ and $\mathbf{w}$. One is less than 180° and one is greater than 180°. We take *the angle between* $\mathbf{v}$ *and* $\mathbf{w}$ to be the angle which is less then 180°.

Theorem 2.5.16 Let $\mathbf{v}$ and $\mathbf{w}$ be linearly independent vectors in $\mathbb{E}^n$. The angle between $\mathbf{v}$ and $\mathbf{w}$ is the unique solution θ to the equation

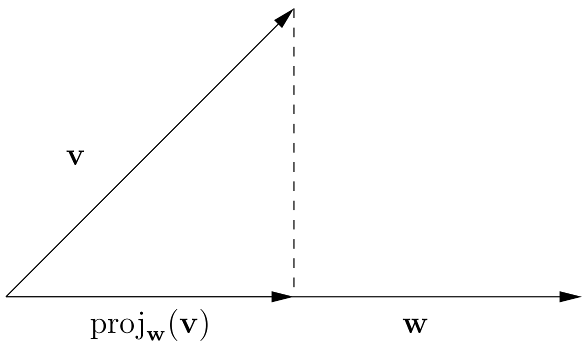
$$\cos\theta = \frac{\langle \mathbf{v}, \mathbf{w}\rangle}{||\mathbf{v}||\,||\mathbf{w}||}, \quad 0^\circ < \theta < 180^\circ. \tag{2.1}$$

Proof. This is just the law of cosines, which states that $||\mathbf{v}-\mathbf{w}||^2 = ||\mathbf{v}||^2 + ||\mathbf{w}||^2 - 2||\mathbf{v}||\,||\mathbf{w}||\cos\theta$. (Draw a picture!)

One of the most important procedures in geometric linear algebra is producing vectors that are orthogonal to a given set of vectors. This leads us to define one of the most important operations on vectors in $\mathbb{E}^n$.

Definition 2.5.17 Let $\mathbf{v}$ and $\mathbf{w}$ be linearly independent vectors in $\mathbb{E}^n$. The *projection of* $\mathbf{v}$ *onto* $\mathbf{w}$ is defined by

$$\text{proj}_{\mathbf{w}}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{w}\rangle}{||\mathbf{w}||}\,\frac{\mathbf{w}}{||\mathbf{w}||}. \tag{2.2}$$



Observe that $\mathbf{w}/\|\mathbf{w}\|$ is a unit vector in the direction of $\mathbf{w}$. Hence, the projection of $\mathbf{v}$ on $\mathbf{w}$ has the same direction as $\mathbf{w}$ if $\langle \mathbf{v}, \mathbf{w}\rangle > 0$, and the direction of $-\mathbf{w}$ if $\langle \mathbf{v}, \mathbf{w}\rangle < 0$. Of course, if $\langle \mathbf{v}, \mathbf{w}\rangle = 0$ then $\mathbf{v}$ and $\mathbf{w}$ are orthogonal and the projection is just the zero vector. Note that the norm of $\text{proj}_{\mathbf{w}}(\mathbf{v})$ is $\|\text{proj}_{\mathbf{w}}(\mathbf{v})\| = \frac{|\langle \mathbf{v},\mathbf{w}\rangle|}{||\mathbf{w}||}$.

Exercise 2.5.18 Show that $\mathbf{v} - \text{proj}_{\mathbf{w}}(\mathbf{v})$ is orthogonal to $\mathbf{w}$.

The exercise above allows us to construct a vector orthogonal to a single vector in $\mathbb{E}^n$. Next, let us consider the problem of constructing a vector orthogonal to a collection of $n-1$ linearly independent vectors.

First, let's look in 2 dimensions. Given a vector $\mathbf{v} = (v_1, v_2) \in \mathbb{E}^2$, $v \neq 0$, we set $\mathbf{v}^\perp = (v_2, -v_1)$. Then, $\langle \mathbf{v}, \mathbf{v}^\perp\rangle = 0$, that is, these vectors are orthogonal.

Now, let's look in 3 dimensions. In $\mathbb{E}^3$, we have the special notion of the *cross product* of two vectors. This can be defined by using determinants, but for the moment, we define it as follows. The cross product, ×, is a map that takes two vectors in $\mathbb{E}^3$ and produces a third vector in $\mathbb{E}^3$. If $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ in $\mathbb{E}^3$, the formula for $\mathbf{v}\times\mathbf{w}$ is given by $\mathbf{v}\times\mathbf{w} = (v_2w_3 - v_3w_2, v_3w_1 - v_1w_3, v_1w_2 - v_2w_1)$. Observe that, unlike the scalar product, the cross product produces a vector, not a scalar.

Exercise 2.5.19

i. Show that $\mathbf{v}\times\mathbf{w} \neq \mathbf{0}$ if and only if $\mathbf{v}$ and $\mathbf{w}$ are linearly independent.

ii. Show that $\langle \mathbf{v}, \mathbf{v}\times\mathbf{w}\rangle = \langle \mathbf{w}, \mathbf{v}\times\mathbf{w}\rangle = \mathbf{0}$.

iii. Show that $\mathbf{w}\times\mathbf{v} = -(\mathbf{v}\times\mathbf{w})$.

Exercise 2.5.20

i. Show that $||\mathbf{v} \times \mathbf{w}|| = ||\mathbf{v}|| \, ||\mathbf{w}|| \sin\theta$, where θ is the angle between $\mathbf{v}$ and $\mathbf{w}$.

ii. Show that $\|\mathbf{v} \times \mathbf{w}\|$ is the area of the parallelogram spanned by $\mathbf{v}$ and $\mathbf{w}$.

We have shown that given a non-zero vector $\mathbf{v} \in \mathbb{E}^2$, we can find a vector $\mathbf{v}^\perp$ which is non-zero and orthogonal to $\mathbf{v}$. Obviously, the pair $\{\mathbf{v}, \mathbf{v}^\perp\}$ is a basis for $\mathbb{E}^2$. Next, given two linearly independent vectors $\mathbf{v_1}, \mathbf{v_2} \in \mathbb{E}^3$, we constructed $\mathbf{v_3} = \mathbf{v_1} \times \mathbf{v_2}$ orthogonal to both of the original vectors. The set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a basis for $\mathbb{E}^3$. Let us investigate how this process can be generalized to $\mathbb{E}^n$.

Determinants can be used in an interesting way in the aforegoing process. For example, given a vector $\mathbf{v} = (v_1, v_2) \in \mathbb{E}^2$, we write the matrix

$$\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 \\ v_1 & v_2 \end{pmatrix}.$$

Without being concerned about vectors in the first row and numbers in the second row, we can take the “determinant” of this matrix by expanding according to the first row and obtain $v_2\mathbf{e}_1 - v_1\mathbf{e}_2$, which is the vector $\mathbf{v}^\perp = (v_2, -v_1)$. Similarly, if we have linearly independent vectors $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$ in $\mathbb{E}^3$, we can consider

$$\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}.$$

Taking the “determinant” of this matrix by expanding according to the first row, we get $(v_2w_3 - v_3w_2)\mathbf{e}_1 + (v_3w_1 - v_1w_3)\mathbf{e}_2 + (v_1w_2 - v_2w_1)\mathbf{e}_3$, which is $\mathbf{v} \times \mathbf{w}$.

We can generalize this to n dimensions.

Theorem 2.5.21 Let $V = \mathbb{E}^n$. Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$ is a collection of linearly independent vectors in V, where $\mathbf{v}_j = (v_{j1}, v_{j2}, \ldots, v_{jn})$. Consider the matrix

$$\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \\ v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(n-1)1} & v_{(n-1)2} & \cdots & v_{(n-1)n} \end{pmatrix}.$$

Let $\mathbf{v}$ be the vector obtained by taking the “determinant” of this matrix with respect to the first row. Then $\mathbf{v}$ is nonzero and is orthogonal to each of the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$. Moreover, the collection $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}, \mathbf{v}\}$ is a basis for V.

Proof. The vector $\mathbf{v}$ obtained by expanding with respect to the first row is simply the vector of cofactors $\mathbf{v} = (C_{11}, C_{12}, \ldots, C_{1n})$. If we replace the first row by $\mathbf{v}$, we obtain the matrix

$$A = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{(n-1)1} & v_{(n-1)2} & \cdots & v_{(n-1)n} \end{pmatrix}.$$

We prove first that the vector $\mathbf{v}$ is orthogonal to the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. To see this, choose i with $1 \leq i \leq n-1$. Then $\langle \mathbf{v}_i, \mathbf{v} \rangle = v_{i1}C_{11} + v_{i2}C_{12} + \cdots + v_{in}C_{1n}$. This is the determinant of the matrix obtained by replacing the first row of A by the vector $\mathbf{v}_i$. By Lemma 2.4.17, this determinant is 0. Next, we establish that $\mathbf{v}$ is not the $\mathbf{0}$ vector. To do this, we note that if $C_{1j} = 0$ for all j, and we replace $\mathbf{v}$ by any vector $\mathbf{w}$ in V, then the determinant of the resulting matrix will be 0. But since $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$ form a linearly independent set in V, we can extend this set to a basis in V with a vector $\mathbf{v}_n$. Replacing the vector $\mathbf{v}$ by $\mathbf{v}_n$, we get a matrix whose determinant is not 0. This is a contradiction.

Given $n-1$ linearly independent vectors in $\mathbb{E}^n$, the above theorem produces a vector that is orthogonal to each of these vectors. If the original set of vectors was mutually orthogonal, the new set of n vectors will be mutually orthogonal. This mutually orthogonal set of vectors will be a basis for $\mathbb{E}^n$, however, because we have norms at our disposal, we can go one step further with the following definition.

Definition 2.5.22 Let $V = \mathbb{E}^n$. We say that $\mathbf{v} \in V$ is a *normalized vector* or *unit vector*, if $\|v\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = v_1^2 + v_2^2 + \ldots + v_n^2 = 1$. An *orthonormal set* in V is a collection $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ of linearly independent vectors in V such that $\langle \mathbf{v}_i \mid \mathbf{v}_j \rangle = 0$ if $i \neq j$, and $\langle \mathbf{v}_i \mid \mathbf{v}_i \rangle = 1$ for all $i, j = 1, \ldots, k$. If $k = n$, the orthonormal set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is called an *orthonormal basis* for V.

Example 2.5.23 The collection $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is an orthonormal basis for $\mathbb{E}^n$.

Exercise 2.5.24 If, in the previous theorem, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$ forms an orthonormal set in V, show that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}, \mathbf{v}\}$ is an orthonormal basis for V. (Hint: Consider A^tA and AA^{-1}.)

Exercise 2.5.25 Suppose that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$ are linearly independent vectors in $\mathbb{E}^n$. Take a point $\mathbf{p}_0$ in $\mathbb{R}^n$ and consider the hyperplane $\mathbf{H}$ through $\mathbf{p}_0$ spanned by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$. If $\mathbf{v}$ is the vector determined in Theorem 2.5.21, show that $\mathbf{H} = \{\mathbf{p} \in \mathbb{R}^n \mid \langle \mathbf{p} - \mathbf{p}_0, \mathbf{v} \rangle = 0\}$. Specialize this to obtain formulas for a line in $\mathbb{R}^2$ and for a plane in $\mathbb{R}^3$.

Exercise 2.5.26 Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$ be linearly independent vectors in $\mathbb{E}^n$. Let $\mathbf{p}_0$ be a point of $\mathbb{R}^n$. Let $\mathbf{H}$ be the hyperplane through $\mathbf{p}_0$ spanned by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$. If $\mathbf{p}$ is any point in $\mathbb{R}^n$, show that the distance from $\mathbf{p}$ to $\mathbf{H}$, that is, $\inf\{\|\mathbf{p} - \mathbf{q}\| \mid \mathbf{q} \in \mathbf{H}\}$, is given by the length of the vector $\mathrm{proj}_{\mathbf{v}}(\mathbf{p} - \mathbf{p}_0)$ where $\mathbf{v}$ is the vector obtained in Theorem 2.5.21. Specialize this to obtain formulas for the distance from a point to a line in $\mathbb{R}^2$ and from a point to a plane in $\mathbb{R}^3$.

Exercise 2.5.27

i. Find a formula for the distance from a point to a line in $\mathbb{R}^n$.

ii. Find the distance between two nonintersecting lines in $\mathbb{R}^3$.

iii. Find the distance between two nonintersecting planes in $\mathbb{R}^5$.

Continuing our theme of attempting to produce vectors that are orthogonal to each other or to other sets of vectors, we finish this section with a general procedure for turning sets of linearly independent vectors into sets of mutually orthogonal vectors. This process is known as the *Gram-Schmidt Orthogonalization* process. Specifically, given a set of k linearly independent vectors in $\mathbb{E}^n$, the Gram-Schmidt Orthogonalization process produces a set of k mutually orthogonal vectors that span the same subspace as the original k vectors. Moreover, the Gram-Schmidt process allows us to extend such a mutually orthogonal set to a mutually orthogonal basis for $\mathbb{E}^n$. As a last step, dividing each vector by its norm will produce an orthonormal basis for $\mathbb{E}^n$.

We begin with a set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ of linearly independent vectors and proceed as follows. Let $\tilde{\mathbf{v}}_1 = \mathbf{v}_1$. We continue to find vectors $\tilde{\mathbf{v}}_k$ by taking $\mathbf{v}_k$ and subtracting the projections on the vectors already constructed. More explicitly, we let

$$\begin{aligned} \tilde{\mathbf{v}}_2 &= \mathbf{v}_2 - \mathrm{proj}_{\tilde{\mathbf{v}}_1}(\mathbf{v}_2), && (2.3)\\ \tilde{\mathbf{v}}_3 &= \mathbf{v}_3 - \mathrm{proj}_{\tilde{\mathbf{v}}_1}(\mathbf{v}_3) - \mathrm{proj}_{\tilde{\mathbf{v}}_2}(\mathbf{v}_3), && (2.4)\\ &\vdots && (2.5)\\ \tilde{\mathbf{v}}_k &= \mathbf{v}_k - \sum_{i=1}^{k-1} \mathrm{proj}_{\tilde{\mathbf{v}}_i}(\mathbf{v}_k). && (2.6) \end{aligned}$$

It is easy to check that this set of vectors is pairwise orthogonal.

Exercise 2.5.28 Check it, and, in addition, show that $\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_k$ span the same subspace as $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$.

On the other hand, suppose that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is a mutually orthogonal set of nonzero vectors with $k < n$. We can complete this to a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$ using Theorem 2.1.36. If we now apply the Gram-Schmidt orthogonalization process to this basis, we get a mutually orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \tilde{\mathbf{v}}_{k+1}, \ldots, \tilde{\mathbf{v}}_n\}$, which does not alter our original mutually orthogonal set.

Exercise 2.5.29 Prove this last statement.

The last step is to try to work with vectors whose norms are equal to one. Note that if $\mathbf{v}$ is a nonzero vector in $\mathbb{E}^n$, then $\mathbf{v}$ can be converted to a vector of norm 1, that is, a unit vector, by dividing $\mathbf{v}$ by $\|\mathbf{v}\|$.

Exercise 2.5.30 Consider the vectors $\mathbf{v}_1 = (1, 1, -1, 0)$, $\mathbf{v}_2 = (1, 0, 0, -1)$, and $\mathbf{v}_3 = (0, 1, 1, 1)$ in $\mathbb{E}^4$.

i. Use the Gram-Schmidt orthogonalization process on these three vectors to produce a set of three mutually orthogonal vectors that span the same subspace.

ii. Extend the set of three vectors produced in part *i* to a mutually orthogonal basis for $\mathbb{E}^4$.

iii. Normalize your basis so that it becomes an orthonormal basis for $\mathbb{E}^4$.

Exercise 2.5.31 Show that, given linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in $\mathbb{E}^n$, we can transform this collection into an orthonormal set $\{\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k\}$, which spans the same subspace. In addition, the set $\{\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k\}$ can be completed to an orthonormal basis for $\mathbb{E}^n$.

2.6 Independent Projects

2.6.1 Groups In Chapter **??**, we have introduced commutative rings with 1, integral domains, and fields. In this chapter, we have introduced vector spaces and algebras over a field, and we have defined groups in the context of the symmetric group. This project gives more details and exercises about groups.

Definition 2.6.1 Let G be a set with an ILC $\circ : G \times G \to G$ (see Definition 0.2.21). The pair $(G, \circ)$ is a *group* if,

a. for all $a, b, c \in G$, we have $(a \circ b) \circ c = a \circ (b \circ c)$ (Associativity);

b. there is an *identity element* $e \in G$ such that for each $a \in G$, $e \circ a = a \circ e = a$ (Identity); and

c. for every element $a \in G$, there is an element $a^{-1} \in G$ called the *inverse of* a such that $a \circ a^{-1} = e = a^{-1} \circ a$ (Inverses).

Here is a set of elementary exercises about groups.

Exercise 2.6.2 Suppose that $(G, \circ)$ is a group. Prove the following statements.

i. Show that the identity is unique.

ii. Show that inverses are unique.

iii. If a, b are elements of G, show that the equations $a \circ x = b$ and $x \circ a = b$ can be solved uniquely.

iv. If $a, b, c \in G$ and $a \circ b = a \circ c$, show that $b = c$.

v. If $a, b \in G$, show that $(a \circ b)^{-1} = b^{-1} \circ a^{-1}$.

Remark 2.6.3 It is often convenient to omit the ILC $\circ$ when writing a product. Thus we can write $a \circ b$ as ab. In particular, if $(G, \circ)$ is a group and k is a positive integer, we can write a^k for $a \circ a \circ \cdots \circ a$ (k times). We can also write $a^{-k} = (a^{-1})^k$. Given all of this, we still occasionally write $a \circ b$ when we feel it is useful.

Examples 2.6.4 Here are some elementary examples of groups:

1. the ordinary integers $(\mathbb{Z}, +)$;
2. the integers $\{1, -1\}$ under multiplication;
3. the rational numbers $(\mathbb{Q}, +)$;
4. the nonzero rational numbers $(\mathbb{Q}^\times, \cdot)$; and
5. for an integer $n \geq 2$, $(\mathbb{Z}_n, +)$.

Exercise 2.6.5 Decide whether or not the following are groups:

i. $(\mathbb{N}, +)$;

ii. $(\{0\}, +)$;

iii. $(\mathbb{Z}, \cdot)$;

iv. $(\mathbb{Q}, \cdot)$;

v. $\mathbb{N}$ with ILC $a \circ b = a^b$;

vi. $(\mathbb{Z}, -)$;

vii. $(\{1\}, \cdot)$;

viii. $(\mathbb{Z}_n \setminus \{\bar{0}\}, \cdot)$.

Here are some more complicated examples.

Example 2.6.6 The rotations of a regular n-gon about the center form a group denoted by C_n. The ILC in this case is the composition of rotations. Let r be a counterclockwise rotation through $(360/n)^\circ$, that is, $2\pi/n$ radians. The rotation group of a regular n-gon consists of I (the identity rotation), r, r^2 (the counterclockwise rotation through $(2 \cdot 360/n)^\circ$), r^3, $\ldots$, r^{n-1}. Note that $r^n = I$.

Examples 2.6.7 1. The symmetric group S_n of permutations of n objects is a group whose ILC is composition of functions.

2. The alternating group A_n of even permutations in S_n is a group under composition of functions.

Examples 2.6.8 1. The set $M_n(F)$ of $n \times n$ matrices over a field F with addition as the ILC is a group.

2. The set $GL_n(F) = \{A \in M_n(F) \mid \det A \neq 0\}$ is a group under multiplication of matrices. The group $GL_n(F)$ is called the $n \times n$ *general linear group* over F.

Remark 2.6.9 Many of the groups we encounter arise from considering a single operation in a structure that has more than one operation. For example, forgetting the multiplication in a ring or the scalar multiplication in a vector space and remembering only the addition leaves us with a group. In such situations, we will continue to denote the ILC by "+," the identity element by "0," and the inverse of an element a by "$-a$." We will also write ka instead of a^k for $a \in A$, $k \in \mathbb{Z}$.

Examples 2.6.10 1. If R is a ring, then $(R, +)$ is a group.

2. If V is a vector space over a field F, then $(V, +)$ is a group.

Definition 2.6.11 Suppose that $(G, \circ)$ is a group. If the number of elements in G is finite, we write $|G| = n$ where n is the number of elements of G, and we call n the *order of* G. If G has an infinite number of elements, we say that G is an *infinite group*.

Exercise 2.6.12 *i.* Find the orders of the groups $\mathbb{Z}_n$, C_n, S_n, and A_n.

ii. Find the orders of $M_n(F)$ and $GL_n(F)$ when $F = \mathbb{Z}_p$.

Many of the examples of groups above, including all of the ones in Examples 2.6.4 and 2.6.10, fit a category of groups that are particularly easy to work with, that is, abelian or commutative groups.

Definition 2.6.13 A group $(G,\circ)$ is *abelian* or *commutative* if $a \circ b = b \circ a$ for all $a, b \in G$.

In the examples above, the groups S_n for $n \geq 3$, A_n for $n \geq 4$, and $GL_n(F)$ for $n \geq 2$ are not abelian. The simplest groups to analyze are cyclic groups.

Definition 2.6.14 If G is a group, and there exists an element $a \in G$ such that $G = \{a^k \mid k \in \mathbb{Z}\}$, then G is called the *cyclic group* generated by a.

Note that we do not assume in the definition of a cyclic group that the various powers a^k are distinct.

Exercise 2.6.15

i. If G is a cyclic group of order n generated by a, show that $a^n = e$.

ii. Show that a cyclic group of order n generated by a is also generated by a^k when k and n are relatively prime.

iii. Show that cyclic groups are abelian and that in a cyclic group of order n, $(a^k)^{-1} = a^{n-k}$ for $1 \leq k \leq n-1$.

Examples 2.6.16

1. The integers, $\mathbb{Z}$, under addition form an infinite cyclic group generated by 1.
2. The even integers, $2\mathbb{Z}$, form an infinite cyclic group under addition generated by 2.
3. Let p be a prime, and let $G = \{p^k \mid k \in \mathbb{Z}\}$. Then G is an infinite cyclic group under multiplication generated by p.
4. The group C_n of rotations of a regular n-gon is a cyclic group of order n generated by r.
5. The group $(\mathbb{Z}_n, +)$ is a cyclic group of order n generated by $\bar{1}$.

Exercise 2.6.17 Show that the group of rational numbers under addition is not a cyclic group.

Exercise 2.6.18 Can a cyclic group be uncountable?

Definition 2.6.19 Let G be a group and let $a \in G$. If there exists $m \in \mathbb{N}$ such that $a^m = e$, we say that a has *finite order in* G. If no such m exists, we say that a has *infinite order in* G. If a is of finite order, then $n = \min\{m \in \mathbb{N} \mid a^m = e\}$ is called the *order of* a *in* G.

Note that order of an element a is 1 if and only if a is the identity element.

Exercise 2.6.20 *i.* If G is a finite group show that the order of an element in G is less then or equal to the order of G.

ii. Find the order of the elements in C_{12} and S_4.

iii. Does there exist a group with elements of both finite and infinite order?

We next need the notion of a subgroup.

Definition 2.6.21 Let $(G, \circ)$ be a group. A non-empty subset H of G is a *subgroup* if the pair $(H, \circ)$ is a group. If H is a subgroup of G, with $H \neq G$ and $H \neq \{e\}$, we say that H is a *proper subgroup of* G.

Remark 2.6.22 Thus, we require that $\circ$ is an ILC on H, that the identity of the group e is an element of H, and if an element a is in H, then a^{-1} is in H. Observe that associativity is automatic for the elements of H because associativity holds for the elements of G. In this situation, we say that associativity is *inherited* from G.

Exercise 2.6.23 Show that a subgroup of a cyclic group is a cyclic group. That is, if G is the cyclic group generated by a and H is a subgroup of G, then there exists an element $k \in \mathbb{N}$ such that a^k generates H. This means that for every $h \in H$, there exists $j \in \mathbb{Z}$ such that $(a^k)^j = h$.

Exercise 2.6.24 Show that every element in a group generates a cyclic subgroup.

Examples 2.6.25

1. The groups $\{e\}$ and G are subgroups of G.
2. The group $(2\mathbb{Z}, +)$, the even integers, is a subgroup of $(\mathbb{Z}, +)$, the additive group of integers.
3. The group $(n\mathbb{Z}, +)$, the integer multiples of n, is a subgroup of $(\mathbb{Z}, +)$.
4. The group $\{I, r^3, r^6\}$ is a subgroup of C_9, the group of rotations of a regular 9-gon. Note that this is an example of Exercise 2.6.23.
5. The groups $\{I, (12)\}$, $\{I, (23)\}$ and $\{I, (13)\}$ are subgroups of S_3.
6. The set $SL_n(F) = \{A \in GL_n(F) \mid \det(A) = 1\}$ is a subgroup of $GL_n(F)$. This is called the $n \times n$ *special linear group* over F.
7. The group A_n is a subgroup of S_n.

Exercise 2.6.26

i. Suppose that G is a group and H is a nonempty subset of G. Show that H is a subgroup of G iff for every $a, b \in H$, $a^{-1}b \in H$.

ii. Suppose that G is a finite group and H is a nonempty subset of G. Show that H is a subgroup of G iff H is closed under multiplication.

Exercise 2.6.27 Let G be a group, and let H be a subgroup of G. Fix $a \in G$, and define $aHa^{-1} = \{aba^{-1} \mid b \in H\}$. Show that aHa^{-1} is a subgroup of G.

Exercise 2.6.28 Find all subgroups of C_{12} and S_4.

Definition 2.6.29 Suppose that G_1 and G_2 are groups. A map $\phi : G_1 \to G_2$ is a *homomorphism* if $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in G_1$. That is, a homomorphism preserves multiplication. In general, a homomorphism of algebraic structures is a map that preserves all the operations of that structure. For example, a homomorphism of rings (or fields) preserves addition *and* multiplication.

If a homomorphism ϕ is a surjection, then ϕ is called an *epimorphism.* If a homomorphism ϕ is an injection, then ϕ is called a *monomorphism.* If a homomorphism ϕ is a bijection, then ϕ is called an *isomorphism.*

Group homomorphisms have many of the properties of linear transformations.

Proposition 2.6.30 Let G_1 and G_2 be groups and $\phi : G_1 \to G_2$ a homomorphism.

i. If e_1 is the identity in G_1, then $\phi(e_1) = e_2$, the identity in G_2.

ii. If $a \in G_1$, then $\phi(a^{-1}) = (\phi(a))^{-1}$.

Proof. You do it.

Example 2.6.31 Let $G_1 = G_2 = (\mathbb{Z}, +)$. For $n \in \mathbb{N}$ with $n \geq 2$, define $\phi_n(a) = na$. Then ϕ_n is a homomorphism, and in fact ϕ_n is a monomorphism. If we let $G_2 = (n\mathbb{Z}, +)$, then ϕ_n is an isomorphism.

Exercise 2.6.32 *i.* Let G be a finite cyclic group of order n generated by a. Define $\phi : G \to G$ by $\phi(a^j) = a^{2j}$. Show that ϕ is a homomorphism. Determine those values of n for which ϕ is a monomorphism, epimorphism, or isomorphism.

ii. Let n be a natural number with $n \geq 2$. Define $\phi : \mathbb{Z} \to \mathbb{Z}_n$ by $\phi(k) = k \pmod{n}$. Show that ϕ is an epimorphism.

For groups, we have a situation which is analogous to the kernel of a linear transformation in vector spaces.

Definition 2.6.33 Let G_1, G_2 be groups and $\phi : G_1 \to G_2$ a homomorphism. The *kernel of* ϕ is the subset of G defined by

$$\ker\phi = \{x \in G_1 \mid \phi(x) = e_2, \text{ the identity in } G_2\}.$$

Exercise 2.6.34

i. Show that $\ker\phi$ is a subgroup of G_1.

ii. Show that $\phi(G_1)$ is a subgroup of G_2.

iii. Find $\ker\phi$ in the homomorphisms given in Example 2.6.31 and Exercise 2.6.32 above.

The kernel of a homomorphism $\phi : G_1 \to G_2$ is a subgroup with special properties. In particular, if $a, b \in G_1$ and $b \in \ker\phi$, then $aba^{-1} \in \ker\phi$.

Definition 2.6.35 Let G be a group and H a subgroup of G. The subgroup H is called a *normal subgroup of* G if, for each $a \in G$, we have $aHa^{-1} = H$.

Exercise 2.6.36

i. Let G_1 and G_2 be groups, and let $\phi : G_1 \to G_2$ be a homomorphism. Show that $\ker\phi$ is a normal subgroup of G_1.

ii. Show that any subgroup of an abelian group is a normal subgroup.

iii. Show that A_n is a normal subgroup of S_n.

iv. Show that if G is a finite group and H is a subgroup of G, then H is a normal subgroup of G if and only if $aHa^{-1} \subseteq H$.

We finish this project by considering some important subgroups of $GL_n(F) = \{x \in M_n(F) \mid \det x \neq 0\}$.

Exercise 2.6.37 Prove that $GL_n(F)$ is a group with the ILC given by multiplication of matrices.

Exercise 2.6.38 *i.* Show that the following are subgroups of $GL_n(F)$.

ii. Determine which of the following are normal subgroups of $GL_n(F)$.

1. The $n \times n$ *special linear group over* F,

$$SL_n(F) = \{x \in GL_n(F) \mid \det x = 1\}.$$

2. The *upper triangular* matrices in $GL_n(F)$,
$$B = \{(b_{ij}) \in GL_n(F) \mid b_{ij} = 0 \text{ if } i > j\}.$$
3. The *upper triangular unipotent* matrices in $GL_n(F)$,
$$N = \{(b_{ij}) \in B \mid b_{jj} = 1,\ \ j = 1, \ldots, n\}.$$
4. The $n \times n$ *diagonal* matrices in $GL_n(F)$,
$$A = \{(a_{ij} \mid a_{ij} = 0 \text{ if } i \neq j\}.$$
5. The $n \times n$ *orthogonal group over* F,
$$O(n, F) = \{x \in GL_n(F) \mid x^t x = I\}.$$

2.6.2 Orthogonal Transformations in Euclidean Space The orthogonal transformations in $\mathbb{R}^n$ play an important role in analysis.

Definition 2.6.39 An *orthogonal transformation* on $\mathbb{R}^n$ is a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ with the property that $\langle T\mathbf{v}, T\mathbf{w}\rangle = \langle \mathbf{v}, \mathbf{w}\rangle$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. We say that T preserves the bilinear form $\langle \cdot\, , \cdot \rangle$.

Exercise 2.6.40 Show that an orthogonal linear transformation is distance preserving. That is, for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, $\|T\mathbf{v} - T\mathbf{w}\| = \|\mathbf{v} - \mathbf{w}\|$. In particular, for any $\mathbf{v} \in \mathbb{R}^n$, we have $\|T\mathbf{v}\| = \|\mathbf{v}\|$.

Exercise 2.6.41 Let M be the matrix of a linear transformation on $\mathbb{R}^n$ relative to the standard basis. Show that, for any pair of vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, we have $\langle M\mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{v}, {}^tM\mathbf{w}\rangle$. Show in particular that, if M is the matrix of an orthogonal linear transformation relative to the standard basis, then $M^tM = {}^tMM = I$. That is, $M \in O(n, \mathbb{R})$ (see Exercise 2.6.38).

Example 2.6.42 Let
$$M = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$
be the matrix of an orthogonal linear transformation on $\mathbb{E}^2$ relative to the standard basis, that is, $M \in O(2, \mathbb{R})$. From the exercise above, it follows that $a_{11}^2 + a_{21}^2 = a_{12}^2 + a_{22}^2 = a_{11}^2 + a_{12}^2 = a_{21}^2 + a_{22}^2 = 1$. Also, $a_{11}a_{12} + a_{21}a_{22} = a_{11}a_{21} + a_{12}a_{22} = 0$. It is now immediate that there is some θ with $0 \leq \theta < 2\pi$, such that
$$M = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$
or
$$M = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Proposition 2.6.43 If $M \in O(n, \mathbb{R})$, then $\det M = \pm 1$.

Proof. This follows from the fact that $\det {}^tM = \det M$, so that if $M \in O(n, \mathbb{R})$, then $(\det M)^2 = \det M \det {}^tM = \det M^tM = \det I = 1$.

The collection $SO(n, \mathbb{R}) = \{M \in O(n, \mathbb{R}) \mid \det M = 1\}$ is a subgroup of $O(n, \mathbb{R})$ called the *special orthogonal group*. The elements of $SO(n, \mathbb{R})$ are called generalized rotations.

The above proposition might lead one to believe that the special orthogonal group makes up half of the orthogonal group. This is good intuition, and it can be made precise in the following manner. Let

$$R_0 = \begin{pmatrix} -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

which may be interpreted geometrically as a reflection through the hyperplane $H = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n \mid x_1 = 0\}$. Note that $R_0 \in O(n, \mathbb{R})$ and that $\det R_0 = -1$. With this particular orthogonal transformation so identified, we may write

$$O(n, \mathbb{R}) = SO(n, \mathbb{R}) \cup R_0 \cdot SO(n, \mathbb{R}),$$

where $R_0 \cdot SO(n, \mathbb{R}) = \{R_0 M \in O(n, \mathbb{R}) \mid M \in SO(n, \mathbb{R})\}$. The elements of the set $R_0 \cdot SO(n, \mathbb{R})$ are called generalized reflections and have determinant equal to -1.

Exercise 2.6.44

i. If $M \in O(2, \mathbb{R})$ and $M = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, show that M is a counter-clockwise rotation around the origin through an angle θ. Obviously, $\det M = 1$.

ii. If $M \in O(2, \mathbb{R})$ and $M = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} -\cos\theta & \sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, show that M is a reflection across a line through the origin. Here, $\det M = -1$. In particular, determine the angle that the line of reflection makes with the positive x-axis.

iii. Finally, show that $SO(2, \mathbb{R}) = \left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \,\middle|\, 0 \le \theta < 2\pi \right\}$ and that $O(2, \mathbb{R})$ is the union of the matrices from parts *i* and *ii.*

Exercise 2.6.45 Let $n \ge 3$ be a natural number, and consider the regular n-gon centered at the origin in $\mathbb{R}^2$ with one of the vertices at the point $(1, 0)$. We define the *dihedral group* D_{2n} to be the subgroup of $O(2, \mathbb{R})$ that sends this n-gon to itself.

i. What is the order of D_{2n}?

ii. Write down the elements of D_{2n} as matrices with respect to the standard basis.

iii. Identify C_n as a subgroup of D_{2n}.

It is possible to write some “rotations” and “reflections” in $\mathbb{R}^n$ very explicitly. For example, in $O(4, \mathbb{R})$ the matrix

$$M = \begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\phi & -\sin\phi \\ 0 & 0 & \sin\phi & \cos\phi \end{pmatrix}$$

is the composition of a rotation around the x_1-x_2 plane through the angle ϕ with a rotation around the x_3-x_4 plane through the angle θ. This is an example of a generalized rotation.

As an example of a generalized reflection, if we have $n - 1$ pairwise orthogonal non-zero vectors in $\mathbb{R}^n$, we can produce an n-th non-zero vector v orthogonal to the original vectors by Theorem 2.5.21. In this case reflection through the hyperplane spanned by the original set of $n - 1$ orthogonal vectors is given by $T_v(w) = w - 2\frac{\langle v,w\rangle}{\langle v,v\rangle} v$.

Exercise 2.6.46

i. Show that T_v is an orthogonal transformation that is a generalized reflection on $\mathbb{R}^n$.

ii. Find the matrix of T_v relative to the standard basis.

When $n = 3$, we can be quite precise about the nature of the elements of $O(3, \mathbb{R})$. If T is an element of $SO(3, \mathbb{R})$, then we can find a line through the origin such that T is a rotation around that line. Furthermore, any reflection in $O(3, \mathbb{R})$ can be written as a rotation around a line in $\mathbb{R}^3$ through the origin combined with reflection through the origin, that is, multiplication by $-I$.

We conclude this project with the symmetry groups of the regular polyhedra in $\mathbb{R}^3$. The regular polyhedra are the regular tetrahedron, the cube, the regular octahedron, the regular dodecahedron, and the regular icosahedron. Since the octahedron is dual to the cube and the icosahedron is dual to the dodecahedron, we need only work with the tetrahedron, the cube, and the dodecahedron. In each case, we can obtain an upper bound on the number of symmetries by proceeding as follows. Each vertex must be mapped to a vertex and the sides adjacent to a vertex must remain adjacent to that vertex, although they can be permuted after the symmetry map is applied. For a tetrahedron, this gives an upper bound of $4 \times 6 = 24$ possible symmetries. In this case, the symmetries are in one to one correspondence with the permutations of the vertices, and the symmetry group is S_4.

Exercise 2.6.47

i. Write the 24 orthogonal matrices that represent the symmetries of the tetrahedron relative to the standard basis.

ii. Show that the rotations in the symmetry group of the tetrahedron form the group A_4.

iii. Which of the reflections in the symmetry group of a regular tetrahedron can be realized as reflections through a plane?

Now consider the symmetry group of the cube with vertices at the eight points $(\pm 1, \pm 1, \pm 1)$. The rotations of this cube can be realized as rotations around the x-axis, rotations around the y-axis, rotations around the z-axis, rotations around the four diagonals of the cube, and finally rotations around the six lines through the origin and the midpoints of the opposite edges.

Exercise 2.6.48

i. Show that rotations around a coordinate axis have order 4, rotations around a diagonal have order 3, and rotations around a line through the origin connecting the midpoints of opposite edges have order 2.

ii. Write the matrices of the 24 rotational symmetries of the cube.

For example, rotations around the z-axis are I,

$$R = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$R^2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ and}$$

$$R^3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

If we rotate around the diagonal adjoining $(1,1,1)$ and $(-1,-1,-1)$, we obtain the non-identity matrices $R = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $R^2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$.

If we rotate around the line connecting $(1,0,1)$ and $(-1,0,-1)$, then we obtain the non-identity matrix $R = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

We obtain, the 24 reflections in the symmetry group of a cube by multiplying the 24 rotations by $-I$. We conclude that the symmetry group of the cube has order 48.

Exercise 2.6.49

i. Show that the group of rotations of a dodecahedron has order 60 and is isomorphic to A_5. Show that the full symmetry group of a regular dodecahedron has order 120.

ii. Write the matrices for the 60 rotational symmetries of a regular dodecahedron.

Chapter 3

Metric Spaces

> *...la notion d'espace métrique fut introduite en 1906 par M. Fréchet, et dévelopée quelques années plus tard par F. Hausdorff dans sa Mengenlehre. Elle acquit une grande importance aprés 1920, d'une part á la suite des travaus fondamentaux de S. Banach et de son école sur les espaces normés et leurs applications á l'Analyse fonctionnelle, de l'autre en raison de l'intérêt que présente la notion de valeur absolue en Arithmétique et en Géométrie algébrique (où notamment la complétion par rapport á une valeur absolue se montre trés féconde).*
>
> *– N. Bourbaki, Topologie Générale, Book 3*

3.1 Introduction

We have already encountered the notion of distance in n-dimensional Euclidean space. All that this involves is the repeated use of the Pythagorean theorem. If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are elements in $\mathbb{R}^n$, then in Chapter 2 we defined

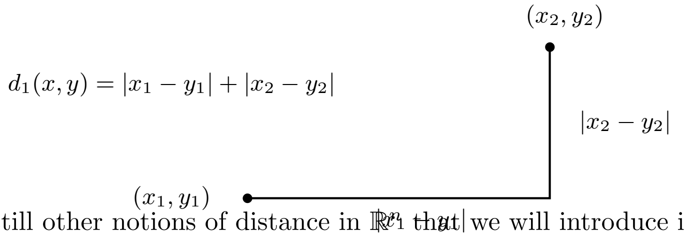
$$d(x, y) = ((x_1 - y_1)^2 + \cdots + (x_n - y_n)^2)^{1/2}.$$

Thus, if $\|x\| = (\sum_{j=1}^{n} x_j^2)^{1/2}$, then $d(x, y) = \|x - y\|$.

In this chapter, we generalize the notion of distance in $\mathbb{R}^n$ to that of a metric on any set S. This new notion of distance will allow us to define open sets, closed sets, and compact sets as we did for the real and complex numbers.

Remark 3.1.1 Note that the metric d on $\mathbb{R}^n$ defined above is called the *usual metric on* $\mathbb{R}^n$.

It should be mentioned that this is not the only way to define the distance between two points in $\mathbb{R}^n$. For example, in $\mathbb{R}^2$, we could define $d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$. This is sometimes called the "taxicab metric." This is what happens when you are driving on city streets and are not allowed to drive through buildings or across people's lawns. The distance d_1 is illustrated below.



There are still other notions of distance in $\mathbb{R}^n$ that we will introduce in short order.

3.2 Definition and Basic Properties of Metric Spaces

Definition 3.2.1 A *metric space* is a pair (X, d) where X is a set and $d : X \times X \to \mathbb{R}$ is a map satisfying the following properties.

a. For $x_1, x_2 \in X$, $d(x_1, x_2) \geq 0$,
and $d(x_1, x_2) = 0$ if and only if $x_1 = x_2$, (positive definite).

b. For any $x_1, x_2 \in X$, we have $d(x_1, x_2) = d(x_2, x_1)$, (symmetric).

c. For any $x_1, x_2, x_3 \in X$, we have

$$d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2),$$

(triangle inequality).

Exercise 3.2.2

i. Draw a triangle and figure out why the triangle inequality is so named.

ii. Replace the triangle inequality by the inequality

$$d(x_1, x_2) \leq d(x_1, x_3) + d(x_2, x_3)$$

for any $x_1, x_2, x_3 \in X$. Show that symmetry (property 3.2.1.b) follows from this version of the triangle inequality and positive definiteness (property 3.2.1.a).

Example 3.2.3 Observe that we have proved in Exercise 2.5.11 that the usual metric on $\mathbb{R}^n$ satisfies this definition.

Exercise 3.2.4 On $\mathbb{C}^n = \{z = (z_1, z_2, \ldots, z_n) \mid z_j \in \mathbb{C}\}$, we define

$$\|z\| = \left(\sum_{j=1}^{n} |z_j|^2 \right)^{1/2}$$

and, for $z, w \in \mathbb{C}^n$, we define $d(z, w) = \|z - w\|$. Show that d is a metric on $\mathbb{C}^n$.

Exercise 3.2.5 Let X be any nonempty set and, for $x_1, x_2 \in X$, define

$$d(x_1, x_2) = \begin{cases} 0 & \text{if } x_1 = x_2, \\ 1 & \text{if } x_1 \neq x_2. \end{cases}$$

Show that d is a metric on X. This is called the *discrete metric*, the pair (X, d) is referred to as a *discrete metric space*. It is designed to disabuse people of the notion that every metric looks like the usual metric on $\mathbb{R}^n$. The discrete metric is very handy for producing counterexamples.

Exercise 3.2.6 Let (X, d) be a metric space, and let Y be a proper subset of X. Show that (Y, d') is a metric space, where we define $d'(y_1, y_2) = d(y_1, y_2)$. We call d' the *inherited metric* on Y.

Expanding on Remark 3.1.1, we introduce an important collection of metrics on $\mathbb{R}^n$ in the next few paragraphs. Pay attention, as these are key examples for future developments.

Let p be a real number such that $p \geq 1$. For $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we define

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p} .$$

As usual, if $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, we define $d_p(x, y) = \|x - y\|_p$. Note that if $p = 1$ and $n = 2$, this is the taxicab metric that we encountered in Remark 3.1.1. Note further that if $p = 2$, this is the usual Euclidean distance on $\mathbb{R}^n$. To show that d_p is a metric on $\mathbb{R}^n$, we need the following inequality:

Theorem 3.2.7 (Hölder's Inequality) Suppose p, q are real numbers greater than 1 such that $1/p+1/q = 1$. Suppose $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, then

$$\sum_{k=1}^{n} |x_k y_k| \le \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q}$$

Proof. The proof is based on the following inequality. Suppose a and b are nonnegative real numbers and p, q are as in the hypothesis of the theorem. Then $ab \le \frac{a^p}{p} + \frac{b^q}{q}$. This is proved with elementary calculus. Consider the function $y = x^{p-1}$, for $x \ge 0$. Then the inverse function is $x = y^{q-1}$, for $y \ge 0$. We have $\int_0^a x^{p-1}dx + \int_0^b y^{q-1}dy = \frac{a^p}{p} + \frac{b^q}{q}$. A look at the graph of the functions reveals immediately that this sum is greater than or equal to ab, where equality holds if and only if $b = a^{p-1}$, which is equivalent to saying $b^q = a^p$.

Using this inequality, we get

$$\sum_{k=1}^{n} \frac{|x_k|}{||x||_p} \frac{|y_k|}{||y||_q} \le \sum_{k=1}^{n} \frac{|x_k|^p}{p||x||_p^p} + \sum_{k=1}^{n} \frac{|y_k|^q}{q||y||_q^q} = 1/p + 1/q = 1.$$

Exercise 3.2.8 Prove that d_p is a metric on $\mathbb{R}^n$ for $p > 1$. *Hint:* The triangle inequality is the only hard part. The proof of the triangle inequality depends on Hölder's Inequality. To begin, observe that

$$||x+y||_p^p = \sum_i |x_i + y_i|^p \le \sum_i |x_i + y_i|^{p-1}|x_i| + \sum_i |x_i + y_i|^{p-1}|y_i|$$

Now apply Hölder.

Exercise 3.2.9 Note that Hölder's inequality only works for $p, q > 1$. Prove the triangle inequality for the d_1 metric.

We also define a metric for $p = \infty$. That is, if $x = (x_1, x_2, \ldots, x_n)$, we set $\|x\|_\infty = \max_{1\le j\le n} |x_j|$, and define

$$d_\infty(x, y) = \max_{1\le j\le n} |x_j - y_j| = \|x - y\|_\infty.$$

Exercise 3.2.10 Prove that d_∞ defines a metric on $\mathbb{R}^n$.

Definition 3.2.11 The metric space $(\mathbb{R}^n, d_p)$, for $1 \le p \le \infty$, is denoted by $\ell_n^p(\mathbb{R})$.

Exercise 3.2.12 Show that everything we have just done for $\mathbb{R}^n$ can also be done for $\mathbb{C}^n$. This yields a collection of spaces $\ell_n^p(\mathbb{C})$.

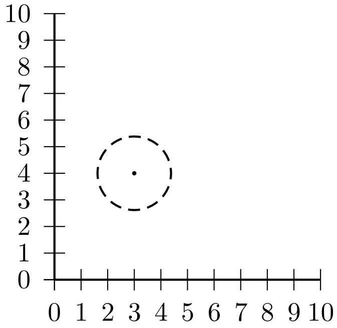
There is a lengthy menu attached to the study of metric spaces. For example, we need to deal with such concepts as open sets, closed sets, compact sets, accumulation points, isolated points, boundary points, interior, closure, and other things. To understand metric spaces fully, the reader must deal not only with these ideas, but with the relationships among them. Most of these ideas have a setting in the context of general topological spaces.

3.3 Topology of metric spaces

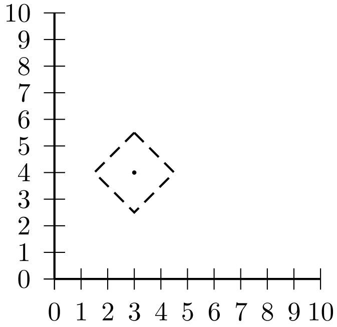
A fundamental notion in the study of metric spaces is that of an open ball.

Definition 3.3.1 Suppose that (X, d) is a metric space and $x_0 \in X$. If $r \in \mathbb{R}$, with $r > 0$, the *open ball of radius r around* x_0 is the subset of X defined by $B_r(x_0) = \{x \in X \mid d(x, x_0) < r\}$. The *closed ball of radius r around* x_0 is the subset of X defined by $\overline{B}_r(x_0) = \{x \in X \mid d(x, x_0) \leq r\}$.

Example 3.3.2 In $\mathbb{R}^2$, with the usual metric, a ball of radius 3/2 around the point $(3, 4)$ looks like this:

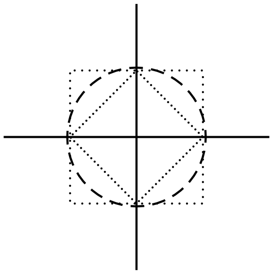


Example 3.3.3 In $\mathbb{R}^2$, a ball of radius 3/2 around the point $(3, 4)$ in the d_1 metric looks like this:



Definition 3.3.4 Suppose that V is a vector space with a metric d. The *unit ball in V* is the ball of radius 1 with center at $\mathbf{0}$, that is, $B_1(\mathbf{0})$. This definition is usually only interesting when the metric arises from a "norm" (see Exercises 3.2.4 and 3.2.12).

We want to analyze the nature of the *unit ball* in $\ell_n^p(\mathbb{R})$, that is, the set of all points $x \in \mathbb{R}^n$ such that $\|x\|_p < 1$. For the moment, let's take $n = 2$ and consider the cases of $\ell_2^1(\mathbb{R})$, $\ell_2^2(\mathbb{R})$, and $\ell_2^\infty(\mathbb{R})$. The pictures of the unit balls in these spaces are shown below. This leads to an interesting diagram.



Exercise 3.3.5 If $1 \leq p < q$, show that the unit ball in $\ell_n^p(\mathbb{R})$ is contained in the unit ball in $\ell_n^q(\mathbb{R})$.

Exercise 3.3.6 Choose p with $1 \leq p \leq \infty$, and let $\varepsilon > 0$. Show that $B_\varepsilon(0) = \{\varepsilon \cdot x \mid x \in B_1(0)\}$.

Exercise 3.3.7 Consider a point $x \in \mathbb{R}^2$ that lies outside the unit ball in $\ell_2^1(\mathbb{R})$ and inside the unit ball in $\ell_2^\infty(\mathbb{R})$. Is there a p between 1 and ∞ such that $\|x\|_p = 1$? Do the same problem in $\mathbb{R}^n$.

Next, we look at open sets.

Definition 3.3.8 Let (X, d) be a metric space and suppose that $A \subseteq X$. The set A is an *open set* in X if, for each $a \in A$, there is an $r > 0$ such that $B_r(a) \subseteq A$.

Notice that the radius r depends on the point a. Also, observe that the empty set $\varnothing$ and the whole space X are both open sets.

Exercise 3.3.9 Prove that, for any $x_0 \in X$ and any $r > 0$, the "open ball" $B_r(x_0)$ is open. So now we can legitimately call an "open" ball an open set.

Exercise 3.3.10 Prove that the following are open sets.

i. The "first quadrant," that is, $\{(x, y) \in \mathbb{R}^2 \mid x > 0 \text{ and } y > 0\}$, in the usual metric;

ii. any subset of a discrete metric space.

Example 3.3.11 Let $X = [-1, 1]$ with the metric inherited from $\mathbb{R}$. What do open balls and open sets look like in this metric space? If $r \leq 1$, then $B_r(0) = (-r, r)$, just as in $\mathbb{R}$. If $r > 1$, then $B_r(0) = [-1, 1] = X$. This does not look like an open set, and in fact, as a subset of $\mathbb{R}$, it is not open. However, in this metric space, it is the whole space, and we have seen that X is always open as a subset of itself.

Exercise 3.3.12 Let $X = [-1, 1]$ with the inherited metric as above. Describe the open balls $B_r(1)$ for various values of r.

Exercise 3.3.13 Let (X, d) be a metric space, and let Y be an open set in X. Show that every open set in (Y, d'), where d' is the inherited metric, is also open in X.

Open sets behave nicely under certain set-theoretic operations.

Theorem 3.3.14

i. If $\{A_j\}_{j \in J}$ is a family of open sets in a metric space (X, d), then

$$\bigcup_{j \in J} A_j$$

is an open set in X;

ii. if $A_1, A_2, \ldots, A_n$ are open sets in a metric space (X, d), then

$$\bigcap_{j=1}^{n} A_j$$

is an open set in X.

Proof.

i. Suppose that $x \in \cup_{j \in J} A_j$. Then $x \in A_k$ for some $k \in J$. Since A_k is open, there is a real number $r > 0$ such that $B_r(x) \subseteq A_k$. But then, $B_r(x) \subseteq \cup_{j \in J} A_j$. ☠

ii. Suppose $x \in \cap_{j=1}^{n} A_j$. Then $x \in A_j$ for each $j = 1, 2, \ldots, n$. Since A_j is open, for each j, there exists a radius r_j such that $B_{r_j}(x) \subseteq A_j$. Let $r = \min_{1 \leq j \leq n}\{r_j\}$. Then $r > 0$ and $B_r(x) \subseteq \cap_{j=1}^{n} A_j$.

We can now say that the collection of open sets is closed under the operations of arbitrary union and finite intersection.

Exercise 3.3.15

i. There can be problems with infinite intersections. For example, let $A_n = B_{1/n}((0,0))$ in $\mathbb{R}^2$ with the usual metric. Show that

$$\bigcap_{n=1}^{\infty} A_n$$

is not open.

ii. Find an infinite collection of distinct open sets in $\mathbb{R}^2$ with the usual metric whose intersection is a nonempty open set.

Thus infinte interesections of open sets may or may not be open.

If there are open sets in a metric space, can closed sets be far behind?

Definition 3.3.16 Let (X, d) be a metric space and suppose that $A \subseteq X$. We say that A is a *closed set* in X if ${}^{c}A$ is open in X. (Recall that ${}^{c}A = X \setminus A$ is the complement of A in X.)

Exercise 3.3.17 Show that the following are closed sets.

i. The x-axis in $\mathbb{R}^2$ with the usual metric;

ii. the whole space X in any metric space;

iii. the empty set in any metric space;

iv. a single point in any metric space;

v. any subset of a discrete metric space;

vi. a closed ball $\overline{B}_r(x_0)$ in any metric space.

Example 3.3.18 Let $X = (-1, 1)$ with the metric inherited from $\mathbb{R}$. If $r < 1$, then $\overline{B}_r(0) = [-r, r]$, just as in $\mathbb{R}$. If $r \geq 1$, then $B_r(0) = (-1, 1) = X$. Despite first appearances, this is again a closed set in X. Note also that $\overline{B}_{\frac{1}{2}}(\frac{1}{2}) = [0, 1)$, another unusual-looking closed set in this metric space.

Exercise 3.3.19 Let (X, d) be a metric space, and let Y be a closed set in X. Show that every closed set in (Y, d'), where d' is the inherited metric, is also closed in X.

Exercise 3.3.20 Show that $\mathbb{Q}$ as a subset of $\mathbb{R}$ with the usual metric is neither open nor closed in $\mathbb{R}$. Of course, if the metric space is simply $\mathbb{Q}$ with the usual metric, then $\mathbb{Q}$ is both open and closed in $\mathbb{Q}$.

Here is a basic theorem about closed sets.

Theorem 3.3.21

i. Suppose that (X, d) is a metric space and that $\{A_j\}_{j \in J}$ is a collection of closed sets in X. Then

$$\bigcap_{j \in J} A_j$$

is a closed set in X;

ii. if $A_1, A_2, \ldots, A_n$ are closed sets in X, then

$$\bigcup_{j=1}^{n} A_j$$

is a closed set in X.

Proof. Use Theorem 3.3.14 and De Morgan's laws.

Exercise 3.3.22 Let (X, d) be a metric space. Let A be an open set in X, and let B be a closed set in X. Show that $A \setminus B$ is open and $B \setminus A$ is closed.

So, a set is closed iff its complement is open, and a set is open iff its complement is closed. However, most of time, most sets in a metric spaces are neither open nor closed. There is a different way to characterize closed sets. First, we need the notion of an accumulation point. From here on, we shall simply refer to a metric space X and suppress the notation d for the metric.

Definition 3.3.23 Suppose that A is a subset of a metric space X. A point $x_0 \in X$ is an *accumulation point* of A if, for every $r > 0$, we have $(B_r(x_0) \setminus \{x_0\}) \cap A \neq \varnothing$.

Thus, if x_0 is an accumulation point of A, there are points of A (other than x_0) that are arbitrarily close to x_0. Note that, x_0 may or may not be an element of A. For example, for $\mathbb{R}$ with the usual metric, 1 and 0 are accumulation points of the open interval $(0, 1)$ as well as all of the points in the interval itself.

Definition 3.3.24 Suppose that A is a subset of a metric space X. A point $x_0 \in A$ is an *isolated point* of A if there is an $r > 0$ such that $B_r(x_0) \cap A = \{x_0\}$.

Definition 3.3.25 Suppose that A is a subset of a metric space X. A point $x_0 \in X$ is a *boundary point* of A if, for every $r > 0$, $B_r(x_0) \cap A \neq \varnothing$ and $B_r(x_0) \cap {}^cA \neq \varnothing$. The *boundary of* A is the set of boundary points of A, and is denoted by ∂A.

We need some examples.

Examples 3.3.26

i. Let $A = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 < 1\}$. We take the usual metric on $\mathbb{R}^3$. The set of accumulation points of A is $B^3 = \{(x, y, z) \mid x^2 + y^2 + z^2 \leq 1\}$ and is called *the closed unit ball* in $\mathbb{R}^3$ with respect to the usual metric. The set A has no isolated points, and $\partial A = S^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}$. The set S^2 is called the 2-*sphere* in $\mathbb{R}^3$ with respect to the usual metric.

ii. Let $A = \{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + x_2^2 + \ldots + x_n^2 < 1\}$. We take the usual metric in $\mathbb{R}^n$. The set of accumulation points of A is $B^n = \{(x_1, x_2, \ldots, x_n) \mid x_1^2 + x_2^2 + \ldots + x_n^2 \leq 1\}$. The set A is called *the open unit ball* with respect to the usual metric and the set B^n is called *the closed unit ball* in $\mathbb{R}^n$ with respect to the usual metric. The set A has no isolated points and $\partial A = S^{n-1} = \{(x_1, x_2, \ldots, x_n) \mid x_1^2 + x_2^2 + \ldots + x_n^2 = 1\}$. The set S^{n-1} is called the $(n-1)$-*sphere* in $\mathbb{R}^n$ with respect to the usual metric.

iii. Let $A = \mathbb{Q} \subseteq \mathbb{R}$ with the usual metric. Then every point in $\mathbb{R}$ is an accumulation point of A, the set A has no isolated points, and $\partial A = \mathbb{R}$.

iv. If A is any subset of a discrete metric space X, then A has no accumulation points. Every point in A is an isolated point, and $\partial A = \varnothing$.

v. Let $A = \{\frac{1}{n} \mid n \in \mathbb{N}\} \subseteq \mathbb{R}$ with the usual metric. Then every point of A is an isolated point and a boundary point, the point 0 is the only accumulation point of the set, and the set A is neither open nor closed.

Now, we have another means of identifying closed sets, that is, accumulation points.

Theorem 3.3.27 Suppose A is a subset of a metric space X. Then A is closed iff A contains all its accumulation points.

Proof. If A is the empty set, then A has no accumulation points. Suppose that A is a non-empty closed set and that x_0 is an accumulation point of A. If $x_0 \notin A$, then $x_0 \in {}^cA$, which is open. Hence, there is an $r > 0$ such that $B_r(x_0) \subseteq {}^cA$, and this contradicts the definition of accumulation point. Conversely, suppose that A contains all its accumulation points and that $x_0 \in {}^cA$. Then x_0 is not an accumulation point of A, and hence there exists $r > 0$ such that $B_r(x_0) \cap A = \varnothing$. This means that ${}^cA$ is open, and so A is closed.

In a discrete metric space any subset is both open and closed. This is not generally the case. For example, in the case of $\ell_n^p(\mathbb{R})$ and $\ell_n^p(\mathbb{C})$, most subsets are neither open nor closed.

Exercise 3.3.28 Find an uncountable number of subsets of $\ell_n^p(\mathbb{R})$ and $\ell_n^p(\mathbb{C})$ that are neither open nor closed.

If a set A in a metric space X is not closed and we wish that it were, then we can do something about it.

Definition 3.3.29 Suppose that A is a nonempty subset of a metric space X. The *closure of* A is the intersection of all the closed sets which contain A.

The closure of any set A exists, since there are always closed sets that contain A, for example X. The closure of A is a closed set since it is the intersection of closed sets. So the closure of A is the "smallest" closed set that contains A. We denote the closure of a set A by $\overline{A}$. Obviously, $A \subseteq \overline{A}$ and $A = \overline{A}$ iff A is closed.

Examples 3.3.30

i. Let $A = \{(x, y, z) \in \mathbb{R}^3 \mid x > 0, y > 0, z > 0\}$. If $\mathbb{R}^3$ has the usual metric, then $\overline{A} = \{(x, y, z) \in \mathbb{R}^3 \mid x \geq 0, y \geq 0, z \geq 0\}$.

ii. Let $\mathbb{Q}^n = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n \mid x_j \in \mathbb{Q} \text{ for } 1 \leq j \leq n\}$. If $\mathbb{R}^n$ has the usual metric, then $\overline{\mathbb{Q}^n} = \mathbb{R}^n$.

iii. Let X be a discrete metric space and let A be any subset of X. Then $\overline{A} = A$.

It should not come as a surprise that the notions of closure and accumulation point are intimately related.

Exercise 3.3.31 Suppose that A is a subset of a metric space X. Show that $\overline{A} = A \cup \{\text{accumulation points of } A\}$.

Exercise 3.3.32 Suppose A is a subset of a metric space X. Prove or disprove: $\overline{A} = A \cup \partial A$.

Exercise 3.3.33 Suppose A is a subset of a metric space X. Prove that $\partial A = \overline{A} \cap \overline{{}^cA}$.

Exercise 3.3.34 Let X be a metric space and let $x_0 \in X$. Suppose that $r > 0$. Prove or disprove: $\overline{B_r(x_0)} = \{x \in X \mid d(x, x_0) \leq r\}$.

Exercise 3.3.35 For definitions and notations for this exercise, see Project 2.1.

i. Consider the set of 2×2 matrices over $\mathbb{R}$, that is, $M_2(\mathbb{R})$. Make this into a metric space by identifying it with $\mathbb{R}^4$ with the usual metric. Show that $GL_2(\mathbb{R})$ is an open subset of $M_2(\mathbb{R})$, and that $\overline{GL_2(\mathbb{R})} = M_2(\mathbb{R})$.

ii. Show that $SL_2(\mathbb{R})$ is a closed subset of $GL_2(\mathbb{R})$

Exercise 3.3.36 Let A be a subset of a metric space X and let x_0 be an isolated point of A. Show that x_0 is in the boundary of A if and only if x_0 is an accumulation point of ${}^cA$.

Corresponding to the notion of closure is the idea of the *interior* of a set.

Definition 3.3.37 Let A be a subset of a metric space X. The *interior* of A is the union of all open sets which are contained in A.

The interior of A is the "largest" open set contained in A. We denote the interior of A by A°. Obviously $A^\circ \subseteq A$ and $A^\circ = A$ iff A is open.

Examples 3.3.38

i. Let $X = \mathbb{R}^3$ with the usual metric and $A = \{(x, y, z) \mid z \geq 0\}$. Then $A^\circ = \{(x, y, z) \mid z > 0\}$;

ii. let X be a discrete metric space and let A be any subset of X. Then $A^\circ = A$ and $\overline{A} = A$, so that $A = A^\circ = \overline{A}$.

Exercise 3.3.39 Show that, in the usual metric on $\mathbb{R}$, the interior of $\mathbb{Q}$ is empty, that is, $\mathbb{Q}^\circ = \varnothing$, but the the interior of $\overline{\mathbb{Q}}$ is $\mathbb{R}$, that is, $(\overline{\mathbb{Q}})^\circ = \mathbb{R}$.

Exercise 3.3.40 Look at combinations of interior, closure, and boundary and determine how many different possibilities result. For this exercise only, let "I" stand for interior, "B" stand for boundary, and "C" stand for closure. Let X be a metric space and let $A \subseteq X$. How many possible sets can be made from A with these operations? For example, $I(I(A)) = I(A)$, but $C(I(A))$ is not necessarily A. Is it $C(A)$? Explore all possibilities of applying combinations of I,C, and B. Hint: There are only a finite number.

Another important concept in the theory of metric spaces is that of diameter.

Definition 3.3.41 Let A be a nonempty subset of a metric space X. The *diameter of* A is

$$\operatorname{diam}(A) = \sup_{x,y \in A} d(x, y).$$

Note that we may have $\operatorname{diam}(A) = \infty$.

Exercise 3.3.42

i. Show that the diameter of a set is 0 iff the set consists of a single point.

ii. Suppose A is a nonempty subset of a metric space X. Show that $\operatorname{diam}(A) = \operatorname{diam}(\overline{A})$.

Definition 3.3.43 Let A be a nonempty subset of $\mathbb{R}^n$. We say that A is *convex* if, given any two points $\mathbf{p}, \mathbf{q} \in A$, the "line segment" with endpoints $\mathbf{p}$ and $\mathbf{q}$, that is, the set

$$\{(1 - t)\mathbf{p} + t\mathbf{q} \mid t \in \mathbb{R}, 0 \leq t \leq 1\},$$

is a subset of A.

Example 3.3.44 The unit ball B^n contained in $\mathbb{R}^n$, in the usual metric, is a convex set.

Exercise 3.3.45 Show that the unit ball $\ell_n^p(\mathbb{R})$, for $1 \leq p \leq \infty$, is a convex set in $\mathbb{R}^n$.

Definition 3.3.46 Let A be a subset of $\mathbb{R}^n$ with the usual metric. The *convex hull* of A is the intersection of all convex sets containing A. The *closed convex hull* of A is the intersection of all closed convex sets containing A.

Exercise 3.3.47 Let A be a nonempty subset of $\mathbb{R}^n$ and let C be the convex hull of A.

i. Prove or disprove the following statement. The closed convex hull of A is $\overline{C}$.

ii. Show that the diameter of A is the diameter of C.

Remark 3.3.48 The concept of convex set in $\mathbb{R}^n$ does not involve a metric in $\mathbb{R}^n$. However a particular metric is often used to define subsets of $\mathbb{R}^n$ that may or may not be convex.

Exercise 3.3.49

i. Describe of the closed convex hull of the unit ball in $\ell_n^p(\mathbb{R})$ for $1 \leq p \leq \infty$.

ii. Suppose $0 < p < 1$. For $\mathbf{x} \in \mathbb{R}^n$, define,

$$\|\mathbf{x}\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}}.$$

Define $S_p = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_p \leq 1\}$. Determine whether S_p is convex. If not, find the closed convex hull of S_p.

Example 3.3.50 We now branch out in a slightly different direction. Suppose that X is a set and and $F = \mathbb{R}$ or $\mathbb{C}$. Denote by $\mathcal{B}(X, F)$ the set of all bounded functions from X to F. Thus, $f \in \mathcal{B}(X, F)$ iff there is a real number M such that $|f(x)| \leq M$ for all $x \in X$. For $f \in \mathcal{B}(X, F)$, we define $\|f\| = \sup_{x \in X} |f(x)|$ (the "sup norm"). For $f, g \in \mathcal{B}(X, F)$, we define $d(f, g) = \sup_{x \in X} |f(x) - g(x)|$ (the "sup metric"). It follows easily from the definition of sup that d is a metric on $\mathcal{B}(X, F)$. In this example, an open ball of radius r around a function f is the collection of all functions which lie within an "r-strip" around f.

Exercise 3.3.51

i. Let $F = \mathbb{R}$ or $\mathbb{C}$. Show that $\mathcal{B}(X, F)$, with d as defined above, is a metric space.

ii. For $f, g \in \mathcal{B}(X, F)$, define $(f + g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$. Also, for $\alpha \in F$ define $(\alpha f)(x) = \alpha f(x)$. Show that, with these operations, $\mathcal{B}(X, F)$ is a commutative algebra with 1 over F (see Definition 2.3.7). Of course, scalar multiplication is simply multiplication by a constant function.

This is a step up in our examples of metric spaces. While previous examples are important, spaces of functions are the most significant examples of metric spaces in analysis.

3.4 Convergence and Completeness

Our next big idea is convergence in a metric space. When we discussed the convergence of a sequence in $\mathbb{R}$ or $\mathbb{C}$, we used the absolute value to measure the distance between two points in one of these fields. Here, in a general metric space, we can use the metric to accomplish the same thing.

Definition 3.4.1 Suppose $(a_n)_{n \in \mathbb{N}}$ is a sequence of points in a metric space X. We say that a point $L \in X$ is the *limit* of the sequence $(a_n)_{n \in \mathbb{N}}$ as n goes to infinity if, for any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $d(a_n, L) < \varepsilon$ whenever $n \geq N_\varepsilon$. When the limit exists, we say that $(a_n)_{n \in \mathbb{N}}$ *converges to* L, and we write

$$\lim_{n \to \infty} a_n = L.$$

Sometimes, we simply say that $(a_n)_{n \in \mathbb{N}}$ *converges in* X without mentioning L explicitly.

As in Chapter 1, we have a concept of Cauchy sequences in a metric space.

Definition 3.4.2 (See 1.5.3, 1.6.22.) Let X be a metric space and let $(a_n)_{n \in \mathbb{N}}$ be a sequence in X. We say that $(a_n)_{n \in \mathbb{N}}$ is a *Cauchy sequence* if, for any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $d(a_n, a_m) < \varepsilon$ whenever $n, m \geq N_\varepsilon$.

It may be that a sequence in a metric space is a Cauchy sequence even though it does not converge. For example, as we observed in Chapter 1, Cauchy sequences in $\mathbb{Q}$ with the usual metric do not necessarily converge in $\mathbb{Q}$. This leads us to the following exercise.

Exercise 3.4.3 Suppose that X is a metric space and that the sequence $(a_n)_{n\in\mathbb{N}}$ converges in X. Show that for any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that $d(a_n, a_m) < \varepsilon$ whenever $n, m \geq N_\varepsilon$. Thus, a convergent sequence is a Cauchy sequence.

Exercise 3.4.4 Let $(a_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in a discrete metric space X. Show that there exists $N \in \mathbb{N}$ such that $d(a_n, a_m) = 0$, that is, $a_n = a_m$ for all $n, m \geq N$. Hence, the sequence is convergent. Such a sequence is called *eventually constant*. Note that an eventually constant sequence in any metric space is convergent, and in fact, it converges to the "eventual constant."

There is a standard litany associated to the notions of convergent sequence and Cauchy sequence. For example, from Exercise 3.4.3, we see that in any metric space, a convergent sequence is a Cauchy sequence. In $\mathbb{R}$ or $\mathbb{C}$ with the usual metric, every Cauchy sequence converges. In $\mathbb{Q}$ with the usual metric, many Cauchy sequences do not converge. The best kinds of metric spaces (at least our favorite kinds) are the ones in which "convergent sequence" and "Cauchy sequence" are synonymous.

Definition 3.4.5 Suppose that X is a metric space. We say that X is a *complete metric space* if every Cauchy sequence in X converges.

Examples 3.4.6 The following metric spaces are complete. (If this seems repetitive, don't worry about it.) The proofs of (*i*) and (*ii*) are in Chapter 1.

i. $\mathbb{R}$ with the usual metric;

ii. $\mathbb{C}$ with the usual metric;

iii. any discrete metric space.

Examples 3.4.7 The following metric spaces are *not* complete.

i. $\mathbb{Q}$ with the usual metric;

ii. any proper nonempty open subset of $\mathbb{R}$ with the inherited metric.

Exercise 3.4.8 Prove that a subset Y of a complete metric space X is also complete metric space with the inherited metric if and only if Y is closed as a subset of X.

Exercise 3.4.9 Show that, for $1 \leq p \leq \infty$, the spaces $\ell_n^p(\mathbb{R})$ and $\ell_n^p(\mathbb{C})$ are complete metric spaces.

We now turn to an investigation of convergence in the spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with the usual metrics. Our approach here is very similar to the one that we took in $\mathbb{R}$. One big difference is that, since there is no notion of order in $\mathbb{R}^n$, the idea of monotonicity has no meaning. However, we will use it one coordinate at a time. Similarly, we need to talk about bounded sets in $\mathbb{R}^n$.

Definition 3.4.10 A subset A of a metric space X is *bounded* if $X = \varnothing$ or there exists a point $x \in X$ and $r > 0$ such that $A \subseteq B_r(x)$.

Exercise 3.4.11 Let $F = \mathbb{R}$ or $\mathbb{C}$, and let $X = F^n$ with the usual metric. Show that a subset A of X is bounded if and only if each of the n sets of the form $A_i = \{x \in F \mid \exists a = (a_1, a_2, \ldots, a_n) \in A \text{ with } a_i = x\}$ are bounded as subsets of F.

The following lemmas and theorems for $\mathbb{R}^n$ and $\mathbb{C}^n$ will be proved for $\mathbb{R}^n$ and left as exercises for $\mathbb{C}^n$.

Lemma 3.4.12 Every bounded sequence in $\mathbb{R}^n$ (or $\mathbb{C}^n$) with the usual metric has a convergent subsequence.

Proof. Let $(a_m)_{m\in\mathbb{N}}$ be a bounded sequence in $\mathbb{R}^n$. Write $a_m = (a_{m,1}, a_{m,2}, \ldots, a_{m,n})$. We prove the Lemma by induction on n. For $n = 1$, this is the content of Lemma 1.6.15. Assume the lemma is true for $n-1$. Let a'_m be the $(n-1)$-tuple $(a_{m,1}, a_{m,2}, \ldots, a_{m,n-1})$. Then $(a'_m)_{m\in\mathbb{N}}$ is a bounded sequence in $\mathbb{R}^{n-1}$. By the induction hypothesis $(a'_m)_{m\in\mathbb{N}}$ has a convergent subsequence in $\mathbb{R}^{n-1}$. Label this convergent subsequence $(a'_{m_j})_{j\in\mathbb{N}}$. Now the sequence $(a_{m_j,n})_{j\in\mathbb{N}}$ is a bounded sequence in $\mathbb{R}$ and hence has a convergent subsequence which we shall not name. Again, taking the corresponding subsequence of $(a_{m_j})_{j\in\mathbb{N}}$, we get a convergent subsequence of the original subsequence $(a_m)_{m\in\mathbb{N}}$.

Exercise 3.4.13

i. For practice, carry out the above proof in $\mathbb{C}^n$.

ii. Prove the above lemma by proceeding coordinate by coordinate. You will notice that the indexing gets quite messy.

Theorem 3.4.14 (Bolzano-Weierstrass) If A is a bounded infinite subset of $\mathbb{R}^n$ or $\mathbb{C}^n$, then A has an accumulation point.

Proof. (We'll do $\mathbb{R}^n$. You do $\mathbb{C}^n$.) Since A is infinite there exists a sequence $(x_k)_{k\in\mathbb{N}}$ in A, where $x_k \neq x_j$ if $k \neq j$. Then $(x_k)_{k\in N}$ is a bounded sequence in $\mathbb{R}^n$ and by Lemma 3.4.12 has a convergent subsequence. If this subsequence converges to x_0, then x_0 is an accumulation point of A.

One of the most important contexts in which to discuss the convergence of sequences is when we consider sequences of functions. There is more than one notion of what it means for a sequence of functions to converge. Below, we discuss two of the most important of these notions, namely pointwise convergence and uniform convergence. We do this in the case of sequences of bounded functions from a set X to $\mathbb{R}$ or $\mathbb{C}$, as in Example 3.3.50.

The most naïve notion of convergence for a sequence of functions is pointwise convergence.

Definition 3.4.15 Let X be a set, and let $F = \mathbb{R}$ or $\mathbb{C}$. Consider a sequence of functions $(f_n)_{n\in\mathbb{N}}$, where $f_n : X \to F$ is a bounded function for each $n \in \mathbb{N}$. We say that a function $f : X \to F$ is the *pointwise limit* of the sequence $(f_n)_{n\in\mathbb{N}}$ if, for every $x \in X$, $\lim_{n\to\infty} f_n(x) = f(x)$.

Example 3.4.16 Let $f_n : [0,1] \to \mathbb{R}$ be given by

$$f_n(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq 1 - \frac{1}{n}, \\ nx - (n-1) & \text{if } 1 - \frac{1}{n} < x \leq 1. \end{cases}$$

Note that for all $x < 1$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ is eventually constant; namely, if $n > \frac{1}{1-x}$, then $f_n(x) = 0$. When $x = 1$, $f_n(1) = 1$ for all $n \in \mathbb{N}$. Thus we can conclude that the pointwise limit of the sequence $(f_n)_{n\in\mathbb{N}}$ is the function $f : [0,1] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x = 1. \end{cases}$$

(The astute reader will have noticed that each of the functions f_n is continuous, while the pointwise limit function f is not—more on this later.)

Example 3.4.17 Let $f_n : (0,1) \to \mathbb{R}$ be given by

$$f_n(x) = \begin{cases} 0 & \text{if } 0 < x < \frac{1}{n}, \\ \frac{1}{x} & \text{if } \frac{1}{n} \leq x < 1. \end{cases}$$

Note that for all $x \in (0,1)$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ is eventually constant; namely, if $n > \frac{1}{x}$, then $f_n(x) = \frac{1}{x}$. Thus we can conclude that the poinwise limit of the sequence $(f_n)_{n\in\mathbb{N}}$ is the function $f : (0,1) \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$. (The astute reader will have noticed that, similarly to the previous example, each of the functions f_n is bounded, while the pointwise limit function f is not.)

Exercise 3.4.18 For the following sequences $(f_n)_{n\in\mathbb{N}}$ of functions, where $f_n : [0, 2\pi] \to \mathbb{R}$ for all $n \in \mathbb{N}$, find all values of $x \in [0, 2\pi]$ such that the sequence $(f_n(x))_{n\in\mathbb{N}}$ converges, and find the pointwise limit function $f : [0, 2\pi] \to \mathbb{R}$ if it exists.

i. $f_n(x) = \sin\left(\frac{x}{n}\right)$

ii. $f_n(x) = \sin(nx)$

iii. $f_n(x) = \sin^n x$

The other major notion of convergence for sequences of functions that we will discuss is uniform convergence. This notion of convergence utilizes the metric on $\mathcal{B}(X, F)$ defined in Example 3.3.50.

Definition 3.4.19 Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $\mathcal{B}(X, F)$, where $F = \mathbb{R}$ or $\mathbb{C}$. The sequence is said to *converge uniformly* to a function $f \in \mathcal{B}(X, F)$ provided that, given $\varepsilon > 0$, $\exists N_\varepsilon \in \mathbb{N}$ such that $\sup_{x\in X}|f_n(x) - f(x)| < \varepsilon$ for $n \geq N_\varepsilon$.

Remark 3.4.20 Note that uniform convergence is convergence in the metric space $\mathcal{B}(X, F)$. However, pointwise convergence is not in general given by convergence in a metric space.

Exercise 3.4.21 Show that if a sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to a function f, then it converges pointwise to the function f.

The converse to the preceding idea is false, as we indicate in the following exercise.

Exercise 3.4.22 Let $f_n(x) = x^n$ for $n \in \mathbb{N}$.

i. Show that the sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to the function $f(x) = 0$ on the interval $(-1, 1)$.

ii. Show that if we restrict to the domain $[-\frac{1}{2}, \frac{1}{2}]$, the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to the function $f(x) = 0$.

iii. Show that the sequence $(f_n)_{n\in\mathbb{N}}$ does *not* converge uniformly on the domain $(-1, 1)$.

We now ask whether a Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{B}(X, F)$ converges uniformly to its pointwise limit f.

Theorem 3.4.23 The spaces $\mathcal{B}(X, \mathbb{R})$ and $\mathcal{B}(X, \mathbb{C})$ are complete metric spaces.

Proof. As above, we consider $\mathcal{B}(X, F)$ where $F = \mathbb{R}$ or $\mathbb{C}$. Suppose that $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{B}(X, F)$. Then, for every $x \in X$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy (and hence convergent) sequence in F because $|f_n(x) - f_m(x)| \leq \sup_{y\in X}|f_n(y) - f_m(y)|$. Thus, the sequence $(f_n)_{n\in\mathbb{N}}$ has a pointwise limit $f : X \to F$. We want to show that the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f. To this end, let $\varepsilon > 0$. Choose $N \in \mathbb{N}$ such that $\sup_{y\in X}|f_n(y) - f_m(y)| < \varepsilon/2$ when $n, m \geq N$. Fix $x \in X$ and choose an integer $N(x) \geq N$ such that $|f_{N(x)}(x) - f(x)| < \varepsilon/2$. Then $|f_n(x) - f(x)| \leq |f_n(x) - f_{N(x)}(x)| + |f_{N(x)}(x) - f(x)| < \varepsilon$ if $n \geq N$.

To complete the proof, we must show that the function f is bounded, that is, $f \in \mathcal{B}(X, F)$. But, from the above inequality, it follows that $|f(x)| < |f_N(x)| + \varepsilon$ for all $x \in X$.

3.5 Continuous functions

We now turn to one of the most important ideas about functions on metric spaces.

Definition 3.5.1 Let (X,d) and (X',d') be metric spaces. A function $f : X \to X'$ is *continuous at the point* $x_0 \in X$ if, for any $\varepsilon > 0$, there is a $\delta > 0$ such that $d'(f(x), f(x_0)) < \varepsilon$ whenever $x \in X$ and $d(x, x_0) < \delta$.

This is the old familiar ε-δ definition. It is simply the statement that

$$\lim_{x \to x_0} f(x) = f(x_0).$$

More generally, the limit of a function $f(x)$ at x_0 is $L \in X'$, and we write

$$\lim_{x \to x_0} f(x) = L,$$

if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d'(f(x), L) < \varepsilon$ whenever $0 < d(x, x_0) < \delta$.

Exercise 3.5.2 Suppose that X and X' are metric spaces as above and that $x_0 \in X$. Show that f is continuous at x_0 iff for every sequence $(x_n)_{n \in \mathbb{N}}$ in X which converges to x_0 in X, we have

$$\lim_{n \to \infty} f(x_n) = f(x_0)$$

in X'.

Note that another way of saying that f is continuous at x_0 is the following: given $\varepsilon > 0$, there exists $\delta > 0$ such that $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$.

Exercise 3.5.3 Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial function, where $\mathbb{R}$ has the usual metric. Show that f is continuous.

In discussing continuity, one must be careful about the domain of the function. For example, define $f : \mathbb{R} \to \mathbb{R}$ by the equation

$$f(x) = \begin{cases} 0 & \text{if } x \notin \mathbb{Q}, \\ 1 & \text{if } x \in \mathbb{Q}. \end{cases}$$

Then, f is not continuous at any point of $\mathbb{R}$. However, suppose we restrict f to be a function from $\mathbb{Q}$ to $\mathbb{Q}$. This means that $f(x) = 1$ on $\mathbb{Q}$ and is continuous at every point of $\mathbb{Q}$.

Exercise 3.5.4 Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1/q & \text{if } x = p/q \text{ (reduced to lowest terms, } x \neq 0), \\ 0 & \text{if } x = 0 \text{ or } x \notin \mathbb{Q}. \end{cases}$$

Show that f is continuous at 0 and any irrational point. Show that f is not continuous at any nonzero rational point.

Continuity is called a *pointwise property* or *local property* of a function f, that is, as in Exercise 3.5.4, a function may be continuous at some points, but not at others. We often deal with functions $f : X \to X'$ which are continuous at every point of X. In this case, we simply say that f is *continuous* without reference to any particular point.

Theorem 3.5.5 Suppose that (X,d) and (X',d') are metric spaces. Then a function $f : X \to X'$ is continuous iff for any open set $V \subset X'$, the set $f^{-1}(V)$ is an open set in X.

Proof. First suppose that f is continuous. Let V be an open set in X'. Suppose $x_0 \in f^{-1}(V)$. Take $\varepsilon > 0$ such that $B_\varepsilon(f(x_0)) \subset V$. Then there exists $\delta > 0$ such that $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$, and so $B_\delta(x_0) \subseteq f^{-1}(B_\varepsilon(f(x_0))) \subseteq f^{-1}(V)$. So $f^{-1}(V)$ is open.

The second half of the proof is easy. You do it.

Corollary 3.5.6 Suppose that (X, d), (X', d'), and (X'', d'') are metric spaces, and $f : X \to X'$ and $g : X' \to X''$ are continuous. Then $g \circ f : X \to X''$ is continuous.

Proof. This follows immediately from the theorem.

Exercise 3.5.7

Prove Corollary 3.5.6 directly from the definition, that is, without using Theorem 3.5.5.

Exercise 3.5.8

i. Let X and X' be metric spaces and assume that X has the discrete metric. Show that any function $f : X \to X'$ is continuous.

ii. Let $X = \mathbb{R}$ with the usual metric and let X' be a discrete metric space. Describe all continuous functions from X to X'.

Exercise 3.5.9 Suppose that (X, d) and (X', d') are metric spaces and that $f : X \to X'$ is continuous. For each of the following statements, determine whether or not it is true. If the assertion is true, prove it. If it is not true, give a counterexample.

i. If A is an open subset of X, then $f(A)$ is an open subset of X';

ii. if A is a closed subset of X, then $f(A)$ is a closed subset of X';

iii. if B is a closed subset of X', then $f^{-1}(B)$ is a closed subset of X;

iv. if A is a bounded subset of X, then $f(A)$ is a bounded subset of X';

v. if B is a bounded subset of X', then $f^{-1}(B)$ is a bounded subset of X;

vi. if $A \subseteq X$ and x_0 is an isolated point of A, then $f(x_0)$ is an isolated point of $f(A)$;

vii. if $A \subseteq X$, $x_0 \in A$, and $f(x_0)$ is an isolated point of $f(A)$, then x_0 is an isolated point of A;

viii. if $A \subseteq X$ and x_0 is an accumulation point of A, then $f(x_0)$ is an accumulation point of $f(A)$;

ix. if $A \subseteq X$, $x_0 \in X$, and $f(x_0)$ is an accumulation point of $f(A)$, then x_0 is an accumulation point of A.

Exercise 3.5.10 Do any of your answers in the previous exercise change if we assume X and X' are complete?

Definition 3.5.11 Let (X, d) and (X', d') be metric spaces. A function $f : X \to X'$ is a *homeomorphism* if

a. f is a bijection,

b. f is continuous, and

c. f^{-1} is also continuous.

Example 3.5.12 The function $\tan : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ is a homeomorphism, with inverse $\tan^{-1} : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$.

Example 3.5.13 Let $X = [0, 1)$ with the induced metric from $\mathbb{R}$, and let $X' = \mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$ with the induced metric from $\mathbb{C}$. The function $f : X \to X'$, $f(x) = e^{2\pi ix}$ is a continuous bijection whose inverse is not continuous.

Exercise 3.5.14 If (X, d) is a metric space, then the function $I(x) = x$ is a homeomorphism from X to itself.

Exercise 3.5.15 Let $X = \mathbb{R}$ with the discrete metric, and let $X' = \mathbb{R}$ with the usual metric. Show that the function $I : X \to X'$, $I(x) = x$ is a continuous bijection but is not a homeomorphism.

Theorem 3.5.16 Suppose $1 \leq p < q \leq \infty$. Then the identity map $I(x) = x$ from $\ell_n^p(\mathbb{R})$ to $\ell_n^q(\mathbb{R})$ is a homeomorphism.

Proof. From Exercise 3.3.5, the image under I of unit ball in $\ell_n^p(\mathbb{R})$ is contained in the unit ball in $\ell_n^q(\mathbb{R})$. Furthermore, by Exercise 3.3.5, the image under I of $B_\varepsilon(0) \subset \ell_n^p(\mathbb{R})$ is contained in $B_\varepsilon(0) \subset \ell_n^q(\mathbb{R})$. Thus, if $\varepsilon > 0$, choosing $\delta = \varepsilon$ shows that I is continuous at 0.

Now take $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and suppose $\max_{1 \leq i \leq n}\{|x_i|\} \leq 1$. Then $|x_1/n| + \cdots + |x_n/n| \leq 1$. This shows that the ball of radius $1/n$ in the ℓ_n^∞ metric is contained in the ball of radius 1 in the ℓ_n^1 metric. In particular, this last fact shows that if we take the unit ball in $\ell_n^q(\mathbb{R})$ and multiply each coordinate by a factor of $1/n$, then the resulting set of points is contained in the unit ball in $\ell_n^p(\mathbb{R})$. By Exercise 3.3.5, this means that the image under I^{-1} of $B_{\frac{1}{n}}(0) \subset \ell_n^q(\mathbb{R})$ is contained in $B_1(0) \subset \ell_n^p(\mathbb{R})$. Similarly, the image under I^{-1} of $B_{\frac{\varepsilon}{n}}(0) \subset \ell_n^q(\mathbb{R})$ is contained in $B_\varepsilon(0) \subset \ell_n^p(\mathbb{R})$. Thus, if $\varepsilon > 0$, choosing $\delta = \frac{\varepsilon}{n}$ shows that I^{-1} is continuous at 0.

A similar argument to the above works at all other points.

Exercise 3.5.17 Show that $\ell_n^p(\mathbb{C})$ and $\ell_n^q(\mathbb{C})$ are homeomorphic.

Exercise 3.5.18 Let (X, d) be a metric space, and for any $x, y \in X$, let $d'(x, y) = \frac{d(x,y)}{1+d(x,y)}$.

i. Show that d' defines a metric on X.

ii. Show that the identity map $I : (X, d) \to (X, d')$, $I(x) = x$, is a homeomorphism.

iii. If (X, d') is complete, is (X, d) necessarily complete?

This exercise is intended to illustrate that, without additional structure, metric spaces can be twisted, expanded, or shrunken without disturbing the open sets too badly.

Definition 3.5.19 Let (X, d) and (X', d') be metric spaces. A homeomorphism $f : X \to X'$ is an *isometry* if

$$d'(f(x_1), f(x_2)) = d(x_1, x_2)$$

for all $x_1, x_2 \in X$.

Exercise 3.5.20 Suppose that, instead, we had define an isometry to be a bijection $f : X \to X'$ such that $d'(f(x_1), f(x_2)) = d(x_1, x_2)$ for all $x_1, x_2 \in X$. Show that with this definition, any isometry is a homeomorphism.

Exercise 3.5.21 Let $X = \mathbb{R}^2$ with the usual metric. Show that the following functions are isometries from X to itself

1. Translation by the vector (a, b) in $\mathbb{R}^2$: $T_{(a,b)}(x, y) = (x + a, y + b)$ for fixed $a, b \in \mathbb{R}$
2. Counterclockwise rotation about the origin by an angle θ: $R_\theta(x, y) = (x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$ for fixed $\theta \in \mathbb{R}$

3. Reflection over a line through the origin making an angle θ with the x-axis: $S_\theta(x, y) = (x \cos 2\theta + y \sin 2\theta, x \sin 2\theta - y \cos 2\theta)$ for fixed $\theta \in \mathbb{R}$

Exercise 3.5.22 Let $\mathbb{R}^n$ have the usual metric. Show that the function $D_a : \mathbb{R}^n \to \mathbb{R}^n$ given by $D_a(x_1, x_2, \ldots, x_n) = (ax_1, ax_2, \ldots, ax_n)$ for fixed $a \in \mathbb{R}$ is an isometry if and only if $a = \pm 1$.

Exercise 3.5.23 In this exercise, we consider isometries from $\mathbb{R}$ to itself in the usual metric.

i. Is $f(x) = x^3$ a bijection? A homeomorphism? An isometry?

ii. Is $f(x) = x + \sin x$ a bijection? A homeomorphism? An isometry?

iii. Find all isometries from $\mathbb{R}$ to itself.

Exercise 3.5.24 For definitions and notations for this exercise, see Project 2.1. Let (X, d) be a metric space. Let G be the collection of all homeomorphisms from X to X. Prove that, under composition of functions, G is a group, and the collection of all isometries is a subgroup of G.

Definition 3.5.25 Suppose that (X, d) is a metric space, and let $F = \mathbb{R}$ or $\mathbb{C}$. Define $\mathcal{BC}(X, F)$ to be the subset of $\mathcal{B}(X, F)$ consisting of continuous functions from X to F. We take the metric on $\mathcal{BC}(X, F)$ to be the induced metric from $\mathcal{B}(X, F)$. If X is compact, then all continuous functions from X to F are bounded (see Exercise 3.6.9 below); so, when X is compact, we will sometimes write $\mathcal{C}(X, F)$ in place of $\mathcal{BC}(X, F)$.

Theorem 3.5.26 The space $\mathcal{BC}(X, F)$ is a complete metric space.

Proof. Suppose that $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{BC}(X, F)$. Then by Theorem 3.4.23, $(f_n)_{n\in\mathbb{N}}$ converges to a function $f \in \mathcal{B}(X, F)$. All we need to show is that f is a continuous function. Now, given $\varepsilon > 0$, there exists N such that $\sup_{x\in X} |f_n(x) - f(x)| < \varepsilon/3$ whenever $n \geq N$. Fix $x_0 \in X$. Then, for any $x \in X$ and $n \geq N$,

$$\begin{aligned} |f(x) - f(x_0)| &= |f(x) + (-f_n(x) + f_n(x)) + (-f_n(x_0) + f_n(x_0)) - f(x_0)| \\ &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)| + |f_n(x_0) - f(x_0)| \\ &< \varepsilon/3 + |f_n(x) - f_n(x_0)| + \varepsilon/3. \end{aligned}$$

Since f_n is continuous, we can choose $\delta > 0$ such that $|f_n(x) - f_n(x_0)| < \varepsilon/3$ whenever $d(x, x_0) < \delta$. But then $|f(x) - f(x_0)| < \varepsilon$ when $d(x, x_0) < \delta$, so f is continuous. ∎

Remark 3.5.27 So we have proved (see Theorems 3.4.23 and 3.5.26) that the uniform limit of bounded functions is a bounded function and the uniform limit of bounded continuous functions is a bounded continuous function. We will find these facts very useful in doing analysis.

Exercise 3.5.28 Let (X, d) be a metric space, and let $F = \mathbb{R}$ or $\mathbb{C}$. Recall that $\mathcal{B}(X, F)$ is a commutative algebra with 1 over F (see Exercise 3.3.51). Show that $\mathcal{BC}(X, F)$ is a subalgebra of $\mathcal{B}(X, F)$, that is, $\mathcal{BC}(X, F)$ is a vector subspace of $\mathcal{B}(X, F)$ that is closed under pointwise multiplication.

Exercise 3.5.29 In Exercise 3.4.22, we saw that the pointwise limit of the sequence of functions $(f_n)_{n\in\mathbb{N}}$, $f_n(x) = x^n$ on $(-1, 1)$, is continuous even though the convergence was not uniform. Now consider the same sequence of functions defined on $[-1, 1]$. Find the pointwise limit of the sequence $(f_n)_{n\in\mathbb{N}}$ and show that it is not continuous.

Exercise 3.5.30 Define a sequence of functions $f_n : (0, 1) \to \mathbb{R}$ by

$$f_n(x) = \begin{cases} \frac{1}{q^n} & \text{if } x = \frac{p}{q} \text{ (reduced to lowest terms, } x \neq 0\text{),} \\ 0 & \text{otherwise,} \end{cases}$$

for $n \in \mathbb{N}$. Find the pointwise limit f of the sequence $(f_n)_{n\in\mathbb{N}}$ and show that $(f_n)_{n\in\mathbb{N}}$ converges to f uniformly.

There is an additional property of continuous functions which is important for future applications.

Definition 3.5.31 Let (X, d) and (X', d') be metric spaces, and let f be a continuous function from X to X'. We say that f is *uniformly continuous* if, given $\varepsilon > 0$, there exists $\delta > 0$ such that, for any pair $x, y \in X$, we have $d'(f(x), f(y)) < \varepsilon$ whenever $d(x, y) < \delta$.

So, f is uniformly continuous if it is continuous at every point and, for a given $\varepsilon > 0$, we can find a corresponding δ that is independent of the point.

Exercise 3.5.32 *i.* Show that a polynomial function $p(x)$ on $\mathbb{R}$ is uniformly continuous if and only if $\deg(p(x)) < 2$.

ii. Show that $f(x) = \sin(x)$ is uniformly continuous on $\mathbb{R}$.

Exercise 3.5.33 Let $X = (0, \infty)$ and determine whether the following functions are uniformly continuous on X:

i. $f(x) = \frac{1}{x}$;

ii. $f(x) = \sqrt{x}$;

iii. $f(x) = \ln(x)$;

iv. $f(x) = x\ln(x)$.

Exercise 3.5.34 Show that any linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ is uniformly continuous.

3.6 Compactness and Connectedness

In Chapter 1, we considered compact subsets of $\mathbb{R}$ and $\mathbb{C}$. Now, with the topology we have in a metric space, we can generalize this idea.

Definition 3.6.1 Let A be a subset of a metric space X. A family $\{U_j\}_{j\in J}$ of open subsets of X is called an *open covering* (or *open cover*) of A if

$$A \subseteq \bigcup_{j\in J} U_j.$$

If $\{U_j\}_{j\in J}$ is an open cover of A, we say that this cover has a *finite subcovering* or (*finite subcover*) if there is a finite subcollection $U_{j_1}, U_{j_2}, \ldots, U_{j_n}$ satisfying

$$A \subseteq \bigcup_{k=1}^{n} U_{j_k}.$$

Examples 3.6.2

i. Let $A = (0, 1) \subseteq \mathbb{R}$ with the usual metric. For $j \in \mathbb{N}$, $j \geq 2$, define $U_j = (\frac{1}{j}, 1)$. Then $A \subseteq \cup_{j\in\mathbb{N}} U_j$, but there is no finite subcover.

ii. Let $B = [0, \infty) \subseteq \mathbb{R}$ with the usual metric. For $j \in \mathbb{N}$, define $U_j = (-1, j)$. Then $B \subseteq \cup_{j\in\mathbb{N}} U_j$, but there is no finite subcover.

iii. Let X be a discrete metric space. For any point $j \in X$, set $U_j = \{j\}$. Then $\{U_j\}_{j\in X}$ is an open cover of X which has a finite subcover iff X is a finite set.

iv. We have seen in Theorem 1.6.38 that if A is a closed and bounded set in $\mathbb{R}$ with the usual metric, then every open cover of A has a finite subcover.

Definition 3.6.3 (See Definitions 1.6.41 and 1.8.20.) Let A be a subset of a metric space X. We say that A is *compact* if every open covering of A has a finite subcovering.

Recall that the Heine-Borel Theorem in $\mathbb{R}$ or $\mathbb{C}$ states that a subset of $\mathbb{R}$ or $\mathbb{C}$ with the usual metric is compact if and only if it is closed and bounded. The statement of the Heine-Borel theorem is certainly not true in a general metric space. For example, take $\mathbb{R}$ with the discrete metric. Then, $\mathbb{R}$ is closed and bounded in this metric. Take an open covering consisting of the individual points in $\mathbb{R}$. This covering does not have a finite subcovering.

For emphasis, we note that the definition insists that for every open covering, there must be a finite subcovering. For example, given any subset A of a metric space X, we have that $\{X\}$ is an open covering which is already finite. So while this particular open covering has a finite subcovering, this does not necessarily imply that other open coverings have finite subcoverings.

Hence, in a general metric space, the closed bounded sets are not necessarily compact. However, we do have one half of the statement of the Heine-Borel theorem in general metric spaces.

Theorem 3.6.4 If a subset A of a metric space X is compact, then A is closed and bounded.

Proof. Recall that a set in a metric space is bounded if and only if it is contained in a ball of finite radius with center at some point. If A is non-empty, take a point $a \in A$ and consider the open covering $\{B_n(a) \mid n \in \mathbb{N}\}$. Since A is compact, this cover has a finite subcovering, and in fact there is an integer N such that $A \subseteq B_N(a)$. Hence, A is bounded.

To prove that A is closed, we assume that x_0 is an accumulation point of A and prove that $x_0 \in A$. Suppose not. Then for each $a \in A$, let $r_a = d(a, x_0)/2$. But the collection $\{B_{r_a}(a) \mid a \in A\}$ is an open cover of A and hence has a finite subcover $\{B_{r_1}(a_1), B_{r_2}(a_2), \ldots, B_{r_n}(a_n)\}$. Let $r = \min\{r_1, r_2, \ldots, r_n\}$. Then $B_r(x_0) \cap B_{r_j}(a_j) = \varnothing$ for all j. Hence, $B_r(x_0) \cap A = \varnothing$, which contradicts the definition of accumulation point. Hence, $x_0 \in A$.

Corollary 3.6.5 If A is a compact set in a metric space X, then every infinite subset of A has an accumulation point in A.

Proof. Suppose that A is a compact set and that C is an infinite subset of A with no accumulation point in A. Then, for each $a \in A$, there is an open ball $B(a)$ centered at a such that $(B(a) \setminus \{a\}) \cap C = \varnothing$. The collection $\{B(a) \mid a \in A\}$ covers A. So, by compactness, we can extract a finite subcover, $\{B(a_1), \ldots, B(a_n)\}$. Thus, $C \subset A \subset B(a_1) \cup \cdots \cup B(a_n)$, and each $B(a_j)$ contains at most one element of C (at its center). This implies that C has at most n elements, which is a contradiction.

Corollary 3.6.6 Let A be a compact set in a metric space. Then, every infinite sequence in A has a subsequence that converges to a point in A.

Exercise 3.6.7 Prove that the Heine-Borel theorem holds in $\mathbb{R}^n$ and $\mathbb{C}^n$ with the usual metrics. (Hint: See the proof of Theorem 1.8.21.)

Exercise 3.6.8

i. Show that a finite union of compact sets is compact.

ii. Give an example of a countable union of compact sets that is not compact.

iii. Show that a closed subset of a compact set is compact.

Exercise 3.6.9

i. Let $f : X \to X'$ be a continuous map of metric spaces. Show that if $A \subseteq X$ is compact, then $f(A) \subseteq X'$ is compact.

ii. Suppose that X is a compact metric space. Show that a continuous function $f : X \to \mathbb{R}$ ($\mathbb{R}$ with the usual metric) is bounded.

iii. Suppose that X is a compact metric space. Show that a continuous function $f : X \to \mathbb{R}$ ($\mathbb{R}$ with the usual metric) attains a maximum and minimum value on X.

Exercise 3.6.10 Suppose X and X' are metric spaces with X compact.

i. If $f : X \to X'$ is continuous on X, show that f is uniformly continuous on X.

ii. If $f : X \to X'$ is a continuous bijection, show that f is a homeomorphism.

Exercise 3.6.11 (Dini's Theorem) Let X be a compact metric space. Suppose f and $(f_n)_{n\in\mathbb{N}}$ are real-valued continuous functions on X. Suppose that, for each $x \in X$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ is a monotonic sequence converging to $f(x)$. Show that $(f_n)_{n\in\mathbb{N}}$ converges to f uniformly.

Exercise 3.6.12 Suppose that A and B are nonempty subsets of a metric space X. The *distance* between A and B is defined by

$$d(A, B) = \inf\{d(a, b) \mid a \in A, b \in B\}.$$

We say that $d(A, B)$ is *assumed* if there exists $a_0 \in A$ and $b_0 \in B$ such that $d(A, B) = d(a_0, b_0)$. Determine whether or not the distance between A and B is necessarily assumed in (*i*)–(*iii*).

i. A is closed and B is closed;

ii. A is compact and B is closed;

iii. A is compact and B is compact.

iv. What happens in the above cases if we assume X is complete?

Exercise 3.6.13 Let X be a metric space, let $A \subset X$ be compact, and let $U \subseteq X$ be an open set containing A. Show that there exists an open set $W \subseteq X$ containing A such that $\overline{W}$ is compact and $\overline{W} \subseteq U$. (Hint: Consider the previous exercise with A and ∂U.)

At this point we introduce an alternate notion of compactness.

Definition 3.6.14 (See Definition 1.6.45.) A subset A of a metric space X is *sequentially compact* if every sequence in A has a subsequence that converges to an element of A.

Exercise 3.6.15 If X is a metric space, and $A \subset X$, we say that A is *totally bounded* if, for any $\varepsilon > 0$, A can be covered finite number of balls of radius ε. Show that a sequentially compact metric space is totally bounded.

One of the most important facts about metric spaces is that compactness and sequential compactness are equivalent. We have already proved (see Corollary 3.6.6) that compactness implies sequential compactness. To prove the converse, we need the following lemma.

Lemma 3.6.16 Let X be a metric space. If $A \subset X$ has the property that every infinite subset of A has an accumulation point in X, then there exists a countable collection of open sets $\{U_i \mid i \in \mathbb{N}\}$ such that, if V is any open set in X and $x \in A \cap V$, then there is some U_i such that $x \in U_i \subset V$.

Proof. We claim that, for each $n \in \mathbb{N}$, there is a finite set of points $x_{n,1}, \dots, x_{n,N(n)}$ in A such that the set of open balls $B_{\frac{1}{n}}(x_{n,1}), B_{\frac{1}{n}}(x_{n,2}), \dots, B_{\frac{1}{n}}(x_{n,N(n)})$ covers A. If A is finite, this is clearly true, so we assume A is infinite.

Suppose our claim is false. Then, there exists $n \in \mathbb{N}$ such that no finite collection of balls of radius $\frac{1}{n}$ centered at points of A can cover A. For each $k \in \mathbb{N}$, define an infinite sequence of points of A inductively as follows. Take $y_1 \in A$. Then $\{B_{\frac{1}{n}}(y_1)\}$ does not cover A. So choose $y_2 \in A \setminus B_{\frac{1}{n}}(y_1)$. Then $\{B_{\frac{1}{n}}(y_1), B_{\frac{1}{n}}(y_2)\}$ does not cover A and $d(y_1, y_2) \geq \frac{1}{n}$. Assume $y_1, \dots, y_k$ have been chosen such that $\{B_{\frac{1}{n}}(y_1), \dots, B_{\frac{1}{n}}(y_k)\}$ does not cover A, and $d(y_i, y_j) \geq \frac{1}{n}$ for all $i \neq j$. Choose $y_{k+1} \in A \setminus (B_{\frac{1}{n}}(y_1) \cup \cdots \cup B_{\frac{1}{n}}(y_k))$. The infinite sequence $(y_k)_{k \in \mathbb{N}}$ does not have an accumulation point anywhere, which contradicts our assumption about A.

Taking all these balls $\{B_{\frac{1}{n}}(x_{n,j}) \mid n \in \mathbb{N} \text{ and } 1 \leq j \leq N(n)\}$ gives the required countable collection.

Exercise 3.6.17 Verify that the above collection satisfies the conclusion of the lemma.

Exercise 3.6.18 Let X be a metric space. If $A \subset X$ has the property that every infinite subset of A has an accumulation point in A, show that for any open covering of A, there exists a countable subcovering.

Now comes a major Theorem.

Theorem 3.6.19 In any metric space, a subset A is compact if and only if it is sequentially compact.

Proof. We have already proved above that compactness implies sequential compactness.

For the converse, suppose that $A \subset X$ is sequentially compact. Then any infinite subset of A contains a countable subset, which defines a sequence in A. By sequential compactness, this sequence has a subsequence that converges to a point $a \in A$. Since this point is clearly an accumulation point of A, we can apply Lemma 3.6.16 and Exercise 3.6.18 to conclude that, for any open cover $\mathcal{U}$ of A, we can find a countable subcover $\mathcal{U}'$.

From this open cover $\mathcal{U}'$, we wish to extract a finite subcover. Let $\mathcal{U}' = \{U_j \mid j \in \mathbb{N}\}$. Suppose that, for each n, the collection $\{U_1, U_2, \dots, U_n\}$ does not cover A. Then, for each n, there exists $x_n \in A \setminus (U_1 \cup \cdots \cup U_n)$. This defines a sequence $(x_n)_{n \in \mathbb{N}}$ in A which by sequential compactness has a convergent subsequence with limit $x \in A$. Since $\mathcal{U}'$ covers A, x must be contained in U_N for some N. But then, U_N contains infinitely many elements of the sequence, and hence contains some x_m with $m > N$. This is a contradiction.

Exercise 3.6.20

i. Show that a compact metric space is complete.

ii. Show that a totally bounded complete metric space is compact. (See Exercise 3.6.15.)

An immediate consequence of Theorem 3.6.19 is the Heine-Borel Theorem in $\mathbb{R}^n$ and $\mathbb{C}^n$.

Theorem 3.6.21 (Heine-Borel) A nonempty subset A of $\mathbb{R}^n$ (or $\mathbb{C}^n$) with the usual metric is compact iff it is closed and bounded.

Proof. Exercise.

Compact sets in $\mathbb{R}^n$ with the usual metric have many interesting properties, some of which are illustrated in the following exercises.

Exercise 3.6.22 Let B be a compact convex subset of $\mathbb{R}^n$ with the usual metric. Define the *nearest point function* $p : {}^cB \to B$ as follows: For $x \in {}^cB$ we set $p(x)$ to be closest point to x that lies in B. Show that

i. the function $p(x)$ is well defined;

ii. the point $p(x)$ lies in the boundary of B;

iii. the function $p(x)$ is surjective onto the boundary of B.

In the next exercise, we continue with the terminology of the preceding exercise. Define the *supporting hyperplane* at $p(x)$ to be the hyperplane through $p(x)$ orthogonal to the vector $p(x)-x$. Define the *supporting half-space* at $p(x)$ to be the set $H_{p(x)} = \{y \in \mathbb{R}^n \mid (y - p(x)) \cdot (p(x) - x) \geq 0\}$. (Note that the supporting hyperplane and the supporting half-space really depend on x, not just on $p(x)$.)

Exercise 3.6.23

i. Show that, for each $x \in {}^cB$, the set B is a subset of $H_{p(x)}$.

ii. Show that $B = \displaystyle\bigcap_{x \in {}^cB} H_{p(x)}$.

iii. Does the above process work when B is a closed convex unbounded subset of $\mathbb{R}^n$ with the usual metric?

Here are a few more interesting facts and ideas about metric spaces. The first involves the notion of separability.

Definition 3.6.24 Let (X, d) be a metric space. A subset $A \subseteq X$ is said to be *dense* in X if $\overline{A} = X$.

Example 3.6.25 *i.* In the usual metric, $\mathbb{Q}$ is dense in $\mathbb{R}$.

ii. The "dyadic numbers," that is, the set $D = \left\{\frac{a}{2^n} \in \mathbb{Q} \mid a, n \in \mathbb{Z}\right\}$, are dense in $\mathbb{R}$ in the usual metric.

Exercise 3.6.26 *i.* Show that in any metric space X, X is dense in X.

ii. Show that in any discrete metric space X, the only dense subset of X is X itself.

iii. Show that if the only dense subset of a metric space X is X itself, then X is discrete.

Definition 3.6.27 Let (X, d) be a metric space. We say that X is *separable* if there exists a countable subset of X that is dense in X.

Example 3.6.28 The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with the usual metric are separable. As a countable dense subset, we can take the collection of all points in $\mathbb{R}^n$ whose coordinates are rational numbers, or the set of all points in $\mathbb{C}^n$ whose coordinates have the property that the real and imaginary parts are rational numbers.

Theorem 3.6.29 If (X, d) is a compact metric space, then X is separable.

Proof. For each $n \in \mathbb{N}$, consider the collection of open balls $\{B_{1/n}(x) \mid x \in X\}$. This is an open covering of X, and hence, there is a finite subcovering $\mathcal{U}_n$. Take the union over all $n \in \mathbb{N}$ of the centers of the balls in $\mathcal{U}_n$. This is a countable collection of points in X that is obviously dense.

Exercise 3.6.30 Suppose X and X' are metric spaces with X separable. Let $f : X \to X'$ be a continuous surjection. Show that X' is separable.

As shown in Example 3.6.28, separable metric spaces do not have to be compact. Many of the important metric spaces which occur in analysis are separable, but there are some very important examples of non-separable metric spaces.

Exercise 3.6.31 Find a metric d on $\mathbb{R}$ such that $(\mathbb{R}, d)$ is not separable.

Exercise 3.6.32 Determine the conditions, if they exist, for which the following metric spaces are separable.

i. $\mathcal{B}(X, F)$

ii. $\mathcal{BC}(X, F)$

Another important idea in metric spaces is connectedness. It has a funny definition because we begin by defining a non-connected set.

Definition 3.6.33 Let X be a metric space and let $A \subset X$. We say that A is *not connected* (or *disconnected*) if there exist open sets $U, V \subset X$ such that

a. $U \cap A \neq \varnothing$ and $V \cap A \neq \varnothing$,

b. $(U \cap A) \cap (V \cap A) = \varnothing$,

c. $A = (U \cap A) \cup (V \cap A)$.

We say that A is *disconnected* by the open sets U and V.

Definition 3.6.34 Let X be a metric space and $A \subset X$. We say A is *connected* if A is not disconnected.

Exercise 3.6.35

i. Show that a subset of a discrete metric space is connected iff its cardinality is at most 1.

ii. Show that a finite subset of any metric space is connected iff its cardinality is at most 1.

iii. Show that a subset A of $\mathbb{R}$ in the usual metric is connected iff A is an interval.

iv. Show that a convex subset of $\mathbb{R}^n$ with the usual metric is a connected set.

The basic theorem about connected sets is the following.

Theorem 3.6.36 Let X, X' be metric spaces and $f : X \to X'$ a continuous function. If A is a connected subset of X, then $f(A)$ is a connected subset of X'. That is, the continuous image of a connected set is connected.

Proof. Let U and V be open sets in X', and assume that U and V disconnect $f(A)$. Then, $f^{-1}(U)$ and $f^{-1}(V)$ are open sets in X which disconnect A.

Corollary 3.6.37 (Intermediate Value Theorem) Let X be a metric space, and take $\mathbb{R}$ with the usual metric. Let $f : X \to \mathbb{R}$ be a continuous function. Let A be a connected subset of X and let $I = f(A)$. Then I is an interval in $\mathbb{R}$, and if $x_0 \in I$ there exists $a_0 \in A$ such that $f(a_0) = x_0$.

Exercise 3.6.38 Take $\mathbb{R}$ with the usual metric, and let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^n$ for $n \in \mathbb{N}$. If b is a positive real number, show that there exists a unique positive real number a such that $a^n = b$. (Hint: Use the Corollary.)

Exercise 3.6.39

i. Consider a "punctured open ball," that is, a set of the from $B_r^\times(a) = B_r(a) \setminus \{a\}$, in $\mathbb{R}^n$ with the usual metric. For which values of n is $B_r^\times(a)$ connected?

ii. Let $B_r^\times(a)$ be a punctured open ball in $\mathbb{C}^n$. For which values of n is $B_r^\times(a)$ connected?

iii. Show that $GL(2, \mathbb{R})$ with the metric inherited from $M_2(\mathbb{R})$ as in Exercise 3.3.35 is not a connected set. (Hint: use the fact that the determinant is a continuous function.)

iv. Show that $GL(2, \mathbb{C})$ with the metric inherited from $M_2(\mathbb{C})$ is a connected set.

If a metric space X is not connected, then it can be decomposed into subsets called connected components.

Definition 3.6.40 If X is a metric space and x_0 is in X, then the *connected component of* x_0 *in* X is the union of the connected sets that contain x_0.

Exercise 3.6.41

i. Let X be a metric space and take $x_0 \in X$. Show that the connected component of x_0 is a connected set in X.

ii. Show that if A is a connected subset of X that contains x_0, then A is contained in the connected component of x_0.

iii. Show that if A is a connected subset of a metric space, then $\overline{A}$ is connected. Deduce that connected components are closed.

Examples 3.6.42

i. Let $X = \mathbb{R}^\times$, the set of nonzero real numbers with the usual metric. This metric space has two connected components, namely, the positive real numbers and the negative real numbers.

ii. The connected components of $GL(2, \mathbb{R})$ with the usual metric are $GL^+(2, \mathbb{R}) = \{x \in GL(2, \mathbb{R}) \mid \det x > 0\}$ and $GL^-(2, \mathbb{R}) = \{x \in GL(2, \mathbb{R}) \mid \det x < 0\}$.

Exercise 3.6.43 Let $O(n, \mathbb{R})$ and $SO(n, \mathbb{R})$ be metric spaces with the metric inherited from $GL(n, \mathbb{R})$. Show that $O(n, \mathbb{R})$ is not connected and that $SO(n, \mathbb{R})$ is connected.

Definition 3.6.44 A metric space X is *totally disconnected* if the connected component of each point is the point itself.

Example 3.6.45 A discrete metric space X is totally disconnected.

Exercise 3.6.46

i. Find an example of a metric space which is totally disconnected but not discrete.

ii. Find an example of a complete metric space which is totally disconnected but not discrete.

3.7 The Contraction Mapping Theorem and its Applications

Definition 3.7.1 Let X be a metric space and f a map from X to X. We say that f is a *contraction mapping of* X if there exists a real number α, with $0 < \alpha < 1$, such that $d(f(x), f(y)) \leq \alpha d(x, y)$ for every pair $x, y \in X$.

Exercise 3.7.2 Show that a contraction mapping is continuous.

Exercise 3.7.3 Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial function. Give conditions on f such that f is a contraction mapping.

Exercise 3.7.4 Let $T : \ell_n^p(\mathbb{R}) \to \ell_n^p(\mathbb{R})$, $1 \leq p \leq \infty$, be a linear transformation. When is T a contraction mapping?

Definition 3.7.5 Let X be a metric space and f a map from X to X. A point $x_0 \in X$ is a *fixed point* of f if $f(x_0) = x_0$.

Exercise 3.7.6

i. Find a continuous function $f : \mathbb{R} \to \mathbb{R}$ that does not have a fixed point.

ii. Find a continuous function $f : (0,1) \to (0,1)$ that does not have a fixed point.

iii. Let $f : [0,1] \to [0,1]$ be continuous. Show that f has a fixed point.

Theorem 3.7.7 (Contraction Mapping Theorem) Let X be a nonempty complete metric space and let $f : X \to X$ be a contraction mapping with constant α. Then f has a unique fixed point $x_0 \in X$.

Proof. Let x_1 be any element of X. Define $x_2 = f(x_1)$, $x_3 = f(x_2) = f(f(x_1)) = f^2(x_1)$, and in general, $x_n = f^{n-1}(x_1)$. Then, if $n > m$, we have

$$\begin{aligned}
d(x_m, x_n) &= d(f^{m-1}(x_1), f^{n-1}(x_1)) \\
&\leq \alpha^{m-1} d(x_1, f^{n-m}(x_1)) \\
&\leq \alpha^{m-1}(d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{n-m}, x_{n-m+1})) \\
&\leq \alpha^{m-1}(d(x_1, x_2) + \alpha d(x_1, x_2) + \cdots + \alpha^{n-m-1} d(x_1, x_2)) \\
&\leq \frac{\alpha^{m-1}}{1-\alpha} d(x_1, x_2).
\end{aligned}$$

It follows that $(x_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in X that converges since X is complete. Let $x_0 = \lim_{n\to\infty} x_n$. From the continuity of f, it follows that $f(x_0) = f(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x_0$.

Exercise 3.7.8 Show that x_0 is the unique fixed point of f.

Exercise 3.7.9 We say that a function $f : \mathbb{R}^n \to \mathbb{R}^N$ satisfies a *Lipschitz condition* if there exists a constant C such that $\|f(x) - f(y)\| \leq C\|x - y\|$ for all $x, y \in \mathbb{R}^n$.

i. Show that f satisfies a Lipschitz condition with constant C for $0 < C < 1$ if and only if f is a contraction mapping.

ii. Show that if f satisfies a Lipschitz condition with any constant, then f is continuous.

iii. For each $C \in (0, \infty)$, find an example of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the Lipschitz condition with constant C.

iv. Let $B = B_1(0)$ be the unit ball in $\mathbb{R}^n$. For each $C > 1$, find an example of a function $f : B \to B$ that satisfies the Lipschitz condition with constant C.

v. Find an example of a continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$ that does not satisfy any Lipschitz condition.

The following theorem, sometimes called *Picard's Theorem*, gives a direct application of the contraction mapping theorem to a problem in analysis.

Theorem 3.7.10 (Picard's Theorem.) Let B be a ball of radius r in $\mathbb{R}^2$ with center at (x_0, y_0). Suppose that $f : B \to \mathbb{R}$ is a continuous function that satisfies a Lipschitz condition in the second variable, that is, there is a constant C such that $|f(x,y) - f(x,y')| \leq C|y - y'|$ for all $(x,y), (x,y') \in B$. Then, there exists a $\delta > 0$ such that the differential equation $dy/dx = f(x,y)$ has a unique solution $y = \phi(x)$, satisfying $\phi(x_0) = y_0$, in the interval $|x - x_0| < \delta$.

Proof. Without loss of generality, we can assume that f is bounded on B, that is, there exists a constant M such that $|f(x,y)| \leq M$ for all $(x,y) \in B$. Take a $\delta > 0$ such that $C\delta < 1$ and $\{(x,y) \mid |x - x_0| \leq \delta, |y - y_0| \leq M\delta\} \subset B$. We now work inside the space $X = \{\phi \in \mathcal{C}([x_0 - \delta, x_0 + \delta]) \mid |\phi(x) - y_0| \leq M\delta\}$. If we give X the sup metric, then by Theorem 3.5.26 and Exercise 3.6.9, X is a complete metric space. Now, take the mapping $T : X \to X$ defined by $T\phi(x) = y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt$. It is obvious that $T\phi \in X$ and that $d(T\phi, T\phi') \leq C\delta d(\phi, \phi')$. Thus, T is a contraction mapping on X, and there is a unique function $\phi \in X$ such that $T\phi = \phi$. It is easy to check that the solutions to the differential equation are precisely the fixed points of T, so the proof is complete.

The contraction mapping theorem can also be applied to systems of differential equations, see for example [Kolmogorov-Fomin].

The following exercise assumes the reader is familiar with some basic concepts in calculus, especially integration theory.

Exercise 3.7.11 Take $\mathcal{C}([0,1], \mathbb{R})$ with the sup metric, and let $k(x,y) : [0,1] \times [0,1] \to \mathbb{R}$ be a continuous function satisfying $\sup_{0 \leq x \leq 1} \int_0^1 |k(x,y)|\, dy < 1$. Given a function $g(x) \in \mathcal{C}([0,1], \mathbb{R})$, show that there is a unique solution $f(x) \in \mathcal{C}([0,1], \mathbb{R})$ to the equation

$$f(x) - \int_0^1 k(x,y) f(y)\, dy = g(x).$$

3.8 Stone-Weierstrass Theorem

In this section, we undertake a closer examination of the space $\mathcal{C}(X, F)$, that is, the collection of continuous functions on a compact metric space X with values in the field F, where $F = \mathbb{R}$ or $\mathbb{C}$. Recall that the distance between two functions f and g in $\mathcal{C}(X, F)$ is given by $d(f,g) = \sup_{x \in X} |f(x) - g(x)|$, the "sup metric." Our investigation will lead to a nice characterization of certain dense subsets of $\mathcal{C}(X, F)$.

Definition 3.8.1 Let A be a collection of functions from a set X to F. We say that the collection A *separates points* if, for every pair of distinct points $x_1, x_2 \in X$, there is a function $f \in A$ such that $f(x_1) \neq f(x_2)$.

Example 3.8.2 If $X = [0,1]$, then $\mathcal{C}(X, \mathbb{R})$ separates points. This is easy to see just by drawing a picture.

Exercise 3.8.3

i. Show that polynomial functions in $\mathcal{C}([0,1], \mathbb{R})$ separate points.

ii. Does the class of functions $\{\sin(2\pi n x) \mid n \in \mathbb{N}\}$ in $\mathcal{C}([0,1], \mathbb{R})$ separate points?

Definition 3.8.4 A real polynomial function $f : \mathbb{R}^n \to \mathbb{R}$ is a finite linear combination of expressions of the form $x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}$ where $m_1, m_2, \ldots, m_n$, are non-negative integers. The coefficients of a polynomial may be taken from $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$. The resulting set of polynomials is denoted by $\mathbb{Z}[x_1, \ldots, x_n]$, $\mathbb{Q}[x_1, \ldots, x_n]$, and $\mathbb{R}[x_1, \ldots, x_n]$, respectively.

Example 3.8.5 A typical polynomial in $\mathbb{R}[x_1, x_2, x_3, x_4]$ looks like $\sqrt{2} x_1^3 x_2 x_3^2 x_4 + \pi x_1 x_2^5 x_4^{15} - 11 x_1 x_4$.

Exercise 3.8.6

i. Show that $R[x_1, x_2, \ldots, x_n]$ is a commutative ring with 1 for $R = \mathbb{Z}, \mathbb{Q}$, or $\mathbb{R}$. Find the units (invertible elements) in each of these rings.

ii. Find the possible images of a polynomial function in $\mathbb{R}[x]$.

iii. Find the possible images of a polynomial function in $\mathbb{R}[x_1, x_2]$.

Theorem 3.8.7 (Weierstrass) Let A be a compact set in $\mathbb{R}^n$. Then every continuous function $f : A \to \mathbb{R}$ is the uniform limit of a sequence of real polynomials in $\mathbb{R}[x_1, \ldots, x_n]$.

Theorem 3.8.8 (Stone) Let X be a compact metric space. Let A be an algebra of continuous, real valued functions on X, and suppose that A separates points. Then $\overline{A}$, the closure of A in $\mathcal{C}(X, \mathbb{R})$ under the sup metric, sometimes called the *uniform closure* of A, either coincides with $\mathcal{C}(X, \mathbb{R})$ or with $\mathcal{C}_{x_0}(X, \mathbb{R}) = \{f \in \mathcal{C}(X, \mathbb{R}) \mid f(x_0) = 0\}$, for some point $x_0 \in X$.

Remark 3.8.9 When $\overline{A} = \mathcal{C}(X, \mathbb{R})$, we say that A is *uniformly dense* in $\mathbb{R}$.

Exercise 3.8.10 Show that Stone's theorem implies Weierstrass's theorem. (Hint: Let $A = \mathbb{R}[x_1, x_2, \ldots, x_n]$.)

Before we attempt the proof of Stone's theorem, it will be helpful to gather some preliminary lemmas.

Lemma 3.8.11 Let A be an algebra of real-valued, continuous functions on a compact metric space X. Then, for $f \in A$, $|f|$ is in the uniform closure of A. That is, there is a sequence $(f_n)_{n \in \mathbb{N}}$ in A such that $(f_n)_{n \in \mathbb{N}}$ converges uniformly to $|f|$.

Proof. Since X is compact, we know that f is bounded (see Exercise 3.6.9). Choose $C \in \mathbb{R}$, $C > 0$, such that $|f(x)| \leq C$ for all $x \in X$. Let $u = \frac{1}{C} f$. Then $u \in A$, and $d(u, 0) \leq 1$. Now we construct a sequence $(w_n)_{n \in \mathbb{N}}$ in A converging uniformly to $|u|$.

Let $w_0 = 0$, and define w_n inductively by the relation

$$w_{n+1} = w_n + \frac{u^2 - w_n^2}{2}.$$

Before proceeding further, notice that if we formally take limits in n, we would have a relation of the form $w = w + (u^2 - w^2)/2$, which would imply that $w^2 = u^2$. With a little luck, we may also show that $w \geq 0$ and hence $w = |u|$.

First notice that $0 \leq w_1 - w_0 = w_1 = u^2/2 \leq u^2 \leq |u|$. Now suppose $w_k - w_{k-1} \geq 0$, and $w_k \leq |u|$ for $1 \leq k \leq n$. Then $w_n \geq 0$, and $w_{n+1} - w_n = \frac{u^2 - w_n^2}{2} = \frac{|u| + w_n}{2}(|u| - w_n) \geq 0$. Also, keeping in mind that $|u| \leq 1$, we have

$$\begin{aligned} 0 \leq w_{n+1} &= w_n + \frac{u^2 - w_n^2}{2} = w_n + (\frac{|u| + w_n}{2})(|u| - w_n) \\ &\leq w_n + |u| - w_n = |u|. \end{aligned}$$

Hence, by induction, $(w_n)_{n \in \mathbb{N}}$ is an increasing sequence of functions, and $0 \leq w_n \leq |u|$ for all n. Now, as suggested in the beginning of the proof, we let w be the pointwise limit of the sequence $(w_n)_{n \in \mathbb{N}}$. Then, $w = |u|$, and by Dini's Theorem (Theorem 3.6.11), we know that the sequence $(w_n)_{n \in \mathbb{N}}$ converges uniformly to $|u|$.

Definition 3.8.12 Let V be a vector space of real valued continuous functions on a metric space X. We say that V is a *lattice* if $|f| \in V$ whenever $f \in V$.

Exercise 3.8.13 Let V be a lattice on a metric space X. If f, g are in V, set $f \wedge g = \min(f, g)$ and $f \vee g = \max(f, g)$. Show that $f \wedge g$, $f \vee g \in V$.

Lemma 3.8.14 Let X be a compact metric space and L a lattice of continuous functions on X. Suppose that, for any $x, y \in X$ with $x \neq y$ and $a, b \in \mathbb{R}$, there is a function $f_{xy} \in L$ satisfying $f_{xy}(x) = a$ and $f_{xy}(y) = b$. Then, for each $f \in \mathcal{C}(X, \mathbb{R})$, there is a sequence $(f_n)_{n \in \mathbb{N}}$ in L such that $(f_n)_{n \in \mathbb{N}}$ converges uniformly to f.

Proof. Take $f \in \mathcal{C}(X, \mathbb{R})$ and $\varepsilon > 0$. For any $x, y \in X$, we identify the function f_{xy} and the sets U_{xy} and V_{xy} as follows. Let $a = f(x)$ and $b = f(y)$. Take $f_{xy} \in L$ such that $f_{xy}(x) = a$ and $f_{xy}(y) = b$. We take $U_{xy} = \{z \in X \mid f_{xy}(z) < f(z) + \varepsilon\}$ and $V_{xy} = \{z \in X \mid f(z) - \varepsilon < f_{xy}(z)\}$. Notice that for any $x, y \in X$, the sets U_{xy} and V_{xy} are open and, in addition, both contain x and y.

Fix y. Then by compactness, there exists a finite number of points $x_1, x_2, \ldots, x_n$ such that $\{U_{x_1y}, U_{x_2y}, \ldots, U_{x_ny}\}$ covers X. Set $h_y = \min(f_{x_1y}, f_{x_2y}, \ldots, f_{x_ny})$. By Exercise 3.8.13, we have $h_y \in L$ and $h_y(z) < f(z) + \varepsilon$ for all $z \in X$. Notice that $f(z) - \varepsilon < h_y(z)$ for $z \in V_y = \bigcap_{i=1}^{n} V_{x_iy}$.

Now let $y \in X$ vary, and for each y, construct h_y and V_y as above. By compactness, we can select an open cover $\{V_{y_1}, V_{y_2}, \ldots, V_{y_m}\}$ of X. Put $l = \max(h_{y_1}, h_{y_2}, \ldots, h_{y_m})$. Then $l \in L$ and $f(z) - \varepsilon < l(z) < f(z) + \varepsilon$.

Finally, to construct $(f_n)_{n \in \mathbb{N}}$, we let $\varepsilon = 2^{-n}$ and choose f_n to be the function l constructed above.

We are ready to return to the proof of Stone's theorem.

Proof. (of Theorem 3.8.8) There are two cases to consider. First, suppose that, for each $x_0 \in X$, there is an $f \in A$ such that $f(x_0) \neq 0$. Take $x_1, x_2 \in X$ such that $x_1 \neq x_2$. Then there is a function $f \in A$ so that $f(x_1) \neq 0$ and $f(x_1) \neq f(x_2)$. To see this, take functions $h, g \in A$ such that $g(x_1) \neq g(x_2)$ and $h(x_1) \neq 0$. Then choose f as follows.

- If $g(x_1) \neq 0$, let $f = g$.
- If $g(x_1) = 0$ and $h(x_1) \neq h(x_2)$, let $f = h$.
- If $g(x_1) = 0$ and $h(x_1) = h(x_2)$, let $f = g + h$.

If $f(x_2) \neq 0$, let $u(x) = f(x)/f(x_2) - (f(x)/f(x_2))^2$. Then $u \in A$, $u(x_1) \neq 0$ and $u(x_2) = 0$.

Hence, we can find f_1 and f_2 in A such that $f_1(x_1) = 1$, $f_1(x_2) = 0$, $f_2(x_1) = 0$, and $f_2(x_2) = 1$. Now, for any $a, b \in \mathbb{R}$, take $f_{x_1x_2}(x) = af_1(x) + bf_2(x)$. Then $f_{x_1x_2}(x_1) = a$ and $f_{x_1x_2}(x_2) = b$. From Lemma 3.8.11, we have that $\overline{A}$, the uniform closure of A, is a lattice. From Lemma 3.8.14, $\overline{A} = \mathcal{C}(X, \mathbb{R})$. This concludes the proof in the first case.

Now we turn to the case when there is an element $x_0 \in X$ such that $f(x_0) = 0$ for all $f \in A$. Let $A' = \{g \in \mathcal{C}(X, \mathbb{R}) \mid g(x) = c + f(x) \text{ for some } c \in \mathbb{R} \text{ and } f \in A\}$. We have that A' is an algebra satisfying the conditions for the first part of the theorem. In particular, if $h(x) \in \mathcal{C}_{x_0}(X, \mathbb{R})$ and $\varepsilon > 0$, then there is a function $f \in A$ and $c \in \mathbb{R}$ such that $\sup_{x \in X} |h(x) - c - f(x)| < \varepsilon$. Looking at x_0, we see that $|c| < \varepsilon$. Hence $\sup_{x \in X} |h(x) - f(x)| < 2\varepsilon$.

Exercise 3.8.15 Let X, Y be compact metric spaces. Let $A = \{(x, y) \mapsto \sum_{i=1}^{n} f_i(x)g_i(y) \mid f_i \in \mathcal{C}(X, \mathbb{R}), \text{ and } g_i \in \mathcal{C}(Y, \mathbb{R}), 1 \leq i \leq n\}$.

i. Show that A is an algebra.

ii. Show that A is uniformly dense in $\mathcal{C}(X \times Y, \mathbb{R})$.

Exercise 3.8.16

i. Prove the complex version of the Stone-Weierstrass theorem:

> Let X be a compact metric space. Let A be an algebra of continuous complex-valued functions on X with the property that if $f \in A$ then its complex conjugate $\overline{f}$ is in A. Assume that A separates points and that there is no point of $x \in X$ such that $f(x) = 0$ for all $f \in A$. Then $\overline{A} = \mathcal{C}(X, \mathbb{C})$.

ii. A *trigonometric polynomial* from $\mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$ to $\mathbb{C}$ is a function of the form $f(e^{i\theta}) = \sum_{j=-n}^{n} a_j e^{ij\theta}$, where the coefficients are in $\mathbb{C}$. Show that the set of trigonometric polynomials is uniformly dense in $\mathcal{C}(\mathbb{T}, \mathbb{C})$.

3.9 The Completion of a Metric Space

Obviously, complete metric spaces play a special role among all metric spaces. We now present a procedure through which any metric space can be embedded as a dense subset of a complete metric space.

Theorem 3.9.1 Let (X, d) be a metric space. Then there exists a complete metric space $(\tilde{X}, \tilde{d})$, and an injection $\phi : X \to \tilde{X}$, such that

1. $\phi : X \to \phi(X)$ is an isometry, and
2. $\phi(X)$ is dense in $\tilde{X}$.

Proof. Consider the set X' of all Cauchy sequences in X. We define an equivalence relation on X' by saying that $(x_n)_{n\in\mathbb{N}}$ is equivalent to $(y_n)_{n\in\mathbb{N}}$ if $\lim_{n\to\infty} d(x_n, y_n) = 0$.

Exercise 3.9.2 Prove that this is an equivalence relation.

Let $\tilde{X}$ be the set of equivalence classes. We denote the equivalence class of a Cauchy sequence $(x_n)_{n\in\mathbb{N}}$ by $\{(x_n)_{n\in\mathbb{N}}\}$. We first define a metric on $\tilde{X}$. Let $\{(x_n)_{n\in\mathbb{N}}\}$ and $\{(x'_n)_{n\in\mathbb{N}}\}$ be elements of $\tilde{X}$. We note that $(d(x_n, x'_n))_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. This follows from the fact that $|d(x_n, x'_n) - d(x_m, x'_m)| \le d(x_n, x_m) + d(x'_n, x'_m)$. We set

$$\tilde{d}(\{(x_n)_{n\in\mathbb{N}}\}, \{(x'_n)_{n\in\mathbb{N}}\}) = \lim_{n\to\infty} d(x_n, x'_n).$$

This limit exists by the Cauchy criterion (see Theorem 1.6.24).

Exercise 3.9.3 Show that $\tilde{d}$ is well-defined.

Now define $\phi : X \to \tilde{X}$ by $\phi(x) = \{(x_k)_{k\in\mathbb{N}}\}$ where $x_k = x$ for all $k \in \mathbb{N}$. It is clear that ϕ is an isometry from X to $\phi(X)$.

There are two things left to do. First, show $\phi(X)$ is dense in $\tilde{X}$, and second, show that $(\tilde{X}, \tilde{d})$ is complete.

Let $\tilde{x} = \{(x_n)_{n\in\mathbb{N}}\} \in \tilde{X}$. Pick $\varepsilon > 0$. Since the sequence $(x_n)_{n\in\mathbb{N}}$ is Cauchy in X, there exists an integer N such that $d(x_N, x_m) < \varepsilon$ if $m \ge N$. Now consider the class of the constant sequence $\phi(x_N)$. Then $\tilde{d}(\tilde{x}, \phi(x_N)) = \lim_{n\to\infty} d(x_n, x_N) \le \varepsilon$ and hence $\phi(X)$ is dense in $\tilde{X}$.

To show that $\tilde{X}$ is complete, take a Cauchy sequence $(\tilde{y}_n)$ in $\tilde{X}$. Remember, each $\tilde{y}_n$ is an equivalence class of Cauchy sequences in X. For each $n \in \mathbb{N}$, by density, choose $\tilde{z}_n \in \phi(X)$ such that $\tilde{d}(\tilde{y}_n, \tilde{z}_n) < \frac{1}{n}$. Then $\tilde{d}(\tilde{z}_n, \tilde{z}_m) \le \tilde{d}(\tilde{z}_n, \tilde{y}_n) + \tilde{d}(\tilde{y}_n, \tilde{y}_m) + \tilde{d}(\tilde{y}_m, \tilde{z}_m) < \frac{1}{n} + \tilde{d}(\tilde{y}_n, \tilde{y}_m) + \frac{1}{m}$. This implies that $(\tilde{z}_n)_{n\in\mathbb{N}}$ is Cauchy in $\tilde{X}$. Let $x_n = \phi^{-1}(\tilde{z}_n)$. Then, since ϕ is an isometry, $(x_n)_{n\in\mathbb{N}}$ is Cauchy in X. Let $\tilde{y}$ be the element of $\tilde{X}$ defined by the equivalence class of this Cauchy sequence, that is, $\tilde{y} = \{(x_n)_{n\in\mathbb{N}}\}$. Then, $\tilde{d}(\tilde{y}_n, \tilde{y}) \le \tilde{d}(\tilde{y}_n, \tilde{z}_n) + \tilde{d}(\tilde{z}_n, \tilde{y}) < \frac{1}{n} + \tilde{d}(\tilde{z}_n, \tilde{y})$. Observe that $\tilde{d}(\tilde{z}_n, \tilde{y}) = \lim_{k\to\infty} d(x_n, x_k)$. Since $(x_n)_{n\in\mathbb{N}}$ is Cauchy in X, for n and k large, $d(x_n, x_k)$ can be made arbitrarily small. Thus, $(\tilde{y}_n)_{n\in\mathbb{N}}$ converges to $\tilde{y}$. ☠

Definition 3.9.4 The metric space $\tilde{X}$ in the above theorem is called the *completion* of X.

Exercise 3.9.5 If (X, d) is already a complete metric space, show that (X, d) and $(\tilde{X}, \tilde{d})$ are isometric.

Exercise 3.9.6 Prove that $(\tilde{X}, \tilde{d})$ is unique up to isometry. That is, if (X', d') is a complete metric space such that X is isometric to a dense subset of X', then $(\tilde{X}, \tilde{d})$ and (X', d') are isometric.

Remark 3.9.7 One might ask at this point, "Why did we write Chapter 1 at all?" Why not just take the rational numbers with the usual metric and complete them by the above process to get the real numbers? Sorry folks, but in the proof of the above theorem, we used the fact that the real numbers are complete. In Project **??**, we will have a simple, yet significant example of the completion of a metric space, namely, the p-adic completion of $\mathbb{Q}$ relative to a prime p. This emphasizes the fact that while $\mathbb{R}$ is the most familiar example of a completion of $\mathbb{Q}$ with respect to a metric, there are in fact infinitely many other completions of $\mathbb{Q}$.

3.10 Independent Projects

3.10.1 The p-adic completion of $\mathbb{Q}$The simplest example of the completion of an incomplete metric space is called the p-adic completion of $\mathbb{Q}$. The p in this case refers to a prime integer p, and the metric is that defined below. This metric plays a significant role in analysis, number theory, theoretical physics, and other areas.

Definitions and basic properties

Definition 3.10.1 Let p be a prime in $\mathbb{Z}$. For $r \in \mathbb{Q}^\times$, we write $r = p^k(\frac{a}{b})$, where a and b are relatively prime integers not divisible by p. Define the *p-adic absolute value* $|\cdot|_p$ on $\mathbb{Q}$ by

$$|r|_p = p^{-k} \text{ if } r \neq 0 \quad \text{and} \quad |0|_p = 0.$$

Exercise 3.10.2 Show that $|\cdot|_p$ has the following properties for all $r, s \in \mathbb{Q}$:

i. $|r|_p \geqslant 0$, and $|r|_p = 0$ if and only if $r = 0$;

ii. $|rs|_p = |r|_p \cdot |s|_p$;

iii. $|r + s|_p \leq \max(|r|_p, |s|_p)$;

iv. $|r + s|_p = \max(|r|_p, |s|_p)$ if $|r|_p \neq |s|_p$.

Note that *i* and *ii* are familiar properties of the usual absolute value on $\mathbb{Q}$, while *iii*, known as *the non-Archimedean Triangle Inequality*, is stronger than the usual triangle inequality on $\mathbb{Q}$, which asserts that

$$|r + s| \leq |r| + |s|, \qquad r, s \in \mathbb{Q}.$$

The absolute value $|\cdot|_p$ gives a metric on $\mathbb{Q}$ defined by

$$d_p(r, s) = |r - s|_p, \qquad r, s \in \mathbb{Q}.$$

Exercise 3.10.3

i. Show that d_p is a metric.

ii. Find a Cauchy sequence in $\mathbb{Q}$ relative to d_p that does not converge in $\mathbb{Q}$. That is, $\mathbb{Q}$ is not complete with respect to d_p.

We denote by $\mathbb{Q}_p$ the completion of $\mathbb{Q}$ with respect to the metric d_p. We can define addition and multiplication on $\mathbb{Q}_p$ such that $\mathbb{Q}_p$ becomes a field. Recall that elements of $\mathbb{Q}_p$ are equivalence classes of Cauchy sequences from $\mathbb{Q}$ relative to d_p. The process of turning $\mathbb{Q}_p$ into a field proceeds exactly as in the case of the real numbers (see section 1.5).

Definition 3.10.4 Addition and multiplication on $\mathbb{Q}_p$ are defined as

$$\{(a_n)_{n\in\mathbb{N}}\} + \{(b_n)_{n\in\mathbb{N}}\} = \{(a_n + b_n)_{n\in\mathbb{N}}\}, and$$
$$\{(a_n)_{n\in\mathbb{N}}\} \cdot \{(b_n)_{n\in\mathbb{N}}\} = \{(a_n b_n)_{n\in\mathbb{N}}\}.$$

Next, we must extend $|\cdot|_p$ to $\mathbb{Q}_p$. Observe that, if $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{Q}$ with respect to d_p, then $(|a_n|_p)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. So if $\{(a_n)_{n\in\mathbb{N}}\} \in \mathbb{Q}_p$, then the absolute value on $\mathbb{Q}_p$ can be defined by

$$|\{(a_n)_{n\in\mathbb{N}}\}|_p = \lim_{n\to\infty} |a_n|_p$$

Note that if $\lim\limits_{n\to\infty} |a_n|_p \neq 0$, then the sequence $(|a_n|_p)_{n\in\mathbb{N}}$ is eventually constant and hence converges to the eventual constant.

Exercise 3.10.5

i. Show that addition, multiplication, and $|\cdot|_p$ are well-defined on $\mathbb{Q}_p$.

ii. Show that $\mathbb{Q}_p$ is a field with the operations given above.

iii. Show that $|\cdot|_p$ on $\mathbb{Q}_p$ satisfies the same properties as it does on $\mathbb{Q}$ (see (3.10.2)).

iv. Show that the image of $\mathbb{Q}_p$ under $|\cdot|_p$ is the same as that of $\mathbb{Q}$ under $|\cdot|_p$, that is, $\{p^k \mid k \in \mathbb{Z}\} \cup \{0\}$.

v. Show that $\mathbb{Q}_p$ cannot be made into an ordered field.

Definition 3.10.6 The field $\mathbb{Q}_p$ with $|\cdot|_p$ is called a *p-adic field*. It is also called the *p-adic completion of* $\mathbb{Q}$.

3.10.2 The additive structure of $\mathbb{Q}_p$

We begin by defining several sets in $\mathbb{Q}_p$ that play an important role in our study of p-adic fields.

Definition 3.10.7 Define the following subsets of $\mathbb{Q}_p$:

a. $R_p = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\}$;

b. $\wp = \{x \in R_p \mid |x|_p < 1\} = \{x \in R_p \mid |x|_p \leq \frac{1}{p}\}$; and

c. $U_p = \{x \in R_p \mid |x|_p = 1\}$.

The set R_p is called the *ring of integers in* $\mathbb{Q}_p$. The set $\wp$ is called the *maximal ideal* in R_p. The set U_p is called the *group of units* in R_p.

Exercise 3.10.8 Show that R_p is a commutative ring with 1.

Proposition 3.10.9 The set $\wp$ is a subgroup of R_p, and $R_p = \bigcup_{0\leq k\leq p-1} k + \wp$.

Proof. It follows from the non-Archimedean triangle inequality that $\wp$ is an additive subgroup of R_p. Let $x \in R_p$. If $|x|_p < 1$, then $x \in \wp$. Suppose $|x|_p = 1$. Since $\mathbb{Q}$ is dense in $\mathbb{Q}_p$, there is some $r \in \mathbb{Q}$ such that $r = a/b$ with $(a, b) = (a, p) = (b, p) = 1$ and $|r - x|_p < 1$. Hence, $x + \wp = r + \wp$. Since p and b are relatively prime, there exists an integer k with $0 < k \leq p-1$ such that p divides $a - kb$. Hence, $|a - kb|_p < 1$, and also $|\frac{a-kb}{b}|_p < 1$ by Exercise 3.10.8 since $p \nmid b$. Thus, $|k - \frac{a}{b}|_p < 1$. It follows that $k + \wp = r + \wp = x + \wp$ so that $x \in k + \wp$. ∎

Exercise 3.10.10

i. Show that U_p is, in fact, the set of *units* in R_p, that is the set of elements in R_p that have multiplicative inverses in R_p.

ii. Show that U_p is a group under multiplication.

iii. Show that $\wp$ is an *ideal* in R_p, that is, if a is in $\wp$ and $x \in R_p$, then $ax \in \wp$.

iv. Show that $\wp$ is a *maximal ideal* in R_p. That is, if $x \in U_p$, then the smallest ideal containing x and $\wp$ is all of R_p.

v. For $n \in \mathbb{Z}$, define $\wp^n = p^n R_p = \{p^n x \mid x \in R_p\} = \{x \in \mathbb{Q}_p \mid |x| \leq p^{-n}\}$. Show that $\wp^n$ is a subgroup of $(\mathbb{Q}_p, +)$.

vi. Show that $\wp^n \setminus \wp^{n+1} = p^n U_p$.

vii. Show that, if $n > 0$, $\wp^n$ is an ideal in R_p, that is, if $a \in \wp^n$ and $x \in R_p$, then $ax \in \wp^n$.

viii. Show that $\mathbb{Q}_p = \bigcup_{n\in\mathbb{Z}} \wp^n$.

ix. Show that $\mathbb{Q}_p^\times = \bigcup_{n\in\mathbb{Z}} p^n U_p$.

Definition 3.10.11 If n is an integer, the set $p^n U_p$ is called a *shell* in $\mathbb{Q}_p$.

3.10.3 The topological structure of $\mathbb{Q}_p$

We now consider the topology on $\mathbb{Q}_p$ determined by the metric d_p associated with $|\cdot|_p$.

Exercise 3.10.12 If $x_0 \in \mathbb{Q}_p$ and $r > 0$, show that there is an integer n such that $B_r(x_0) = B_{p^{-n}}(x_0) = x_0 + \wp^{n+1} = \{x \in \mathbb{Q}_p \mid |x - x_0|_p < p^{-n}\} = \{x \in \mathbb{Q}_p \mid |x - x_0|_p \leq p^{-n-1}\}$.

This shows that the open balls in $\mathbb{Q}_p$ are simply cosets of some power of $\wp$.

Proposition 3.10.13 For each $n \in \mathbb{Z}$, the subsets $\wp^n$ and $p^n U_p$ are both open and closed in $\mathbb{Q}_p$.

Proof. First, consider $p^n U_p$ for some $n \in \mathbb{Z}$. If $x \in p^n U_p$, then $|x|_p = p^{-n}$. If $k > n$, then the ball $x + \wp^k$ is contained in $p^n U_p$ by Exercise 3.10.5.*iii*. This proves that $p^n U_p$ is open. Now, consider $\wp^n$. If $x \in \wp^n$ and $k > n$, then the ball $x + \wp^k$ is contained in $\wp^n$. Hence $\wp^n$ is open. To show that $\wp^n$ is closed, notice that $\mathbb{Q}_p \setminus \wp^n = \bigcup_{k<n} p^k U_p$, which is open. Finally, $p^n U_p$ is the complement of $\wp^{n+1}$ in $\wp^n$ so that $p^n U_p$ is closed by Exercise 3.3.22.

Corollary 3.10.14 If $n \in \mathbb{Z}$ and $x \in \mathbb{Q}_p$, then $x + \wp^n$ is both open and closed.

Corollary 3.10.15 Any open set A in $\mathbb{Q}_p$ can be written as a disjoint union of cosets of the subgroups $\wp^n$, $n \in \mathbb{Z}$.

Proof. If A is empty, then we are done, so suppose that it is not. Suppose further that A is bounded. Then the set S of integers n such that A contains some coset of $\wp^n$ is bounded below. By the Well Ordering Principle, applied to a suitable shift of S, we see that S has a least element n_0. Let $A_0 = a_0 + \wp^{n_0}$ be a coset of $\wp^{n_0}$ contained in A. By Corollary 3.10.14, A_0 is closed, so $A \setminus A_0$ is open. If $A \setminus A_0$ is empty, then we are done. Otherwise, repeat to get A_1, and so on.

Exercise 3.10.16 Prove that this algorithm terminates, so that we have written A as a disjoint union of cosets $A_0, A_1, \ldots$ of the desired form.

Exercise 3.10.17 Explain how to reduce the case of general A to the case of bounded A. (Hint: Consider the intersection of an arbitrary open set A with cosets of R_p.)

It is now easy to prove some of the basic topological properties of $\mathbb{Q}_p$.

Exercise 3.10.18 Show that the ring of integers R_p is a maximal compact subring of $\mathbb{Q}_p$.

Exercise 3.10.19 Show that the field $\mathbb{Q}_p$ has the Bolzano-Weierstrass property, that is, if A is a bounded infinite subset of $\mathbb{Q}_p$, then A has an accumulation point in $\mathbb{Q}_p$.

Exercise 3.10.20 Show that $\mathbb{Q}_p$ has the Heine-Borel property, that is, if A is a closed, bounded subset of $\mathbb{Q}_p$, then A is compact.

Exercise 3.10.21 Show that the field $\mathbb{Q}_p$ is a locally compact field, that is, every point in $\mathbb{Q}_p$ has a neighborhood whose closure is compact. In fact, show that every point in $\mathbb{Q}_p$ has a neighborhood that is both open and compact.

We now introduce some exercise concerning the cosets of $\wp$ in R_p

Exercise 3.10.22 Recall from Proposition **??** that $R_p = \cup_{0\leq k\leq p-1}(k+\wp)$. Show that these p cosets are disjoint.

Exercise 3.10.23 Show that $U_p = \cup_{1\leq k\leq p-1}(k+\wp)$.

We introduce an algebraic structure on the collection of cosets as follows. We define $(k+\wp)+(j+\wp) = (k+j)+\wp$, and $(k+\wp)(j+\wp) = kj+\wp$.

Exercise 3.10.24 Show that the addition and multiplication so defined are well-defined.

Exercise 3.10.25 Show that the collection $F = \{k+\wp \mid 0 \leq k \leq p-1\}$ is a field with these two operations.

Exercise 3.10.26 Show that F is isomorphic to $\mathbb{Z}_p$, the field of integers modulo p.

Exercise 3.10.27 For a fixed $n \in \mathbb{Z}$, show that

$$p^n U_p = \bigcup_{1\leqslant k\leqslant p-1} (p^n k + \wp^{n+1}).$$

Exercise 3.10.28 The ring of ordinary integers $\mathbb{Z}$ is dense in R_p relative to $|\cdot|_p$.

Definition 3.10.29 The *valuation map* $\nu : \mathbb{Q}_p \to \mathbb{Z}\cup\{+\infty\}$ is defined by the following rule:

$$\begin{aligned} p^{-\nu(x)} &= |x|_p \quad \text{if } x \neq 0, \\ \nu(0) &= +\infty. \end{aligned}$$

(See Exercise 3.10.5.*iv*.

With this definition, we can now write:

a. $R_p = \{x \in \mathbb{Q}_p \mid \nu(x) \geq 0\}$;

b. $\wp = \{x \in R_p \mid \nu(x) > 0\}$;

c. $U_p = \{x \in R_p \mid \nu(x) = 0\}$.

We now consider the convergence of infinite series in $\mathbb{Q}_p$. The situation here is simpler than that in Section 1.9.

In real and complex analysis, determining whether or not an infinite series converges can be a delicate matter. The p-adic case is different.

Theorem 3.10.30 Let $a_n \in \mathbb{Q}_p$ for all $n \in \mathbb{N}$. Then $\displaystyle\sum_{n=1}^{\infty} a_n$ converges in $\mathbb{Q}_p$ if and only if $\displaystyle\lim_{n\to\infty} a_n = 0$.

Proof. The "only if" part is clear, just as in the real and complex cases. (See Theorem 1.9.9.)

Now suppose that $\displaystyle\lim_{n\to\infty} a_n = 0$. This means that, given $k \geq 0$, we can pick $N \in \mathbb{N}$ such that $|a_n|_p < p^{-k}$ for all $n > N$. Thus, for all $m > n > N$

$$|s_m - s_n|_p = |a_{n+1} + \cdots + a_m|_p \leq \max_{n+1\leq i\leq m} |a_i|_p < p^{-k},$$

the first inequality following from the non-Archimedean Triangle Inequality. Therefore, the sequence $(s_n)_{n\in\mathbb{N}}$ of partial sums is Cauchy, and so it must converge by the completeness of $\mathbb{Q}_p$.

From the decomposition $\mathbb{Q}_p^\times = \bigcup_{n\in\mathbb{Z}} p^n U_p$ into shells (see Exercise 3.10.10.*ix*, one can express any non-zero x as an infinite series $x = \sum a_k p^k$, where the $a_k \in \{0, 1, \ldots, p-1\}$ are uniquely determined, and there only finitely many $k < 0$ (possibly none) for which $a_k \neq 0$. In fact, the first non-zero term in the series is the one corresponding to the valuation of x, and one can write x in a *p-adic expansion*:

$$x = \sum_{k=\nu(x)}^{\infty} a_k p^k,$$

where $a_{\nu(x)} \neq 0$.

It follows immediately from Theorem 3.10.30 that the p-adic expansion of x converges to x.

Exercise 3.10.31 Find the 5-adic expansion of $x = \frac{1}{3}$ in $\mathbb{Q}_5$.

Exercise 3.10.32 Let x and y be the two square roots of 11 in $\mathbb{Q}_5$. Find the 5-adic expansions of x and y. (What is the 5-adic expansion of $x + y$?)

Exercise 3.10.33 Show that there is no x in $\mathbb{Q}_5$ such that $x^2 = 7$.

Exercise 3.10.34 Show that a rational number has a periodic p-adic expansion and determine the length of the period.

3.10.4 Fundamental Theorem of AlgebraHere is a proof of the Fundamental Theorem of Algebra promised in Chapter ??.

Exercise 3.10.35 Let P be a polynomial of positive degree with coefficients in $\mathbb{C}$. Show that there exists $z_0 \in \mathbb{C}$ such that $|P(z_0)| \leq |P(z)|$ for all $z \in \mathbb{C}$. Then show that, by considering the polynomial $P(z + z_0)$, we may assume $z_0 = 0$.

Theorem 3.10.36 (The Fundamental Theorem of Algebra) The field $\mathbb{C}$ is algebraically closed, that is, any nonconstant polynomial with coefficients in $\mathbb{C}$ has a root in $\mathbb{C}$.

Proof. Let $P \in \mathbb{C}[z]$ be a polynomial of positive degree. By Exercise 3.10.35, we may assume $P(z)$ has a minimum at 0. There exists $n \geq 1$ and $a, b \in \mathbb{C}$ with $b \neq 0$ such that

$$P(z) = a + bz^n + z^{n+1}Q(z),$$

where $Q \in \mathbb{C}[z]$. Suppose that $P(0) = a \neq 0$, and choose an n-th root w of $-a/b$ in $\mathbb{C}$.

By continuity, there exists t with $0 < t < 1$ such that $t|w^{n+1}Q(tw)| < |a|$. Now, we have

$$\begin{aligned} P(tw) &= a + b(tw)^n + (tw)^{n+1}Q(tw) \\ &= (1 - t^n)a + (tw)^{n+1}Q(tw) \end{aligned}$$

because $bw^n = -a$. Hence,

$$\begin{aligned} |P(tw)| &\leq (1 - t^n)|a| + t^{n+1}|w^{n+1}Q(tw)| \\ &< (1 - t^n)|a| + t^n|a| = |a| = |P(0)|. \end{aligned}$$

This is a contradiction, and hence we must have $P(0) = a = 0$.

Exercise 3.10.37 Find 10 other proofs of the fundamental theorem of algebra.

Starting with any field F we wish to define an algebraic closure of F. We first define an algebraic extension of F.

Definition 3.10.38 Let F be a field, and let E be a field containing F as a subfield of E. We say that E is an *algebraic extension* of F if given $\alpha \in E$, there exists a nonzero polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$.

Definition 3.10.39 Let F a field. A field E containing F is an *algebraic closure* of F, if E is an algebraic extension of F and E is algebraically closed.

The following sequence of statements leads to the existence and uniqueness, up to isomorphism, of an algebraic closure of F.

Facts 3.10.40

1. The field F is contained in an algebraically closed field E.
2. There is an extension E of F that is both algebraically closed and algebraic over F.
3. Suppose F is a field and E is an algebraic extension of F. Let σ be a monomorphism (injective homomorphism) of F into an algebraically closed field L. Then σ can be extended to a monomorphism of E into L.
4. If L and L' are algebraically closed fields that are algebraic over F, then there exists an isomorphism $\tau : L \to L'$ such that τ is the identity on F.

Exercise 3.10.41 Prove the above statements. Use Lang's *Algebra*, Chapter 5, if you must.

Exercise 3.10.42 Show that if F is algebraically closed, then the algebraic closure of F is F.

Remark 3.10.43 The Fundamental Theorem of Algebra shows that $\mathbb{C}$ is algebraically closed and, in fact, that $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$. (You should prove, by writing down polynomials, that $\mathbb{C}$ is an algebraic extension of $\mathbb{R}$.)

Exercise 3.10.44

i. Show that $\mathbb{A}_{\mathbb{R}}$, the field of real algebraic numbers, is not algebraically closed.

ii. Show that $\mathbb{A}$, the field of algebraic numbers, is the algebraic closure of $\mathbb{Q}$.

Chapter 4

Differentiation

> *The actual state of [infinitesimal calculus] at German universities still shows the same colorful diversity. On the one hand, there is the rigorous approach that begins with a six weeks long Dedekind treatment, and then deduces as natural and necessary consequences the concrete rules of differentiation and integration from the general concepts of number and function. On the other, there is the intuitive treatment that lets reign the magic of the differential, and even in the last hour of a two semester course does not dissipate by the sunshine of a clear conception of limit, the fog surrounding the indivisibles. And, in between, there are the shadings of hundreds of diagonals that one can construct between the two perpendicular directions of thought.*
>
> – *Otto Toeplitz, Jahresbericht dur Deutschen Mathematiker (1927)*

4.1 Review of Differentiation in One Variable

We assume that the reader is familiar with the standard properties of the derivative in one variable, and we will not review the computational aspects of elementary calculus. However, we shall establish rigorously those properties of the derivative in one variable that stem from the completeness of the real numbers. Many of the aspects of differentiation which occur in several variables are motivated by, and rely on, results in one variable.

We begin by defining the derivative of a real-valued function of one variable at a point.

Definition 4.1.1 Let $[a, b]$ be an interval in $\mathbb{R}$, and consider $f : [a, b] \to \mathbb{R}$. We say that f is *differentiable* at a point $x \in (a, b)$ if there exists $L \in \mathbb{R}$ such that

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

Observe that this definition can be phrased in the following way. The function f is differentiable at $x \in (a, b)$ if there exists $L \in \mathbb{R}$ such that

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

The number L is called the *derivative* of f at x, and is denoted by $f'(x)$ or $Df(x)$.

Exercise 4.1.2 If L exists, show that it is unique.

Exercise 4.1.3 Show that f is differentiable at $x \in (a, b)$ iff there exists a constant L such that

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

Exercise 4.1.4 Show that f is differentiable at $c \in (a, b)$ iff $\lim_{x \to c} \frac{f(x)-f(c)}{x-c}$ exists, and show that when this limit does exist, it is equal to $f'(c)$.

The reader should be familiar with the derivative of a function at a point. Differentiability is a *pointwise property* of functions, that is, it is possible for a function to be differentiable at one point and nowhere else (see Example 4.1.6 below).

Theorem 4.1.5 Suppose $f : [a, b] \to \mathbb{R}$ is differentiable at a point $x \in (a, b)$. Then f is continuous at x.

Proof. Take $\varepsilon = 1$. Then there exists a $\delta > 0$ such that $|f(x+h) - f(x) - f'(x)h| < \varepsilon|h| = |h|$ whenever $|h| < \delta$. It follows from the triangle inequality that

$$|f(x+h) - f(x)| < |h| + |f'(x)||h| = (1 + |f'(x)|)|h|$$

when $|h| < \delta$. Letting $h \to 0$, we get the result.

Example 4.1.6 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$ if x is rational and 0 if x is irrational. This function is discontinuous at every nonzero x. On the other hand, f is continuous at $x = 0$, and

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

The expression whose limit we are evaluating is equal to either h or 0 depending on whether h is rational or irrational, respectively. Thus, the limit as h approaches 0 is 0, and thus $f'(0) = 0$. Hence f is differentiable at $x = 0$ and nowhere else.

Exercise 4.1.7 Generalize the function from Exercise 3.5.4. Let $r \geq 1$, and set

$$f_r(x) = \begin{cases} \frac{1}{q^r} & \text{if } x = \frac{p}{q} \text{ in lowest terms, and } x \neq 0, \text{ and} \\ 0 & \text{if } x = 0 \text{ or } x \text{ is irrational.} \end{cases}$$

i. Show that for any $r \geq 1$, f_r is continuous at 0 and the irrational numbers and is discontinuous at the nonzero rationals.

ii. If $1 \leq r \leq 2$, show that f_r is not differentiable at any irrational point. (Hint: Use Theorem 1.3.9.)

iii. For which r is f_r differentiable at $x = 0$?

To settle all discussions about the relationship between differentiability and continuity, consider the following example.

Example 4.1.8 We want to create a continuous function on the interval $[0, \infty)$ that is not differentiable at any point in that interval. Define

$$f_1(x) = \begin{cases} x & \text{if } x \leq 1/2 \\ 1 - x & \text{if } 1/2 \leq x \leq 1 \end{cases}$$

and extend periodically to $[0, \infty)$ by $f_1(x+1) = f_1(x)$. Then define for all $n \geq 2$, $f_n(x) = \frac{1}{2} f_{n-1}(2x)$. Let $S_m(x) = \sum_{n=1}^{m} f_n(x)$. Then S_m is a continuous function on $[0, \infty)$.

Exercise 4.1.9 Show that the sequence $(S_m)_{m\in\mathbb{N}}$ converges uniformly to a continuous function S.

Exercise 4.1.10 Show that S is not differentiable at any point in $(0,\infty)$.

Theorem 4.1.11 *Suppose $f : [a,b] \to \mathbb{R}$ and $g : [a,b] \to \mathbb{R}$ are both differentiable at $x \in (a,b)$. Then for any $\alpha, \beta \in \mathbb{R}$, $\alpha f + \beta g$ is differentiable at x. Also, the product fg and the quotient $\frac{f}{g}$ are differentiable at x (for $\frac{f}{g}$ we must have $g(x) \neq 0$). Then we have*

i. $(\alpha f + \beta g)'(x) = \alpha f'(x) + \beta g'(x)$;

ii. $(fg)'(x) = f(x)g'(x) + f'(x)g(x)$; and

iii. $\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x)-f(x)g'(x)}{[g(x)]^2}$.

Proof. Look in a rigorous calculus book.

Before proceeding further, we want to create a setting which will prevail throughout the theory of differentiation. We said above that differentiability is a pointwise property. Generally speaking, we will assume that a function is not only differentiable at a point, but at all points in a neighborhood of a given point. It is rarely the case that we deal with functions which are differentiable only at a single point.

We want to pay special attention to the derivative of a composition of functions, sometimes known as the *chain rule*. The proof here takes a little care and will require even more care in several variables.

Theorem 4.1.12 (Chain Rule) *Let f be differentiable at a point a and let g be differentiable at $f(a)$. Then $g \circ f$ is differentiable at a and $D(g \circ f)(a) = (Dg)(f(a))Df(a)$.*

Proof. Let $b = f(a)$, $L = Df(a)$ and $M = Dg(b)$. Set

$$\begin{aligned} F(x) &= f(x) - f(a) - L(x-a), \\ G(y) &= g(y) - g(b) - M(y-b), \\ H(x) &= (g\circ f)(x) - (g \circ f)(a) - ML(x-a). \end{aligned}$$

By hypothesis, we know that

$$\lim_{x\to a} \frac{|F(x)|}{|x-a|} = \lim_{y\to b} \frac{|G(y)|}{|y-b|} = 0.$$

To prove the theorem, we must show that

$$\lim_{x\to a} \frac{|H(x)|}{|x-a|} = 0.$$

Notice that $H(x) = G(f(x)) + MF(x)$. Now,

$$\frac{|MF(x)|}{|x-a|} = |M| \frac{|F(x)|}{|x-a|} \to 0$$

as $x \to a$. For the remaining term, it follows from above that given $\varepsilon > 0$, there exists a $\delta > 0$ such that $|y-b| < \delta$ implies $|G(y)| < \varepsilon|y-b|$. The continuity of f at a implies that there exists a $\delta' > 0$ such that $|f(x) - b| < \delta$ when $|x-a| < \delta'$. Hence, if $|x-a| < \delta'$, we have $|G(f(x))| < \varepsilon|f(x)-b|$. But $|f(x)-b| \leq |F(x)| + |L||x-a|$, so

$$\frac{|G(f(x))|}{|x-a|} \to 0$$

as $x \to a$.

Exercise 4.1.13 Give a critique of the following supposed proof of the chain rule:

$$\begin{aligned}
\lim_{x\to a}\frac{g(f(x))-g(f(a))}{x-a} &= \lim_{x\to a}\frac{g(f(x))-g(f(a))}{f(x)-f(a)}\frac{f(x)-f(a)}{x-a}\\
&= \left(\lim_{x\to a}\frac{g(f(x))-g(f(a))}{f(x)-f(a)}\right)\left(\lim_{x\to a}\frac{f(x)-f(a)}{x-a}\right)\\
&= Dg(f(a))Df(a).
\end{aligned}$$

Suppose that $[a, b]$ is a closed interval in $\mathbb{R}$ and $f : [a, b] \to \mathbb{R}$ is continuous. Assume f is differentiable on (a, b). Since f is continuous, we know that f assumes a maximum and minimum value on $[a, b]$. This observation leads to the following familiar fact from elementary calculus.

Theorem 4.1.14 Suppose that f satisfies the hypotheses above and f assumes a local maximum or minimum at a point $c \in (a, b)$. Then $f'(c) = 0$.

Proof. Assume that f has a local maximum at c. There exists $\varepsilon > 0$ such that if $|x-c| < \varepsilon$ then $f(x) \leq f(c)$. It follows that if $x \in (c-\varepsilon, c)$ then $\frac{f(x)-f(c)}{x-c} \geq 0$ and if $x \in (c, c+\varepsilon)$ then $\frac{f(x)-f(c)}{x-c} \leq 0$. Thus, in evaluating $f'(c) = \lim_{x\to c}\frac{f(x)-f(c)}{x-c}$, the former inequality tells us that this limit, if it exists, must be greater than or equal to 0, and the latter shows that if it exists, it must be less than or equal to 0. Since we know by hypothesis that f is differentiable at c, the limit does exist and can only equal 0. The proof for a local minimum is similar.

We now take up various versions of the Mean Value Theorem. The Mean Value Theorem can be regarded as the most important theorem in analysis both in one and several variables. The statements in the following theorem are often called *Rolle's Theorem*, the *Mean Value Theorem* (MVT), and the *Generalized Mean Value Theorem.* We state them all here so that the reader will be in familiar territory. We prove only the Generalized Mean Value Theorem, which immediately implies the other two. The Generalized Mean Value Theorem is often called *Cauchy's Mean Value Theorem.*

Theorem 4.1.15 (Mean Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous and suppose that f is differentiable on (a, b).

i. If $f(a) = f(b)$, then there exists $c \in (a, b)$ such that $f'(c) = 0$.

ii. In any case, there exists $c \in (a, b)$ such that $f'(c)(b-a) = f(b) - f(a)$.

iii. If g satisfies the same hypotheses as f, then there exists $c \in (a, b)$ such that $(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c)$.

Proof. To prove *iii*, set $h(x) = (f(b) - f(a))g(x) - (g(b) - g(a))f(x)$. Then h is continuous on $[a, b]$ and differentiable on (a, b). Note that $h(a) = h(b)$. If h is constant, we are done. If not, h assumes a max or min at some point $c \in (a, b)$. Theorem 4.1.14 says that $h'(c) = 0$. The conclusion follows.

The Mean Value Theorem has some serious applications. The first is important for the Fundamental Theorem of Calculus.

Corollary 4.1.16 Suppose f is continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant.

Proof. Given any two points x and y in $[a, b]$ with $x < y$, there exists a point $c \in (x, y)$ such that $f(y)-f(x) = (y-x)f'(c) = 0$. Hence, $f(x) = f(y)$.

Corollary 4.1.17 Suppose f is continuous on $[a, b]$ and differentiable on (a, b). If $f'(c) > 0$ for all $c \in (a, b)$, then f is monotonic increasing.

Proof. For $x, y \in [a, b]$ with $x < y$, there exists $c \in (x, y)$ such that $f(y) - f(x) = (y - x)f'(c) > 0$.

Exercise 4.1.18 Suppose f is continuous on $[a, b]$ and differentiable on (a, b), and $f'(c) < 0$ for all $c \in (a, b)$. Show that f is monotonic decreasing.

Exercise 4.1.19 (L'Hôpital's Rule) Let (a, b) be any open interval in $\mathbb{R}$, and suppose f and g are differentiable on (a, b). Take $c \in (a, b)$, and suppose $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$ and $\lim_{x\to c} \frac{f'(x)}{g'(x)}$ exists. Show that $\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f'(x)}{g'(x)}$.

We now pause to present one of the all time favorite examples in one variable differentiation.

Example 4.1.20 Let

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

Then

$$f'(x) = \begin{cases} 2x \sin(1/x) - \cos(1/x) & \text{when } x \neq 0, \\ 0 & \text{when } x = 0. \end{cases}$$

So even though $f'(0)$ exists, f' is not continuous at 0.

Things are not really as bad as they seem, because, although the derivative may not be continuous, it does have the intermediate value property.

Theorem 4.1.21 (Intermediate Value Theorem for Derivatives) Let $f : [a, b] \to \mathbb{R}$ be continuous and let f be differentiable on (a, b). Suppose that $[c, d] \subseteq (a, b)$, $f'(c) < 0$, and $f'(d) > 0$. Then there exists a point $x \in (c, d)$ such that $f'(x) = 0$.

Proof. Since f is continuous on $[c, d]$, it assumes both a maximum and minimum value. Since $f'(c) < 0$, there exists a point $x \in (c, d)$ such that $f(x) < f(c)$ and, since $f'(d) > 0$, there exists a point $y \in (c, d)$ such that $f(y) < f(d)$. Hence the minimum does not occur at either c or d. The conclusion follows.

Exercise 4.1.22 There is a point in the previous proof which requires attention. It is related to Corollary 4.1.17 and the exercise which follows it. In the theorem above, it is assumed that $f'(c) < 0$. One might be inclined to think that this means that f is decreasing in a neighborhood of c. To show that this is not true, consider the function

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

Show that f has a positive derivative at $x = 0$ but is not increasing in any neighborhood of $x = 0$.

There is no reason to stop at one derivative. Once we get started, we can continue taking derivatives as long as the function allows us. Most of the functions encountered in elementary calculus such as polynomials, rational functions, trigonometric functions, exponential functions, logarithmic functions, hyperbolic functions, and algebraic functions, are differentiable infinitely often as long as nothing untoward happens in the domain. The above functions make up more or less the entire list of functions considered in elementary calculus. We assume that the reader knows how to differentiate them. The following definition is useful throughout analysis.

Definition 4.1.23 Let f be a continuous function from (a,b) to $\mathbb{R}$. If k is an integer greater than or equal to 1, we say that $f \in C^k(a,b)$ if f has k derivatives at each point in (a,b) and each of these derivatives is continuous on (a,b). We informally refer to such a function f as being C^k on (a,b). We denote the k-th derivative of f by $f^{(k)}$. By convention, we take $f^{(0)} = f$. We say that $f \in C^\infty(a,b)$ if f has derivatives of all orders on (a,b), and we refer to such a function as being C^∞ on (a,b). If U is any open set in $\mathbb{R}$, the expressions $C^k(U)$ and $C^\infty(U)$ are defined similarly.

Exercise 4.1.24 Suppose $f : [a,b] \to \mathbb{R}$ is C^1 on (a,b). Let $[c,d] \subseteq (a,b)$. Then there exists a constant M such that for all $x, y \in [c,d]$, we have $|f(y) - f(x)| \leq M|y - x|$.

Exercise 4.1.25 For $n \in \mathbb{N}$, find the maximum value of k for which the function $f(x) = |x|^n$ is in $C^k(\mathbb{R})$?

Exercise 4.1.26 Let

$$f(x) = \begin{cases} e^{-1/x^2} & \text{when } x \neq 0 \\ 0 & \text{when } x = 0. \end{cases}$$

i. Show that $f \in C^\infty(\mathbb{R})$.

ii. Using L'Hôpital's rule, or anything you wish, show that $f^{(k)}(0) = 0$ for all $k \geq 0$.

The higher derivatives of a function f sometimes allow us to approximate f with polynomials rather than mere linear functions.

Corollary 4.1.27 (Taylor's Theorem) Suppose $f \in C^{k+1}(a,b)$ and $x_0 \in (a,b)$. Then, for any $x \in (a,b)$, we can write

$$\begin{aligned} f(x) \quad &= \quad f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \\ &\quad + \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \frac{f^{(k+1)}(c)}{(k+1)!}(x - x_0)^{k+1} \end{aligned}$$

where c is some point between x and x_0.

Proof. Without loss of generality, we assume that $x > x_0$. For $t \in [x_0, x]$, define a polynomial $T_k(x,t)$ in the variable x by

$$T_k(x,t) = f(t) + f'(t)(x - t) + \frac{f''(t)}{2!}(x - t)^2 + \cdots + \frac{f^{(k)}(t)}{k!}(x - t)^k.$$

Define $R_k(x,t) = f(x) - T_k(x,t)$. Note that $R_k(x, x_0) \in C^{k+1}(a,b)$, $R_k^{(i)}(x_0, x_0) = 0$ for $0 \leq i \leq k$, and $R_k^{(k+1)}(x, x_0) = f^{(k+1)}(x)$ for all $x \in (a,b)$. For $t \in [x_0, x]$, $x \in (a,b), x > x_0$, set

$$Q(t) = (x - t)^{k+1} \frac{R_k(x, x_0)}{(x - x_0)^{k+1}} - R_k(x,t).$$

It is clear that Q is continuous on $[x_0, x]$, differentiable on (x_0, x), and $Q(x_0) = Q(x) = f(x)$. It is easy to see that

$$Q'(t) = -(k+1)(x - t)^k \frac{R_k(x, x_0)}{(x - x_0)^{k+1}} + \frac{(x - t)^k}{k!} f^{(k+1)}(t).$$

Hence, by Rolle's Theorem, there exists $c \in (x_0, x)$ such that

$$(k+1)(x - c)^k \frac{R_k(x, x_0)}{(x - x_0)^{k+1}} = \frac{(x - c)^k}{k!} f^{(k+1)}(c).$$

What Taylor's Theorem allows us to do is to approximate a C^{k+1} function in the neighborhood of a point by a polynomial of degree k. Usually, the *remainder term*,

$$R_k(x, x_0) = \frac{f^{(k+1)}(c)}{(k+1)!}(x - x_0)^{k+1},$$

is reasonably small because of the $(k+1)!$ in the denominator. The expansion in Taylor's Theorem is called the *Taylor expansion of f about x_0*.

Exercise 4.1.28 Find the Taylor expansions of the following functions about the indicated points to at least 6 terms.

i. $f(x) = \sin(x)$ about $x_0 = \pi$;

ii. $f(x) = \frac{1}{x-1}$ about $x_0 = -1$;

iii. $f(x) = e^{-1/x^2}$ about $x_0 = 0$;

iv. $f(x) = \sqrt{x^2+1}$ about $x_0 = 2$.

Exercise 4.1.29

i. Suppose that $f \in C^1(a, b)$ and $f'(x_0) = 0$ for some $x_0 \in (a, b)$. Show that f may or may not have a local extremum at x_0.

ii. Suppose that $f \in C^2(a, b)$, $f'(x_0) = 0$, and $f''(x_0) > 0$. Show that f has a local minimum at x_0. Formulate and prove an analogous statement in case $f''(x_0) < 0$.

iii. Suppose that $f \in C^{k+1}(a, b)$, $f^{(i)}(x_0) = 0$ for $1 \leq i \leq k$, and $f^{(k+1)}(x_0) \neq 0$. Under what conditions can you say that f must have a local extremum at x_0?

As another application of the Mean Value Theorem, we present the theorem of Liouville. This theorem can sometimes be used to determine whether a given irrational real number is algebraic or transcendental.

Theorem 4.1.30 (Liouville's Theorem) Let $\alpha \in \mathbb{R}$ be algebraic of degree $n \geq 2$. Then there exists $C = C(\alpha)$ depending on α such that $|\alpha - p/q| > C/q^n$ for all $p/q \in \mathbb{Q}$.

Proof. Let $f(x) \in \mathbb{Z}[x]$ be an irreducible polynomial of degree n such that $f(\alpha) = 0$. Let $p/q \in \mathbb{Q}$ and assume that $\alpha < p/q$. By the Mean Value Theorem, there exists c with $\alpha < c < p/q$ such that $f(p/q) - f(\alpha) = f(p/q) = (p/q - \alpha)f'(c)$. By the irreducibility of $f(x)$, we have $f(p/q) \neq 0$, and therefore $f'(c) \neq 0$. Assume that $0 < p/q - \alpha < 1$. Choose $d > 0$ such that $|f'(x)| < 1/d$ for $x \in [\alpha, p/q]$. Then $|f(p/q)| = |p/q - \alpha|\,|f'(c)| \neq 0$. Since f has integer coefficients, $|q^n f(p/q)| \in \mathbb{Z}$, so it is greater than or equal to 1. Hence, $1 \leq q^n|p/q - \alpha|(1/d)$. It follows that $|\alpha - p/q| \geq d/q^n$ when $0 < p/q - \alpha < 1$.

Exercise 4.1.31

i. Modify the above proof to cover the case when $p/q < \alpha$.

ii. Show that there exists a constant C such that $|\alpha - p/q| > C/q^n$ for all $p/q \in \mathbb{Q}$.

Exercise 4.1.32 Suppose $n \geq 2$ and α is real algebraic of degree n. Show that, if $r > n$, then the function f_r of Exercise 4.1.7 is differentiable at α.

In elementary calculus, the derivative is often motivated through a discussion of the tangent line to a curve at a point. This is accomplished using secant lines which approximate the tangent line and then taking limits. The actual definition of the tangent line to the graph of a function at a point is: If the function is differentiable at the point, the tangent line at that point is the line through the point whose slope is the derivative of the function at that point. We can think of the tangent line as the *best linear approximation* at that point. This is the idea that motivates the concept of derivative in several variables.

4.2 Differential Calculus in $\mathbb{R}^n$

We now turn to a study of the properties of differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. Throughout this chapter, we use the Euclidean norm in $\mathbb{R}^n$.

Our definition of the derivative will be stated in terms of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$, and we will need the following ideas.

Proposition 4.2.1 Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Then there exists a constant $C > 0$ such that for every $x \in \mathbb{R}^n$, $\|Tx\| \leq C\|x\|$.

Proof. Consider $S^{n-1} = \{x \in R^n \mid \|x\| = 1\}$. This is a compact set by Heine-Borel, and thus the function $S^{n-1} \to \mathbb{R}$, $x \mapsto \|Tx\|$, is continuous and hence attains a maximum C on S^{n-1} by Exercise **??**. It follows that for any nonzero $x \in \mathbb{R}^n$, $\frac{\|Tx\|}{\|x\|} = \|T(\frac{x}{\|x\|})\| \leq C$.

Definition 4.2.2 We define the *operator norm* of a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ to be the constant C identified in the proposition above, and we denote this by $\|T\| = C$.

Exercise 4.2.3 If $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear, and $\|T\| = 0$, show that $T = 0$.

Remark 4.2.4 We use the same symbol $\|\cdot\|$ to indicate the operator norm for a linear transformation as for the Euclidean norm in $\mathbb{R}^n$, but the interpretation will always be clear from the context.

The definition of a differentiable function is motivated by the discussion at the conclusion of the previous section about the derivative of a function $f : \mathbb{R} \to \mathbb{R}$. As in the case of the derivative of functions of one variable, we form a difference quotient, which means that we must divide by something. Since division in $\mathbb{R}^n$ does not make sense for $n > 1$, we need to keep the divisor in $\mathbb{R}$.

Definition 4.2.5 Suppose $U \subseteq \mathbb{R}^n$ is an open set. A function $f : U \to \mathbb{R}^m$ is *differentiable at* $x \in U$ if there is a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ such that

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

First, note that the $h \to 0$ in $\mathbb{R}^n$. Notice also that the norm sign in the numerator denotes the Euclidean norm in $\mathbb{R}^m$ while the norm sign in the denominator denotes the Euclidean norm in $\mathbb{R}^n$. If we use the norm sign for an element of $\mathbb{R}$, it indicates the usual absolute value on $\mathbb{R}$. We write $T = Df(x)$ and we call this the *derivative* of f at x. We say that f is *differentiable on* U if f is differentiable at each point in U.

Thus the derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ at a point is a linear transformation. It may be difficult to think of this as a generalization of the slope of a tangent line as it is in one variable. However, if one thinks of the tangent line in one variable as the best linear approximation to a function at a point, we can think of the derivative in $\mathbb{R}^n$ as a generalization of this concept; that is,

$$f(x_0) + Df(x_0)(x - x_0)$$

provides the "best linear approximation" to the function f at the point $x_0 \in \mathbb{R}^n$.

There are many theorems about derivatives of functions of several variables which are analogous to those in one variable.

Theorem 4.2.6 Suppose U is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ is differentiable at a point $x_0 \in U$. Then f is continuous at x_0.

Proof. Take $\varepsilon = 1$. Then there exists a $\delta > 0$ such that

$$\|f(x_0 + h) - f(x_0) - Df(x_0)h\| < \varepsilon\|h\| = \|h\|$$

whenever $\|h\| < \delta$. It follows from the triangle inequality that

$$\|f(x_0 + h) - f(x_0)\| < \|h\| + \|Df(x_0)h\| \leq \|h\| + \|Df(x_0)\| \cdot \|h\| = (1 + \|Df(x_0)\|)\|h\|$$

when $\|h\| < \delta$.

When $m = 1$ and f is a real-valued function, this leads to a special situation.

Definition 4.2.7 Let $U \subseteq \mathbb{R}^n$ be an open set, and let $f : U \to \mathbb{R}$ be differentiable on U. Let $x \in U$ and let $v \in \mathbb{R}^n$ be a unit vector. The *directional derivative of f at x in the direction v* is defined as

$$D_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.$$

In the particular case when $v = e_j$, a standard basis vector, we obtain the *partial derivative in the jth direction*

$$D_j f(x) = D_{e_j} f(x) = \lim_{t \to 0} \frac{f(x + te_j) - f(x)}{t}.$$

Exercise 4.2.8

i. Let $U \subseteq \mathbb{R}^n$ be an open set and let $f : U \to \mathbb{R}^m$ be differentiable. Write $f = (f_1, f_2, \ldots, f_m)$, where $f_k : U \to \mathbb{R}$ is the k^{th} coordinate function of f. Let $v \in \mathbb{R}^n$ be a unit vector. If $x \in U$, define

$$D_v f(x) = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.$$

Show that $D_v f(x)$ exists if and only if $D_v f_k(x)$ exists for all k, $1 \leq k \leq m$, and, in this case, $D_v f(x) = (D_v f_1(x), D_v f_2(x), \ldots, D_v f_m(x))$.

ii. Explain why it is useful for us to have required v to be a unit vector.

Remark 4.2.9 Note that, in Definition 4.2.7, these directional derivatives are real honest-to-goodness derivatives of functions of one variable, that is, they represent the rate of change of a function in a particular direction. The partial derivatives play a special role, as we shall see below. It is worth observing that the classical notation for $D_j f(x)$ is $\frac{\partial f}{\partial x_j}(x)$. All sorts of theorems and properties can be stated much more easily with the notation $D_j f(x)$.

Exercise 4.2.10

i. Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x_1, x_2, \ldots, x_n) = x_k$ for some k, for $1 \leq k \leq n$. Show that f is differentiable at any point and that $Df(x) = f$ for all $x \in \mathbb{R}^n$.

ii. Find $D_v f(x)$ for any unit vector v in $\mathbb{R}^n$.

Remark 4.2.11 The map f in the above exercise is called the *projection onto the k-th coordinate* and is denoted p_k. More generally, if $m \leq n$, we can pick indices $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and define a projection $p : \mathbb{R}^n \to \mathbb{R}^m$ by $p(x_1, x_2, \ldots, x_n) = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$.

Exercise 4.2.12 Show that any such p as above is differentiable and find its derivative.

All of the statements below are easy exercises, but we prove one or two just to show that we are working at this.

Proposition 4.2.13 If U is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ is differentiable, then the derivative of f is unique.

Proof. Suppose T and T' are derivatives of f at $x \in U$. Then, for $h \in \mathbb{R}^n$,

$$\begin{aligned}\frac{\|Th - T'h\|}{\|h\|} &= \frac{\|(f(x+h) - f(x) - Th) - (f(x+h) - f(x) - T'h)\|}{\|h\|} \\ &\leq \frac{\|f(x+h) - f(x) - Th\|}{\|h\|} + \frac{\|f(x+h) - f(x) - T'h\|}{\|h\|}.\end{aligned}$$

Thus $\|Th - T'h\|/\|h\| \to 0$ as $h \to 0$.

But $T - T'$ is a linear map, so by the definition of the norm of an operator,

$$\max_{\|h\|=M} \frac{\|Th - T'h\|}{\|h\|} = \|T - T'\|.$$

Because this is a constant independent of M for $M \neq 0$ we must have $\|T - T'\| = 0$, that is, $T = T'$.

Exercise 4.2.14 If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map and $x \in \mathbb{R}^n$, then f is differentiable at x and $Df(x) = f$.

This exercise says something which is almost tautological, but it says something nonetheless. That is, if $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear and $h \in \mathbb{R}^n$, then at any point $x \in \mathbb{R}^n$, $Df(x)h = f(h)$. How does this all work for a linear function $f : \mathbb{R} \to \mathbb{R}$? Notice first that if $f : \mathbb{R} \to \mathbb{R}$ is a linear map, then $f(0)$ must be equal to 0. Moreover, there is an element $a \in \mathbb{R}$ such that $f(x) = ax$, for all $x \in \mathbb{R}$. Conversely, given an element $a \in \mathbb{R}$ we can construct a linear map $f_a : \mathbb{R} \to \mathbb{R}$ defined by $f_a(x) = ax$. In elementary calculus, we use this correspondence to treat derivatives of functions of one variable as numbers at each point, or as functions on $\mathbb{R}$, rather than as linear maps $\mathbb{R} \to \mathbb{R}$ at each point. In our present case, instead of saying that if $f_a(x) = ax$, then $f'_a(x) = a$, we have $Df_a(x) = f_a$ for all x. Observe that this discussion tells us that if a function is already linear, then the best linear approximation is the function itself.

Proposition 4.2.15 Let U be an open set in $\mathbb{R}^n$ and $f, g : U \to \mathbb{R}^m$ be differentiable on U. Then $f + g$ is differentiable on U, and $D(f + g)(x) = Df(x) + Dg(x)$ for $x \in U$.

Proof. For $x \in U$,

$$\begin{aligned}&\frac{\|(f+g)(x+h) - (f+g)(x) - (Df(x) + Dg(x))h\|}{\|h\|} \\ &\leq \frac{\|f(x+h) - f(x) - Df(x)h\|}{\|h\|} + \frac{\|g(x+h) - g(x) - Dg(x)h\|}{\|h\|}.\end{aligned}$$

Both expressions on the right go to zero as h goes to zero.

The reader might ask about the product rule. The question depends on the type of multiplication being used. Let us try the easiest case in which f and g map U to $\mathbb{R}$.

Proposition 4.2.16 If U is an open set in $\mathbb{R}^n$, and $f : U \to \mathbb{R}$ and $g : U \to \mathbb{R}$ are real-valued functions which are differentiable on U, then fg is differentiable on U. For $x \in U$, we have $D(fg)(x) = f(x)Dg(x) + g(x)Df(x)$.

Proof. Before starting the proof, we observe that $f(x)Dg(x)$ really makes sense, since $f(x)$ is a real scalar and $Dg(x)$ is a linear map, and we can always multiply linear maps by scalars. So let's go:

$$\begin{aligned}
&\frac{1}{\|h\|}\|f(x+h)g(x+h)-f(x)g(x)-(f(x)Dg(x)+g(x)Df(x))h\| \\
&\quad= \quad \frac{1}{\|h\|}\|(f(x+h)[g(x+h)-g(x)]-f(x)Dg(x)h)+(g(x)[f(x+h)-f(x)]-g(x)Df(x)h)\| \\
&\quad\leq \quad \frac{1}{\|h\|}\big(\|(f(x+h)-f(x))(g(x+h)-g(x))\|+\|f(x)(g(x+h)-g(x)-Dg(x)h)\| \\
&\qquad\qquad + \|g(x)(f(x+h)-f(x)-Df(x)h)\|\big).
\end{aligned}$$

By the definition of derivative, the second and third terms vanish as $h \to 0$. For the first term, we have $\frac{1}{\|h\|}\|(f(x+h)-f(x))(g(x+h)-g(x))\| = \frac{\|f(x+h)-f(x)\|}{\|h\|}\|g(x+h)-g(x)\|$.

Exercise 4.2.17 Finish the proof by showing $\frac{\|f(x+h)-f(x)\|}{\|h\|}\|g(x+h)-g(x)\|$ goes to zero as $h \to 0$.

We now turn to the Chain Rule which is a very important theorem in the calculus of several variables. The reader will note that the proof is pretty much the same as the proof in one variable (Theorem 4.1.12).

Theorem 4.2.18 (Chain Rule) Suppose that U is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ is differentiable on U. Let V be an open set in $\mathbb{R}^m$ such that $f(U) \subseteq V$. Suppose $g : V \to \mathbb{R}^p$ is differentiable on V. Then $g \circ f : U \to \mathbb{R}^p$ is differentiable on U, and, for any $a \in U$ we have $D(g \circ f)(a) = Dg(f(a)) \circ Df(a)$.

Proof. Let $b = f(a)$, $L = Df(a)$, and $M = Dg(b)$. Set

$$\phi(x) = f(x) - f(a) - L(x-a),$$

$$\psi(y) = g(y) - g(b) - M(y-b),$$

and

$$\rho(x) = (g \circ f)(x) - (g \circ f)(a) - ML(x-a).$$

Then, our hypotheses tell us that

$$\lim_{x \to a} \frac{\|\phi(x)\|}{\|x-a\|} = \lim_{y \to b} \frac{\|\psi(y)\|}{\|y-b\|} = 0,$$

and we need to show that

$$\lim_{x \to a} \frac{\|\rho(x)\|}{\|x-a\|} = 0.$$

But

$$\rho(x) = \psi(f(x)) + M\phi(x)$$

(write out the definitions). Then

$$\frac{\|M\phi(x)\|}{\|x-a\|} \leq \|M\| \frac{\|\phi(x)\|}{\|x-a\|} \to 0$$

as $x \to a$. By the above limit and the continuity of f, for any $\varepsilon > 0$, there is $\delta > 0$ such that $\|x-a\| < \delta$ ensures $\|\psi(f(x))\| < \varepsilon\|f(x)-b\|$. But $\|f(x)-b\| \leq \|\phi(x)\| + \|L\|\|x-a\|$, so

$$\frac{\|\psi(f(x))\|}{\|x-a\|} \to 0$$

as $x \to a$ as well.

4.3 The Derivative as a Matrix of Partial Derivatives

We now proceed to some of the more computational aspects of differentiation theory in several variables. Let us consider the form of a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. If $f : \mathbb{R}^n \to \mathbb{R}^m$, we can write $f(x) = (f_1(x), f_2(x), \ldots, f_m(x))$, where for each k with $1 \leq k \leq m$, $f_k : \mathbb{R}^n \to \mathbb{R}$. These functions f_k are the *component functions* of f, and there is an obvious fact about them.

Corollary 4.3.1 (to the Chain Rule) If f is differentiable, then f_k is differentiable for each k.

Proof. Observe that $f_k = p_k \circ f$ and use the chain rule.

Exercise 4.3.2 Show that if each f_k is differentiable on an open set $U \subseteq \mathbb{R}^n$, then $f = (f_1, f_2, \ldots, f_m)$ is differentiable on U and $Df = (Df_1, Df_2, \ldots, Df_m)$.

How does one interpret $(Df_1, Df_2, \ldots, Df_m)$ as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$? For $x \in U$ the expression

$$(Df_1(x), Df_2(x), \ldots, Df_m(x))$$

is the linear transformation whose value at $y \in \mathbb{R}^n$ is given by

$$[(Df_1(x), Df_2(x), \ldots, Df_m(x))](y) = (Df_1(x)(y), Df_2(x)(y), \ldots, Df_m(x)(y)).$$

What do partial derivatives have to do with all this? We know that linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ can be represented by matrices with respect to any bases we choose in the respective spaces. Let us work first with a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We know that

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

If we think of $Df(x)$ as a $1 \times n$ matrix of the form $(a_1(x), a_2(x), \ldots, a_n(x))$ and write $h = (h_1, h_2, \ldots, h_n)$, then

$$Df(x)h = (a_1(x), a_2(x), \ldots, a_n(x)) \cdot (h_1, h_2, \ldots, h_n),$$

where the dot signifies the dot product in $\mathbb{R}^n$. Now consider the case $h = te_j$. It follows from Definition 4.2.7 that $a_j(x) = D_j f(x)$. This is summarized in the following statement.

Proposition 4.3.3 If $U \subseteq \mathbb{R}^n$ is open and $f : U \to \mathbb{R}$ is differentiable on U, then

$$Df(x) = (D_1 f(x), D_2 f(x), \ldots, D_n f(x)).$$

For $h \in \mathbb{R}^n$,

$$Df(x)h = D_1 f(x)h_1 + D_2 f(x)h_2 + \cdots + D_n f(x)h_n.$$

Exercise 4.3.4 Let $U \subseteq \mathbb{R}^n$ be open, and let $f = (f_1, f_2, \ldots, f_m) : U \to \mathbb{R}^m$ be differentiable on U. If $x \in U$ and (a_{ij}) is the matrix of $Df(x)$ with respect to the standard basis, then $a_{ij} = D_j f_i(x)$.

Definition 4.3.5 Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}^m$ be differentiable on U, and let $x \in U$. The matrix $(D_j f_i(x))_{1 \leq i \leq m, 1 \leq j \leq n}$ is called the *Jacobian matrix* of f at x.

In the case that $m = 1$, the vector-valued function $(D_1 f(x), D_2 f(x), \ldots, D_n f(x))$ is called the *gradient vector* and is denoted $\nabla f(x)$. Thus, by Proposition 4.3.3, $Df(x)(h) = \nabla f(x) \cdot h$.

Exercise 4.3.6

i. Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable on U. If $x \in U$ and v is a unit vector in $\mathbb{R}^n$, then $D_v f(x) = \nabla f(x) \cdot v$.

ii. Show that the maximum value of the directional derivative of f at a point x is in the direction of $\nabla f(x)$, and the value of this directional derivative is $\|\nabla f(x)\|$.

The existence of the partial derivatives at a point, or even in the neighborhood of a point, does not assure the differentiability of the function at the point. Actually, it is much worse than that. As shown in the following example, the existence of the directional derivative at a point in every direction does not assure that the function is differentiable at the point. For differentiability, one must approach the given point from every direction along all sorts of paths, while the directional derivative is taken along straight-line paths through the point.

Example 4.3.7 Let

$$f(x,y) = \begin{cases} 0 & \text{when } (x,y) = (0,0), \\ \frac{xy}{x^2+y^2} & \text{otherwise.} \end{cases}$$

Then $D_1 f(0,0) = D_2 f(0,0) = 0$, but nonetheless f is not continuous at the origin, and hence not differentiable at $(0,0)$.

Exercise 4.3.8

i. Let

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

Show that f has a directional derivative in every direction at the origin, but that f is not continuous at the origin, and hence not differentiable.

ii. Let

$$f(x,y) = \begin{cases} 0 & \text{when } (x,y) = (0,0), \\ (x^2+y^2)\sin\left(\frac{1}{\sqrt{x^2+y^2}}\right) & \text{otherwise.} \end{cases}$$

Show that $D_1 f$ and $D_2 f$ exist at the origin but are not continuous.

The previous examples and exercises illustrate a few pathologies. Nevertheless, if the partial derivatives all exist and are continuous in a neighborhood of a point, then the function is differentiable at that point.

Theorem 4.3.9 *Let U be an open set in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}^m$ be a function with the property that $D_j f_i$ is continuous on U for $1 \le i \le m$, $1 \le j \le n$. Then f is differentiable on U and, as we might expect, $Df(x) = (D_j f_i(x))_{1\le i\le m, 1\le j\le n}$.*

Proof. We prove it first for $n = 2$ and $m = 1$. The proof uses the Mean Value Theorem for functions of one variable. For $x = (x_1, x_2) \in U$ and $h = (h_1, h_2)$, write

$$\begin{aligned} &\frac{\|f(x+h) - f(x) - (Df(x))h\|}{\|h\|} \\ &= \frac{\|(f(x_1+h_1, x_2+h_2) - f(x_1,x_2) - (D_1 f(x_1,x_2), D_2 f(x_1,x_2))(h_1,h_2)\|}{\|h\|} \\ &= \frac{\|f(x_1+h_1, x_2+h_2) - f(x_1, x_2+h_2) - D_1 f(x_1,x_2)h_1 + f(x_1, x_2+h_2) - f(x_1,x_2) - D_2 f(x_1,x_2)h_2\|}{\|h\|} \\ &\le \frac{\|f(x_1+h_1, x_2+h_2) - f(x_1, x_2+h_2) - D_1 f(x_1,x_2)h_1\|}{\|h\|} \\ &\quad + \frac{\|f(x_1, x_2+h_2) - f(x_1,x_2) - D_2 f(x_1,x_2)h_2\|}{\|h\|} \end{aligned}$$

Now, by the Mean Value Theorem for functions of one variable, there exists $\xi_1 \in (x_1, x_1 + h_1)$ (or $\xi_1 \in (x_1 + h_1, x_1)$ if $h_1 < 0$) and $\xi_2 \in (x_2, x_2 + h_2)$ (or $\xi_2 \in (x_2 + h_2, x_2)$ if $h_2 < 0$) such that

$$f(x_1 + h_1, x_2 + h_2) - f(x_1, x_2 + h_2) = D_1 f(\xi_1, x_2 + h_2) h_1$$

and

$$f(x_1, x_2 + h_2) - f(x_1, x_2) = D_2 f(x_1, \xi_2) h_2.$$

Thus the above sequence of inequalities continues as

$$\begin{aligned} & \frac{|h_1|}{\|h\|}\|D_1 f(\xi_1, x_2 + h_2) - D_1 f(x_1, x_2)\| + \frac{|h_2|}{\|h\|}\|D_2 f(x_1, \xi_2) - D_2 f(x_1, x_2)\| \\ \leq \ & \|D_1 f(\xi_1, x_2 + h_2) - D_1 f(x_1, x_2)\| + \|D_2 f(x_1, \xi_2) - D_2 f(x_1, x_2)\|, \end{aligned}$$

and this goes to zero as $h \to 0$ since $D_1 f$ and $D_2 f$ are continuous at (x_1, x_2).

The general case for arbitrary n and m is easy to complete by adding and subtracting enough times and using the Mean Value Theorem over and over again.

So far, we have not computed a lot of derivatives because it is awkward to compute the linear transformation associated to the definition. With this last theorem, it becomes much easier to compute the derivative of a function given by explicit formulas. For example, let $f : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by

$$f(x, y, z) = (x^2 y + e^{xz}, \sin(xyz)).$$

Then, f is differentiable on all of $\mathbb{R}^3$ and

$$Df(x, y, z) = \begin{pmatrix} 2xy + ze^{xz} & x^2 & xe^{xz} \\ yz\cos(xyz) & xz\cos(xyz) & xy\cos(xyz) \end{pmatrix}.$$

Thus,

$$Df(1, 0, -1) = \begin{pmatrix} -e^{-1} & 1 & e^{-1} \\ 0 & -1 & 0 \end{pmatrix},$$

which represents a linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^2$. The more general expression above assigns to each point $(x, y, z) \in \mathbb{R}^3$ a linear transformation $Df(x, y, z) \in \mathcal{L}(\mathbb{R}^3, \mathbb{R}^2)$. Thus, we can think of Df as a function

$$Df : \mathbb{R}^3 \to \mathcal{L}(\mathbb{R}^3, \mathbb{R}^2).$$

This last example should totally demystify the idea of derivatives of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. There is of course some interesting theory related to differentiation, and we are in the midst of that exposition. However, to compute partial derivatives, and hence derivatives, requires nothing more than being able to differentiate functions of one variable, which you learned in elementary calculus. In fact, as we have commented before, there really are not that many functions that you can differentiate explicitly. So you will notice that all of the examples involve polynomials, rational functions, trigonometric functions, logarithmic functions, etc.

Exercise 4.3.10 For any function $f : \mathbb{C} \to \mathbb{C}$, we can write f in terms of its real and imaginary parts as $f(z) = u(x, y) + iv(x, y)$, where u and v are functions from $\mathbb{R}^2$ to $\mathbb{R}$, and z is written in the form $z = x + iy$.

A function $f : \mathbb{C} \to \mathbb{C}$ (with the usual metric on $\mathbb{C}$) is *complex differentiable* at $z_0 \in \mathbb{C}$ if

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. A function f is *analytic* on an open set $U \subseteq \mathbb{C}$ if f is differentiable at each point of U.

i. Suppose f is analytic on an open set $U \subseteq \mathbb{C}$. Show that u and v are differentiable on U considered as a subset of $\mathbb{R}^2$.

ii. Suppose f is analytic on an open set $U \subseteq \mathbb{C}$. Show that $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$, and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$. These are called the *Cauchy-Riemann equations.*

iii. If $U \subseteq \mathbb{C}$ is an open set, and u and v are continuously differentiable on U and satisfy the Cauchy-Riemann equations, show that $f(z) = u(x, y) + iv(x, y)$ is analytic on U.

iv. Find an example of a function $f : \mathbb{C} \to \mathbb{C}$ that is differentiable at one point but not in a neighborhood of that point.

Exercise 4.3.11 Let $f : \mathbb{C} \to \mathbb{C}$ be given by $f(z) = e^z$, which can be written $f(x+iy) = e^x \cos y + ie^x \sin y$ (see Definition 1.7.12). Show that f is analytic on $\mathbb{C}$, and that $f'(z) = f(z)$.

Exercise 4.3.12 Let $z_0 \in \mathbb{C}$, and define $f : \mathbb{C} \to \mathbb{C}$ by $f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$, where $a_n \in \mathbb{C}$ for all n. Let r be the radius of convergence of this power series (see Section 1.9) and suppose that $r > 0$.

i. Show that $f(z)$ is analytic on $B_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| < r\}$. (Hint: Show that the series can be differentiated term-by-term inside the radius of convergence.)

ii. Show that the radius of convergence of the power series for $f'(z)$ is equal to r.

Exercise 4.3.13 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = \sqrt{|x| + |y|}$. Find those points in $\mathbb{R}^2$ at which f is differentiable.

Exercise 4.3.14 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that $|f(x)| \leq \|x\|^\alpha$ for some $\alpha > 1$. Show that f is differentiable at 0.

Exercise 4.3.15 Let $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x, y) = x \cdot y$.

i. Show that f is differentiable on $\mathbb{R}^n \times \mathbb{R}^n$.

ii. Show that $Df(a, b)(x, y) = a \cdot y + b \cdot x$.

4.4 The Mean Value Theorem

Next we consider the Mean Value Theorem for functions of several variables. As in the case of functions of one variable, the Mean Value Theorem relates the average rate of change of a function in a specified direction to the instantaneous rate of change at a particular point as measured by the derivative. In the case of a function of one variable, there is little choice about the so-called specified direction. But, when we consider functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, we find that it is necessary to specify a direction both in the domain and in the range of the function in order to give a proper interpretation to the Mean Value Theorem.

Theorem 4.4.1 Let U be an open set in $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be differentiable on U. Let x, y be two distinct points in U such that the line segment joining x to y lies entirely in U. Then, there exists $\xi \in (0, 1)$ such that

$$f(y) - f(x) = Df(z)(y - x),$$

where $z = (1 - \xi)x + \xi y$.

Proof. Define

$$F(t) = f((1 - t)x + ty).$$

Then F is continuous on $[0, 1]$ and differentiable on $(0, 1)$, so there exists $\xi \in (0, 1)$ such that

$$F(1) - F(0) = F'(\xi).$$

The left-hand side is $f(y) - f(x)$, and by the chain rule,

$$F'(\xi) = Df((1-\xi)x + \xi y)(y - x).$$

Note that $Df(z)(y-x) = D_u f(z)\|y-x\|$, where u is the unit vector in the direction of the vector $y - x$.

Exercise 4.4.2 Let U be a convex open set in $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be differentiable on U. Show that if $Df(x) = 0$ for all $x \in U$, then f is constant on U.

We note again that the Mean Value Theorem for real-valued functions of several variables that we have just proved is really a one-variable theorem. That is, to make sense of the mean value property, it was essential that we move away from a point x in exactly one direction, namely, the straight-line direction from x to y. It is this idea that motivates the statement of a Mean Value Theorem for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. To retain the one-variable nature of the Mean Value Theorem, in addition to having a straight-line direction implicitly chosen for us in the domain (namely, the direction $y - x$ in $\mathbb{R}^n$), we must also explicitly choose a direction in $\mathbb{R}^m$ in order to make sense of the mean value property.

Theorem 4.4.3 Let U be an open subset in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}^m$ be differentiable on U. For any two distinct points $x, y \in U$ such that the line segment joining x to y lies entirely in U, and any vector $v \in \mathbb{R}^m$, there exists $\xi \in (0,1)$ such that

$$v \cdot (f(y) - f(x)) = v \cdot (Df(z)(y - x)),$$

where $z = (1 - \xi)x + \xi y$.

Exercise 4.4.4 Prove this.

We note in connection to the discussion above that if v is a unit vector, the expression $v \cdot (f(y) - f(x))$ is the component of $f(y) - f(x)$ in the direction of the vector v. A similar statement is true for the expression $v \cdot (Df(z)(y - x))$.

Exercise 4.4.5 Let U be a convex open set in $\mathbb{R}^n$, and let $f : U \to \mathbb{R}^m$ be differentiable on U. Show that if $Df(x) = 0$ for all $x \in U$, then f is constant on U.

The significance of the multivariable version of the Mean Value Theorem is that the direction vector v is arbitrary. However, in general, there is no single ξ that will satisfy the conclusion of the Mean Value Theorem for all v simultaneously.

Exercise 4.4.6 Show that the function $f : \mathbb{R} \to \mathbb{R}^2$ given by $f(x) = (\cos x, \sin x)$ does not satisfy the property that, given any $x, y \in \mathbb{R}$ with $x < y$, there exists $z \in (x, y)$ such that $f(y) - f(x) = Df(z)(y - x)$. (Hint: think about the periodic nature of f.)

Exercise 4.4.7 Let U be an open set in $\mathbb{R}^n$, and let $f : U \to \mathbb{R}^m$ be differentiable on U with continuous partial derivatives. (In the next section, we call such a function C^1.) Suppose that B is a compact, convex subset of U. Then there exists a constant M such that for any two points $x, y \in B$, we have $\|f(y) - f(x)\| \leq M\|y - x\|$. This is analogous to Exercise 4.1.24 for functions of one variable.

4.5 Higher-Order Partial Derivatives and Taylor's Theorem

The next natural question is, "What about second derivatives?" There are two ways to look at this. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at a point x, then $Df(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. That is, it is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. If we think of the symbol Df as a map from $\mathbb{R}^n$ (or a subset of $\mathbb{R}^n$) to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, then what is the derivative of the derivative? Think about that while we move to a more tractable problem.

We have seen that, for a function $f : \mathbb{R}^n \to \mathbb{R}$, the partial derivative $D_j f$ can be thought of as a function from $\mathbb{R}^n$ to $\mathbb{R}$. Consequently, it would make sense to consider the partial derivatives of the functions $D_j f$. For example, suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is defined by $f(x, y, z) = x^2 y + e^{xz}$. Then, $D_1 f(x, y, z) = 2xy + ze^{xz}$. Then, $D_3(D_1 f)(x, y, z) = e^{xz} + zxe^{xz}$. Looking at this from another perspective, we next compute $D_3 f(x, y, z)$. We get $D_3 f(x, y, z) = xe^{xz}$ and $D_1(D_3 f)(x, y, z) = e^{xz} + xze^{xz}$. So, for this f, we see that

$$D_3(D_1 f)(x, y, z) = D_1(D_3 f)(x, y, z).$$

The functions $D_3(D_1 f)$ and $D_1(D_3 f)$ are examples of *mixed partial derivatives* of f.

For a general setting, consider a function $f = (f_1, f_2, \ldots, f_m)$ from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$. Then, we define the mixed partial derivative

$$D_{ij} f_k = D_i(D_j f_k)$$

assuming that the $D_j f_k$ has partial derivatives for all j, k. The question is, "Under what conditions will we have $D_{ij} f_k = D_{ji} f_k$ for all i, j?"

Theorem 4.5.1 Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable on U. Suppose that $D_{ij} f$ and $D_{ji} f$ exist and are continuous on U. Then $D_{ij} f(x) = D_{ji} f(x)$ for all $x \in U$.

Proof. It is enough to prove the theorem for $f : \mathbb{R}^2 \to \mathbb{R}$. Let $x = (a, b) \in U$, and let h be small enough that $B_{2h}(x)$ is contained in U. We consider the second differences for computing the partial derivatives of f. Set

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

By the Mean Value Theorem, there exist ξ and ξ' between a and $a + h$ such that

$$f(a + h, b + h) - f(a, b + h) = hD_1 f(\xi, b + h)$$

and

$$f(a + h, b) - f(a, b) = hD_1 f(\xi', b).$$

In turn this gives ξ'' between ξ and ξ' and η between b and $b + h$ such that

$$D_1 f(\xi, b + h) - D_1 f(\xi', b) = hD_{21} f(\xi'', \eta) = hA(h).$$

For the next step, we rewrite

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

and proceed similarly. From this we get $A(h) = D_{12} f(\xi^*, \eta^*)$ where ξ^* and η^* are obtained similarly. If we let h go to 0, the continuity of the mixed partials now implies the result.

□

Exercise 4.5.2 Let

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

Show that f is differentiable everywhere. Show that $D_{12} f(0, 0)$ and $D_{21} f(0, 0)$ exist, but $D_{12} f(0, 0) \neq D_{21} f(0, 0)$.

In one variable calculus, higher derivatives are defined by differentiating the derivative considered as a function on some set. In the present situation, where we have a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the derivative is a map from $\mathbb{R}^n$ to the space $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$. From our discussion of linear algebra, we know that $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ can be identified with the space $\mathbb{R}^{mn}$. We can then intepret the second derivative at a point as an element of

$$\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)) \cong \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{mn}).$$

This space has dimension mn^2, and the entries of the matrix representing the derivative are the partial derivatives $D_{ij}f_k$. We could continue this process and define the ℓ^{th} derivative, but this would not be particularly useful for this text and we confine ourselves to the first and second derivatives.

On the other hand, it will be useful to talk about all orders of differentiation for the partial dervatives. For example, if $f : \mathbb{R}^5 \to \mathbb{R}$ has partial derivatives of order 3 on an open set $U \subseteq \mathbb{R}^n$, that is, $D_1^3 f(x)$, $D_2^2 D_4 f(x)$, $D_2 D_4 D_2 f(x)$, etc. all exist, we can consider the situation similar to that in Theorem 4.5.1 in which two of these are equal.

Definition 4.5.3 Let $U \subseteq \mathbb{R}^n$ be an open set and $f : U \to \mathbb{R}^m$. The function f is said to be in $C^k(U)$ (or C^k on U) if f has all continuous partial derivatives up to and including order k on U. The function f is said to be in $C^\infty(U)$ or *smooth* on U if f has all partial derivatives of all orders on U.

Exercise 4.5.4 Suppose $U \subseteq \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is C^k on U. Let $\ell \leq k$ and $i_1, i_2, \ldots, i_\ell$ be a collection of integers between 1 and n. Show that for any σ in S_ℓ, the symmetric group on ℓletters,

$$D_{i_1, i_2, \ldots, i_\ell} f = D_{\sigma(i_1), \sigma(i_2), \ldots, \sigma(i_\ell)} f.$$

If f is C^k on an open set U in $\mathbb{R}^n$, then every partial derivative of order k can be written in the form $D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_n^{\alpha_n} f$, where the α_i are nonnegative integers and $\sum_{i=1}^n \alpha_i = k$. There is an interesting combinatorial problem that arises here.

Exercise 4.5.5

i. How many ways can one partition k into an ordered sum of n nonnegative summands?

ii. How many ways can one partition k into n nonnegative summands if the order is ignored?

Exercise 4.5.6 This is an exercise in high school algebra. Consider the polynomial in n variables,

$$p(x_1, x_2, \ldots, x_n) = (x_1 + x_2 + \cdots + x_n)^k.$$

Show that, upon expanding this in monomials, the coefficient of $x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$, where $\sum_{i=1}^n k_i = k$, is given by

$$\frac{k!}{k_1! k_2! \cdots k_n!}.$$

This expression is called a *multinomial coefficient* and is denoted by $\binom{k}{k_1, k_2, \ldots, k_n}$.

Exercise 4.5.7 Recall Exercise 4.3.12. Let $z_0 \in \mathbb{C}$, and define $f : \mathbb{C} \to \mathbb{C}$ by $f(z) = \sum_{n=0}^\infty a_n (z - z_0)^n$, where $a_n \in \mathbb{C}$ for all n. Let r be the radius of convergence of this power series and supposed that $r > 0$. Show that f is infinitely differentiable on $B_r(z_0)$.

Recall that in the Mean Value Theorem for functions $f : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^n$, we stated that for points $x, y \in U$, there is an element z on the line joining x to y such that $f(y) = f(x) + Df(z)(y - x)$. For this to hold, the line joining x and y must lie in U. If we assume that f is C^{k+1}, we would like to further expand f in terms of its partial derivatives of order j where $j \leq k + 1$. How would we expect such an expression to look? First, let us write $y = x + th$, where $h = (h_1, h_2, \ldots, h_n)$ is a unit vector in $\mathbb{R}^n$ and

$t > 0$. Then, the directional derivative of f at x in the direction of the vector h is $D_h f(x) = \nabla f(x) \cdot h$. If we iterate this directional derivative r times, we get the expression

$$D_h^r f(x) = \sum_{\alpha_1 + \cdots + \alpha_n = r} \binom{r}{\alpha_1, \ldots, \alpha_n} h_1^{\alpha_1} \cdots h_n^{\alpha_n} D_1^{\alpha_1} \cdots D_n^{\alpha_n} f(x).$$

We can now write

$$f(y) = f(x) + \sum_{r=1}^{k} \frac{D_h^r f(x) t^r}{r!} + R_k(y),$$

where, as one might suspect,

$$R_k(y) = f(y) - f(x) - \sum_{r=1}^{k} \frac{D_h^r f(x) t^r}{r!}.$$

Theorem 4.5.8 (Taylor's Theorem) Let U be a convex open set in $\mathbb{R}^n$, let $f : U \to \mathbb{R}$ be a C^{k+1} function, and let $x, y \in U$ with $y \neq x$. Write $y = x + th$, where h is a unit vector in $\mathbb{R}^n$, and $t > 0$. Then there exists $s \in \mathbb{R}$, $0 < s < t$, such that

$$R_k(y) = \frac{D_h^{k+1} f(x + sh) t^{k+1}}{(k+1)!}.$$

This is, of course, the multivariable analog of Corollary 4.1.27.

Proof. Apply Corollary 4.1.27 to the function $F(a) = f(x + ah)$.

Definition 4.5.9 The polynomial

$$f(x) + \sum_{r=1}^{k} \frac{D_h^r f(x) t^r}{r!}$$

is called the *Taylor polynomial of degree k* for the function f at x in the direction h.

The Taylor polynomial may be regarded as a reasonable approximation to f in a neighborhood of the point x in the direction h because the remainder term vanishes to order $k + 1$ at x and has a $(k + 1)!$ in the denominator.

Exercise 4.5.10 Compute the Taylor polynomial of degree 3 for the following functions at the specified point (in an arbitrary direction h).

i. $f(x, y, z) = \frac{1}{xyz}$ at $(1, 1, 1)$.

ii. $f(x, y, z) = e^{xy+yz}$ at $(0, 0, 0)$.

Exercise 4.5.11 Let U be a convex open set in $\mathbb{R}^n$, let $x \in U$, and let $f : U \to \mathbb{R}$ be a C^{k+1} function. Show that the Taylor polynomial of f at x is the best polynomial approximation to f at x by proving that if P is a polynomial of degree k such that

$$\lim_{t \to 0} \left| \frac{f(x + th) - P(t)}{t^k} \right| = 0,$$

then P is the Taylor polynomial of degree k of f at x in the direction h.

4.6 Hypersurfaces and Tangent Hyperplanes in $\mathbb{R}^n$

As stated in Section 4.1, the derivative of a function of one variable $f : \mathbb{R} \to \mathbb{R}$ that is differentiable at a point $x = c$ is the slope of the tangent line to the curve $y = f(x)$ at $(c, f(c))$. The equation of the line tangent to the curve at that point can be written $y = f(c) + f'(c)(x - c)$.

This situation may be regarded as a special case of a tangent line to a general curve in $\mathbb{R}^2$. The curve $y = f(x)$ may be considered as the level curve $\{(x, y) \in \mathbb{R}^2 \mid F(x, y) = 0\}$ of the function $F : \mathbb{R}^2 \to \mathbb{R}$ given by $F(x, y) = f(x) - y$.

We now assume that $F : \mathbb{R}^2 \to \mathbb{R}$ is any C^1 function. Consider the curve $F(x, y) = k$ in $\mathbb{R}^2$. Let (x_0, y_0) be a point on this curve, and assume that the gradient $\nabla F(x_0, y_0) \neq (0, 0)$. The tangent line to the curve at the point (x_0, y_0) is the line through the point perpendicular to $\nabla F(x_0, y_0)$. The equation of this line is then

$$\nabla F(x_0, y_0) \cdot (x - x_0, y - y_0) = 0, \text{ or}$$
$$D_1F(x_0, y_0)(x - x_0) + D_2F(x_0, y_0)(y - y_0) = 0.$$

Exercise 4.6.1 Show that in the case $F(x, y) = f(x) - y$, the equation of the tangent line is the same as that given in the first paragraph above.

Exercise 4.6.2

i. Given the function $F(x, y) = x^2 + y^2$, consider the curve $F(x, y) = 3$, that is, $x^2 + y^2 = 3$. Find the equation of the tangent line at each point of this curve.

ii. Given the function
$$F(x, y) = \begin{cases} x^2 \sin^2(1/y) & \text{if } y \neq 0, \\ 0 & \text{if } y = 0, \end{cases}$$
consider the curve $F(x, y) = 1$. On what domain is F a C^1 function? At what points is the gradient vector nonzero? Find an equation for the tangent line at all points where the gradient vector is nonzero.

How does this generalize to N-dimensional space? We consider a C^1 function $F : \mathbb{R}^N \to \mathbb{R}$.

Definition 4.6.3 A *smooth hypersurface* in $\mathbb{R}^N$ is a set of points defined by the equation

$$F(x_1, x_2, \ldots, x_N) = k$$

with the property that $\nabla F \neq 0$ at each point in this set.

Examples 4.6.4

i. Let $F : \mathbb{R}^N \to \mathbb{R}$ be defined by $F(x_1, x_2, \ldots, x_N) = x_1^2 + x_2^2 + \cdots + x_N^2$. Then the smooth hypersurface defined by $F(x_1, x_2, \ldots, x_N) = 1$ is the unit sphere in $\mathbb{R}^N$.

ii. Let $F : \mathbb{R}^3 \to \mathbb{R}$ be defined by $F(x, y, z) = x^2 + y^2 - z$. Then the equation $F(x, y, z) = 0$ gives a smooth hypersurface in $\mathbb{R}^3$ that is called a *paraboloid*.

Definition 4.6.5 Let S be a smooth hypersurface in $\mathbb{R}^N$ defined by $F : \mathbb{R}^N \to \mathbb{R}$, and let $c = (c_1, c_2, \ldots, c_N) \in S$. We define the *tangent hyperplane* to S at c to be the hyperplane through the point c normal to the vector $\nabla F(c)$, that is, the hyperplane $\nabla F(c) \cdot (x - c) = 0$.

Exercise 4.6.6 Let $F : \mathbb{R}^3 \to \mathbb{R}$ be given by

$$F(x, y, z) = \begin{cases} \frac{1}{x} + \frac{1}{y} + \frac{1}{z} & \text{if } x \neq 0, y \neq 0, z \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

i. For what values of k does the equation $F(x,y,z)=k$ define a smooth hypersurface?

ii. For those values of k, find the equation of the tangent hyperplane at every point of the hypersurface.

Exercise 4.6.7 Let $F:\mathbb{R}^4\to\mathbb{R}$ be given by $F(x,y,z,w)=x^2\sin y+y^2\sin z-zwe^{xy}$.

i. For what values of k does the equation $F(x,y,z,w)=k$ define a smooth hypersurface?

ii. For those values of k, find the equation of the tangent hyperplane at every point of the hypersurface.

In some cases the discussion above can be made simpler. Suppose that the hypersurface defined by $F(x_1,x_2,\ldots,x_N)=0$ may be realized as the graph of a C^1 function $f:\mathbb{R}^n\to\mathbb{R}$, where $N=n+1$. Specifically, we suppose the surface can be defined by $x_{n+1}=f(x_1,x_2,\ldots,x_n)$, that is, the level set $\{(x_1,x_2,\ldots,x_n,x_{n+1})\in\mathbb{R}^{n+1}\mid F(x_1,x_2,\ldots,x_n,x_{n+1})=0\}$, where

$$F(x_1,x_2,\ldots,x_n,x_{n+1})=f(x_1,x_2,\ldots,x_n)-x_{n+1}.$$

A point C on this hypersurface can be written as $C=(c,f(c))$, where $c=(c_1,c_2,\ldots,c_n)$ is a point in $\mathbb{R}^n$. Using the ideas developed above, we note that the tangent hyperplane to this hypersurface at the point C is defined by $\nabla F(C)\cdot(X-C)=0$, where $X=(x_1,\ldots,x_n,x_{n+1})$. Writing this out in coordinates, we get

$$D_1f(c)(x_1-c_1)+D_2f(c)(x_2-c_2)+\cdots+D_nf(c)(x_n-c_n)-(x_{n+1}-f(c))=0,$$

which can be rearranged to give

$$x_{n+1}=f(c)+D_1f(c)(x_1-c_1)+D_2f(c)(x_2-c_2)+\cdots+D_nf(c)(x_n-c_n).$$

Exercise 4.6.8 Show that this tangent hyperplane is the *n-dimensional affine subspace* of $\mathbb{R}^{n+1}$ through the point $C=(c,f(c))$ spanned by the set of vectors $\{v_1,v_2,\ldots,v_n\}$, where for each j with $1\le j\le n$, we have

$$v_j=(0,0,\ldots,0,0,1,0,0,\ldots,0,0,D_jf(c)),$$

where the 1 in the vector v_j occurs in the jth place.

Exercise 4.6.9 Let $f:\mathbb{R}^n\to\mathbb{R}$ be a C^1 function, and let $S\subseteq\mathbb{R}^{n+1}$ be the graph of f. Fix a point $c\in\mathbb{R}^n$, and let T be the tangent hyperplane to S at the point $C=(c,f(c))\in S$.

i. For any $v\in\mathbb{R}^n$, define $\gamma_v:\mathbb{R}\to\mathbb{R}^n$ by $\gamma_v(t)=tv+c$. Let $\phi_v:\mathbb{R}\to\mathbb{R}^{n+1}$ be defined by $\phi_v(t)=(\gamma_v(t),f(\gamma_v(t)))$. Show that $C+\phi_v'(0)$ lies in T.

ii. Show that every vector $X\in T$ can be written in the form $X=C+V$, where $V=\phi_v'(0)$ for some $v\in\mathbb{R}^n$. Hence, every vector in the tangent hyperplane may be realized as the tangent vector to a curve in S.

4.7 Max-Min Problems

We now consider the problem of finding maximum and/or minimum values of a function on various subsets of $\mathbb{R}^n$ using properties of the derivative. The first observation is that, in order to discuss the notion of maximum or minimum, we need an order on the range of the function. Thus, we will restrict our attention to real-valued functions for the rest of this discussion.

The second observation is that there are two types of maxima and minima, namely, global and local.

Definition 4.7.1 Let $B\subseteq\mathbb{R}^n$ be any set, and let $f:B\to\mathbb{R}$ be a function. If there exists $p\in B$ such that $f(x)\le f(p)$ for all $x\in B$, then we say that $f(p)$ is the *global maximum* of f on B. Similarly, if there exists $q\in B$ such that $f(x)\ge f(q)$ for all $x\in B$, then we say that $f(q)$ is the *global minimum* of f on B.

Definition 4.7.2 Let $B \subseteq \mathbb{R}^n$ be any set, and let $f : B \to \mathbb{R}$ be a function. We say that f assumes a *local maximum* at a point $p \in B$ if there exists $r > 0$ such that $f(x) \leq f(p)$ for all $x \in B_r(p) \cap B$. Similarly, we say that f assumes a *local minimum* at a point $q \in B$ if there exists $r > 0$ such that $f(x) \geq f(q)$ for all $x \in B_r(q) \cap B$.

Note in particular that if $p \in B$ is a point at which f attains a global maximum, then f automatically has a local maximum at p, and similarly for minima. This implies that our search for global maxima and minima will begin with the search for local maxima and minima, a search to which we may naturally apply the tools of analysis. Note further that global maxima and global minima need not exist. If they do exist, they need not be unique. Indeed, a global maximum or minimum may even occur at an infinite number of points as in the case of a constant function. Finally, note that we do know of one very important special case in which global maxima and minima are guaranteed to exist: namely, when B is compact and $f : B \to \mathbb{R}$ is continuous.

Let us begin our search for the local maxima and minima of a function $f : B \to \mathbb{R}$. A local maximum or minimum can be found in one of two places: in the interior of B, which is an open subset of $\mathbb{R}^n$, or on the boundary of B, which is typically a hypersurface or the intersection of several hypersurfaces.

Let $p \in B$ be a point at which f has a local maximum or minimum. If such a point p occurs in the interior of B, and f is differentiable at p, then we will see that $Df(p) = 0$. If p occurs in the interior of B, but f is not differentiable at p, we must explore the behavior of the function in a neighborhood of the point p using various estimation techniques. Finally, if p is on the boundary of B, and the boundary of B may be realized as a hypersurface or the intersection of hypersurfaces, then the theory of Lagrange multipliers (see Section 4.8) may be used to determine the point p. Note the natural correspondence with the one-variable case, where maxima and minima can occur at three types of points: critical points ($f' = 0$), singular points (f' does not exist), and endpoints.

We begin the analysis by considering $f : U \to \mathbb{R}$, where U is an open subset of $\mathbb{R}^n$.

Theorem 4.7.3 Let U be an open set in $\mathbb{R}^n$, and $f : U \to \mathbb{R}$. If f has a local maximum at p, and f is differentiable at p, then $Df(p) = 0$.

Proof. Write $p = (p_1, p_2, \ldots, p_n)$. For each $j = 1, 2, \ldots, n$, define

$$f_j(x) = f(p_1, p_2, \ldots, p_{j-1}, x, p_{j+1}, \ldots, p_n).$$

The hypotheses of the theorem imply that, as a function of the single variable x, f_j is differentiable at p_j and has a local maximum at p_j. Hence $D_j f(p) = f_j'(p_j) = 0$. Since f is differentiable, this implies $Df(p) = 0$.

Definition 4.7.4 As in the one-variable case, a point where Df vanishes is called a *critical point* of f.

We note that, as in the case of one variable, f need not have either a local maximum or a local minimum at a critical point.

Example 4.7.5 The function $f(x, y) = xy$ has vanishing derivative at the origin, but has neither a maximum nor minimum there, since f is positive in the first and third quadrants and negative in the second and fourth quadrants.

Example 4.7.6 The following example illustrates the care that we must take in identifying critical points. Let

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

This function has all partial derivatives equal to zero at the origin, yet is not even continuous there.

Exercise 4.7.7 Let $p = (x_0, y_0)$ be a point in the plane, and let $ax + by + c = 0$ be the equation of a line in the plane. Verify that the distance from p to the this line is given by

$$\frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}$$

using the max-min technique discussed above. Find a similar formula for the distance from a point to a hyperplane in $\mathbb{R}^n$, and verify it.

Exercise 4.7.8 Let $p_j = (x_j, y_j)$, $j = 1, \ldots, m$, be m points in $\mathbb{R}^2$ with at least two distinct x_j's. Given a line $y = mx + b$, define $E(m, b) = \sum_{j=1}^{m}(y_j - (mx_j + b))^2$. Find the values of m and b that minimize this sum. For the values of m and b that minimize the function E, the line $y = mx + b$ is called the *ordinary least squares approximation* to the data $p_1, \ldots, p_m$.

Exercise 4.7.9 Let $p_j = (x_j, y_j)$, $j = 1, \ldots, m$, be m points in $\mathbb{R}^2$ with at least two distinct x_j's. Given a line ℓ with equation $ax + by + c = 0$, with a and b not both zero, denote by $d(p_j, \ell)$ the distance from the point p_j to the line ℓ. Consider the function $E(a, b, c) = \sum_{j=1}^{m} d(p_j, \ell)$. Find values of a, b, and c that minimize this function.

Exercise 4.7.10 Given a point $p = (x_0, y_0, z_0)$ with $x_0, y_0, z_0 > 0$, find an equation for a plane passing through this point that cuts off a tetrahedron of least volume in the first octant.

Once we have identified the critical points of a function f, we might then ask if there is a convenient way to determine whether f actually assumes a local maximum or minimum value at these points. One surefire way is to check the behavior of the function in a neighborhood of the critical point directly using inequalities. The next theorem shows that there is a test, similar to the second-derivative test for functions of one variable, for determining whether a function assumes a maximum or minimum value at a critical point.

Let U be an open set in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be twice differentiable at a point $x \in U$. We define the *Hessian* of f at x to be the *quadratic form* $H_x : \mathbb{R}^n \to \mathbb{R}$ defined by

$$H_x(v) = D(Df)(x)(v)(v).$$

In terms of the $n \times n$ matrix of second-order partial derivatives, denoted as

$$A_x = \begin{pmatrix} D_{11}f(x) & D_{12}f(x) & \cdots & D_{1n}f(x) \\ D_{21}f(x) & D_{22}f(x) & \cdots & D_{2n}f(x) \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1}f(x) & D_{n2}f(x) & \cdots & D_{nn}f(x) \end{pmatrix},$$

we can write $H_x(v) = {}^t v A_x v$.

Remark 4.7.11 We say that the quadratic form H_x is *positive definite* if $H_x(v) \geq 0$ for all $v \in \mathbb{R}^n$, and $H_x(v) = 0$ iff $v = 0$. Negative definiteness is defined similarly.

Lemma 4.7.12 Let U be an open set in $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be a C^2 function. If H_p is positive definite at a point $p \in U$, then H_x is positive definite for x in a neighborhood of p, and similarly when H_p is negative definite.

Proof. It is enough to prove the statement about positive definiteness. Let $m = \inf_{\|v\|=1} H_p(v)$; then $H_p(v) \geq m\|v\|^2$ for any vector v, because v is the product of $\|v\|$ and a unit vector. Since the $D_{ij}f$ which form the coefficients of H are continuous in x, for x sufficiently close to p, $H_x(v) \geq \frac{1}{2}m\|v\|^2$ for any vector v.

Exercise 4.7.13 Show that $m > 0$ to conclude the proof of the lemma.

Theorem 4.7.14 (Second derivative test for extrema) Let U be an open set in $\mathbb{R}^n$, let $f : U \to \mathbb{R}$ be a C^2 function, and let p be a critical point of f. If H_p is positive definite, then f assumes a local minimum at p, and if H_p is negative definite, then f assumes a local maximum at p.

Proof. It is enough to prove the statement about minima, since we can replace f by $-f$ for maxima. Given any sufficiently small vector h, the Taylor formula for f at p, to second order, is

$$f(p+h) = f(p) + Df(p) \cdot h + \frac{1}{2} H_{p+th}(h) t^2$$

for some $t \in [0, 1]$. Since p is a critical point, $Df(p) = 0$. By the lemma above, H_{p+th} is positive definite for h sufficiently small, so we have $f(p+h) \geq f(p)$, which proves that f assumes a local minimum at p.

Example 4.7.15 Consider the function $f(x, y) = xy$. The only critical point of f is $(0, 0)$. The Hessian matrix at this point is

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and the associated quadratic form is

$$H_{(0,0)}(u, v) = 2uv.$$

Thus, the second derivative test does not apply. Nonetheless, it is easy to see that f has no local extremum at $(0, 0)$.

Exercise 4.7.16 For each of the following functions $f : \mathbb{R}^2 \to \mathbb{R}$, find the critical points and compute the Hessian at each such point. Use the second derivative test to determine whether the critical points are local maxima or local minima, if possible. If the test does not apply, determine the nature of the critical point by other means.

i. $f(x, y) = x^2 + y^2$.

ii. $f(x, y) = x^4 + y^4$.

iii. $f(x, y) = x^2 - 2xy + y^2$.

iv. $f(x, y) = x^2 - y^4$.

v. $f(x, y) = (1 - xy)^2 + x^2$.

Exercise 4.7.17 Let $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}$ be C^2. Let $p \in U$ be a critical point of f and suppose that the matrix of H_p has negative determinant. Show that f does not have a local extremum at p.

The critical points of the type in the above exercise are referred to as *saddle points.*

Exercise 4.7.18 Find the critical points of $f(x, y) = x^3 + 8y^3 - 6xy - 2$. For each, determine if it is a local maximum, local minimum, or saddle point, if possible.

Exercise 4.7.19 Let $p_0 = (x_0, y_0, z_0)$ be a point in the first octant in $\mathbb{R}^3$, that is, $x_0, y_0, z_0 > 0$. Consider planes through p_0 that intersect the x, y, and z axes at $p_1 = (a, 0, 0)$, $p_2 = (0, b, 0)$, and $p_3 = (0, 0, c)$, respectively, with $a, b, c > 0$. Find the values of a, b, and c that minimize the area of the triangle with vertices p_1, p_2, and p_3.

4.8 Lagrange Multipliers

We now turn to the case where we wish to find the extreme values of a function $f : \mathbb{R}^n \to \mathbb{R}$ restricted to a hypersurface in $\mathbb{R}^n$ or the intersection of several hypersurfaces. We will first deal with the case of a single hypersurface S. The equation for the hypersurface S is sometimes referred to as a *constraint*, and the overall technique we develop here is called optimization with constraints.

In the case that S can be realized as the graph of a function $g : \mathbb{R}^{n-1} \to \mathbb{R}$, we are reduced to the problem of finding extreme values of the function $h : \mathbb{R}^{n-1} \to \mathbb{R}$ defined by

$$h(x_1, \ldots, x_{n-1}) = f(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1}))$$

on the open set $\mathbb{R}^{n-1}$, which we have previously done.

Even if the equation defining S cannot be solved explicitly for one of the variables, it may be still possible to describe S by an unconstrained set of $n-1$ variables by parametrization. While we will not pursue a full discussion of the notion of parametrization here, we will give a typical example of this approach.

Example 4.8.1 Let $S = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}$ be the unit sphere in $\mathbb{R}^3$. Suppose we wish to find the maxima and minima of the function $f(x, y, z) = x + y + z$ on S. We may parametrize S by using spherical coordinates as follows. Recall that, in general, spherical coordinates are given by

$$\begin{aligned} x &= \rho \cos\theta \sin\phi, \\ y &= \rho \sin\theta \sin\phi, \\ z &= \rho \cos\phi. \end{aligned}$$

In the specific case at hand, however, we have $\rho = 1$, so S is parametrized by the two unconstrained variables θ and ϕ.

Written in spherical coordinates, the function f becomes $f_0(\theta, \phi) = \cos\theta \sin\phi + \sin\theta \sin\phi + \cos\phi$. This parameterization causes problems at points where the Jacobian of the change of variables map above is singular, and this occurs exactly when $\rho = 0$ or $\sin\phi = 0$. In particular, we cannot determine whether points with $\sin\phi = 0$ are critical with this parameterization.

We have

$$\begin{aligned} \frac{\partial f_0}{\partial \theta} &= -\sin\theta \sin\phi + \cos\theta \sin\phi = \sin\phi(\cos\theta - \sin\theta), \\ \frac{\partial f_0}{\partial \phi} &= \cos\theta \cos\phi + \sin\theta \cos\phi - \sin\phi = \cos\phi(\cos\theta + \sin\theta) - \sin\phi. \end{aligned}$$

Note that $\partial f_0/\partial\theta = 0$ when $\cos\theta = \sin\theta$ or $\sin\phi = 0$. The former occurs when $\cos\theta = \sin\theta = \pm\frac{\sqrt{2}}{2}$. In this case, $\cos\theta + \sin\theta = \pm\sqrt{2}$, so $\partial f_0/\partial\phi = 0$ when $\pm\sqrt{2}\cos\phi - \sin\phi = 0$, that is, when $\tan\phi = \pm\sqrt{2}$.

We have thus shown that points with $\sin\theta = \cos\theta = \pm\frac{\sqrt{2}}{2}$ and $\tan\phi = \pm\sqrt{2}$ (where the signs are the same) are among the critical points of f.

Extreme value problems of the type discussed in Example 4.8.1 can be approached in a somewhat different fashion, by the method of *Lagrange multipliers*. The basic idea works as follows. Let S be the level set of a C^1 function $g : \mathbb{R}^n \to \mathbb{R}$. We have already observed that $\nabla g(x)$ is normal to the tangent hyperplane at any point $x \in S$. Now suppose that a function $f : \mathbb{R}^n \to \mathbb{R}$, when restricted to S, has an extreme value at the point $a \in S$. Then, for any C^1 curve $\phi : \mathbb{R} \to S$ with $\phi(0) = a$, the function $f \circ \phi : \mathbb{R} \to \mathbb{R}$ has an extreme value at 0, and hence

$$0 = (f \circ \phi)'(0) = \nabla f(\phi(0)) \cdot \phi'(0) = \nabla f(a) \cdot \phi'(0).$$

In other words, $\nabla f(a)$ is normal to the vector tangent to the curve ϕ at the point a. Since this is true for any such curve ϕ, $\nabla f(a)$ is normal to the tangent hyperplane of S at a. Since $\nabla g(a)$ is also normal to the tangent hyperplane of S at a, this implies that $\nabla f(a)$ is proportional to $\nabla g(a)$, so we can write $\nabla f(a) = \lambda \nabla g(a)$ for some $\lambda \in \mathbb{R}$. The scalar λ is known as a *Lagrange multiplier*. Before proving a theorem about Lagrange multipliers, let us study the example above using this new idea.

Example 4.8.2 Let $S = \{(x, y, z) \mid x^2+y^2+z^2 = 1\}$ be the unit sphere in $\mathbb{R}^3$. Suppose we wish to find the maxima and minima of the function $f(x, y, z) = x+y+z$ on S. We observe first that $\nabla f(x, y, z) = (1, 1, 1)$. The surface S can be realized as the level set $g(x, y, z) = 0$ for the function $g(x, y, z) = x^2 + y^2 + z^2 - 1$, which has gradient $\nabla g(x, y, z) = (2x, 2y, 2z)$. To identify the potential points $a = (x_0, y_0, z_0)$ on the surface S where f attains an extreme value, we set up the Lagrange multiplier equation $\nabla f(a) = \lambda \nabla g(a)$, which becomes

$$(1, 1, 1) = \lambda(2x_0, 2y_0, 2z_0).$$

This gives us three equations in our four unknowns, namely,

$$\begin{aligned} 2\lambda x_0 &= 1, \\ 2\lambda y_0 &= 1, \\ 2\lambda z_0 &= 1. \end{aligned}$$

But we must not forget that we have the original constraint, namely, that $g(a) = 0$. And hence, there is a fourth equation,

$$x_0^2 + y_0^2 + z_0^2 = 1.$$

Solving this system of four equations in four unknowns gives $\lambda = \pm\frac{\sqrt{3}}{2}$, which leads to $x_0 = \pm\frac{\sqrt{3}}{3}$, $y_0 = \pm\frac{\sqrt{3}}{3}$, $z_0 = \pm\frac{\sqrt{3}}{3}$. The point $a_1 = (+\frac{\sqrt{3}}{3}, +\frac{\sqrt{3}}{3}, +\frac{\sqrt{3}}{3})$ is the point where f attains its maximum value of $\sqrt{3}$ when restricted to S, and the point $a_2 = (-\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3})$ is the point where f attains its minimum value of $-\sqrt{3}$ when restricted to S.

Finally, it is worth noting that in this case, the precise value of λ is irrelevant and was merely used in an auxiliary fashion to help find the potential extrema.

Let us affirm this method with a theorem.

Theorem 4.8.3 Let U be an open set in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be C^1. Let $g : U \to \mathbb{R}$ be C^1 and let S be the hypersurface defined by $g(x) = 0$. If $f|_S$ has a local maximum or minimum at a point $a \in S$, and $\nabla g(a) \neq 0$, then there exists $\lambda \in \mathbb{R}$ such that $\nabla f(a) = \lambda \nabla g(a)$.

Proof. By Exercise 4.9.14 in the next section, every vector in the tangent hyperplane to S at a may be realized as the tangent vector $\phi'(0)$ to some C^1 curve $\phi : \mathbb{R} \to S$ with $\phi(0) = a$. By the discussion above, $\nabla f(a)$ and $\nabla g(a)$ both lie in the affine space normal to the tangent hyperplane. By Exercise 4.9.13, which you will also do in the next section, this space is one-dimensional, so we are done.

Example 4.8.4 Let us use the method of Lagrange multipliers to determine the points on the ellipse $x^2 + 4y^2 = 4$ that are closest to and farthest from the point $(1, 0)$. The square of the distance from the point (x, y) to this point is given by $f(x, y) = (x - 1)^2 + y^2$ and we wish to optimize this function subject to the constraint

$$g(x, y) = x^2 + 4y^2 - 4 = 0.$$

We have $\nabla f(x, y) = (2(x - 1), 2y)$ and $\nabla g(x, y) = (2x, 8y)$, so we consider the equations

$$\begin{array}{rcl} 2(x-1) &=& \lambda(2x) \\ 2y &=& \lambda(8y) \end{array}$$

It is easy to see that the only points on the ellipse satisfying these equations for some λ are $(\pm 2, 0)$. Plugging into f we conclude that $(2, 0)$ is the nearest point and $(-2, 0)$ is the farthest point.

Exercise 4.8.5 Let $S = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}$ be the unit sphere in $\mathbb{R}^3$. Find the maxima and minima of the function $f(x, y, z) = x^3 + y^3 + z^3$ on S.

Exercise 4.8.6 Consider the function $P : (0, \infty) \times (0, \infty) \to \mathbb{R}$ given by $P(L, C) = \alpha L^a C^b$, where α, a, b are positive constants, and $a + b = 1$. Let $R : (0, \infty) \times (0, \infty) \to \mathbb{R}$ be given by $R(L, C) = \beta_1 L + \beta_2 C$ for positive constants β_1 and β_2.

i. Maximize P subject to the constraint $R(L, C) = \kappa_1$, where κ_1 is a positive constant.

ii. Minimize R subject to the constraint $P(L, C) = \kappa_2$, where κ_2 is a positive constant.

In economics, the function P is known as the Cobb-Douglas production function.

Exercise 4.8.7 Let $x_1, x_2, \ldots, x_n$ be positive real numbers. Prove the *arithmetic-geometric mean inequality,*

$$(x_1 x_2 \cdots x_n)^{\frac{1}{n}} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Hint: Consider the function $f(x_1, x_2, \ldots, x_n) = \frac{x_1+x_2+\cdots+x_n}{n}$ subject to the constraint $x_1 x_2 \cdots x_n = c$, a constant.

Exercise 4.8.8 If a triangle has side lengths x, y, and z, so its perimeter is $2s = x + y + z$, its area A satisfies $A^2 = s(s - x)(s - y)(s - z)$. (This is Heron's Formula from classical geometry.) Show that, among all triangles with given perimeter, an equilateral triangle has the largest area.

There is also a max-min theorem with several constraints.

Theorem 4.8.9 Let U be an open set in $\mathbb{R}^n$, let $f : U \to \mathbb{R}$ be C^1, and let $g_1, g_2, \ldots, g_m : U \to \mathbb{R}$ be C^1, where $m < n$. Let S be the intersection of the hypersurfaces S_i defined by $g_i(x) = 0$. If $f|_S$ has a local maximum or minimum at a point $a \in S$, and the vectors $\nabla g_1(a), \nabla g_2(a), \ldots, \nabla g_m(a)$ form a linearly independent set, then $\nabla f(a)$ is a linear combination of $\nabla g_1(a), \nabla g_2(a), \ldots, \nabla g_m(a)$.

If $\nabla f(a) = \lambda_1 \nabla g_1(a) + \lambda_2 \nabla g_2(a) + \cdots + \lambda_m \nabla g_m(a)$ as in the theorem, the scalars $\lambda_1, \lambda_2, \ldots, \lambda_m$ are called *Lagrange multipliers*. We do not prove this theorem here, but we present an example and some exercises to illustrate the theory.

Example 4.8.10 Given the line defined by $P(x, y) = y - (mx + k) = 0$, and an ellipse defined by $E(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0$, we wish to find the minimum distance between a point on the line and a point on the ellipse. Equivalently, we minimize the square distance

$$d(x_1, y_1, x_2, y_2) = (x_1 - x_2)^2 + (y_1 - y_2)^2$$

subject to the constraints

$$g_1(x_1, y_1, x_2, y_2) = P(x_1, y_1) = 0,$$

$$g_2(x_1, y_1, x_2, y_2) = E(x_2, y_2) = 0.$$

We assume that the line does not intersect the ellipse. The reader should verify that this will occur when $|k| > |b|$ and $m^2 < \frac{k^2 - b^2}{a^2}$. These conditions should also emerge from the solution below.

We have

$$\begin{aligned}
\nabla d &= (2(x_1 - x_2), 2(y_1 - y_2), -2(x_1 - x_2), -2(y_1 - y_2)), \\
\nabla g_1 &= (-m, 1, 0, 0), \\
\nabla g_2 &= \left(0, 0, \frac{2x_2}{a^2}, \frac{2y_2}{b^2}\right).
\end{aligned}$$

We first note that ∇g_1 and ∇g_2 are everywhere linearly independent, so by Theorem 4.8.9, if (x_1, x_2, y_1, y_2) is a maximum or minimum value for d subject to the constraints g_1 and g_2, then the following system of six equations in six unknowns must be satisfied.

$$\begin{aligned}
2(x_1 - x_2) &= -\lambda_1 m, \\
2(y_1 - y_2) &= \lambda_1, \\
-2(x_1 - x_2) &= \lambda_2 \frac{2x_2}{a^2}, \\
-2(y_1 - y_2) &= \lambda_2 \frac{2y_2}{b^2}, \\
y_1 &= mx_1 + k, \\
\frac{x_2^2}{a^2} + \frac{y_2^2}{b^2} - 1 &= 0.
\end{aligned}$$

Exercise 4.8.11 Solve the above system.

Exercise 4.8.12 Consider the plane defined by $P(x, y, z) = Ax + By + Cz + D = 0$ and the ellipsoid defined by $E(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0$.

i. Find conditions on A, B, C, D, a, b, c such that the plane and the ellipsoid do not intersect.

ii. Find the minimum distance between the plane and the ellipsoid when they do not intersect.

Exercise 4.8.13 Let v and w be vectors in $\mathbb{R}^n$. Find the maximum and minimum values of $f(v, w) = v \cdot w$ subject to the constraints $\|v\| = \|w\| = 1$.

Exercise 4.8.14 Consider two nonparallel planes in $\mathbb{R}^3$. Find the point on their line of intersection closest to the origin in $\mathbb{R}^3$.

Exercise 4.8.15 In the situation of Theorem 4.8.9, what happens if the number of constraints exceeds or equals the number of variables, that is, if $m \geq n$?

4.9 Implicit and Inverse Function Theorems

Let n and m be positive integers, and let $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$. We attack the problem of determining a set of conditions under which we can solve $f = 0$ explicitly for the first m variables, which we denote with y's, in terms of the remaining n variables, which we denote with x's. Thus, if we write f in the form $f(y_1, y_2, \ldots, y_m; x_1, x_2, \ldots, x_n)$ as a function of $m + n$ variables, we would like to produce functions $\phi_1, \phi_2, \ldots, \phi_m$ from an open set in $\mathbb{R}^n$ to $\mathbb{R}$ such that, on some open set in $\mathbb{R}^m \times \mathbb{R}^n$, the assignment $y_j = \phi_j(x_1, x_2, \ldots, x_n)$ solves the equation $f(y_1, y_2, \ldots, y_m; x_1, x_2, \ldots, x_n) = 0$.

Of course, we can expect some conditions on differentiability, nonvanishing of the determinant of a Jacobian matrix, and other properties. We begin by considering the case when $m = 1$ and n is arbitrary. Here, the proof involves only simple results from basic calculus. This development can be completed by induction on m, but the techniques are somewhat tedious.

Example 4.9.1 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(y, x) = y^2 + x^2 - 1$. Here, the equation $f(y, x) = 0$ gives us the unit circle in $\mathbb{R}^2$. As is ordinarily discussed in elementary calculus courses, this analysis produces two functions that serve our purpose here. We wish to solve for y in terms of x, and this can be done with either of the equations $y = \sqrt{1 - x^2}$, for $-1 \leq x \leq 1$, or $y = -\sqrt{1 - x^2}$, for $-1 \leq x \leq 1$. Note that, in either case, we have $f(y, x) = 0$. Also, $\frac{\partial f}{\partial y} = 2y \neq 0$ when $y \neq 0$. This condition about the nonvanishing of the derivative is one that will be required in the next theorem.

Theorem 4.9.2 Let f be a C^1 function from an open set in $\mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}$. Let (y_0, x_0) be a point in this open set such that $f(y_0, x_0) = 0$, and

$$\frac{\partial f}{\partial y}(y_0, x_0) \neq 0.$$

Then, there exist open sets $V \subseteq \mathbb{R}$ and $U \subseteq \mathbb{R}^n$, such that $(y_0, x_0) \in V \times U$, and to every $x \in U$, there exists a unique $\phi(x)$ in V such that $f(\phi(x), x) = 0$, and $\phi : U \to \mathbb{R}$ is C^1 on U. Furthermore,

$$\frac{\partial \phi}{\partial x_j}(x) = -\frac{\frac{\partial f}{\partial x_j}(\phi(x), x)}{\frac{\partial f}{\partial y}(\phi(x), x)}.$$

Proof. We can assume that $\frac{\partial f}{\partial y}(y_0, x_0) > 0$. Since f is C^1, there is an open set $U' \subseteq \mathbb{R}^n$ containing x_0 and an open interval $V = (y_0 - \varepsilon, y_0 + \varepsilon) \subseteq \mathbb{R}$ such that $\frac{\partial f}{\partial y}(y, x) > 0$ for $x \in U', y \in V$. Then $f(y_0 + \varepsilon, x_0) > 0$, and $f(y_0 - \varepsilon, x_0) < 0$. By continuity of f, there exists an open set $U \subseteq U' \subseteq \mathbb{R}^n$ containing x_0 such that, for all $x \in U$, $f(y_0 - \varepsilon, x) < 0$ and $f(y_0 + \varepsilon, x) > 0$.

Fix $x \in U$. By the Intermediate Value Theorem, there exists $y \in V$ such that $f(y, x) = 0$. The function $g(y) = f(y, x)$ satisfies $g'(y) = \frac{\partial f}{\partial y}(y, x) > 0$ for $y \in V$, so by Rolle's Theorem, the value of y for which $f(y, x) = 0$ is unique. We set $\phi(x) = y$.

The continuity of ϕ at x_0 follows from the fact that we can choose ε in the above construction to be arbitrarily small. This same argument holds for any $x \in U$, which proves the continuity of ϕ.

Since $f(\phi(x), x) = 0$, formally, from the chain rule, we get

$$0 = \frac{\partial}{\partial x_j}[f(\phi(x), x)] = \frac{\partial f}{\partial y}(\phi(x), x)\frac{\partial \phi}{\partial x_j}(x) + \frac{\partial f}{\partial x_j}(\phi(x), x).$$

Thus,

$$\frac{\partial \phi}{\partial x_j}(x) = -\frac{\frac{\partial f}{\partial x_j}(\phi(x), x)}{\frac{\partial f}{\partial y}(\phi(x), x)}.$$

Using the formal expression for the derivative given above, we can write the difference quotient for the derivative of ϕ, subtract this expression, and show that the limit of the difference goes to zero.

A continuation of this proof to yield the Implicit Function Theorem for general m is outlined in Osgood []. If the reader enjoys a stiff climb over rocks and across streams, they might wish to pursue this proof. As an alternative, they might wish to consult Folland's Appendix [].

We now make an about-face and move directly to a proof of the Inverse Function Theorem. This seems to be the more common approach in current mathematics texts. We first review the single-variable case, to remind the reader of the nature of the result.

Suppose that U is an open set in $\mathbb{R}$ and $f : U \to \mathbb{R}$ is C^1. Take a point $x_0 \in U$. We saw earlier in the chapter that if $f'(x_0) \neq 0$, then f is monotonic in an open interval I around x_0. This, of course, implies that f is one-to-one on I. Moreover, $f(I)$ is an open interval J contained in $\mathbb{R}$, and $f^{-1} : J \to I$ is C^1 and $(f^{-1})'(y) = (f'(f^{-1}(y)))^{-1}$. It is worth remarking at this point that this one-variable theorem requires the continuity of the derivative. See, for example, Exercise 4.1.22.

The *Inverse Function Theorem* is the generalization of this result to functions $f : \mathbb{R}^n \to \mathbb{R}^n$. Essentially, the theorem says that if such an f is C^1 and has a nonsingular derivative at a point x_0, then, in some neighborhood of x_0, f is invertible, and f^{-1} is also C^1. We approach this through a sequence of lemmas and corollaries, which can be combined to provide a proof of the Inverse Function Theorem.

Lemma 4.9.3 Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^n$ be C^1. Take $x_0 \in U$ and suppose that $Df(x_0)$ is nonsingular. Then there exists a neighborhood W of x_0 and a constant $c > 0$ such that

$$\|f(y) - f(x)\| \geq c\|y - x\| \text{ for all } x, y \in W.$$

Proof. For any nonsingular linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, we know that $\|T(y) - T(x)\| \leq \|T\|\|y - x\|$. It follows immediately that

$$\|T(y) - T(x)\| \geq \|T^{-1}\|^{-1}\|y - x\|.$$

Take $c = \|Df(x_0)^{-1}\|^{-1}/2$. Suppose that $f_1, f_2, \ldots, f_n$ are the coordinate functions of f. Of course, these are C^1, so that there exists a convex neighborhood W of x_0 such that $\|Df_i(y) - Df_i(x_0)\| \leq c/n$ for $y \in W$ and all i. Now the Mean Value Theorem implies that, given $x, y \in W$, there exists a point ξ_i on the line segment joining x and y such that

$$f_i(y) - f_i(x) = Df_i(\xi_i)(y - x).$$

Consequently, for $x, y \in W$ and each i, we have

$$\|f_i(y) - f_i(x) - Df_i(x_0)(y - x)\| \leq \frac{c}{n}\|y - x\|.$$

It follows immediately that

$$\|f(y) - f(x) - Df(x_0)(y - x)\| \leq c\|y - x\|.$$

Now, using the triangle inequality, we get

$$\|f(y) - f(x)\| \geq c\|y - x\| \text{ for } x, y \in W.$$

Corollary 4.9.4 Let $U \subseteq \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^n$ be C^1. Take $x_0 \in U$ and suppose that $Df(x_0)$ is nonsingular. Then there exists a neighborhood W of x_0 such that $f|_W$ is one-to-one.

Proof. Exercise.

Corollary 4.9.5 Let $U \subseteq \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^n$ be C^1. Take $x_0 \in U$ and suppose that $Df(x_0)$ is nonsingular. Then there exists a neighborhood V of x_0 such that $f(V)$ is open, and $f|_V : V \to f(V)$ is a homeomorphism.

Proof. Using the previous lemma we can pick a neighborhood W of x_0 such that $\overline{W} \subseteq U$ and for some constant c, $\|f(y) - f(x)\| \geq c\|y - x\|$ for all $x, y \in \overline{W}$, and finally $Df(x)$ is nonsingular for $x \in \overline{W}$. Let V be any open ball contained in W and let $S = \partial\overline{V}$. Given a point $x \in V$ with $y = f(x) \notin f(S)$, since $f(S)$ is compact, the distance from y to $f(S)$, which we denote by d, is greater than zero. To show that $f(V)$ is open, we establish that $f(V)$ contains $B_{d/2}(y)$. To see this, take $z \in B_{d/2}(y)$. Then $\|z - y\| < d/2$. Moreover, the distance from z to $f(\overline{V})$, which equals

$$\inf_{x \in \overline{V}} \|z - f(x)\|,$$

is less than $d/2$. Since the distance from y to $f(S)$ is equal to d, it follows from the triangle inequality that the distance from z to $f(S)$ is greater than $d/2$. For $x \in \overline{V}$, we define the function

$$g(x) = \|z - f(x)\|^2 = \sum_{i=1}^{n}(z_i - f_i(x))^2.$$

We want to minimize this function. We know that there exists $x_1 \in \overline{V}$ such that

$$\|z - f(x_1)\|^2 = d(z, f(\overline{V}))^2.$$

From the previous inequalities, it follows that $x_1 \notin S$, so that $x_1 \in V$. So the minimum of g occurs at x_1. This implies that

$$0 = D_j g(x_1) = -2 \sum_{i=1}^{n} (z_i - f_i(x_1)) D_j f_i(x_1).$$

It follows immediately that $Df(x_1)(z - f(x_1)) = 0$, and since $Df(x_1)$ is nonsingular, $z = f(x_1)$. Since f is invertible on V by the previous corollary, f is a homeomorphism.

Theorem 4.9.6 (Inverse Function Theorem) Let U be an open set in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}^n$ be C^1. Let $x_0 \in U$ be such that $Df(x_0)$ is nonsingular. Then there exists a neighborhood V of x_0 such that

i. $f : V \to f(V)$ is a bijection;

ii. $f(V)$ is an open set in $\mathbb{R}^n$;

iii. $f^{-1} : f(V) \to V$ is C^1 and $Df^{-1}(f(x)) = (Df(x))^{-1}$ for $x \in V$.

Proof. The first two statements follow from Corollary 4.9.4 and Corollary 4.9.5.

Now where are we? We have proved that $f : V \to f(V)$ is one-to-one and that $f(V)$ is open. We consider the map $f^{-1} : f(V) \to V$. By the last corollary, this map is continuous. We want to prove that it is C^1. A formal computation using the chain rule shows that $Df^{-1}(f(x))$, if it exists, is equal to $(Df(x))^{-1}$. To complete the proof, we take $y = f(x) \in f(V)$ and consider the difference quotient

$$\frac{\|f^{-1}(z) - f^{-1}(y) - Df(x)^{-1}(z - y)\|}{\|z - y\|}.$$

We can write this as

$$\frac{\|Df(x)^{-1}[Df(x)(f^{-1}(z) - f^{-1}(y)) - (z - y)]\|}{\|z - y\|},$$

which is less than or equal to

$$\|Df(x)^{-1}\| \frac{\|Df(x)(f^{-1}(z) - f^{-1}(y)) - (z - y)\|}{\|z - y\|}.$$

And using the inequality of Lemma 4.9.3, this is less than or equal to

$$\frac{\|Df(x)^{-1}\|}{c} \frac{\|Df(x)(f^{-1}(z) - x) - (z - f(x))\|}{\|f^{-1}(z) - x\|}.$$

Finally, since f^{-1} is continuous, this last expression goes to zero as z goes to y.

The function Df^{-1} is continuous because f is C^1 and, by the exercise below, the map $A \mapsto A^{-1}$ is continuous from $GL_n(\mathbb{R}) \to GL_n(\mathbb{R})$.

Exercise 4.9.7 Prove that the map $f : GL_n(\mathbb{R}) \to GL_n(\mathbb{R})$ given by $f(A) = A^{-1}$ is continuous.

We now use the Inverse Function Theorem to outline a proof of the Implicit Function Theorem in a sequence of exercises.

Theorem 4.9.8 (Implicit Function Theorem) Let f be a C^1 function from an open set in $\mathbb{R}^m \times \mathbb{R}^n$ to $\mathbb{R}^m$. Let (y_0, x_0) be a point in this open set such that $f(y_0, x_0) = 0$, and the matrix

$$L = \left(\frac{\partial f_i}{\partial y_j}(y_0, x_0) \right)_{i,j=1,\ldots,m}$$

is nonsingular. Then there exist open sets $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ such that $(y_0, x_0) \in V \times U$, and to every $x \in U$, there exists a unique $y = \phi(x)$ in V such that $f(\phi(x), x) = 0$, and ϕ is C^1 on U.

Proof. To begin, we construct the function $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ defined by $F(y, x) = (f(y, x), x)$.

Exercise 4.9.9 Show that $\det DF(y_0, x_0) = \det L$.

By applying the Inverse Function Theorem to F, we obtain neighborhoods $U' \subseteq \mathbb{R}^n$ of x_0 and $V \subseteq \mathbb{R}^m$ of y_0 such that F has a C^1 inverse on the open set $W = F(V \times U')$. Let $U = \{x \in U' \mid (0, x) \in W\}$.

Exercise 4.9.10 Show that there exists a C^1 function $\Phi : W \to \mathbb{R}^m$ such that $F^{-1}(y, x) = (\Phi(y, x), x)$ on W.

Exercise 4.9.11 Show that $f(\Phi(y, x), x) = y$.

Exercise 4.9.12 Let $\phi(x) = \Phi(0, x)$. Show that ϕ satisfies the conclusion of the theorem.

This concludes the proof of the implicit function theorem.

Exercise 4.9.13 Let $F : \mathbb{R}^{n+1} \to \mathbb{R}$ be a C^1 function, and let S be the hypersurface defined by $F(x) = 0$. Suppose that $\nabla F(x) \neq 0$ for all $x \in S$, and fix $x_0 \in S$. Show that the tangent hyperplane to S at x_0 is spanned by n linearly independent vectors. *Hint*: Use the Implicit Function Theorem, and apply Exercise 4.6.8.

Exercise 4.9.14 Let $F : \mathbb{R}^{n+1} \to \mathbb{R}$ be a C^1 function, and let S be the hypersurface defined by $F(x) = 0$. Suppose that $\nabla F(x) \neq 0$ for all $x \in S$, and fix $x_0 \in S$. Show that every vector in the tangent hyperplane to S at x_0 is the tangent vector to some C^1 curve in S through x_0. *Hint*: Use the Implicit Function Theorem, and apply Exercise 4.6.9.

Exercise 4.9.15 Let $f : \mathbb{R}^3 \setminus \{(0,0,0)\} \to \mathbb{R}^3 \setminus \{(0,0,0)\}$ be given by

$$f(x, y, z) = \left(\frac{x}{x^2 + y^2 + z^2}, \frac{y}{x^2 + y^2 + z^2}, \frac{z}{x^2 + y^2 + z^2} \right).$$

Show that f is locally invertible at every point in $\mathbb{R}^3 \setminus \{(0,0,0)\}$. Find an explicit formula for f^{-1}.

Exercise 4.9.16 Consider the equations

$$ab^2 + cde + a^2d = 3 \quad \text{and} \quad ace^3 + 2bd - b^2e^2 = 2.$$

Determine which pairs of variables can be solved for in terms of the other three near the point $(a, b, c, d, e) = (1, 1, 1, 1, 1)$.

We now present a result, Lemma 4.9.19, related to the previous theorems that will be of assistance in the change-of-variables theorem for multiple integrals. Given a function $\phi : \mathbb{R}^n \to \mathbb{R}^n$ we want to define the best linear approximation to ϕ at a point using the Taylor polynomials of its real-valued coordinate functions. Let $U \subseteq \mathbb{R}^n$ be an open set and let $\phi : U \to \mathbb{R}^n$ be a C^1 function. For $y \in U$, define a function $T^y(x)$ by $T^y(x) = (T_1^y(x), T_2^y(x), \ldots, T_n^y(x))$ where

$$T_j^y(x) = \phi_j(y) + \sum_{k=1}^{n} D_k\phi_j(y)(x_k - y_k).$$

This is the first-order Taylor polynomial of ϕ_j in the direction of the unit vector $\frac{x-y}{\|x-y\|}$, evaluated at $t = \|x - y\|$ in the notation of Definition 4.5.9.

Lemma 4.9.17 Let $K \subseteq U$ be compact. Then

$$\lim_{x \to y} \frac{\|T^y(x) - \phi(x)\|}{\|x-y\|} = 0$$

uniformly in y, for $y \in K$.

Proof. It is enough to prove the analogous statement for each component function separately. By the Mean Value Theorem, there exists $\xi \in [0,1]$ such that

$$\begin{aligned} |T_j^y(x) - \phi_j(x)| &= \left| \sum_{k=1}^{n} (D_k\phi_j(y) - D_k\phi_j(y + \xi(x-y)))(x_k - y_k) \right| \\ &\leq \|x-y\| \|D\phi_j(y) - D\phi_j(y + \xi(x-y))\| \end{aligned}$$

by the Cauchy-Schwarz inequality. Dividing both sides by $\|x-y\|$, we get

$$\frac{|T_j^y(x) - \phi_j(x)|}{\|x-y\|} \leq \|D\phi_j(y) - D\phi_j(y + \xi(x-y))\|.$$

The right-hand side goes to zero as $x \to y$ uniformly in y, for $y \in K$, since ϕ is C^1 and K is compact.

Definition 4.9.18 We define a *generalized rectangle* in $\mathbb{R}^n$ to be a set of the form $R = I_1 \times I_2 \times \cdots \times I_n$, where I_i, $i = 1, \ldots, n$, are bounded intervals in $\mathbb{R}$. If the intervals are all open, we call R an *open generalized rectangle* and if the intervals are all closed, we call R a *closed generalized rectangle*. In the particular case when the intervals are of the form $I_i = [a_i, b_i)$, we refer to R as a *half-open generalized rectangle*.

Lemma 4.9.19 Let $U \subseteq \mathbb{R}^n$ be open, and suppose that $\phi : U \to \phi(U)$ is C^1, one-to-one onto its image, and has C^1 inverse. Let $R \subseteq U \subseteq \mathbb{R}^n$ be a generalized rectangle with nonempty interior. For $y \in U$ and $\lambda > 0$, we denote by λR^y the generalized rectangle with center y similar to R with sides scaled by λ. For $0 < \varepsilon < 1$ and $h > 0$, we define

$$\begin{aligned} R_1 = R_1(h,y) &= (1-\varepsilon)hR^y \\ R_2 = R_2(h,y) &= hR^y \\ R_3 = R_3(h,y) &= (1+\varepsilon)hR^y \end{aligned}$$

Then, for each compact set $K \subseteq U$, there exists a number $h_0(K) > 0$ such that, if $0 < h < h_0(K)$ and $y \in K$ is such that $R_2(h,y) \subseteq U$, then

$$T^y(R_1(h,y)) \subseteq \phi(R_2(h,y)) \subseteq T^y(R_3(h,y)).$$

Proof. Since T^y has a continuously differentiable inverse, there exists a constant $C > 0$ such that

$$\|x - z\| \leq C\|T^y(x) - T^y(z)\|$$

for all $x, y, z \in K$. If we apply this estimate to the previous lemma, we have

$$\frac{\|T^y(x) - \phi(x)\|}{\|T^y(x) - \phi(y)\|} \longrightarrow 0$$

as $x \to y$ uniformly in y. Then

$$\sup_{x \in \partial R_2(h,y)} \frac{\|T^y(x) - \phi(x)\|}{\|T^y(x) - \phi(y)\|} \longrightarrow 0$$

as $h \to 0$, so there exists a constant $h_0(K) > 0$ such that, for $0 < h < h_0(K)$,

$$\phi(\partial R_2) \subseteq T^y(R_3) \setminus \overline{T^y(R_1)}.$$

It follows that $\phi(R_2) \subseteq T^y(R_3)$.

Since $\phi(R_2)$ has nonempty interior and contains $\phi(y) = T^y(y)$, there exists some $h' > 0$ such that $T^y(h'R^y) \subseteq \phi(R_2)$. Let

$$h'' = \sup\{h' \mid T^y(h'R^y) \subseteq \phi(R_2)\}.$$

Then $T^y(h''R^y) \cap \phi(\partial R_2) \neq \varnothing$, and hence $h'' > (1-\varepsilon)h$. Thus, $T^y(R_1) \subseteq \phi(R_2)$.

4.10 Independent Projects

4.10.1 Term-by-Term Differentiation The question we discuss here is the following. Let $U \subseteq \mathbb{R}$ be an open set, and let $[a,b] \subset U$. If we have a series of real-valued, differentiable functions $\sum_{k=1}^{\infty} f_k$ on U that converges on $[a,b]$ to f, is the series differentiable, and if so, does

$$f' = \sum_{k=1}^{\infty} f_k'$$

on $[a,b]$? The answer is generally no.

However, if the series $\sum_{k=1}^{\infty} f_k'$ converges uniformly on $[a,b]$, the answer is yes. Let $S_n(x) = \sum_{k=1}^{n} f_k(x)$, $R_n(x) = f(x) - S_n(x)$, and $f^*(x) = \sum_{k=1}^{\infty} f_k'(x)$ for $x \in [a,b]$, and take $\varepsilon > 0$. Then $S_n'(x) = \sum_{k=1}^{n} f_k'(x)$ on $[a,b]$. Since the series of derivatives converges uniformly, there exists $N \in \mathbb{N}$ such that, for all $n > N$,

1. $|S_n'(x) - f^*(x)| \leq \varepsilon/8$ on $[a,b]$;
2. $|S_{n+p}'(x) - S_n'(x)| \leq \varepsilon/8$ on $[a,b]$, for all $p \geq 0$.

Take $n > N$ and $p \geq 0$.

Exercise 4.10.1 In this exercise, you will show that for all $x \in [a,b]$ and $h \neq 0$ such that $x+h \in [a,b]$, we have

$$\left|\frac{R_n(x+h) - R_n(x)}{h}\right| \leq \frac{\varepsilon}{8}.$$

i. Assume there exists x and h as above such that

$$\left|\frac{R_n(x+h) - R_n(x)}{h}\right| = \frac{\varepsilon}{8} + \eta,$$

where $\eta > 0$. By suitable maneuvering, show that

$$\begin{aligned}
&\left|\frac{R_n(x+h) - R_n(x)}{h}\right| \\
&= \left|\Big(\frac{S_{n+p}(x+h) - S_n(x+h)}{h} - \frac{S_{n+p}(x) - S_n(x)}{h}\Big)\right. \\
&\qquad \left. + \Big(\frac{f(x+h) - S_{n+p}(x+h)}{h} - \frac{f(x) - S_{n+p}(x)}{h}\Big)\right| = \frac{\varepsilon}{8} + \eta.
\end{aligned}$$

ii. Fixing n, show that if we take p large enough,

$$\left|\frac{f(x+h) - S_{n+p}(x+h)}{h}\right| < \frac{1}{2}\eta, and \left|\frac{f(x) - S_{n+p}(x)}{h}\right| < \frac{1}{2}\eta.$$

This yields the inequality

$$\left|\frac{S_{n+p}(x+h)-S_n(x+h)}{h}-\frac{S_{n+p}(x)-S_n(x)}{h}\right|>\frac{\varepsilon}{8}.$$

iii. By applying the Mean Value Theorem, show that

$$|S'_{n+p}(x_1)-S'_n(x_1)|>\frac{\varepsilon}{8}$$

for some point x_1 between x and $x+h$, which contradicts (2).

We now have

$$\left|\frac{f(x+h)-f(x)}{h}-\frac{S_n(x+h)-S_n(x)}{h}\right|=\left|\frac{R_n(x+h)-R_n(x)}{h}\right|\leq\frac{\varepsilon}{8}$$

for all $x\in[a,b]$ and $h\neq 0$ such that $x+h\in[a,b]$. Since S_n is differentiable at x, there exists $\delta>0$ such that if $0<|h|<\delta$, then

$$\left|\frac{S_n(x+h)-S_n(x)}{h}-S'_n(x)\right|\leq\frac{\varepsilon}{8}.$$

Exercise 4.10.2 Show that

$$\left|\frac{f(x+h)-f(x)}{h}-S'_n(x)\right|\leq\frac{\varepsilon}{8}+\frac{\varepsilon}{8}=\frac{\varepsilon}{4},$$

and use this inequality to conclude that $f'(x)=f^*(x)$.

Exercise 4.10.3 Show that the series for f must converge uniformly on $[a,b]$.

Exercise 4.10.4 By giving a counterexample, show that if the series for f^* does not converge uniformly, then the conclusion can be false. (*Hint:* Try finding a series of differentiable functions that converges to $f(x)=|x|$ on $[-1,1]$.)

Exercise 4.10.5 How can one apply this result to Exercise 4.3.12?

Exercise 4.10.6

i. Let $f(x)=e^x$. Use the fact that $y=f(x)$ satisfies the differential equation $y'=y$ to find the power series for f.

ii. Let $f(x)=\sin x$ and $g(x)=\cos x$. Using the fact that both $y=f(x)$ and $y=g(x)$ satisfy the differential equation $y''=-y$, find the power series for f and g. (

iii. Hint: f is an odd function, and g is an even function.)

Exercise 4.10.7 Show that the argument is considerably simplified if you use the Fundamental Theorem of Calculus at an appropriate juncture. (If you don't remember the Fundamental Theorem, see the next chapter.)

4.10.2 Leibniz's Rule In this project, we use elementary properties of the Riemann integral in one variables, including properties of improper integrals. We treat these topics in greater detail in the next chapter.

The classical theorem of Leibniz about differentiation under the integral sign states the following. Let U be an open set in $\mathbb{R}^2$, and let $R = [a,b] \times [c,d]$ be a closed rectangle contained in U. Let $f(x,y)$ be a continuous, real-valued function on U such that $D_2 f(x,y)$ exists and is continuous on U, and let

$$F(y) = \int_a^b f(x,y)\, dx.$$

Then the derivative of F exists and

$$F'(y) = \int_a^b D_2 f(x,y)\, dx.$$

Exercise 4.10.8 Show that the above conclusion is equivalent to the statement

$$\lim_{h \to 0} \Big(\frac{F(y+h) - F(y)}{h} - \int_a^b D_2 f(x,y)\, dx \Big) = 0.$$

Exercise 4.10.9 Use the Mean Value Theorem to show

$$F(y+h) - F(y) = \int_a^b D_2 f(x, y + \theta h)\, dx$$

for some $0 \leq \theta \leq 1$.

Exercise 4.10.10 Use continuity (uniform continuity) to show that $F'(y) = \int_a^b D_2 f(x,y)\, dx$.

There is a more general version of Leibniz's rule using variable bounds of integration. Let $f(x,t)$ and $D_2 f(x,t)$ be continuous on the domain $[a,b] \times V$ w;here $V \subset \mathbb{R}$ is open. Assume $\alpha, \beta : V \to [a,b]$ are C^1 functions. Define

$$\Phi(t) = \int_{\alpha(t)}^{\beta(t)} f(x,t)\, dx.$$

Exercise 4.10.11 Show that Φ is differentiable on V, and

$$\Phi'(t) = f(\beta(t),t)\beta'(t) - f(\alpha(t),t)\alpha'(t) + \int_{\alpha(t)}^{\beta(t)} \frac{\partial f}{\partial t}(x,t)\, dx.$$

The next exercise presents an interesting result that can be proved in several ways. This result is

$$\int_0^\infty \frac{\sin(x)}{x}\, dx = \frac{\pi}{2}.$$

Exercise 4.10.12 Let $f(x,t) = e^{-xt}\sin(x)/x$ for $t, x > 0$, and let

$$\Phi(t) = \int_0^\infty f(x,t)\, dx.$$

i. Show that $\Phi(t) \to 0$ as $t \to \infty$.

ii. Use integration by parts twice to show that

$$\Phi'(t) = -\frac{1}{1+t^2}.$$

iii. Conclude that

$$\Phi(t) = \frac{\pi}{2} - \tan^{-1}(t)$$

for $t > 0$.

iv. Use integration by parts to show that

$$\left| \int_r^\infty f(x,t)\, dx \right| \leq \frac{2}{r}$$

for all $r > 0$ and $t > 0$.

v. Show that

$$\lim_{t \to 0^+} \int_0^r f(x,t)\, dx = \int_0^r \frac{\sin(x)}{x}\, dx.$$

vi. Show that

$$\int_0^\infty \frac{\sin(x)}{x}\, dx = \frac{\pi}{2}.$$

An interesting article about Leibniz's rule by H. Flanders can be found in the American Mathematical Monthly, Jun–July 1973.

Chapter 5

Integration on $\mathbb{R}^n$

The integral of a function over a region R is designed for numerous applications. The most basic of these is the determination of the "area under a curve" of a function of one real variable defined on a closed interval $[a, b] \subset \mathbb{R}$. The idea moves in a number of directions, such as volume in $\mathbb{R}^n$, arc length of a curve, flux through a surface, and other applications. Whereas differentiation is designed to describe the behavior of a function at a point or in the neighborhood of a point, integration is intended to study the cumulative properties of a function on a larger domain.

5.1 The Riemann Integral in One Variable: Definitions

We begin this chapter on integration in $\mathbb{R}^n$ with a discussion of the Riemann integral in $\mathbb{R}$. We start with a closed, bounded interval $[a, b] \subset \mathbb{R}$ and a bounded function $f : [a, b] \to \mathbb{R}$. As we proceed through the definition, we will derive conditions on the function f that ensure the existence of the Riemann integral.

Definition 5.1.1 A *partition* P of the interval $[a, b]$ is a finite set of points

$$P = \{a_0, a_1, \ldots, a_k\}$$

such that

$$a = a_0 < a_1 < \cdots < a_{k-1} < a_k = b.$$

The *mesh* of the partition P is

$$|P| = \max_{1 \leq i \leq k} (a_i - a_{i-1}).$$

Let f be a bounded function on $[a, b]$ with $m \leq f(x) \leq M$ for $x \in [a, b]$.

Definition 5.1.2 Let $[a, b] \subset \mathbb{R}$, and let $f : [a, b] \to \mathbb{R}$ be a bounded function. If P is a partition of $[a, b]$, we set $m_i = \inf_{x \in [a_{i-1}, a_i]} f(x)$, and $M_i = \sup_{x \in [a_{i-1}, a_i]} f(x)$. We define the *lower sum* of f on $[a, b]$ relative to the partition P to be

$$L(f, P) = \sum_{i=1}^{k} m_i (a_i - a_{i-1}),$$

and the *upper sum* of f on $[a, b]$ relative to P to be

$$U(f, P) = \sum_{i=1}^{k} M_i (a_i - a_{i-1}).$$

Exercise 5.1.3 Let f be a bounded function on $[a, b]$ with $m \leq f(x) \leq M$ for $x \in [a, b]$. Given a partition P of $[a, b]$, show that $m(b - a) \leq L(f, P) \leq U(f, P) \leq M(b - a)$.

Example 5.1.4 Let $[a, b] = [-1, 1]$, and let $f(x) = x^2$. Consider the partition $P = \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$. Then $m_1 = m_4 = \frac{1}{4}$, $m_2 = m_3 = 0$, $M_1 = M_4 = 1$, $M_2 = M_3 = \frac{1}{4}$, so that $L(f, P) = \frac{1}{4}$, and $U(f, P) = \frac{5}{4}$. Now consider the partition $P' = \{-1, -\frac{3}{4}, -\frac{1}{2}, -\frac{1}{4}, 0, \frac{1}{3}, \frac{2}{3}, 1\}$. Then $L(f, P) = \frac{349}{864}$, and $U(f, P) = \frac{853}{864}$.

Exercise 5.1.5 Let $f(x) = \sin x$ on $[0, 2\pi]$. Compute the upper and lower sums for f relative to the partitions $P = \{\frac{k\pi}{4} \mid k = 0, 1, \ldots, 8\}$ and $P' = \{0, 1, 2, 3, 4, 5, 6, 2\pi\}$.

Exercise 5.1.6 Let $[a, b] = [0, 2]$, $f(x) = \begin{cases} -1 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } 1 \leq x \leq 2. \end{cases}$ For every partition P of $[0, 2]$, determine $U(f, P)$ and $L(f, P)$.

Definition 5.1.7 Let $P = \{a_j \mid j = 0, 1, 2, \ldots, k\}$ be a partition of $[a, b]$. A partition $P' = \{a'_j \mid j = 0, 1, 2, \ldots, k'\}$, where $k' \geq k$, is called a *refinement* of P if $P \subseteq P'$.

Exercise 5.1.8 Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Let P be a partition of $[a, b]$, and let P' be a refinement of P. Show that $L(f, P) \leq L(f, P') \leq U(f, P') \leq U(f, P)$.

Exercise 5.1.9 Let $f : [a, b] \to \mathbb{R}$ be a bounded function, and let P_1 and P_2 be any two partitions of $[a, b]$. Show that $L(f, P_1) \leq U(f, P_2)$.

We are now ready to define the lower and upper integrals of f on $[a, b]$.

Definition 5.1.10 Let $f : [a, b] \to \mathbb{R}$ be a bounded function.

1. We define the *lower integral* of f on $[a, b]$ to be
$$\underline{\int_a^b} f = \text{lub}\{L(f, P) \mid P \text{ is a partition of } [a, b]\}.$$
2. We define the *upper integral* of f on $[a, b]$ to be
$$\overline{\int_a^b} f = \text{glb}\{U(f, P) \mid P \text{ is a partition of } [a, b]\}.$$

Note that the set of lower sums is bounded above by $M(b - a)$, and the set of upper sums is bounded below by $m(b - a)$, and therefore the lower and upper integrals exist for any bounded function.

Exercise 5.1.11 Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Show that $\underline{\int_a^b} f \leq \overline{\int_a^b} f$.

Exercise 5.1.12 Let $[a, b] = [0, 2]$, and let $f(x) = \begin{cases} -1 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } 1 \leq x \leq 2. \end{cases}$ Show that $\underline{\int_0^2} f = \overline{\int_0^2} f = 0$.

Exercise 5.1.13 Let $[a, b] = [0, 1]$, and let $f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$ Find $\underline{\int_0^1} f$ and $\overline{\int_0^1} f$.

Definition 5.1.14 Let f be a bounded function on an interval $[a, b]$. We say that f is *Riemann integrable* on $[a, b]$ if $\underline{\int_a^b} f = \overline{\int_a^b} f$, and we denote this common value by $\int_a^b f$, which we call the *Riemann integral* of f on $[a, b]$.

We often refer to a function that is Riemann integrable as simply *integrable*.

Exercise 5.1.15 Let $f : [a, b] \to \mathbb{R}$ be a constant function. Show that f is Riemann integrable on $[a, b]$ and compute the value of its integral.

5.2 The Riemann Integral in One Variable: Properties

Theorem 5.2.1 Let f be a bounded function on an interval $[a, b]$. Then f is integrable on $[a, b]$ if and only if, given $\varepsilon > 0$, there exists a partition P of $[a, b]$ such that $U(f, P) - L(f, P) < \varepsilon$.

Proof. This useful criterion follows directly from the definitions of integrability and the upper and lower integrals.

Exercise 5.2.2 Let $f : [a, b] \to \mathbb{R}$ be bounded on $[a, b]$.

i. Show that f is integrable on $[a, b]$ if and only if, given $\varepsilon > 0$, there exists a $\delta > 0$ such that if P is a partition with $|P| < \delta$, then $U(f, P) - L(f, P) < \varepsilon$.

ii. Suppose f is integrable on $[a, b]$, and we have a sequence of partitions $(P_k)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}|P_k| = 0$. Show that $\int_a^b f = \lim_{k\to\infty} L(f, P_k) = \lim_{k\to\infty} U(f, P_k)$.

Theorem 5.2.3 Suppose $f : [a, b] \to \mathbb{R}$ is continuous. Then f is Riemann integrable on $[a, b]$.

Proof. Fix $\varepsilon > 0$. Since f is continuous on $[a, b]$, a compact set in $\mathbb{R}$, it follows from Exercise 3.6.10.*i* that f is uniformly continuous on $[a, b]$. So there is a $\delta > 0$ such that $|x_1 - x_2| < \delta$ implies $|f(x_1) - f(x_2)| < \frac{\varepsilon}{b-a}$. Choose a partition P such that $|P| < \delta$. Then

$$\begin{aligned} U(f, P) - L(f, P) &= \sum_{i=1}^{k}(M_i - m_i)(x_i - x_{i-1}) \\ &\leq \frac{\varepsilon}{b-a}\sum_{i=1}^{k}(x_i - x_{i-1}) \\ &= \varepsilon. \end{aligned}$$

The following exercises give the standard properties of the Riemann integral.

Exercise 5.2.4 Suppose that $f, g : [a, b] \to \mathbb{R}$ are integrable on $[a, b]$, and $\alpha \in \mathbb{R}$. Show that

i. $\int_a^b (f + g) = \int_a^b f + \int_a^b g$;

ii. $\int_a^b \alpha f = \alpha \int_a^b f$;

iii. if $c \in [a, b]$, then f is integrable on $[a, c]$ and $[c, b]$, and $\int_a^b f = \int_a^c f + \int_c^b f$;

iv. if $f(x) \leq g(x)$ for all $x \in [a, b]$, then $\int_a^b f \leq \int_a^b g$; and

v. $|f|$ is integrable, and $\left|\int_a^b f\right| \leq \int_a^b |f|$.

Exercise 5.2.5 Show that the converse to part *v* above is false by giving an example of a function f such that $|f|$ is integrable on some $[a, b]$, but f is not.

Exercise 5.2.6 Which of the statements in Exercise 5.2.4 are true for lower or upper integrals?

Example 5.2.7 Let $f(x) = x^2$ on $[-1, 1]$. Then f is integrable on $[-1, 1]$ by Theorem 5.2.3, and we compute $\int_{-1}^{1} f$ as follows. By Exercise 5.2.2.*ii*, it suffices to consider partitions whose mesh goes to zero, and we make our computation easier by choosing regular partitions. That is, for each $k \in \mathbb{N}$, we let $P_k = \{-1, -1 + \frac{1}{k}, -1 + \frac{2}{k}, \cdots, -1 + \frac{2(k-1)}{k}, 1\}$. Note that $|P_k| = \frac{1}{k}$. Furthermore, it suffices to compute the upper integral. We have

$$\begin{aligned} U(f, P_k) &= \sum_{i=1}^{2k} M_i \cdot \frac{1}{k} \\ &= 2\sum_{i=1}^{k} M_{k+i} \cdot \frac{1}{k} \\ &= 2\sum_{i=1}^{k} \left(\frac{i}{k}\right)^2 \frac{1}{k} \\ &= \frac{2}{k^3} \sum_{i=1}^{k} i^2 \\ &= \frac{2}{k^3} \frac{k(k+1)(2k+1)}{6} \end{aligned}$$

by a simple high school formula. Then

$$\int_{-1}^{1} f = \lim_{k\to\infty} \frac{2}{k^3} \frac{k(k+1)(2k+1)}{6} = \frac{2}{3}.$$

Exercise 5.2.8

i. Let $f(x) = x^3$ on $[0, 2]$. Compute $\int_0^2 f$.

ii. Let

$$f(x) = \begin{cases} \frac{1}{q} & \text{when } x = \frac{p}{q} \in \mathbb{Q} \text{ in lowest terms, } x \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that f is integrable, and compute $\int_0^1 f$.

Exercise 5.2.9 Suppose f is Riemann integrable on $[a, b]$. Take a partition P, and choose a point $a_i' \in [a_{i-1}, a_i]$. The *Riemann sum* of f with respect to P and the selection of points a_i' is $\sum_{i=1}^{k} f(a_i')(a_i - a_{i-1})$. Show that, for any sequence of partitions $(P_j)_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty}|P_j| = 0$, any set of associated Riemann sums converges to $\int_a^b f$.

Theorem 5.2.10 Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of Riemann integrable functions defined on $[a, b]$. Suppose that $(f_k)_{k\in\mathbb{N}}$ converges uniformly to a function f on $[a, b]$. Then f is Riemann integrable, and $\int_a^b f = \lim_{k\to\infty} \int_a^b f_k$.

Proof. First from Theorem 3.4.23, we know that f is bounded. Given $\varepsilon > 0$, there exists k such that $|f_k(x) - f(x)| < \frac{\varepsilon}{3(b-a)}$ for all $x \in [a, b]$. Since f_k is integrable on $[a, b]$, there exists a partition P of $[a, b]$ such that $U(f_k, P) - L(f_k, P) < \frac{\varepsilon}{3}$. Then

$$U(f, P) - L(f, P) \leq |U(f, P) - U(f_k, P)| + |U(f_k, P) - L(f_k, P)| + |L(f, P) - L(f_k, P)| < \varepsilon.$$

So f is integrable by Theorem 5.2.1, and it is clear that $\lim_{k\to\infty} \int_a^b f_k = \int_a^b f$.

5.3 The Fundamental Theorem of Calculus and Its Consequences

The next theorem is often regarded as the most important theorem in integral calculus in one variable. This seems to be based on two facts: one is that it allows some interesting applications of the integral that are important both in mathematics and other sciences; the other is that it allows explicit computation of the integrals of a few functions. These few functions seem capable of providing an indefinite set of questions on calculus exams. As we shall see, this type of application depends on finding a function F whose derivative is f. This is a truly special, limited class of elementary functions that are usually listed in the endpapers of calculus textbooks. On the other hand, we know from Theorem 5.2.3 that every continuous function is integrable, leaving us with the prospect of eternally dancing on a very small dance floor with a limited number of partners.

Theorem 5.3.1 (Fundamental Theorem of Calculus) Let f be Riemann integrable on $[a, b]$. For $x \in [a, b]$, define $F(x) = \int_a^x f$. If f is continuous at a point $x \in (a, b)$, then F is differentiable at x, and $F'(x) = f(x)$.

Proof. Fix $\varepsilon > 0$. Consider the difference quotient

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h}\left(\int_a^{x+h} f - \int_a^x f\right).$$

Assume first that h is positive. Then the difference quotient is equal to $\frac{1}{h}\int_x^{x+h} f$. Since f is continuous at x, we can choose h so small that $|f(x+t) - f(x)| < \varepsilon$ for $t \in [0, h]$. Then

$$(f(x) - \varepsilon)h < \int_x^{x+h} f < (f(x) + \varepsilon)h.$$

So

$$f(x) - \varepsilon < \frac{1}{h}\int_x^{x+h} f < f(x) + \varepsilon.$$

Since ε was arbitrary, it follows that the difference quotient converges to $f(x)$. For h negative, the argument is similar.

Corollary 5.3.2 Suppose f is continuous on $[a, b]$, and G is continuous on $[a, b]$ and differentiable on (a, b) with $G'(x) = f(x)$ for $x \in (a, b)$. Then $\int_a^b f = G(b) - G(a)$. Moreover, if F is defined as above, then $F(x) = G(x) - G(a)$.

Proof. Since $G' = F'$ on (a, b), it follows from Corollary 4.1.16 that there exists a constant C such that $F(x) - G(x) = C$. It is clear that $C = -G(a)$.

Remark 5.3.3 In the language of the preceding corollary, the function G is often referred to as a *primitive* or an *antiderivative* of f. A substantial portion of many single-variable calculus courses consists of a search for primitives.

For the next corollary, we need to introduce the traditional notation for the integral. That is, if f is integrable on $[a, b]$, we write $\int_a^b f = \int_a^b f(x)\,dx = \int_a^b f(t)\,dt$, where x, t, and any other symbol are called "dummy" variables. For the moment, these expressions should be thought of in terms of the bookkeeping they allow us to do. We will give a proper definition of the differential dx in Chapter 6.

Corollary 5.3.4 (Change of variables in $\mathbb{R}$) Let f be continuous on $[a,b]$, and let $\phi : [c,d] \to [a,b]$ be continuous on $[c,d]$ and C^1 on (c,d) such that $\phi(c) = a$, $\phi(d) = b$, and $\phi'(x) > 0$ for all $x \in (c,d)$. Then

$$\int_a^b f(u)\,du = \int_c^d f(\phi(x))\phi'(x)\,dx.$$

Remark 5.3.5 To emphasize the lack of a role that these dummy variables play, we could also write the conclusion as $\int_a^b f = \int_c^d (f \circ \phi) \cdot \phi'$.

Exercise 5.3.6 Prove this statement using the chain rule and the fundamental theorem of calculus.

Exercise 5.3.7 Prove this statement without using the fundamental theorem of calculus.

The change of variables theorem is what allows the classic integration technique known as "integration by substitution." We illustrate this with the following examples.

Example 5.3.8 Suppose we wish to evaluate $\int_c^d \sqrt{\alpha x + \beta}\,dx$ with $\alpha > 0$ and $d > c \geq -\frac{\beta}{\alpha}$. The function $g(x) = \sqrt{\alpha x + \beta}$ is not among the small number of functions that have an obvious antiderivative. However, we can write $g(x) = f(\phi(x))\phi'(x)$, where $f(u) = \frac{1}{\alpha}\sqrt{u}$, $\phi(x) = \alpha x + \beta$, and $\phi'(x) = \alpha$. The change of variables formula then tells us that

$$\begin{aligned}\int_c^d \sqrt{\alpha x + \beta}\,dx &= \int_c^d f(\phi(x))\phi'(x)\,dx \\ &= \int_{\phi(c)}^{\phi(d)} f(u)\,du, \\ &= \int_{\alpha c + \beta}^{\alpha d + \beta} \frac{1}{\alpha}\sqrt{u}\,du.\end{aligned}$$

The advantage of having used the change of variables formula is that $f(u) = \frac{1}{\alpha}\sqrt{u}$ is a function one of whose primitives everyone knows to be $G(u) = \frac{2}{3\alpha}u^{\frac{3}{2}}$. Corollary 5.3.2 now tells us that this integral evaluates to

$$G(\alpha d + \beta) - G(\alpha c + \beta) = \frac{2}{3\alpha}\left((\alpha d + \beta)^{\frac{3}{2}} - (\alpha c + \beta)^{\frac{3}{2}}\right).$$

Example 5.3.9 Suppose we wish to evaluate $\int_0^b \frac{1}{1+u^2}\,du$. The thorough reader will recall this example from Project 4.10.2.

Our goal here is to find a function $u = u(x)$ satisfying the conditions of Corollary 5.3.4 such that $\int_{u^{-1}(0)}^{u^{-1}(b)} \frac{1}{1+u(x)^2}u'(x)\,dx$ can be computed directly. The technique of trigonometric substitution from a standard single-variable calculus course suggests the substitution $u(x) = \tan x$. Then $u'(x) = \sec^2 x$, and $\frac{1}{1+u(x)^2} = \frac{1}{1+\tan^2 x} = \frac{1}{\sec^2 x}$, so the expression $\frac{u'(x)}{1+u(x)^2} = 1$. Our integral then becomes

$$\int_{\tan^{-1}(0)}^{\tan^{-1}(b)} 1\,dx = \tan^{-1}(b) - \tan^{-1}(0) = \tan^{-1}(b).$$

Exercise 5.3.10 Use implicit differentiation to show that if $\tan y = x$ then $y' = \frac{1}{1+x^2}$.

Exercise 5.3.11 Let $\alpha > 0$, and let $f : [a,b] \to \mathbb{R}$ be continuous. Show that $\alpha \int_a^b f(x)\,dx = \int_{\alpha a}^{\alpha b} f\left(\frac{x}{\alpha}\right)\,dx$

Another classic technique is that of integration by parts.

Corollary 5.3.12 (Integration by parts) Suppose f and g are C^1. Show that $\int_a^b f(x)g'(x)\,dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)\,dx$.

Proof. Use the product rule and Corollary 5.3.2.

Corollary 5.3.13 (Mean value theorem for integrals) Let f be continuous on $[a,b]$. Then there exists a point $c \in (a,b)$ such that $f(c) = \frac{1}{b-a}\int_a^b f$.

Proof. Apply the mean value theorem for derivatives to the function $F(x) = \int_a^x f$.

Remark 5.3.14 The quantity $\frac{1}{b-a}\int_a^b f$ is called the *average value of f on* $[a,b]$.

Exercise 5.3.15 Let $f : [a,b] \to \mathbb{R}$ be a bounded integrable function. For each $n \in \mathbb{N}$, we define

$$A_n = \frac{1}{n}\sum_{i=1}^{n} f\left(a + \frac{(2i-1)(b-a)}{2n}\right).$$

Show that $\lim_{n\to\infty} A_n$ exists and is equal to the average value of f. (*Hint*: What does the formula for A_n represent?)

Exercise 5.3.16 A train travels from Chicago to St. Louis in 4 hours. Show that at some point the speed of the train must exceed 60 miles per hour.

5.4 Principal Value Integrals

So far, in this chapter, we have worked with bounded, real-valued functions defined on a closed, bounded interval in $\mathbb{R}$. For many reasons, in mathematics and its applications, it is useful and even necessary to extend the definition of integrals to both unbounded functions and unbounded intervals. Here, we present some cases of this extension, and the reader can create additional cases from these. We have already used these ideas in Project 4.10.2 in the previous chapter.

First, we consider the case of a bounded, real-valued function f defined on an unbounded interval of the form $[a,\infty)$.

Definition 5.4.1 Suppose that $f : [a,\infty) \to \mathbb{R}$ is a bounded function, and suppose further that for all $b > a$, the function f is integrable on $[a,b]$. The *principal value integral*, $\int_a^\infty f$, is defined by $\lim_{b\to\infty}\int_a^b f$ if this limit exists.

Example 5.4.2 Let $f(x) = \frac{1}{x^\alpha}$ on $[1,\infty)$, where $\alpha > 0$. Then, for any $b > 1$, f is integrable on $[1,b]$, and

$$\int_1^b \frac{1}{x^\alpha}\,dx = \begin{cases} \frac{b^{1-\alpha}-1}{1-\alpha} & \text{if } \alpha \neq 1, \\ \ln b & \text{if } \alpha = 1. \end{cases}$$

If $\alpha > 1$, then the principal value integral exists and is equal to $-\frac{1}{1-\alpha}$. If $\alpha \leq 1$, the limit diverges, so the principal value integral does not exist.

Exercise 5.4.3 Decide whether the following principal value integrals exist, and if so, evaluate them.

i. $\int_0^\infty e^{-x}\,dx$

ii. $\int_0^\infty \sin x\,dx$

iii. $\int_0^\infty \frac{\sin x}{x}\,dx$

iv. $\int_0^\infty \left|\frac{\sin x}{x}\right|\,dx$

v. $\int_2^\infty \frac{1}{x \ln x}\,dx$

Remark 5.4.4 You will notice in the above exercise that the principal value integral exists in some cases because we are simply extending the length of the interval from bounded to unbounded while the function itself is nonnegative on the entire interval, whereas in other cases the principal value integral exists due to cancelation in different parts of the interval of integration.

Exercise 5.4.5 Define the principal value integral $\int_{-\infty}^b f$ including any necessary hypotheses on f.

The case of the principal value integral of a bounded real-valued function on $\mathbb{R}$ is more subtle.

Exercise 5.4.6

i. Show that $\lim_{a\to\infty}\int_{-a}^a \sin x\,dx = 0$.

ii. Show that $\lim_{a\to\infty}\int_{-a}^a \cos x\,dx$ does not exist.

iii. Show that $\lim_{a\to\infty}\int_{-a}^a \sin(x+b)\,dx$ exists if and only if $b = \pi n$ for some $n \in \mathbb{Z}$.

Since the functions we are attempting to integrate in the preceding exercise are essentially the same, that is, they are horizontal translates of each other, the values of their integrals should be the same, assuming these integrals exist.

So how do we proceed? Suppose f is a bounded function on $\mathbb{R}$, and suppose further that f is integrable on any closed interval $[a,b]$. There are two potential impediments to f being integrable on $\mathbb{R}$: the first is the behavior of f toward $+\infty$, and the second is the behavior of f toward $-\infty$. We ensure that neither of these impediments is too great by defining the principal value integral of f on $\mathbb{R}$ as follows.

Definition 5.4.7 Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a bounded function, and suppose further that f is integrable on any closed bounded interval in $\mathbb{R}$. The *principal value integral* of f on $\mathbb{R}$ is defined as $\int_{-\infty}^{\infty} f = \int_{-\infty}^0 f + \int_0^\infty f$, provided that the latter two principal value integrals exist.

Exercise 5.4.8 Show that if the principal value integral of f on $\mathbb{R}$ exists, then $\int_{-\infty}^\infty f = \lim_{a\to\infty}\int_{-a}^a f$.

Next, we consider unbounded functions on bounded intervals. We treat an important special case first, from which the general theory will follow.

Definition 5.4.9 Let $f : (a,b] \to \mathbb{R}$ be a function. Suppose further that for all $c \in (a,b)$, f is bounded and integrable on $[c,b]$. The *principal value integral* of f on $(a,b]$ is $\int_a^b f = \lim_{c\to a^+}\int_c^b f$, provided this limit exists.

Example 5.4.10 Let $f(x) = \frac{1}{x^\alpha}$ on $(0,1]$, where $\alpha > 0$. Then, for any $c \in (0,1)$, f is integrable on $[c,1]$, and

$$\int_c^1 \frac{1}{x^\alpha}\,dx = \begin{cases} \frac{1-c^{1-\alpha}}{1-\alpha} & \text{if } \alpha \neq 1, \\ -\ln c & \text{if } \alpha = 1. \end{cases}$$

If $\alpha < 1$, then the principal value integral exists and is equal to $\frac{1}{1-\alpha}$. If $\alpha \geq 1$, the limit diverges, so the principal value integral does not exist.

Exercise 5.4.11 Decide whether the following principal value integrals exist, and if so, evaluate them.

i. $\int_{-\frac{\pi}{2}}^0 \tan x\,dx$

ii. $\int_0^{\frac{\pi}{2}} \frac{1}{\sin^2 x}\,dx$

iii. $\int_0^1 \ln x\,dx$

Exercise 5.4.12 Define the principal value integral $\int_a^b f$ for a function $f : [a, b) \to \mathbb{R}$, where f is bounded and integrable on $[a, c]$ for any $c \in (a, b)$.

Exercise 5.4.13 Define the principal value integral $\int_a^b f$ for a function $f : [a, c) \cup (c, b] \to \mathbb{R}$, where f is bounded and integrable on $[a, c - \varepsilon]$ and on $[c + \eta, b]$ for any $\varepsilon, \eta > 0$.

Exercise 5.4.14 Criticize the following computation:

$$\int_{-1}^{1} \frac{1}{x^2}\, dx = \left[\frac{x^{-1}}{-1}\right]_{-1}^{1} = -1 - 1 = -2.$$

The class of principal value integrals we have discussed above are sometimes called *improper integrals.* As we have seen, there are many variations on establishing a value for an improper integral. If the procedure has a certain element of symmetry in it, then cancelation may play a role. For example, $\int_{-1}^{1} \frac{1}{x}\, dx$ has no value according to any of our previous definitions. However, if we assert that $\int_{-1}^{1} \frac{1}{x}\, dx = \lim_{\varepsilon \to 0} \left(\int_{-1}^{-\varepsilon} \frac{1}{x}\, dx + \int_{\varepsilon}^{1} \frac{1}{x}\, dx \right)$, then the value of this improper integral is 0. This idea of cancelation through symmetry plays a major role in some areas of advanced analysis.

Finally, we note that the ideas presented above can be generalized to treat integrals of unbounded functions on unbounded intervals. This idea will be used in Project 5.11.1 on the Gamma function.

We will see in the next section that there are necessary and sufficient conditions on a bounded function f for it to be integrable. However, there is little point to proving it in one variable when we will need to prove it in several variables.

5.5 The Riemann Integral in Several Variables: Definitions

In developing a theory of integration, we build off of the fundamental notion of the volume of a generalized rectangle. When we did this for a function of a single variable, we considered rectangles in $\mathbb{R}$, namely, intervals, whose volumes were their lengths. Such an interval formed the base for a rectangle in $\mathbb{R}^2$, whose height was determined by function values on this interval. The volume of this rectangle was base times height, and the integral of the function was approximated by volumes of such rectangles. We say this here merely to emphasize the distinction between volume in the domain $\mathbb{R}$ (in this case, length), and volume in $\mathbb{R}^2$ (in this case, area). We note further that the notion of volume for the base is a nonnegative quantity, while we are allowing the height to be a signed quantity, and hence the volume in $\mathbb{R}^2$ is also a signed quantity.

In trying to generalize this to functions of several variables, we therefore need to consider the volumes of generalized rectangles in $\mathbb{R}^n$, as well as the notion of a signed volume in $\mathbb{R}^{n+1}$, which will be given by the volume of the base times the signed height. We begin by defining the volume of generalized rectangles in $\mathbb{R}^n$.

Definition 5.5.1 Let $R \subset \mathbb{R}^n$ be a generalized rectangle of the form $R = I_1 \times I_2 \times \cdots \times I_n$, where I_i is a bounded interval with endpoints a_i and b_i, with $a_i \leq b_i$, for each $i = 1, 2, \ldots, n$. Then the *volume* of R in $\mathbb{R}^n$ is

$$v(R) = \prod_{i=1}^{n} (b_i - a_i).$$

For convenience, we also define $v(\varnothing) = 0$.

Definition 5.5.2 A *partition* P of a closed generalized rectangle $R \subset \mathbb{R}^n$ is a finite collection $P_1, \ldots, P_k$ of pairwise disjoint open generalized rectangles contained in R such that

$$R = \bigcup_{i=1}^{k} \overline{P_i}.$$

The *mesh* of the partition P is

$$|P| = \max_{1 \le i \le k} \operatorname{diam}(P_i).$$

Remark 5.5.3 This new definition supersedes, but is consistent with, Definition 5.1.1. Given a partition $P = \{a = a_0 < a_1 < \cdots < a_k = b\}$ of $[a, b]$ in the sense of Definition 5.1.1, we obtain a partition in the new sense consisting of the open generalized rectangles (which in this case are open intervals) (a_0, a_1), (a_1, a_2), $\ldots$, (a_{k-1}, a_k).

Exercise 5.5.4 Let $R \subset \mathbb{R}^n$ be a generalized rectangle. Show that if $v(R) = 0$, then there do not exist partitions of R.

Exercise 5.5.5 Let $R \subset \mathbb{R}^n$ be a generalized rectangle with $v(R) > 0$. Let P_i and P_j be two subrectangles from a partition of R. Show that $v(\overline{P_i} \cap \overline{P_j}) = 0$.

Exercise 5.5.6 Let $R \subset \mathbb{R}^n$ be a generalized rectangle with $v(R) > 0$, and let $R_1, R_2, \ldots, R_N$ be a finite collection of subrectangles of R, each with nonzero volume. Show that there exists a partition P of R that satisfies the following property. For every i and every subrectangle S in P, either S is a subrectangle of R_i or $S \cap R_i = \varnothing$.

Let R be a generalized rectangle in $\mathbb{R}^n$. Let $f : R \to \mathbb{R}$ be a bounded function. We define the integral of f over the generalized rectangle R by the same method we used in the one-variable case.

Definition 5.5.7 Let $R \subset \mathbb{R}^n$ be a generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function. If P is a partition of R, we set $m_i = \inf_{x \in \overline{P_i}} f(x)$, and $M_i = \sup_{x \in \overline{P_i}} f(x)$. We define the *upper sum* of f on R relative to the partition P to be

$$U(f, P) = \sum_{i=1}^{k} M_i \cdot v(P_i),$$

and the *lower sum* of f on R relative to P to be

$$L(f, P) = \sum_{i=1}^{k} m_i \cdot v(P_i).$$

Exercise 5.5.8 Let f be a bounded function on R with $m \le f(x) \le M$ for all $x \in R$. Given a partition P of R, show that $m \cdot v(R) \le L(f, P) \le U(f, P) \le M \cdot v(R)$.

Exercise 5.5.9 Let $R \subset \mathbb{R}^2$ be the generalized rectangle $R = [0, 1] \times [0, 2]$. Let $P = \{P_1 = (0, \frac{1}{2}) \times (0, 1), P_2 = (0, \frac{1}{2}) \times (1, 2), P_3 = (\frac{1}{2}, 1) \times (0, \frac{2}{3}), P_4 = (\frac{1}{2}, 1) \times (\frac{2}{3}, \frac{4}{3}), P_5 = (\frac{1}{2}, 1) \times (\frac{4}{3}, 2)\}$, and let $Q = \{Q_1 = (0, \frac{1}{2}) \times (0, \frac{2}{3}), Q_2 = (0, \frac{1}{2}) \times (\frac{2}{3}, \frac{4}{3}), Q_3 = (0, \frac{1}{2}) \times (\frac{4}{3}, 2), Q_4 = (\frac{1}{2}, 1) \times (0, \frac{2}{3}), Q_5 = (\frac{1}{2}, 1) \times (\frac{2}{3}, \frac{4}{3}), Q_6 = (\frac{1}{2}, 1) \times (\frac{4}{3}, 2)\}$. Find the mesh of these partitions. Let $f : R \to \mathbb{R}$, $f(x, y) = x^2 - y^2$. Compute the upper and lower sums for f relative to the partitions P and Q.

Definition 5.5.10 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and let $P = \{P_j \mid j = 1, 2, \ldots, k\}$ be a partition of R. A partition $P' = \{P'_j \mid j = 1, 2, \ldots, k'\}$ of R, where $k' \ge k$, is called a *refinement* of P if every rectangle in P' is contained in a rectangle in P.

Exercise 5.5.11 Let $f : R \to \mathbb{R}$ be a bounded function. Let P be a partition of R, and let P' a refinement of P. Show that $L(f, P) \le L(f, P') \le U(f, P') \le U(f, P)$.

Exercise 5.5.12 Find a common refinement of the partitions P and Q given in Exercise 5.5.9 and verify the inequalities in Exercise 5.5.11.

Exercise 5.5.13 Let $f : R \to \mathbb{R}$ be a bounded function, and let P and Q be any two partitions of R. Show that $L(f, P) \le U(f, Q)$.

We are now ready to define the lower and upper integrals of f on R.

Definition 5.5.14 Let $R \subset \mathbb{R}^n$ be closed generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function.

1. We define the *lower integral* of f on R to be
$$\underline{\int_R} f = \text{lub}\{L(f, P) \mid P \text{ is a partition of } R\}.$$
2. We define the *upper integral* of f on R to be
$$\overline{\int_R} f = \text{glb}\{U(f, P) \mid P \text{ is a partition of } R\}.$$

Note that the set of lower sums is bounded above by $M \cdot v(R)$, and the set of upper sums is bounded below by $m \cdot v(R)$, and therefore the lower and upper integrals exist for any such bounded function.

Exercise 5.5.15 For any bounded function $f : R \to \mathbb{R}$, we have $\underline{\int_R} f \leq \overline{\int_R} f$.

Exercise 5.5.16 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and define $f : R \to \mathbb{R}$ by
$$f(x) = \begin{cases} 0 & \text{if } x = (x_1, x_2, \cdots, x_n), \text{ where } x_i \in \mathbb{Q} \text{ for all } i, \\ 1 & \text{otherwise.} \end{cases}$$
Find $\underline{\int_R} f$ and $\overline{\int_R} f$.

In studying the Riemann integral in several variables, it is often useful to work with partitions that have a regularity condition.

Definition 5.5.17 Let $R = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$ be a closed generalized rectangle in $\mathbb{R}^n$. For each i, $1 \leq i \leq n$, we take a partition $P_i = \{a_i = a_{i,0}, a_{i,1}, \ldots, a_{i,k_i} = b_i\}$, where $a_{i,0} < a_{i,1} < \cdots < a_{i,k_i}$. Note that P_i is a partition of the closed interval $[a, b]$ in the sense of Definition 5.1.1. With this, we define a *regular partition* P of R by $P = P_1 \times P_2 \times \cdots \times P_n$, consisting of the $k_1 \cdot k_2 \cdot \cdots \cdot k_n$ subrectangles of the form $(a_{1,j_1}, a_{1,j_1+1}) \times (a_{2,j_2}, a_{2,j_2+1}) \times \cdots \times (a_{n,j_n}, a_{n,j_n+1})$ with $0 \leq j_i < k_i$ for $1 \leq i \leq n$.

Exercise 5.5.18 Show that
$$\underline{\int_R} f = \text{lub}\{L(f, P) \mid P \text{ is a regular partition of } R\},$$
and
$$\overline{\int_R} f = \text{glb}\{U(f, P) \mid P \text{ is a regular partition of } R\}.$$

Definition 5.5.19 Let f be a bounded function on a closed generalized rectangle $R \subset \mathbb{R}^n$. We say that f is *integrable* on R if $\underline{\int_R} f = \overline{\int_R} f$, and we denote this common value by $\int_R f$, which we call the *Riemann integral* of f on R.

5.6 The Riemann Integral in Several Variables: Properties

The following 3 theorems are perfect analogues of the corresponding results in one variable. The careful reader should verify that the proofs carry over from before.

Theorem 5.6.1 Let f be a bounded function on a closed generalized rectangle $R \subset \mathbb{R}^n$. Then f is integrable on R if and only if, given $\varepsilon > 0$, there exists a partition P of R such that $U(f, P) - L(f, P) < \varepsilon$.

Theorem 5.6.2 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle. Suppose $f : R \to \mathbb{R}$ is continuous. Then f is Riemann integrable on R.

Theorem 5.6.3 Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of Riemann integrable functions defined on a closed generalized rectangle $R \subset \mathbb{R}^n$. Suppose that $(f_k)_{k\in\mathbb{N}}$ converges uniformly to a function f on R. Then f is Riemann integrable on R, and $\int_R f = \lim_{k\to\infty} \int_R f_k$.

The results of many of the exercises from Section 5.2 carry over to several variables as well. We summarize these results in the statements below.

Proposition 5.6.4 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function.

1. f is integrable on R if and only if, given $\varepsilon > 0$, there exists a $\delta > 0$ such that if P is a partition of R with $|P| < \delta$, then $U(f, P) - L(f, P) < \varepsilon$.

2. Suppose f is integrable on R, and we have a sequence of partitions $(P_k)_{k\in\mathbb{N}}$ of R such that $\lim_{k\to\infty}|P_k| = 0$. Then $\int_R f = \lim_{k\to\infty} L(f, P_k) = \lim_{k\to\infty} U(f, P_k)$.

Proposition 5.6.5 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, let $f, g : R \to \mathbb{R}$ be integrable functions, and let $\alpha \in \mathbb{R}$. Then

1. $\int_R (f + g) = \int_R f + \int_R g$;

2. $\int_R \alpha f = \alpha \int_R f$;

3. if $f(x) \leq g(x)$ for all $x \in R$, then $\int_R f \leq \int_R g$; and

4. $|f|$ is integrable, and $\left|\int_R f\right| \leq \int_R |f|$.

Exercise 5.6.6 Check that your one-variable proofs of the above statements from Exercises 5.2.2 and 5.2.4 carry over to several variables.

Note that the extension of the result of Exercise 5.2.4.*iii* can be carried out in a number of ways. We leave these to the imagination of the reader.

5.7 The Computation of Integrals in Several Variables and Fubini's Theorem

In computing integrals in several variables, we begin by introducing a method that can use the fundamental theorem of calculus in an effective way. The basic idea is to iterate integrals one variable at a time and combine the results to obtain the integral in several variables. This is one place where our use of the traditional notation and its dummy variables will be particularly advantageous. We illustrate the basic principle with the following simple example.

Example 5.7.1 Let $R = [1, 3] \times [2, 4] \subset \mathbb{R}^2$, and let $f : R \to \mathbb{R}$ be given by $f(x, y) = x^2y^3$. We proceed in the same manner that we did with partial derivatives. That is, we hold one variable constant, and integrate with respect to the other variable. Symbolically, we write $\int_R f = \int_1^3 \int_2^4 x^2y^3 \, dy \, dx$, by which we mean $\int_1^3 \left(\int_2^4 x^2y^3 \, dy\right) dx$. This indicates that we first compute the integral with respect to y (the inner integral) on the interval $[2, 4]$ by holding x constant and applying the fundamental theorem. This inner integral will depend on x and should be thought of as a function of x, and we can compute the integral of this function (the outer integral) over the interval $[1, 3]$ using the fundamental theorem a second time.

We now compute

$$\int_1^3 \int_2^4 x^2y^3\,dy\,dx = \int_1^3 \left[\frac{1}{4}x^2y^4\right]_{y=2}^{y=4} dx = \int_1^3 60x^2\,dx = 520.$$

Note that, in computing the inner integral, if we consider $g(y) = f(x, y) = x^2y^3$ for fixed x, then $G(y) = \frac{1}{4}x^2y^4$ has derivative $G'(y) = g(y)$, which is what allows us to apply the fundamental theorem the way we did.

Exercise 5.7.2 Show that $\int_2^4 \int_1^3 x^2y^3\,dx\,dy = 520$.

Exercise 5.7.3 Convince yourself that you do not want to show that $\int_R x^2y^3 = 520$ using the definition of the integral.

Did we just get lucky? Or does the function $f(x, y) = x^2y^3$ have properties that allow us to compute integrals by iterating one variable at a time? Observe that we knew that $\int_R x^2y^3$ existed because $f(x, y) = x^2y^3$ is continuous on R.

Example 5.7.4 On $R = [0, 1] \times [0, 1]$, we define

$$f(x, y) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 2y & \text{if } x \text{ is irrational.} \end{cases}$$

If we attempt to iterate integrals as in the preceding example, we get

$$\int_0^1 \int_0^1 f(x, y)\,dy\,dx = \int_0^1 1\,dx = 1,$$

because of the following separate computations: if x is irrational, $\int_0^1 2y\,dy = \left[y^2\right]_0^1 = 1$, and if x is rational, $\int_0^1 1\,dy = [y]_0^1 = 1$. On the other hand, if we attempt to iterate our integrals in the other order, $\int_0^1 \int_0^1 f(x, y)\,dx\,dy$, we discover that unless $y = \frac{1}{2}$, the function of one variable

$$g(x) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 2y & \text{if } x \text{ is irrational,} \end{cases}$$

is not integrable on $[0, 1]$, and thus the inner integral does not exist.

Exercise 5.7.5 With R and f as in the above example, show that $\underline{\int_R f} = \frac{3}{4}$, and $\overline{\int_R f} = \frac{5}{4}$.

Clearly, we need to put some restrictions on f before applying this iteration technique. It turns out that the integrability of f is sufficient. Now we can state an actual theorem that says when it is possible to compute the integral of a function on a rectangle in $\mathbb{R}^n$ by iteration. We state the following with stronger hypotheses than necessary because the routine application of this theorem is to continuous functions, where these stronger hypotheses are met. Exercise 5.7.8 will discuss the weaker hypotheses for which the result also holds.

It will be notationally convenient to reintroduce our dx notation for evaluating Riemann integrals in $\mathbb{R}^n$. That is, if f is a Riemann integrable function on a closed generalized rectangle $R \subset \mathbb{R}^n$, we write $\int_R f = \int_R f(x)\,dx$.

Theorem 5.7.6 (Fubini's Theorem) Let $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ be closed generalized rectangles, and let $f : R_1 \times R_2 \to \mathbb{R}$ be an integrable function such that, for any $x \in R_1$, the function $g_x : R_2 \to \mathbb{R}$ defined by $g_x(y) = f(x, y)$ is integrable. Then, the function $G : R_1 \to \mathbb{R}$ defined by $G(x) = \int_{R_2} g_x(y)\,dy$ is integrable, and

$$\int_{R_1 \times R_2} f = \int_{R_1} G(x)\,dx = \int_{R_1} \left(\int_{R_2} g_x(y)\,dy \right) dx = \int_{R_1} \left(\int_{R_2} f(x, y)\,dy \right) dx.$$

Proof. We make use of the result of Exercise 5.5.18 on the sufficiency of regular partitions. Let P_1 be a regular partition of R_1, and let P_2 be a regular partition of R_2. We denote by P the corresponding regular partition of $R_1 \times R_2$. Given a generalized rectangle $S \in P$, we can write $S = S_1 \times S_2$, where $S_1 \in P_1$, and $S_2 \in P_2$. Then

$$\begin{aligned} L(f, P) &= \sum_{S \in P} m_S(f) v(S) \\ &= \sum_{S_1 \in P_1} \left(\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) v(S_2) \right) v(S_1). \end{aligned}$$

Now, for any $x \in S_1$, we have $m_{S_1 \times S_2}(f) \leq m_{\{x\} \times S_2}(f) = m_{S_2}(g_x)$, and hence

$$\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) v(S_2) \leq \sum_{S_2 \in P_2} m_{S_2}(g_x) v(S_2) \leq \int_{R_2} g_x(y)\, dy = G(x).$$

Since x was arbitrary, $\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) v(S_2) \leq m_{S_1}(G)$. This implies

$$\begin{aligned} L(f, P) &= \sum_{S_1 \in P_1} \left(\sum_{S_2 \in P_2} m_{S_1 \times S_2}(f) v(S_2) \right) v(S_1) \\ &\leq \sum_{S_1 \in P_1} m_{S_1}(G) v(S_1) \\ &= L(G, P_1). \end{aligned}$$

A similar statement for upper sums gives us the inequality

$$L(f, P) \leq L(G, P_1) \leq U(G, P_1) \leq U(f, P).$$

The assumption that f is integrable implies that the outer terms of this inequality can be made arbitrarily close by an appropriate choice of P, and hence G is integrable on R_1 by Theorem 5.6.1, with $\int_{R_1} G = \int_{R_1 \times R_2} f$.

Exercise 5.7.7 Let $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ be closed generalized rectangles, and let $f : R_1 \times R_2 \to \mathbb{R}$ be an integrable function such that, for any $y \in R_2$, the function $h_y : R_1 \to \mathbb{R}$ defined by $h_y(x) = f(x, y)$ is integrable. Show that the function $H : R_2 \to \mathbb{R}$ defined by $H(y) = \int_{R_1} h_y(x)\, dx$ is integrable, and

$$\int_{R_1 \times R_2} f = \int_{R_2} H(y)\, dy = \int_{R_2} \left(\int_{R_1} h_y(x)\, dx \right) dy = \int_{R_2} \left(\int_{R_1} f(x, y)\, dx \right) dy.$$

Exercise 5.7.8 Show that we can weaken the hypotheses of Fubini's Theorem as follows and obtain an analogous result.

Let $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ be closed generalized rectangles, and let $f : R_1 \times R_2 \to \mathbb{R}$ be an integrable function. For each $x \in R_1$, define the function $g_x : R_2 \to \mathbb{R}$ by $g_x(y) = f(x, y)$. Then, the functions $G_L : R_1 \to \mathbb{R}$ defined by $G_L(x) = \underline{\int_{R_2}} g_x(y)\, dy$ and $G_R : R_1 \to \mathbb{R}$ defined by $G_R(x) = \overline{\int_{R_2}} g_x(y)\, dy$ are integrable, and

$$\int_{R_1} G_L(x)\, dx = \int_{R_1 \times R_2} f = \int_{R_1} G_R(x)\, dx.$$

Exercise 5.7.9 Let $R_1 \subset \mathbb{R}^n$ and $R_2 \subset \mathbb{R}^m$ be closed generalized rectangles, and let $R = R_1 \times R_2 \subset \mathbb{R}^{n+m}$. Let $g : R_1 \to \mathbb{R}$ and $h : R_2 \to \mathbb{R}$ be integrable functions. Show that the function $f : R \to \mathbb{R}$ given by $f(x, y) = g(x)h(y)$ is integrable on R, and

$$\int_R f = \left(\int_{R_1} g \right) \cdot \left(\int_{R_2} h \right).$$

5.8 Sets of Measure Zero and the Riemann Integrability Criterion

We are led naturally to the following question. What are necessary and sufficient conditions for a function f to be integrable? The following theorem, sometimes called the Riemann Integrability Criterion, gives a useful characterization of integrable functions in terms of their continuity.

Theorem 5.8.1 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function. Then f is Riemann integrable on R if and only if f is continuous except on a set of measure zero.

Obviously, we need to explain what it means for a set to have measure zero.

Definition 5.8.2 Let A be a subset of $\mathbb{R}^n$. We say that A has *measure zero* if for every $\varepsilon > 0$, there exists a countable collection $\{R_k\}_{k\in\mathbb{N}}$ of open generalized rectangles in $\mathbb{R}^n$ such that

$$A \subset \bigcup_{k=1}^{\infty} R_k,$$

and

$$\sum_{k=1}^{\infty} v(R_k) < \varepsilon.$$

Exercise 5.8.3

i. Show that the empty set has measure zero.

ii. Show that any countable set has measure zero.

iii. Show that any subset of a set of measure zero has measure zero.

iv. Show that a countable union of sets of measure zero has measure zero.

Exercise 5.8.4 Show that the Cantor Set (see Exercise 1.6.37) has measure zero.

Exercise 5.8.5

i. Show that a generalized rectangle R with $v(R) = 0$ has measure zero.

ii. Show that a generalized rectangle R with $v(R) > 0$ does not have measure zero.

Before returning to the proof of the Riemann Integrability Criterion, we need the idea of the oscillation of a function at a point, which is a quantitative measure of how badly discontinuous a function is at that point.

Definition 5.8.6 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function. Fix $x \in R$. Given any $\delta > 0$, define $M(f, x, \delta) = \sup_{y\in B_\delta(x)\cap R}\{f(y)\}$, and $m(f, x, \delta) = \inf_{y\in B_\delta(x)\cap R}\{f(y)\}$. Then the *oscillation* of f at x is defined to be $o(f, x) = \lim_{\delta\to 0}[M(f, x, \delta) - m(f, x, \delta)]$.

Exercise 5.8.7 Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle, and let $f : R \to \mathbb{R}$ be a bounded function. Show that f is continuous at a point $x \in R$ if and only if $o(f, x) = 0$.

Exercise 5.8.8 Compute the oscillations of the following functions at the point $x = 0$.

i.

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

ii.

$$f(x) = \begin{cases} \sin\frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

We are now ready to prove the Riemann Integrability Criterion.

Proof. (of Theorem 5.8.1) Suppose that f is continuous except on a set $B \subset R$ of measure zero, and $|f| \leq M$ on R. Let $\varepsilon > 0$. We cover B with a countable collection $\{U_i\}_{i \in \mathbb{N}}$ of open generalized rectangles in $\mathbb{R}^n$ such that $\sum_{i=1}^{\infty} v(U_i) < \varepsilon$. For each point $x \in R \setminus B$, we can find an open generalized rectangle $V_x \subset \mathbb{R}^n$ containing x such that $\sup_{y \in V_x} f(y) - \inf_{y \in V_x} f(y) < \varepsilon$. The union of these two sets of open generalized rectangles covers the closed generalized rectangle R, which is compact. Hence, we can extract a finite subcovering that covers R. Because this is a cover by a finite collection of generalized rectangles, we can choose a partition P of R such that each generalized rectangle of P is a subset of the closure of one of these open generalized rectangles by Exercise 5.5.6. Then $U(f,P) - L(f,P) < 2M\varepsilon + v(R)\varepsilon$, where the first term accounts for the subrectangles contained in the U_is, and the second term accounts for those contained in the V_xs. Thus, by Theorem 5.6.1, f is integrable on R.

Now suppose f is Riemann integrable on R. Again, we let $B \subset R$ be the set of discontinuities of f. We note that $B = \bigcup_{k=1}^{\infty} B_{1/k}$, where $B_{1/k} = \left\{x \in R \mid o(f,x) \geq \frac{1}{k}\right\}$. Since B is a countable union of sets of this form, it suffices to show that each of these sets has measure zero by 5.8.3.

Fix $k \in \mathbb{N}$, and choose $\varepsilon > 0$. By Theorem 5.6.1, we can find a partition P of R such that $U(f,P) - L(f,P) < \varepsilon$. Note that the interiors of the subrectangles P_i of P do not cover R, and may not cover $B_{1/k}$. However, the boundaries of the P_is form a set of measure zero, so we may cover these with a countable collection $\{U_j\}_{j \in \mathbb{N}}$ of open generalized rectangles in $\mathbb{R}^n$ with $\sum_{j=1}^{\infty} v(U_j) < \varepsilon$. Let S be the subcollection of P consisting of those P_is such that $P_i \cap B_{1/k} \neq \varnothing$. The interiors of the elements of S along with the U_is cover $B_{1/k}$. The total volume of the generalized rectangles in the first collection is

$$\begin{aligned} \sum_{P_i \in S} v(P_i) &\leq \sum_{P_i \in S} k(M_{P_i}(f) - m_{P_i}(f))v(P_i) \\ &\leq k \sum_{P_i \in P} (M_{P_i}(f) - m_{P_i}(f))v(P_i) \\ &< k\varepsilon, \end{aligned}$$

and the total volume of the generalized rectangles in the second collection is, by definition, less than ε. So we have covered $B_{1/k}$ by a countable collection of generalized rectangles with total volume at most $(k+1)\varepsilon$, and thus $B_{1/k}$ has measure zero.

5.9 Integration over Bounded Regions in $\mathbb{R}^n$

Let Ω be a bounded region in $\mathbb{R}^n$, and let $f : \Omega \to \mathbb{R}$ be a bounded function. We wish to define $\int_\Omega f$, but we will need certain restrictions, which are necessary in the Riemann theory of integration. The first is a consideration of the boundary of Ω. It is necessary that this boundary be "nice," so that we can essentially disregard it.

Definition 5.9.1 Let $A \subset \mathbb{R}^n$. We say that A has *content zero* if, given $\varepsilon > 0$, there is a finite collection of open generalized rectangles $R_1, R_2, \ldots, R_k$ such that $A \subset \bigcup_{i=1}^{k} R_i$, and $\sum_{i=1}^{k} v(R_i) < \varepsilon$.

Exercise 5.9.2

1. If $A \subset \mathbb{R}^n$ has content zero, then A has measure zero.

2. If $A \subset \mathbb{R}^n$ has measure zero, and A is a compact set, then A has content zero. Thus, content zero and measure zero are equivalent for compact subsets of $\mathbb{R}^n$, and in particular for the boundaries of bounded subsets of $\mathbb{R}^n$.

We assume that $\Omega \subset \mathbb{R}^n$ is a bounded region, and the "nice" property referred to above is that the boundary of Ω has content zero. By Exercise 3.3.33, we know that the boundary of Ω is closed and hence compact, so by the exercise above, it is enough that the boundary have measure zero. We choose a closed generalized rectangle R that contains $\overline{\Omega}$. The function f, which was originally defined as a function $\Omega \to \mathbb{R}$, can be extended to a bounded function $\widetilde{f} : R \to \mathbb{R}$ by setting

$$\widetilde{f}(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ 0 & \text{if } x \in R \setminus \Omega. \end{cases}$$

Now observe that, according to Theorem 5.8.1, $\widetilde{f}$ is Riemann integrable on R if and only if $\widetilde{f}$ is continuous on R except on a set of measure zero.

Where is $\widetilde{f}$ discontinuous? There are two potential types of points. On the one hand, if f is discontinuous at a point in Ω, then $\widetilde{f}$ will be discontinuous there too. On the other hand, it is possible for $\widetilde{f}$ to be discontinuous on the boundary of Ω, even if f is nicely behaved near the boundary. Thus, if the union of the boundary of Ω and the set of discontinuities of f has measure zero, then $\widetilde{f}$ is integrable on R. For example, let $\Omega = B_1(0) \subset \mathbb{R}^n$ be the open unit ball, and let $f : \Omega \to \mathbb{R}$ be the constant function equal to 1 on Ω. Now let $R = [-1,1] \times [-1,1] \times \cdots \times [-1,1]$. Then, $\widetilde{f}$ defined as above is discontinuous on $\partial\Omega = S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$.

Definition 5.9.3 Let $\Omega \subset \mathbb{R}^n$ be a bounded region whose boundary has content zero, and let $f : \Omega \to \mathbb{R}$ be a bounded function that is continuous on Ω except on a set of measure zero. Let $R \subset \mathbb{R}^n$ be a closed generalized rectangle containing Ω. Let

$$\widetilde{f}(x) = \begin{cases} f(x) & \text{if } x \in \Omega, \\ 0 & \text{if } x \in R \setminus \Omega. \end{cases}$$

Then we define $\int_\Omega f = \int_R \widetilde{f}$.

Exercise 5.9.4 Show that the value of the integral is independent of the choice of the closed generalized rectangle R.

Exercise 5.9.5 Let $\Omega \subset \mathbb{R}^n$ be a bounded region whose boundary has content zero, and let $f : \Omega \to \mathbb{R}$ be a bounded function that is continuous on Ω except on a set of measure zero. Let $\Omega_1, \Omega_2, \ldots, \Omega_N$ be subsets of Ω such that $\Omega = \bigcup_{i=1}^N \overline{\Omega_i}$, the sets $\overline{\Omega_i} \cap \overline{\Omega_j}$ have content zero if $i \neq j$, and $\partial\Omega_i$ has content zero for each $i = 1, 2, \ldots, N$. Show that $\int_\Omega f = \sum_{i=1}^N \int_{\Omega_i} f$.

Exercise 5.9.6 Let Ω be the unit disk in $\mathbb{R}^2$, and let $R = [-1,1] \times [-1,1]$. Let $f : \Omega \to \mathbb{R}$ be the constant function 1, and let $\widetilde{f} : R \to \mathbb{R}$ be defined as above. For each $n \in \mathbb{N}$, let P_n be the partition of R into n^2 congruent squares each of volume $\frac{1}{n^2}$. Compute $\int_\Omega f$ from the definition using these partitions. (*Hint:* Use Gauss's theorem on the number of lattice points inside a circle of radius $\sqrt{N}$.)

Such a computation from first principles is perhaps not the most productive approach to computing integrals in several variables. We reexamine this same example using the tool of Fubini's theorem.

Example 5.9.7 Let Ω be the closed unit disk in $\mathbb{R}^2$, and let $f : \Omega \to \mathbb{R}$ the constant function 1. Again, it is convenient to choose $R = [-1,1] \times [-1,1]$. Since f is continuous on Ω and $\partial\Omega$ has content zero, we know that $\int_\Omega f$ exists, and Fubini's theorem tells us that

$$\begin{aligned} \int_\Omega f &= \int_R \widetilde{f} \\ &= \int_{-1}^1 \int_{-1}^1 \widetilde{f}(x,y)\, dx\, dy. \end{aligned}$$

For each fixed value of y, we compute the inner integral $\int_{-1}^{1} \widetilde{f}(x,y)\,dx$. That is, we are integrating the function $\widetilde{f}$ considered as a function of the single variable x along a line segment parallel to the x-axis. For a fixed value of y, we have

$$\widetilde{f}(x,y) = \begin{cases} 1 & \text{if } -\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $\int_{-1}^{1} \widetilde{f}(x,y)\,dx = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} 1\,dx = 2\sqrt{1-y^2}$. We can now use this formula to evaluate the outer integral, and get, by a standard trigonometric substitution,

$$\int_{-1}^{1} 2\sqrt{1-y^2}\,dy = \pi.$$

The real point of this second example is that what we have effectively done is to parametrize the boundary of Ω, and to determine the limits of integration in the inner integrals of a Fubini-type computation. In this particular example, we were fortunate in that the domain of integration was a convex set, so that we could parametrize the boundary using numerical limits of integration in one direction and simple functional expressions in the other.

Exercise 5.9.8 Let Ω be the triangle in $\mathbb{R}^2$ with vertices, $(0,0)$, $(0,1)$, and $(1,1)$, and let $f : \Omega \to \mathbb{R}$ be given by $f(x,y) = \sin y^2$. Let $R = [0,1] \times [0,1]$, and define $\widetilde{f} : R \to \mathbb{R}$ as above.

i. Show that $\partial\Omega$ has measure zero.

ii. Compute $\int_\Omega f = \int_R \widetilde{f}$ using Fubini's Theorem.

iii. Observe that, though the integral exists and Fubini's theorem allows us to iterate the integrals in either order, the order of integration is nonetheless relevant to obtaining a successful computation.

The ordinary treatment of multiple integration in advanced calculus books focuses principally on the determination of the boundaries of integration for various regions in the plane and in three-dimensional space. In the following set of exercises, this will be exemplified in computing multiple integrals in closed form. While this is an important aspect of studying multiple integrals, especially in computing volumes of regions in higher-dimensional space, the deeper aspects of analysis can be addressed much more efficiently through the use of the Lebesgue integral.

With this in mind, it is nonetheless important to recognize that in our study of "vector calculus" in the next chapter, we will rely almost totally on the Riemann integral.

Definition 5.9.9 Let Ω be a bounded region in $\mathbb{R}^n$ with $\partial\Omega$ having content zero. The *n-volume* of Ω is $v_n(\Omega) = \int_\Omega 1$.

Exercise 5.9.10 Let Ω and Ω' be bounded regions in $\mathbb{R}^n$ with $\partial\Omega$ and $\partial\Omega'$ having content zero. Show that if $\Omega \subseteq \Omega'$, then $v_n(\Omega) \leq v_n(\Omega')$.

Exercise 5.9.11 Show that $v_n(\Omega) = v_n(T(\Omega))$, where T is a translation on $\mathbb{R}^n$.

Exercise 5.9.12 Show that the n-volume of a generalized rectangle in $\mathbb{R}^n$ is the product of its side lengths.

Exercise 5.9.13 Find the 2-volume of a right triangle whose legs are parallel to the coordinate axes.

Exercise 5.9.14 Find the 2-volume of the parallelogram in $\mathbb{R}^2$ with vertices $(0,0)$, (a,b), $(a+c, b+d)$, and (c,d).

Exercise 5.9.15 In this exercise, we will show that the n-volume of the image of a generalized rectangle $R \subset \mathbb{R}^n$ under an invertible linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ is equal to $|\det T| \cdot v_n(R)$.

i. Show that it suffices to prove this result when T is an elementary linear transformation. (See Exercise 2.3.12.)

ii. Use the previous set of exercises, together with Fubini's theorem, to prove the result when T is an elementary linear transformation.

Exercise 5.9.16 Let P be a plane in $\mathbb{R}^3$ parallel to the xy-plane. Let Ω be a closed, bounded set in the xy-plane with 2-volume B. Pick a point Q in P and construct a pyramid by joining each point in Ω to Q with a straight line segment. Find the 3-volume of this pyramid.

Exercise 5.9.17 Find the volume of the generalized tetrahedron in $\mathbb{R}^n$ bounded by the coordinate hyperplanes and the hyperplane $x_1 + x_2 + \cdots + x_n = 1$.

5.10 Change of Variables

We have illustrated the change of variables theorem in one variable in Corollary 5.3.4. The situation for several variables is considerably more complicated and involves not only the properties of differentiation in $\mathbb{R}^n$ but also the change of volume under nonsingular linear transformations. In the one-variable case, the image of a closed interval under a monotonic C^1 function is very well understood and easily parametrized. In several variables, the notion of monotonic is meaningless, and the difficulties connected to determining the volume of the image of a region (even a rectangle) under a C^1 map are already illustrated in Lemma 4.9.19.

The main result of this section is the following.

Theorem 5.10.1 (Change of Variables) Let $V \subset \mathbb{R}^n$ be a bounded open set, and let $U \subseteq \mathbb{R}^n$ be an open set such that $\overline{V} \subset U$. Let $\phi : U \to \mathbb{R}^n$ be C^1 on U and one-to-one on V, with $D\phi(x)$ invertible on V. Suppose that ∂V and $\partial\phi(V)$ have content zero. Then a bounded real-valued function f is Riemann integrable on $\phi(V)$ if and only if $f \circ \phi$ is Riemann integrable on V, and in this case,

$$\int_{\phi(V)} f(y)\, dy = \int_V (f \circ \phi)(x) |\det D\phi(x)|\, dx.$$

Our first step is to prove the change of variables theorem when V is a rectangle. The first step to doing that is to study the change of volume of a rectangle under a C^1 function.

Lemma 5.10.2 Let $R \subset \mathbb{R}^n$ be a generalized rectangle, let $U \subseteq \mathbb{R}^n$ be an open set containing $\overline{R}$, and let $\phi : U \to \mathbb{R}^n$ be C^1 on U, and one-to-one on R, with $D\phi(x)$ invertible on R. Then $v_n(\phi(R)) = \int_R |\det D\phi(x)|\, dx$.

Proof. It is easy to see that the rectangle R can be partitioned into N congruent subrectangles $R_1, R_2, \ldots, R_N$ that are similar to R, with centers $y_1, y_2, \ldots, y_N$, respectively. Recalling the notation of Lemma 4.9.19, we let $0 < \varepsilon < 1$ and define $R_i' = (1-\varepsilon)R_i$, and $R_i'' = (1+\varepsilon)R_i$, for $1 \le i \le N$.

By Lemma 4.9.19, if N is large enough, then

$$T^{y_i}(R_i') \subseteq \phi(R_i) \subseteq T^{y_i}(R_i'')$$

for all i, where T^{y_i} is defined as in the lemma. Hence

$$v_n(T^{y_i}(R_i')) \le v_n(\phi(R_i)) \le v_n(T^{y_i}(R_i'')).$$

By Exercises 5.9.11 and 5.9.15,

$$(1-\varepsilon)^n J(y_i) v_n(R_i) \le v_n(\phi(R_i)) \le (1+\varepsilon)^n J(y_i) v_n(R_i),$$

where $J(x) = |\det D\phi(x)|$.

Exercise 5.10.3

i. Show that $\phi(\partial R_i)$ has content 0. Hint: Use Lemma 4.9.19 and the above dissection technique.

ii. Show that the boundary of a generalized rectangle is the union of finitely many closed generalized rectangles with volume zero.

iii. Show that $\partial\phi(R_i) = \phi(\partial R_i)$. Conclude that $\partial\phi(R_i)$ has content zero, and that $\sum_{i=1}^N v_n(\phi(R_i)) = v_n(\phi(R))$.

Hence, summing over i, we have

$$(1-\varepsilon)^n \sum_{i=1}^N J(y_i)v_n(R_i) \le v_n(\phi(R)) \le (1+\varepsilon)^n \sum_{i=1}^N J(y_i)v_n(R_i).$$

Letting the mesh of our partition tend to zero, we have

$$(1-\varepsilon)^n \int_R J(x)\,dx \le v_n(\phi(R)) \le (1+\varepsilon)^n \int_R J(x)\,dx.$$

Letting $\varepsilon \to 0$ gives the desired result.

We are now ready to prove the theorem in the case when the domain is a generalized rectangle in $\mathbb{R}^n$.

Theorem 5.10.4 Let $R \subset \mathbb{R}^n$ be a generalized rectangle, let $U \subset \mathbb{R}^n$ be an open set containing $\overline{R}$, and let $\phi : U \to \mathbb{R}^n$ be C^1 on U and one-to-one on R, with $D\phi(x)$ invertible on R. Then a bounded real-valued function f on $\phi(R)$ is Riemann integrable on $\phi(R)$ if and only if $f \circ \phi$ is Riemann integrable on R, and in this case,

$$\int_{\phi(R)} f(y)\,dy = \int_R (f\circ\phi)(x)|\det D\phi(x)|\,dx.$$

Proof. The equivalence of the two integrability conditions follows immediately from the fact that ϕ is invertible and that the Jacobian $J(x) = |\det D\phi(x)|$ is everywhere nonzero on R. Let $P = \{R_1, R_2, \ldots, R_N\}$ be a partition of R. Let $m_i = \inf_{R_i}(f\circ\phi)$, and let $M_i = \sup_{R_i}(f\circ\phi)$. Then

$$\int_{\phi(R_i)} m_i \le \int_{\phi(R_i)} f \le \int_{\phi(R_i)} M_i.$$

By Lemma 5.10.2,

$$\int_{R_i} m_i J(x)\,dx \le \int_{\phi(R_i)} f \le \int_{R_i} M_i J(x)\,dx.$$

Let $s_P = \sum_{i=1}^N \int_{R_i} m_i J(x)\,dx$, and let $S_P = \sum_{i=1}^N \int_{R_i} M_i J(x)\,dx$. Then by Exercise 5.9.5,

$$s_P \le \int_{\phi(R)} f \le S_P.$$

It is also clear that

$$s_P \le \int_R (f\circ\phi)(x)J(x)\,dx \le S_P.$$

So it suffices to show that as $|P| \to 0$, $S_P - s_P \to 0$. Let C be a constant bounding $J(x)$ on R. Then

$$S_P - s_P \le C \cdot \sum_{i=1}^N (M_i - m_i)v(R_i).$$

Since $f \circ \phi$ is integrable on R, the right-hand side goes to zero as $|P| \to 0$. ∎

Finally, we are ready to prove the full-scale change of variables theorem.

Proof. (of Theorem 5.10.1) Let $\varepsilon > 0$. By Exercise 3.6.13, there exists an open set W such that $\overline{V} \subset W$ and $\overline{W}$ is compact and contained in U. Let C be a bound on $J(x) = |\det D\phi(x)|$ on $\overline{\phi(W)}$, and let M be a bound on $|f|$ on this same set. Let R be a closed generalized rectangle containing $\overline{V}$, and let P be a partition of R such that if S is a subrectangle of P, and $S \cap \overline{V} \neq \varnothing$, then $S \subseteq W$. This is possible by Exercise 3.6.12, since the distance from $\overline{V}$ to ∂W must be positive, and so we need only choose rectangles that are sufficiently small. We write $\mathscr{S}_1 = \{S \in P \mid S \subseteq V\}$, and $\mathscr{S}_2 = \{S \in P \mid S \cap \partial V \neq \varnothing\}$. Since ∂V has content 0, by Exercise 5.5.6, we may choose P such that $\sum_{S \in \mathscr{S}_2} v(S) < \varepsilon$.

We may write $V = A \cup B$, where $A = \bigcup_{S \in \mathscr{S}_1} S$, and $B = V \setminus A$. Then by Exercise 5.9.5,

$$\int_{\phi(V)} f = \int_{\phi(A)} f + \int_{\phi(B)} f.$$

Since A is a finite union of generalized rectangles, $\int_{\phi(A)} f = \int_A (f \circ \phi)(x)J(x)\,dx$, by Theorem 5.10.4. The second integral on the right can be bounded using Lemma 5.10.2 and our bounds on f and J as follows:

$$\begin{aligned}
\left|\int_{\phi(B)} f\right| &\leq M \cdot \int_{\phi(B)} 1 \\
&= M \cdot v_n(\phi(B)) \\
&\leq M \cdot v_n\left(\bigcup_{S \in \mathscr{S}_2} \phi(S)\right) \\
&= M \cdot \sum_{S \in \mathscr{S}_2} v_n(\phi(S)) \\
&\leq M \cdot \sum_{S \in \mathscr{S}_2} C \cdot v_n(S) \\
&< MC\varepsilon.
\end{aligned}$$

Similarly,

$$\int_V (f \circ \phi)(x)J(x)\,dx = \int_A (f \circ \phi)(x)J(x)\,dx + \int_B (f \circ \phi)(x)J(x)\,dx.$$

The second integral on the right can be straightforwardly bounded by $MC\varepsilon$. So

$$\begin{aligned}
\left|\int_{\phi(V)} f - \int_V (f \circ \phi)(x)J(x)\,dx\right| &= \left|\left(\int_{\phi(A)} f + \int_{\phi(B)} f\right) - \left(\int_A (f \circ \phi)(x)J(x)\,dx + \int_B (f \circ \phi)(x)J(x)\,dx\right)\right| \\
&= \left|\left(\int_A (f \circ \phi)(x)J(x)\,dx + \int_{\phi(B)} f\right) - \left(\int_A (f \circ \phi)(x)J(x)\,dx + \int_B (f \circ \phi)(x)J(x)\,dx\right)\right| \\
&= \left|\int_{\phi(B)} f - \int_B (f \circ \phi)(x)J(x)\,dx\right| \\
&< 2MC\varepsilon.
\end{aligned}$$

Since ε was arbitrary, the result follows. ∎

Example 5.10.5 We compute the area of the circle C of radius a centered at the origin in $\mathbb{R}^2$. (For comparison, see Example 5.9.7.) In the usual "rectangular" coordinates, this area is given by $A = \int_C 1 = \int_{-a}^{a} \int_{-\sqrt{a^2-x^2}}^{\sqrt{a^2-x^2}} 1\, dy\, dx$. The evaluation of the second iterated integral involves a change of variables in one variable, in particular, a trigonometric substitution. Instead, we compute this area using a change of variables known as polar coordinates.

Let R be the rectangle $(0, a) \times (0, 2\pi) \subset \mathbb{R}^2$. We define a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ by $\phi(r, \theta) = (r \cos\theta, r \sin\theta)$. Then

$$D\phi(r, \theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix},$$

so that $J(r, \theta) = |\det D\phi(r, \theta)| = |r|$. Note that ϕ is C^1 on all of $\mathbb{R}^2$, ϕ is one-to-one on R (though not on $\overline{R}$), and J is nonzero on R (though not on $\overline{R}$). Furthermore, ∂R has content 0.

The image of R under ϕ is $C \setminus \{(x, 0) \mid x \geq 0\}$, which differs from C by a set of content zero, so that the integral of any integrable function on C is the same as that on $\phi(R)$. (See Exercise 5.9.5.) So by Theorem 5.10.1,

$$\begin{aligned} A &= \int_C 1 \\ &= \int_{\phi(R)} 1 \\ &= \int_R J(r, \theta)\, dr\, d\theta \\ &= \int_0^{2\pi} \int_0^a r\, dr\, d\theta \\ &= \pi a^2. \end{aligned}$$

Exercise 5.10.6 Compute the volume of the ball of radius a centered at the origin in $\mathbb{R}^3$ using the following change of variables, known as spherical coordinates. Let $R = (0, a) \times (0, 2\pi) \times (0, \pi) \subset \mathbb{R}^3$, and let $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ be defined by $\phi(r, \theta_1, \theta_2) = (r \cos\theta_1 \sin\theta_2, r \sin\theta_1 \sin\theta_2, r \cos\theta_2)$.

Exercise 5.10.7

1. Find the volume in $\mathbb{R}^3$ of the intersection of the two "infinite cylinders"

$$C_1 = \{(x, y, z) \in \mathbb{R}^3 \mid y^2 + z^2 \leq 1\}$$

and

$$C_2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + z^2 \leq 1\}.$$

2. Find the volume in $\mathbb{R}^3$ of the intersection of the three "infinite cylinders" C_1, C_2 (as above), and

$$C_3 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 \leq 1\}.$$

Exercise 5.10.8 Find the volume in $\mathbb{R}^4$ of the intersection of the two "infinite hypercylinders"

$$H_1 = \{(x, y, z, w) \in \mathbb{R}^4 \mid x^2 + y^2 \leq 1\}$$

and

$$H_2 = \{(x, y, z, w) \in \mathbb{R}^4 \mid w^2 + z^2 \leq 1\}.$$

5.11 Projects

5.11.1 The Gamma Function The gamma function is in a class of functions occurring frequently in applications of mathematics. These are called “Special Functions” or sometimes “Transcendental Functions.” For a detailed and interesting exposition on this topic of Transcendental Functions, see Erdélyi et al., *Higher Transcendental Functions*, McGraw-Hill.

If s is a real number, $s > 0$, we define

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1}\, dt.$$

Of course, this is an improper integral for two reasons:

1. the domain of integration is unbounded; and
2. if $0 < s < 1$, then the integrand is unbounded in a neighborhood of 0.

Exercise 5.11.1 Show, by a simple use of the definition, that both parts of this improper integral converge if $s > 0$.

Exercise 5.11.2 Show that $\Gamma(1) = 1$.

Exercise 5.11.3 Show that for any $s > 0$,

$$s\Gamma(s) = \Gamma(s+1).$$

(Hint: Use and prove the validity of a generalization of integration by parts to improper integrals.)

Exercise 5.11.4 Show that $\Gamma(n+1) = n!$ for all nonnegative integers n.

Remark 5.11.5 This exercise gives further justification for the surprising fact that $0! = 1$.

Exercise 5.11.6 Generalize Leibniz’s rule to prove that Γ is infinitely differentiable with respect to s on the interval $(0, \infty)$.

Exercise 5.11.7 Find the point in $(0, 2)$ at which Γ assumes a minimum value, and determine this minimum value.

We now wish to present an exercise in which we compute one special value of the gamma function which will contribute to our computation of the volume of the unit ball in $\mathbb{R}^n$ in the next project.

Proposition 5.11.8 Show that

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

Proof. First note that

$$\Big(\int_{-\infty}^{\infty} e^{-x^2}\, dx\Big)^2 = \int_{-\infty}^{\infty} e^{-x^2}\, dx \int_{-\infty}^{\infty} e^{-y^2}\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\, dx\, dy.$$

If we change to polar coordinates, this last integral is

$$\int_0^{2\pi}\int_0^\infty e^{-r^2} r\, dr\, d\theta = 2\pi \int_0^\infty e^{-r^2} r\, dr = \pi \int_0^\infty e^{-u}\, du = \pi.$$

∎

Exercise 5.11.9 Show that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Exercise 5.11.10 Show that $\lim_{s\to 0^+}\Gamma(s) = +\infty$.

Next, using the functional equation $s\Gamma(s) = \Gamma(s+1)$, we want to extend the definition of Γ to negative real numbers. For values of s with $-1 < s < 0$, we define $\Gamma(s) = \Gamma(s+1)/s$. For example, $\Gamma(-\frac{1}{2}) = \frac{\Gamma(\frac{1}{2})}{-\frac{1}{2}} = -2\sqrt{\pi}$. Arguing inductively, we may continue this process to define the gamma function for all negative real values of s, not including the negative integers.

Exercise 5.11.11 Show that $\lim_{s\to 0^-}\Gamma(s) = -\infty$.

Exercise 5.11.12 Show by induction that if $n \in \mathbb{N}$, then we have the following limits. If n is odd, the limit as s approaches $-n$ from below of $\Gamma(s)$ is $+\infty$, and the limit as s approaches $-n$ from above is $-\infty$. If n is even, the limit as s approaches $-n$ from below of $\Gamma(s)$ is $-\infty$, and the limit as s approaches $-n$ from above is $+\infty$.

Exercise 5.11.13 Construct a graph of the gamma function.

5.11.2 Volume of the Unit Ball Here is a lesson in developing spherical coordinates in $\mathbb{R}^n$ and the associated volume element in these coordinates. This is, of course, a generalization of the formulas for polar coordinates and the volume element for $\mathbb{R}^2$. That is, if (x, y) is a point in $\mathbb{R}^2 \setminus \{(0, 0)\}$, we can write

$$x = r\cos\theta, \quad y = r\sin\theta,$$

where $r = \sqrt{x^2 + y^2}$, and θ is the unique solution in $[0, 2\pi)$ of the above pair of equations. The volume element

$$dx\,dy = r\,dr\,d\theta$$

can be derived from the change of variables formula or simply a geometric argument about sectors of circles in the plane. Both results that we derive are obtained by an iteration of polar coordinates.

Take a point $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, and assume that $(x_1, x_2) \neq (0, 0)$. First, we write (x_1, x_2) in polar coordinates. That is, $x_1 = r_1\cos\theta_1$ and $x_2 = r_1\sin\theta_1$, where $0 < r_1 < \infty$ and $0 \leq \theta_1 < 2\pi$. The related volume element as stated above is $dx_1\,dx_2 = r_1\,dr_1\,d\theta_1$. For the next step, take the pair (x_3, r_1) and treat it as a pair of Euclidean coordinates. Thus, we have

$$x_3 = r_2\cos\theta_2, \quad r_1 = r_2\sin\theta_2,$$

where

$$r_2 = \sqrt{x_3^2 + r_1^2} = \sqrt{x_3^2 + x_2^2 + x_1^2}.$$

Here, we have $0 < r_2 < \infty$ and $0 < \theta_2 < \pi$. Two things happen here. First of all, $(r_2, \theta_1, \theta_2)$ are the familiar spherical coordinates in $\mathbb{R}^3$. That is, $\rho^2 = r_2^2 = x_3^2 + x_2^2 + x_1^2$, $x_1 = r_2\sin\theta_2\cos\theta_1 = \rho\sin\phi\cos\theta$, and $x_2 = r_2\sin\theta_2\sin\theta_1 = \rho\sin\phi\sin\theta$. Second, the volume element for (x_3, r_1) can be written as $dx_3\,dr_1 = r_2\,dr_2\,d\theta_2$. By employing Fubini's Theorem, we can write

$$dx_3\,dx_2\,dx_1 = dx_3\,(r_1\,dr_1\,d\theta_1) = r_1 r_2\,dr_2\,d\theta_2\,d\theta_1.$$

By combining r_1 and r_2, we reach the expression

$$dx_1\,dx_2\,dx_3 = r_2^2\sin\theta_2\,dr_2\,d\theta_2\,d\theta_1,$$

and in the familiar spherical coordinates, this is

$$dx_1\,dx_2\,dx_3 = \rho^2\sin\phi\,d\rho\,d\phi\,d\theta,$$

where $0 < \rho < \infty$, $0 < \phi < \pi$, and $0 \leq \theta < 2\pi$.

Exercise 5.11.14 Show that for four dimensions, we have the following change of coordinates:

$$\begin{aligned} x_1 &= r_3 \sin\theta_3 \sin\theta_2 \cos\theta_1 \\ x_2 &= r_3 \sin\theta_3 \sin\theta_2 \sin\theta_1 \\ x_3 &= r_3 \sin\theta_3 \cos\theta_2 \\ x_4 &= r_3 \cos\theta_3, \end{aligned}$$

where $r_3 = x_4^2 + x_3^2 + x_2^2 + x_1^2$.

Exercise 5.11.15 Show that the four-dimensional volume element in spherical coordinates is

$$\rho^3 \sin^2\theta_1 \sin\theta_2 \, d\rho \, d\theta_1 \, d\theta_2 \, d\theta_3,$$

where the range is $0 < \rho < \infty$, $0 \le \theta_1 < 2\pi$, and $0 < \theta_2, \theta_3 < \pi$.

Exercise 5.11.16 Compute the surface area of the unit sphere S^3 in $\mathbb{R}^4$.

Exercise 5.11.17 Generalize the process developed above to prove the following formulas for spherical coordinates in $\mathbb{R}^n$:

$$\begin{aligned} x_1 &= \rho \sin\theta_{n-1} \sin\theta_{n-2} \cdots \sin\theta_2 \cos\theta_1 \\ x_2 &= \rho \sin\theta_{n-1} \sin\theta_{n-2} \cdots \sin\theta_2 \sin\theta_1 \\ x_3 &= \rho \sin\theta_{n-1} \sin\theta_{n-2} \cdots \sin\theta_3 \cos\theta_2 \\ &\vdots \\ x_{n-1} &= \rho \sin\theta_{n-1} \cos\theta_{n-2} \\ x_n &= \rho \cos\theta_{n-1}, \end{aligned}$$

where $0 < \rho < \infty$, $0 \le \theta_1 < 2\pi$, and $0 < \theta_j < \pi$ for $2 \le j \le n-1$.

We now have a formula for spherical coordinates in n dimensions.

Exercise 5.11.18 Show that the associated volume element in n dimensions is

$$dx_1 \, dx_2 \cdots dx_n = \rho^{n-1} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-1} \, d\rho \, d\theta_1 \, d\theta_2 \cdots d\theta_{n-1}.$$

At this point, it is possible to compute integrals for functions $f : \mathbb{R}^n \to \mathbb{R}$ written in spherical coordinates as $f(\rho, \theta_1, \theta_2, \ldots, \theta_{n-1})$. In particular, the $(n-1)$-dimensional volume element

$$d\sigma = \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-1} \, d\theta_1 \, d\theta_2 \cdots d\theta_{n-1}$$

allows us to compute integrals over the hypersurface S^{n-1}. (Note that the n-dimensional volume element can then be written as $\rho^{n-1} \, d\sigma \, d\rho$.) So if we have $f : S^{n-1} \to \mathbb{R}$, we can compute such an integral as follows:

$$\int_{S^{n-1}} f(\sigma) \, d\sigma = \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} f(\theta_1, \ldots, \theta_{n-1}) \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-1} \, d\theta_1 \, d\theta_2 \cdots d\theta_{n-1}.$$

Exercise 5.11.19 Compute this integral when f is the characteristic function of S^{n-1}.

Exercise 5.11.20 Compute the volume of the unit ball in $\mathbb{R}^n$ by computing

$$\int_0^1 \int_{S^{n-1}} \rho^{n-1} \, d\sigma \, d\rho.$$

To complete this project, we compute the volume of the n-ball through a process called "the use of auxiliary functions in analysis."

Exercise 5.11.21 Take the function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = e^{-|x|^2}$. Show that

$$\int_{\mathbb{R}^n} e^{-|x|^2}\, dx = \pi^{n/2}$$

by integrating in Euclidean coordinates.

Exercise 5.11.22 By integrating in spherical coordinates, show that

$$\int_{\mathbb{R}^n} e^{-|x|^2}\, dx = \int_0^\infty \int_{S^{n-1}} e^{-\rho^2} \rho^{n-1}\, d\sigma\, d\rho = \frac{1}{2}\Big(\int_{S^{n-1}} d\sigma\Big)\Gamma(\frac{n}{2}).$$

Conclude that the "surface area" of S^{n-1} is

$$\frac{\pi^{n/2}}{\frac{1}{2}\Gamma\left(\frac{n}{2}\right)}.$$

Exercise 5.11.23 Finally, show that

$$\mathrm{vol}(B_n) = \frac{\pi^{n/2}}{\frac{n}{2}\Gamma(\frac{n}{2})}.$$

Now that we have the volume of B_n, we would like to make some interesting observations. For instance, here is a table of these values for small n.

n	$\mathrm{vol}(B_n)$
1	2
2	2π
3	$\frac{4}{3}\pi$
4	$2\pi^2$
$\vdots$	$\vdots$

Exercise 5.11.24 Continue this table and determine the value of n for which $\mathrm{vol}(B_n)$ stops increasing and starts decreasing.

Exercise 5.11.25 Show that after reaching its maximum, $vol(B_n)$ is a monotonic decreasing sequence with the property

$$\lim_{n\to\infty} \mathrm{vol}(B_n) = 0.$$

If you want to have a nightmare, try to picture the unit ball inside a hypercube of side length 2, with the ball tangent to the hypercube at the center of each face. The volume of this hypercube is 2^n, which goes to ∞ as $n \to \infty$, while the volume of the n-ball, computed as above, goes to 0 as $n \to \infty$.

Exercise 5.11.26 *i.* Fix $r > 0$, and let $B_r^n(0)$ be the ball of radius r centered at the origin in $\mathbb{R}^n$. Show that $\lim_{n\to\infty} \mathrm{vol}(B_r^n(0)) = 0$.

ii. Fix $r > 0$, and let $C_r^n(0)$ be the hypercube with side length $2r$ centered at the origin in $\mathbb{R}^n$. Compute $\lim_{n\to\infty} \mathrm{vol}(C_r^n(0))$ for various values of r.

We have computed the volume of the unit ball in $\mathbb{R}^n$ in the ℓ^2 metric. One might ask what happens to the volume of the unit ball if we take the ℓ^p metric for $p \neq 2$.

Exercise 5.11.27 Let $1 \leq p \leq \infty$.

i. Compute the volume in $\mathbb{R}^n$ of the unit ball in the ℓ^p metric.

ii. Determine whether this volume goes to 0 as $n \to \infty$.

For further reading, consult Lecture III in *Lectures on the Geometry of Numbers* by Carl Ludwig Siegel.

Chapter 6

Vector Calculus and the Theorems of Green, Gauss, and Stokes

The goal of this chapter is to prove the integral theorems of Green, Gauss, and Stokes. These theorems are of great use in applications of mathematics to physics and many other fields. We begin by giving an exposition that relies on restrictions of the domain of application and the use of symbols whose definition is tailored to these restrictive conditions. In fact, our presentation is sufficient for most of the applications; however, the correct setting for these theorems lies in the calculus of differential forms, which will constitute the remainder of this chapter. Our approach will allow us to breathe life into the theory through the constant use of examples.

We begin with a careful treatment of curves in $\mathbb{R}^n$ and the study of the arc length of curves. After that, we state in order the theorems of Green, Gauss, and Stokes, and give references to what might be called the "classical" proofs. These proofs can be found, for example, in advanced calculus books by W. Kaplan or G. Folland.

6.1 Curves in $\mathbb{R}^n$

To begin our discussion of curves in $\mathbb{R}^n$, we will consider continuous images of closed intervals in $\mathbb{R}$. We would also like these functions to be differentiable, and we would like them to have nice behavior at the endpoints at these intervals.

Definition 6.1.1 Let $f : [a, b] \to \mathbb{R}^n$ be a function. The *right derivative* of f at a is the unique linear map $T : \mathbb{R} \to \mathbb{R}^n$ such that

$$\lim_{\substack{h\to 0,\\ h>0}} \frac{\|f(x+h) - f(x) - Th\|}{h} = 0,$$

provided such a T exists. We write $D_+f(a)$ for T. We define the *left derivative* $D_-f(b)$ of f at b similarly.

We use the preceding notion to define what it means for a function $f : [a, b] \to \mathbb{R}^n$ to be C^1.

Definition 6.1.2 Let $f : [a, b] \to \mathbb{R}^n$ be a function. We say that f is C^1 on $[a, b]$ if it is C^1 on (a, b), right differentiable at a, left differentiable at b, $\lim_{\substack{x\to a,\\ x>a}} Df(x) = D_+f(a)$, and $\lim_{\substack{x\to b,\\ x<b}} Df(x) = D_-f(b)$.

We are ready to proceed with our main agenda.

Definition 6.1.3 Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$. A *smooth parametrized curve* in $\mathbb{R}^n$ is a C^1 map $\phi : [a, b] \to \mathbb{R}^n$ with the property that $D\phi(t) \neq 0$ for $t \in (a, b)$, $D_+\phi(a) \neq 0$, and $D_-\phi(b) \neq 0$.

Example 6.1.4 If $f : [a, b] \to \mathbb{R}$ is a C^1 function, then the function $\phi : [a, b] \to \mathbb{R}^2$ given by $\phi(t) = (t, f(t))$ is a parametrization of the graph of f.

Example 6.1.5 The function $\phi : [0, 2\pi] \to \mathbb{R}^2$ given by $\phi(\theta) = (\cos\theta, \sin\theta)$ is a parametrization of the unit circle in $\mathbb{R}^2$.

Definition 6.1.6 Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$. A *piecewise smooth parametrized curve* in $\mathbb{R}^n$ is a continuous map $\phi : [a, b] \to \mathbb{R}^n$ for which there exists a partition $P = \{a_0, a_1, \ldots, a_k\}$ of $[a, b]$ such that ϕ restricts to a smooth parametrized curve on each subinterval $[a_i, a_{i+1}]$ of the partition.

Example 6.1.7 Suppose we wish to find a parametrization of the square in $\mathbb{R}^2$ with vertices $A = (0, 0)$, $B = (1, 0)$, $C = (1, 1)$, and $D = (0, 1)$. Because of the corners of the square, we will not be able to find such a smooth parametrized curve, but are able to give a piecewise smooth parametrized curve as follows.

$$\phi : [0, 4] \to \mathbb{R}^2, \quad \phi(t) = \begin{cases} (t, 0) & \text{if } 0 \le t < 1, \\ (1, t-1) & \text{if } 1 \le t < 2, \\ (3-t, 1) & \text{if } 2 \le t < 3, \text{ and} \\ (0, 4-t) & \text{if } 3 \le t \le 4. \end{cases}$$

Definition 6.1.8 A piecewise smooth parametrized curve $\phi : [a, b] \to \mathbb{R}^n$ is *closed* if $\phi(a) = \phi(b)$.

Definition 6.1.9 A piecewise smooth parametrized curve ϕ is *simple* if $\phi(s) \neq \phi(t)$ for $s \neq t$, with the obvious exception that we allow $\phi(a) = \phi(b)$ when ϕ is closed.

Note that the examples above are all simple, and the circle and square are closed.

Exercise 6.1.10

i. Give an example of a smooth parametrized curve that is closed but not simple.

ii. Give an example of a smooth parametrized curve that is neither closed nor simple.

Example 6.1.5 above shows that we can parametrize the unit circle with the function $\phi : [0, 2\pi] \to \mathbb{R}^2$ given by $\phi(\theta) = (\cos\theta, \sin\theta)$, which traces out the circle at constant speed in a counterclockwise direction starting at the point $(1, 0)$. All of these qualifications should tell us that there are many ways to parametrize the unit circle. For example, the function $\psi : [0, \pi] \to \mathbb{R}^2$ given by $\psi(\theta) = (\cos 2\theta, \sin 2\theta)$ traces out the same circle in the same direction from the same starting point but at double the speed of ϕ.

Exercise 6.1.11 For each of the following sets of conditions, give a parametrization of the unit circle in $\mathbb{R}^2$.

i. Counterclockwise, constant speed, starting at $(0, 1)$.

ii. Clockwise, constant speed, starting at $(1, 0)$.

iii. Counterclockwise, non-constant speed, starting at $(1, 0)$.

As functions, these parametrized curves are obviously distinct, but what they have in common is that they all trace out the same set in $\mathbb{R}^2$. We formalize this idea in the following definition.

Definition 6.1.12 Given a piecewise smooth parametrized curve $\phi : [a, b] \to \mathbb{R}^n$, the *path* C_ϕ of ϕ is the image of ϕ in $\mathbb{R}^n$, that is, C_ϕ is the path from $\phi(a)$ to $\phi(b)$ defined by the piecewise smooth parametrization ϕ.

Exercise 6.1.13 Consider the spiral $\phi : [0, 8\pi] \to \mathbb{R}^3$, $\phi(\theta) = (\cos\theta, \sin\theta, \theta)$. Find a second parametrization of C_ϕ. Find a third.

We have seen in the case of the unit circle that a parametrized curve has the distinguishing characteristics of direction, speed, and in the case of a closed curve, the starting point. (Note that the path of a non-closed curve can only be parametrized starting from one end or the other, while the path of a closed curve can be parametrized starting from any point.) As we will see, it is often the case in vector calculus that important calculations are independent of the speed of a parametrization, but do depend on the direction.

Definition 6.1.14 Two non-closed piecewise smooth parametrized curves $\phi : [a, b] \to \mathbb{R}^n$ and $\psi : [c, d] \to \mathbb{R}^n$ are *equivalent* if there exists a continuous, piecewise C^1, strictly monotonic function $\gamma : [a, b] \to [c, d]$ such that $\gamma(a) = c$ and $\gamma(b) = d$ and $(\psi \circ \gamma)(t) = \phi(t)$ for all $t \in [a, b]$.

Exercise 6.1.15 Show that if ϕ and ψ are equivalent, then $C_\phi = C_\psi$.

Exercise 6.1.16 Show that the property of being simple is preserved under equivalence.

Exercise 6.1.17 Which of the parametrizations that you constructed in Exercise 6.1.11 are equivalent? Are any equivalent to the parametrization from Example 6.1.5?

Definition 6.1.18 Let $\phi : [a, b] \to \mathbb{R}^n$ be a piecewise smooth parametrized curve. The *opposite parametrization* of ϕ is the piecewise smooth parametrized curve $-\phi : [a, b] \to \mathbb{R}^n$ given by $-\phi(t) = \phi(a + b - t)$.

Exercise 6.1.19 Let $\phi : [a, b] \to \mathbb{R}^n$ be a non-closed piecewise smooth parametrized curve. Show that ϕ and $-\phi$ are not equivalent.

Theorem 6.1.20 Let $\phi : [a, b] \to \mathbb{R}^n$ and $\psi : [c, d] \to \mathbb{R}^n$ be simple, non-closed, piecewise smooth parametrized curves with $C_\phi = C_\psi$. Then ψ is equivalent to either ϕ or $-\phi$.

Proof. We prove the result for smooth curves; the general result follows by taking refinements of partitions.

We suppose first that $\phi(a) = \psi(c)$ and $\phi(b) = \psi(d)$. By assumption, ψ is bijective with its image $C_\psi = C_\phi$, so we can define a map $\gamma : [a, b] \to [c, d]$ by $\gamma = \psi^{-1} \circ \phi$. The map γ is a composition of two homeomorphisms and hence a homeomorphism. We need to show that γ is C^1. Let $t_0 \in (a, b)$ and let $s_0 = \gamma(t_0) \in (c, d)$, and write $\psi(s) = (\psi_1(s), \psi_2(s), \ldots, \psi_n(s))$. Since $\psi'(s_0) \neq 0$, without loss of generality, we may assume that $\psi_1'(s_0) \neq 0$. Choose an open interval U around s_0 such that $\psi_1'(s) \neq 0$ on U. We now define a map $\Psi : U \times \mathbb{R}^{n-1} \to \mathbb{R}^n$ by $\Psi(s, s_2, s_3, \ldots, s_n) = (\psi_1(s), \psi_2(s) + s_2, \ldots, \psi_n(s) + s_n)$. The Jacobian matrix of Ψ is

$$D\Psi(s, s_2, \ldots, s_n) = \begin{pmatrix} \psi_1'(s) & 0 & \cdots & 0 \\ \psi_2'(s) & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_n'(s) & 0 & \cdots & 1 \end{pmatrix}.$$

This matrix has determinant $\psi_1'(s)$, which by assumption is nonzero on $U \times \mathbb{R}^{n-1}$. Thus, by the inverse function theorem, Ψ has a C^1 inverse $\Psi^{-1} : \Psi(U \times \mathbb{R}^{n-1}) \to U \times \mathbb{R}^{n-1}$. By continuity of ϕ, there exists an interval V around t_0 such that $\phi(V) \subseteq U$. If we write $\pi_1 : U \times \mathbb{R}^{n-1} \to U$ for the projection onto U, then we can write $\gamma\big|_V = \pi_1 \circ \Psi^{-1} \circ \phi\big|_V$, which is a composition of three C^1 functions and hence itself C^1.

If instead $\phi(a) = \psi(d)$ and $\phi(b) = \psi(c)$, we may replace ϕ by $-\phi$ and apply the above argument.

The preceding results show that given a simple, non-closed curve, there are exactly two equivalence classes of parametrizations. These two equivalence classes correspond to the two directions in which one can traverse the curve. This allows us to make the following definition.

Definition 6.1.21 An *oriented piecewise smooth curve* is an equivalence class of piecewise smooth parametrized curves.

In the case of closed or non-simple curves, there may be more than two equivalence classes. This problem can be addressed by defining the concatenation of curves.

Definition 6.1.22 Let $\phi_1 : [a_1, b_1] \to \mathbb{R}^n$ and $\phi_2 : [a_2, b_2] \to \mathbb{R}^n$ be piecewise smooth parametrized curves such that $\phi_2(a_2) = \phi_1(b_1)$. The *concatenation* $\phi = \phi_1 + \phi_2$ of ϕ_1 and ϕ_2 is the piecewise smooth parametrized curve $\phi_1 + \phi_2 : [a_1, b_1 + b_2 - a_2] \to \mathbb{R}^n$ given by

$$\phi(t) = \begin{cases} \phi_1(t) & \text{if } a_1 \leq t \leq b_1, \\ \phi_2(t - b_1 + a_2) & \text{if } b_1 < t \leq b_1 + b_2 - a_2. \end{cases}$$

Example 6.1.23 Note that in Example 6.1.7, the curve ϕ is already written as the concatenation of four other curves.

There are many ways that a curve can fail to be simple, but in practice, we will only consider curves that are concatenations of finitely many simple curves.

Exercise 6.1.24 Let $\phi : [0, 5] \to \mathbb{R}^2$ be the piecewise smooth parametrized curve given by

$$\phi(t) = \begin{cases} \left(-1 + \cos\left(\frac{5\pi}{4} + \frac{\pi}{2}t\right), 1 + \sin\left(\frac{5\pi}{4} + \frac{\pi}{2}t\right)\right) & \text{if } 0 \leq t < 1, \\ \left((\sqrt{2} - 1)\cos\left(\frac{5\pi}{4} - \frac{\pi}{2}t\right), (\sqrt{2} - 1)\sin\left(\frac{5\pi}{4} - \frac{\pi}{2}t\right)\right) & \text{if } 1 \leq t < 4, \text{ and} \\ \left(-1 + \cos\left(-\frac{7\pi}{4} + \frac{\pi}{2}t\right), -1 + \sin\left(-\frac{7\pi}{4} + \frac{\pi}{2}t\right)\right) & \text{if } 4 \leq t \leq 5. \end{cases}$$

i. Sketch the path C_ϕ in $\mathbb{R}^2$.

ii. How many oriented piecewise smooth curves have path C_ϕ?

iii. Give a parametrization of C_ϕ that is not equivalent to ϕ or $-\phi$.

Finally, it is worth noting that there are still infinitely many different equivalence classes of parametrizations for a closed curve, because we have infinitely many choices of the common starting and ending point. However, given a particular starting and ending point, a simple closed curve again has exactly two equivalence classes of parametrizations. In fact, it is the orientation that has a larger role to play than the choice of the starting and ending point (see Exercise 6.2.5 in the next section).

The next step in our analysis of curves is to find a way to compute the length of a curve. This, in turn, will lead us to a means to define the integral of a function on a curve.

Definition 6.1.25 Let $\phi : [a, b] \to \mathbb{R}^n$ be a (piecewise) smooth parametrized curve. Let $P = \{t_0, t_1, \ldots, t_n\}$, with $a = t_0 < t_1 < \cdots < t_m = b$, be a partition of $[a, b]$ such that any points at which $D\phi$ is discontinuous are included in P. A *polygonal approximation* to ϕ is the concatenation of the oriented line segments $\lambda_i : [0, 1] \to \mathbb{R}^n$, $\lambda_i(t) = (1 - t)\phi(t_{i-1}) + t\phi(t_i)$, as i runs from 1 to m.

Exercise 6.1.26 If $\phi : [a, b] \to \mathbb{R}^n$ is a (piecewise) smooth parametrized curve, show that the length of the polygonal approximation defined by the partition $a = t_0 < t_1 < \cdots < t_m = b$ is $\sum_{i=1}^m \|\phi(t_i) - \phi(t_{i-1})\|$.

We can rewrite this polygonal approximation as follows. If we write $\phi(t) = (\phi_1(t), \phi_2(t), \ldots, \phi_n(t))$, then $\|\phi(t_i) - \phi(t_{i-1})\|^2 = \sum_{j=1}^n (\phi_j(t_i) - \phi_j(t_{i-1}))^2$, and this is equal to $\sum_{j=1}^n (t_i - t_{i-1})^2 \left(\frac{\phi_j(t_i) - \phi_j(t_{i-1})}{(t_i - t_{i-1})}\right)^2$. By the mean-value theorem, there exists $t_i^{(j)} \in (t_{i-1}, t_i)$ such that $\frac{\phi_j(t_i) - \phi_j(t_{i-1})}{(t_i - t_{i-1})} = \frac{d\phi_j}{dt}(t_i^{(j)})$. Thus we can write the length of the polygonal approximation to ϕ as

$$\sum_{i=1}^m (t_i - t_{i-1}) \sqrt{\sum_{j=1}^n \left(\frac{d\phi_j}{dt}(t_i^{(j)})\right)^2}.$$

If we set $F(t) = \|\phi'(t)\| = \sqrt{\sum_{j=1}^n \frac{d\phi_j}{dt}^2}$, then the above expression is bounded by $L(F, P)$ and $U(F, P)$, where P is the partition $P = (t_0, t_1, \ldots, t_m)$. Since F is continuous (except at a finite number of points), it is integrable, and thus the lengths of the polygonal approximations converge to the value of the integral of F.

Definition 6.1.27 The *length* of the (piecewise) smooth parametrized curve ϕ is $\ell(\phi) = \int_a^b \|\phi'(t)\| \, dt$.

Exercise 6.1.28

i. Show that $\ell(\phi) = \ell(-\phi)$.

ii. Show that if ϕ and ψ are equivalent, then $\ell(\phi) = \ell(\psi)$. Thus, we may speak of the length of a smooth curve, independent of its orientation.

iii. Show that if ϕ and ψ are equivalent, and $f : U \to \mathbb{R}$ is a continuous function on some open set $U \subset \mathbb{R}^n$ containing $C_\phi = C_\psi$, then

$$\int_a^b f(\phi(t)) \|D\phi(t)\| \, dt = \int_c^d f(\psi(t)) \|D\psi(t)\| \, dt.$$

Example 6.1.29 Consider the spiral ϕ from Exercise 6.1.13. We compute

$$\phi'(\theta) = (-\sin\theta, \cos\theta, 1),$$

and hence

$$\begin{aligned} \ell(\phi) &= \int_0^{8\pi} \sqrt{(-\sin\theta)^2 + (\cos\theta)^2 + 1^2} \, d\theta \\ &= \int_0^{8\pi} \sqrt{2} \, d\theta \\ &= 8\sqrt{2}\pi \end{aligned}$$

Exercise 6.1.30 Show using your parametrization from Exercise 6.1.13 that the length of the spiral is still $8\sqrt{2}\pi$.

Exercise 6.1.31 Compute the lengths of the following curves.

i. $\phi : [0,1] \to \mathbb{R}^3$ given by $\phi(t) = (at, bt, ct)$, where a, b, and c are real constants, not all of which are zero.

ii. $\phi : [0,1] \to \mathbb{R}^2$ given by $\phi(t) = (t, t^3)$.

iii. $\phi : [0,1] \to \mathbb{R}^2$ given by $\phi(t) = (t^3, t^3)$. Why is your result unsurprising?

Definition 6.1.32 Let $\phi : [a,b] \to \mathbb{R}^n$ be a (piecewise) smooth parametrized curve. We say that ϕ is an *arc-length parametrization* if $\|\phi'(t)\| = 1$ wherever ϕ' exists.

Exercise 6.1.33 Let $\phi : [a,b] \to \mathbb{R}^n$ be an arc-length parametrization.

i. Show that $\ell(\phi) = b - a$.

ii. Show that if $[c,d] \subseteq [a,b]$, and $\phi_{[c,d]} : [c,d] \to \mathbb{R}^n$ is the restriction of ϕ to $[c,d]$, then $\ell(\phi_{[c,d]}) = d - c$.

Exercise 6.1.34 Let $\phi : [a,b] \to \mathbb{R}^n$ be a (piecewise) smooth parametrized curve such that, for every closed subinterval $[c,d] \subseteq [a,b]$, the restriction $\phi_{[c,d]} : [c,d] \to \mathbb{R}^n$ of ϕ to $[c,d]$ satisfies $\ell(\phi_{[c,d]}) = d - c$. Show that ϕ is an arc-length parametrization.

Exercise 6.1.35 Show that every (piecewise) smooth oriented curve admits a unique arc-length parametrization.

Exercise 6.1.36

i. Find the arc-length parametrization for the spiral from Exercise 6.1.13.

ii. Show that the parametrization of the square in Example 6.1.7 is an arc-length parametrization.

iii. Give an arc-length parametrization for the curve in Exercise 6.1.24. Show that with this parametrization, the curve is smooth.

iv. In order to get a sense of how intractable these types of problems can be more generally, attempt to find the arc-length parametrization of the curve $\phi : [0,1] \to \mathbb{R}^2$ given by $\phi(t) = (t, t^2)$.

6.2 Line Integrals in $\mathbb{R}^n$ and Differential 1-Forms

Line integrals have important applications in physics and elsewhere. We define them here and allow them to lead us to a discussion of differential 1-forms. With these ideas in place, we will be able to approach the first of the classical theorems of vector calculus, namely Green's theorem.

In the last section (Exercise 6.1.28), we saw how to integrate a real-valued function over a parametrized curve. In this section, we will define the line integral as the integral of a vector-valued function over a curve. In particular, if $\phi : [a,b] \to \mathbb{R}^n$ is our parametrized curve, and if $U \subseteq \mathbb{R}^n$ is an open set containing the path C_ϕ, we will integrate continuous functions of the form $F : U \to \mathbb{R}^n$. Such functions are often referred to as vector fields on U because they associate to each point in U a vector in $\mathbb{R}^n$.

Definition 6.2.1 Let $\phi : [a,b] \to \mathbb{R}^n$ be a piecewise smooth parametrized curve, let $U \subseteq \mathbb{R}^n$ be an open set containing C_ϕ, and let $F : U \to \mathbb{R}^n$ be a continuous function which is written in terms of its coordinate functions as $F = (F_1, \ldots, F_n)$. The *line integral* of F on ϕ is denoted by $\int_\phi F$ and is defined to be

$$\int_\phi F = \int_a^b (F \circ \phi)(t) \cdot \phi'(t)\, dt = \int_a^b [F_1(\phi(t))\phi_1'(t) + \cdots + F_n(\phi(t))\phi_n'(t)]\, dt.$$

Remark 6.2.2 There are two important things to note here. First, our function F takes values in $\mathbb{R}^n$ in part so that the dot product in the definition makes sense. Second, because we have taken the dot product before integrating, we are in fact integrating a single-variable function $(F \circ \phi) \cdot \phi' : [a,b] \to \mathbb{R}$, which is something we know how to do from elementary calculus.

Exercise 6.2.3 If $\phi : [a,b] \to \mathbb{R}^n$ and $\psi : [c,d] \to \mathbb{R}^n$ are equivalent piecewise smooth parametrized curves, then $\int_\phi F = \int_\psi F$.

Exercise 6.2.4 Let $\phi_1 : [a_1, b_1] \to \mathbb{R}^n$ and $\phi_2 : [a_2, b_2] \to \mathbb{R}^n$ be piecewise smooth parametrized curves such that $\phi_2(a_2) = \phi_1(b_1)$, and let ϕ be their concatenation. Show that $\int_\phi F = \int_{\phi_1} F + \int_{\phi_2} F$.

Exercise 6.2.5 Let $\phi : [a,b] \to \mathbb{R}^n$ be a simple, closed, piecewise smooth parametrized curve, and let ψ be another parametrization of C_ϕ. Show that $\int_\phi F = \pm \int_\psi F$. This shows that the line integral over a closed curve depends only on the orientation of the curve, not on the starting and ending point.

Exercise 6.2.6 Let $\phi : [a,b] \to \mathbb{R}^n$ be a piecewise smooth parametrized curve. Recall that the opposite parametrization of ϕ is the piecewise smooth parametrized curve $-\phi : [a,b] \to \mathbb{R}^n$ given by $(-\phi)(t) = \phi(a + b - t)$. Show that $\int_{-\phi} F = -\int_\phi F$.

Remark 6.2.7 The preceding exercise justifies the convention from one-variable integration that if f is integrable on $[a,b]$, then $\int_b^a f = -\int_a^b f$. This highlights the significance of the orientation of curves.

Example 6.2.8 Let $\phi : [0,\pi] \to \mathbb{R}^2$ be given by $\phi(\theta) = (\cos\theta, \sin\theta)$. Observe that the image of this parametrized curve is the top half of the unit circle, parametrized counterclockwise. Then $\phi'(\theta) = (-\sin\theta, \cos\theta)$. We compute the line integrals for four different functions on ϕ.

1. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $F(x, y) = (1, 0)$. Then

$$\begin{aligned}\int_\phi F &= \int_0^\pi (F_1(\cos\theta, \sin\theta)(-\sin\theta) + F_2(\cos\theta, \sin\theta)(\cos\theta))\,d\theta\\ &= \int_0^\pi -\sin\theta\,d\theta\\ &= \cos\pi - \cos 0\\ &= -2.\end{aligned}$$

2. Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $G(x, y) = (0, 1)$. Then

$$\begin{aligned}\int_\phi G &= \int_0^\pi (G_1(\cos\theta, \sin\theta)(-\sin\theta) + G_2(\cos\theta, \sin\theta)(\cos\theta))\,d\theta\\ &= \int_0^\pi \cos\theta\,d\theta\\ &= \sin\pi - \sin 0\\ &= 0.\end{aligned}$$

3. Let $H : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $H(x, y) = (x, y)$. Then

$$\begin{aligned}\int_\phi H &= \int_0^\pi (H_1(\cos\theta, \sin\theta)(-\sin\theta) + H_2(\cos\theta, \sin\theta)(\cos\theta))\,d\theta\\ &= \int_0^\pi (-\cos\theta\sin\theta + \sin\theta\cos\theta)\,d\theta\\ &= \int_0^\pi 0\,d\theta\\ &= 0.\end{aligned}$$

4. Let $I : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $I(x, y) = (-y, x)$. Then

$$\begin{aligned}\int_\phi I &= \int_0^\pi (I_1(\cos\theta, \sin\theta)(-\sin\theta) + I_2(\cos\theta, \sin\theta)(\cos\theta))\,d\theta\\ &= \int_0^\pi (\sin^2\theta + \cos^2\theta)\,d\theta\\ &= \int_0^\pi 1\,d\theta\\ &= \pi.\end{aligned}$$

Exercise 6.2.9 Compute the line integrals of the four functions in the example above for the following parametrized curves.

i. $\psi : [-1, 1] \to \mathbb{R}^2$ given by $\psi(t) = (-t, 0)$.

ii. $\rho : [0, \pi] \to \mathbb{R}^2$ given by $\rho(\theta) = (\cos\theta, -\sin\theta)$.

iii. $\sigma : [0, 2] \to \mathbb{R}^2$ given by $\sigma(t) = (t, t)$

iv. $\tau : [0, 2] \to \mathbb{R}^2$ given by $\tau(t) = (t, t^2)$.

Exercise 6.2.10 Find a piecewise smooth parametrization ϕ of the boundary of the triangle with vertices $(0, 0)$, $(1, 0)$, and $(1, 1)$ that starts and ends at $(0, 0)$ and traverses the vertices in the order given.

i. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $F(x, y) = (y^2, x)$. Find $\int_\phi F$.

ii. Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $G(x, y) = (y^2, 2xy)$. Find $\int_\phi G$.

The value of a line integral $\int_\phi F$ obviously depends on both the curve ϕ and the function F. We explore how these two components interact. At the infinitesimal level, we are taking the dot product of the vector that is the value of the function F with the vector that is the derivative of ϕ at each point. The value of this dot product, then, will tell us the extent to which these two vectors are aligned. Vectors in the same direction (that is, those that form an angle of less than $\frac{\pi}{2}$) will make a positive contribution towards the value of the integral, while those in opposite direction (that is, those that form an angle of more than $\frac{\pi}{2}$) will make a negative contribution. Perpendicular vectors make a contribution of zero. The integral represents the total cumulative effect of these contributions.

We now consider the different behaviors of the four line integrals in Example 6.2.8. Recall that in all four, the curve ϕ was the upper half of the unit circle, traversed in a counterclockwise direction. What aspects of this curve are revealed by each of these four line integrals?

In part (a), the function F is a constant function whose vector value is $(1, 0)$. If we interpret this as measuring the change of ϕ in the positive x direction, the line integral tells us that the accumulated change is -2, which matches our intuition about the diameter of the unit circle.

In part (b), the function G is again a constant function whose vector value is $(0, 1)$. If we interpret this as measuring the change of ϕ in the positive y direction, the line integral tells us that the accumulated change is 0. We note that this does not say that the curve has not moved at all in the y direction, but rather than from start to finish, the net change is zero.

In part (c), the function H is not constant. At each point in $\mathbb{R}^2$, the vector value of H is pointing radially outward from the origin. The line integral is therefore a measure of whether the curve is moving towards or away from the origin. Since our curve ϕ is always a constant distance from the origin, there is no radial change towards or away from the origin, and our dot product is identically zero. This behavior should be distinguished from part (b), where our answer was zero only by cancellation.

In part (d), the function I is again not constant. At each point in $\mathbb{R}^2$, the function I is producing a vector $(-y, x)$ that is orthogonal to the input (x, y). In this sense, I is measuring change not toward or away from the origin, but rather angular measure along circles of fixed radii, and in particular in a counterclockwise direction. Since our curve ϕ is a half-circle parametrized in a counterclockwise direction, the line integral measures the accumulated angular change, which is, unsurprisingly, π.

In these examples we have measured our curve ϕ by taking the dot product of its derivative vector with a vector-valued function at each point. We can free ourselves from the implicit bonds of coordinate geometry by recognizing these dot products as evaluations of linear maps on these tangent vectors. This naturally leads us to the following definition.

Definition 6.2.11 Let $\Omega \subseteq \mathbb{R}^n$. A *differential 1-form* on Ω is a map $\omega : \Omega \to \mathcal{L}(\mathbb{R}^n, \mathbb{R})$. The linear map in $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ associated with the point $x \in \Omega$ is denoted ω_x.

Remark 6.2.12 Thus, if ω is a differential 1-form, it takes elements of Ω and returns linear maps, while ω_x is a linear map which takes elements of $\mathbb{R}^n$ and returns elements of $\mathbb{R}$.

Remark 6.2.13 Because $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ is a finite dimensional real vector space of dimension n, we can pick a basis to identify it with $\mathbb{R}^n$. We then give $\mathbb{R}^n$ the usual metric. This allows us to define what it means for a differential 1-form $\omega : \Omega \to \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ to be continuous, differentiable, smooth, etc.

This definition should not look entirely unfamiliar. In Chapter 4, we identified the derivative of a function $f : U \to \mathbb{R}$ as a function that assigned to each point $x \in U$ the linear map $Df(x) : \mathbb{R}^n \to \mathbb{R}$. This linear map, applied to a vector v, represented the directional derivative of f in the direction v. Thus, we see that the derivatives of differentiable real-valued functions on $\mathbb{R}^n$ form a ready class of differential 1-forms.

Definition 6.2.14 Let $U \subset \mathbb{R}^n$ be an open set, and let $f : U \to \mathbb{R}$ be a differentiable function. The *differential* of f is the differential 1-form df given by $df_x = Df(x)$.

Remark 6.2.15 Note that in some sense, the preceding definition is merely a change in notation and terminology from our work in Chapter 4.

Definition 6.2.16 For each $i = 1, \ldots, n$, we identify a special differential 1-form by the name dx_i, which takes a point $x = (x_1, \ldots, x_n)$ to the linear map $\pi_i : \mathbb{R}^n \to \mathbb{R}$ which is the ith coordinate projection, that is, if $v \in \mathbb{R}^n$, then $\pi_i(v) = \pi_i(v_1, \ldots, v_n) = v_i$. In the cases that $n = 2$ or $n = 3$, it is traditional to write $dx = dx_1$, $dy = dx_2$, and $dz = dx_3$.

Exercise 6.2.17 Let $\pi_i : \mathbb{R}^n \to \mathbb{R}$ be the ith coordinate projection function. Show that $dx_i = d\pi_i$.

Remark 6.2.18 Observe that differential 1-forms on a set $\Omega \subset \mathbb{R}^n$ can be added together and be multiplied not just by real scalars but by arbitrary real-valued functions on Ω. That is, if ω_1 and ω_2 are differential 1-forms on Ω, then it is obvious what we mean by $\omega_1 + \omega_2$. Similarly, if $f : \Omega \to \mathbb{R}$ is a function, it is clear what differential 1-form we mean by $f\omega_1$.

Exercise 6.2.19 Let $\Omega \subseteq \mathbb{R}^n$, and let ω be a differential 1-form on Ω. Show that there exist unique functions $F_i : \Omega \to \mathbb{R}$ for $1 \leq i \leq n$ such that $\omega = F_1 dx_1 + \cdots + F_n dx_n$. Show that ω is continuous if and only if the functions F_i are continuous for each i.

Exercise 6.2.20 Let $U \subseteq \mathbb{R}^n$ be an open set, and let $f : U \to \mathbb{R}$ be a differentiable function. Show that

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_n} dx_n,$$

or in other symbols,

$$df_x = D_1 f(x) dx_1 + \cdots + D_n f(x) dx_n.$$

The implication of Exercise 6.2.19 is that there is a natural correspondence between differential 1-forms and vector-valued functions. We restate this more precisely as follows. If ω is a differential 1-form on a set $\Omega \subseteq \mathbb{R}^n$, then there exists a unique function $F : \Omega \to \mathbb{R}^n$ such that for all $x \in \Omega$ and $v \in \mathbb{R}^n$, we have

$$\omega_x(v) = F(x) \cdot v.$$

On the other hand, given such a function F, we can define a differential 1-form ω in this way. For example, if we consider the function $I : \mathbb{R}^2 \to \mathbb{R}^2$ given by $I(x, y) = (-y, x)$ from Example 6.2.8, we can write down the corresponding differential 1-form as $\omega = -y\,dx + x\,dy$. This correspondence allows us to rewrite the definition of line integrals in the language of differential 1-forms.

Definition 6.2.21 Let $\Omega \subseteq \mathbb{R}^n$, and let ω be a continuous differential 1-form on Ω. Let $\phi : [a, b] \to \Omega$ be a piecewise smooth parametrized curve. The *integral* of ω over ϕ is defined to be:

$$\int_\phi \omega = \int_a^b \omega_{\phi(t)}(\phi'(t))\,dt.$$

Remark 6.2.22

1. If $\omega = F_1 dx_1 + \cdots + F_n dx_n$, then

$$\begin{aligned}\int_\phi \omega &= \int_a^b [F_1(\phi(t))\phi_1'(t) + \cdots + F_n(\phi(t))\phi_n'(t)]\,dt \\ &= \int_\phi F.\end{aligned}$$

2. Because we can interpret the integral of a 1-form as a line integral of a function, Exercise 6.2.3 implies that $\int_\phi \omega = \int_\psi \omega$ if ϕ is equivalent to ψ.

Normally at this point, we would pose several exercises in which you would compute integrals of differential 1-forms. Please note, however, that the remark above shows that you have already done this back in Exercises 6.2.9 and 6.2.10. We also saw in the previous section how to find the length of a curve using integrals. Can we interpret the length of a parametrized curve in terms of differential 1-forms? Yes, as the following important example illustrates.

Example 6.2.23 Let $\phi : [a,b] \to \mathbb{R}^n$ be a simple smooth parametrized curve, and let $\Omega = C_\phi$. The *length 1-form* $\lambda : \Omega \to \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ is defined by the formula

$$\lambda_{\phi(t)}(v) = \frac{1}{\|\phi'(t)\|}[D\phi(t)](v).$$

The integral of this 1-form is

$$\int_\phi \lambda = \int_a^b \frac{1}{\|\phi'(t)\|}[D\phi(t)](\phi'(t))\,dt = \int_a^b \frac{1}{\|\phi'(t)\|}\|\phi'(t)\|^2\,dt = \int_a^b \|\phi'(t)\|\,dt,$$

which is precisely the length of the curve as given in Definition 6.1.27.

In single-variable calculus, the fundamental theorem (in the form of Corollary 5.3.2) related the integral of the derivative of a function to the values of that function. In essence, it expressed the total change in the value of a function over an interval as the accumulation of instantaneous changes. It is not unreasonable to expect a similar result to pertain to line integrals. That is, if the 1-form whose integral we are evaluating can be recognized as the differential of a function, we should expect that integral to represent the accumulated instantaneous changes of that function along the curve, and to be equal to the difference of the values of the function at the endpoints of the curve.

Theorem 6.2.24 (Fundamental Theorem for Line Integrals) Let $U \subseteq \mathbb{R}^n$ be an open set, and let $f : U \to \mathbb{R}$ be a C^1 function. Let $\phi : [a,b] \to U$ be a piecewise smooth curve. Then

$$\int_\phi df = f(\phi(b)) - f(\phi(a)).$$

Proof. Without loss of generality, assume that ϕ is a smooth curve. Applying the single-variable fundamental theorem of calculus gives

$$\begin{aligned}\int_\phi df &= \int_a^b [D_1 f(\phi(t))\phi_1'(t) + \cdots + D_n f(\phi(t))\phi_n'(t)]\,dt \\ &= \int_a^b \left[\frac{d}{dt}(f\circ\phi)\right](t)\,dt \\ &= (f\circ\phi)(b) - (f\circ\phi)(a).\end{aligned}$$

Corollary 6.2.25 Let $U \subseteq \mathbb{R}^n$ be an open set, let $f : U \to \mathbb{R}$ be a C^1 function, and let $\omega = df$. If ϕ and ψ are two smooth curves in U with the same beginning and end points, then

$$\int_\phi \omega = \int_\psi \omega.$$

Proof. Exercise.

Note the significance of this corollary. The conclusion would be true for any two equivalent parametrizations ϕ and ψ. In this corollary, though, because of our strong condition on the differential 1-form, ϕ and ψ are not necessarily equivalent parametrizations of the same curve. They merely need to start and end at the same points.

Exercise 6.2.26 Let $U \subseteq \mathbb{R}^n$ be a connected open set, and let ω be a continuous differential 1-form on U. Suppose that $\int_\phi \omega = \int_\psi \omega$ whenever ϕ and ψ begin and end at the same points. Show that there exists a C^1 function $f : U \to \mathbb{R}$ such that $\omega = df$. (Hint: Consider how the fundamental theorem of calculus allows us to construct the antiderivative of a continuous function of a single variable as an integral.)

Exercise 6.2.27 Show that the condition that $\int_\phi \omega = \int_\psi \omega$ whenever ϕ and ψ begin and end at the same points is equivalent to the condition that $\int_\rho \omega = 0$ whenever $\rho : [a, b] \to U$ parametrizes a closed curve. (Hint: Consider the trivial curve $\tau : [0, 0] \to U$ with $\tau(0) = \rho(a) = \rho(b)$.)

Example 6.2.28 We consider the vector-valued functions from Exercise 6.2.10 written in the new language of differential 1-forms. We wish to determine for each of these differential 1-forms ω whether there exists a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $\omega = df$.

1. Let $\omega = y^2\, dx + x\, dy$. You showed in Exercise 6.2.10 that $\int_\phi \omega \neq 0$. Since ϕ parametrizes a closed curve, it follows that $\omega \neq df$ for any function f.

2. Let $\omega = y^2\, dx + 2xy\, dy$. You showed in Exercise 6.2.10 that $\int_\phi \omega = 0$. This is no guarantee that the integral around every closed curve is zero, but it at least admits the possibility that such an f exists. If such an f does exist, we should be able to reconstruct it as follows. We fix a point, the origin, and determine the value of f at other points by integrating ω along straight-line curves from the origin.

   Let $(x, y) \in \mathbb{R}^2$, and let $\psi : [0, 1] \to \mathbb{R}^2$ be the straight-line curve from the origin to (x, y) given by $\psi(t) = (tx, ty)$. Then $\psi'(t) = (x, y)$. So

   $$\begin{aligned}\int_\psi \omega &= \int_0^1 (\psi_2(t)^2, 2\psi_1(t)\psi_2(t)) \cdot (x, y)\, dt \\ &= \int_0^1 (t^2y^2, 2t^2xy) \cdot (x, y)\, dt \\ &= \int_0^1 3t^2(xy^2)\, dt \\ &= xy^2.\end{aligned}$$

   Thus, if there exists a function f such that $\omega = df$, then $f(x, y) = xy^2$ must be such a function. Since we can easily compute $df = y^2\, dx + 2xy\, dy = \omega$, we discover that our f is indeed such a function.

Exercise 6.2.29 We consider the vector-valued functions from Example 6.2.8 written in the new language of differential 1-forms. For each, determine whether there exists a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $\omega = df$, and if so, determine such a function. (Hint: Use the computations in Example 6.2.8 and Exercise 6.2.9.)

i. $\omega = dx$

ii. $\omega = dy$

iii. $\omega = x\, dx + y\, dy$

iv. $\omega = -y\, dx + x\, dy$

We have shown that we have a special type of differential 1-form, whose integral is "independent of the path."

Definition 6.2.30 Let $U \subseteq \mathbb{R}^n$ be a connected open set, and let ω be a continuous differential 1-form on U. We say that ω is *exact* if there exists a C^1 function $f : U \to \mathbb{R}$ such that $\omega = df$. Such a function f is called a *primitive* for ω.

Exercise 6.2.31 Let ω be an exact differential 1-form on a connected open set $U \subseteq \mathbb{R}^n$ with primitive f.

i. Show that for any constant $C \in \mathbb{R}$, the function $f + C$ is also a primitive for ω.

ii. Show that if g is any primitive of ω, then $g = f + C$ for some constant $C \in \mathbb{R}$.

Exercise 6.2.32 Let $U \subseteq \mathbb{R}^2$ be a convex open set, and let $\omega = P(x, y)\, dx + Q(x, y)\, dy$ be a C^1 differential 1-form on U. Show that ω is exact if and only if $\dfrac{\partial P}{\partial y} = -\dfrac{\partial Q}{\partial x}$ as follows:

i. Assuming ω is exact with primitive f, use results about second-order partial derivatives.

ii. Assuming $\dfrac{\partial P}{\partial y} = -\dfrac{\partial Q}{\partial x}$, construct a primitive f. (Hint: First, determine the collection of functions $g : U \to \mathbb{R}$ such that $\dfrac{\partial g}{\partial x} = P(x, y)$. Then, show that there exists such a function f with $\dfrac{\partial f}{\partial y} = Q(x, y)$.)

Exercise 6.2.33 Construct a primitive for the following exact differential 1-forms on $\mathbb{R}^2$.

i. $\omega = x^3 y\, dx + \left(\frac{1}{4}x^4 + y^2\right) dy$

ii. $\omega = \left(y^2 e^{xy^2} + xe^{x^2}\right) dx + \left(2xye^{xy^2} - 2y\right) dy$

We finish this section with an application of differential 1-forms to ordinary differential equations.

Many first-order ordinary differential equations can be written in the form $P(x, y) + Q(x, y)\dfrac{dy}{dx} = 0$. We can sometimes solve such an equation by considering the differential 1-form $\omega = P(x, y)\, dx + Q(x, y)\, dy$ on $\mathbb{R}^2$. Suppose that ω is exact, that is, there exists a C^1 function f such that $df = \omega$. If y is considered as a function of x, we have

$$\begin{aligned} \frac{d}{dx} f(x, y) &= \frac{\partial f}{\partial x}(x, y) + \frac{\partial f}{\partial y}(x, y)\frac{dy}{dx} \\ &= P(x, y) + Q(x, y)\frac{dy}{dx}. \end{aligned}$$

Thus, y is a solution to the differential equation if and only if $\dfrac{d}{dx} f(x, y) = 0$, that is, $f(x, y) = c$ for some constant c. In other words, y is a solution to the differential equation if and only if the graph of y is a level curve of the function f. Because of the relationship between the differential equation and the differential 1-form ω, many textbooks present the original differential equation in the form $P(x, y)\, dx + Q(x, y)\, dy = 0$.

Exercise 6.2.34 Recall that a separable differential equation is one that can be written in the form $q(y)\dfrac{dy}{dx} = p(x)$.

i. Show that the differential 1-form ω associated with this differential equation is exact.

ii. Use the above method to solve the separable differential equation

$$\frac{dy}{dx} = 4x^3 y$$

with initial condition $y = 2$ when $x = 0$.

6.3 Green's Theorem in the Plane

One of the major and most impressive theorems in vector calculus is Green's theorem, which relates the integral of a differential 1-form along a closed curve to the integral of a related function on the region bounded by the curve. It actually produces a rather astounding result. We state it first for curves in $\mathbb{R}^2$.

Lemma 6.3.1 Let $U \subseteq \mathbb{R}^2$ be an open set, and let $\Omega \subset U$ be a compact region with the property that there exists an interval $[a, b]$ and piecewise smooth functions $f_1, f_2 : [a, b] \to \mathbb{R}$ such that $f_1(x) < f_2(x)$ for all $x \in [a, b]$, and

$$\Omega = \{(x, y) \mid x \in [a, b], f_1(x) \le y \le f_2(x)\}.$$

Let ϕ be a "counterclockwise" parametrization of $\partial\Omega$, and let $P : U \to \mathbb{R}$ be a C^1 function. Then

$$\int_\Omega -\frac{\partial P}{\partial y} = \int_\phi P\,dx.$$

Proof. Because of the conditions on f_1, f_2, and Ω above, it is possible to parametrize the $\partial\Omega$, which we will proceed to do. By Exercise 6.2.5, we may choose a preferred parametrization of $\partial\Omega$. We construct ϕ as follows.

1. Let $\phi_1 : [a, b] \to \mathbb{R}^2$ be given by $\phi_1(t) = (t, f_1(t))$.
2. Let $\phi_2 : [f_1(b), f_2(b)] \to \mathbb{R}^2$ be given by $\phi_2(t) = (b, t)$.
3. Let $\phi_3 : [a, b] \to \mathbb{R}^2$ be given by $\phi_3(t) = (t, f_2(t))$.
4. Let $\phi_4 : [f_1(a), f_2(a)] \to \mathbb{R}^2$ be given by $\phi_4(t) = (a, t)$.

Let ϕ be the piecewise smooth curve obtained by concatenating ϕ_1, ϕ_2, $-\phi_3$, and $-\phi_4$, in that order. The meaning of the word "counterclockwise" in the statement of the theorem should be obvious.

We first work out the left-hand side using Fubini's theorem.

$$\begin{aligned}
\int_\Omega -\frac{\partial P}{\partial y} &= \int_a^b \int_{f_1(x)}^{f_2(x)} -\frac{\partial P}{\partial y}\,dy\,dx \\
&= -\int_a^b (P(x, f_2(x)) - P(x, f_1(x)))\,dx \\
&= \int_a^b P(x, f_1(x))\,dx - \int_a^b P(x, f_2(x))\,dx.
\end{aligned}$$

The right-hand side can be broken up into the sum of four line integrals.

$$\int_\phi P\,dx = \int_{\phi_1} P\,dx + \int_{\phi_2} P\,dx - \int_{\phi_3} P\,dx - \int_{\phi_4} P\,dx.$$

The second and fourth integrals are zero, since there is no change in x along a vertical line. The first integral is

$$\int_{\phi_1} P\,dx = \int_a^b P(t, f_1(t))\,dt,$$

and the third integral is

$$\int_{\phi_3} P\,dx = \int_a^b P(t, f_2(t))\,dt.$$

So

$$\int_\phi P\,dx = \int_a^b P(t, f_1(t))\,dt - \int_a^b P(t, f_2(t))\,dt = \int_\Omega -\frac{\partial P}{\partial y}.$$

Lemma 6.3.2 Let $U \subseteq \mathbb{R}^2$ be an open set, and let $\Omega \subset U$ be a compact region with the property that there exists an interval $[c, d]$ and piecewise smooth functions $g_1, g_2 : [c, d] \to \mathbb{R}$ such that $g_1(y) < g_2(y)$ for all $y \in [c, d]$, and

$$\Omega = \{(x, y) \mid y \in [c, d], g_1(y) \leq x \leq g_2(y)\}.$$

Let ϕ be a "counterclockwise" parametrization of $\partial\Omega$, and let $Q : U \to \mathbb{R}$ be a C^1 function. Then

$$\int_\Omega \frac{\partial Q}{\partial x} = \int_\phi Q\,dy.$$

Proof. Exercise.

Theorem 6.3.3 (Special Case of Green's Theorem in the Plane) Let $U \subseteq \mathbb{R}^2$ be an open set, and let $\Omega \subset U$ be a compact region with the property that there exist intervals $[a, b]$ and $[c, d]$ and piecewise smooth functions $f_1, f_2 : [a, b] \to \mathbb{R}$, $g_1, g_2 : [c, d] \to \mathbb{R}$, such that $f_1(x) < f_2(x)$ for all $x \in [a, b]$, $g_1(y) < g_2(y)$ for all $y \in [c, d]$, and

$$\Omega = \{(x, y) \mid x \in [a, b], f_1(x) \leq y \leq f_2(x)\} = \{(x, y) \mid y \in [c, d], g_1(y) \leq x \leq g_2(y)\}.$$

Let ϕ be a "counterclockwise" parametrization of $\partial\Omega$, and let $P : U \to \mathbb{R}$ and $Q : U \to \mathbb{R}$ be C^1 functions. Then

$$\int_\Omega \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) = \int_\phi (P\,dx + Q\,dy).$$

Proof. Add the results of the previous two lemmas.

Remark 6.3.4

1. Note that there is a slight asymmetry in the two lemmas. The occurrence of the minus sign is governed by the chosen orientation of the boundary. Had we chosen a "clockwise" parametrization, the signs would have been reversed.
2. The conditions that we imposed on Ω may seem strong. However, the types of regions for which the result of Green's Theorem applies may be broadened to include any regions that may be decomposed into a finite union of regions of the type described in the theorem. Most regions that occur in applications fall into this large class.

One of the most significant applications of Green's Theorem is the following remarkable result. If Ω is a bounded region in the plane whose boundary can be parametrized in the fashion above, then we can compute the area of Ω as follows. If we can find a pair of functions P and Q such that $\dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y} = 1$, then the left-hand integral in the statement of Green's Theorem becomes $\displaystyle\int_\Omega 1$, which is equal to the area of Ω. There are many such choices of P and Q, such as $P(x, y) = -y$, and $Q(x, y) = 0$, or, for example, $P(x, y) = 0$, and $Q(x, y) = x$. A particularly nice choice is the pair of functions $P(x, y) = -\dfrac{1}{2}y$, and $Q(x, y) = \dfrac{1}{2}x$. In this case, Green's Theorem gives us

$$\text{Area}(\Omega) = \int_\Omega 1 = \int_\Omega \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) = \int_{\partial\Omega} \left(-\frac{1}{2}y\,dx + \frac{1}{2}x\,dy\right).$$

This particular way of evaluating the area of Ω permits the construction of a *planimeter*, an instrument used to compute area whose working principle is based on this equation. If you wheel a planimeter around the boundary of a lake, the area of the lake just pops out.

Exercise 6.3.5 Use the method described above to find the area of the region bounded by the ellipse $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$.

Exercise 6.3.6 Use the method described above to find the area of the region bounded by the folium of Descartes $x^3 + y^3 = 3xy$ in the first quadrant.

The most important step in our proof of the special case of Green's theorem is the application of the fundamental theorem of calculus. In fact, if we take one step back, we can view the special case of Green's theorem as being a version of the fundamental theorem of calculus just one dimension higher. That is, each of them compares the behavior of a function on a region to the behavior of a closely related function on the boundary of that region. In the case of the fundamental theorem, the region is a closed interval $[a, b]$ in $\mathbb{R}$, and the boundary of that region is the set $\{a, b\}$. If the function being measured on the boundary is F, then the function being measured on the interval is $f = F'$. In the case of Green's theorem, the region is a compact set Ω in $\mathbb{R}^2$, and the boundary is a piecewise smooth closed curve ϕ. If the "function" being measured on the boundary ϕ is the differential 1-form $\omega = P\,dx + Q\,dy$, then the function being measured on the set Ω is $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}$.

Can we phrase both of these results in a common language? Yes! Unsurprisingly, this language is the language of differential forms. Let us spend a minute rephrasing the fundamental theorem of calculus purely in the language of differential forms.

Definition 6.3.7 Let $\Omega \subseteq \mathbb{R}^n$. A *differential 0-form* on Ω is a function $f : \Omega \to \mathbb{R}$.

Definition 6.3.8 Let $U \subseteq \mathbb{R}^n$, and let f be a C^1 differential 0-form on U. The *differential* of f is the differential 1-form df on U.

Remark 6.3.9 Note that in some sense, the preceding definition is merely a change in notation and terminology from our work in the previous section. (See Definition 6.2.14.)

Just as differential 1-forms can be integrated over sets that are essentially 1-dimensional, that is, piecewise smooth curves, differential 0-forms can be integrated over sets that are fundamentally 0-dimensional in nature, that is, finite sets. How should one define the integral of a differential 0-form on a finite set? By evaluating the function at each point of that set and "adding." Before we can define the integral of a differential 0-form, however, we must address the issue of orientation.

Recall from Exercise 6.2.6 that if ϕ is a piecewise smooth parametrized curve and ω is a differential 1-form, then $\int_{-\phi} \omega = -\int_\phi \omega$. This reflects the fact that integration of differential 1-forms depends not just on the underlying path C_ϕ as a subset of $\mathbb{R}^n$, but on the orientation of that path, which is encoded in the parametrization. Similarly, it will be important to assign an orientation to the sets on which we integrate differential 0-forms.

Definition 6.3.10 Let $X \subset \mathbb{R}^n$ be a finite set. An *orientation* on X is a function $O : X \to \{\pm 1\}$.

Definition 6.3.11 Let $X \subset \mathbb{R}^n$ be a finite set with orientation O, and let f be a differential 0-form on X. The *integral* of f on the oriented set X is defined to be $\int_X f = \sum_{x \in X} O(x) \cdot f(x)$.

Exercise 6.3.12 Let $X = \{1, 2, \ldots, n\}$ have orientation $O(x) = (-1)^x$. Let $f : X \to \mathbb{R}$ be given by $f(x) = \left(\frac{1}{2}\right)^x$. Find $\int_X f$.

We are now ready to restate the fundamental theorem of calculus. We do so using the language of differential 0-forms and differential 1-forms, as well as oriented intervals and their oriented boundaries. Intervals in $\mathbb{R}$ are endowed with natural orientations coming from the ordering on the real numbers. If $\Omega = [a, b] \subset \mathbb{R}$ with $a < b$, then the boundary is the set $\partial\Omega = \{a, b\}$, and the natural orientation on this boundary is $O : \partial\Omega \to \{\pm 1\}$ given by $O(a) = -1$, $O(b) = +1$.

Theorem 6.3.13 (Fundamental Theorem of Calculus) Let $\Omega = [a, b]$, and let f be a continuous differential 0-form on $[a, b]$ that is C^1 on (a, b). Then

$$\int_{\Omega} df = \int_{\partial\Omega} f.$$

These same ideas immediately apply to give a reformulation of the fundamental theorem for line integrals. Let $U \subseteq \mathbb{R}^n$ be an open set, let $\phi : [a, b] \to U$ be a piecewise smooth curve, and let $\Omega = C_\phi$, the path of ϕ. The parametrization of ϕ gives a natural "orientation" to the set Ω, as well as to the boundary $\partial\Omega$, namely $O : \partial\Omega \to \{\pm 1\}$ given by $O(\phi(a)) = -1$ and $O(\phi(b)) = +1$.

Theorem 6.3.14 (Fundamental Theorem for Line Integrals) Let $U \subseteq \mathbb{R}^n$ be an open set, let $\phi : [a, b] \to U$ be a piecewise smooth curve, and let $\Omega = C_\phi$. Let f be a C^1 differential 0-form on U. Then

$$\int_{\Omega} df = \int_{\partial\Omega} f.$$

What about Green's Theorem? The right-hand side of Green's Theorem in the plane is already expressed in the form $\int_{\partial\Omega} \omega$, that is, the integral of a differential 1-form ω over the boundary $\partial\Omega$ of a 2-dimensional region. If we are to recognize Green's Theorem as the 2-dimensional version of the Fundamental Theorem of Calculus, we would like to recognize the function $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}$ as the "differential 2-form" $d\omega$. We will explore the formalization of these ideas in the next two sections. Once we do, we will see that we can restate Green's Theorem in the Plane as follows.

Theorem 6.3.15 (Green's Theorem in the Plane) Let Ω be a compact region in $\mathbb{R}^2$ whose boundary $\partial\Omega$ is the path of a piecewise smooth curve. Let ω be a continuous differential 1-form on Ω that is C^1 on the interior Ω°. Then

$$\int_{\Omega} d\omega = \int_{\partial\Omega} \omega.$$

Finally, the same sort of reasoning by which we can view the fundamental theorem for line integrals as a generalization of the fundamental theorem of calculus will lead us to a version of Green's Theorem for 2-dimensional surfaces in $\mathbb{R}^n$.

6.4 Surfaces in $\mathbb{R}^n$

Our immediate task is to define and characterize 2-dimensional surfaces in $\mathbb{R}^n$. Traditionally, these objects are referred to simply as surfaces. When it comes time to discuss more general objects of dimension greater than 2, we will introduce new vocabulary.

In the same manner that curves are essentially the images of intervals under continuous maps, with some differentiability hypotheses, surfaces will be the images of rectangles under continuous maps with some differentiability hypotheses. In the case of curves, the boundary of a domain was a pair of points, which led to two possible behaviors: closed curves, where the two were mapped to the same point, and non-closed curves, where they were not. The boundary of a rectangle, however, is the union of four line segments whose images are potentially much more complicated. To give a sense of the various behaviors of the boundaries of surfaces, we give some examples before the formal definition.

Example 6.4.1 Let $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$, and let $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ be linearly independent vectors in $\mathbb{R}^3$. We may parametrize the parallelogram spanned by the vectors $\mathbf{u}$ and $\mathbf{v}$ from the point $\mathbf{x}$ by

$$\begin{aligned}&\psi : [0, 1] \times [0, 1] \to \mathbb{R}^3,\\ &\qquad \psi(s, t) = \mathbf{x} + s\mathbf{u} + t\mathbf{v} = (x_1 + su_1 + tv_1, x_2 + su_2 + tv_2, x_3 + su_3 + tv_3).\end{aligned}$$

Example 6.4.2 Let $\Delta = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq y \leq 1, 0 \leq x \leq y\}$. We may parametrize this triangle by

$$\psi : [0,1] \times [0,1] \to \mathbb{R}^2,$$
$$\psi(s,t) = (s, st).$$

Example 6.4.3 Let $C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1, |z| \leq 1\}$. We may parametrize this cylinder by

$$\psi : [0,1] \times [0,1] \to \mathbb{R}^3,$$
$$\psi(s,t) = (\cos(2\pi s), \sin(2\pi s), -1 + 2t).$$

Example 6.4.4 Let $A = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1 - z^2, 0 \leq z \leq 1\}$. We may parametrize this cone by

$$\psi : [0,1] \times [0,2\pi] \to \mathbb{R}^3,$$
$$\psi(s,t) = (s\cos t, s\sin t, 1 - s)$$

In each of these examples, the function ψ is C^1 on the interior of the rectangle. Moreover, just as our curves were required to have nonzero derivative so that they were locally injective and the parametrization was never stationary, here the derivative of ψ has full rank at every point in the interior.

On the hand, the boundaries behave much differently in each example. In the first example, the boundary is mapped injectively onto the boundary of the parallelogram. In particular, the restriction of ψ to each side of the rectangle parametrizes a curve. In the second example, three of the sides of the rectangle parametrize curves, but the fourth side is collapsed to a point. In the third example, the top and bottom sides parametrize separate closed curves, while the left and right sides parametrize curves with the same path. The fourth example combines the behaviors of the second and third examples.

Definition 6.4.5 Let $R = [a, b] \times [c, d]$ be a closed rectangle in $\mathbb{R}^2$. A *smooth parametrized surface* in $\mathbb{R}^n$ is a continuous map $\psi : R \to \mathbb{R}^n$ such that ψ is C^1 on R°, and for each point $(s, t) \in R^\circ$, the map $D\psi(s, t)$ has rank 2 (see Remark 2.2.19), that is, $D_1\psi_i(s, t)$ and $D_2\psi_i(s, t)$ are linearly independent vectors in $\mathbb{R}^n$. Furthermore, let $\phi_1 : [a, b] \to \mathbb{R}^n$ be given by $\phi_1(s) = \psi(s, c)$, let $\phi_2 : [c, d] \to \mathbb{R}^n$ be given by $\phi_2(t) = \psi(b, t)$, let $\phi_3 : [a, b] \to \mathbb{R}^n$ be given by $\phi_3(s) = \psi(a + b - s, d)$, and let $\phi_4 : [c, d] \to \mathbb{R}^n$ be given by $\phi_4(t) = \psi(a, c + d - t)$. For each i, we require that ϕ_i be either a piecewise smooth parametrized curve or a constant function, and, for each pair of distinct i and j, the images of ϕ_i and ϕ_j must either intersect nowhere, intersect at one or both endpoints, or coincide. Moreover, if they coincide, they must give opposite orientations of their shared path.

The last condition in the above definition requires some explanation. For the purposes of integration, we will require our surfaces to be "orientable," and it is this condition that will ensure that they are. In particular, this condition will assure that when we integrate along the boundary curve, the portions the integral corresponding to two coincident sides will cancel.

We offer the following informal description of orientation of surfaces in $\mathbb{R}^3$ to help guide the reader's intuition in the matter. (Those whose geometric intuition in $\mathbb{R}^4$ is strong should be forewarned that this description fails in higher dimensions.) It is easy to imagine what we mean by the two faces of a rectangle sitting in $\mathbb{R}^3$. The plane containing the rectangle separates $\mathbb{R}^3$ into two halves, each of which corresponds to a face of the rectangle. We can imagine these two faces being painted with two different colors. Likewise we will say that a surface is orientable if it has two faces that can be "painted" with two different colors. If our surface is correctly (in the sense of the definition above) parametrized by a rectangle that has been painted in two different colors, then this parametrization transfers these two colors to the two faces of our surface. The following example shows what can go wrong without the last condition.

Example 6.4.6 The most famous example of a non-orientable surface is the Möbius band. Consider the following parametrization.

$$\psi : [0,2\pi] \times [-1,1] \to \mathbb{R}^3,$$
$$\psi(\theta,t) = (2\cos\theta, 2\sin\theta, 0) + t(\cos\theta\cos\frac{\theta}{2}, \sin\theta\cos\frac{\theta}{2}, \sin\frac{\theta}{2}).$$

This parametrization identifies the left and right sides of the rectangle. Their common path is the line segment $\{(x, 0, 0) \mid 1 \leq x \leq 3\}$, but ϕ_2 and ϕ_4 parametrize this segment in the same direction. If we color the faces of the rectangle, then the colors transferred to the Möbius band by ψ will not match up along this line segment.

Next, we define the parametrized boundary of a smooth parametrized surface.

Definition 6.4.7 Let $R = [a, b] \times [c, d]$, and let $\psi : R \to \mathbb{R}^n$ be a smooth parametrized surface. Let $\phi_1, \ldots, \phi_4$ be the functions which parametrize the boundary of R as in Definition 6.4.5. The *parametrized boundary* of the surface is the concatenation, $\phi_1 + \phi_2 + \phi_3 + \phi_4$, of these curves.

This is an unusual definition. Let us consider the implications of this definition for Example 6.4.3 above. According to our definition, the parametrized boundary of ψ is curve that starts at $(1, 0, -1)$, traverses the circle $C_1 = \{(x, y, z) \mid x^2 + y^2 = 1, z = -1\}$ counterclockwise as viewed from the positive z-axis, goes up the line segment from $(1, 0, -1)$ to $(1, 0, 1)$, traverses the circle $C_2 = \{(x, y, z) \mid x^2 + y^2 = 1, z = 1\}$ clockwise as viewed from the positive z-axis, and finally back down the line segment from $(1, 0, 1)$ to $(1, 0, -1)$. The two things to note here are that we are including the parametrization as part of this definition, and that the doubly parametrized line segment is part of the parametrized boundary, even though we would normally consider the boundary of the cylinder to consist only of the two circles. Indeed, the parametrized boundary clearly depends on the choice of parametrization ψ, whereas in more advanced courses on differential geometry, the subset $C \subset \mathbb{R}^3$ will be considered to have a boundary $\partial C = C_1 \cup C_2$ independent of any parametrization.

Our choice to distinguish the parametrized boundary from the usual boundary will be of immediate use when we integrate over the boundary. The integral of a 1-form over the parametrized boundary is equal to its integral over the usual boundary by Exercise 6.2.6. Integrating over the parametrized boundary will simply some of our proofs, while integrating over the usual boundary often simplifies calculations.

Example 6.4.8 Let $R = [0, 1] \times [0, 2\pi]$, and let $\phi : R \to \mathbb{R}^2$ be given by $\phi(s, t) = (s \cos t, s \sin t)$. This is the parametrization of the unit disk that we used in Example 5.10.5 to define polar coordinates.

Example 6.4.9 Let $R = [0, 2\pi] \times [0, \pi]$, and let $\phi : R \to \mathbb{R}^3$ be given by $\phi(s, t) = (\cos s \sin t, \sin s \sin t, \cos t)$. This is the parametrization of the unit sphere coming from spherical coordinates (see Exercise 5.10.6 and Project 5.11.2).

Exercise 6.4.10 Verify that $D\phi(s, t)$ has rank 2 for all $(s, t) \in R^\circ$ in the above three examples.

Exercise 6.4.11 Find a parametrization of each of the following surfaces.

i. $T^2 = \{(x, y, z, w) \in \mathbb{R}^4 \mid x^2 + y^2 = 1, z^2 + w^2 = 1\}$

ii. $T = \left\{(x, y, z) \in \mathbb{R}^3 \mid \left(2 - \sqrt{x^2 + y^2}\right)^2 + z^2 = 1\right\}$

One of the most basic things one might like to know about a surface is the "surface area." We first discuss the special case of surfaces in $\mathbb{R}^3$ and then see how this generalizes to surfaces in $\mathbb{R}^n$.

The first approach one might think to use is a sort of "polyhedral approximation." Suppose we have a piecewise smooth parametrized surface defined by a single rectangle $R = [a, b] \times [c, d]$ and a single function $\phi : R \to \mathbb{R}^3$. If we take a regular partition P of R and subdivide each subrectangle of P into a pair of triangles, then the images under ϕ of the vertices of each triangle form a triangle in $\mathbb{R}^3$, and together, these triangles form a polyhedral surface in $\mathbb{R}^3$ with triangular faces. The sum of the triangular areas is an approximation to the area of our surface.

Unfortunately, such approximations do not converge even for some extremely simple examples.

Example 6.4.12 (Schwarz's Lantern) Consider the lateral surface of a right circular cylinder C in $\mathbb{R}^3$ with height 1 and radius 1. We triangulate C as follows. First, we subdivide C into m congruent bands, each of height $\frac{1}{m}$. We then approximate each band with $2n$ congruent isosceles triangles arranged such the bases of n of them form a regular n-gon on one end of the band, and the bases of the other n form a regular n-gon on the other end. The area of one of these triangles is $\sin\frac{\pi}{n}\sqrt{\frac{1}{m^2}+\left(1-\cos\frac{\pi}{n}\right)^2}$, so that the area of the "lantern" is

$$A(m,n) = 2mn\sin\frac{\pi}{n}\sqrt{\frac{1}{m^2}+\left(1-\cos\frac{\pi}{n}\right)^2}.$$

We might hope that as m and n tend to infinity, $A(m,n) \to 2\pi$. Indeed, if we first let $n \to \infty$, and then let $m \to \infty$, we get

$$\lim_{m\to\infty}\left(\lim_{n\to\infty} A(m,n)\right) = 2\pi.$$

However, for any fixed $n > 1$,

$$\lim_{m\to\infty} A(m,n) = \infty.$$

Even if we let m and n tend to infinity together, we do not necessarily get the right answer. For instance, for any positive integer c,

$$\lim_{n\to\infty} A(cn^2,n) = 2\pi\sqrt{1+\frac{\pi^4c^2}{4}},$$

which is strictly greater than 2π.

We can try to understand the failure of this example by contrasting it with the case of curves. For curves, we designed a polygonal approximation and used the Mean Value Theorem to show that each segment of the polygonal approximation was parallel to the curve at some nearby point. In the above example, on the other hand, if we allow m to grow with n fixed, the triangles, far from becoming closer to parallel to the surface, in fact become closer to perpendicular. This example reinforces our claim in Section 4.4 that the Mean Value Theorem is really a one-variable theorem.

This first failed approach is not without its lessons. In the modified approach that follows, we will guarantee that the approximating polygons will be parallel to the surface at at least one point by considering tangent vectors.

We again assume for simplicity that we have a surface defined by a single rectangle $R = [a,b] \times [c,d]$ and a piecewise smooth function $\phi : R \to \mathbb{R}^3$. Let $P_1 = \{a = a_0 < a_1 < \cdots < a_k = b\}$ be a partition of $[a,b]$, let $P_2 = \{c = c_0 < c_1 < \cdots < c_\ell = d\}$ be a partition of $[c,d]$, and let $P = P_1 \times P_2$ be the corresponding regular partition of R (see Definition 5.5.17). For each lower-left vertex (a_i, c_j), $0 \le i < k$, $0 \le j < \ell$, we consider the associated point on the surface $x_{ij} = \phi(a_i, c_j)$ and the two tangent vectors $u_{ij} = D_1\phi(a_i, c_j)$ and $v_{ij} = D_2\phi(a_i, c_j)$. We then consider the parallelogram spanned by the vectors u_{ij} and v_{ij} at x_{ij}, that is, the parallelogram with vertices $x_{ij}, x_{ij} + u_{ij}, x_{ij} + u_{ij} + v_{ij}$, and $x_{ij} + v_{ij}$. This parallelogram is in fact tangent to the surface at the point x_{ij} by construction.

What is the area of this parallelogram? Since we have specialized to the case of surfaces in $\mathbb{R}^3$, we know from Exercise 2.5.20 that the area is the norm of the cross product of the two tangent vectors, $A_{ij} = \|u_{ij} \times v_{ij}\|$. Summing over the points of our partition, we get an approximation to the surface area given by

$$\begin{aligned} A_P &= \sum_{i=0}^{k-1}\sum_{j=0}^{\ell-1} A_{ij} \\ &= \sum_{i=0}^{k-1}\sum_{j=0}^{\ell-1} \|u_{ij} \times v_{ij}\|. \end{aligned}$$

If we let $f : R \to \mathbb{R}$ be given by $f(s,t) = \|D_1\phi(s,t) \times D_2\phi(s,t)\|$, we can easily see that $L(f,P) \le A_P \le U(f,P)$. Since ϕ is continuously differentiable except possibly at finitely many points, the function f is integrable on $\mathbb{R}$, and hence, $\lim_{\|P\|\to 0} A_P = \int_R f$. Thus, the area of surface is

$$A = \int_R \|D_1\phi \times D_2\phi\| = \int_a^b \int_c^d \|D_1\phi(s,t) \times D_2\phi(s,t)\|\,dt\,ds.$$

Example 6.4.13 Let $R = [0,1] \times [0,2\pi]$, and let $\phi : R \to \mathbb{R}^3$ be given by $\phi(s,t) = (s\cos t, s\sin t, 1-s)$. This is the parametrization of a cone from Example 6.4.4. To find the surface area, we first compute the derivative of ϕ.

$$\begin{aligned} D_1\phi(s,t) &= (\cos t, \sin t, -1), \\ D_2\phi(s,t) &= (-s\sin t, s\cos t, 0). \end{aligned}$$

So

$$D_1\phi(s,t) \times D_2\phi(s,t) = (s\cos t, s\sin t, s\cos^2 t + s\sin^2 t) = (s\cos t, s\sin t, s),$$

and hence

$$\|D_1\phi(s,t) \times D_2\phi(s,t)\| = \sqrt{s^2\cos^2 t + s^2\sin^2 t + s^2} = \sqrt{2s^2} = \sqrt{2}s.$$

Thus, the surface area of this cone is

$$A = \int_0^1 \int_0^{2\pi} \sqrt{2}s \, dt \, ds = \sqrt{2}\pi.$$

Exercise 6.4.14 Use the parametrization from Example 6.4.9 to compute the surface area of the unit sphere in $\mathbb{R}^3$.

Exercise 6.4.15 Use your parametrization from Exercise 6.4.11.*ii* to compute the surface area of the torus T in $\mathbb{R}^3$.

Exercise 6.4.16

i. Let

$$S = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + z^2 = 1, y^2 + z^2 = 1\}.$$

Find a parametrization of S and compute the surface area.

ii. Let

$$S' = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + z^2 = 1, y^2 + z^2 = 1, x^2 + y^2 = 1\}.$$

Find a parametrization of S' and compute the surface area.

Exercise 6.4.17 Let $R = [a,b] \times [c,d] \subset \mathbb{R}^2$ be a rectangle, let $f : R \to \mathbb{R}$ be a C^1 function, and let $S = \{(s,t,f(s,t)) \mid (s,t) \in R\} \subset \mathbb{R}^3$. Show that the surface area of S is

$$A = \int_R \sqrt{1 + (D_1 f(s,t))^2 + (D_2 f(s,t))^2}.$$

Exercise 6.4.18 Let $\phi : [a,b] \to \mathbb{R}^3$ be a curve whose path lies in the half of the xz-plane with positive x-coordinate. Let S be the surface obtained by revolving C_ϕ about the z-axis. Show that the surface area of S is

$$A = 2\pi \int_a^b \phi_1(t)\|\phi'(t)\| \, dt = \int_\phi (2\pi\phi_1(t))\lambda.$$

We may use these same ideas to find a formula for the area of a surface in $\mathbb{R}^n$ for $n > 3$. However, the formula we used for the area of a parallelogram depended on the peculiar existence of the cross product in $\mathbb{R}^3$. We will instead refer back to more elementary considerations and write the area of the parallelogram spanned by vectors $\mathbf{u}$ and $\mathbf{v}$ as $\|\mathbf{u}\|\|\mathbf{v}\|\sin\theta$, where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$. By Theorem 2.5.16, $\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$, so we can rewrite the area of the parallelogram as

$$\begin{aligned} \|\mathbf{u}\|\|\mathbf{v}\| \sin\left(\cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right)\right) &= \|\mathbf{u}\|\|\mathbf{v}\|\sqrt{1 - \left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right)^2} \\ &= \sqrt{\|\mathbf{u}\|^2\|\mathbf{v}\|^2 - (\mathbf{u}\cdot\mathbf{v})^2}. \end{aligned}$$

By the same considerations as in $\mathbb{R}^3$, the area of a surface in $\mathbb{R}^n$ parametrized by a function $\phi : R \to \mathbb{R}^n$ can be approximated by the area of parallelograms that are tangent to points at the surface and given exactly as

$$A = \int_R \sqrt{\|D_1\phi\|^2 \cdot \|D_2\phi\|^2 - (D_1\phi \cdot D_2\phi)^2}.$$

Exercise 6.4.19 Use your parametrization from Exercise 6.4.11.i to compute the area of the torus T^2 in $\mathbb{R}^4$.

6.5 Differential 2-Forms

In the previous section, we defined surfaces in $\mathbb{R}^n$ and devised a means of computing their surface areas. In this section, we will see that, just as the length of a curve was the integral of a special differential 1-form on the curve, the area of a surface can be viewed as the integral of a "differential 2-form" on the surface. This viewpoint will lead us to an analogue of Green's Theorem for surfaces in $\mathbb{R}^n$, as we suggested at the end of Section 6.3.

Just as a differential 1-form is designed to measure a single vector at each point in its domain, a differential 2-form is designed to measure a pair of vectors. Thus, at the infinitesimal level, the integration of a differential 1-form along a curve involved applying a linear form to a single vector, namely, a tangent vector to the curve. Similarly, the integration of a differential 2-form over a surface will involve applying an alternating bilinear form to a pair of vectors that span the tangent plane to the surface.

The prototypical example of an alternating bilinear form is the one in $\mathbb{R}^2$ that assigns to an ordered pair of vectors the "signed area" of the parallelogram spanned by those two vectors. Given two vectors $u = (a, b)$ and $v = (c, d)$ in $\mathbb{R}^2$, recall from Exercise 5.9.14 that the area of the parallelogram spanned by u and v is $|ad - bc|$. Of course, the map that assigns the scalar $|ad - bc|$ to the pair (u, v) is not bilinear, but the closely related map $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by $L(u, v) = ad - bc$ is. Notice that while the original formula did not depend on the order of the two vectors, this new signed area does. If we switch u and v, then $L(v, u) = bc - ad = -L(u, v)$, and this is what we mean by "alternating."

We have already defined bilinear forms in Chapter 2, but we restate the definition here for convenience.

Definition 6.5.1 A map $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a *bilinear form* if it satisfies the following conditions.

1. $L(a_1v_1 + a_2v_2, w) = a_1L(v_1, w) + a_2L(v_2, w), \quad \forall a_1, a_2 \in \mathbb{R}, \forall v_1, v_2, w \in \mathbb{R}^n$
2. $L(v, b_1w_1 + b_2w_2) = b_1L(v, w_1) + b_2L(v, w_2), \quad \forall b_1, b_2 \in \mathbb{R}, \forall v, w_1, w_2 \in \mathbb{R}^n$

Exercise 6.5.2 Decide whether or not the following maps are bilinear forms.

1. $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by $L((x_1, y_1), (x_2, y_2)) = x_1 + x_2 + y_1 + y_2$
2. $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by $L((x_1, y_1), (x_2, y_2)) = x_1x_2$
3. $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by $L((x_1, y_1), (x_2, y_2)) = x_1x_2 + y_1y_2$
4. $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by $L((x_1, y_1), (x_2, y_2)) = x_1y_2 - x_2y_1$
5. $L : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ given by $L((x_1, y_1, z_1), (x_2, y_2, z_2)) = x_1y_2 - x_2y_1$

Definition 6.5.3 A bilinear form $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is said to be *alternating* if

$$L(v, w) = -L(w, v), \quad \forall v, w \in \mathbb{R}^n.$$

We denote by $\mathcal{A}_n$ the collection of alternating bilinear forms on $\mathbb{R}^n$.

Exercise 6.5.4 Of the maps from the previous exercise which are bilinear forms, determine which are alternating.

Exercise 6.5.5 Show that the collection $\mathcal{A}_n$ of alternating bilinear forms on $\mathbb{R}^n$ is a real vector space.

Remark 6.5.6 Note that our original bilinear forms of interest from Chapter 2, namely inner products, are not alternating, but rather symmetric. In fact, every bilinear form can be uniquely written as a sum of an alternating form and a symmetric form, so we are in some sense working with the complementary category of forms here.

Exercise 6.5.7 Show that if $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is an alternating bilinear form, then there exists a scalar $c \in \mathbb{R}$ such that $L((x_1, y_1), (x_2, y_2)) = c(x_1y_2 - x_2y_1)$.

Exercise 6.5.8 Show that $\dim(\mathcal{A}_n) = \dfrac{n(n-1)}{2}$. (Hint: Construct a basis using the form from the previous exercise as a template.)

This vector space $\mathcal{A}_n$ is the 2-dimensional analogue of the vector space $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ of linear forms on $\mathbb{R}^n$, which are the infinitesimal building blocks for measuring 1-dimensional aspects of Euclidean space. In other words, the elements of $\mathcal{A}_n$ will be the infinitesimal building blocks for measuring 2-dimensional aspects of Euclidean space. It should not be surprising, then, that we can combine two linear forms to obtain a bilinear form.

Definition 6.5.9 Let $S, T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$. The *wedge product* of S and T is the alternating bilinear form $S \wedge T : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by $(S \wedge T)(v, w) = S(v)T(w) - T(v)S(w)$.

Exercise 6.5.10

i. Verify that $S \wedge T$ is an alternating bilinear form.

ii. Show that $T \wedge S = -S \wedge T$.

iii. Show that $S \wedge S = 0$.

Exercise 6.5.11 Show that the bilinear form $L : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ from Exercise 6.5.7 given by $L((x_1, y_1), (x_2, y_2)) = c(x_1y_2 - x_2y_1)$ is the wedge product of two linear forms.

Exercise 6.5.12 Find a basis for $\mathcal{A}_n$ consisting of wedge products of linear forms.

We are now ready to define differential 2-forms.

Definition 6.5.13 Let $\Omega \subseteq \mathbb{R}^n$. A *differential 2-form* on Ω is a map $\omega : \Omega \to \mathcal{A}_n$, where $\mathcal{A}_n$ is the collection of alternating bilinear forms on $\mathbb{R}^n$. The bilinear map in $\mathcal{A}_n$ associated with the point $x \in \Omega$ is denoted ω_x.

Remark 6.5.14 Because $\mathcal{A}_n$ is a finite dimensional real vector space of dimension $k = \frac{n(n-1)}{2}$, we can pick a basis to identify it with $\mathbb{R}^k$. We then give $\mathbb{R}^k$ the usual metric. This allows us to define what it means for a differential 2-form $\omega : \Omega \to \mathcal{A}_n$ to be continuous, differentiable, smooth, etc.

Example 6.5.15 Let $L \in \mathcal{A}_n$. The map $\omega : \mathbb{R}^n \to \mathcal{A}_n$ given by $\omega_x = L$ is a constant differential 2-form.

Example 6.5.16 Let $\phi : [a, b] \times [c, d] \to \mathbb{R}^n$ be a simple smooth parametrized surface in $\mathbb{R}^n$ with image S. The *area form* of ϕ is the differential 2-form $\alpha : S \to \mathcal{A}_n$ given by

$$\alpha_{\phi(s,t)} = \frac{D_1\phi(s,t)}{\|D_1\phi(s,t)\|} \wedge \frac{D_2\phi(s,t)}{\|D_2\phi(s,t)\|},$$

that is,

$$\alpha_{\phi(s,t)}(v, w) = \frac{D_1\phi(s,t)(v)}{\|D_1\phi(s,t)\|}\frac{D_2\phi(s,t)(w)}{\|D_2\phi(s,t)\|} - \frac{D_2\phi(s,t)(v)}{\|D_2\phi(s,t)\|}\frac{D_1\phi(s,t)(w)}{\|D_1\phi(s,t)\|}.$$

The above example shows that just as two linear forms can be combined using the wedge product to give an alternating bilinear form, two differential 1-forms can be combined pointwise using the wedge product to give a differential 2-form. In fact, this is the first of two essential methods for constructing differential 2-forms.

Definition 6.5.17 Let $\Omega \subseteq \mathbb{R}^n$, and let $\omega, \eta : \Omega \to \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ be differential 1-forms. The *wedge product* of ω and η, denoted $\omega \wedge \eta$, is the differential 2-form defined by

$$(\omega \wedge \eta)_x(v, w) = \omega_x(v)\eta_x(w) - \omega_x(w)\eta_x(v).$$

Example 6.5.18 Recall that the differential 1-forms dx and dy on $\mathbb{R}^2$ are given by $dx_a(v) = v_1$ and $dy_a(v) = v_2$ for $a \in \mathbb{R}^2$ and $v = (v_1, v_2) \in \mathbb{R}^2$. The differential 2-form $dx \wedge dy$ on $\mathbb{R}^2$ is thus given by

$$(dx \wedge dy)_a(v, w) = v_1 w_2 - v_2 w_1,$$

where $v = (v_1, v_2) \in \mathbb{R}^2$ and $w = (w_1, w_2) \in \mathbb{R}^2$. Note that this is the constant form corresponding to the alternating bilinear form L from Exercise 6.5.7 with $c = 1$.

Exercise 6.5.19 Use Exercise 6.5.7 to show that every bilinear form ω on a set $\Omega \subseteq \mathbb{R}^2$ is of the form $\omega = f\, dx \wedge dy$, where $f : \Omega \to \mathbb{R}$ is a function.

Exercise 6.5.20 Recall that the differential 1-form dx_i on $\mathbb{R}^n$ is given by $(dx_i)_a(v) = v_i$, where $a \in \mathbb{R}^n$ and $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$. Use Exercise **??** to show that every bilinear form ω on a set $\Omega \subseteq \mathbb{R}^n$ can be written as

$$\omega = \sum_{1 \leq i < j \leq n} f_{ij} dx_i \wedge dx_j,$$

where the $f_{ij} : \Omega \to \mathbb{R}$ are functions.

Exercise 6.5.21 Compute the given wedge product.

i. $(x\, dx + y\, dy) \wedge (y\, dx - x\, dy)$

ii. $(y \cos x\, dx - x \sin x\, dy) \wedge (y \sin x\, dx + x \cos x\, dy)$

iii. $(x^2 y\, dx + xyz\, dy + xy^2\, dz) \wedge (xyz\, dx + yz^2\, dy + dz)$

We are now prepared to integrate differential 2-forms. For simplicity, we make the following definition for the integral of a differential 2-form on a smooth surface. It is clear how to extend the definition to piecewise smooth surfaces.

Definition 6.5.22 Let $\phi : [a, b] \times [c, d] \to \mathbb{R}^n$ be a smooth parametrized surface with image S. Let $\omega : S \to \mathcal{A}_n$ be a continuous differential 2-form on S. The *integral* of ω over ϕ is defined to be

$$\int_\phi \omega = \int_a^b \int_c^d \omega_{\phi(s,t)}(D_1\phi(s, t), D_2\phi(s, t))\, dt\, ds.$$

We first show that this integral is independent of the parametrization of S.

Proposition 6.5.23 Let $R = [a, b] \times [c, d]$ and $R' = [a', b'] \times [c', d']$ be rectangles in $\mathbb{R}^2$. Suppose $\phi : R \to \mathbb{R}^n$ and $\psi : R' \to \mathbb{R}^n$ are equivalent parametrizations of the surface S. Let $\omega : S \to \mathcal{A}_n$ be a continuous differential 2-form. Then $\int_\phi \omega = \int_\psi \omega$.

Proof. Let $\gamma : R \to R'$ be the C^1 homeomorphism such that $\det D\gamma > 0$ and $\psi \circ \gamma = \phi$. Using the chain rule, we can write

$$\begin{aligned}
\int_\phi \omega &= \int_a^b \int_c^d \omega_{\phi(s,t)}(D_1\phi(s,t), D_2\phi(s,t))\, dt\, ds \\
&= \int_a^b \int_c^d \omega_{\psi(\gamma(s,t))}(D_1(\psi \circ \gamma)(s,t), D_2(\psi \circ \gamma)(s,t))\, dt\, ds \\
&= \int_a^b \int_c^d \omega_{\psi(\gamma(s,t))}(D_1\psi(\gamma(s,t)) \cdot D_1\gamma_1(s,t) + D_2\psi(\gamma(s,t)) \cdot D_1\gamma_2(s,t), \\
&\qquad\qquad D_1\psi(\gamma(s,t)) \cdot D_2\gamma_1(s,t) + D_2\psi(\gamma(s,t)) \cdot D_2\gamma_2(s,t))\, dt\, ds.
\end{aligned}$$

For notational simplicity, we let $L = \omega_{\psi(\gamma(s,t))}$, $v_1 = D_1\psi(\gamma(s,t))$, $v_2 = D_2\psi(\gamma(s,t))$, and $a_{ij} = D_i\gamma_j$ for $1 \leq i, j, \leq 2$. By the definition of an alternating bilinear form, we can write the integrand as

$$\begin{aligned}
L(a_{11}v_1 + a_{12}v_2, a_{21}v_1 + a_{22}v_2) &= L(a_{11}v_1, a_{21}v_1) + L(a_{11}v_1, a_{22}v_2) + L(a_{12}v_2, a_{21}v_1) + L(a_{12}v_2, a_{22}v_2) \\
&= L(a_{11}v_1, a_{22}v_2) - L(a_{21}v_1, a_{12}v_2) \\
&= (a_{11}a_{22} - a_{21}a_{12})L(v_1, v_2) \\
&= \det D\gamma(s,t) L(v_1, v_2).
\end{aligned}$$

Since we assumed $\det D\gamma > 0$, it follows that $\det D\gamma = |\det D\gamma|$. Thus, by change of variables,

$$\begin{aligned}
\int_\phi \omega &= \int_a^b \int_c^d \omega_{\psi(\gamma(s,t))}(D_1\psi(\gamma(s,t)), D_2\psi(\gamma(s,t))) \cdot |\det D\gamma(s,t)|\, dt\, ds \\
&= \int_{a'}^{b'} \int_{c'}^{d'} \omega_{\psi(s',t')}(D_1\psi(s',t'), D_2\psi(s',t'))\, dt'\, ds' \\
&= \int_\psi \omega.
\end{aligned}$$

Remark 6.5.24 Since the integral of a differential 2-form is independent of the parametrization, we will often write $\int_S \omega$ for $\int_\phi \omega$ when ϕ parametrizes S.

Exercise 6.5.25

i. Let $\phi : [a,b] \times [c,d] \to \mathbb{R}^n$ be a simple smooth parametrized surface in $\mathbb{R}^n$. Let α be the area form of ϕ defined in Example 6.5.16. Show that $\int_\phi \alpha$ is the surface area of ϕ.

ii. Show that the area form is independent of parametrization, so that the surface area is independent of parametrization.

Example 6.5.26 Let $a, b, c > 0$, and let $\phi : [0,a] \times [0,b] \to \mathbb{R}^3$ be given by $\phi(s,t) = (s, t, c - \frac{c}{b}t)$. This parametrizes the rectangle S in $\mathbb{R}^3$ whose vertices are $(0,0,c)$, $(a,0,c)$, $(a,b,0)$, and $(0,b,0)$. We have $D_1\phi(s,t) = (1,0,0)$, and $D_2\phi(s,t) = (0,1,-\frac{c}{b})$.

We can compute

$$\begin{aligned}
\int_S dx \wedge dy &= \int_0^a \int_0^b (dx \wedge dy)\left((1,0,0), \left(0,1,-\frac{c}{b}\right)\right)\, dt\, ds \\
&= \int_0^a \int_0^b (1 \cdot 1 - 0 \cdot 0)\, dt\, ds \\
&= ab.
\end{aligned}$$

This is the area of the projection of the rectangle S onto the xy-plane, and this makes sense because $dx \wedge dy$ should measure the "xy-ness" of the surface.

Exercise 6.5.27 Let S be the rectangle from the above example. Compute the following integrals.

i. $\int_S dx \wedge dz$

ii. $\int_S dy \wedge dz$

iii. $\int_S dy \wedge dx$

iv. $\int_S \alpha$

As the example and exercise above show, different differential 2-forms can be used to measure different 2-dimensional aspects of a surface.

Exercise 6.5.28 Let $S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1, z \geq 0\}$ be the upper hemisphere of the unit sphere in $\mathbb{R}^3$. For each of the following integrals, first predict what the integral will be, and then do the computation to verify your prediction.

i. $\int_S dx \wedge dy$

ii. $\int_S dy \wedge dx$

iii. $\int_S dx \wedge dz$

iv. $\int_S \alpha$

v. $\int_S \sqrt{1 - x^2 - y^2}\, dx \wedge dy$

Chapter A

Sets and Functions

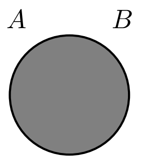
> *Dans la présente Note, on va essayer de préciser une terminologie propre à l'étude des ensembles abstraits. Cette étude a pour but principal d'étendre les propriétés des ensembles linéaires à des ensembles de plus en plus généraux, et par voie de conséquence, de disséquer ces propriétés et d'en rechercher pour chacune la véritable origine. Il en résulte que le fond des démonstrations est pour ainsi dire donné d'avance et que la difficulté consiste en grande partie a préciser dans quel cadre elles viendront se placer. Adopter une terminologie, c'est donc tracer d'avance toute la théorie. Au fur et à mesure du développement de cette théorie, la terminolgie a varié et variera encore. Mais il n'est peut-être pas inutile d'en proposer une, adaptée à la situation présente.*
>
> *– M. M. Fréchet,*
>
> *Extrait des Comptes rendus du Congrès des Sociétés savantes en 1924.*

A.1 Sets and Elements

You are probably familiar with the notion of a set as a "collection", or a "bunch", or maybe even a "set" of objects. Formally, we begin our discussion of sets with two undefined terms, that is, "set" and "membership in a set." So we might say that a set is a thing which is a collection of other things called the elements of the set. In practice, this sort of "definition by synonym" suffices for most mathematicians. If A is a set, we write $x \in A$ to denote membership in A, and we say that *x is an element of the set A*.

A.2 Equality, Inclusion, and Notation

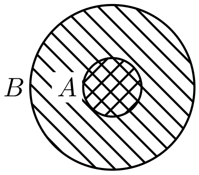
If A and B are sets, we say that $A = B$ (*A is equal to B*) if they have the same elements. That is, if $x \in A$, then $x \in B$, and conversely if $x \in B$, then $x \in A$. This is used to prove that two sets, which at first glance might not appear to be equal, are indeed equal. Examples of this are given below, and many more are in the exercises. There is one very special set that plays an important role – the *empty set* which contains no elements. The empty set is denoted $\varnothing$.



$A = B$

$\varnothing$

The next idea is that of a subset. We say that A is a *subset* of B if for any $x \in A$, we have $x \in B$. If A is a subset of B, we write $A \subseteq B$. We also say that B *contains* A (some people even say that B is a *superset* of A, but that is the only time you will see that word in this book). When $A \subseteq B$, it may be the case that B is actually "bigger" than A, that is, there is an element $b \in B$ such that b is not in A, or symbolically, $b \notin A$. In this case, we say that A is a *proper subset* of B, and, if we wish to emphasize this, we write $A \subsetneq B$. However, keep in mind that when we write $A \subseteq B$, A can certainly be a proper subset of B.



$A \subsetneq B$

Exercise A.2.1 If A is a set, show that $A \subseteq A$.

Exercise A.2.2 If A and B are sets, show that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

Exercise A.2.3 Suppose that A, B, and C are sets. If $A \subseteq B$ and $B \subseteq C$, show that $A \subseteq C$.

Exercise A.2.4 Show that if A is a set, then $\varnothing \subseteq A$.

To be fair, we should observe that all of this is a bit fuzzy logically and may even seem to be tautological. Nonetheless, if you assume the appropriate properties for the symbol $\in$, and you practice enough, you will feel comfortable with this whole business.

There are two *quantifiers* which we use regularly throughout this book. The first is $\forall$ which reads "for all," and the second is $\exists$ which reads "there exists." Also, the phrase "such that" will often be replaced by the symbol $\ni$, and we abbreviate the phrase "if and only if" by iff.

We usually just assume that all of the sets we consider are contained in some "big" set which is large enough to include all the objects we need. This big set or *universal set* is often denoted by the symbol X. Nevertheless, it is possible for a set to be "too big" (see Section A.8). When a quantifier appears without a domain, as in the definition of equality, we mean to consider all objects in the current universe as our domain. Don't get the mistaken idea that the elements of the universal set X must all look the "same". For example, X can contain equilateral polygons, purple Buicks, real numbers, fried green tomatoes, etc.

There is an abbreviated notation for the set of all objects $x \in X$ that satisfy some condition $P(x)$. This notation means that P is a proposition which is either true or false depending on the value of x. For the set of all $x \in X$ such that $P(x)$ is true, we write $\{x \in X \mid P(x)\}$. We may write simply $\{x \mid P(x)\}$ which again is meant to imply that we take all x from some designated universe. For example, "the set x such that x is an even integer" is not sufficiently precise about the universe. It would be better to say, for example, "x in the real numbers such that x is an even integer."

There will be cases when we list the elements of a set. If the set is small enough, for instance, the set of the first five letters of the alphabet, we write $A = \{a, b, c, d, e\}$. If the set is very large (maybe even infinite), but there is no ambiguity, we may simply list the first few elements of the set and describe the set to the reader. For example, we write the natural numbers as $\mathbb{N} = \{1, 2, 3, 4, \ldots, n, \ldots\}$. This familiar "dot, dot, dot" signifies that you should use your brain and continue as indicated.

A.3 The Algebra of Sets

This section is about taking subsets of a universal set X and putting them together in different ways to create new subsets of X. In fact, that's what most of mathematics is all about, building new things from old things. As was the case in Sections A.1 and A.2, most students will have seen this material, so let's cut right to the chase.

Definition A.3.1 Let A and B be sets. The *union* of A and B, denoted $A \cup B$, is defined by

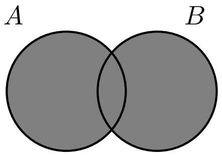
$$A \cup B = \{x \in X \mid x \in A \text{ or } x \in B\}.$$

Note that the "or" in this definition is inclusive (as opposed to "either–or"). That is, even if x is an element of both A and B, then x is still an element of $A \cup B$.

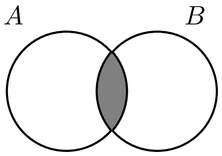
Definition A.3.2 Let A and B be sets. The *intersection* of A and B, denoted $A \cap B$, is defined by

$$A \cap B = \{x \in X \mid x \in A \text{ and } x \in B\}.$$

Here, there is no doubt about what "and" means. "And" is "and."



$A \cup B$

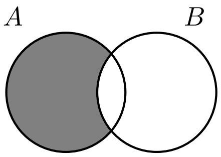


$A \cap B$

Definition A.3.3 Let A and B be sets. We say that A and B are *disjoint* if $A \cap B = \varnothing$. If $\mathcal{C}$ is a collection of sets, any pair of which are disjoint, then the elements of $\mathcal{C}$ are said to be *pairwise disjoint*.

Definition A.3.4 Let A and B be sets. The *difference* of A and B, denoted $A \setminus B$, and read "A minus B", is defined by

$$A \setminus B = \{x \in A \mid x \notin B\}.$$



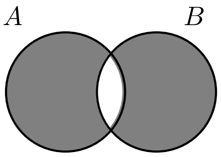
$A \setminus B$

At this point, it is useful to remark that union and intersection are obviously commutative, that is $A \cup B = B \cup A$ and $A \cap B = B \cap A$. However, difference is not commutative. For example, let $A = \{a\}$ and $B = \varnothing$. The reader may find it amusing to experiment with the difference of various pairs of sets.

The cure for the non-commutativity of the difference is provided by the symmetric difference.

Definition A.3.5 Let A and B be sets. The *symmetric difference* of A and B, denoted $A \triangle B$, is defined by

$$A \triangle B = (A \setminus B) \cup (B \setminus A).$$

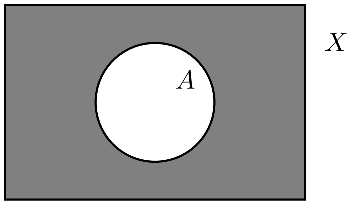


$A \triangle B$

Obviously, the symmetric difference is a commutative operation. Finally, we define the complement of a subset A of a universal set X.

Definition A.3.6 Let $A \subseteq X$. The *complement* of A (in X), denoted ${}^cA$, is defined by

$$ {}^cA = X \setminus A. $$

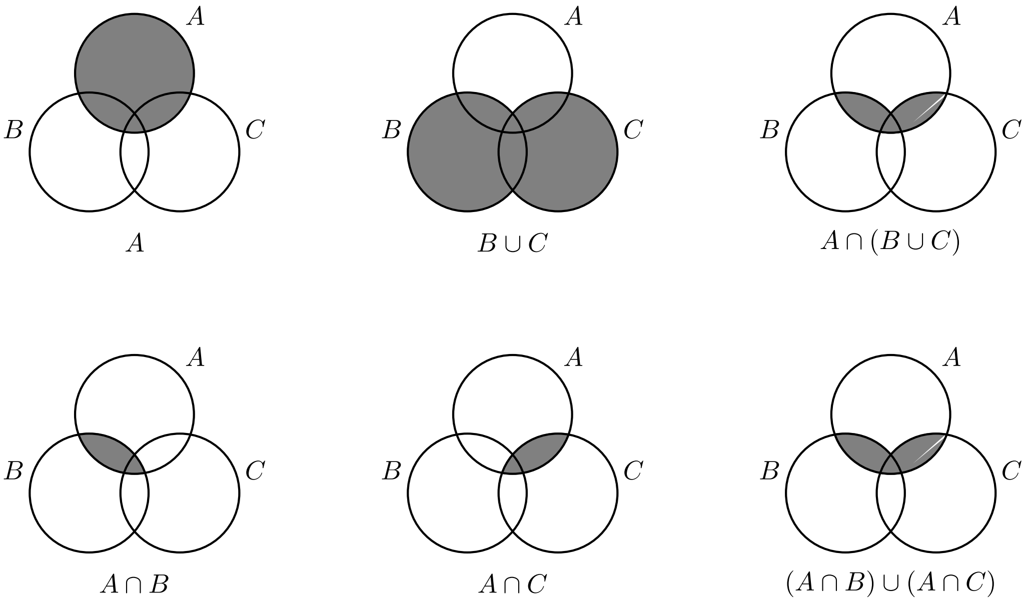


${}^cA$

There are many identities among sets that result from using the above operations. We illustrate a couple and then assign a multitude of problems for practice.

Example A.3.7 This example shows that intersection is distributive over union. That is $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Take $x \in X$. Then $x \in A \cap (B \cup C)$ iff $x \in A$ and $x \in B \cup C$ iff ($x \in A$ and $x \in B$) or ($x \in A$ and $x \in C$). Now this means that $x \in A \cap B$ or $x \in A \cap C$, that is $x \in (A \cap B) \cup (A \cap C)$. Notice that in this proof, we simply replace symbols by words and use the common understandings of these words.



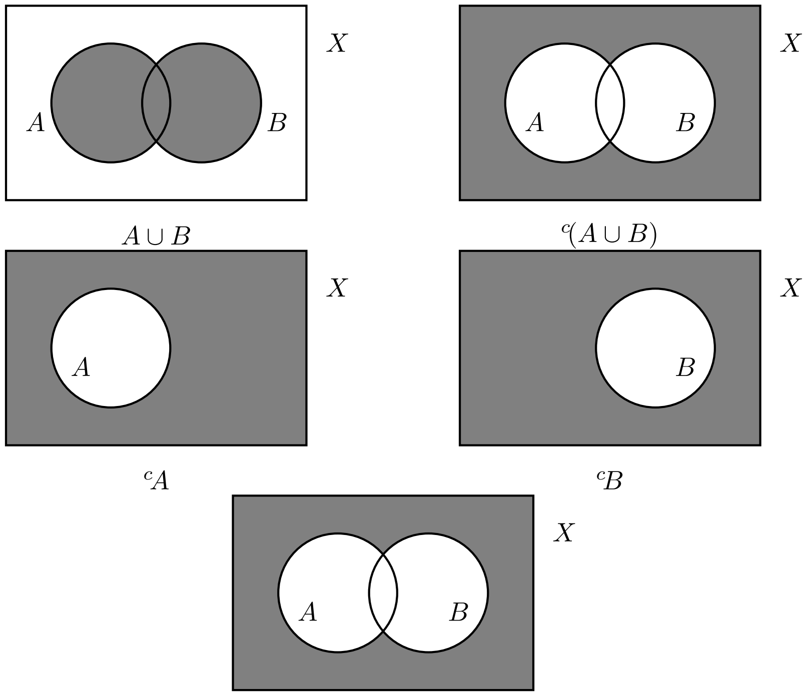
A $\quad$ $B \cup C$ $\quad$ $A \cap (B \cup C)$

$A \cap B$ $\quad$ $A \cap C$ $\quad$ $(A \cap B) \cup (A \cap C)$

Illustration of the proof of the example

Example A.3.8 (DeMorgan's Law I) ${}^c(A \cup B) = {}^cA \cap {}^cB$

Take $x \in X$. Then $x \in {}^c(A \cup B)$ iff $x \in X \setminus (A \cup B)$ iff $x \in X$ and $x \notin A \cup B$ iff ($x \in X$ and $x \notin A$) and ($x \in X$ and $x \notin B$) iff $x \in {}^cA$ and $x \in {}^cB$ iff $x \in {}^cA \cap {}^cB$.



$A \cup B$

${}^{c}(A \cup B)$

${}^{c}A$

${}^{c}B$

${}^{c}A \cap {}^{c}B$

Illustration of the proof of the example.

As the book progresses, occasionally we will need to show, for certain pairs of sets A and B, that $A \subseteq B$ or perhaps that $A = B$. Such inclusions and equalities can be difficult to prove. The following list of exercises will help the reader to develop skills in this direction.

Exercise A.3.9 Prove the following equalities. As in A.3.7 and A.3.8, diagrams will illuminate the situation, but they will not suffice for proof. The sets A, B, C are subsets of some universe X.

i. $A \cup (B \cup C) = (A \cup B) \cup C$ (Associative law for union).

ii. $A \cap (B \cap C) = (A \cap B) \cap C$ (Associative law for intersection).

iii. $A \triangle (B \triangle C) = (A \triangle B) \triangle C$ (Associative law for symmetric difference).

iv. $A \cup \varnothing = A$ (The empty set is an identity for union).

v. $A \triangle \varnothing = A$ (The empty set is an identity for symmetric difference).

vi. $A \cap X = A$ (The universe is an identity for intersection).

vii. $A \cup B = \varnothing$ iff $A = \varnothing$ and $B = \varnothing$.

viii. $A \cap B = X$ iff $A = X$ and $B = X$.

ix. $A \triangle B = \varnothing$ iff $A = B$.

x. $A \cap (B \triangle C) = (A \cap B) \triangle (A \cap C)$ (Distributive law of intersection over symmetric difference).

xi. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributive law of union over intersection).

xii. ${}^{c}(A \cap B) = {}^{c}A \cup {}^{c}B$ (DeMorgan's Law II).

It is obvious that the concepts of union and intersection can be extended to any finite number of sets $A_1, A_2, \ldots, A_n$. These are written $\cup_{i=1}^{n} A_i$ and $\cap_{i=1}^{n} A_i$. More generally, we can extend union and intersection to any collection of sets. We select an *index set* I, and write $\cup_{i \in I} A_i$ and $\cap_{i \in I} A_i$. Here, I could be finite or infinite. This is discussed later in the chapter.

A.4 Cartesian Products, Counting, and Power Sets

The Cartesian product of two sets may be familiar from the example of the coordinate or Euclidean plane. This is the set of all pairs (x, y) of real numbers where x denotes the "first coordinate" and y denotes the "second coordinate." The symbol (x, y) is called an *ordered pair*. This is not equal to the ordered pair (y, x), unless $y = x$. That is, the position of the coordinates makes a difference. For example, in the coordinate plane, the point $(1, 2)$ is not the same as the point $(2, 1)$, whereas the sets $\{1, 2\}$ and $\{2, 1\}$ are the same since they have the same elements and order is irrelevant. There is a formal definition of ordered pair, namely $(a, b) = \{\{a\}, \{a, b\}\}$. We are more concerned with a working principle. We say that two ordered pairs (x, y) and (x', y') are equal iff $x = x'$ and $y = y'$.

Definition A.4.1 Let A and B be sets. The *Cartesian product* of A and B, denoted $A \times B$, is defined by

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

Thus, the Cartesian product of A and B is the set of all ordered pairs in which the first coordinate comes from the set A, and the second coordinate comes from the set B. Notice that we have created a new universal set of ordered pairs of elements of X. This will cause no difficulty for the moment.

Example A.4.2 If $A = \{a, b, c\}$ and $B = \{1, 2, 3\}$, then $A \times B = \{(a, 1), (a, 2),$
$(a, 3), (b, 1), (b, 2), (b, 3), (c, 1), (c, 2), (c, 3)\}$.

Exercise A.4.3 Write out $B \times A$ where A and B are as in the example above. Observe that the elements of $B \times A$ are different from those of $A \times B$.

Exercise A.4.4 Prove that $A \times \varnothing = \varnothing \times A = \varnothing$.

Exercise A.4.5 Suppose $A \neq \varnothing$ and $B \neq \varnothing$. Show $A \times B = B \times A$ iff $A = B$.

There is a fundamental counting principle that accompanies the Cartesian product.

Theorem A.4.6 (Fundamental Counting Principle) If A has m elements and B has n elements, then $A \times B$ has mn elements.

This is simple to prove by drawing little trees or using some other artifice. A formal proof by induction is very straightforward and will be given as an exercise later in the chapter. This counting principle is the basis for most of the combinatorial formulas in finite probability theory. We will have occasion to use this formula in only a few instances since most of the sets with which we deal in analysis have an infinite number of elements.

We discuss the terms "finite set" and "infinite set" in Section A.8. We denote the number of elements in a set A by ${}^{\#}A$. So our fundamental counting principle says

$${}^{\#}(A \times B) = ({}^{\#}A)({}^{\#}B).$$

To generalize this fundamental counting principle to a finite number of finite sets, we must first define the Cartesian product of these sets. Suppose that $A_1, A_2, \ldots, A_n$ are subsets of X.

Definition A.4.7 The *n-fold Cartesian product* is $A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) \mid a_j \in A_j \text{ for } 1 \leq j \leq n\}$. This is the set of ordered n-tuples, with each coordinate coming from the appropriate subset.

Exercise A.4.8 If A_1 has k_1 elements, A_2 has k_2 elements, $\ldots$, A_n has k_n elements, show that ${}^{\#}(A_1 \times A_2 \times \cdots \times A_n) = ({}^{\#}A_1)\ ({}^{\#}A_2)\ \cdots\ ({}^{\#}A_n) = k_1 k_2 \cdots k_n$. Hint: this can be proved drawing pictures but a formal proof is better.

Another counting principle has to do with the union of two sets. When counting the number of elements in $A \cup B$, we cannot simply add $^{\#}A$ and $^{\#}B$ since the intersection might be non-empty (that is, not the empty set), so we would be counting the number of elements in the intersection twice.

Exercise A.4.9 Inclusion-Exclusion Principle

i. If A and B are finite sets and $A \cap B = \varnothing$, show that

$$^{\#}(A \cup B) = {}^{\#}A + {}^{\#}B.$$

ii. If A and B are finite sets, show that

$$^{\#}(A \cup B) = {}^{\#}A + {}^{\#}B - {}^{\#}(A \cap B).$$

iii. Do it for three sets; that is if A,B, and C are finite sets, show that

$$^{\#}(A \cup B \cup C) = {}^{\#}A + {}^{\#}B + {}^{\#}C - {}^{\#}(A \cap B) - {}^{\#}(A \cap C) - {}^{\#}(B \cap C) + {}^{\#}(A \cap B \cap C).$$

iv. Generalize the previous exercise to any finite number of finite sets.

The next thing to look at is the collection of all subsets of a given set. The idea here is to start with a universe X and study all the subsets of X.

Exercise A.4.10

i. Let $X = \varnothing$. Write a list of the subsets of X. (If your list doesn't contain any elements, try again.)

ii. Let $X = \{1\}$. Write a list of the subsets of X.

iii. Let $X = \{1, 2\}$. Write a list of the subsets of X.

iv. On the basis of this information, make a conjecture about the number of subsets of a set with n elements.

Definition A.4.11 Let X be a set. The *power set* of X, denoted $\wp(X)$, is the collection of all subsets of X.

Here is the counting principle that goes with $\wp(X)$.

Theorem A.4.12 If X is a set with n elements, then $\wp(X)$ has 2^n elements.

Proof. Enumerate the elements of X: $x_1, x_2, \ldots, x_n$. Given a subset A of X, we construct a sequence $c_1, c_2, \ldots, c_n$ of 0's and 1's as follows. Let $c_i = 1$ if $x_i \in A$ and $c_i = 0$ if $x_i \notin A$. Thus, the subset A corresponds to a unique sequence of length n consisting of 0's and 1's. Similarly, given a sequence of 0's and 1's of length n, one can construct a unique subset of X. But how many sequences of length n are there consisting of 0's and 1's? By the fundamental counting principle, there are 2^n such sequences.

A.5 Some Sets of Numbers

The set of natural numbers is the familiar collection $\mathbb{N} = \{1, 2, 3, \ldots, n, \ldots\}$. It would be possible to rigorously develop the properties of the natural numbers deriving from the Peano postulates. We choose not to do this here, but we refer the interested reader to [La].

We do wish to take a more formal approach towards another familiar set of numbers, namely the integers. The integers form the collection $\{0, 1, -1, 2, -2, \ldots\}$, which we study in elementary arithmetic. We denote the integers by the symbol $\mathbb{Z}$ (from the German word *Zahlen*). The operations in the integers are addition $(+)$ and multiplication $(\cdot)$, and here are the rules. We expect that the reader is well versed in the arithmetic of the integers, but we are stating these properties explicitly for two reasons. First, these properties are used in arithmetic from the earliest grades, but are seldom justified. Second, these properties will be used to describe other algebraic structures that we will meet later.

Rules of Arithmetic in $\mathbb{Z}$ A.5.1

$$\left.\begin{array}{ll}\text{(A1)} & \text{If } a, b \in \mathbb{Z}, \text{ then } a + b \in \mathbb{Z}. \\ \text{(M1)} & \text{If } a, b \in \mathbb{Z}, \text{ then } a \cdot b \in \mathbb{Z}.\end{array}\right\} \text{Closure}$$

$$\left.\begin{array}{ll}\text{(A2)} & \text{If } a, b, c \in \mathbb{Z}, \text{ then } a + (b + c) = (a + b) + c. \\ \text{(M2)} & \text{If } a, b, c \in \mathbb{Z}, \text{ then } a \cdot (b \cdot c) = (a \cdot b) \cdot c.\end{array}\right\} \text{Associativity}$$

$$\left.\begin{array}{ll}\text{(A3)} & \text{If } a, b \in \mathbb{Z}, \text{ then } a + b = b + a. \\ \text{(M3)} & \text{If } a, b \in \mathbb{Z}, \text{ then } a \cdot b = b \cdot a.\end{array}\right\} \text{Commutativity}$$

$$\left.\begin{array}{ll}\text{(A4)} & \exists\, 0 \in \mathbb{Z} \ni \forall a \in \mathbb{Z},\ a + 0 = 0 + a = a. \\ \text{(M4)} & \exists\, 1 \in \mathbb{Z} \ni \forall a \in \mathbb{Z},\ a \cdot 1 = 1 \cdot a = a.\end{array}\right\} \text{Identities}$$

$$\left.\begin{array}{ll}\text{(A5)} & \forall a \in \mathbb{Z},\ \exists -a \in \mathbb{Z} \ni a + (-a) = (-a) + a = 0.\end{array}\right\} \text{Additive inverses}$$

In general, elements in $\mathbb{Z}$ do not have multiplicative inverses in $\mathbb{Z}$. That is, given an element $a \in \mathbb{Z}$, we cannot necessarily find another element $b \in \mathbb{Z}$ such that $ab = 1$. However, some integers do have multiplicative inverses, namely 1 and -1.

The operations of addition and multiplication are tied together by the distributive law.

(D) If $a, b, c \in \mathbb{Z}$, then $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

Without the distributive law, there would be no connection between addition and multiplication. The richness of the structure is embodied in the interaction between the two operations.

Let's stop and investigate some of the implications of these 10 axioms.

Facts A.5.2

1. Additive identities are unique.

   Proof. Suppose that 0 and $0'$ are additive identities. Then $0 = 0 + 0' = 0'$.

2. Multiplicative identities are unique.

   Proof. Exercise. (Hint: Use the same technique as above.)

3. Additive inverses are unique.

   Proof. Suppose that $a \in \mathbb{Z}$ and $a + a' = 0$. Then $-a + (a + a') = -a + 0 = -a$. On the other hand, by associativity ((A2)), we have $-a + (a + a') = ((-a) + a) + a' = 0 + a' = a'$. Thus, $a' = -a$.

4. (Cancellation for addition) If $a, b, c \in \mathbb{Z}$ and $a + b = a + c$, then $b = c$.

   Proof. If $a + b = a + c$, then $-a + (a + b) = -a + (a + c)$. By associativity ((A2)), $((-a) + a) + b = ((-a) + a) + c$, and hence $0 + b = 0 + c$, from which we conclude that $b = c$.

5. If $a \in \mathbb{Z}$, then $a \cdot 0 = 0$.

   Proof. We can write

$$\begin{aligned} a \cdot 0 &= a \cdot (0+0) \\ (a \cdot 0) + 0 &= a \cdot 0 + a \cdot 0 \end{aligned}$$

   by properties of the additive identity and the distributive law. Now cancel to get $a \cdot 0 = 0$.

   This is really quite something, and it emphasizes the role of the distributive law. What we have here is multiplication by the additive identity reproducing the additive identity. We have more interaction between multiplication and addition in the following statements.

6. If $a \in \mathbb{Z}$, then $(-1) \cdot a = -a$.

   Proof. We can write $a + (-1) \cdot a = 1 \cdot a + (-1) \cdot a = (1 + (-1)) \cdot a = 0 \cdot a = 0$. But additive inverses are unique, so $-a = (-1) \cdot a$.

   Notice that this really says something. That is, the left-hand expression, $(-1) \cdot a$, represents the additive inverse of the multiplicative identity multiplied by a. The right-hand side, $-a$, on the other hand, represents the additive inverse of a.

Notice that, when convenient, we drop the dot which signifies multiplication.

Exercise A.5.3 If $a, b \in \mathbb{Z}$, then $(-a)b = a(-b) = -(ab)$.

Exercise A.5.4 If $a, b \in \mathbb{Z}$, then $(-a)(-b) = ab$.

Now, what other properties do the integers have? In the integers, cancellation for multiplication doesn't follow from the first 10 axioms. Cancellation for multiplication should be familiar; many texts introduce it as an additional axiom for the integers in the following form.

(C) If $a, b, c \in \mathbb{Z}$ with $a \neq 0$ and $ab = ac$, then $b = c$.

Exercise A.5.5 Why is $a = 0$ excluded?

However, we will see shortly that because the integers are also ordered, cancellation in the integers is a consequence of the order properties.

Exercise A.5.6 Cancellation can be phrased in another way. Show that the statement "if $a, b \in \mathbb{Z}$ and $ab = 0$, then either $a = 0$ or $b = 0$" is equivalent to cancellation.

What else do we have for the integers? We have inequalities. The $<$ sign should be familiar to you. It is subject to the following *rules of order.*

(O1) If $a, b \in \mathbb{Z}$, then one and only one of the following holds: $a < b$, $a = b$, or $b < a$. (Trichotomy)

(O2) If $a, b, c \in \mathbb{Z}$ with $a < b$ and $b < c$, then $a < c$. (Transitivity)

(O3) If $a, b, c \in \mathbb{Z}$ and $a < b$, then $a + c < b + c$. (Addition)

(O4) If $a, b, c \in \mathbb{Z}$, $a < b$, and $0 < c$, then $ac < bc$. (Multiplication by positive elements)

We adopt the usual notation and terminology. That is, if $a < b$, we say that "a is less than b." If $a < b$ or $a = b$, we say that "a is less than or equal to b" and write $a \leq b$. If $a < b$ we may also write $b > a$ and say that "b is greater than a." The statement $b \geq a$ is now self-explanatory.

Here are some examples of recreational exercises and facts which go with the order axioms. For these statements and the following exercises, let $a, b, c \in \mathbb{Z}$.

Facts A.5.7

1. $a > 0$ iff $-a < 0$.

   Proof. Suppose $a > 0$. Add $-a$ to both sides.

2. If $a > 0$ and $b > 0$, then $ab > 0$.

   Proof. Suppose $a > 0$. Then, since $b > 0$, $ab > 0 \cdot b = 0$.

3. If $a > 0$ and $b < 0$, then $ab < 0$.

   Proof. Suppose $a > 0$ and $b < 0$. Then $-b > 0$ and $a(-b) = -(ab) > 0$. So $ab < 0$.

4. If $a < 0$ and $b < 0$, then $ab > 0$.

   Proof. If $a < 0$ and $b < 0$, then $-a > 0$ and $-b > 0$. Hence $(-a)(-b) = ab > 0$.

5. If $a \neq 0$, then $a^2 > 0$.

   Proof. If a is greater then 0, use Fact 2. If a is less then 0, use Fact 4.

6. $1 > 0$.

   Proof. $1 = 1^2$.

7. If $a > b$ and $c < 0$, then $ac < bc$.

   Proof. If $a > b$, then $a - b > 0$. Since $-c > 0$, $(-c)(a - b) = -ac + bc > 0$. Hence, $bc > ac$.

8. If $a > b$, then $-a < -b$.

   Proof. Let $c = -1$.

Are you having fun yet? Good, try these exercises.

Exercise A.5.8 Suppose that $0 < a$ and $0 < b$. Show that $a < b$ iff $a^2 < b^2$.

Exercise A.5.9 Suppose that $a < 0$ and $b < 0$. Show that $a < b$ iff $b^2 < a^2$.

Exercise A.5.10 Show that $2ab \leq a^2 + b^2$.

The set $\mathbb{N}$ of natural numbers is the set of positive elements in $\mathbb{Z}$, that is, the set of elements which are greater than 0. It is clear that $\mathbb{N}$ is closed under addition and multiplication. If we add trichotomy, these properties lead to an alternate characterization of order.

Exercise A.5.11 Suppose now that we have only the first 10 axioms for $\mathbb{Z}$ as well as the cancellation property (C). Let P be a set of integers with the following properties.

1. If $a \in \mathbb{Z}$, then one and only one of the following holds: $a \in P$, $a = 0$, or $-a \in P$.

2. If $a, b \in P$, then $a + b \in P$ and $ab \in P$.

For $a, b \in \mathbb{Z}$, define $a < b$ if $b - a \in P$. Show that this relation satisfies $(\mathsf{O}1)$–$(\mathsf{O}4)$. Moreover, if we have a relation that satisfies $(\mathsf{O}1)$–$(\mathsf{O}4)$, and we define $P = \{a \in \mathbb{Z} \mid a > 0\}$, then show that P satisfies properties 1 and 2 above.

Exercise A.5.12 Show that the cancellation property (C) can be proved using the axioms for addition and multiplication and the order axioms.

So far, the integers have five axioms for addition, four for multiplication, one for the distributive law, and four for order. There is one more axiom which plays a crucial role. It is called the *Well-Ordering Principle.* This Principle assures us that 1 is the smallest positive integer. This should not come as a surprise but we do need something to confirm this. In the rational numbers, which we construct in the next section, the first fourteen axioms are satisfied, but there is actually no smallest positive element. Thus, we need to introduce the Well-Ordering Principle as an axiom for $\mathbb{Z}$.

(**A.5.13** WO) Well-Ordering Principle for $\mathbb{Z}$ If A is a nonempty subset of the positive integers, then A has a least element. That is, there exists an element $a_0 \in A$, such that for all $a \in A$, $a_0 \leq a$.

That does it! We now have the 15 properties, and they completely characterize the integers. (For a proof of this, see Project 2 in this chapter.) Most of the work with the Well-Ordering Principle will be done later. However, here are a couple of facts which follow immediately from the Well-Ordering Principle.

Facts A.5.14

1. There are no integers between 0 and 1.

   Proof. Let $A = \{a \in \mathbb{Z} \mid 0 < a < 1\}$. If $A \neq \varnothing$, then it has a least element a_0 which is in A. So, $0 < a_0 < 1$, and, by property (O4), $0 < a_0^2 < a_0$. But then $a_0^2 \in A$ and a_0 is not the least element.

2. (Mathematical Induction) Let A be a set of positive integers such that $1 \in A$, and if $k \in A$, then $k + 1 \in A$. Then A is the set of all positive integers.

   Proof. Suppose there exists a positive integer which is not in A, and let A' be the set of all such positive integers. Then A' is a nonempty subset of the positive integers, and hence has a least element c. Now $c > 1$ since $1 \in A$, and there is no integer between 0 and 1. So $c - 1$ is an integer greater than 0. Since $c - 1 < c$, it follows that $c - 1 \in A$. And, so, $(c - 1) + 1 = c$ is also in A, which is a contradiction.

Exercise A.5.15 If n and k are non-negative integers with $n \geq k$, we define the *binomial coefficient* $\binom{n}{k}$ by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $n! = n(n-1)\cdots 2 \cdot 1$, and we set $0! = 1$ (this will be explained later in the book when we discuss the Gamma function). Prove the *Binomial Theorem*: If $a, b \in \mathbb{Z}$ and n is a positive integer, then

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

(Use Mathematical Induction.)

Remark A.5.16 Observe that the binomial coefficient $\binom{n}{k}$ represents the number of ways of choosing k objects from n objects where order does not matter. The binomial coefficient $\binom{n}{k}$ is the number of subsets of k elements in a set with n elements. Of course the binomial theorem implies that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$, the total number of subsets of a set with n elements.

Exercise A.5.17 *i.* Prove by induction that if A and B are finite sets, A with n elements and B with m elements, then $A \times B$ has nm elements.

ii. Prove by induction the corresponding result for a collection of k finite sets, where $k > 2$.

A.6 Equivalence Relations and the Construction of $\mathbb{Q}$

Next we turn to the idea of a relation on a set. Here is the formal definition of a relation.

Definition A.6.1 A *relation* on a set X is a subset R of $X \times X$.

For example, we can define a relation on $\mathbb{Z}$ by setting R equal to $\{(a,b) | a,b \in \mathbb{Z} \text{ and } a < b\}$.

Equivalence relations are everywhere in mathematics, and we really mean that. What an equivalence relation does is take a set and partition it into subsets in a special way. Some equivalence relations appear to be very natural, some appear to be supernatural, and others appear to make no sense at all.

Definition A.6.2 Let X be a set. An *equivalence relation* on X is a relation R on X such that

(ER1) For all $a \in X$, $(a,a) \in R$. (Reflexive)

(ER2) For $a, b \in X$, if $(a,b) \in R$, then $(b,a) \in R$. (Symmetric)

(ER3) For $a, b, c \in X$, if $(a,b), (b,c) \in R$, then $(a,c) \in R$. (Transitive)

The "twiddle" notation ($\sim$) is often used in mathematics. That is, if $(a,b) \in R$, we write $a \sim b$. Then the definition of equivalence relation becomes

(ER1) For all $a \in X$, $a \sim a$. (Reflexive)

(ER2) For $a, b \in X$, if $a \sim b$ then $b \sim a$. (Symmetric)

(ER3) For $a, b, c \in X$, if $a \sim b$ and $b \sim c$, then $a \sim c$. (Transitive)

Again, speaking loosely, we can refer to $\sim$ as an equivalence relation on X.

Exercise A.6.3 Let R be a relation on X, which satisfies

a. For all $a \in X$, $(a,a) \in R$, and

b. for $a, b, c \in X$ if $(a,b), (b,c) \in R$, then $(c,a) \in R$.

Show that R is an equivalence relation.

Example A.6.4 The most basic example of an equivalence relation is equality. That is, $a \sim b$ iff $a = b$. Prove this, but please don't write anything.

Example A.6.5 If A and B are triangles in the plane, write $A \sim B$ if and only if A is similar to B.

Example A.6.6 Let n be an integer greater than or equal to 2. If $a, b \in \mathbb{Z}$, we say that $a \sim b$ iff $a - b$ is a multiple of n, that is, n divides $a - b$.

This last example requires a little more elucidation. So, we present a brief discussion about divisibility in $\mathbb{Z}$.

Definition A.6.7 Suppose that a and b are integers. We say that a *divides* b, written $a|b$, if there is an element $c \in \mathbb{Z}$ such that $b = ac$. The number a is called a *divisor* of b.

We need the following facts about divisibility.

Facts A.6.8

1. If $a \in \mathbb{Z}$, then $a|a$.
2. If $a|b$ then $a| - b$.

3. If $a|b$ and $b|c$, then $a|c$.

These facts are easy to prove. For example, if $a|b$ and $b|c$, there are integers h and k such that $b = ha$ and $c = kb$. But then $c = (hk)a$, and that does it.

Exercise A.6.9 Show that, if $a \in \mathbb{Z}$, then $a|0$.

Exercise A.6.10 Show that, if a and b are integers such that $a|b$ and $b|a$, then $a = \pm b$.

Exercise A.6.11 Show that, if $c|a$ and $c|b$, and $s, t \in \mathbb{Z}$, then $c|(sa + tb)$.

There is one other type of integer which should be familiar to the reader.

Definition A.6.12 Let p be a positive integer greater than or equal to 2. We say that p is *prime* if the only positive divisors of p are 1 and p.

If n is a positive integer greater than 2 which is not prime, then n is called *composite*. So, if n is composite there exist integers a and b both greater than or equal to 2, such that $n = ab$.

Exercise A.6.13 Let n be a positive integer greater than or equal to 2. Then there exists a prime p such that p divides n.

The partitioning into subsets relative to an equivalence relation comes about as follows. If $a \in X$, we write $C(a) = \{b \in X \mid b \sim a\}$. $C(a)$ is called *the class of* a or *the equivalence class containing* a. Here are the properties of equivalence classes.

Theorem A.6.14 (Properties of equivalence classes)

1. $a \in C(a)$.

   Proof. Reflexivity.

2. If $a \sim b$, then $C(a) = C(b)$.

   Proof. Transitivity.

3. If a is not equivalent b ($a \nsim b$), then $C(a) \cap C(b) = \varnothing$.

   Proof. If $c \in C(a) \cap C(b)$, then $c \sim a$ and $c \sim b$, so $a \sim b$. So $C(a) \cap C(b) \neq \varnothing$ iff $C(a) = C(b)$.

4. $\bigcup_{a \in X} C(a) = X$.

   Proof. Use 1 above.

This all means that an equivalence relation on a set X partitions X into a collection of pairwise disjoint subsets. Although this looks quite special, it's really not that impressive. For example, take a set X and break it up into pairwise disjoint nonempty subsets whose union is all of X. Then, for $a, b \in X$, define $a \sim b$ if a and b are in the same subset.

Exercise A.6.15 Prove that this is an equivalence relation on X.

One more example of an equivalence relation will prove useful for future developments. This is a method for constructing the rational numbers $\mathbb{Q}$ from the integers $\mathbb{Z}$ using the properties discussed in the last section. We consider the set $F = \{(a, b) \mid a, b \in \mathbb{Z} \text{ and } b \neq 0\}$. We are thinking (for example) of the pair $(2, 3)$ as the fraction $2/3$. For $(a, b), (c, d) \in F$, we define $(a, b) \sim (c, d)$ if $ad = bc$. Thus, for instance, $(2, 3) \sim (8, 12) \sim (-6, -9)$.

Exercise A.6.16 Show that $\sim$ is an equivalence relation on F.

The set of equivalence classes determined by this equivalence relation is called the *rational numbers* and is denoted by $\mathbb{Q}$. You should be extremely happy about this since it explains all that business about equivalent fractions that you encountered in elementary school. What a relief!

We have several things to do with this example. First, we have to add and multiply rational numbers, that is, add and multiply equivalence classes. The fundamental principle to be established here is that, when we add or multiply equivalence classes, we do it by selecting an element from each equivalence class and adding or multiplying these. We must be certain that the result is independent of the representatives that we choose in the equivalence classes. For simplicity, we denote the class of (a, b) by $\{(a, b)\}$ rather than $C((a, b))$.

For $\{(a, b)\}, \{(c, d)\} \in \mathbb{Q}$, we define $\{(a, b)\} + \{(c, d)\} = \{(ad + bc, bd)\}$ and $\{(a, b)\} \cdot \{(c, d)\} = \{(ac, bd)\}$. What we must establish is the fact that if $(a, b) \sim (a', b')$ and $(c, d) \sim (c', d')$, then $(ad + bc, bd) \sim (a'd' + b'c', b'd')$ and $(ac, bd) \sim (a'c', b'd')$. All this requires is a little elementary algebra, but, for your sake, we'll actually do one and you can do the other. Of course, we do the easier of the two and leave the more complicated one for you. So, here goes: $(a, b) \sim (a', b')$ means that $ab' = a'b$, and $(c, d) \sim (c', d')$ means that $cd' = c'd$. Multiplying the first equality by cd', and then substituting $cd' = c'd$ on the right hand side of the resulting equation, we get the desired equality $acb'd' = a'c'bd$.

Exercise A.6.17 You do addition. It's messy.

When we are defining some operation which combines equivalence classes, we often do this by choosing representatives from each class and then showing that it doesn't make any difference which representatives are chosen. We have a formal name for this. We say that the operation under consideration is *well-defined* if the result is independent of the representatives chosen in the equivalence classes.

Throughout this book, we will encounter equivalence relations on a regular basis. You will be fortunate enough to have the opportunity to prove that these are actually equivalence relations.

What properties are satisfied by addition and multiplication as defined above? For example, what about the associativity of addition? We must prove that $(\{(a, b)\}+\{(c, d)\})+\{(e, f)\} = \{(a, b)\}+(\{(c, d)\}+\{(e, f)\})$. Well,

$$\begin{aligned}(\{(a, b)\} + \{(c, d)\}) + \{(e, f)\} &= \{(ad + bc, bd)\} + \{(e, f)\} \\ &= \{((ad + bc)f + (bd)e, (bd)f)\}.\end{aligned}$$

Now we use associativity and distributivity in $\mathbb{Z}$ to rearrange things in an appropriate fashion. This gives $\{(((ad)f + (bc)f) + (bd)e, (bd)f)\}$, and using the acrobatics of parentheses, we get $\{(a(df) + b(cf + de), b(df))\} = \{(a, b)\} + (\{(c, d)\} + \{(e, f)\})$. This is all rather simple. To prove various properties of addition and multiplication in $\mathbb{Q}$, we reduce this to known properties from $\mathbb{Z}$.

Exercise A.6.18

i. Prove the associative law for multiplication in $\mathbb{Q}$.

ii. Prove the commutative laws for addition and multiplication in $\mathbb{Q}$.

iii. Show that $\{(0, 1)\}$ is an additive identity in $\mathbb{Q}$.

iv. Show that $\{(1, 1)\}$ is a multiplicative identity in $\mathbb{Q}$.

v. Show that $\{(-a, b)\}$ is an additive inverse for $\{(a, b)\}$.

vi. Prove the distributive law for $\mathbb{Q}$.

Notice here that if $\{(a,b)\} \neq \{(0,1)\}$, that is, $a \neq 0$, then $\{(a,b)\} \cdot \{(b,a)\} = \{(1,1)\}$. Thus, in $\mathbb{Q}$, we have multiplicative inverses for nonzero elements.

Let's tidy this up a bit. First of all, we have no intention of going around writing rational numbers as equivalence classes of ordered pairs of integers. So let's decide once and for all to write the rational number $\{(a,b)\}$ as a/b. Most of the time this fraction will be reduced to lowest terms, but, if it is not reduced to lowest terms, it will certainly be in the same equivalence class as a fraction which is reduced to lowest terms. With this, addition and multiplication of rational numbers have their usual definition:

$$\begin{aligned} \frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd}, \\ \frac{a}{b} \cdot \frac{c}{d} &= \frac{ac}{bd}. \end{aligned}$$

Now consider the axioms for the integers (A1)–(A5), (M1)–(M4), and (D). All of these hold for the rational numbers, and there is another multiplicative property, multiplicative inverses.

(M5) If $a \neq 0$, then there is an element a^{-1} such that $aa^{-1} = a^{-1}a = 1$.

The operations of addition and multiplication are sometimes called *binary operations* or *internal laws of composition*

Definition A.6.19 Let R be a non-empty set. An *internal law of composition (ILC)* on R is a map $\circ : R \times R \to R$. If $a, b \in R$ then we usually write $\circ((a,b)) = a \circ b$

Of course the more properties that are satisfied by internal laws of composition, the better life gets.

Definition A.6.20 A set with two internal laws of composition, $+$ and $\cdot$, that satisfy (A1)–(A5), (M1)–(M4), and (D) is called a *commutative ring with* 1. If, in addition, cancellation (C) holds for multiplication, the commutative ring with 1 is called an *integral domain.* If (M5) also holds, the structure is called a *field*.

Note that the word "commutative" in this definition refers not to the commutativity of addition but to the commutativity of multiplication. Thus, in our latest terminology, $\mathbb{Z}$ is a integral domain and $\mathbb{Q}$ is a field. What about cancellation for multiplication? This followed from order in $\mathbb{Z}$, but for $\mathbb{Q}$ (or any field for that matter) cancellation for multiplication holds automatically.

Exercise A.6.21 Prove this.

Exercise A.6.22 Let X be a nonempty set and $R = \wp(X)$. Show that R with symmetric difference as addition and intersection as multiplication is a commutative ring with 1. When is R a field?

There is another definition which will prove useful in our discussions about these various algebraic structures.

Definition A.6.23 Suppose that R is a commutative ring with 1. A subset R_0 of R is a *subring* if R_0 is a ring itself with the same operations of addition and multiplication as in R. We don't necessarily require that R_0 have a multiplicative identity and in this case we call R_0 simply a *commutative ring.*

The same idea can be used to define *subintegral domain.* Finally, if F is a field and F_0 is a subset of F, we say that F_0 is a *subfield* if it is a field with the same operations of addition and multiplication as in F.

Exercise A.6.24

i. Let R be a ring and R_0 a non empty subset of R, show that R_0 is a subring if for any $a, b \in R_0$ we have $a - b$ and ab in R_0.

ii. If F is a field and F_0 is non-empty subset of F, are the properties in *i* enough to ensure that F_0 is a subfield?

What about order in $\mathbb{Q}$? It is simple to extend the order from $\mathbb{Z}$ to $\mathbb{Q}$. We do this using the notion of a set of positive elements. We say that $a/b \in \mathbb{Q}$ is positive if $ab > 0$ in $\mathbb{Z}$.

Exercise A.6.25 Show that the above notion of positivity in $\mathbb{Q}$ satisfies the properties in Exercise A.5.11, or equivalently, the properties of order given in (O1)–(O4).

Definition A.6.26 An integral domain or field in which there is an order relation satisfying (O1)–(O4) is called an *ordered integral domain* or *ordered field* respectively. See Project 1.3 for more about this.

Remark A.6.27 Note that the natural numbers $\mathbb{N}$ may be regarded as a subset of $\mathbb{Z}$, and in turn the integers $\mathbb{Z}$ may be regarded as a subset of $\mathbb{Q}$.

So what is this all about? We have rules for the integers, and the same rules, along with (M5), are satisfied by the rational numbers. Actually, there are lots of structures other than the integers and the rational numbers which have operations of addition, multiplication, and, sometimes, an order relation.

We want to give two more examples before we leave this section. First, let n be a positive integer greater than or equal to 2 and consider the equivalence relation given in Example A.6.6. What are the equivalence classes? For example, take $n = 5$. Then we have 5 classes. They are

$$\begin{aligned}
C(0) = \overline{0} &= \{0, 5, -5, 10, -10, \ldots\} \\
C(1) = \overline{1} &= \{1, 6, -4, 11, -9, \ldots\} \\
C(2) = \overline{2} &= \{2, 7, -3, 12, -8, \ldots\} \\
C(3) = \overline{3} &= \{3, 8, -2, 13, -7, \ldots\} \\
C(4) = \overline{4} &= \{4, 9, -1, 14, -6, \ldots\}.
\end{aligned}$$

Note that, in this example, we have simplified the notation of equivalence class by writing the equivalence class $C(a)$ by $\bar{a}$. Observe that $\overline{5} = \overline{0}$, $\overline{6} = \overline{1}$, etc. In general, for an arbitrary n, we will have n classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$. These are called *the equivalence classes modulo n*, or, for short, *mod n*. Moreover, for any integer a, we denote the equivalence class in which a lies by $\overline{a}$. Of course, it is always true that $\overline{a}$ is equal to one of the classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$. Let's define addition and multiplication mod n.

Definition A.6.28 Denote the set of equivalence classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$ by $\mathbb{Z}_n$. For $\overline{a}, \overline{b} \in \mathbb{Z}_n$, define $\overline{a} + \overline{b} = \overline{a+b}$ and $\overline{a}\overline{b} = \overline{ab}$.

Exercise A.6.29

i. Show that addition and multiplication in $\mathbb{Z}_n$ are well-defined.

ii. Show that, with these operations, $\mathbb{Z}_n$ is a commutative ring with 1.

iii. Show that $\mathbb{Z}_n$ cannot satisfy the order axioms no matter how $>$ is defined.

iv. Show that $\mathbb{Z}_2$ is a field but $\mathbb{Z}_4$ is not.

v. For p prime show that $\mathbb{Z}_p$ is a field.

The second example is the real numbers denoted by $\mathbb{R}$. A construction and complete discussion of the real numbers is given in Chapter 1. For the moment, however, it will suffice to say that the real numbers are an ordered field which contains $\mathbb{Q}$ and has one additional property called the least upper bound property. In Chapter 2, we use the real numbers as an example without being concerned with this additional property.

A.7 Functions

If you think equivalence relations are everywhere, wait until you see functions. We would all be better off if functions were introduced in kindergarten and studied regularly thereafter. The concept of a function is one of the most important ideas in mathematics. We give the informal definition first because it is much closer to the way people think about functions in practice.

Informally, a function from a set A to a set B is a correspondence between elements of A and elements of B such that each element of A is associated to exactly one element of B. This includes the familiar numerical functions of calculus, where, most often, the sets A and B are the real numbers or subsets thereof. But it also includes many examples which have nothing to do with the concept of numbers.

Example A.7.1 Given any set A, there is a unique function from A to A which assigns each element of A to itself. This is called the *identity function* on A.

What functions do is take elements of a given set and push them into another set (or maybe even the same set). The requirement is that to each element of the first set must correspond exactly to one element of the second. This does not preclude having two distinct elements of the first set correspond to the same element of the second set.

Example A.7.2 Let A and B be nonempty sets and choose a fixed element $b \in B$. Define a function from A to B by letting every element of A correspond to b. This is called a *constant function*.

Before we go too far with the informal idea, let's give a more formal definition for the notion of function.

Definition A.7.3 Let A and B be nonempty sets. A *function* from A to B is a subset of $A \times B$ such that each element of A occurs exactly once as a first coordinate.

This, of course, is an entirely useless definition, but it does carry with it the idea expressed informally above. That is, to each element of A there corresponds exactly one element of B. When you think of functions you will hardly ever think of ordered pairs. The informal notion of a correspondence satisfying certain properties should be your guide. Here's the notation we use. If A and B are sets and f is a function from A to B, we write $f : A \to B$. If $a \in A$, we write $f(a)$ for the corresponding element of B. So, if we were to write this as an ordered pair, we would write $(a, f(a))$.

Exercise A.7.4 How would you formulate the definition of function if either A or B were the empty set?

Example A.7.5 Take $A = \{a, b, c, d, e\}$ and $B = \{1, 2, 3, 4\}$. Now consider functions from A to B, that is assign a number to each letter. For example one such function is $(a, 1)$, $(b, 2)$, $(c, 3)$, $(d, 4)$, $(e, 2)$.

Exercise A.7.6 Determine all the functions from A to B in the previous example.

Exercise A.7.7

i. If A has n elements and B has m elements, how many functions are there from A to B?

ii. Let $B = \{0, 1\}$. Use the conclusion of part *i.* to give an alternate proof of Theorem A.4.12.

This is a convenient place to state the *pigeonhole principle*.

Theorem A.7.8 (Pigeonhole Principle) Suppose that m and n are positive integers with $n > m$. If n objects are distributed in m boxes, then some box must contain at least two objects. In terms of functions, the pigeon hole principle can be stated as follows. Suppose that A is a set with n elements and B is a set with m elements. If $f : A \to B$ is a function, there are two distinct elements of A that correspond to the same element of b.

Exercise A.7.9 Prove this any way you choose.

We turn next to the language of functions. Here is a list. Let f be a function from A to B. That is, $f : A \to B$.

Definitions A.7.10

a. The set A is called the *domain* of f.

b. If $A' \subseteq A$, we define $f(A') = \{b \in B \mid \exists a \in A' \text{ with } f(a) = b\}$. The set $f(A')$ is called the *image of A' in B under f*. In particular, $f(A)$ is the *image of A in B under f*. This is commonly called the *image* of f.

c. Note that there is no reason in the world for thinking that $f(A) = B$. If $f(A) = B$, we say that f is *onto* or *surjective* (f is a *surjection*). In general, B is called the *range* of f. Note that, if we change the range to $f(A)$, then f is surjective. That is, a function is always surjective onto its image.

d. Along with the property of being a function, f may also have the property that each element in the image of f corresponds to exactly one element in A. This can be written as follows. For any $a, a' \in A$ if $f(a) = f(a')$, then $a = a'$. A function with this property is called *one-to-one* or *injective* (f is an *injection*).

e. A function that is one-to-one and onto (that is, injective and surjective) is called *bijective* (f is a *bijection*). A bijection between two sets is often called a *one-to-one correspondence* between the sets.

f. Two functions f and g are the same, or equal, when they have the same domain, same range, same image, and the sets of pairs $\{(a, f(a))\}$ and $\{(a, g(a))\}$ are identical.

Exercise A.7.11 Determine which of your functions in Exercise A.7.6 are surjective. Notice that there are no injections or bijections. Why is this?

There is a way to combine functions which is very useful for many purposes.

Definition A.7.12 Suppose A, B, and C are sets and $f : A \to B$ and $g : B \to C$ are functions. The *composition* of f and g is a function $g \circ f : A \to C$ defined by $(g \circ f)(a) = g(f(a))$.

Of course, you met composition of functions in elementary calculus, and you enjoyed learning the chain rule. There is one very useful property concerning composition of functions, that is, composition of functions is *associative*.

Theorem A.7.13 If A, B, C, and D are sets and $f : A \to B$, $g : B \to C$, and $h : C \to D$ are functions, then $h \circ (g \circ f) = (h \circ g) \circ f$.

Proof. Suppose $a \in A$. Then $(h \circ (g \circ f))(a) = h((g \circ f)(a)) = h(g(f(a))) = (h \circ g)(f(a)) = ((h \circ g) \circ f)(a)$.

Let's stop for a minute with the definitions and consider some numerical examples. The sets we work with will be the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. (You may know what the real and complex numbers are, but we will construct them in Chapter 1.) Often when writing a function, we will specify a rule which tells us how to associate an element of the range to an element of the domain.

Examples A.7.14

i. $f : \mathbb{N} \to \mathbb{N}$, $f(n) = 2n$.

ii. $f : \mathbb{Z} \to \mathbb{Z}$, $f(n) = n + 6$.

iii. $f : \mathbb{N} \to \mathbb{Q}$, $f(n) = n$.

iv. Write the real numbers in terms of their decimal expansions. As usual, we do not allow a real number to end in all 9's repeating. Let $f : \mathbb{R} \to \mathbb{N}$ be defined by $f(x)$ equals the third digit after the decimal point (this is called the *Michelle* function).

Exercise A.7.15 Determine which of the above functions are surjective, injective, or bijective.

We mentioned above the so-called identity function, and we assign a symbol to it.

Definition A.7.16 Let A be a set. We give a symbol for the identity function defined in Example A.7.1. The identity function $I_A : A \to A$ is defined by $I_A(a) = a$ for $a \in A$.

Now suppose A and B are sets, and $f : A \to B$ is a bijection. Since each element of B comes from only one element of A under the function f, we can define a function $f^{-1} : B \to A$ which sends every element of B back to where it came from.

Definition A.7.17 Let $f : A \to B$ be a bijection. Then the *inverse* of f is the function $f^{-1} : B \to A$ defined as follows. If $b \in B$, then we set $f^{-1}(b) = a$ where a is the unique element of A such that $f(a) = b$.

Exercise A.7.18 Show that $f^{-1} \circ f = I_A$ and $f \circ f^{-1} = I_B$.

Exercise A.7.19 Suppose A, B, and C are sets and $f : A \to B$ and $g : B \to C$ are bijections. Show that $g \circ f$ is a bijection. Compute $(g \circ f)^{-1} : C \to A$.

Exercise A.7.20 Given $f : A \to B$ suppose there exist $g, h : B \to A$ so that $f \circ g = I_B$ and $h \circ f = I_A$. Show that f is a bijection and that $g = h = f^{-1}$.

Exercise A.7.21 Let $\mathbb{R}_+$ be the positive real numbers and define $f : \mathbb{R}_+ \to \mathbb{R}_+$ by $f(x) = x^2$. Show that f is a bijection from $\mathbb{R}_+$ to $\mathbb{R}_+$ and find f^{-1}. If we expand the domain and include all real numbers, what happens?

Exercise A.7.22 Define $f : \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} \dfrac{n}{2}, & \text{if } n \text{ is even} \\[2ex] \dfrac{1-n}{2}, & \text{if } n \text{ is odd.} \end{cases}$$

Show that f is a bijection.

Even if a function $f : A \to B$ is not a bijection, we can still take a subset $A' \subseteq A$ and consider its image $f(A') \subseteq B$. Moreover, we can take a subset B' of B and consider the *preimage* of B' in A.

Definition A.7.23 Suppose A and B are sets and $f : A \to B$ is a function. If $B' \subseteq B$, then the *preimage* of B' in A is defined as $f^{-1}(B') = \{a \in A \mid f(a) \in B'\}$.

So, $f^{-1}(B')$ is everything in A that is pushed into B' by the function f. Let's make a couple of quick observations about the empty set.

1. $f(\varnothing) = \varnothing$.
2. $f^{-1}(\varnothing) = \varnothing$.
3. More generally, if $B' \subseteq B$ and $B' \cap f(A) = \varnothing$, then $f^{-1}(B') = \varnothing$.

Take heed: given any subset $B' \subseteq B$ its preimage $f^{-1}(B')$ always exists. This has nothing to do with whether or not f is a bijection.

There are four basic results on images and preimages.

Theorem A.7.24 Suppose A and B are sets and $f : A \to B$ is a function. Let $A_1, A_2 \subseteq A$ and $B_1, B_2 \subseteq B$. Then

i. $f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$;

ii. $f(A_1 \cap A_2) \subseteq f(A_1) \cap f(A_2)$;

iii. $f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$;

iv. $f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2)$.

Proof. The proof is standard stuff. We will prove *iii.* Take $x \in f^{-1}(B_1 \cup B_2)$. Then $f(x) \in B_1 \cup B_2$ so $f(x) \in B_1$ or $f(x) \in B_2$. Hence $x \in f^{-1}(B_1)$ or $x \in f^{-1}(B_2)$. That is, $x \in f^{-1}(B_1) \cup f^{-1}(B_2)$ and so $f^{-1}(B_1 \cup B_2) \subseteq f^{-1}(B_1) \cup f^{-1}(B_2)$. Actually, you can read this argument backwards to show that $f^{-1}(B_1) \cup f^{-1}(B_2) \subseteq f^{-1}(B_1 \cup B_2)$. Thus, finally, the sets are equal. This is terribly boring, but you should do *i* and *iv* to discipline yourself. On the other hand, *ii* is more interesting.

Exercise A.7.25 Find an example to show that equality does not necessarily hold in *ii.*

Exercise A.7.26 Show that equality holds in *ii* of Theorem A.7.24 if f is an injection. In fact, if equality holds in *ii* for all subsets $A_1, A_2 \subseteq A$ then f is an injection.

Exercise A.7.27 Let A and B be sets and let $f : A \to B$ be a function. Suppose that $\{A_i\}_{i \in I}$ is a collection of subsets of A and $\{B_j\}_{j \in J}$ is a collection of subsets of B. Show that

i. $f(\bigcup_{i \in I} A_i) = \bigcup_{i \in I} f(A_i)$;

ii. $f(\bigcap_{i \in I} A_i) \subseteq \bigcap_{i \in I} f(A_i)$;

iii. $f^{-1}(\bigcup_{j \in J} B_j) = \bigcup_{j \in J} f^{-1}(B_j)$;

iv. $f^{-1}(\bigcap_{j \in J} B_j) = \bigcap_{j \in J} f^{-1}(B_j)$.

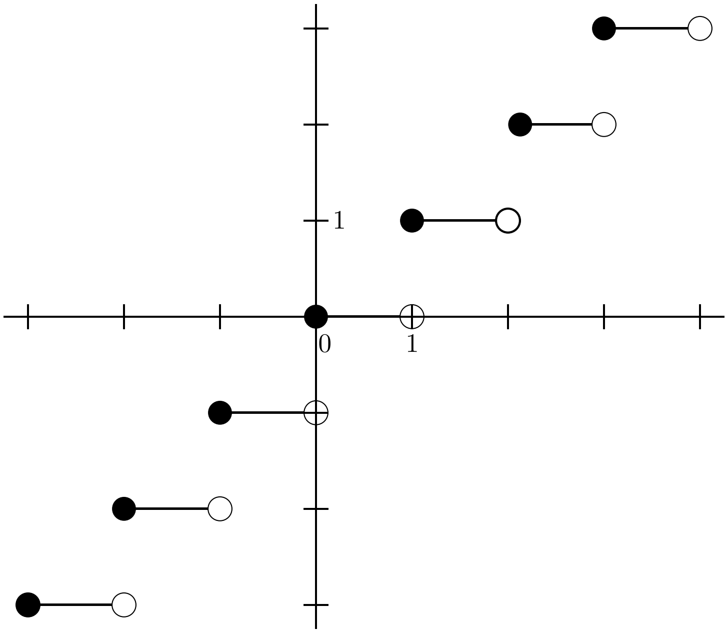
Note that, in this exercise, the number of sets in the union and intersection is not necessarily finite.

To close this section, we consider two important examples of functions.

Definition A.7.28 The *greatest integer function* $[\cdot] : \mathbb{R} \to \mathbb{R}$ is defined by $[x]$ equals the largest integer that is less than or equal to x.

Example A.7.29 $[n] = n$ for $n \in \mathbb{Z}$; $[17.5] = 17$; $[\sqrt{2}] = 1$; $[\pi] = 3$; $[-e] = -3$, etc.

Here is the graph of the greatest integer function.



Exercise A.7.30 Express the Michelle function (Example A.7.14) in terms of the greatest integer function. Graph the Michelle function.

Now we define polynomial functions. Polynomial functions are perfect examples of functions that fit into the "What is my rule?" category. Here is a polynomial function, with its rule. Let $p(x)$ be the function given by $p(x) = x^2 + 2x + 1$. You can plug in numbers for x and get values for $p(x)$. For instance, $p(0) = 1$ and $p(1) = 4$, and $p(\sqrt{2}) = 3 + 2\sqrt{2}$. Let's be a little more formal here.

Definition A.7.31 A *polynomial function* $p : \mathbb{R} \to \mathbb{R}$ is a function whose rule of correspondence is given by an expression of the form

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

where n is a natural number or zero, and the *coefficients* $a_n, a_{n-1}, \ldots, a_1, a_0$ are in $\mathbb{R}$. The function $p(x) = 0$ is a polynomial function. If $p(x) \neq 0$, then there is a largest power of x (including x^0) with a nonzero coefficient. This power is called the *degree* of the polynomial function. We make the convention that the degree of the polynomial function 0 is $-\infty$.

As an ordered pair, a polynomial function is written $(x, p(x))$. In the above definition, we say that p is a *polynomial function with real coefficients.* Notice that we could restrict the coefficients to be integers or rational numbers. In these cases we could restrict the domain to be the integers or rational numbers respectively.

Examples A.7.32

i. $p(x) = 0$.

ii. $p(x) = 17x^2 + 2x - 7$.

iii. $p(x) = (\sqrt{2} - 1)x^{83} + \pi x^{17} + \sqrt[3]{2}x^{11} + ex^7 + 6$.

You should have experience in adding and multiplying polynomial functions, so we won't go into details here. Here are a couple of interesting exercises.

Exercise A.7.33 If p and q are polynomial functions with real coefficients, then $\deg(pq) = \deg(p) + \deg(q)$. To accommodate the zero polynomial we use the convention $-\infty + k = -\infty$ for any k.

Exercise A.7.34 If p and q are polynomial functions with real coefficients and $\deg(p) \neq \deg(q)$, then $\deg(p+q) = \max(\deg(p), \deg(q))$. In any case, $\deg(p+q) \leq \max(\deg(p), \deg(q))$.

Exercise A.7.35 Show that the set of polynomial functions from $\mathbb{R}$ to $\mathbb{R}$ is an integral domain.

Another important type of function is that of a sequence. Sequences will occur regularly throughout the remainder of the text.

Definition A.7.36 Let X be a nonempty set. A *sequence* in X is a function $f : \mathbb{N} \to X$.

Thus, respecting the order in $\mathbb{N}$, we write a sequence

$$f(1), f(2), \ldots, f(n), \ldots$$

or $(x_1, x_2, \ldots, x_n, \ldots)$. We will also adopt the notation $(x_n)_{n \in \mathbb{N}}$.

Remark A.7.37 It is also useful to have sequences indexed by the non-negative integers, or even the set of all integers. So, for example, we might have $(x_0, x_1, \ldots, x_n, \ldots)$ or $(x_n)_{n \geq 0}$. If the index set is $\mathbb{Z}$, we write $(\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots)$, or $(x_n)_{n \in \mathbb{Z}}$.

Exercise A.7.38 If A_1 and A_2 are subsets of a universal set X, show that there is a bijection between the Cartesian product $A_1 \times A_2$ and the set of all functions $f : \{1, 2\} \to X$ such that $f(1) \in A_1$ and $f(2) \in A_2$. Do the same for any finite number of subsets of X.

This exercise is the beginning of our study of the axiom of choice, which comes up later in the chapter.

A.8 Other basic ideas

Finally, we come to a serious discussion of infinite sets. There are great pitfalls involved in any discussion of set theory, and our basic goal is to avoid these pitfalls while still having appropriate definitions, ideas and facts. Of course, in analysis, most of the sets we deal with are infinite. In fact, most of them contain the integers. Moreover, any discussion of continuity and change involves infinite sets. So of course in our usual perverse manner, we define finite sets first.

Definition A.8.1 A set A is *finite* if A is empty or there exists $n \in \mathbb{N}$ such that there is a bijection $f : A \to \{1, 2, \ldots, n\}$, where $\{1, 2, \ldots, n\}$ is the set of all natural numbers less than or equal to n. In this case, we say A has n elements.

Exercise A.8.2 If A is a finite set and B is a subset of A, show that B is a finite set. In addition show that if B is a proper subset then the number of elements in B is less then the number of elements in A.

There is a natural and useful characteristic of finite sets:

Theorem A.8.3 If A is a finite set and B is a proper subset of A, then there is no bijection between B and A.

Proof. Suppose A has n elements and B has m elements with $m < n$. Then the Pigeonhole Principle tells us that, for any function from A to B, there is an element of B which is the image of two different elements of A.

Exercise A.8.4 Show that the following are finite sets:

i. The English alphabet.

ii. The set of all possible twelve letter words made up of letters from the English alphabet.

iii. The set of all subsets of a finite set.

This approach to things makes the definition of infinite sets quite simple:

Definition A.8.5 An *infinite* set is a set that is not finite.

The notion of cardinality of a set is very important. Of course, most authors don't define cardinality. Instead, they say what it means for two sets to have the same cardinal number. This will do for our purposes.

Definition A.8.6 The *cardinal number* of a finite set A is the number of elements in A, that is, the cardinal number of A is the natural number n if there is a bijection between A and $\{k \in \mathbb{N} \mid 1 \leq k \leq n\}$.

Definition A.8.7 A set A has *cardinality* $\aleph_0$ (pronounced "aleph null" or "aleph naught") if it can be put in one-to-one correspondence with $\mathbb{N}$, that is there is a bijection between the set and $\mathbb{N}$. In general, two sets have the same cardinality if they can be put in one-to-one correspondence with each other.

Example A.8.8 The set $\mathbb{N}$ has cardinality $\aleph_0$ (this should not come as a surprise).

Although we will not see one for a while, be assured that there are infinite sets with cardinality other than $\aleph_0$.

Example A.8.9 The set $\mathbb{N} \cup \{0\}$ has cardinality $\aleph_0$ because the function $f : \mathbb{N} \cup \{0\} \to \mathbb{N}$ given by $f(n) = n + 1$ is a bijection.

Example A.8.10 The set $\mathbb{Z}$ has cardinality $\aleph_0$ because the function $f : \mathbb{Z} \to \mathbb{N}$ given by

$$f(z) = \begin{cases} 2z + 2 & \text{if } z \geq 0 \\ -2z - 1 & \text{if } z < 0 \end{cases}$$

is a bijection.

There is a very useful theorem which asserts the existence of a one-to-one correspondence between two sets. This relieves us of the burden of constructing a bijection between two sets to show that they have the same cardinality.

Theorem A.8.11 (Schröder-Bernstein) If A and B are sets, and there exist injections $f : A \to B$ and $g : B \to A$, then there exists a bijection between A and B.

Proof. First, we divide A into three disjoint subsets. For each $x \in A$, consider the list of elements

$$S_x = \{x, g^{-1}(x), f^{-1} \circ g^{-1}(x), g^{-1} \circ f^{-1} \circ g^{-1}(x), \dots\}.$$

The elements of this sequence are called *predecessors* of x. Notice that in S_x, we start with $x \in A$. Then $g^{-1}(x) \in B$ if $g^{-1}(x)$ exists (x may not be in the image of g). For each $x \in A$, exactly one of the three following possibilities occurs.

1. The list S_x is infinite.
2. The last term in the list is an element of A. That is, the last term is of the form $y = f^{-1} \circ g^{-1} \circ \cdots \circ g^{-1}(x)$, and $g^{-1}(y)$ does not exist (i.e. y is not in the image of g). In this case, we say that S_x *stops in A*.

3. The last term in the list is an element of B. That is, the last term is of the form $z = g^{-1} \circ f^{-1} \circ \cdots \circ g^{-1}(x)$ and $f^{-1}(z)$ does not exist (i.e. z is not in the image of f). In this case, we say that S_x *stops in B*.

Let the corresponding subsets of A be denoted by A_1, A_2, A_3. Similarly, define the corresponding subsets of B. That is

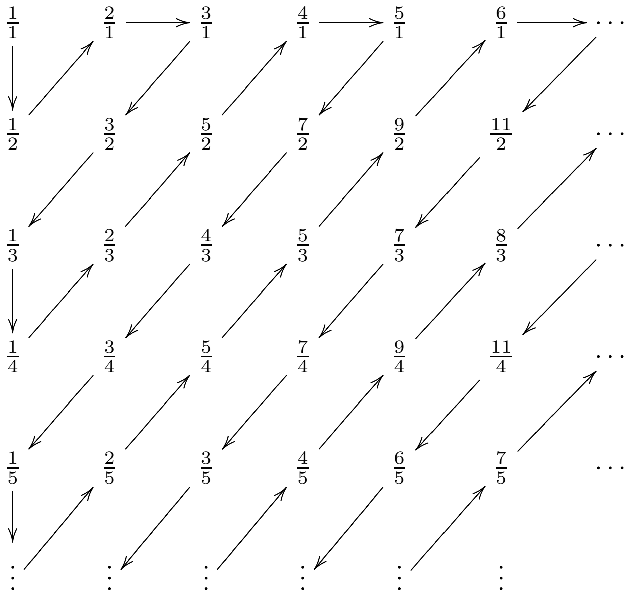
$$\begin{aligned}
B_1 &= \{y \in B \mid y \text{ has infinitely many predecessors } \}, \\
B_2 &= \{y \in B \mid \text{the predecessors of } y \text{ stop in } A\}, \text{ and} \\
B_3 &= \{y \in B \mid \text{the predecessors of } y \text{ stop in } B\}.
\end{aligned}$$

Now observe that $f : A_1 \to B_1$, $g : B_1 \to A_1$ are both bijections. Also, $g : B_2 \to A_2$ and $f : A_3 \to B_3$ are bijections.

Exercise A.8.12 Suppose A, B, and C are subsets of a set X such that $A \subseteq B \subseteq C$. Show that if A and C have the same cardinality, then A and B have the same cardinality.

Example A.8.13 $\mathbb{Q}_+$ has cardinality $\aleph_0$ (recall that $\mathbb{Q}_+$ denotes the positive rational numbers). Here are three proofs:

1. This is a very common and very sloppy proof. However the underlying idea will stand us in good stead.



To find a bijection between $\mathbb{N}$ and $\mathbb{Q}_+$, we write all the positive fractions in a grid, with all fractions with denominator 1 in the first row, all fractions with denominator 2 in the second row, all fractions with denominator 3 in the third row, etc. Now go through row by row and throw out all the fractions that aren't written in lowest terms. Then, starting at the upper left hand corner, trace a path through all the remaining numbers as above.

We can count along the path we drew, assigning a natural number to each fraction. So $\frac{1}{1} \to 1$, $\frac{1}{2} \to 2$, $\frac{2}{1} \to 3$, $\frac{3}{1} \to 4$, $\frac{3}{2} \to 5$, etc. This is a bijection. Therefore, $\mathbb{Q}_+$ is countable. Although this is a very common proof, the bijection is not at all obvious. It is very difficult to see, for example, which rational number corresponds to $1,000,000$.

2. In this proof, we'll make use of the Schröder-Bernstein Theorem. It is easy to inject $\mathbb{N}$ into $\mathbb{Q}_+$; simply send n to n. The injection from $\mathbb{Q}_+$ to $\mathbb{N}$ will be the one we used in Example A.7.14:**??** (where $\frac{a}{b}$ is sent to $adb_{(11)}$). Each number which is the image of a fraction has one and only one d in it, so it is easy to see which fraction is represented by a given integer. According to Schröder-Bernstein, two injections make a bijection, so $\mathbb{Q}_+$ is countable.

3. Write each positive fraction in lowest terms and factor the numerator and denominator into primes, so that $\frac{p}{q} = \frac{p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}}{q_1^{\beta_1} q_2^{\beta_2} \cdots q_m^{\beta_m}}$, with $p_i \neq q_j$. If by chance p or q is 1, and can't be factored, write it as 1^1. Then let $f : \mathbb{Q}_+ \to \mathbb{N}$ be defined by

   $$f\left(\frac{p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}}{q_1^{\beta_1} q_2^{\beta_2} \cdots q_m^{\beta_m}}\right) = p_1^{2\alpha_1} p_2^{2\alpha_2} \cdots p_n^{2\alpha_n} q_1^{2\beta_1 - 1} q_2^{2\beta_2 - 1} \cdots q_m^{2\beta_m - 1}.$$

   In particular, note that if $a \in \mathbb{Q}_+$ is an integer, then $f(a) = a^2$.

   Exercise A.8.14 Verify that f is a bijection.

   Exercise A.8.15 Suppose that $N = 10^k$ for some integer k. Find $p/q \in \mathbb{Q}$ such that $f(p/q) = N$.

Exercise A.8.16 Use any one of the above three proofs to show that $\mathbb{Q}$ is countable.

Exercise A.8.17 Show that the natural numbers are an infinite set.

Exercise A.8.18 Show that any set that has the same cardinal number as $\mathbb{N}$ is an infinite set.

Note: A set is called *countable*, or sometimes *denumerable*, if it has cardinality $\aleph_0$ (that is, if it is in one-to-one correspondence with the natural numbers.) The term countable is used in several ways. Many people use it to refer to infinite sets which are in one-to-one correspondence with $\mathbb{N}$, while others include finite sets when they say countable. This is not something to get disturbed about. Usually, when we refer to a countable set, we mean countably infinite (cardinality $\aleph_0$) . When we refer to a finite set, we will generally say "A is a finite set."

Exercise A.8.19 Show that a subset of a countable set is countable or finite.

Exercise A.8.20 Show that the set of all polynomial functions with integer coefficients is a countable set.

Theorem A.8.21 If A is an infinite set, then A has a countable subset.

Proof. Take any infinite set A and choose an element a_1 in A. Let $A_1 = A \setminus \{a_1\}$. By the definition of infinite set, A_1 is infinite. So we choose a_2 in A_1 and define $A_2 = A \setminus \{a_1, a_2\}$. Since A is not finite, we can continue to choose elements. Thus, if we have chosen $a_1, \ldots, a_n$, we consider $A_n = A \setminus \{a_1, \ldots, a_n\}$. Since A is infinite, we can choose an element a_{n+1} in A_n. Continuing inductively, we obtain our desired countable subset. Note that this countable set may be all of A.

Remark A.8.22 There is some discussion among mathematicians as to whether the preceding proof involves the Axiom of Choice. The Axiom of Choice in its fullest form will be discussed below. However, one can make the argument that it requires some sort of choice mechanism to pick an element from a non-empty set. The technique that we use in the proof of Theorem A.8.21 is sometimes referred to as "the countable Axiom of Choice."

We could pursue an alternate definition of an infinite set. In fact, we could define infinite sets first and then say that a finite set is a set that is not infinite. We use Theorem A.8.21 as motivation for the following.

Redefinition A.8.23 A set is *infinite* if there is a bijection between the set and one of its proper subsets.

Redefinition A.8.24 A *finite* set is a set which is not infinite.

To show the equivalence of the two definitions, recall that in Theorem A.8.3 we showed there is no bijection between a finite set and any of its proper subsets. This means that if a set is infinite by our new definition, it is not finite (hence, infinite) by the old definition too. Next, let's show that any set which is infinite by the old definition is bijective with one of its proper subsets.

Proof. Say A is an infinite set and $B \subseteq A$ is countable. Then we can write $B = \{b_1, b_2, \ldots, b_n, \ldots\}$. Now define $f : A \to A \setminus \{b_1\}$ as follows: for $a \in A \setminus B$, $f(a) = a$, and for $b_i \in B$, $f(b_i) = b_{i+1}$. Thus f is a bijection between A and $A \setminus \{b_1\}$. Therefore, our definitions are equivalent.

We now turn to operations involving infinite sets.

Facts A.8.25

1. If A_1 and A_2 are countable sets, then $A_1 \cup A_2$ is a countable set.
2. If A_1, A_2, …, A_n are countable sets, then $\cup_{j=1}^{n} A_j$ is a countable set.
3. Let $\{A_j\}_{j\in\mathbb{N}}$ be a countable collection of countable sets. Then $\cup_{j\in\mathbb{N}} A_j$ is a countable set.

Proof. We prove 3 only. You can prove the other two (or deduce them from 3).

Write $A_j = \{a_{j,1}, a_{j,2}, \ldots, a_{j,n}, \ldots\}$. We use the diagonal process, as in Example A.8.13. Simply write

$$A_1 : a_{1,1}, a_{1,2}, \ldots, a_{1,n}, \ldots$$

$$A_2 : a_{2,1}, a_{2,2}, \ldots, a_{2,n}, \ldots$$

$$\vdots$$

$$A_m : a_{m,1}, a_{m,2}, \ldots, a_{m,n}, \ldots$$

$$\vdots$$

Now count diagonally, ignoring repetitions.

Now let's take a look at Cartesian products. It is clear from the ideas presented above that if A_1 and A_2 are countable, then $A_1 \times A_2$ is countable.

Exercise A.8.26

i. Show that if A_1, A_2, …, A_n are countable, then $A_1 \times A_2 \times \cdots \times A_n$ is countable.

ii. What can you say about the countable Cartesian product of countable sets?

Next we look at the power set $\wp(A)$ for any set A.

Theorem A.8.27 If A is any set (including the empty set), there is no bijection between A and $\wp(A)$.

Proof. This is clear if A is the empty set. Suppose that there is a bijection between A and $\wp(A)$. If $a \in A$, let P_a be the subset of A associated with it. Now consider the set $B = \{a | a \notin P_a\}$. The set B must be associated to some element of A, which we creatively call b, so that $B = P_b$. Is b in B? For b to be in B, we must have that $b \notin P_b$. But $B = P_b$, so therefore b is not in B. But then $b \in P_b$, which means that b is in B. This is a contradiction. Therefore, there is no bijection between A and $\wp(A)$.

Definition A.8.28 If A is a countably infinite set, then the cardinality of $\wp(A)$ is denoted by $\mathbf{c}$.

Exercise A.8.29 Show that the definition of the cardinal number $\mathbf{c}$ does not depend on the choice of the countably infinite set A. That is if A and B are countably infinite sets then there is a bijection between $\wp(A)$ and $\wp(B)$.

Remark A.8.30 At this point, we observe that if A is a countable set, $A = \{a_1, a_2, \ldots, a_n, \ldots\}$, then $\wp(A)$ is in one-to-one correspondence with the set of all functions from A to the set $\{0, 1\}$. This correspondence is defined as follows: If B is a subset of A, then we define the map $f_B : A \to \{0, 1\}$ by $f_B(a_j) = 1$ if a_j is in B, 0 if a_j is not in B. In accordance with the notation of Theorem A.4.12, we will write $\mathbf{c} = 2^{\aleph_0}$. Observe that f_B can be viewed as a binary expansion of a real number between 0 and 1.

Exercise A.8.31 Suppose that A is a nonempty set. Show that $\wp(A)$ is in one to one correspondence with the set of all functions from A to $\{0, 1\}$.

One of the most important sets of numbers that we deal with in this book is the collection of real numbers $\mathbb{R}$. In Chapter 3, we will go through the formal construction of the real numbers from the rational numbers. For the present discussion, we can just consider the set of real numbers to be the set of all terminating or infinite decimals with the convention that no decimal expansion can terminate in all 9's. There are two things to show about the reals. The first is the proof due to Cantor that the reals are uncountable, and the second is that the cardinality of the real numbers is in fact $\mathbf{c}$.

Theorem A.8.32 The set of all infinite decimals between 0 and 1 is not countable.

Proof. We first note that the decimal expansion is unique with the exception of those that end in all nines. In this case, we always round up the digit which occurs before the sequence of nines. To prove that this set is not countable, we assume that it is, and list the real numbers between 0 and 1 vertically.

$$a_1 = 0.a_{1,1}a_{1,2} \ldots a_{1,n} \ldots$$

$$a_2 = 0.a_{2,1}a_{2,2} \ldots a_{2,n} \ldots$$

$$\vdots$$

$$a_m = 0.a_{m,1}a_{m,2} \ldots a_{m,n} \ldots$$

$$\vdots$$

We now proceed using a process similar to the one used in the proof of Theorem A.8.27 to produce a real number between 0 and 1 which is not on our list. We construct a number $b = 0.b_1b_2 \ldots b_n \ldots$ by proceeding diagonally down the list as follows: if $a_{1,1} = 1$, take $b_1 = 2$. If $a_{1,1} \neq 1$, take $b_1 = 1$. Next, if $a_{2,2} = 1$, take $b_2 = 2$. If $a_{2,2} \neq 1$, take $b_2 = 1$. Continuing this process, we see that the decimal $b = 0.b_1b_2 \ldots b_n \ldots$ cannot be on our list, since it differs from each number we list in at least one digit. Consequently, the real numbers between 0 and 1 are not countable.

Theorem A.8.33 The cardinality of the real numbers between 0 and 1 is $\mathbf{c} = 2^{\aleph_0}$.

Proof. To write down an exact bijection between $\wp(\mathbb{N})$ and the real numbers between 0 and 1 requires some care. The standard way to do this is to write all real numbers between 0 and 1 in their binary expansion in such a way that no expansion terminates in all ones. In considering the corresponding subsets of $\mathbb{N}$, we first remove two specific subsets of $\wp(\mathbb{N})$. We remove the two collections $A_f = \{C \in \wp(\mathbb{N}) \mid C \text{ is finite}\}$ and $A_{cf} = \{D \in \wp(\mathbb{N}) \mid {}^cD \text{ is finite}\}$. The collection $\wp(\mathbb{N}) \setminus (A_f \cup A_{cf})$ is in one-to-one correspondence with all binary expansions which have an infinite number of ones but do not terminate in all ones. We get the required bijection by Remark A.8.30.

We can place A_f into one-to-one correspondence with the set of all finite binary expansions with 0 in the first place, and A_{cf} can be put into one-to-one correspondence with the set of all finite binary expansions with 1 in the first place.

Exercise A.8.34 Write down these last two bijections explicitly.

Exercise A.8.35

i. Prove that the countable union of sets of cardinality $\mathbf{c}$ again has cardinality $\mathbf{c}$.

ii. Prove that the set of all real numbers has cardinality $\mathbf{c}$.

iii. Prove that the set of irrational numbers in $\mathbb{R}$ has cardinality $\mathbf{c}$.

How big do cardinal numbers get? For instance, the power set of $\mathbb{R}$ is "bigger than" $\mathbf{c}$. In fact, the power set of $\mathbb{R}$ can be identified with the set of all maps from $\mathbb{R}$ into $\{0, 1\}$ just as we did above for the power set of $\mathbb{N}$. Thus, we have $\wp(\mathbb{R}) = 2^{\mathbf{c}}$. We sometimes denote $2^{\mathbf{c}}$ by $\mathbf{f}$.

The following theorem is interesting and useful.

Theorem A.8.36 There is a bijection between the unit interval and the unit square.

Proof. Let

$$I = [0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$$

and $I^2 = [0, 1] \times [0, 1]$. This seems like a great time to use Schröder-Bernstein. The function $f : I \to I^2$ defined by $f(x) = (x, 0)$ is an injection. Define the function $g : I^2 \to I$ by the rule $g((a_0.a_1a_2 \ldots a_n \ldots, b_0.b_1b_2 \ldots b_n \ldots)) = (0.a_0b_0a_1b_1a_2b_2 \ldots a_nb_n \ldots)$, where $a_0a_1a_2 \ldots a_n \ldots$ and $b_0.b_1b_2 \ldots b_n \ldots$ are decimal expansions of the coordinates of any point in I^2 (of course, the decimal expansion is prohibited from ending in all 9s). The function $g : I^2 \to I$ is an injection. Therefore, there is a bijection between I and I^2.

A.9 Axiom of Choice

Definition A.9.1 A *partially ordered set* is a set X with a relation $\leq$ which is reflexive, transitive, and anti-symmetric (that means that if $a \leq b$ and $b \leq a$, then $a = b$). A *totally ordered set* is a partially ordered set with the additional property that, for any two elements $a, b \in X$, either $a \leq b$ or $b \leq a$. A *well-ordered set* is a totally ordered set in which any non-empty subset has a least element.

Example A.9.2

1. $(\mathbb{N}, \leq)$ is a totally ordered set, as are $(\mathbb{Z}, \leq)$, $(\mathbb{Q}, \leq)$ and $(\mathbb{R}, \leq)$.

2. Let X be a set, and let $\wp(X)$ be the collection of all subsets of X. Then $(\wp(X), \subseteq)$ is a partially ordered set.

Definition A.9.3 Let Y be a subset of a partially ordered set X. An *upper bound* for Y is an element $a \in X$ such that $y \leq a$ for all $y \in Y$. A *least upper bound* for Y is an element $b \in X$ such that b is an upper bound for Y and if a is an upper bound for Y, then $b \leq a$. The least upper bound is sometimes abbreviated lub, and is also denoted as sup (supremum). You can figure out what a *lower bound* and *greatest lower bound (*glb*)* are. The greatest lower bound is also denoted by inf (infimum).

Observe that a subset of a partially ordered set may not have an upper bound or a lower bound.

Exercise A.9.4 If a subset Y of a partially ordered set X has an upper bound, determine whether or not Y must have a least upper bound. If Y has a least upper bound, determine whether or not this least upper bound is unique.

Definition A.9.5 In a partially ordered set, an element b is *maximal* if $a \geq b$ implies $a = b$.

We turn now to one of the major topics of this chapter, the axiom of choice, and various logically equivalent statements. For many years, there has been considerable discussion among mathematicians about the use of the axiom of choice and the seemingly contradictory results that come along with it. We find it indispensable in obtaining a number of results in mathematics.

The Axiom of Choice A.9.6 Given a collection $\mathcal{C}$ of sets which does not include the empty set, there exists a function $\phi : \mathcal{C} \to \cup_{C \in \mathcal{C}} C$ with the property that $\forall A \in \mathcal{C}$, $\phi(A) \in A$.

Another way of looking at this is as follows. Suppose $\{A_i\}_{i \in I}$ is a collection of non-empty sets indexed by an index set I. A *choice function* is then defined as a map $\phi : I \to \bigcup_{i \in I} A_i$ such that $\phi(i) \in A_i$. The axiom of choice can then be rephrased.

The Axiom of Choice A.9.7 For every collection of nonempty sets there exists a choice function.

The axiom of choice is equivalent to a number of other very useful statements which are not at all obvious. Here they are, in no particular order.

Let X be a partially ordered set. The collection $\wp(X)$ can be partially ordered by inclusion, see A.9.2. This partial ordering on $\wp(X)$ is used in some of the statements below.

Hausdorff Maximality Principle A.9.8 Every partially ordered set X contains a totally ordered subset that is maximal with respect to the ordering on $\wp(X)$.

Zorn's Lemma A.9.9 If a non-empty partially ordered set has the property that every non-empty totally ordered subset has an upper bound, then the partially ordered set has a maximal element.

Well-Ordering Principle A.9.10 Every set can be well-ordered.

The following lemma is slightly complicated, but it will allow us to prove the equivalence of the above statements with little trouble.

Lemma A.9.11 Suppose that $(X, \leq)$ is a non-empty partially ordered set such that every non-empty totally ordered subset has a least upper bound. If $f : X \to X$ is such that $f(x) \geq x$ for all $x \in X$, then there is some $w \in X$ such that $f(w) = w$.

Proof. First we reduce to the case when X contains a least element, call it b. In fact, if X is nonempty choose any $b \in X$ and replace X by $X' = \{x \in X \mid x \geq b\}$. It is clear that X' is stable under f (that is $f(X') \subseteq X'$) and has the same properties as X. We call a subset Y of X "admissible" if

1. $b \in Y$

2. $f(Y) \subseteq Y$

3. Every lub of a totally ordered subset of Y belongs to Y.

X is certainly admissible, and the intersection of any family of admissible sets is admissible. Let W be the intersection of all admissible sets. The set $\{x|b \leq x\}$ is admissible, so if $y \in W$, then $b \leq y$.

We will now construct a totally ordered subset of W with the property that its least upper bound is a fixed point of f. Consider the set $P = \{x \in W|$ if $y \in W$ and $y < x$ then $f(y) \leq x\}$. Note that P is non-empty since $b \in P$. First we show that any element of P can be compared to any element of W and hence P is totally ordered.

Now fix an $x \in P$ and define $A_x = \{z \in W|z \leq x$ or $z \geq f(x)\}$. We would like to show that A_x is admissible.

1. Obviously, $b \in A_x$ since $b \leq x$.

2. Suppose $z \in A_x$. There are three possibilities. If $z < x$, $f(z) \leq x$ by the conditions of P, so $f(z) \in A_x$. If $z = x$, $f(z) = f(x) \geq f(x)$ so $f(z) \in A_x$. If $z \geq f(x)$, then $f(z) \geq z \geq f(x)$ so $f(z) \in A_x$.

3. Finally, let Y be a totally ordered non-empty subset of A_x, and let y_0 be the lub of Y in X. Then $y_0 \in W$, since W is admissible. If $z \leq x$ for all $z \in Y$ then $y_0 \leq x$ and hence $y_0 \in A_x$. Otherwise $z \geq f(x)$ for some $z \in Y$, which implies $y_0 \geq f(x)$, so $y_0 \in A_x$.

Thus, A_x is admissible.

Since A_x is an admissible subset of W, $A_x = W$. Put another way, if $x \in P$ and $z \in W$, then either $z \leq x$ or $z \geq f(x) \geq x$, and thus P is totally ordered. Therefore P has a least upper bound, call it x_0. Again $x_0 \in W$ and $f(x_0) \in W$ because W is admissible. We will now show $f(x_0) = x_0$. First we claim $x_0 \in P$. Indeed, if $y \in W$ and $y < x_0$, then there exists $x \in P$ with $y < x \leq x_0$, whence $f(y) \leq x \leq x_0$. Let $y \in W$ and suppose $y < f(x_0)$. As we saw above $A_{x_0} = W$, so we have $y \leq x_0$. If $y = x_0$, then $f(y) = f(x_0) \leq f(x_0)$. If $y < x_0$, then $f(y) \leq x_0 \leq f(x_0)$. In either case, we find $f(x_0) \in P$. Hence $f(x_0) \leq x_0 \leq f(x_0)$.

Whew!

Theorem A.9.12 (1) The Axiom of Choice, (2) Hausdorff Maximality Principle, (3) Zorn's Lemma, and (4) Well-Ordering Principle are all equivalent.

Proof. We will show that (1) implies (2), which implies (3), which implies (4), which implies (1), and then we will be done.

(1) ⇒ (2)

Take a non-empty partially ordered set $(E, \leq)$. Make $\mathcal{E}$, the family of totally ordered subsets of E, into a partially ordered set under inclusion. We wish to show that $\mathcal{E}$ has a maximal element (i.e., an element which is not smaller than any other element). So we will assume the opposite and reach a contradiction by applying Lemma A.9.11. We must first check to see if the lemma is applicable: Suppose $\mathcal{F}$ is a totally ordered subset of $\mathcal{E}$. Then it has a least upper bound, namely $\cup_{F \in \mathcal{F}} F$. Now, for a given $e \in \mathcal{E}$, let $S_e = \{x \in \mathcal{E}|e \subseteq x, e \neq x\}$. Then S_e can never be the empty set, because that would mean that e is maximal. So we apply the axiom of choice by defining a function $f : \{S_e|e \in \mathcal{E}\} \to \mathcal{E}$ with the property that $f(S_e) \in S_e$. Now define $g : \mathcal{E} \to \mathcal{E}$ by $g(e) = f(S_e)$. This gives us that $e \subsetneq g(e)$ for all $e \in \mathcal{E}$, contradicting the lemma.

(2) ⇒ (3)

Again, consider a partially ordered set $(E, \leq)$. Now let x be the upper bound for E_0, a maximal totally ordered subset of E. Suppose that there is some $y \in E$ such that $y > x$. Then $E_0 \cup \{y\}$ is a totally ordered set containing E_0, contradicting our assumption of maximality.

Exercise A.9.13 Now you finish the proof. Show that Zorn's Lemma implies the Well Ordering Principle, and that the Well Ordering Principle implies the Axiom of Choice.

A.10 Independent Projects

A.10.1 Basic Number Theory The following statements present a number of facts about elementary number theory. Prove all of these. If you don't understand some of the words, find a number theory book and look them up. Most of these facts will be used in Chapter **??** when we discuss p-adic numbers. The notation $a \equiv b \pmod{c}$ (pronounced "a is *congruent* to b *modulo* c") means that $c|(a-b)$.

1. The division algorithm: if $a, b \in \mathbb{Z}$ and $b \neq 0$, then there is a unique pair $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < |b|$.
2. If M is a subset of $\mathbb{Z}$ which is closed under subtraction and contains a nonzero element, then $M = \{np | n \in \mathbb{Z}\}$, where p is the least positive element of M.
3. If the greatest common divisor of a and b is $d = (a, b)$, then there exist $s, t \in \mathbb{Z}$ such that $d = sa + tb$.
4. Euclid's lemma: If p is prime and $p|ab$, then $p|a$ or $p|b$.
5. If $(a, c) = 1$, and $c|ab$, then $c|b$.
6. If $(a, c) = 1$, $a|m$ and $c|m$, then $ac|m$.
7. If $a > 0$ then $(ab, ac) = a(b, c)$.
8. $\mathbb{Z}$ has unique factorization, that is, if n is an integer greater than or equal to 2, then there exist unique distinct primes $p_1, \ldots, p_k$ and exponents $\alpha_1, \ldots, \alpha_k$ greater than or equal to one such that $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$.
9. If $a \equiv b \pmod{m}$, then $-a \equiv -b \pmod{m}$, $a + x \equiv b + x \pmod{m}$, and $ax \equiv bx \pmod{m}$ for every $x \in \mathbb{Z}$.
10. If $(c, m) = 1$ and $ca \equiv cb \pmod{m}$, then $a \equiv b \pmod{m}$.
11. If $(c, m) = 1$, then $cx \equiv b \pmod{m}$ has a unique solution x modulo m.
12. If p is prime and $c \not\equiv 0 \pmod{p}$, then $cx \equiv b \pmod{p}$ has a unique solution x modulo p.
13. If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then $a + c \equiv b + d \pmod{m}$ and $ac \equiv bd \pmod{m}$.
14. If $a, b, c \in \mathbb{Z}$ and $d = (a, b)$, then $ax + by = c$ has a solution in integers x, y if and only if $d|c$.
15. If $[a, b]$ is the least common multiple of a and b, then $m[a, b] = [ma, mb]$ when $m > 0$.
16. If $ca \equiv cb \pmod{m}$ and $d = (c, m)$, then $a \equiv b \pmod{\frac{m}{d}}$.
17. If $m, a, b \in \mathbb{Z}$, the congruence $ax \equiv b \pmod{m}$ is solvable if and only if $(a, m)|b$. There are exactly (a, m) solutions distinct modulo m.
18. If $a, b, s, t \in \mathbb{Z}$ are such that $sa + tb = 1$, then $(a, b) = 1$.

    Now suppose that P is the set of integers between 1 and $m - 1$, inclusive, which are relatively prime to m. A *reduced residue system* modulo m is a set of integers such that each of integer in P is congruent modulo m to exactly one of the elements in this set.
19. The number of elements in a reduced residue system modulo m is independent of the representatives chosen.
20. If p is a prime and ϕ denotes Euler's ϕ function (where $\phi(a)$ is the number of integers between 0 and a which are relatively prime to a), then $\phi(p^n) = p^n - p^{n-1} = p^n(1 - \frac{1}{p})$.
21. The number of elements in a reduced residue system modulo m is $\phi(m)$.

22. If $a_1, \ldots, a_{\phi(m)}$ is a reduced residue system modulo m and $(\kappa, m) = 1$, then $\kappa a_1, \ldots, \kappa a_{\phi(m)}$ is a reduced residue system modulo m.

23. If m is a positive integer and $(\kappa, m) = 1$, then $\kappa^{\phi(m)} \equiv 1 \pmod{m}$.

24. If $d_1, \ldots, d_k$ are the positive divisors of n, then $\sum_{i=1}^{k} \phi(d_i) = n$.

A.10.2 The Complete Independence of Axiom Systems The rules of arithmetic and order which characterize the integers are also known as *axioms.* In general, an axiom is an assumption or rule that we accept without proof. In fact, we made cancellation for multiplication an axiom for an integral domain precisely because we could not prove it from the other axioms of addition and multiplication. A group of axioms is called an *axiom system.*

There are a number of questions we can ask about a given axiom system S. First, is it *consistent*? That is, is there a model for S? For instance, the axioms for a field have a model, namely the integers modulo 2.

The next question we could ask is whether any axiom A in S is *independent*. What we mean is, could we replace A with its negation, $\overline{A}$, and still have a consistent system? (Symbolically, we would represent our new axiom system as $(S - A) + \overline{A}$.)

Exercise: What is the negation of "P(b), for all $b \in B$"? What about the negation of "P(b), for some $b \in B$"?

Consider the axioms for an equivalence relation. Clearly, they are consistent, because the relation of equality on any set is a model. Suppose we remove axiom 1) (reflexivity) and replace it by its negation, $\bar{1}$), by which we mean, $(a, a) \notin R$ for some a in A. Can we come up with a model for $\bar{1}$), 2), and 3)? To do this, pick $a \in A$ and do not include $(a, a) \in R$. For the remaining elements you can fix it up so that 2 and 3 work.

Exercise: State $\bar{2}$) and $\bar{3}$) for the equivalence relation axioms (non-symmetry and non-transitivity.) How is non-symmetry different from anti-symmetry?

The axiom system S is called *independent* if each of its axioms is independent. It is called *completely independent* if, for any subset S_1 of S, the system $(S - S_1) + \overline{S_1}$ is consistent.

Exercise A.10.1 Show that the axioms for an equivalence relation are completely independent. You can do this by providing models for {1, 2, 3}; {$\bar{1}$, 2, 3}; {1, $\bar{2}$, 3}; {1, 2, $\bar{3}$}; {$\bar{1}$, $\bar{2}$, 3}; {$\bar{1}$, 2, $\bar{3}$}; {1, $\bar{2}$, $\bar{3}$}; and {$\bar{1}$, $\bar{2}$, $\bar{3}$}. Your models can, but need not, be based on relations you've seen before. Or, you could invent a relation on a set which satisfies the necessary axioms. For example, a model for {1, $\bar{2}$, $\bar{3}$} could be the relation on $\{a, b, c\}$ defined by $(a, a), (b, b), (c, c), (a, b), (b, c) \in R$.

A.10.3 Ordered Integral Domains This project is designed to show that any ordered integral domain contains a copy of the integers. Thus, in particular, any ordered field such as the rationals or real numbers contains a copy of the integers. Let $(R, +, \cdot, <)$ be an ordered integral domain.

Definition A.10.2 An inductive set in R is a subset S of R such that

a. $1 \in S$,

b. if $x \in S$, then $x + 1 \in S$.

Example A.10.3 *i.* R is an inductive subset of R.

ii. $S = \{x \in R \mid x \geq 1\}$ is an inductive subset of R.

Now define N to be the intersection of all the inductive subsets of R. It is clear that N is an inductive subset of R. Of course, N is supposed to be the natural numbers. Since of all the axioms for a commutative ring with 1, as well as the order axioms hold in R, we can use them freely in N. The following facts are easy to prove, so prove them.

Facts A.10.4 1. Suppose that S is a non-empty subset of N such that $1 \in S$ and if $x \in S$ then $x+1 \in S$, show that $S = N$.

2. Show that N is closed under addition.

3. Show that N is closed under multiplication. Hint: fix $x \in N$ and look at the set $M_x = \{y \in N \mid xy \in N\}$. Then $1 \in M_x$. If $y \in M_x$, then $x(y+1) = xy + x$, and xy is in N by the induction hypothesis. Since N is closed under addition, $xy + x \in N$. Hence $y + 1 \in M_x$ and $M_x = N$.

4. Show that the well ordering principle holds in N.

   This is all fine, but where do we get the integers? Well, of course, we just tack on 0 and the negative natural numbers. Before nodding your head and shouting "Hooray!", you must show that this new set $Z = N \cup \{0\} \cup -N$ is closed under multiplication and addition.

5. Show that Z is closed under addition.
   This is a little tricky and requires the following fact. If $m, n \in N$ then $m - n \in Z$. In particular if $m \in N$ then $m - 1 \in Z$.

6. Show that Z is closed under multiplication.

   So we have that Z is an ordered integral domain in which the positive elements are well ordered.

7. Show that Z and the integers, $\mathbb{Z}$, are order isomorphic. That is there exists a bijection $\phi : Z \to \mathbb{Z}$ such that

   (a) $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in Z$,

   (b) $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in Z$,

   (c) if $x < y$ in Z, then $\phi(x) < \phi(y)$ in $\mathbb{Z}$.