Analysis I

Trevor M. Leslie

Contents

Introduction

Defining what 'analysis' means as a mathematical discipline is a bit of a difficult task, largely because of its vast scope. However, it is perhaps useful to have a working definition of the subject of these notes before beginning in earnest. Therefore we'll give a brief discussion of what analysis is, with the caveat that it just a first approximation. Analysis involves drawing qualitative (and sometimes quantitative) conclusions about functions when only limited information is initially available. What exactly does 'limited information' mean, then? One of the most relevant situations here is when an exact formula for a function is not available or not useful. Alternatively, deriving a formula may involve a computation that is impractical or impossible. The basic tools of analysis are designed so that it is still possible in many cases to extract useful information.

The main concepts of analysis are convergence, continuity, differentiation, and integration. The reader should be familiar with each of these concepts from courses in Calculus. However, these concepts are often treated in a less-than-rigorous manner in Calculus courses; even when rigor is not lacking, these concepts are introduced in a very limited context, which needs to be expanded before one can go deeper into analysis. These notes aim to put the main theorems of Calculus on a rigorous footing and broaden their scope somewhat. However, the reader should be aware that the context provided by these notes is only the 'tip of the iceberg', so to speak. The immense power that analysis can bring to mathematical and physical problems is only really accessible after several more passes.

In the interest of rigor, it would be appealing to develop the theory 'from scratch', so to speak. However, these notes are intended to accompany a one-semester course; therefore such a development is practically impossible. We choose a starting point that is compatible with the goal of developing the main concepts in analysis, while sacrificing as little rigor as possible. We assume to begin with that the reader is familiar with the arithmetic of the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, and the rational numbers $\mathbb{Q}$. Furthermore, we will take a 'naive' rather than 'axiomatic' approach to set theory (these are terms of art). However, the shortcuts we take here can be justified (and are subtle enough that most readers will not notice); the interested reader can fill in the gaps by taking courses in logic and in algebra.

While we assume knowledge of the basic properties of $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$, we assume essentially nothing about the set $\mathbb{R}$ of real numbers. This is because $\mathbb{R}$ is a more complicated set. Indeed, starting from $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$, can you write down a definition of $\mathbb{R}$? Almost any way you try to define it will implicitly make use of limits. For example, suppose we try to define $\mathbb{R}$ in terms of infinite decimals. The question then arises: What exactly is an infinite decimal? The natural definition is as a limit or supremum of finite decimals. But then we must ask what exactly is meant by a 'limit', and whether the supremum is guaranteed to exist. In short, rigorously defining $\mathbb{R}$ starting from $\mathbb{Q}$ is far from straightforward. Giving the 'right' definition of $\mathbb{R}$ will be the subject of this second chapter.

When writing these notes, the author referred frequently to several textbooks, including Walter Rudin's *Principles of Mathematical Analysis*, Stephen Abbott's *Understanding Analysis*, James Munkres' *Topology*, and Andrew Browder's *Mathematical Analysis: An Introduction*. A few sections in these notes follow various parts of the above texts rather closely. No copyright violation is intended; however, as these notes will be posted publicly, please email me if you are a publisher or author who happens upon these notes and has any objections to the presentation.

Finally, these notes are a work in progress. Please email lesliet@usc.edu for any corrections or suggestions for improvements.

Part 1

Basics and Preliminaries

CHAPTER 1

Naive Set Theory

In this chapter, we build the basic set-theoretic terminology and machinery needed for the rest of the course. For a combination of reasons (time and space considerations, pedagogical sensibility, expertise of the instructor), We take the approach of 'Naive Set Theory' rather than an axiomatic approach. Basically, all this means is that we assume that the notion of a 'set' is intuitively clear. The study of what exactly a set is belongs to the realm of logic and mathematical foundations. To illustrate that this issue is a non-triviality, consider the following.

Russell's Paradox: Let R be the set of all sets that are not members of themselves. Is $R \in R$?

Well-known Layman's Reformulation: Suppose a barber cuts the hair of exactly those people who do not cut their own hair. Does the barber cut his or her own hair?

Both of these questions seem impossible to answer. The logical resolution is that such an object R cannot actually be a set, and that such a barber cannot exist. That is, the extent to which one can abstractly manipulate sets to create new ones is not without limitations. Therefore, when constructing new sets, one has to use a bit of care.

The above might induce some existential worry in the minds of more dramatic readers. What *is* a set, anyway? What are the 'rules' of mathematics? Has math been deceiving me all this time? Fortunately, logicians have got us covered. They have carefully and painstakingly crafted a set of axioms on which all of mathematics can be based. Unfortunately, acquiring a thorough understanding of these axioms and their consequences is a major undertaking. We will forgo a discussion of these axioms in these notes. Instead, we assure the reader that the logic we use is justified by the axioms. The reader who wishes for a more careful and systematic treatment of set theory should consider taking a course in logic or mathematical foundations.

1. Sets and Set Operations

1.1. Sets and Subsets.

DEFINITION 1.1. A *set* is any collection of objects, or *elements*. Usually a set is denoted by a capital letter, such as A, and its elements are denoted by lowercase letters, such as x. The notation $x \in A$ is equivalent to the statement "x is an element of the set A". If A and B are sets and every element of A belongs to B (i.e., $x \in A$ implies $x \in B$), then we say that A is a *subset* of B, or that A is *contained in* B, and we write $A \subset B$. We say that two sets are the same, or *equal*, if $B \subset A$ and $A \subset B$; in this case we write $A = B$.

Note that all the notation in the definition above is 'reversible': $x \in A$ is the same as $A \ni x$, and $A \subset B$ is the same as $B \supset A$.

If $A \subset B$ and $A \neq B$, we sometimes write $A \subsetneq B$. This is more specific than $A \subset B$, since the latter notation allows for the possibility that $A = B$. However, the notation $A \subset B$ is usually used unless we wish to explicitly stress the non-equality.

An emphatically different notation is $A \not\subset B$. This means that $A \subset B$ *fails*, i.e. A is *not* a subset of B, i.e. there exists $a \in A$ such that $a \notin B$. Do not confuse the two notations $A \not\subset B$ and $A \subsetneq B$.

We assume familiarity with the following sets:

- The integers $\mathbb{Z}$.
- The natural numbers (i.e. positive integers) $\mathbb{N}$.
- The rational numbers $\mathbb{Q}$.

Later, we will also assume familiarity with the usual algebraic operations on these sets (e.g., addition, multiplication, etc.).

Sometimes it is useful to write down all the elements of a set, either explicitly, by suggesting a pattern, or by specifying a rule. The usual notation here is illustrated in the following examples.

$$A = \{1, 2, \tfrac{7}{2}, 9, 3, 14\}$$

$$B = \{2, 4, 6, \ldots, 100\}.$$

$$C = \{1, 4, 7, 10, \ldots\}.$$

$$D = \{x \in A : x \text{ is an even integer}\} = \{2, 14\}.$$

In the last example, the colon is read "such that". The notation used here is called *set-builder notation*. It should be familiar and will not be explained further.

DEFINITION 1.2. The *empty set* is the set with zero elements, usually denoted either by $\emptyset$ or $\{\}$. Any set that is not empty is called *nonempty*.

Note that if A is any set whatsoever, then $\emptyset \subset A$, since the statement "every element of $\emptyset$ is an element of A" is vacuously true. If $\emptyset \subsetneq A \subsetneq B$, then A is said to be a *proper subset* of B.

1.2. Collections of Sets. Certain kinds of sets often carry slightly different terminology and notation. For example, it may be the case that each element of a set is itself a set! In this case the "set of sets" is often referred to as a *collection* of sets, and is usually denoted with a scripted capital letter, such as $\mathcal{A}$.

EXAMPLE 1.3. If $A = \{1, 2, 3\}$, $B = \{1, 2\}$, $C = \{1\}$, then $\mathcal{A} = \{A, B, C\} = \{\{1, 2, 3\}, \{1, 2\}, \{1\}\}$ is a collection of sets. To clarify the notation, note that we write (for example) $B \in \mathcal{A}$, but $B \subset A$; we also write (for example) $1 \in B$. This is consistent with the definition of the symbols $\in$ and $\subset$.

EXAMPLE 1.4. Given a set A, one can form its *power set* $\mathcal{P}(A)$, which consists of all subsets of A. For example, if $A = \{1, 2, 3\}$, then

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

1.3. Binary Set Operations.

DEFINITION 1.5. Given two sets A and B, their *union* $A \cup B$ is defined to be the set consisting of all elements belonging either to A or to B. The *intersection* $A \cap B$ of A and B is defined to be the set of elements belonging to *both* A and B.

$$A \cup B = \{x : x \in A \text{ or } x \in B\}, \qquad A \cap B = \{x : x \in A \text{ and } x \in B\},$$

If $A \cap B = \emptyset$, we say that A and B are *disjoint*. If $A \cap B \neq \emptyset$, then we say that A and B *intersect*.

DEFINITION 1.6. Let A and B be sets. The *relative complement* of A in B, denoted $B \backslash A$, is defined to be the set of elements of B which are not in A.

$$B \backslash A = \{x \in B : x \notin A\}.$$

Sometimes one works within the context of some very large set X, and all other sets under consideration are understood to be subsets of X. In this case (and if X is clear from context), we define the *absolute complement* of A by

$$A^c = \{x \in X : x \notin A\}.$$

Note that when A and B are both subsets of a large set such as the set X in the definition above, one has

$$B\backslash A = B \cap A^c.$$

In practice, one often drops the qualifiers 'relative' and 'absolute' from the terminology.

EXERCISE 1.1. Let A and B be subsets of another set X. Prove the following statements.

(a) $A \cap B = A\backslash(A\backslash B)$
(b) $A \subset B$ if and only if $X\backslash A \supset X\backslash B$.

Note: For any sets A and B, we have

$$A = (A\backslash B) \cup (A \cap B).$$

In particular, $A = (A\backslash B) \cup B$ if and only if $B \subset A$ (in which case $A \cap B = B$).

The intersection and union satisfy the following basic properties.

- (Commutativity) $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
- (Associativity) $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$
- (Distributivity) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

In light of the commutativity and associativity, we may define the union of n sets inductively by

$$A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_{n-1} \cup A_n := (\cdots((A_1 \cup A_2) \cup A_3) \cup \cdots \cup A_{n-1}) \cup A_n,$$

and similarly for intersections.

DEFINITION 1.7. An *ordered pair* in a set X is a pair of elements of X listed in an order. We use the notation (x_1, y_1) to denote ordered pairs. We say that two ordered pairs (x_1, y_1) and (x_2, y_2) are *equal* if $x_1 = x_2$ and $y_1 = y_2$.

DEFINITION 1.8. Given two sets A and B, their pairwise *product* $A \times B$ consists of all ordered pairs (a, b) in $A \cup B$, such that $a \in A$ and $b \in B$:

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\}.$$

1.4. General Unions and Intersections.

DEFINITION 1.9. If $\mathcal{A}$ is a collection of subsets of a set X, then the *union* of all sets in $\mathcal{A}$ consists of all elements of every set in $\mathcal{A}$:

$$\bigcup_{A\in\mathcal{A}} A = \{x \in X : x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

The *intersection* of all sets in $\mathcal{A}$ consists of those elements of X which are in every set of $\mathcal{A}$:

$$\bigcap_{A\in\mathcal{A}} A = \{x \in X : x \in A \text{ for every } A \in \mathcal{A}\}.$$

If $\mathcal{A} = \{A_1, \ldots, A_n\}$, then clearly

$$\bigcup_{A\in\mathcal{A}} A = A_1 \cup \cdots \cup A_n,$$

and similarly for the intersection. We will introduce some more convenient notation later on.

A rather annoying example which is worth mentioning explicitly is the case where $\mathcal{A} = \emptyset$. In this case the union over $\mathcal{A}$ is $\emptyset$, while the intersection over $\mathcal{A}$ is all of X. The reader can check that these statements are (vacuously) true.

General unions and intersections satisfy properties analogous to those listed in the previous subsection for binary unions and intersections (commutativity, associativity, distributivity). These will be used without mention throughout the notes. However, we point out two particularly useful equalities:

THEOREM 1.10 (DeMorgan's Laws). *If $\mathcal{A}$ is a collection of sets and X is another set, then the following two equalities hold.*

$$X\backslash\Big(\bigcup_{A\in\mathcal{A}} A\Big) = \bigcap_{A\in\mathcal{A}} (X\backslash A). \tag{1}$$

$$X\backslash\Big(\bigcap_{A\in\mathcal{A}} A\Big) = \bigcup_{A\in\mathcal{A}} (X\backslash A). \tag{2}$$

PROOF. Assume x is an element on the left side of (1). Then $x \in X$, but x does not belong to any $A \in \mathcal{A}$. Thus $x \in X\backslash A$, for every $A \in \mathcal{A}$; that is, x belongs to the set on the right side of (1). This shows that LHS$\subset$RHS for (1). To prove the opposite inclusion, assume that x belongs to the set on the right side of (1). Then $x \in X\backslash A$ for every $A \in \mathcal{A}$. Thus $x \in X$, but $x \notin A$ for any $A \in \mathcal{A}$, i.e. $x \notin \bigcup_{A\in\mathcal{A}} A$. Therefore x belongs to the set on the left side of (1). This shows that LHS = RHS for (1).

To prove (2), we use (1). Both sets in (2) are subsets of X, therefore

$$X\backslash\Big(\bigcup_{A\in\mathcal{A}} (X\backslash A)\Big) = \bigcap_{A\in\mathcal{A}} X\backslash(X\backslash A) = \bigcap_{A\in\mathcal{A}} A.$$

Taking complements in X one more time, we thus obtain

$$\bigcup_{A\in\mathcal{A}} (X\backslash A) = X\backslash\Big(\bigcap_{A\in\mathcal{A}} A\Big),$$

as claimed. □

2. Relations

DEFINITION 2.1. A *relation* R from a set A to a set B is a subset of $A \times B$. We sometimes write aRb to mean that $(a, b) \in R$. If A is a set, a relation 'on A' means a relation between A and itself, i.e. a subset of $A \times A$.

DEFINITION 2.2. Let R be a relation on a set A. Then

- R is said to be *reflexive* if aRa for every $a \in A$.
- R is said to be *symmetric* if aRb implies bRa.
- R is said to be *antisymmetric* if aRb and bRa imply that $a = b$.
- R is said to be *transitive* if aRb and bRc together imply that aRc.

Though you may not realize it, several familiar concepts can be described using relations. Since the notation aRb may be new, we allow some redundancy in the following definitions for the sake of clarity.

DEFINITION 2.3. A *partial order* $\preceq$ on a set A is a relation on A that satisfies the following properties:

- (Reflexivity) $a \preceq a$, for all $a \in A$.
- (Antisymmetry) If $a \preceq b$ and $b \preceq a$, then $a = b$.
- (Transitivity) If $a \preceq b$ and $b \preceq c$, then $a \preceq c$.

A *total order* $\leq$ on a set A is a partial order which has the following additional property:

- (Comparability) If $a, b \in A$, then at least one of the statements $a \leq b$ or $b \leq a$ must hold.

EXAMPLE 2.4. The usual meaning 'less than or equal to' of the symbol $\leq$ constitutes a total ordering on $\mathbb{Q}$, or any subset thereof. For small sets, the relation can be given explicitly. If $A = \{1, 2\}$, then $\leq$ is the set $\{(1, 1), (1, 2), (2, 2)\}$.

If $\mathcal{A}$ is a collection of sets, then $\subset$ is a partial ordering on $\mathcal{A}$; we also say that $\mathcal{A}$ is *partially ordered by inclusion*. However, $\subset$ may or may not be a total order on $\mathcal{A}$. For example, if $\mathcal{A} = \{\{1, 2\}, \{1\}, \{2\}\}$,

then neither of the statements $\{1\} \subset \{2\}$ or $\{2\} \subset \{1\}$ is true (the sets $\{1\}$ and $\{2\}$ are not *comparable*); therefore $\mathcal{A}$ is not ordered by inclusion. However, the collection $\mathcal{B} = \{\{1\}, \{1,2\}, \{1,2,3\}\}$ is (totally) ordered by inclusion. If a collection $\mathcal{A}$ is totally ordered by inclusion, it is said to be a collection of *nested sets*.

REMARK 2.5. Technically, any symbol R can be used to denote an order relation (partial or total). However, it is often useful to have compact notation for the situation where aRb but $a \neq b$. Whenever the symbols $\preceq$, $\leq$, $\subset$ are used to denote (partial or total) orders, it should be understood that the corresponding symbols $\prec$, $<$, $\subsetneq$ specify non-equality of the related elements.

The symbols usually used for order relations are also 'reversible': $b \succeq a$ means $a \preceq b$; $b \geq a$ means $a \leq b$, etc.

DEFINITION 2.6. An *equivalence relation* $\sim$ on a set A is a relation on A that satisfies the following properties.

- (Reflexivity) $a \sim a$ for all $a \in A$.
- (Symmetry) $a \sim b$ implies $b \sim a$.
- (Transitivity) If $a \sim b$ and $b \sim c$, then $a \sim c$.

If A is the set of all people on the planet, then the relation defined by 'Person 1$\sim$ Person 2 if and only if Person 1 and Person 2 have the same birthday' is an equivalence relation.

DEFINITION 2.7 (Function). A *function* f from A to B is a relation between two sets A and B that satisfies the following properties:

- For each $a \in A$, we have afb for some $b \in B$.
- If afb and afc hold, then $b = c$.

We almost always use the notation $f(a) = b$ instead of afb. A is called the *domain*, and B is called the *codomain*; we express this by writing $f : A \to B$.

The definition of 'function' above can easily be seen to be equivalent to the following probably more familiar variant: Let A and B be sets, and suppose that to each $x \in A$ there is exactly one associated element $f(x) \in B$. Then the association f is called a *function* from A to B. If $f : A \to B$ and $g : A \to B$ be functions, then f and g are *equal* (as functions) if and only if $f(a) = g(a)$ for all $a \in A$. It is easy to check that this is equivalent to equality of f and g as relations.

We now enter into a much deeper discussion of functions.

3. Functions

3.1. Basic Definitions and Notation.

DEFINITION 3.1 (Some Important Kinds of Functions).

(a) Given a set A, the map $\mathrm{id}_A : A \to A$ defined by $\mathrm{id}_A(a) = a$ for all $a \in A$ is called the *identity* map on A.
(b) If $A \subset B$, the map $\iota : A \to B$ defined by $\iota(a) = a$ for all $a \in A$ is called the *inclusion* map from A to B.
(c) If $f : A \to B$ is a function and $C \subset A$, then the function $f\big|_C : C \to B$ defined by $(f\big|_C)(c) = f(c)$ for all $c \in C$ is called the *restriction* of f to C.
(d) If $f : A \to B$ is a function such that for all $a \in A$ we have $f(a) \in D \subset B$, then the map $g : A \to D$ defined by $g(a) = f(a)$ is referred to as the function formed by *restricting the codomain* of f to D. (There is no standard notation for this function.)
(e) If $f : A \to B$ and $g : B \to C$ are functions, then the *composition* of f and g is the function $g \circ f : A \to C$ defined by $(g \circ f)(a) = g(f(a))$ for all $a \in A$. (If $f : A \to B$ and $g : B' \to C$ are functions with $B \subset B'$, we use the notation $g \circ f$ to denote $(g\big|_B) \circ f$.)

DEFINITION 3.2. Let $f : A \to B$ be a function.

(a) If $E \subset A$, then the *image* $f(E)$ of E under f is defined by
$$f(E) = \{f(x) : x \in E\}.$$
The image $f(A)$ of the entire domain A is called the *image* of f, or the *range* of f. The elements of $f(A)$ are called the *values* of f. If $f(A) = B$, we say that f maps A *onto* B, or that $f : A \to B$ is *surjective*.

(b) If $G \subset B$, then the *inverse image* $f^{-1}(G)$ of G under f is defined as
$$f^{-1}(G) = \{x \in A : f(x) \in G\}.$$
The inverse image is sometimes called the *preimage*. If $y \in B$, then we define $f^{-1}(y) = f^{-1}(\{y\})$. If $f^{-1}(y)$ contains at most one element of A for each $y \in B$, then f is said to be *injective*, or a *one-to-one* mapping of A into B.

(c) If f is one-to-one and onto (injective and surjective), then we say that f is *bijective*.

REMARK 3.3. A few comments are in order:

- Another formulation of injectivity is the following: f is injective if and only if $f(x) = f(y)$ implies $x = y$.
- If $f : A \to B$ is any function, then the function formed by restricting the codomain of f to $f(A)$ is a surjection. In particular, such a restriction results in a bijection if the original function f is injective.
- If $\iota : C \to A$ is the inclusion map, then $f\big|_C = f \circ \iota$.
- The image of a function $f : A \to B$ is sometimes denoted by $\operatorname{Im} f$ rather than by $f(A)$.

One additional, especially important remark: In general, one can't 'cancel' f and f^{-1}, the way the notation might tempt one to do. Convince yourself of this by looking at the following Example.

EXAMPLE 3.4. Let $A = \{1, 2, 3\}$, $B = \{4, 5, 6\}$ and define $f : A \to B$ by $f(1) = f(2) = f(3) = 5$. Then
$$f^{-1}(f(\{1, 2\})) = f^{-1}(\{5\}) = \{1, 2, 3\} \neq \{1, 2\}$$
$$f(f^{-1}(\{4, 5\})) = f(\{1, 2, 3\}) = \{5\} \neq \{4, 5\}.$$

However, if $f : A \to B$ is a function and $C \subset A$ and $D \subset B$, we *do* have that[1]
$$f^{-1}(f(C)) \supset C \qquad \text{and} \qquad f(f^{-1}(D)) \subset D.$$

In fact, a little more is true:

EXERCISE 3.1. Let $f : A \to B$ be a function. Prove the following statements:

(a) f is injective if and only if $f^{-1}(f(C)) = C$ for every subset C of A.
(b) f is surjective if and only if $f(f^{-1}(D)) = D$ for every subset D of B.

EXERCISE 3.2. Let $f : A \to B$ and $g : B \to C$ be functions.

(a) Prove that if f and g are both injective, then so is $g \circ f$.
(b) Prove that if f and g are both surjective, then so is $g \circ f$.
(c) Prove that if $g \circ f$ is surjective, then so is g.
(d) Argue that surjectivity of $g \circ f$ does not imply surjectivity of f, by providing explicit examples of functions f and g for which $g \circ f$ is surjective but f is not. You should explicitly demonstrate that your functions have the desired properties.

[1] Proof: If $a \in C$, then $f(a) \in f(C)$, which means that $a \in f^{-1}(f(C))$ by definition of the inverse image. If $b \in f(f^{-1}(D))$, then there exists $a \in f^{-1}(D)$ such that $f(a) = b$. But since $a \in f^{-1}(D)$, we have $f(a) \in D$ by definition of the inverse image. Thus $b \in D$.

(e) Prove that if $g \circ f$ is injective, then so is f.
(f) Argue that injectivity of $g \circ f$ does not imply injectivity of g. Format your answer similarly to part (d).

In parts (d) and (f), your examples will be easiest to construct if you choose A, B, and C to be sets with very few elements.

3.2. Functions and Set Operations. In this section, we summarize the relationship between the image/preimage of a function and the set operations union, intersection, and complement.

PROPOSITION 3.5. *Let $f : X \to Y$ be a function.*

(a) For any collection $\mathcal{A}$ of subsets of X, we have

$$f\left(\bigcup_{A\in\mathcal{A}} A\right) = \bigcup_{A\in\mathcal{A}} f(A), \tag{3}$$

$$f\left(\bigcap_{A\in\mathcal{A}} A\right) \subset \bigcap_{A\in\mathcal{A}} f(A). \tag{4}$$

(b) f is injective if and only if for every nonempty collection $\mathcal{A}$ of subsets of X, we have

$$f\left(\bigcap_{A\in\mathcal{A}} A\right) = \bigcap_{A\in\mathcal{A}} f(A).$$

(c) If $\mathcal{B}$ is a collection of subsets of Y, then

$$f^{-1}\left(\bigcup_{B\in\mathcal{B}} B\right) = \bigcup_{B\in\mathcal{B}} f^{-1}(B). \qquad f^{-1}\left(\bigcap_{B\in\mathcal{B}} B\right) = \bigcap_{B\in\mathcal{B}} f^{-1}(B).$$

PROOF. The proofs of (a) and (c) follow basically by unraveling the definitions and are omitted. We prove only statement (b), considering the contrapositive of both implications. Note that the direction ($\Longleftarrow$) in the proof explains why we cannot in general replace the inclusion with an equality in (4).

($\Longrightarrow$) Let $\mathcal{A}$ be a nonempty collection of subsets of X. For convenience, we write

$$Y_1 = f\left(\bigcap_{A\in\mathcal{A}} A\right), \qquad Y_2 = \bigcap_{A\in\mathcal{A}} f(A).$$

Assume that $Y_2 \neq Y_1$. Then by part (a), we have $Y_2 \not\subset Y_1$; that is, there exists $y \in Y_2$ such that $y \notin Y_1$. Suppose such a y has been chosen. Then $y \in \mathrm{Im}(f)$; therefore choose $x \in X$ such that $y = f(x)$. This x cannot be an element of $\bigcap_{A\in\mathcal{A}} A$; if it were, than we would have $y = f(x) \in Y_1$, contradicting our assumption. Therefore there exists $\widetilde{A} \in \mathcal{A}$ such that $x \notin \widetilde{A}$. But it is still true that $y \in f(\widetilde{A})$, so we can pick $\widetilde{x} \in \widetilde{A}$ such that $f(\widetilde{x}) = y = f(x)$. On the other hand, $x \neq \widetilde{x}$, since $\widetilde{x}$ is an element of $\widetilde{A}$ and x is not. This proves that f is not injective, completing the proof of the implication ($\Longrightarrow$).

($\Longleftarrow$) Assume that f is not injective, i.e. there exist $x_1, x_2 \in X$ and $y \in Y$ such that $f(x_1) = f(x_2) = y$, but $x_1 \neq x_2$. Consider the collection $\mathcal{A} = \{\{x_1\}, \{x_2\}\}$. Then

$$f\left(\bigcap_{A\in\mathcal{A}} A\right) = f(\{x_1\} \cap \{x_2\}) = f(\emptyset) = \emptyset,$$

while

$$\bigcap_{A\in\mathcal{A}} f(A) = f(\{x_1\}) \cap f(\{x_2\}) = \{y\} \cap \{y\} = \{y\}.$$

That is, the equality under consideration does not hold for the collection $\mathcal{A}$. This completes the proof of the implication ($\Longleftarrow$). □

EXERCISE 3.3. Let $f : X \to Y$ be a function. Prove the following statements.

(a) If A and C are subsets of X, then $f(C\backslash A) \supset f(C)\backslash f(A)$.
(b) f is injective if and only $f(C\backslash A) = f(C)\backslash f(A)$ for any two subsets A and C of X.
(c) If B and D are subsets of Y, then $f^{-1}(D\backslash B) = f^{-1}(D)\backslash f^{-1}(B)$.

3.3. Function Inverses.

DEFINITION 3.6 (Function Inverses). Let $f : A \to B$ be a function.

- A *left inverse* for f is a function $g : B \to A$ such that $g \circ f = \mathrm{id}_A$.
- A *right inverse* for f is a function $h : B \to A$ such that $f \circ h = \mathrm{id}_B$.
- A *two-sided inverse* (or simply an *inverse*) for f is a function $k : B \to A$ which is a right inverse and a left inverse.

Note: g is a left inverse for f if and only if f is a right inverse for g.

REMARK 3.7. Let $f : A \to B$ and $g : B \to A$ be such that $g \circ f = \mathrm{id}_A$. Then since id_A is bijective, Exercise 3.2 tells us that g must be surjective and f must be injective. Thus any left inverse is surjective, and any right inverse is injective. Consequently, any two-sided inverse must be bijective.

THEOREM 3.8. *Let $f : A \to B$ be a function.*

(a) f is injective if and only if it has a left inverse.
(b) f is surjective if and only if it has a right inverse.

PROOF. Remark 3.7 above essentially proves the direction ($\Longleftarrow$) for both statements. Indeed, if f has a left inverse g, then f is a right inverse for g, so f is injective. On the other hand, if f has a right inverse h, then f is a left inverse for h, so f is surjective.

Now we prove the direction ($\Longrightarrow$).

(a) Assume f is injective. Then for each $y \in B$, the set $f^{-1}(y)$ contains at most one element. Pick $x_0 \in A$ arbitrarily. Define a function $g : B \to A$ as follows. If $y \in f(A)$, define $g(y)$ to be the (unique) element of the set $f^{-1}(y)$; otherwise put $g(y) = x_0$. We claim that $g \circ f = \mathrm{id}_A$. Indeed, for any $x \in A$, we have $\{x\} = f^{-1}(f(x))$ since f is injective. On the other hand, $g(f(x))$ is by definition the unique element of the set $f^{-1}(f(x)) = \{x\}$, thus $g(f(x)) = x$. That is, $g \circ f = \mathrm{id}_A$.

(b) Assume f is surjective. Then for each $y \in B$, the set $f^{-1}(y)$ is nonempty. Define a function $h : B \to A$ as follows[2]. For each $y \in B$, let $h(y)$ be an arbitrary element of $f^{-1}(y)$. We claim that $f \circ h = \mathrm{id}_B$. Indeed, if $y \in B$, then $h(y) \in f^{-1}(y)$, so that $f(h(y)) = y$. This proves the claim. □

PROPOSITION 3.9. *If $f : A \to B$ has a both left inverse g and a right inverse h, then $g = h$. That is, $g = h$ is actually a two-sided inverse for f.*

PROOF. For any $b \in B$, we have $g(b) = g((f \circ h)(b)) = g(f(h(b))) = (g \circ f)(h(b)) = h(b)$. □

REMARK 3.10. The Proposition above guarantees that if a function f has a two-sided inverse g, then g is unique. Therefore it makes sense to use the notation f^{-1} to denote this two-sided inverse. To be more specific: Normally, f^{-1} is technically a function from $\mathcal{P}(B)$ to $\mathcal{P}(A)$. If f has a two-sided inverse function g, then $f^{-1}(\{b\}) = \{g(b)\}$ for all $b \in B$. Therefore it is harmless to use f^{-1} and g interchangeably in this case when f is bijective. A word of caution is of course in order here, though, since the inverse of a function does *not* always exist. If $f : A \to B$ does not have an inverse, then the notation $f^{-1}(a)$ reverts to its usual meaning as the inverse image of the set $\{a\}$.

[2]For the benefit of readers who know some Logic, we note that this argument tacitly uses what's called the *Axiom of Choice*.

We summarize the relationships between inverses and injectivity/surjectivity/bijectivity as follows:

$$\begin{aligned} f : X \to Y \text{ is injective} &\iff f \text{ has a (necessarily surjective) left inverse } g : Y \to X \\ &\iff f \text{ is a right inverse for } g \\ &\iff f^{-1}(f(C)) = C \text{ for all } C \subset X \\ &\iff f\left(\bigcap_{A\in\mathcal{A}} A\right) = \bigcap_{A\in\mathcal{A}} f(A) \text{ for any nonempty collection } \mathcal{A} \text{ of subsets of } X \\ &\iff f(C\backslash A) = f(C)\backslash f(A) \text{ for any two subsets } A \text{ and } C \text{ of } X. \end{aligned}$$

$$\begin{aligned} f : A \to B \text{ is surjective} &\iff f \text{ has a (necessarily injective) right inverse } g : B \to A \\ &\iff f \text{ is a left inverse for } g \\ &\iff f(f^{-1}(D)) = D \text{ for all } D \subset B. \end{aligned}$$

$$\begin{aligned} f : A \to B \text{ is bijective} &\iff f \text{ has a (necessarily bijective) two-sided inverse } g : B \to A \\ &\iff f \text{ is a two-sided inverse for } g \\ &\iff f^{-1}(f(C)) = C \text{ and } f(f^{-1}(D)) = D \text{ for all } C \subset A \text{ and } D \subset B. \end{aligned}$$

3.4. J-tuples and Cartesian Products.

3.4.1. *J-tuples.*

DEFINITION 3.11. Let J and X be sets. The set of all functions from J to X is denoted X^J. A *J-tuple* of elements of X is a function $x : J \to X$, i.e. an element of X^J. However, when we refer to a function as a 'J-tuple,' we usually denote its values by x_j instead of $x(j)$, and we denote the function itself by $x = (x_j)_{j\in J}$. We use the term *index set* to refer to the domain J.

Technically, the difference between a function and a J-tuple is purely notational. However, in practice, the two are used differently, and certain kinds of sets J are used more frequently than others. Denote

$$J_n := \{1, 2, \ldots, n\}, \quad n \in \mathbb{N}. \tag{5}$$

(This notational convention will be used throughout the notes.) A J_n-tuple is often just called an n-tuple, and several possible notations are commonly used:

$$(x_j)_{j\in J_n} = (x_j)_{j=1}^n = (x_1, \ldots, x_n).$$

Furthermore, the set X^{J_n} of all n-tuples is often just denoted by X^n. For $n \geq 3$, we define the notation $X \times \cdots \times X$ to mean X^n. The set X^2 can be thought of as the same thing as $X \times X$ (but we have already defined the latter). Note that, for example, the set $X^3 = X \times X \times X$ means something slightly different from $(X \times X) \times X$. However, for most purposes, these sets can be treated as if they were the same[3].

[3]We give one caveat in the form of an example. Writing $(a, b) \in A^2 \times A$ implies that $a \in A^2$ and $b \in A$; this is *not* the same as writing $(a, b) \in A \times A^2$, which implies that $a \in A$ and $b \in A^2$. This demonstrates the difference between $(A \times A) \times A$ and $A \times (A \times A)$; the difference between these two sets and $A \times A \times A$ is similar in spirit.

3.4.2. *Sequences and Subsequences.*

DEFINITION 3.12. An $\mathbb{N}$-tuple of elements of X is called a *sequence* in X, and is denoted

$$(x_j)_{j\in\mathbb{N}} = (x_j)_{j=1}^{\infty} = (x_1, x_2, \ldots).$$

REMARK 3.13. The image of a J-tuple $(x_j)_{j\in J}$ is denoted by

$$\{x_j\}_{j\in J} = \{x_j : j \in J\}.$$

For the image of n-tuples and sequences, the following natural notation is also available:

$$\{x_j\}_{j\in J_n} = \{x_j : j \in J_n\} = \{x_j\}_{j=1}^{n} = \{x_1, x_2, \ldots, x_n\}.$$

$$\{x_j\}_{j\in\mathbb{N}} = \{x_j : j \in \mathbb{N}\} = \{x_j\}_{j=1}^{\infty} = \{x_1, x_2, \ldots\}$$

Do not confuse the two notations $(x_j)_{j=1}^{\infty}$ and $\{x_j\}_{j=1}^{\infty}$. The first is a sequence (which is a function); the second is the image of the sequence, which is a set. For example, if $x_i = 2$ for odd i and $x_i = 3$ for even i, then the sequence $(x_i)_{i=1}^{\infty}$ is $(2, 3, 2, 3, 2, 3, \ldots)$, while the set $\{x_i\}_{i=1}^{\infty}$ is just $\{2, 3\}$. Similar remarks hold for ordered n-tuples. Unfortunately, some authors (including myself, previously, and also including Rudin) do not make this distinction, which can lead to a lot of confusion.

DEFINITION 3.14. Let $x = (x_j)_{j=1}^{\infty}$ be a sequence, and let $k = (k_\ell)_{\ell=1}^{\infty}$ be a strictly increasing sequence in $\mathbb{N}$ (i.e., $k(1) < k(2) < k(3) < \cdots$). Then the sequence $y := x \circ k$ is called a *subsequence* of x, and we often write $y = (y_\ell)_{\ell=1}^{\infty} = (x_{k_\ell})_{\ell=1}^{\infty}$.

REMARK 3.15. The definition above is the commonly accepted definition of a subsequence, but it might seem rather opaque at first. In practice, given a sequence $x = (x_1, x_2, x_3, \ldots)$, we think of a subsequence as being formed by deleting some of the entries of x and preserving the original order, i.e. $y = (x_2, x_3, x_5, x_7, \ldots)$. The above definition is just a way of writing this process. However, writing down the following equivalent notations may help clarify the picture:

$$y_\ell = y(\ell) = (x \circ k)(\ell) = x(k(\ell)) = x_{k(\ell)} = x_{k_\ell}.$$

EXAMPLE 3.16. If the sequence $(x_j)_{j=1}^{\infty}$ is given by $(x_j)_{j=1}^{\infty} = (1, 2, 3, 1, 2, 3, 1, 2, 3, \ldots)$, then the subsequence $(y_\ell)_{\ell=1}^{\infty} = (x_{2\ell})_{\ell=1}^{\infty}$ is given by $(y_\ell)_{\ell=1}^{\infty} = (2, 1, 3, 2, 1, 3, \ldots)$.

3.4.3. *Indexed Sets.*

DEFINITION 3.17. If X is a set and X is the image of a J-tuple, i.e., X can be written $X = \{x_j\}_{j\in J}$ for some set J, then we say that X is *indexed by* J, and we refer to the notation $\{x_j\}_{j\in J}$ as an *indexed set*.

Note that the function implicit in this definition is necessarily surjective when considered as a function from $J \to X$. This function could also be considered as a J-tuple of elements in some space Y containing X; if this is the case we do *not* say that Y is indexed by J. The function $J \to X$ (or $J \to Y$) is not in general required to be injective.

The most useful indexed sets are actually indexed collections: $\mathcal{A} = \{A_j\}_{j\in J}$. If A_j is a set for each $j \in J$, we sometimes we refer to $\{A_j\}_{j\in J}$ as a *family of sets indexed by* J.

Indexed families of sets yield convenient notation for unions. If $\mathcal{A} = \{A_j\}_{j\in J}$, we write

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

For sets indexed by J_n (for some $n \in \mathbb{N}$) or by $\mathbb{N}$, we often use the following notation:

$$\bigcup_{j\in J_n} A_j = \bigcup_{j=1}^{n} A_j = A_1 \cup A_2 \cup \cdots \cup A_n,$$

$$\bigcup_{j\in\mathbb{N}} A_j = \bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup \cdots .$$

Entirely similar notation is used for intersections.

EXAMPLE 3.18. Any set can be indexed by itself (although this is rather silly). For example, if $A = \{3, 4, 7\}$, then we can define a function $x : A \to A$ by the identity map (denoted temporarily by x rather than by id_A, so $x(3) = x_3 = 3$, etc.). In this case we have $A = \{x_a\}_{a\in A} = \{a\}_{a\in A}$. The purpose of this example is to show that we don't lose anything if we consider unions, etc. only over indexed sets.

3.4.4. *Cartesian Products.*

DEFINITION 3.19. Let X be a set; let $\mathcal{A} = \{A_j\}_{j\in J}$ be a collection of subsets of X set indexed by J. The *cartesian product* of $\{A_j\}_{j\in J}$ is the set of all J-tuples $(x_j)_{j\in J}$ in X such that $x_j \in A_j$ for each $j \in J$, and is denoted

$$\prod_{j\in J} A_j.$$

The above general definition is useful to have in some contexts; however, we will most often consider it in the special case where $J = J_n$. In that context, the definition reads as follows:

DEFINITION 3.20. Let $\{A_j\}_{j=1}^n$ be a collection of subsets of a set X. The *cartesian product* of the sets $\{A_j\}_{j=1}^n$ is the set of all n-tuples $(x_1, \ldots, x_n)$ in X^n such that $x_j \in A_j$ for each $j \in J_n$.

We use the notation

$$\prod_{j\in J_n} A_j = \prod_{j=1}^{n} A_j = A_1 \times \cdots \times A_n.$$

Notice that this definition of $A_1 \times A_2$ agrees with the one we already have. We discussed the relationship between the sets $X^3 = X \times X \times X$ and $(X \times X) \times X$ above; similar comments hold for $A \times B \times C$ and $(A \times B) \times C$, for example.

4. Cardinality

Our next topic is that of cardinality, which gives us a way to compare the size of two sets. Making sense of this for finite sets is fairly straightforward (the cardinality of a finite set turns out to basically agree with the intuitive notion of the 'number of elements' it contains), but it is more subtle for infinite sets. In fact, cardinality is—strictly speaking—necessary for rigorously definining and distinguishing between 'finite' and 'infinite' in the first place (though we expect that the reader already has a pretty good idea of what these words should mean). We give a definition of cardinality that allows us to treat both finite and infinite sets satisfactorily, and to rigorously define the concepts just mentioned.

4.1. Definitions and Notation.

DEFINITION 4.1. Let A and B be nonempty sets. We use the notation

$$\mathrm{card}(A) \leq \mathrm{card}(B), \quad \mathrm{card}(A) \geq \mathrm{card}(B), \quad \mathrm{card}(A) = \mathrm{card}(B)$$

to mean that there exists a function $f : A \to B$ which is injective, surjective, or bijective, respectively. If there exists an injection $A \to B$ but not a bijection, then we write $\mathrm{card}(A) < \mathrm{card}(B)$; if there exists a surjection but not a bijection, we write $\mathrm{card}(A) > \mathrm{card}(B)$. The notation $\mathrm{card}(A)$ is called the *cardinality* of the set A.

The notation here is reminiscent of that of an ordering. This is mostly a manner of notational convenience. However, we do give some partial justification for the notation that is independently useful. The following Theorem implies, roughly speaking, that $\leq$ 'acts like' an order relation.

THEOREM 4.2. *Let A and B be sets.*

(a) If $A \subset B$, then $\mathrm{card}(A) \leq \mathrm{card}(B)$.
(b) If $\mathrm{card}(A) \leq \mathrm{card}(B)$ and $\mathrm{card}(B) \leq \mathrm{card}(C)$, then $\mathrm{card}(A) \leq \mathrm{card}(C)$.
(c) $\mathrm{card}(A) \leq \mathrm{card}(B)$ if and only if $\mathrm{card}(B) \geq \mathrm{card}(A)$.
(d) At least one of the statements $\mathrm{card}(A) \leq \mathrm{card}(B)$ or $\mathrm{card}(B) \leq \mathrm{card}(A)$ holds.
(e) (Schröder-Bernstein) If $\mathrm{card}(A) \leq \mathrm{card}(B)$ and $\mathrm{card}(B) \leq \mathrm{card}(A)$, then $\mathrm{card}(A) = \mathrm{card}(B)$.

(PARTIAL) PROOF. Statement (a) follows from the fact that the inclusion map $\iota : A \to B$ is an injection. Statement (b) follows from the fact that the composition of two injections is again an injection. Statement (c) follows from the fact that any injection $f : A \to B$ has a (necessarily surjective) left inverse $g : B \to A$, and any surjection $g : B \to A$ has a (necessarily injective) right inverse $f : A \to B$.

Statement (d) says that given any two sets A and B, either there exists an injection from A to B, or there exists an injection from B to A (or both). This statement may seem intuitive; however, for general sets A and B, the proof of (d) requires the use of some set-theoretic machinery that we have not developed[4]. We will not prove statement (d) in general; however, we will treat some special cases below.

We do have the tools to prove the Schröder-Bernstein Theorem, but its proof is not very enlightening, so we omit it also. □

We sometimes use the notation $A \sim B$ as shorthand for $\mathrm{card}(A) = \mathrm{card}(B)$. Note that $\sim$ satisfies the properties of an equivalence relation.

- (Reflexivity) $A \sim A$, for every set A.
- (Symmetry) If $A \sim B$, then $B \sim A$.
- (Transitivity) If $A \sim B$ and $B \sim C$, then $A \sim C$.

Therefore we sometimes say that A and B are *equivalent* if $A \sim B$.

4.2. Generating Equivalences. The following statement gives an extremely useful way to construct sets with the same cardinality from existing ones.

PROPOSITION 4.3. *Let A, B, X, and Y be sets. If $A \cap B = X \cap Y = \emptyset$ and $A \sim X$, $B \sim Y$, then $A \cup B \sim X \cup Y$.*

The proof of this Proposition follows directly from Lemma 4.4 below. A particularly useful case of this Proposition is the case where $A = X$.

LEMMA 4.4. *Assume $A \cap B = X \cap Y = \emptyset$. Let $f : A \to X$ and $g : B \to Y$ be bijections, and define a function $h : (A \cup B) \to (X \cup Y)$ by the following rule:*

$$h(c) = \begin{cases} f(c) & \text{if } c \in A \\ g(c) & \text{if } c \in B. \end{cases}$$

Then h is a bijection.

PROOF. Note that since f and g are bijections, we can define their inverses f^{-1} and g^{-1}, respectively. Define a function $k : X \cup Y \to A \cup B$ by

$$k(z) = \begin{cases} f^{-1}(z) & \text{if } z \in X \\ g^{-1}(z) & \text{if } z \in Y. \end{cases}$$

One can easily check[5] that k is a two-sided inverse for h. □

[4]Namely, it requires Zorn's Lemma. It turns out that Zorn's Lemma is equivalent to the Axiom of Choice. However, unlike the Axiom of Choice, invoking Zorn's Lemma is not something that will go unnoticed.

[5]To see that k is a right inverse, note the following:

- If $z \in X$, then $k(z) = f^{-1}(z) \in A$, so $h(k(z)) = h(f^{-1}(z)) = f(f^{-1}(z)) = z$.
- If $z \in Y$, then $k(z) = g^{-1}(z) \in B$, so $h(k(z)) = h(g^{-1}(z)) = g(g^{-1}(z)) = z$.

The proof that k is a left inverse is entirely similar.

A word of caution: In the context of the previous Lemma, if X and Y are not disjoint, then the function h may not be bijective. If A and B are not disjoint, then the object h as defined above might not even be a function!

Recall that the notation B^A means the set of all functions from A to B. The following Exercise gives a useful way to generate equivalences.

EXERCISE 4.1. Assume that $\mathrm{card}(A) \leq \mathrm{card}(X)$ and $\mathrm{card}(B) \leq \mathrm{card}(Y)$. Prove that $\mathrm{card}(B^A) \leq \mathrm{card}(Y^X)$. Hint: Consider a function $\Phi : B^A \to Y^X$ of the form $\Phi(f) = h \circ f \circ k$, where $k : X \to A$ and $h : B \to Y$ are certain functions. Theorem 3.8 might be useful for your final step.

Recall that for any set A, the notation $\mathcal{P}(A)$ is used to denote the *power set* of A, i.e., the collection of all subsets of A. Here is a useful *non*-equivalence:

PROPOSITION 4.5. *For any set A,* $\mathrm{card}(A) < \mathrm{card}(\mathcal{P}(A))$.

PROOF. Let $g : A \to \mathcal{P}(A)$ be any map; we show it is not surjective. Define $B := \{x \in A : x \notin g(x)\}$. We claim that $B \notin \mathrm{Im}\, g$, which will show that g is not surjective. To prove our claim, we argue by contradiction: Suppose $g(x) = B$ for some $x \in A$. Then x cannot be in B by definition of B. On the other hand, if $x \notin B = g(x)$, then $x \in B$ by definition of B. This contradiction establishes the claim. □

Thinking about cardinality as a notion of (relative) 'size', the Proposition above shows that there is no limit to how 'big' sets can get; given any set A we can always construct a 'bigger' one by taking the power set of A.

EXERCISE 4.2. Prove that for any set A, one has $\mathcal{P}(A) \sim \{0,1\}^A$.

4.3. Finite and Infinite Sets. For each $n \in \mathbb{N}$, let J_n denote the set

$$J_n := \{1, \ldots, n\}.$$

It can be proved rigorously (by a rather tedious induction argument[6]) that for $m, n \in \mathbb{N}$, one has

$$\mathrm{card}(J_m) \leq \mathrm{card}(J_n) \iff m \leq n.$$

Therefore it makes sense to define $\mathrm{card}(J_n) = n$ for all $n \in \mathbb{N}$. We also define $\mathrm{card}(\emptyset) = 0$.

DEFINITION 4.6. A set A is called *finite* if $\mathrm{card}(A) = \mathrm{card}(J_n)$ for some $n \in \mathbb{N}$, or if $A = \emptyset$. A set is called *infinite* if it is not finite.

In light of these conventions, it makes sense to think of the cardinality of a finite set A as the number of elements it contains. However, one should keep in mind that the statement $\mathrm{card}(A) = n$ really means that there exists a bijection $f : J_n \to A$. Furthermore, if A is an infinite set, then the concept of 'number of elements in A' becomes rather vague.

For finite sets, we sometimes use the notation $|A| = \mathrm{card}(A)$.

[6] If $m \leq n$, then the inclusion map $\iota : J_m \to J_n$ is an injection, so $\mathrm{card}(J_m) \leq \mathrm{card}(J_n)$. We prove other direction, the statement "$\mathrm{card}(J_m) \leq \mathrm{card}(J_n)$ implies $m \leq n$", by induction on n. The only function $f : J_m \to J_1$ is defined by the rule $f(j) = 1$ for all $j \in J_m$. If f is injective, then the fact that $f(1) = f(m)$ implies that $m = 1$. Next, assume the statement is true for $n = k \in \mathbb{N}$; we prove it for $n = k+1$. So, we start with an injection $f : J_m \to J_{k+1}$ and try to prove that $m \leq k+1$. If $m = 1$, we are already done, so we assume that $m \geq 2$ without loss of generality. It actually suffices to show that there exists a bijection $g : J_{m-1} \to J_k$. Indeed, assume such a g exists. Then by our inductive hypothesis, we have $m - 1 \leq k$, i.e. $m \leq k + 1$, which is what we are trying to prove. Therefore, let us construct an injection $g : J_{m-1} \to J_k$. We consider two cases. First, if $f(m) = k$, then we can simply define g by $g(j) = f(j)$ for all $j \in J_{m-1}$ (i.e., by restricting both the domain and codomain of f). If $f(m) \neq k$, then define two functions $g : J_{m-1} \to J_k \backslash \{f(m)\}$ and $h : J_{k+1} \backslash \{f(m)\} \to J_k$ as follows. Define $g(j) = f(j)$ for each j; define $h(j) = j$ if $j \neq k$ and $h(k) = f(m)$. Then g is an injection and h is a bijection, so $h \circ g : J_{m-1} \to J_k$ is an injection.

PROPOSITION 4.7. *If A and B are finite, disjoint sets, then $A \cup B$ is finite, and $|A \cup B| = |A| + |B|$.*

PROOF. Since A and B are finite, there exist integers m and n such that

$$A \sim J_m,$$
$$B \sim J_n \sim \{m+1, \dots, m+n\}.$$

Since both $A \cap B$ and $J_m \cap \{m+1, \dots, m+n\}$ are empty, Lemma 4.4 tells us that

$$A \cup B \sim J_m \cup \{m+1, \dots, m+n\} = J_{n+m},$$

i.e. $|A \cup B| = m + n = |A| + |B|$. □

COROLLARY 4.8. *If A and B are any finite sets (not necessarily disjoint), then $A \cup B$ is finite.*

PROOF. $A \cup B$ can also be written as the union of the two finite, disjoint sets $A \backslash B$ and B. □

PROPOSITION 4.9. *Let A and B be finite sets. Then $|A^B| = |A|^{|B|}$, and $|\mathcal{P}(A)| = 2^{|A|}$.*

PROOF. Since A and B are finite, there exist $m, n \in \mathbb{N}$ such that $A \sim J_m$ and $B \sim J_n$. By Exercise 4.1, it follows that $A^B \sim \{1, \dots, m\}^n$, i.e. the set of n-tuples whose entries are the numbers $1, \dots m$. Since there are m choices for each of the n entries, there are $m^n = |A|^{|B|}$ elements in this set: $|A^B| = |A|^{|B|}$. This proves the first statement.

The second statement follows from the first statement, together with Exercise 4.2: Since $\mathcal{P}(A) \sim \{0,1\}^A$, we have $|\mathcal{P}(A)| = |\{0,1\}^A| = |\{0,1\}|^{|A|} = 2^{|A|}$. □

REMARK 4.10. Note that the set $\mathbb{N}$ is *not* finite. Hopefully this doesn't come as a surprise, but a few words about the proof are in order: Suppose (to obtain a contradiction) that $\mathbb{N}$ is finite. Then there exists $n \in \mathbb{N}$ and a surjection $f : J_n \to \mathbb{N}$. On the other hand, it is easy to construct a surjection $g : \mathbb{N} \to J_{n+1}$. (For example, take $g(j) = j$ for $j \in J_{n+1}$, and $g(j) = 1$ for $j > n+1$.) But then $g \circ f$ is a surjection from J_n to J_{n+1}, which is impossible. Therefore $\mathbb{N}$ cannot be finite. However, $\mathbb{N}$ is, in a sense, the 'smallest' infinite set. The following discussion (especially Corollary 4.14 below) makes this statement precise.

DEFINITION 4.11. Let A be a set. We say that A is *countable* if $\mathrm{card}(A) \leq \mathrm{card}(\mathbb{N})$. We say A is *countably infinite* if $\mathrm{card}(A) = \mathrm{card}(\mathbb{N})$. We say A is *uncountable* if it is not countable.

REMARK 4.12. A completely equivalent definition of 'countable' is the following: A set A is countable if it can be indexed by $\mathbb{N}$, i.e. there exists a sequence $(a_j)_{j=1}^{\infty}$ whose image $\{a_j\}_{j=1}^{\infty}$ is equal to A. In this case the sequence $(a_j)_{j=1}^{\infty}$ is called an *enumeration* of the elements of A, since all the elements of A appear in the 'list' $(a_1, a_2, a_3, \dots)$.

PROPOSITION 4.13. *Let A be an infinite set. Then A contains a countably infinite subset.*

PROOF. The basic idea of the proof is to remove elements, one at a time, from A. As we remove them, we are counting them, or *enumerating* them, to construct an injection $f : \mathbb{N} \to A$. The image of f will be our countably infinite subset of A. A more careful version of the above idea is as follows:

Since A is infinite, A is not empty; therefore we can find an element of $A_1 := A$, which we denote a_1. The set $A_2 := A_1 \backslash \{a_1\}$ is still infinite. (If A_2 is finite, then so is $A_1 = A_2 \cup \{a_1\}$, since $A_2 = A_1 \backslash \{a_1\}$ and $\{a_1\}$ are finite disjoint sets.) Therefore we can find an element $a_2 \in A_2$, and $A_3 := A_2 \backslash \{a_2\} = A_1 \backslash \{a_1, a_2\}$ is still infinite. We can continue this process indefinitely, obtaining sets A_k and elements a_k of A_k such that $A_{k+1} = A_k \backslash \{a_k\}$, for each $k \in \mathbb{N}$. Consequently, for $m < n$, we have

$$A_n = A_{n-1} \backslash \{a_{n-1}\} = A_{n-2} \backslash \{a_{n-2}, a_{n-1}\} = \cdots = A_m \backslash \{a_m, \dots, a_{n-1}\},$$

so that $a_m \in A_m \backslash A_n$ whenever $m < n$. Since $a_n \in A_n$, it follows that $a_m \neq a_n$ for $m < n$. Therefore, the function $f : \mathbb{N} \to A$ defined by $f(n) = a_n$ is injective. Let B denote the image of f. Then B is a countably infinite subset of A. □

COROLLARY 4.14. *A set A is infinite if and only if* $\mathrm{card}(A) \geq \mathrm{card}(\mathbb{N})$.

PROOF. If A is infinite, then A contains a countably infinite subset B. Therefore $\mathrm{card}(A) \geq \mathrm{card}(B) = \mathrm{card}(\mathbb{N})$. On the other hand, if $\mathrm{card}(A) \geq \mathrm{card}(\mathbb{N})$, then $\mathrm{card}(A) > \mathrm{card}(J_m)$ for every $m \in \mathbb{N}$, so A is not finite. □

Another useful characterization of infinite sets is the following:

THEOREM 4.15. *A set A is infinite if and only if there exists a proper subset $B \subsetneq A$ such that $A \sim B$.*

PROOF. Suppose A is infinite; choose $a \in A$ and define the proper subset $B = A\backslash\{a\}$ of A. We prove that $A \sim B$. Now, B is infinite, so it contains a countably infinite subset C, and $a \notin C$. The set $C \cup \{a\}$ is still countably infinite. (Indeed, if $f : \mathbb{N} \to C$ is a bijection, define a bijection $g : \mathbb{N} \to C \cup \{a\}$ by $g(1) = a$, $g(n) = f(n-1)$ for $n > 1$.) Since $(B\backslash C) \cap C$ and $(B\backslash C) \cap (C \cup \{a\})$ are both empty, we conclude that

$$B = (B\backslash C) \cup C \sim (B\backslash C) \cup (C \cup \{a\}) = B \cup \{a\} = A.$$

The equivalence $\sim$ follows from Proposition 4.3.

Suppose A is finite. If A contains no proper subsets, then we are done; otherwise let B be any proper subset of A. Then $A\backslash B$ is nonempty, so $|A\backslash B| \geq 1$. Therefore $|A| = |A\backslash B| \cup |B| \geq 1 + |B| > |B|$. In particular, $|A| \neq |B|$, so the statement $A \sim B$ cannot hold. □

We summarize the last few results:

$$\begin{aligned} A \text{ is infinite} &\iff \mathrm{card}(A) \geq \mathrm{card}(\mathbb{N}) \\ &\iff A \text{ contains a countably infinite subset} \\ &\iff A \sim B \text{ for some proper subset } B \text{ of } A. \end{aligned}$$

EXERCISE 4.3. Let A and B be sets, and assume $f : A \to B$ and $g : B \to A$ are injective functions.

(a) Assume additionally that A is finite. Prove that f and g must actually be bijections.
(b) Show by way of an example that both f and g may fail to be bijective if we do not assume that A is finite.

EXERCISE 4.4. Let A and B be sets. Assume A is infinite, B is countable, and A and B are disjoint. Prove that $A \sim A \cup B$. Hint: The strategy of Theorem 4.15 may be useful. You may also use the fact that the union of two countable sets is countable. (A more general statement on unions of countable sets is proved in the next subsection.)

EXERCISE 4.5. Let X and Y be sets. Assume Y is countable and $X\backslash Y$ is infinite. Prove that $X \sim X \cup Y \sim X\backslash Y$. Hint: Each of the equivalences can be done extremely quickly if you use the previous exercise and some set manipulations.

4.4. Countable Sets.

PROPOSITION 4.16. $\mathbb{N} \times \mathbb{N}$ *is countably infinite.*

PROOF. The function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ defined by $f((n,m)) = 2^n \cdot 3^m$ is injective. Indeed, if $2^{n_1}3^{m_1} = 2^{n_2}3^{m_2}$ for some $n_1, n_2, m_1, m_2 \in \mathbb{N}$, then we must have $n_1 = n_2$ and $m_1 = m_2$, by uniqueness of prime decompositions. Therefore $(n_1, m_1) = (n_2, m_2)$, i.e., f is injective. It follows that $\mathrm{card}(\mathbb{N} \times \mathbb{N}) \leq \mathrm{card}(\mathbb{N})$. On the other hand, the function $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, $g((n,m)) = n$ is a surjection, so $\mathrm{card}(\mathbb{N} \times \mathbb{N}) \geq \mathrm{card}(\mathbb{N})$. Thus $\mathrm{card}(\mathbb{N} \times \mathbb{N}) = \mathrm{card}(\mathbb{N})$, as needed. □

If $\mathcal{A}$ is a countable collection of sets, we say that $\bigcup_{A \in \mathcal{A}} A$ is a *countable union*. If $\mathcal{A}$ is finite, we call it a *finite union*. (We use similar terminology for cartesian products.) The next Proposition can be stated briefly as "A countable union of countable sets is countable."

PROPOSITION 4.17. *Let $\{A_j\}_{j=1}^{\infty}$ be a countable collection of countable sets. Then the union*

$$S = \bigcup_{j=1}^{\infty} A_j$$

is also countable.

PROOF. We define a function $f : \mathbb{N} \times \mathbb{N} \to S$ as follows. For each $j \in \mathbb{N}$, let $g_j : \mathbb{N} \to A_j$ be a surjection. Define $f((j,k)) = g_j(k)$ for each $(j,k) \in \mathbb{N} \times \mathbb{N}$. Then f is surjective. Indeed, if $x \in S$, then $x \in A_j$ for some $j \in \mathbb{N}$. But then since g_j is surjective, we have $x = g_j(k) = f(j,k)$ for some $k \in \mathbb{N}$. It follows that $\mathrm{card}(\mathbb{N}) = \mathrm{card}(\mathbb{N} \times \mathbb{N}) \geq \mathrm{card}(S)$. That is, S is countable. □

Note that this Proposition implies rather trivially that $\mathbb{Z}$ is countable. (Write $\mathbb{Z} = \mathbb{N} \cup (-\mathbb{N}) \cup \{0\}$, where $-\mathbb{N}$ denotes the set $\{-n\}_{n=1}^{\infty}$. Consequently, $\mathbb{Q}$ is countable, since the function $f : \mathbb{Z} \times (\mathbb{Z} \backslash \{0\}) \to \mathbb{Q}$ defined by $f((n,m)) = \frac{n}{m}$ is a surjection.

EXERCISE 4.6. A tempting—but incorrect—variation on the argument of the proof of Proposition 4.17 is the following. Point out the error in the argument.

> INCORRECT argument: We construct an injection $f : S \to \mathbb{N} \times \mathbb{N}$ as follows: For each j, let $g_j : A_j \to \mathbb{N}$ be an injection. Then define $f : S \to \mathbb{N} \times \mathbb{N}$ according to the following rule: If $a \in A_j$, define $f(a) = (j, g_j(a))$. Then f is injective, so S is countable.

EXERCISE 4.7. Let X be a countable set.

(a) Prove that $X^{n+1} \sim X^n \times X$ for any $n \in \mathbb{N}$. (This is not difficult, but be careful with the notation.)
(b) Prove inductively that X^n is countable for any $n \in \mathbb{N}$.

Despite the result of this Exercise, we are able rather easily to prove that an infinite product of copies of a set X is rarely countable, even if X is finite. We make this precise in the next subsection.

4.5. Uncountable Sets.

PROPOSITION 4.18. *If X is any set that contains at least two elements and J is countably infinite, then X^J is uncountable. In particular, $\{0,1\}^{\mathbb{N}}$ is uncountable.*

PROOF. Since $\mathrm{card}(X) \geq \mathrm{card}(\{0,1\})$ and $\mathrm{card}(J) \geq \mathrm{card}(\mathbb{N})$, we have

$$\begin{aligned} \mathrm{card}(X^J) &\geq \mathrm{card}(\{0,1\}^{\mathbb{N}}) && \text{by Exercise 4.1} \\ &= \mathrm{card}(\mathcal{P}(\mathbb{N})) && \text{by Exercise 4.2} \\ &> \mathrm{card}(\mathbb{N}) && \text{by Proposition 4.5.} \end{aligned}$$

□

This high-brow argument is nice in that it is concise, but doesn't give too much of an idea of what is going on here. We give a more down-to-earth proof of the following special case.

PROPOSITION 4.19. *$\{0,1\}^{\mathbb{N}}$ is uncountable.*

Remember that X^J denotes the set of functions from J to X. So $\{0,1\}^{\mathbb{N}}$ is the set of sequences in $\{0,1\}$. A typical element of $\{0,1\}^{\mathbb{N}}$ might look like $(0,1,0,0,0,1,0,1,1,\ldots)$, for example. The Proposition says that $\mathrm{card}(\{0,1\}^{\mathbb{N}}) > \mathrm{card}(\mathbb{N})$, i.e., there *does not* exist a surjection from $\mathbb{N}$ to $\{0,1\}^{\mathbb{N}}$. As with many nonexistence claims, this statement is proven by contradiction.

PROOF. Suppose $A := \{0,1\}^{\mathbb{N}}$ is countable, and let $(x_1, x_2, \ldots)$ be an enumeration of the elements of A. Remember that each x_j is a sequence; let a_{jk} denote the kth coordinate of the sequence x_j. Let the sequence $y = (y_\ell)$ be defined by the rule $y_\ell \neq a_{\ell\ell}$, for each $\ell \in \mathbb{N}$. (Since there are only two choices, 0 and 1, for $a_{\ell\ell}$, this definition is not ambiguous.) Then $y \neq x_j$ for any $j \in \mathbb{N}$ (since $y(j) \neq x_j(j) = a_{jj}$). Therefore, y does not appear in the supposed enumeration $(x_1, x_2, \ldots)$, contradicting our assumption. □

REMARK 4.20. The idea of this proof is illustrated in the following diagram:

$$\begin{array}{ccccc} \mathbf{a_{11}} & a_{12} & a_{13} & a_{14} & \cdots \\ a_{21} & \mathbf{a_{22}} & a_{23} & a_{24} & \cdots \\ a_{31} & a_{32} & \mathbf{a_{33}} & a_{34} & \cdots \\ a_{41} & a_{42} & a_{43} & \mathbf{a_{44}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

We construct y by flipping all the entries on the diagonal. For example, if $a_{11} = 1$, $a_{22} = 0$, $a_{33} = 0$, then we choose $y_1 = 0$, $y_2 = 1$, $y_3 = 1$, and so on. This argument is thus called *Cantor's diagonal argument*, after Georg Cantor, who introduced it in 1891.

CHAPTER 2

The Real and Complex Number Systems

1. Ordered Sets and the Least Upper Bound Property

We begin by recalling the definition of an order, and introducing the concept of an ordered set.

DEFINITION 1.1. Let S be a set. An *order* (or *total order*) on S is a relation $\leq$, such that

- (Reflexivity) $a \leq a$, for all $a \in A$.
- (Antisymmetry) If $a \leq b$ and $b \leq a$, then $a = b$.
- (Transitivity) If $a \leq b$ and $b \leq c$, then $a \leq c$.
- (Comparability) If $a, b \in A$, then at least one of the statements $a \leq b$ or $b \leq a$ must hold.

An *ordered set* is a set S together with an order $\leq$, denoted $(S, \leq)$.

Some comments on notation: First, for most of the ordered sets we consider, the specific order $\leq$ is obvious from context. In this case we refer to 'the ordered set S', even when we actually mean $(S, \leq)$. Other structures on sets are treated similarly. Second, we remind the reader that $x < y$ will mean $x \leq y$ and $x \neq y$.

Note that if $(S, \leq)$ is an ordered set and $T \subset S$, then $(T, \leq)$ is also an ordered set.

DEFINITION 1.2. Let E be a subset of an ordered set S, and let α and β be elements of S.

- If $x \leq \beta$ for all $x \in E$, then β is called an *upper bound* for E, and we say that E is *bounded above* if such a β exists.
- If β is an upper bound for E in S and $\beta \in E$, then β is called the *maximum* of E. (This is written $\beta = \max E$.)
- Suppose α is an upper bound for E in S and that $\alpha \leq \gamma$ for any upper bound γ of E in S. Then α is called the *least upper bound* of E in S, or the *supremum* of E in S. For short, we write $\alpha = \sup E$ if this holds. (Usually the set S is clear from context; if not, we can simply write the relevant statement in sentence form.)

The terms *lower bound*, *bounded below*, *minimum*, *greatest lower bound*, and *infimum* are defined analogously. If α is the infimum of a subset E of S, we write $\alpha = \inf E$. If α is the minimum of E, we write $\alpha = \min E$. If E is bounded above and bounded below, we sometimes simply say that E is *bounded* (in S).

We make the following notes about the above definitions.

- Whenever a maximum α of E exists, it is equal to the supremum of E. (More precisely, let E be a subset of an ordered set S. Assume that E has a maximum α. Then $\alpha = \sup E$.) Indeed, if $\gamma < \alpha$, then γ cannot be an upper bound for E, since α is an element of E which is greater than γ. Similarly, whenever a minimum exists, it is equal to the infimum. Note that if E is a finite subset of S, then the maximum and minimum of E automatically exist.
- One way we often prove that an upper bound α for E in S is in fact the *least* upper bound is by proving the contrapositive of the requirement in the definition: If $\gamma < \alpha$, then γ is not an upper bound for E in S. (C.f. both the previous bullet point and the next Example.)

- If X is a set, A is a subset of X, $(S, \leq)$ is an ordered set, and $f : X \to S$ is a function, then the following notations are all considered equivalent:

$$\sup f(A) = \sup\{f(x) : x \in A\} = \sup_{x \in A} f(x) = \sup_{A} f.$$

  Similar conventions are used for the infimum.

REMARK 1.3. Here is a trivial but useful property of the natural numbers: The subset $\mathbb{N}$ of $\mathbb{Q}$ is not bounded above. Indeed, given $q \in \mathbb{Q}$, we can show that there exists $n \in \mathbb{N}$ such that $n > q$. If $q \leq 0$, just take $n = 1$. Otherwise, write $q = \frac{a}{b}$, where $a, b \in \mathbb{N}$. Then $a \in \mathbb{N}$, and $a = \frac{a}{1} > \frac{a}{b} = q$. Thus $n = a$ is a number greater than q.

Below, we will frequently use this fact in the following form: "For any $q \in \mathbb{Q}$, there exists $n \in \mathbb{N}$ such that $n > q$."

EXAMPLE 1.4. Let E denote the set of all rational numbers of the form $2 - \frac{1}{n}$, where $n \in \mathbb{N}$. The supremum of E in $\mathbb{Q}$ is 2. Indeed, $2 - \frac{1}{n} < 2$ for all n, so 2 is an upper bound for E. On the other hand, if q is a rational number less than 2, let n be any natural number greater than $(2 - q)^{-1}$ (which exists, by the remark above). Then $\frac{1}{n} < 2 - q$, so $2 - \frac{1}{n} > q$, which implies that q is not an upper bound for E. Since $2 \notin E$, it follows that E has no maximum. On the other hand, 1 is the minimum of E, therefore also the infimum of E in $\mathbb{Q}$.

This example shows that the least upper bound and greatest lower bound, when they exist, may or may not be elements of the set E. Also note that in this example, the supremum of E exists in $\mathbb{Q}$. There are, however, bounded subsets of $\mathbb{Q}$ for which no least upper bound exists in $\mathbb{Q}$. (Consider for example $\{r \in \mathbb{Q} : r^2 < 2\}$. We will prove later that this set has no least upper bound in $\mathbb{Q}$.) That is, $\mathbb{Q}$ does not have the *least-upper-bound property*, defined below. Actually, this turns out to be the key difference between $\mathbb{Q}$ and $\mathbb{R}$. We will return to this point later.

DEFINITION 1.5. An ordered set S is said to have the *least-upper-bound property* (LUBP) if the following statement holds: "Whenever E is a nonempty subset of S that is bounded above, it follows that E has a least upper bound in S." The *greatest-lower-bound property* (GLBP) is defined similarly.

The greatest-lower-bound property won't make much more of an appearance in these notes, since it turns out to be equivalent to the least upper bound property:

THEOREM 1.6. *Let S be an ordered set. Then S has the least-upper-bound property if and only if it has the greatest-lower-bound property.*

PROOF. We will prove only one direction here, namely that the LUBP implies the GLBP. The other direction is quite similar.

Assume then that S has the LUBP. We need to show that if $B \subset S$ is bounded below, then it has a greatest lower bound. Let L denote the set of lower bounds for B. Since B is bounded below, we know L is nonempty. Furthermore, L is bounded above, by any element of B[1]. Since S has the least-upper-bound property, we may conclude that L has a least upper bound α in S. We claim that α is also the greatest lower bound for B.

First, we recall that every $x \in B$ is an upper bound for L. Then since α is the *least* upper bound, we conclude that $\alpha \leq x$ for all $x \in B$. This says exactly that α is a lower bound for B.

Next, if γ is any lower bound for B, then $\gamma \in L$, so that $\gamma \leq \alpha$, as α is an upper bound for L. Therefore α is the *greatest* lower bound for B. This completes the proof. □

EXERCISE 1.1. Let E, F, and G be nonempty subsets of an ordered set $(S, \leq)$. Prove the following statements.

[1]After all, L consists of lower bounds for B. So if $\gamma \in L$, then $\gamma \leq x$ for all $x \in B$, by definition of lower bound. On the other hand, if we fix $x \in B$, then $\gamma \leq x$ for *every* $\gamma \in L$. So x is an upper bound for L.

(a) If $\alpha \in S$ is a lower bound for E and $\beta \in S$ is an upper bound for E, then $\alpha \leq \beta$.
(b) $\sup E \leq \inf F$ if and only if $x \leq y$ for any $x \in E$, $y \in F$.
(c) If $E \subset G$, then $\sup E \leq \sup G$.

EXERCISE 1.2. Let $(S, \leq)$ be an ordered set, let f and g be functions from X to S and let A be a subset of X. Assume that $f(x) \leq g(x)$ for all $x \in A$, and that furthermore $\sup_A f$ and $\sup_A g$ exist in S. Prove that $\sup_A f \leq \sup_A g$.

2. Fields and Ordered Fields

The concept of a field should be familiar from linear algebra courses. Here we'll just review the definition.

DEFINITION 2.1. A *field* is a set F which has two operations, called *addition* (denoted by $+$) and *multiplication* (denoted either by $\cdot$ or simply by juxtaposition), such that the following *field axioms* hold:

(1) F is closed under addition and multiplication. That is, if x and y are elements of F, then so are $x + y$ and xy.
(2) Addition and multiplication are commutative and associative.
(3) Addition and multiplication each have identity elements (usually denoted 0 and 1, respectively) which are distinct.
(4) Each element of F has both an additive and a multiplicative inverse (with the exception of 0, which does not have a multiplicative inverse). That is, if $x \in F$, there exists $y \in F$ (the additive inverse) such that $x + y = 0$; if additionally $x \neq 0$, then there exists $z \in F$ (the multiplicative inverse) such that $xz = 1$. Usually the additive inverse of x is denoted $-x$ and the multiplicative inverse of $x \neq 0$ is denoted x^{-1} or $1/x$.
(5) The distributive law holds: $x(y + z) = xy + xz$.

Note that we use the notation for addition and multiplication that is usually associated to addition and multiplication of real numbers. Keep in mind that in principle these symbols could take other meanings. Furthermore, if more than one field is under consideration, alternate notation may be used.

DEFINITION 2.2. An *ordered field* is a field F which is also an ordered set, whose order relation $\leq$ satisfies the following:

(1) If $x, y, z \in F$ and $y \leq z$, then $x + y \leq x + z$.
(2) If $x, y \in F$ and $x > 0$, $y > 0$, then $xy > 0$.

If $x > 0$ we say that x is *positive*; if $x < 0$ we say that x is *negative*. (If $x \geq 0$, we say x is *nonnegative*; if $x \leq 0$, we say x is *nonpositive*.)

Examples: Recall that $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are all ordered sets, under the usual ordering. But even though the usual addition and multiplication operations can be defined on $\mathbb{N}$ and $\mathbb{Z}$, neither of these is a field, since, for example, the requirement of multiplicative inverses fails for each. On the other hand, $\mathbb{Q}$ is a field under the usual addition and multiplication operations. In fact, it is an ordered field, as one can check.

Not all fields are ordered fields. For example, the usual addition and multiplication operations on the set $\mathbb{C}$ of complex numbers (introduced later) make $(\mathbb{C}, +, \cdot)$ into a field. However, it is provably impossible to define an order $\leq$ on $(\mathbb{C}, +, \cdot)$ such that $(\mathbb{C}, +, \cdot, \leq)$ is an ordered field.

EXERCISE 2.1. Let A be a nonempty subset of an ordered field $(F, +, \cdot, \leq)$. Assume that $\sup A$ and $\inf A$ exist in F, and let c be any element of F. Define the set $cA := \{ca : a \in A\}$.

(a) Prove that if $c \geq 0$, then $\sup(cA) = c \sup A$.
(b) What is $\sup(cA)$ if $c < 0$? Prove that your answer is correct.

EXERCISE 2.2. Let A and B be nonempty subsets of an ordered field $(F, +, \cdot, \leq)$. Assume $\sup A$ and $\sup B$ exist in F. Define $A+B := \{a+b : a \in A, b \in B\}$. Prove that $\sup(A+B) = \sup A + \sup B$ by filling in the details of the following outline:

- Denote $s = \sup A$, $t = \sup B$. Then $s + t$ is an upper bound for $A + B$.
- Let u be any upper bound for $A + B$, and let a be any element of A. Then $t \leq u - a$.
- We have $s+t \leq u$. Consequently, $\sup(A+B)$ exists in F and is equal to $s+t = \sup A+\sup B$.

EXERCISE 2.3. Let f and g be functions from a set X to an ordered field $(F, +, \cdot, \leq)$. Let A be a subset of X.

(a) Prove that the following inequality holds, assuming the relevant suprema all exist.

$$(*) \qquad \sup_{x\in A}(f(x) + g(x)) \leq \sup_{x\in A} f(x) + \sup_{x\in A} g(x).$$

(b) Show by way of an example that equality might not hold in $(*)$, even if the suprema all exist. (Hint: This is probably easiest if you choose X to be a set with two elements, and $F = \mathbb{Q}$.)

DEFINITION 2.3. We say that an ordered field F has the *Archimedean property* if for every $x, y \in F$, with $x > 0$, there is a positive integer n such that $nx > y$.

Note: In this definition, we think of $x > 0$ as being small and y being (possibly) large.

PROPOSITION 2.4. *$\mathbb{Q}$ has the Archimedean property.*

PROOF #1. Let p, q be rational numbers with $p > 0$. We need to show that $np > q$ for some $n \in \mathbb{N}$. If $q \leq 0$ there is nothing to show, so we assume without loss of generality that $q > 0$. Write $p = a/b$ and $q = c/d$, where $a, b, c, d \in \mathbb{N}$. We seek $n \in \mathbb{N}$ such that $n(a/b) > c/d$, or clearing denominators, $nad > bc$. Since $nad \geq n$ (as $a, d \in \mathbb{N}$ and hence are each at least 1), it suffices to choose $n > bc$, say $n = bc + 1$. This is an integer because b and c are integers. Let's check that this choice works. We have

$$np = (bc + 1)p = (pb)c + p = ac + p = a \cdot \frac{c}{d} \cdot d + p = adq + p \geq q + p > q.$$

Thus $\mathbb{Q}$ has the Archimedean property, as claimed. □

Here is a slightly shorter proof that implicitly incorporates Remark 1.3.

PROOF #2. Let p and q be rational numbers with $p > 0$ (and without loss of generality, $q > 0$ as well). Write $p = \frac{a}{b}$ and $q = \frac{c}{d}$, with $a, b, c, d \in \mathbb{N}$. We want to find $n \in \mathbb{N}$ such that $np > q$, or in other words, $n(\frac{a}{b}) > \frac{c}{d}$. The latter inequality is equivalent to $n > \frac{ac}{bd}$. Since $\frac{ac}{bd}$ is a rational number, we can find $n \in \mathbb{N}$ such that $n > \frac{ac}{bd}$, and then reversing the steps above yields the desired inequality $np = n(\frac{a}{b}) > \frac{c}{d} = q$. □

3. The Problem with $\mathbb{Q}$

Working in $\mathbb{Q}$ rather than $\mathbb{R}$ has its shortcomings. Most likely the reader is already convinced of this fact; however, it might be less clear what the most fundamental shortcomings actually are. First of all, square roots are not guaranteed to exist:

PROPOSITION 3.1. *There is no rational number p such that $p^2 = 2$.*

PROOF. Assume that p is a rational number such that $p^2 = 2$. Since p is rational, we can write $p = \frac{m}{n}$, where m and n are integers with no common factors. The equation $p^2 = 2$ can thus be rewritten as $m^2 = 2n^2$. This implies that m^2 is even, which implies that m is even, which in turn implies that m^2 is actually divisible by 4. But then $m^2/2 = n^2$ is even, so n is even. But evenness of both m and n is incompatible with our initial assumption that m and n have no common factors! We conclude that such a rational number p cannot exist. □

The above proof can be modified to prove the irrationality of many square roots. However, the lack of square roots is not the end of the problem. It turns out that the fundamental shortcoming of $\mathbb{Q}$ is the fact that it does not have the least upper bound property. We prove this fact by giving an explicit example of a set nonempty subset A of $\mathbb{Q}$ whose supremum does not exist in $\mathbb{Q}$. The set we use is the set of rational numbers between 0 and $\sqrt{2}$. However, we have to be a bit careful about how we state and prove our claim, since $\sqrt{2}$ is an object for which we don't yet have a rigorous definition.

THEOREM 3.2. *Define $A = \{r \in \mathbb{Q} : r^2 < 2\}$. Then A has no least upper bound in $\mathbb{Q}$. Consequently, $\mathbb{Q}$ does not have the least-upper-bound property.*

We give the proof after a rather extended discussion into its strategy. Clearly A is nonempty and bounded above; 2 is an upper bound, for example. So proving that A has no least upper bound will prove that $\mathbb{Q}$ does not have the least-upper-bound property. We break up the proof into two steps:

(1) $p \in \mathbb{Q}$ is an upper bound for A if and only if $p^2 > 2$ and $p > 0$.
(2) If $p \in \mathbb{Q}$, $p^2 > 2$ and $p > 0$, then there exists $q \in \mathbb{Q}$ such that $0 < q < p$ and $q^2 > 2$.

Suppose these two statements are proven, and let $p \in \mathbb{Q}$ be an upper bound for A. Then $p^2 > 2$ and $p > 0$ by (1); then by (2) there exists $q \in \mathbb{Q}$ such that $0 < q < p$ and $q^2 > 2$, which implies (by (1) again) that q is an upper bound for A. But since $q < p$, it follows that p is not the least upper bound for A. But p was an arbitrary upper bound for A in $\mathbb{Q}$; therefore no upper bound for A in $\mathbb{Q}$ can be the least upper bound in $\mathbb{Q}$. Thus A has no least upper bound in $\mathbb{Q}$.

Though statement (1) probably looks intuitive, it still requires a bit of untangling. It should be clear that $p^2 > 2$ and $p > 0$ together imply that p is an upper bound for $\mathbb{Q}$. (We'll write this down explicitly in the actual proof, though.) To rigorously show that 'p is an upper bound' implies $p^2 > 2$, though, we need to argue the contrapositive. That is, we need to show that if $p^2 \leq 2$ and $p \in \mathbb{Q}$, then p is not an upper bound for A in $\mathbb{Q}$. We have just shown that $p^2 \neq 2$, so we essentially need to show that if $p^2 < 2$ (and $p > 0$, without loss of generality), then there exists $r \in A$ such that $p < r$. This should look similar to statement (2). It turns out that in order to prove statements (1) and (2), the relevant task is the following:

Given $p \in \mathbb{Q}$ such that $p > 0$, find $q \in \mathbb{Q}$ strictly in between p and $\sqrt{2}$.

The reader might object that the use of $\sqrt{2}$ in this reasoning is 'cheating', since we don't yet have access to the full real number system. To this, my response is the following: 'Cheating' like this is completely fair game when trying to figure out how to complete a problem, provided that the logic of the actual proof doesn't rely on it. In this case, we will see that the quantity $\sqrt{2}$ does not actually appear anywhere in the proof below, even though we use it extensively when devising our strategy.

Let's consider first the case where $p > \sqrt{2}$. We want to find a rational number q such that $\sqrt{2} < q < p$. Our first attempt might be to subtract some small (rational) number $\varepsilon > 0$ from p. But this strategy will only work for some p's; once our p's get very close to $\sqrt{2}$, we will need to choose a different ε. That is, the quantity we subtract from p will itself need to depend on p. So, what's a small, positive rational number that depends on p? we know that $p^2 - 2 > 0$, so we might try something like $q = p - \varepsilon_p$, where ε_p involves the expression $p^2 - 2$ somehow. In order to guarantee that ε_p is small enough, in the sense that we still have $(p - \varepsilon_p)^2 > 2$, we reverse engineer a little more. We need $p - \sqrt{2} > \varepsilon_p > 0$; we try setting $\varepsilon_p = \frac{p^2-2}{r}$, where r is some positive rational number. (Then ε_p will still be a rational number, since $\mathbb{Q}$ is a field.) How large must we take r in order to guarantee that $\frac{p^2-2}{r} < p - \sqrt{2}$? Well, rearranging, we get

$$r > \frac{p^2 - 2}{p - \sqrt{2}} = \frac{(p - \sqrt{2})(p + \sqrt{2})}{p - \sqrt{2}} = p + \sqrt{2}.$$

So we put $r = p + 2$, and $\varepsilon_p = \frac{p^2-2}{p+2}$. Note that the '2' in the equation $r = p + 2$ could have been any rational number greater than $\sqrt{2}$. We suspect now that our choice gives $p > p - \varepsilon_p > \sqrt{2}$, which is

essentially what we need. Actually, the case where $p < \sqrt{2}$ can be dealt with using the *same choice of* q! The reader should take a moment to convince themself that this is to be expected before reading the actual proof.

Having brainstormed the strategy of the proof, we can now write it out fairly succinctly.

PROOF OF THEOREM 3.2. Given $p \in \mathbb{Q}$, $p > 0$, set $q = p - \frac{p^2-2}{p+2}$. We claim that q is a positive rational number, and that q^2 is always between 2 and p^2. Indeed, $q \in \mathbb{Q}$ follows from the fact that $\mathbb{Q}$ is a field. Next,

$$p - q = \frac{p^2 - 2}{p + 2};$$

therefore $p - q$ and $p^2 - 2$ have the same sign. (So if $p^2 > 2$, then $p > q$; if $p^2 < 2$, then $p < q$.) Next, we write

$$q = \frac{p(p+2)}{p+2} - \frac{p^2-2}{p+2} = \frac{2p+2}{p+2} = \frac{2(p+1)}{p+2}.$$

This shows that $q > 0$; furthermore,

$$\begin{aligned} q^2 - 2 &= \frac{4(p+1)^2}{(p+2)^2} - \frac{2(p+2)^2}{(p+2)^2} = \frac{4(p^2+2p+1) - 2(p^2+4p+4)}{(p+2)^2} \\ &= \frac{2p^2-4}{(p+2)^2} = \frac{2(p^2-2)}{(p+2)^2}. \end{aligned}$$

Thus q^2-2 and p^2-2 have the same sign. (So if $p^2 > 2$, then $q^2 > 2$; if $p^2 < 2$, then $q^2 < 2$.) Combining this with the previous step, we conclude that either $p^2 < q^2 < 2$, or $p^2 > q^2 > 2$.

Now we prove that A has no least upper bound in $\mathbb{Q}$, in two steps. First, we claim that if p is an upper bound for A in $\mathbb{Q}$ then $p^2 > 2$. Indeed, we have shown above that if $p \in \mathbb{Q}$, then p^2 cannot be equal to 2; therefore if $p^2 > 2$ fails, we must have $p^2 < 2$. But then q as defined above satisfies $p^2 < q^2 < 2$, $q > 0$, $q \in \mathbb{Q}$. Therefore $q \in A$ and $p < q$. It follows that p is not an upper bound for A. This proves that if p is an upper bound for A in $\mathbb{Q}$, then $p^2 > 2$.

Now let p be any upper bound for A in $\mathbb{Q}$. We show that p is not the *least* upper bound for A in $\mathbb{Q}$. Indeed, if $p^2 > 2$, then q as defined above satisfies $p^2 > q^2 > 2$, $q > 0$, $q \in \mathbb{Q}$. It follows that q is strictly less than p; we claim that additionally q is an upper bound for A. Indeed, if there exists $r \in A$ such that $r > q$, then $r^2 > q^2 > 2$, contradicting the definition of A. Therefore q is an upper bound for A which is strictly less than p. So p is not the least upper bound. We conclude that A has no least upper bound in $\mathbb{Q}$, as desired.

Since A is nonempty (as $1 \in A$) and bounded above (by 2, for instance), but A has no least upper bound in $\mathbb{Q}$, it follows that $\mathbb{Q}$ does not have the least upper bound property. □

REMARK 3.3. The extensive motivation that preceded the proof in this case is a luxury you as a reader won't always have. Without it, however, you might be convinced of the truth of the claimed statement, but without any idea of how the writer came up with the strategy. If reading a proof is to have any benefit to your ability to write similar proofs, you should make sure you understand both how *and why* each step follows logically from the previous ones. This means that you will have to deconstruct some of the proofs you see, in a manner similar to how we motivated this proof.

EXERCISE 3.1. Using the strategies similar to those of the proofs in this section, prove the following statements.

(a) There is no rational number whose square is 20.
(b) The set $A := \{r \in \mathbb{Q} : r^2 < 20\}$ has no least upper bound in $\mathbb{Q}$.

Hint: Most of the solution for both parts can be directly copied from the proof of the corresponding result in this section. The key differences are as follows: In (a), the 'common factor' of m and n from the proof

of Proposition 3.1 needs to be modified in order to reach a contradiction in the present circumstances; in (b), the number q and the associated calculations from the proof of Theorem 3.2 require modification.

4. Definition and Basic Properties of $\mathbb{R}$

THEOREM 4.1. *There exists an ordered field which has the least-upper-bound property. This field is unique up to isomorphism, and it contains $\mathbb{Q}$ as a subfield.*

We will not prove this Theorem. For a proof, we refer the reader to either Rudin's *Principles of Mathematical Analysis* or Pugh's *Real Mathematical Analysis*. (The latter is less concise but easier to read.)

DEFINITION 4.2. $(\mathbb{R}, +, \cdot, \leq)$ is defined to be the field from the Theorem above.

In this section, we prove the following three properties of $\mathbb{R}$:

(1) $\mathbb{R}$ has the Archimedean property.
(2) $\mathbb{Q}$ is dense in $\mathbb{R}$.
(3) nth roots of positive real numbers exist in $\mathbb{R}$.

PROPOSITION 4.3. *$\mathbb{R}$ has the Archimedean property.*

PROOF. Choose real numbers x and y such that $x > 0$. We need to show that $nx > y$ for a sufficiently large integer n. This lends itself well to an argument by contradiction: If $nx \leq y$ for all $n \in \mathbb{N}$, then y is an upper bound for the (nonempty) set $A := \{nx : n \in \mathbb{N}\}$. Since $\mathbb{R}$ has the least upper bound property, A has a least upper bound α in $\mathbb{R}$. Since α is the least upper bound, $\alpha - x$ is not an upper bound for A, i.e. there is an integer n for which $nx > \alpha - x$. But then $(n+1)x > \alpha$, contradicting the fact that α is an upper bound for A. □

Recall that $\mathbb{Q}$ has the Archimedean property as well. This might cause one to wonder whether the LUBP is really necessary when proving that $\mathbb{R}$ has the Archimedean property. Technically, the answer to this question is negative; one does *not* need the LUBP. However, the proof that avoids the LUBP uses the explicit construction of $\mathbb{R}$ from $\mathbb{Q}$. The proof above is much simpler.

PROPOSITION 4.4. *If $x, y \in \mathbb{R}$ and $x < y$, then there exists a $p \in \mathbb{Q}$ such that $x < p < y$.*

This statement is sometimes rephrased as saying that $\mathbb{Q}$ is *dense* in $\mathbb{R}$. We will later meet another, more general definition of the term *dense*, in the context of metric (or topological) spaces. In the limited context of the real line, these two definitions are the same.

Before we write down the proof, we give the idea. For fixed x and y (not equal), it would be easier to prove the existence of a rational number between them if we knew they were far apart. In fact, if we know $y - x > 1$, then we can find an *integer* between them. This is a pretty obvious statement, but it's a helpful step to record explicitly.

LEMMA 4.5. *If x and y are real numbers with $y - x > 1$, then there exists an integer m such that $x < m < y$.*

PROOF. let m be the smallest integer greater than x. Then $m - 1 < x$ and $x < y - 1$, so $m < y$. Thus $x < m < y$. □

Now, back to the discussion of the density of $\mathbb{Q}$ in $\mathbb{R}$. Even if x and y are close together, we can 'zoom in' by multiplying both by some large integer n, so that nx and ny are far apart. Then we can find an integer m between nx and ny, which guarantees that m/n is between x and y. The formal argument is below.

PROOF OF PROPOSITION 4.4. Choose n large enough so that $ny - nx > 1$. This is possible by the Archimedean property of $\mathbb{R}$ because $ny - nx = n(y - x)$ and $y - x > 0$. Let m be the smallest positive integer greater than nx. We claim that $x < m/n < y$; since $nx < m$ by definition of m, it remains to show that $m < ny$. Now, since m is the *smallest* integer greater than nx, we have $m - 1 \leq nx$. On the other hand, our choice of n guarantees that $nx < ny - 1$. Combining these two inequalities gives $m - 1 \leq nx < ny - 1$, or adding 1 to both sides, $m < ny$. This finishes the proof. □

Elements of $\mathbb{R}\backslash\mathbb{Q}$ are called *irrational numbers*. We can already prove that irrational numbers exist. Indeed, we have shown that $\mathbb{Q}$ does not have the least upper bound property; i.e., there exists a subset A of $\mathbb{Q}$ which is nonempty and bounded above, such that A has no least upper bound in $\mathbb{Q}$. But A must have a least upper bound x in $\mathbb{R}$; this x cannot be rational (otherwise it would be least upper bound of A in $\mathbb{Q}$, contrary to assumption), therefore $x \in \mathbb{R}\backslash\mathbb{Q}$.

EXERCISE 4.1. Prove the following statements about rational and irrational numbers.

(a) Assume r is rational and x is irrational. Show that $r + x$ is irrational. Show that rx is irrational unless $r = 0$.
(b) Use the Archimedean property of $\mathbb{R}$ to prove that the set of irrational numbers is dense in $\mathbb{R}$. (Hint: Let x be any positive irrational number. If y and z are real numbers with $z - y > x$, then there exists an integer m such that $y < mx < z$.)

We know that $\mathbb{R}$ is an ordered field with the least upper bound property; therefore the set $A = \{r \in \mathbb{Q} : r^2 < 2\}$ from the previous section has an upper bound in $\mathbb{R}$. Intuitively, it should be clear that the only reasonable candidate for $\sup A$ should be the object that we normally call $\sqrt{2}$. However, note carefully that we have not yet proven that $(\sup A)^2 = 2$. The following Proposition puts our intuition on rigorous footing, and actually establishes the existence of nth roots (not just square roots) for positive real numbers.

PROPOSITION 4.6. *For every real $x > 0$ and every integer $n > 0$ there is one and only one positive real number α such that $\alpha^n = x$.*

A couple of notes before the proof: Whenever a statement claims existence of an object satisfying some properties, the first step in the proof should be to come up with a candidate for that object. (It's pretty hard to prove properties of an object which has no specified identity.) In this case, the object should be the supremum of the set $E := \{t \in \mathbb{R},\, t > 0 : t^n < x\}$. Basically, the logic of (the existence part of) the proof is as follows (note that proving the existence of a *candidate* for the nth root is the very first step).

Step 1. Show that $\alpha = \sup E$ exists.
Since $\mathbb{R}$ has the LUBP, this amounts to showing that E is nonempty and bounded above.
Step 2. Show that if $\beta^n \neq x$, then $\beta \neq \sup E$, in two steps:
(a) If $\beta^n < x$, then β is not an upper bound for E.
(b) If $\beta^n > x$, then β is not the *least* upper bound for E.
Step 3. Conclude from Steps 1 and 2 that $\alpha^n = x$.

We are now ready to begin the proof. We will again be required to make some rather clever choices for various parameters that enter the argument. Some of the motivation for these choices is described in the footnotes. However, the reader should start to prepare to reverse engineer these sorts of choices themself when reading proofs in the future.

PROOF. Before we get to the existence part of the statement, we deal with the much easier uniqueness claim; that is, we show that there can be *at most* one positive real number α satisfying $\alpha^n = x$. Indeed, if α_1 and α_2 are distinct positive numbers, then one is bigger, say α_2, and then $\alpha_1^n < \alpha_2^n$. In particular, α_1^n and α_2^n cannot both be equal to x.

Now we prove existence of such a α, following the outline in the remarks above. As above, define $E := \{t \in \mathbb{R},\, t > 0 : t^n < x\}$.

Step 1. We show that $\alpha = \sup E$ exists. First, we show that E is not empty. Indeed, define $t_1 := \frac{x}{x+1}$. Then $t_1 < 1$ implies $t_1^n < t_1$; since $t_1 < x$ we also have $t_1^n < x$, i.e. $t_1 \in E$. Next, E is bounded above. Indeed, define $t_2 := x + 1$; we show that $t \leq t_2$ for every $t \in E$, i.e. t_2 is an upper bound for E. Actually, we show the contrapositive: whenever $t > t_2$ it follows that $t \notin E$. Indeed, since $t_2 > 1$, we have $t^n > t_2^n > t_2 > x$ in this case. This proves that $x + 1$ is an upper bound for E. Now since $\mathbb{R}$ has the least upper bound property, it follows that E has a least upper bound α in $\mathbb{R}$.

Step 2.(a) We show that if $\beta^n < x$, then β is not an upper bound for E. The statement is obvious if $\beta \leq 0$, so we consider only the nontrivial case $\beta > 0$. To show that β is not an upper bound, we show that there exists $\varepsilon > 0$ such that $\gamma := \beta + \varepsilon$ is an element of E. In particular, we can choose[2]

$$\varepsilon = \min\left\{1, \frac{x - \beta^n}{n(\beta+1)^{n-1}}\right\}.$$

Then

$$\gamma^n - \beta^n < \varepsilon n \gamma^{n-1} \leq \varepsilon n(\beta+1)^{n-1} \leq \frac{x-\beta^n}{n(\beta+1)^{n-1}} \cdot n(\beta+1)^{n-1} = x - \beta^n.$$

Thus $\gamma^n < x$, so $\gamma \in E$. Thus β is not an upper bound for E. We conclude that $\alpha^n \geq x$.

Step 2.(b) We show that if $\beta^n > x$, then β is not the *least* upper bound for E. As in the previous step, we assume without loss of generality that $\beta > 0$. If $\beta^n > x$, put $\gamma = \beta - \varepsilon$ and choose $\varepsilon > 0$ small enough so that $\beta^n - \gamma^n < \beta^n - x$. In particular, we can take[3]

$$\varepsilon = \frac{\beta^n - x}{n\beta^{n-1}}.$$

Then

$$\beta^n - \gamma^n < \varepsilon n \beta^{n-1} = \beta^n - x,$$

and rearranging gives $\gamma^n > x$. Therefore, if $t > \gamma$, then $t^n > \gamma^n > x$, so $t \notin E$. Thus γ is an upper bound for E that is strictly smaller than β, i.e. β is not the least upper bound for E.

Step 3. We conclude. Steps 2(a) and 2(b) show collectively that if $\beta^n \neq x$, then $\beta \neq \sup E$. Therefore $\alpha^n = x$; that is, α is the desired nth root of x. □

EXERCISE 4.2. Assume $a, b \in \mathbb{R}$. Prove that $a \leq b$ if and only if $a \leq b + \varepsilon$ for every $\varepsilon > 0$.

EXERCISE 4.3. Let E be a set of real numbers, and let s be an upper bound for E. Prove that $s = \sup E$ if and only if for every $\varepsilon > 0$ there exists $x \in E$ such that $x > s - \varepsilon$.

EXERCISE 4.4. Let A and B be nonempty sets of real numbers. Decide whether the following statements are true or false. If true, give a proof; if false, give a counterexample.

(a) If $\sup A < \inf B$, then there exists a $c \in \mathbb{R}$ satisfying $a < c < b$ for all $a \in A$ and $b \in B$.

(b) If there exists a $c \in \mathbb{R}$ satisfying $a < c < b$ for all $a \in A$ and $b \in B$, then $\sup A < \inf B$.

[2] We want to choose $\varepsilon > 0$ small enough so that γ is still in E. How small does ε need to be? Well, we want $\gamma = \beta + \varepsilon \in E$, i.e. $(\beta + \varepsilon)^n < x$. We also know that $\beta^n < x$, so it will be good enough to find ε satisfying $(\beta+\varepsilon)^n - \beta^n < x - \beta^n$. Now we try to put something in between the left and right sides of this inequality. To do so, we rewrite the left side using a telescoping sum, then estimate $\beta < \gamma$.

$$\gamma^n - \beta^n = (\gamma - \beta)(\gamma^{n-1} + \gamma^{n-2}\beta + \cdots + \gamma\beta^{n-2} + \beta^{n-1}) < \varepsilon \cdot n \cdot \gamma^{n-1}.$$

Thus, in order for ε to be small enough, we just need $\varepsilon n(\beta+\varepsilon)^{n-1} < x - \beta^n$. Actually, let's make the LHS of this very last inequality just a little bigger, so that we can get ε by itself. As long as $\varepsilon \leq 1$, it suffices to choose ε small enough so that $\varepsilon n(\beta+1)^{n-1} \leq x - \beta^n$.

[3] Reasoning as before, we have

$$\beta^n - \gamma^n < \varepsilon n \beta^{n-1},$$

so it suffices to choose $\varepsilon > 0$ small enough so that $\varepsilon n \beta^{n-1} \leq \beta^n - x$.

5. Subsets of $\mathbb{R}$, and the Extended Real Number System $\overline{\mathbb{R}}$

Though working in $\mathbb{R}$ is certainly much more convenient for most purposes than working in $\mathbb{Q}$, it is rather annoying that the supremum of a set of real numbers is not always defined in $\mathbb{R}$. We correct this deficiency by defining the extended real number system $\overline{\mathbb{R}}$, where the supremum and infimum always exist for nonempty sets.

DEFINITION 5.1. The *extended real number system* $(\overline{\mathbb{R}}, \leq)$ is an ordered set, defined by the following. As a set, $\overline{\mathbb{R}}$ is simply the set formed by adjoining the two symbols $+\infty$ and $-\infty$ to the set $\mathbb{R}$ of real numbers. Within $\mathbb{R}$, the order $<$ remains the same, but we define $-\infty < x < +\infty$ for every $x \in \mathbb{R}$.

We make the following remarks:

- By this definition, $\overline{\mathbb{R}}$, and every subset thereof, is bounded above by $+\infty$. Therefore the hypothesis 'nonempty and bounded above' is equivalent to 'nonempty' when working in $\overline{\mathbb{R}}$.
- Suppose E is a nonempty subset of $\mathbb{R}$ that is not bounded above in $\mathbb{R}$. Then $\sup E$ is not defined in $\mathbb{R}$, but the supremum of E in $\overline{\mathbb{R}}$ is $+\infty$, since $+\infty$ is in fact[4] the *only* upper bound for E in $\overline{\mathbb{R}}$.

In light of the above, it is easy to see that $(\overline{\mathbb{R}}, \leq)$ has the LUBP.

DEFINITION 5.2. Let a, b be elements of $\overline{\mathbb{R}}$ such that $a < b$. We use special notation for the following subsets of $\overline{\mathbb{R}}$:

$$(a,b) = \{x \in \overline{\mathbb{R}} : a < x < b\}, \qquad [a,b) = \{x \in \overline{\mathbb{R}} : a \leq x < b\},$$
$$(a,b] = \{x \in \overline{\mathbb{R}} : a < x \leq b\}, \qquad [a,b] = \{x \in \overline{\mathbb{R}} : a \leq x \leq b\}.$$

If $a \neq -\infty$ and $b \neq +\infty$, then these sets are all called *intervals*; more specifically (a,b) is an *open interval*, $(a,b]$ and $[a,b)$ are *half-open intervals*, and $[a,b]$ is a *closed interval*. For any $c \in \mathbb{R}$, the sets $(c, +\infty)$ and $(-\infty, c)$ are called *open rays* of $\mathbb{R}$, and the sets $[c, \infty)$ and $(-\infty, c]$ are called *closed rays* of $\mathbb{R}$.

Note the following:

- Whenever A is a subset of $\overline{\mathbb{R}}$ that does not contain $-\infty$ or $+\infty$, we always consider it as a subset of $\mathbb{R}$ unless explicitly stated otherwise.
- There is a small risk of confusion between the subset (a,b) of $\mathbb{R}$ and the ordered pair (x,y) in $A \times B$, with $x \in A$ and $y \in B$. However, the distinction should be clear from context as long as the reader realizes that the same notation is used for the two different meanings.

REMARK 5.3. Though working in $\overline{\mathbb{R}}$ has its technical advantages, the introduction of the symbols $+\infty$ and $-\infty$ has the unpleasant side effect of creating lots of extra cases in the proofs of Theorems. Furthermore, the symbols $+\infty$ and $-\infty$ are rarely dealt with directly, but rather as byproducts of the unboundedness of certain sets. Therefore we make the following conventions.

- If E is a subset of $\mathbb{R}$, the statements 'E is bounded above' and 'E is not bounded above' are to be interpreted as 'E is bounded above in $\mathbb{R}$' and 'E is not bounded above in $\mathbb{R}$', respectively.
- If E is a nonempty subset of $\mathbb{R}$ or $\overline{\mathbb{R}}$, the notation $\sup E$ will refer to the supremum in $\overline{\mathbb{R}}$ unless otherwise specified.
- Arithmetic: If $x \in \mathbb{R}$, then $x + \infty = +\infty$, $x - \infty = -\infty$, $\frac{x}{+\infty} = \frac{x}{-\infty} = 0$. If $y \in \overline{\mathbb{R}}$ and $y > 0$, then $y \cdot (+\infty) = +\infty$, $y \cdot (-\infty) = -\infty$; if $z \in \overline{\mathbb{R}}$ and $z < 0$, then $z \cdot (+\infty) = -\infty$, $z \cdot (-\infty) = +\infty$. Despite these conventions, we stress that $\overline{\mathbb{R}}$ is *not* a field. In particular, $\infty - \infty$ is not defined.

[4] Let M be any upper bound in $\overline{\mathbb{R}}$. Clearly $M \neq -\infty$, since $x > -\infty$ for all $x \in E$. Furthermore, $M \notin \mathbb{R}$, since E is by assumption not bounded above in $\mathbb{R}$.

EXERCISE 5.1. Let a and b be real numbers. Show that the following three equalities hold:

$$\bigcap_{x>b}(a,x) = (a,b], \qquad \bigcup_{n=1}^{\infty}[a+\tfrac{1}{n},\ b-\tfrac{1}{n}) = (a,b), \qquad \bigcap_{n=1}^{\infty}(a+n,+\infty) = \emptyset.$$

EXERCISE 5.2. Let $a_1, a_2, \ldots$ be any enumeration of the negative rational numbers; let $b_1, b_2, \ldots$ be any enumeration of the positive rational numbers. Show that the following two equalities hold:

$$\bigcap_{j=1}^{\infty}(a_j, b_j) = \{0\}, \qquad \bigcup_{j=1}^{\infty}(a_j, b_j) = \mathbb{R}.$$

6. The Complex Field

In this section, we define complex numbers. These will be especially important in a second course in Analysis.

DEFINITION 6.1. The set $\mathbb{C}$ of complex numbers is defined simply as $\mathbb{R} \times \mathbb{R}$. The *complex field* $(\mathbb{C}, +, \cdot)$ is defined by the following operations: If $x = (a,b)$, $y = (c,d)$, then

$$x + y = (a,b) + (c,d) = (a+c,\ b+d),$$
$$xy = (a,b)\cdot(c,d) = (ac - bd,\ ad + bc).$$

The additive and multiplicative identities are $(0,0)$ and $(1,0)$, respectively.

In the above equalities, note that the addition and multiplication operations on the far right take place in $\mathbb{R}$; we emphasize that these rules constitute the *definition* of complex addition and multiplication written on the left.

One should check that the addition and multiplication defined here satisfy the field axioms of Definition 2.1. The reader is asked to provide the details in Exercise 6.1 below.

The real field $\mathbb{R}$ can be thought of as a subset (actually a 'subfield') of $\mathbb{C}$ in a natural way via the map $f : \mathbb{R} \to \mathbb{C}$ defined by $f(a) = (a,0)$. We will always dispense with this mapping and instead simply write $\mathbb{R} \subset \mathbb{C}$ and $a = (a,0)$. We will also use the notation $i = (0,1)$. With these conventions, we have

$$i^2 = (0,1)\cdot(0,1) = (0\cdot 0 - 1\cdot 1,\ 0\cdot 1 + 1\cdot 0) = (-1,0) = -1.$$
$$a + bi = (a,0) + (b,0)\cdot(0,1) = (a,0) + (0,b) = (a,b).$$

DEFINITION 6.2. If $a, b \in \mathbb{R}$ and $z = a + bi$, the number $a - bi$ is called its *complex conjugate* and denoted by $\bar{z}$. The numbers a and b are called the *real* and *imaginary parts* of z, and denoted $a = \mathrm{Re}(z)$, $b = \mathrm{Im}(z)$.

REMARK 6.3. Two extremely important identities are the following:

$$\mathrm{Re}(z) = \frac{z + \bar{z}}{2}, \qquad \mathrm{Im}(z) = \frac{z - \bar{z}}{2i}.$$

The reader should take a moment to convince themself that these are true. Another important fact is that $z \in \mathbb{C}$ is a *real* number (i.e., $z = \mathrm{Re}(z)$ and $\mathrm{Im}(z) = 0$) if and only if $z = \bar{z}$.

Two more useful identities are the following:

$$\mathrm{Re}(iz) = -\mathrm{Im}(z), \qquad \mathrm{Im}(iz) = \mathrm{Re}(z).$$

We state the following Proposition without proof:

PROPOSITION 6.4. *Let z and w be complex numbers. Then*

$$\overline{z+w} = \bar{z} + \bar{w}, \qquad \overline{zw} = \bar{z}\bar{w}.$$

Furthermore, if $w \neq 0$, then $\overline{(\frac{z}{w})} = \frac{\bar{z}}{\bar{w}}$.

Note that the number $z\bar{z}$ is always real and nonnegative; in fact, if $z = a + bi$ (with $a, b \in \mathbb{R}$), then

$$z\overline{z} = (a + bi)(a - bi) = a^2 + b^2.$$

Geometrically, the quantity $|z| := (z\overline{z})^{\frac{1}{2}} = \sqrt{a^2 + b^2}$ represents the 'size' of $z = (a, b)$, in the sense that it is a measure of the distance from z to the zero element $0 = (0, 0)$ of the field $(\mathbb{C}, +, \cdot)$. This motivates the following definition.

DEFINITION 6.5. Given $z \in \mathbb{C}$, its *modulus*, or *absolute value*, is defined by $|z| = (z\bar{z})^{\frac{1}{2}}$.

It is easy to see from this definition that for any two complex numbers z, w, we have $|zw| = |z||w|$, and, provided $w \neq 0$, we also have $|\frac{z}{w}| = \frac{|z|}{|w|}$.

EXERCISE 6.1. Prove that the addition and multiplication operations in $(\mathbb{C}, +, \cdot)$ satisfy the field axioms of Definition 2.1.

EXERCISE 6.2. Prove that there exists no order $\leq$ that makes $(\mathbb{C}, +, \cdot, \leq)$ into an ordered field. (Hint: If there were such an ordering, then i would necessarily be either positive or negative.)

EXERCISE 6.3. Let $z = a + bi$ a complex number, with $a, b \in \mathbb{R}$ and $b \neq 0$. Explicitly identify two different complex numbers w_1 and w_2 such that $w_1^2 = w_2^2 = z$. Define w_1 and w_2 in terms of a and b (you may also use $|z| = \sqrt{a^2 + b^2}$). Do not use the complex exponential function, even if you happen to know about it.

CHAPTER 3

More Structures on Sets

1. Vector Spaces

The concept of a vector space is one that should be familiar from Linear Algebra. However, (normed) vector spaces provide useful intuition for the more general 'metric spaces' that will be our primary focus in the following chapters. While the concept of a metric space is strictly more general than that of a normed vector space, many of the most important metric spaces are actually normed vector spaces. Therefore, we give a few definitions and examples related to vector spaces. However, the reader should be aware that there is *much* more to be learned about vector spaces than the limited presentation here suggests. For example, we will not even treat the notions of a basis or the dimension of a vector space; these are fundamental to the theory of vector spaces but do not, in the author's opinion, motivate any fundamental concepts in the metric space theory (at least not to the extent that would justify a lengthy digression into their development at this juncture). Furthermore, with the exception of the definition of a vector space and some preliminary examples, we limit ourselves to the treatment of real and sometimes complex vector spaces.

1.1. Definition of a Vector Space.

DEFINITION 1.1. Let F be a field. A *vector space* V over the field F, or an F-vector space, is a set V, consisting of elements called *vectors*, together with two operations:

(1) Vector addition $V \times V \to V$, denoted by $+$ (e.g. $v + w$, where $v, w \in V$), and
(2) Multiplication of a vector by a scalar[1] $F \times V \to V$, denoted by juxtaposition (e.g. αv, where $\alpha \in F$, $v \in V$), or sometimes by $\cdot$ (e.g., $\alpha \cdot v$).

The vector space operations are required to satisfy the following *vector space axioms*:

- Vector addition is commutative and associative: If $u, v, w \in V$, then $u + v = v + u$ and $(u + v) + w = u + (v + w)$.
- V contains an additive identity, called the *zero vector* and denoted by 0 (sometimes $\mathbf{0}$, 0_V, or $\mathbf{0}_V$ if this heavier notation provides clarification), such that $0 + v = v$ for all $v \in V$.
- Every element v of V has an *additive inverse*, denoted $-v$, such that $v + (-v) = 0$.
- Multiplication by scalars is compatible with multiplication in F, in the sense that $\alpha(\beta v) = (\alpha\beta)v$ whenever $\alpha, \beta \in F$ and $v \in V$.
- If 1 denotes the multiplicative identity in F, then $1v = v$ for any $v \in V$.
- The following distributive laws hold. For any $\alpha, \beta \in F$ and $u, v \in V$, we have

$$\alpha(u + v) = \alpha u + \alpha v \qquad \text{and} \qquad (\alpha + \beta)v = \alpha v + \beta v.$$

An $\mathbb{R}$-vector space is also called a *real vector space*; a $\mathbb{C}$-vector space is also called a *complex vector space*. A *subspace* W of a vector space V is a subset of V which is an F-vector space in its own right, with the same operations as V. (Equivalently, a subspace is required to satisfy the condition that $\lambda v + w$ belongs to W whenever $v, w \in W$ and $\lambda \in F$, where the operations in $\lambda v + w$ are the operations in V.)

It is worth noting for the uninitiated that the notation in the definition above hides a lot of information. For example, consider the equality $(\alpha + \beta)v = \alpha v + \beta v$. On the left, the symbol $+$ denotes addition

[1] In the context of F-vector spaces, elements of F are also called *scalars*.

in the field F; on the right, the same symbol denotes addition in V, *which is a completely different operation!* Therefore some care is warranted (at least when first learning the theory) when parsing statements involving vector space operations.

As noted above, we will assume basic familiarity with vector spaces. However, we do give some examples before proceeding; this gives us a chance to set some notation as well.

1.2. Examples. Let F be a field. The following are examples of F-vector spaces.

EXAMPLE 1.2. F itself is an F-vector space. In this case, the vector space operations are the same as the operations in F.

EXAMPLE 1.3. F^n can be made into an F-vector space, for any $n \in \mathbb{N}$. Vector addition and multiplication by scalars are defined component-wise: For $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n, \lambda \in F$, we define

$$(\alpha_1, \ldots, \alpha_n) + (\beta_1, \ldots, \beta_n) = (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n),$$

$$\lambda(\alpha_1, \ldots, \alpha_n) = (\lambda\alpha_1, \ldots, \lambda\alpha_n).$$

The zero element in F^n is $(0, \ldots, 0)$, where 0 denotes the zero element in F.

EXAMPLE 1.4. If X is any set, then the set F^X of functions from X to F can be made into a vector space, by defining addition of functions and scalar multiples of functions pointwise. (The previous example is technically a special case of this one, with $X = J_n = \{1, \ldots, n\}$, but the former is important enough to mention separately.) Given $f, g \in F^X$ and $\alpha \in F$, we define $f + g$ and αf by

$$(f + g)(x) = f(x) + g(x), \qquad (\alpha f)(x) = \alpha(f(x)), \qquad \text{for all } x \in X.$$

We emphasize that the notation $(f + g)(x) = f(x) + g(x)$ is a *definition* of vector addition. That is, the symbol $+$ on the left side, between the two vectors f and g, is defined in terms of addition in F—the symbol $+$ on the right, between the two elements $f(x)$ and $g(x)$ of F. Similarly, the statement $(\alpha f)(x) = \alpha(f(x))$ is the *definition* of multiplication of elements in F^X by scalars in F.

The zero element in F^X is the function f such that $f(x) = 0$ for all $x \in X$. Note that under this definition, F^X cannot be a field under the usual definition $(fg)(x) = f(x)g(x)$ of pointwise multiplication (unless of course X has only one element). Indeed, if x and y are two distinct elements of X and $f(x) = 0$, $f(y) \neq 0$, then f is not the zero element of F^X, but it cannot have a multiplicative inverse under pointwise multiplication. However, the pointwise multiplication operation does make F^X into what is called a *commutative algebra* over F. Furthermore, even though elements of F^X do not in general have multiplicative inverses, we often use the notation $\frac{1}{f}$ to denote the function taking the value $\frac{1}{f(x)}$ at x, on the restricted domain $X \backslash f^{-1}(0) = \{x \in X : f(x) \neq 0\}$. Of course, $\frac{1}{f}$ is only an element of F^X if $f^{-1}(0)$ is empty. We also write $\frac{g}{f}$ for $g \cdot \frac{1}{f}$.

EXAMPLE 1.5. If V and W are F-vector spaces, then $V \times W$ can be made into an F-vector space in a natural way, by defining

$$c(v_1, w_1) + (v_2, w_2) = (cv_1 + v_2, cw_1 + w_2); \qquad (v_1, w_1), (v_2, w_2) \in V \times W,\ c \in F.$$

REMARK 1.6. The concept of an F-vector space makes sense for any field F, and the definitions and examples given above are completely independent of the structure of F. However, any further development will require specialized treatment that depends on exactly what F is. We will treat both real ($F = \mathbb{R}$) and complex ($F = \mathbb{C}$) vector spaces, with an emphasis on the former.

1.3. Normed Vector Spaces over $\mathbb{R}$ and $\mathbb{C}$.

DEFINITION 1.7. Let V be a real or complex vector space. A *norm* on V is a function $\|\cdot\|$ from V to $\mathbb{R}$, satisfying the following properties:

(1) (Nonnegativity) $\|v\| \geq 0$, and $\|v\| = 0$ if and only if v is the zero vector in V.
(2) (Homogeneity) $\|\alpha v\| = |\alpha|\|v\|$ for any scalar α and any $v \in V$.
(3) (The Triangle Inequality) $\|u + v\| \leq \|u\| + \|v\|$ whenever $u, v \in V$.

A *normed vector space* (or *normed linear space*) over F ($= \mathbb{R}$ or $\mathbb{C}$) is an F-vector space V on which a norm $\|\cdot\|$ is defined. We denote this normed vector space by $(V, \|\cdot\|)$, or simply by V, when the norm is understood from context.

REMARK 1.8. One might wonder what the 'dot' in the notation $\|\cdot\|$ means. The dot is just a placeholder, which says we're going to put something there. We could use similar notation for other functions, e.g., $f(\cdot)$ instead of f. In fact, this notation is sometimes also used if a function is defined on some product space $A \times B$, and its formula depends on the individual components of its argument. For example, consider the function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $g(x, y) = x^2 + 3y$. One can define a new function $h : \mathbb{R} \to \mathbb{R}$ by the formula $h(y) = 5^2 + 3y$, or one can simply use the notation $h = g(5, \cdot)$. More generally, one might wish to use the notation $k = g(x, \cdot)$ to mean that k is defined by the formula $k(y) = x^2 + 3y$, where x is unspecified but fixed.

EXAMPLE 1.9. We can define norms for some of the vector spaces in Example 1.4 as follows. (We use $F = \mathbb{R}$ for definiteness, but similar definitions can be made when $F = \mathbb{C}$.)

- $\mathbb{R}$ itself is a real normed vector space, with the absolute value serving as its norm.
- $\mathbb{R}^n$ is a real normed vector space, with the *Euclidean norm*

$$\|(x_1, \ldots, x_n)\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

  The fact that the triangle inequality holds for this norm requires some justification; this will be provided below.
- We can't make $\mathbb{R}^X$ into a real normed vector space for general sets X. However, we can define a norm on the subspace of *bounded* real-valued functions $B(X; \mathbb{R})$. (A real-valued function is called *bounded* if its image is a bounded subset of $\mathbb{R}$.) The *supremum norm* (or the *sup* norm for short), also called the *uniform norm*, is defined by

$$\|f\|_u = \sup_{x \in X} |f(x)|, \qquad f \in B(X; \mathbb{R}).$$

  Exercise 2.3 in Chapter 2 shows that this norm satisfies the triangle inequality.
- As a special case of the previous point, we consider the uniform norm on $\mathbb{R}^n$. Since J_n is a finite set, the supremum becomes a maximum:

$$\|x\|_u = \max\{|x_1|, \ldots, |x_n|\}.$$

EXERCISE 1.1. Let $\|\cdot\|$ be a norm on a real vector space V. Prove the *reverse triangle inequality*:

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

1.4. Real and Complex Inner Product Spaces.

DEFINITION 1.10. An inner product on an F-vector space V (with $F = \mathbb{R}$ or $\mathbb{C}$) is a function $\langle \cdot, \cdot \rangle$ from $V \times V$ to F, which satisfies the following properties, for any $x, y, z \in V$ and $\lambda \in F$:

- ((Conjugate) Symmetry) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for all $x, y \in V$,
- (Linearity in the first component) $\langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle$,
- (Positive-definiteness) $\langle x, x \rangle \geq 0$ for all $x \in V$, with $\langle x, x \rangle = 0$ if and only if x is the zero vector.

(The complex conjugate in the symmetry requirement is of course unnecessary—though not incorrect—for real inner products.)

A real vector space on which an inner product is defined is called a *real inner product space*. Similarly, a complex vector space on which an inner product is defined is called a *complex inner product space*. We sometimes refer to an inner product defined on a real vector space as a *real inner product*, or an inner product on a complex vector spaces as a *complex inner product*.

EXERCISE 1.2. Prove that any complex inner product is *conjugate linear* in its second argument; that is,

$$\langle x, \lambda y + z\rangle = \overline{\lambda}\langle x, y\rangle + \langle x, z\rangle,$$

for any scalar λ. (Note that this implies that any real inner product is linear in its second argument.)

EXAMPLE 1.11. The *dot product* on $\mathbb{R}^n$ is an inner product: If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, then their dot product $x \cdot y$ is defined by

$$x \cdot y = (x_1, \dots, x_n) \cdot (y_1, \dots, y_n) = x_1y_1 + \cdots + x_ny_n.$$

The symmetry and linearity requirements follow from basic properties of real numbers. The positive definiteness follows from the fact that $r^2 \geq 0$ for all $r \in \mathbb{R}$, with $r^2 > 0$ unless $r = 0$. Note that

$$x \cdot x = x_1^2 + \cdots + x_n^2 = \|x\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Thus Theorem 1.18 below will complete the verification that the Euclidean norm is in fact a norm. We can also define an analogous complex inner product on $\mathbb{C}^n$:

$$\langle (z_1, \dots, z_n), (w_1, \dots, w_n)\rangle = z_1\overline{w}_1 + \cdots + z_n\overline{w}_n.$$

We occasionally also refer to the norm associated to this inner product as the 'Euclidean norm' on $\mathbb{C}^n$.

EXAMPLE 1.12. We have a long way to go before we rigorously define the integral operator. However, the idea of the integral should be familiar from Calculus courses. On a certain class of functions (which we will not specify at this time), the integral is an inner product. Indeed, if f and g are (nice) functions, then (being intentionally ambiguous with the notation), we have $\int f\overline{g}\,\mathrm{d}x = \overline{\int g\overline{f}\,\mathrm{d}x}$, $\int (\lambda f + g)\overline{h}\,\mathrm{d}x = \lambda\int f\overline{h}\,\mathrm{d}x + \int g\overline{h}\,\mathrm{d}x$, and $\int f\overline{f}\,\mathrm{d}x = \int |f|^2\,\mathrm{d}x \geq 0$, with equality if and only if f is the zero element of this class of functions.

Even though we don't have the terminology to be more rigorous here, we point out the example of the integral to show that the dot product is not the only useful inner product.

THEOREM 1.13. *Let $(V, \langle\cdot,\cdot\rangle)$ be a real or complex inner product space. Then $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}$ is a norm on V.*

We prove this Theorem after several preparatory and hopefully familiar results. We use the notation $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ introduced in the statement of the Theorem.

DEFINITION 1.14. Let V be an inner product space with inner product $\langle\cdot,\cdot\rangle$, and let u and v be vectors in V. We say that u and v are *orthogonal* if $\langle u, v\rangle = 0$.

THEOREM 1.15 (Pythagorean Theorem). *Let V be a real or complex inner product space with inner product $\langle\cdot,\cdot\rangle$, and let u and v be orthogonal vectors in V. Then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$.*

PROOF. This follows immediately by expanding $\|u + v\|^2$:

$$\|u + v\|^2 = \langle u + v,\ u + v\rangle = \langle u, u\rangle + \underbrace{\langle u, v\rangle}_{=0} + \underbrace{\langle v, u\rangle}_{=0} + \langle v, v\rangle = \|u\|^2 + \|v\|^2.$$

□

DEFINITION 1.16. Let $(V, \langle\cdot,\cdot\rangle)$ be a real or complex inner product space; let u and v be elements of V, with $u \neq 0$. The *projection of* v onto u, denoted by $\mathrm{proj}_u(v)$, is defined to be the unique vector $w \in V$ such that $w = cu$ for some scalar c, and such that $v - w$ is orthogonal to u.

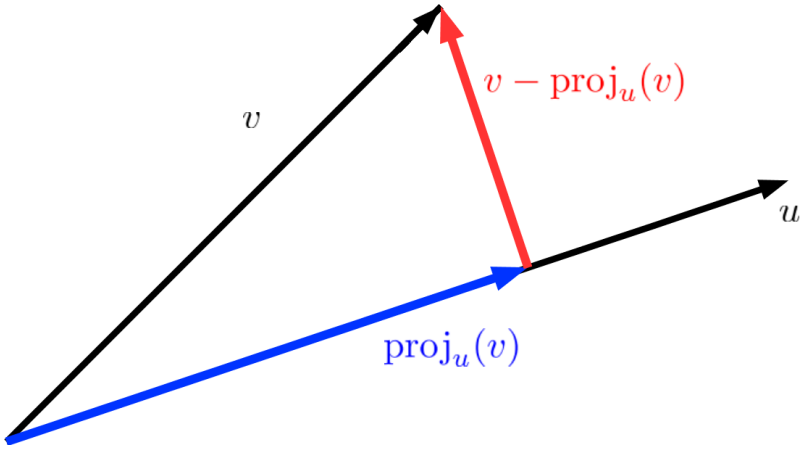

FIGURE 1. Schematic of an orthogonal projection

We can encode the requirements of the definition into the single equation

$$\langle v - cu, u\rangle = 0.$$

A one-line calculation shows that the above equation completely determines c (justifying the word 'unique' in our definition):

$$c = \frac{\langle v, u\rangle}{\|u\|^2}.$$

Thus $\mathrm{proj}_u v = cu$ must be given by the formula

$$\mathrm{proj}_u(v) = \frac{\langle v, u\rangle}{\|u\|^2} u. \tag{6}$$

It is easy to see that this vector does indeed satisfy the requirements of the orthogonal projection: It is manifestly a scalar multiple of u, and

$$\langle v - \mathrm{proj}_u(v), u\rangle = \langle v, u\rangle - \frac{\langle v, u\rangle}{\|u\|^2}\langle u, u\rangle = 0.$$

THEOREM 1.17 (Cauchy-Schwarz inequality). *Let $(V, \langle\cdot,\cdot\rangle)$ be a real or complex inner product space. For any $u, v \in V$, the following inequality holds.*

$$|\langle u, v\rangle| \leq \|u\|\|v\|.$$

PROOF. We break up v into $\mathrm{proj}_u v$ and $v - \mathrm{proj}_u v$, then use the fact that these two vectors are orthogonal to invoke the Pythagorean Theorem:

$$\begin{aligned}
\|v\|^2 &= \|\mathrm{proj}_u v + (v - \mathrm{proj}_u v)\|^2 \\
&= \|\mathrm{proj}_u v\|^2 + \|v - \mathrm{proj}_u v\|^2 \\
&= \left(\frac{|\langle v, u\rangle|}{\|u\|^2}\right)^2 \|u\|^2 + \underbrace{\|v - \mathrm{proj}_u v\|^2}_{\geq 0} \geq \frac{|\langle u, v\rangle|^2}{\|u\|^2}
\end{aligned}$$

Multiplying both sides by $\|u\|^2$ and taking square roots gives the desired inequality. □

Using the Cauchy-Schwarz inequality, we finally prove that $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ is a norm.

THEOREM 1.18. *Let $(V, \langle\cdot,\cdot\rangle)$ be a real or complex inner product space. Then $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}$ is a norm on V.*

PROOF. Expand $\|u+v\|^2$ to get

$$\|u+v\|^2 = \langle u+v,\ u+v\rangle = \langle u,u\rangle + 2\,\mathrm{Re}\langle u,v\rangle + \langle v,v\rangle \le \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2.$$

Take square roots to obtain the triangle inequality. The other requirements for a norm are clear. □

We end this section with two additional identities that are sometimes useful.

PROPOSITION 1.19 (Parallelogram Law). *Let $(V, \langle\cdot,\cdot\rangle)$ be a real or complex inner product space, and let $v, w \in V$. Then*

$$\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2.$$

The proof is a one-line computation, which we leave to the reader. The following identity is also not difficult to prove, but it is nontrivial enough to merit the designation 'Exercise.' (It is Exercise 1.3 below.)

PROPOSITION 1.20 (Polarization identity). *If $(V, \langle\cdot,\cdot\rangle)$ is a real inner product space, then*

$$\langle v, w\rangle = \frac{1}{4}[\|v+w\|^2 - \|v-w\|^2], \quad \textit{for all } v, w \in V.$$

If $(V, \langle\cdot,\cdot\rangle)$ is a complex inner product space, then

$$\langle v, w\rangle = \frac{1}{4}[(\|v+w\|^2 - \|v-w\|^2) + i(\|v+iw\|^2 - \|v-iw\|^2)], \quad \textit{for all } v, w \in V.$$

The upshot of the polarization identities is that any inner product is completely determined by the corresponding norm. Consequently, one can deduce properties of the inner product using only information about the norm. Furthermore, the polarization identities allow us to check immediately whether a given norm is one that arises from an inner product, by checking whether or not the right hand side of the polarization identities satisfy the requirements for an inner product (in particular, the linearity requirement).

EXERCISE 1.3. Prove Proposition 1.20. (Don't skip steps, but try to make the computations as efficient as possible.)

EXERCISE 1.4. Prove that the uniform norm on $\mathbb{R}^2$ cannot arise from an inner product. That is, prove that there is no real inner product $\langle\cdot,\cdot\rangle$ defined on $\mathbb{R}^2$ such that

$$\langle (x,y), (x,y)\rangle = \|(x,y)\|_u^2 = \max\{|x|, |y|\}^2$$

for all $(x, y) \in \mathbb{R}^2$.

2. Metric Spaces

The norm in a vector space give a notion of the 'size' of an element, or its 'distance' from the zero element in that space. Furthermore, one can talk about the 'distance' between two vectors v and u in terms of $\|u-v\|$, the 'distance' of $u-v$ from the zero element. However, the structure of a vector space is rather rigid, and we would like a sensible notion of distance for sets which are not vector spaces. The notion of a *metric space* will serve this need.

DEFINITION 2.1. Let X be a set. A *metric* on X is a function $d : X \times X \to \mathbb{R}$, satisfying the following properties:

- (Nonnegativity) $d(x,y) \ge 0$ for all if $x, y \in X$, with $d(x,y) = 0$ if and only if $x = y$.
- (Symmetry) $d(x,y) = d(y,x)$, for all $x, y \in X$.
- (Triangle Inequality) $d(x,y) \le d(x,z) + d(z,y)$ for all $x, y, z \in X$.

The set X together with the metric d is called a *metric space*, denoted (X, d) (though we simply refer to 'the metric space X' when d is understood from context). Elements of the set X are called *points*. The number $d(x, y)$ is called the *distance* from x to y, and a metric is sometimes called a *distance function*.

EXAMPLE 2.2.

- Any normed vector space is a metric space. More specifically, if V is a normed vector space with norm $\|\cdot\|$, then the function $d : V \times V \to \mathbb{R}$ given by $d(x, y) = \|x - y\|$ is a metric on V. Only the triangle inequality for d is not obvious. But, as one might expect, it follows from the triangle inequality for $\|\cdot\|$: Let x, y, z be points of X. then

  $$d(x, y) = \|x - y\| = \|(x - z) + (z - y)\| \leq \|x - z\| + \|z - y\| = d(x, z) + d(z, y).$$

- Not every metric space is a normed vector space. This is true for rather trivial reasons: If (X, d) is a metric space, X might not even be a vector space. However, even if X is a vector space, d might not define a norm on X. In fact, consider the function $d : X \times X \to \mathbb{R}$ defined by $d(x, y) = 1$ if $x \neq y$, $d(x, x) = 0$. Then d is a metric on *any* set, called the *discrete metric*. Clearly the discrete metric does not satisfy any kind of homogeneity requirement on $\mathbb{R}^n$, for example.

EXERCISE 2.1. Let X be any set. Prove that the discrete metric $d : X \times X \to \mathbb{R}$ (defined by $d(x, y) = 1$ if $x \neq y$ and $d(x, x) = 0$ for $x \in X$) satisfies the triangle inequality and is therefore a metric on X.

EXERCISE 2.2. For each of (a), (b), and (c), determine whether the given function d_j is a metric on $\mathbb{R}$, and prove that your answer is correct.

(a) $d_1(x, y) = \sqrt{|x - y|}$.
(b) $d_2(x, y) = |x - 2y|$.
(c) $d_3(x, y) = \frac{|x-y|}{1+|x-y|}$.

2.1. Open Balls in a Metric Space. It is often useful to talk about the points in a metric space (X, d) which are 'close' to a given point x. We introduce the following notation for the open 'ball' of radius r centered at x in the metric space (X, d):

$$B_{(X,d)}(x, r) := \{y \in X : d(x, y) < r\}.$$

If the metric d, or the set X is understood from context, we simply write $B_X(x, r)$ or $B_d(x, r)$. If both are understood, we simply write $B(x, r)$.

EXAMPLE 2.3. The set $B_{(X,d)}(x, r)$ may or may not 'look' like a ball, even in Euclidean spaces. For example, consider the following metrics on $\mathbb{R}^2$.

- $d_1(x, y) = \|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm. (The metric defined this way is called the *Euclidean metric*.) Then $B_{d_1}(x, r)$ is the set of points $y \in \mathbb{R}^2$ such that $\|x - y\| < r$. This really does look like a ball of radius r centered at x. (Note that it does not include the 'boundary' of the ball.)
- $d_2(x, y) = \|x - y\|_u = \max\{|x_1 - y_1|, |x_2 - y_2|\}$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$. (The notation $\|\cdot\|_u$ was introduced in Example 1.9 and denotes the uniform norm.) Then

  $$\begin{aligned} d_2(x, y) < r &\iff |x_1 - y_1| < r \text{ and } |x_2 - y_2| < r \\ &\iff x_1 - r < y_1 < x_1 + r \text{ and } x_2 - r < y_2 < x_2 + r \\ &\iff y = (y_1, y_2) \in (x_1 - r, x_1 + r) \times (x_2 - r, x_2 + r). \end{aligned}$$

  (Note that on the RHS we have the product of intervals, not ordered pairs.) Thus

  $$B_{d_2}(x, r) = (x_1 - r, x_1 + r) \times (x_2 - r, x_2 + r).$$

This is (the 'interior' of) a square of side length $2r$. For this reason the metric d_2 is sometimes called the *square metric* (as are its analogs in higher-dimensional Euclidean spaces).

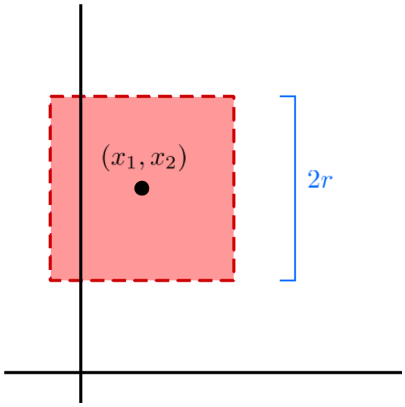

FIGURE 2. The open "ball" of radius r, centered at $(x_1, x_2) \in \mathbb{R}^2$, under the square metric.

EXERCISE 2.3. Consider the function $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, defined by

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2|, \qquad (x = (x_1, x_2), \;\; y = (y_1, y_2)).$$

(a) Prove that d is a metric on $\mathbb{R}^2$.
(b) On a sheet of graph paper, draw the set $B_d((5, 1), 3)$. Use dotted lines to indicate the 'boundary', which is not included in the set you are drawing. (Hint: it may be easier to figure out what the set looks like if you first consider $B_d((0, 0), 3)$.)
(c) On the same graph as in the previous part, draw $B_{d_u}((-3, 2), 1)$, where d_u denotes the square metric.

If (X, d) is a metric space and Y is a subset of X, then d is still a metric on the set Y. (And we always consider Y to be a metric space with the same metric d, unless explicitly noted otherwise.) If $x \in Y$ is 'near the edge' of Y, then $B_{(Y,d)}(x, r)$ might be a smaller set than $B_{(X,d)}(x, r)$. Indeed,

$$B_{(Y,d)}(x, r) = \{z \in Y : d(x, z) < r\} = \{z \in X : d(x, z) < r\} \cap Y = B_{(X,d)}(x, r) \cap Y.$$

EXAMPLE 2.4. For example, let $X = \mathbb{R}^2$, $Y = [-1, 3] \times [2, 4]$, and let d denote the Euclidean metric on $X = \mathbb{R}^2$. Let p denote the point $(3, 4)$, the upper right corner of Y. Then $B_{(Y,d)}(p, 1)$ looks like a quarter of the ball $B_{(X,d)}(p, 1)$. See Figure 3.

EXERCISE 2.4. Let (X, d) be a metric space, and let E be a subset of X. The *diameter* of E in (X, d) is defined by the formula

$$\operatorname{diam}_d(E) = \sup\{d(x, y) : x, y \in E\}.$$

(Usually we just write $\operatorname{diam}(E)$ when d is clear.)

(a) Prove that for any $r > 0$ and $x \in X$, we have $\operatorname{diam}(B(x, r)) \leq 2r$.
(b) If X is any set containing at least two elements and d is the discrete metric, show that $\operatorname{diam}(B(x, r)) = 0$ for any $r \leq 1$, while $\operatorname{diam}(B(x, r)) = 1$ for any $r > 1$.
(c) If $X = \mathbb{R}^n$ for some $n \in \mathbb{N}$ and d is the Euclidean metric, prove that $\operatorname{diam}(B(x, r)) = 2r$.

2.2. Interior Points. We have used the term 'interior' rather loosely in the preceding discussion. We can now clarify what we mean.

DEFINITION 2.5. Let (X, d) be a metric space and let U be a subset of X. A point x in U is called an *interior point* of U with respect to X if there exists $r > 0$ such that $B_X(x, r) \subset U$. The set of all

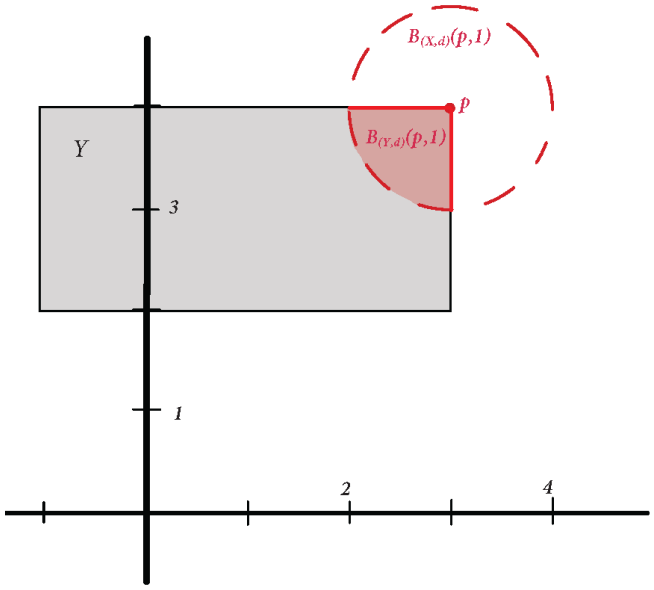

FIGURE 3. The ball $B_{(Y,d)}(p,1)$ (shaded red) is the portion of $B_{(X,d)}(p,1)$ that lies in Y.

interior points of U with respect to X is called the *interior* of U with respect to X, and we will denote it $\mathrm{Int}_X(U)$.

$$\mathrm{Int}_X(U) = \{x \in U : x \text{ is an interior point of } U \text{ with respect to } X\}.$$

If (X,d) is clear from context and it is the only metric space under consideration, we sometimes write U° instead of $\mathrm{Int}_X(U)$.

Note that Int_X preserves inclusions, in the sense that

$$U \subset V \implies \mathrm{Int}_X(U) \subset \mathrm{Int}_X(V).$$

Suppose (X,d) is a metric space and $U \subset Y \subset X$. If x is an interior point of U with respect to X, then (as we will see below) it is also an interior point of U with respect to Y. However, the converse is not true. In Example 2.6 (and Figure 3), for instance, the point p is an interior point of $U = B_Y(p,1)$ with respect to Y, but it is not an interior point of U with respect to X. The following Proposition sheds some light on the situation, and you are asked to think about it some more in Exercise 2.6.

PROPOSITION 2.6. *Let (X,d) be a metric space; assume $U \subset Y \subset X$. Then*

$$\mathrm{Int}_X(U) = \mathrm{Int}_Y(U) \cap \mathrm{Int}_X(Y) \qquad (U \subset Y \subset X). \tag{7}$$

PROOF. We prove only the inclusion $\subset$, leaving the other inclusion as Exercise 2.5. Given $x \in \mathrm{Int}_X(U)$, we can choose $r > 0$ such that $B_X(x,r) \subset U$. But then $B_Y(x,r) = B_X(x,r) \cap Y \subset U \cap Y \subset U$, so x is an interior point of U with respect to Y as well, i.e. $x \in \mathrm{Int}_Y(U)$. On the other hand, $B_X(x,r) \subset U \subset Y$ implies that $x \in \mathrm{Int}_X(Y)$ as well. It follows that $\mathrm{Int}_X(U) \subset \mathrm{Int}_Y(U) \cap \mathrm{Int}_X(Y)$, as needed. □

EXERCISE 2.5. Finish the proof of Proposition 2.6, by proving that

$$\mathrm{Int}_X(U) \supset \mathrm{Int}_Y(U) \cap \mathrm{Int}_X(Y) \qquad (U \subset Y \subset X)$$

EXERCISE 2.6. As in Example 2.4, let $X = \mathbb{R}^2$, $Y = [-1,3] \times [2,4]$, and let d denote the Euclidean metric on $X = \mathbb{R}^2$. Let $p = (3,4)$ and let $q = (2,4)$.

(a) Arguing *directly from the definition of an interior point* (i.e., without using Proposition 2.6, show that q is an interior point of $B_Y(p,2)$ with respect to Y, but q is not an interior point of

$B_Y(p, 2)$ with respect to X. In addition, draw a picture on a piece of graph paper that illustrates the idea of your proof.

(b) Give a short argument that re-establishes your conclusion from (a) but relies instead on Proposition 2.6.

In light of the results above, we must be careful about whether we are considering a set U as a subset of some large metric space (X, d) or a smaller one $Y \subset X$. However, if X is the only metric space under consideration, we often just say 'x is an interior point of U', and leave off the specification 'with respect to X'.

2.3. Open Sets in Metric Spaces.

DEFINITION 2.7. Let (X, d) be a metric space. A subset U of X is said to be *open* in X if every point of U is an interior point of U (with respect to X), that is, if $U = \mathrm{Int}_X(U)$.

The concept of an open set is *tremendously* important, as we will see starting very soon. In this subsection, we answer two questions:

(1) What are the open sets in a metric space (X, d)?
(2) If (X, d) is a metric space and $Y \subset X$, what is the relationship between the open sets in X and the open sets in Y?

Let us preview the answers to both of these questions. First of all, as one might expect from the terminology, the open balls $B_X(x, r)$ of a metric space are in fact open sets of X (Proposition 2.9). And it should be believable that any union of such balls should also be an open set (Proposition 2.11). What might be more surprising is that *any* open set can be written as some union of open balls (Theorem 2.12). That is, the collection of all open sets of X is completely determined by the collection of all open balls; the open balls, in turn, are determined by the metric d. These considerations motivate the following definition.

DEFINITION 2.8. We refer to the collection $\mathcal{T}$ of all open sets of a metric space (X, d) as the *topology* of X *generated* by the metric d (or simply, the *topology* of X, if the metric d is clear from context).

Thus, Question (1) is really asking us about the topology of (X, d); the discussion just above provides the answer and will be justified below. We are able to answer Question (2) as a consequence of our answer to Question (1); see Theorem 2.13 below. Let us now begin our discussion in earnest.

PROPOSITION 2.9. *Let (X, d) be a metric space. Then $B_X(x, r)$ is open in X.*

PROOF. Choose $y \in B_X(x, r)$. We must show that y is an interior point of $B_X(x, r)$. Put $\delta = r - d(x, y)$ (see Figure 4 for some motivation of this choice). We know that δ is positive, since $y \in B_X(x, r)$ implies $d(x, y) < r$. We claim that $B_X(y, \delta) \subset B_X(x, r)$, and that therefore y is an interior point of $B_X(x, r)$. To see this, let z be any point of $B_X(y, \delta)$. To show that $z \in B_X(x, r)$, we estimate its distance from x:

$$d(x, z) \leq d(x, y) + d(y, z) < d(x, y) + \delta = r,$$

where we have used the fact that $z \in B_X(y, \delta)$ to estimate $d(y, z) < \delta$, and we used the definition of δ to obtain the last equality. We have proved that $d(x, z) < r$, which means $z \in B_X(x, r)$, which means $B_X(y, \delta) \subset B_X(x, r)$, which means that y is an interior point of $B_X(x, r)$, which (finally) means that $B_X(x, r)$ is open in X, since y was an arbitrary point of $B_X(x, r)$. □

EXERCISE 2.7. Let (X, d) be a metric space, and let U be a subset of X. Use Proposition 2.9 to prove that $\mathrm{Int}_X(U)$ is open in X.

REMARK 2.10. In any metric space (X, d), the empty set is (vacuously) an open set in X, and X is (trivially) an open set in X.

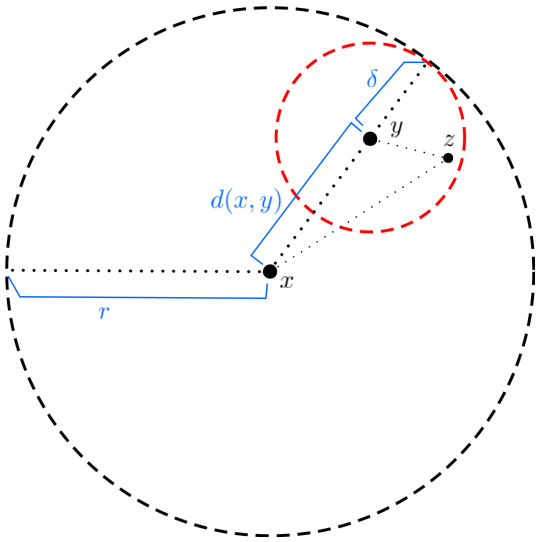

FIGURE 4. An illustration of the proof of Proposition 2.9.

PROPOSITION 2.11. *Let (X, d) be a metric space.*

- *Let $\mathcal{U}$ be a collection of open sets of X. Then the set $\bigcup_{U\in\mathcal{U}} U$ is open in X.*
- *Let $\{U_1, \dots U_n\}$ be a finite collection of open sets of X. Then $\bigcap_{i=1}^n U_i$ is an open set of X.*

PROOF. Denote $V = \bigcup_{U\in\mathcal{U}} U$. If V is empty, then there is nothing to show. Otherwise, let x be a point of V. Then $x \in U$ for some $U \in \mathcal{U}$. Since U is open, it follows that x is an interior point of U, so there exists $r > 0$ such that $B_X(x, r) \subset U \subset V$. Thus x is an interior point of V. Thus every point of V is an interior point, i.e. V is open in X.

Next, denote $W = \bigcap_{i=1}^n U_i$, and (assuming without loss of generality that W is nonempty) pick $x \in W$. Then $x \in U_i$ for every $i \in J_n$; since each U_i is open in X, we can choose $r_i > 0$ such that $B_X(x, r_i) \subset U_i$. Take $r = \min\{r_1, \dots, r_n\}$. Then

$$B_X(x, r) = \bigcap_{i=1}^n B_X(x, r_i) \subset \bigcap_{i=1}^n U_i = W.$$

Thus x is an interior point of W, so W is open in X. □

Note that the above Proposition does *not* claim that an arbitrary intersection of open sets is open. The latter statement is in fact false, as can be seen by the fact that $\bigcap_{n=1}^\infty (-\frac{1}{n}, \frac{1}{n}) = \{0\}$ is not an open set in $\mathbb{R}$, even though each of the sets $(-\frac{1}{n}, \frac{1}{n})$ is.

THEOREM 2.12. *Let (X, d) be a metric space, and let U be a subset of X. Then U is open in X if and only if U can be written as a union of open balls $B_X(x, r_x)$ of X.*

The notation r_x indicates that the radius of each ball may depend on its center point x.

PROOF. ($\implies$) Assume U is open in X. Then for each $x \in U$, there exists $r_x > 0$ such that $B_X(x, r_x) \subset U$. Therefore

$$\bigcup_{x\in U} B_X(x, r_x) \subset U.$$

On the other hand, every x in U is contained in $B_X(x, r_x)$, so U is contained in the union of these sets. That is,

$$U = \bigcup_{x\in U} \{x\} \subset \bigcup_{x\in U} B_X(x, r_x).$$

It follows that

$$\bigcup_{x\in U} B_X(x, r_x) = U,$$

which establishes the direction ($\Longrightarrow$).

($\Longleftarrow$) On the other hand, if U can be written as the union of open balls $B_X(x, r)$, then the fact that U is open in X follows immediately from Proposition 2.9 (open balls are open sets) and the first part of Proposition 2.11 (unions of open sets are open). □

We are now in a position to clarify the relationship between open sets in a metric space (X, d) and those in a smaller metric space (Y, d).

THEOREM 2.13. *Let (X, d) be a metric space, and let U and Y be subsets of X such that $U \subset Y \subset X$. Then U is open in Y if and only if $U = Y \cap V$ for some set V which is open in X.*

This Theorem should not be too surprising in light of the fact that $B_Y(x, r) = B_X(x, r) \cap Y$, as we have already observed. In fact, this observation plays a key role in the proof. It may be helpful to look again at Figure 3 for a reminder of what this prototypical scenario looks like.

PROOF. ($\Longrightarrow$) Assume first that U is open in Y. Then we can write U as a union of open balls $B_Y(x, r_x)$ of Y, by Theorem 2.12. Consequently,

$$U = \bigcup_{x\in U} B_Y(x, r_x) = \bigcup_{x\in U} (B_X(x, r_x) \cap Y) = Y \cap \bigcup_{x\in U} B_X(x, r_x).$$

Define V to be the union $\bigcup_{x\in U} B_X(x, r_x)$ on the right side of this equality. Then V is open in X, as it has been written as a union of open balls of X, and $U = Y \cap V$. This finishes the proof of the forward implication.

($\Longleftarrow$) On the other hand, assume that $U = Y \cap V$, where V is open in X. Assume without loss of generality that U is nonempty. Choose $y \in U$; we show that $y \in \mathrm{Int}_Y(U)$ to conclude. Since $y \in V$ and V is open in X, there exists $r > 0$ such that $B_X(y, r) \subset V$; taking the intersection with Y on both sides of this inclusion, we obtain

$$B_Y(y, r) = B_X(x, r) \cap Y \subset V \cap Y = U,$$

which finishes the argument. □

EXERCISE 2.8. Let (X, d) be a metric space. Assume that $U \subset Y \subset X$, and additionally that Y is open in X. Prove that U is open in Y if and only if U is open in X. (Note: There at least two possible solutions; one uses Theorem 2.13, the other uses Proposition 2.6.)

We collect just one more statement before moving on:

PROPOSITION 2.14. *Let (X, d) be a metric space. If V is an open set and $V \subset U$, then $V \subset \mathrm{Int}_X(U)$. Consequently,*

$$\mathrm{Int}_X(U) = \bigcup_{V\in\mathcal{U}} V; \qquad \mathcal{U} = \{V \in \mathcal{P}(X) : V \text{ is open and } V \subset U\}. \tag{8}$$

PROOF. We start with the first claim. Let V be an open set contained in U. Then $\mathrm{Int}_X(V) \subset \mathrm{Int}_X(U)$. But since V is open, this implies that $V = \mathrm{Int}_X(V) \subset \mathrm{Int}_X(U)$, as claimed.

Now we prove (8). By the first statement of the Proposition, if $V \in \mathcal{U}$, then $V \subset \mathrm{Int}_X(U)$. Therefore $\bigcup_{V\in\mathcal{U}} V \subset \mathrm{Int}_X(U)$. On the other hand, $\mathrm{Int}_X(U) \in \mathcal{U}$, by Exercise 2.7. Therefore $\mathrm{Int}_X(U) \subset \bigcup_{V\in\mathcal{U}} V$. The two inclusions we have proved establish the desired equality. □

2.4. Equivalent Metrics.

DEFINITION 2.15. Let X be a set, and let d_1 and d_2 be metrics on X. If d_1 and d_2 generate the same topology, we say that they are *equivalent metrics*.

EXERCISE 2.9. Suppose X is a finite, nonempty set and suppose d be a metric on X. Let $\mathcal{T}$ denote the topology generated by d. Show that $\mathcal{T} = \mathcal{P}(X)$. Conclude that any metric on X is equivalent to the discrete metric. (Hint: To show that $\mathcal{T} = \mathcal{P}(X)$, start by proving that $\{x\} = B_{(X,d)}(x, r_x)$ for some sufficiently small r_x, for each $x \in X$.)

EXERCISE 2.10. Prove that the Euclidean metric and the square metric are equivalent on $\mathbb{R}^n$.

The topology generated by either of these metrics is referred to as the *standard topology* on $\mathbb{R}^n$. Unless explicitly stated otherwise, we assume that $\mathbb{R}^n$ comes equipped with a metric that generates the standard topology—usually (but not always) the Euclidean metric.

3. Topological Spaces

3.1. Definition of a Topology. In the previous subsection, we characterized the collection of open subsets of a metric space in terms of the open balls of that metric space. A topology, is, more generally, a choice of what subsets of a given set X to call 'open', even if that set X is not a metric space.

In these notes, we will deal exclusively with topologies which are associated to a metric space (or at least a *metrizable* space—see Definition 3.7 below). The following more general definition will nevertheless be useful.

DEFINITION 3.1. Let X be a set. A *topology* on the set X is a subset $\mathcal{T}$ of $\mathcal{P}(X)$, i.e., a collection of subsets of X. The collection $\mathcal{T}$ is required to satisfy certain properties:

(1) $\emptyset, X \in \mathcal{T}$.
(2) $\mathcal{T}$ is closed under (arbitrary) unions: If $\mathcal{U} \subset \mathcal{T}$, then $\bigcup_{U \in \mathcal{U}} U$ is an element of $\mathcal{T}$.
(3) $\mathcal{T}$ is closed under finite intersections: If $U_1, \dots U_n \in \mathcal{T}$, then $\bigcap_{i=1}^n U_i$ is an element of $\mathcal{T}$.

A *topological space* is a set X together with a topology $\mathcal{T}$ on X, denoted $(X, \mathcal{T})$, or simply X when $\mathcal{T}$ is understood. The elements of $\mathcal{T}$ are called *open* subsets of X.

It should be noted immediately that in a metric space (X, d), the 'topology generated by the metric d' (c.f. Definition 2.8) is in fact a topology according to this more general Definition 3.1; Remark 2.10 and Proposition 2.11 provide the justification. What about the role of the 'open balls?' The notion of a *basis* provides us with the corresponding building blocks in this more general setting. See Definition 3.3 below. First, we give an example.

EXAMPLE 3.2. Consider the set $X = \{a, b, c\}$ of three (distinct) elements. We define three different topologies on it as follows. (There are several others that we have not listed.)

$$\mathcal{T}_1 = \{\emptyset, X\}, \quad \mathcal{T}_2 = \mathcal{P}(X) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, X\}, \quad \mathcal{T}_3 = \{\emptyset, \{a\}, \{b, c\}, X\}.$$

The reader can easily check that $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ all satisfy the definition of a topology on X. In fact, $\mathcal{T}_2$ is the same as the topology that arises from equipping X with the discrete metric. It is sometimes called the 'discrete topology.'

An example of a collection $\mathcal{N}$ of $\mathcal{P}(X)$ which is *not* a topology on $X = \{a, b, c\}$ is

$$\mathcal{N} = \{\emptyset, \{a\}, \{b\}, \{b, c\}, X\}.$$

Indeed, $\{a\}$ and $\{b\}$ are both elements of $\mathcal{N}$, but their union $\{a, b\}$ is not.

3.2. Basis of a Topology.

DEFINITION 3.3. A *basis* for a topology $\mathcal{T}$ is a subset $\mathcal{B}$ of $\mathcal{T}$ such that every element of $\mathcal{T}$ can be written as a union of elements of $\mathcal{B}$. If $\mathcal{B}$ is a basis for $\mathcal{T}$, then we refer to $\mathcal{T}$ as the topology *generated* by $\mathcal{B}$.

Theorem 2.12 tells us that in a metric space (X, d), the collection $\mathcal{B} = \{B_X(x, r) : x \in X; r > 0\}$ of all open balls forms a basis for the topology generated by the metric d. Thus, 'the topology generated by $\mathcal{B}$' and 'the topology generated by d' are the same thing in this case.

The following Proposition gives us a way to identify bases of a topology:

PROPOSITION 3.4. *Let X be a set, and let $\mathcal{B}$ be a collection of subsets of X which has the following properties:*

(1) Every $x \in X$ is contained in at least one element B of $\mathcal{B}$.

(2) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, then there exists a $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Then the following collection $\mathcal{T}$ is a topology on X:

$$\mathcal{T} = \left\{ U \in \mathcal{P}(X) : U = \bigcup_{B \in \mathcal{A}} B \text{ for some subcollection } \mathcal{A} \subset \mathcal{B} \right\}, \tag{9}$$

and $\mathcal{B}$ is a basis for $\mathcal{T}$. On the other hand, if $\mathcal{T}$ is a topology on X and $\mathcal{B}$ is a subcollection of $\mathcal{T}$ such that (9) holds, then $\mathcal{B}$ must satisfy properties (1) and (2).

EXERCISE 3.1. Prove Proposition 3.4. The following may be helpful.

- Given $\mathcal{B}$ satisfying the two properties of the Proposition, you can verify that $\mathcal{T}$ as defined in (9) satisfies each of the three points in the definition of a topology.
- In the converse statement, you are given a topology $\mathcal{T}$ and a subcollection $\mathcal{B}$ for which (9) holds. You need to show that $\mathcal{B}$ satisfies (1) and (2) in the Proposition. For (1), start your (short) argument by noting that X must belong to $\mathcal{T}$. For (2), it is helpful to remember that $\mathcal{B} \subset \mathcal{T}$, so $B_1 \cap B_2$ must be an element of $\mathcal{T}$ if B_1 and B_2 are. Argue that if $B_1 \cap B_2 = \bigcup_{B \in \mathcal{A}} B$ for some $\mathcal{A} \subset \mathcal{B}$ and $x \in B_1 \cap B_2$, then x must belong to some $B \in \mathcal{A}$, and this B must be contained in $B_1 \cap B_2$.

REMARK 3.5. If $\mathcal{B}$ is a basis for a topology $\mathcal{T}$ on X, then $\mathcal{T}$ is the 'smallest' topology on X that contains $\mathcal{B}$. After all, as soon as a topology $\mathcal{T}'$ contains $\mathcal{B}$, it contains all unions of elements of $\mathcal{B}$ and therefore all of $\mathcal{T}$. This observation is very helpful in the following common situation: Suppose we are given a topology $\mathcal{T}$ on a set X, generated by a basis $\mathcal{B}$. We would like to know if *another* collection $\mathcal{B}'$ is another basis for $\mathcal{T}$. If $\mathcal{B}'$ satisfies the hypotheses of Proposition 3.4, then it generates some topology $\mathcal{T}'$ on X, which in principle could be different from $\mathcal{T}$, if we don't have further information. However, if we can establish that $\mathcal{B} \subset \mathcal{T}'$ and $\mathcal{B}' \subset \mathcal{T}$, then the argument above will tell us that $\mathcal{T} \subset \mathcal{T}'$ and $\mathcal{T}' \subset \mathcal{T}$; in other words, $\mathcal{T} = \mathcal{T}'$ in this case.

The Remark above establishes the following Proposition:

PROPOSITION 3.6. *Let $\mathcal{T}$ and $\mathcal{T}'$ be topologies on X which are generated by bases $\mathcal{B}$ and $\mathcal{B}'$, respectively. If $\mathcal{B} \subset \mathcal{T}'$ and $\mathcal{B}' \subset \mathcal{T}$, then $\mathcal{T} = \mathcal{T}'$.*

EXERCISE 3.2. Prove that the collection $\mathcal{R}$ of all *open rectangles* of the form

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n), \qquad a_j, b_j \in \mathbb{R},\ a_j < b_j, \text{ for all } j.$$

is a basis for the standard topology on $\mathbb{R}^n$. You are encouraged to use the following (partial) outline:

- Prove that $\mathcal{R}$ satisfies the hypotheses of Proposition 3.4 and is therefore a basis for *some* topology on $\mathbb{R}^n$.

- Show that if $R \in \mathcal{R}$, then R is open with respect to the Euclidean metric.
- Show that if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $r > 0$, then $B(x, r)$ can be written as a union of open rectangles in $\mathcal{R}$. (This is easier than it seems—mimic part of the proof of Theorem 2.12.) Alternatively, use Exercise 2.10.
- Finish by invoking Proposition 3.6.

3.3. Definition of Metrizability. Suppose we are given a topology $\mathcal{T}$ on a set X, which is not *a priori* known to be generated by a metric. A natural question to ask is whether it is possible to *define* a metric that generates $\mathcal{T}$. If the answer is yes, we say that $\mathcal{T}$ is *metrizable*.

DEFINITION 3.7. Let $(X, \mathcal{T})$ be a topological space. If there exists a metric d on X which generates $\mathcal{T}$, then we say that the topology $\mathcal{T}$ is *metrizable*.

EXAMPLE 3.8. The topology $\mathcal{T}_2$ in Example 3.2 is metrizable, since it is the same as the topology generated by the discrete metric. On the other hand, $\mathcal{T}_1$ and $\mathcal{T}_3$ are not metrizable. In fact, there is *no* metrizable topology on any finite nonempty set other than the discrete topology, as we have already seen in Exercise 2.9.

The reader might wonder at this point why we bother to distinguish between the notions of 'metric' and 'metrizable' spaces, since, at the end of the day, they are the same thing. One answer is that the mere *existence* of a metric that generates the topology bestows special properties on that topology. The choice of metric itself may not be nearly as important as the extra structure that it entails. In this case, one can dispense with the metric and simply work with the gifts it has left behind. This will be our viewpoint as we discuss the natural topological structure of the extended real line.

3.4. The Standard Topology of $\overline{\mathbb{R}}$. Let us define the following collections of $\overline{\mathbb{R}}$:

$$\mathcal{B}_1 = \{(a, b) : a, b \in \mathbb{R},\ a < b\}; \qquad \mathcal{B}_2 = \{[-\infty, a) : a \in \mathbb{R}\}; \qquad \mathcal{B}_3 = \{(b, +\infty] : b \in \mathbb{R}\}.$$

Let $\overline{\mathcal{B}} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \mathcal{B}_3$. It should be clear that $\mathcal{B}$ satisfies the hypotheses of Proposition 3.4; therefore it generates a topology $\overline{\mathcal{T}}$ on $\overline{\mathbb{R}}$; we refer to $\overline{\mathcal{T}}$ as the *standard* topology of $\overline{\mathbb{R}}$.

The standard topology $\overline{\mathcal{T}}$ on $\overline{\mathbb{R}}$ is compatible with the standard topology $\mathcal{T}$ on $\mathbb{R}$, in the sense made precise by the following Proposition:

PROPOSITION 3.9. *Let $\mathcal{T}$ and $\overline{\mathcal{T}}$ be as above, and let E be a subset of $\overline{\mathbb{R}}$. Then $E \in \overline{\mathcal{T}}$ if and only if $E \cap \mathbb{R} \in \mathcal{T}$.*

PROOF. The statement can be reframed as saying that

$$\mathcal{T} = \{E \cap \mathbb{R} : E \in \overline{\mathcal{T}}\}.$$

Denote the right side of this equality by $\mathcal{T}'$. We prove the Proposition by establishing that $\mathcal{T} = \mathcal{T}'$.

Suppose $E \in \mathcal{T}$. Then E can be written as a union of the form

$$E = \bigcup_{i \in I} (a_i, b_i)$$

for some index set I and real numbers a_i, b_i, $i \in I$. The set on the right is a subset of $\mathbb{R}$ which is a union of elements of $\mathcal{B}_1$. It follows that $E \in \overline{\mathcal{T}}'$. Thus $\overline{\mathcal{T}} \subset \overline{\mathcal{T}}'$. On the other hand, if $E \in \overline{\mathcal{T}}'$, write

$$E = \bigcup_{i \in I} (a_i, b_i) \cup \bigcup_{j \in J} [-\infty, a_j) \cup \bigcup_{k \in K} (b_k, +\infty],$$

where each a_i, a_j, b_i, b_k is finite. Then

$$E \cap \mathbb{R} = \bigcup_{i \in I} (a_i, b_i) \cup \bigcup_{j \in J} (-\infty, a_j) \cup \bigcup_{k \in K} (b_k, +\infty),$$

and the set on the right is clearly in $\mathcal{T}$. □

Despite the above Proposition, however, it should be clear that the usual Euclidean metric $d(x, y) = |x - y|$ on $\mathbb{R}$ *cannot* be extended to a metric on $\overline{\mathbb{R}}$. (How would we define $d(0, +\infty)$?) On the other hand, there *is* a metric that generates the topology of $\overline{\mathbb{R}}$:

PROPOSITION 3.10. *The standard topology of $\overline{\mathbb{R}}$ is metrizable. A metric that gives rise to the standard topology is the function $\overline{d} : \overline{\mathbb{R}} \times \overline{\mathbb{R}} \to \mathbb{R}$, defined by*

$$\overline{d}(x, y) = |f(x) - f(y)|,$$

where $f : \overline{\mathbb{R}} \to [-1, 1]$ is given by

$$f(x) = \begin{cases} -1, & x = -\infty, \\ \frac{x}{1+|x|}, & x \in \mathbb{R}, \\ 1, & x = +\infty. \end{cases}$$

We do not prove this Proposition, but we give an indication of how it could be done. It is easy to verify that $\overline{d}$ is a metric, directly from the definition. To prove that $\overline{d}$ generates the standard topology on $\overline{\mathbb{R}}$, one can consider the basis $\overline{\mathcal{B}}$ of the standard topology $\overline{\mathcal{T}}$. On can also consider the basis $\mathcal{B}_0$ consisting of open balls with respect to the metric $\overline{d}$, together with the topology $\mathcal{T}_0$ that they generate. Showing that $\mathcal{B}_0 \subset \overline{\mathcal{T}}$ and $\overline{\mathcal{B}} \subset \mathcal{T}_0$ will guarantee us that the statement of the Proposition is true, by Proposition 3.6. We omit the tedious verifications of these inclusions.

Part 2

Analysis on Metric Spaces

CHAPTER 4

Basic Topological Concepts in a Metric Space

1. Fundamental Notions

Recall from the previous Chapter that a subset U of a metric space (X, d) is called *open* if every point of U is an interior point of U with respect to X. That is, if $x \in U$, then $B_X(x, r) \subset U$ for some $r > 0$. We now define and investigate several related concepts.

1.1. Limit Points and Limits of Sequences.

1.1.1. *Definition of a Limit Point.*

DEFINITION 1.1. Let (X, d) be a metric space. If U is an open set of X containing x, we say that U is a *neighborhood* of x in X. The set $B_X(x, \varepsilon)$ is called an *ε-neighborhood* of x in X.

If (X, d) is understood and is the only metric space under consideration, we sometimes refer simply to a 'neighborhood of x', or an 'ε-neighborhood of x', without mention of the underlying space X.

DEFINITION 1.2. Let (X, d) be a metric space; let E be a subset of X. A point x is said to be a *limit point* of E with respect to X if every neighborhood U of x in X intersects $E\backslash\{x\}$. We will denote the set of all limit points of E with respect to X by $\mathrm{Lim}_X(E)$:

$$\mathrm{Lim}_X(E) = \{x \in X : x \text{ is a limit point of } E \text{ with respect to } X\}.$$

A point x is called an *isolated point* of E with respect to X if $x \in E$ and x is not a limit point of E with respect to X.

If (X, d) is understood from context and it is the only metric space under consideration, we sometimes write E' instead of $\mathrm{Lim}_X(E)$.

We make a trivial but useful observation: If $x \notin E$, then $E = E\backslash\{x\}$. So if $x \notin E$, then x is a limit point of E if and only if every neighborhood of x intersects E.

EXAMPLE 1.3. Let A be the subset of $\mathbb{R}$ given by $A = (3, 5) \cup \{9\}$. Then $\mathrm{Lim}_{\mathbb{R}}(A) = [3, 5]$. The point 9 is an isolated point of A. We prove these claims in several steps.

First, 9 is an isolated point of A with respect to $\mathbb{R}$, since $(8, 10)$ is a neighborhood of 9 in $\mathbb{R}$ which does not intersect $A\backslash\{9\}$.

Next, we show that $\mathrm{Lim}_{\mathbb{R}}(A) \subset [3, 5]$ by taking complements. Assume that $x \notin [3, 5]$; we want to show that $x \notin \mathrm{Lim}_{\mathbb{R}}(A)$. We may assume without loss of generality that $x \neq 9$, as we have already shown that 9 is an isolated point (and isolated points cannot be limit points). Then $x \in (-\infty, 3) \cup (5, 9) \cup (9, \infty)$, which is an open set of $\mathbb{R}$ (and thus a neighborhood of x in $\mathbb{R}$) that does not intersect A. Therefore x is not a limit point of A with respect to $\mathbb{R}$. Since x was an arbitrary point of $\mathbb{R}\backslash[3, 5]$, we conclude that $\mathrm{Lim}_{\mathbb{R}}(A) \subset [3, 5]$.

To prove that $\mathrm{Lim}_{\mathbb{R}}(A) \supset [3, 5]$, we argue as follows. Choose $x \in [3, 5]$, and let U be a neighborhood of x. We consider three cases.

- Case 1: $x \in (3, 5)$. Then $U \cap (3, 5)$ is a neighborhood of x, so there exists $r > 0$ such that $B_{\mathbb{R}}(x, r) = (x - r,\ x + r)$ is contained in $U \cap (3, 5)$. Then $U \cap (A\backslash\{x\}) \supset (x, x + r) \ni x + \frac{r}{2}$. So x is a limit point of A with respect to $\mathbb{R}$.
- Case 2: $x = 3$. Then there exists $r \in (0, 2)$ such that $(3 - r,\ 3 + r) \subset U$. Thus $U \cap A \supset (3, 3 + r) \ni 3 + \frac{r}{2}$. Thus 3 is a limit point of A with respect to $\mathbb{R}$.

- Case 3: $x = 5$. Then there exists $r \in (0, 2)$ such that $(5 - r,\ 5 + r) \subset U$. Thus $U \cap A \supset (5 - r, 5) \ni 5 - \frac{r}{2}$. Thus 5 is a limit point of A with respect to $\mathbb{R}$.

One can show in the same manner as above that whenever $a, b \in \mathbb{R}$ and $a < b$, we have

$$\mathrm{Lim}_{\mathbb{R}}((a, b)) = \mathrm{Lim}_{\mathbb{R}}((a, b]) = \mathrm{Lim}_{\mathbb{R}}([a, b)) = \mathrm{Lim}_{\mathbb{R}}([a, b]) = [a, b].$$

Furthermore, for any $c \in \mathbb{R}$, we have

$$\mathrm{Lim}_{\mathbb{R}}((c, +\infty)) = \mathrm{Lim}_{\mathbb{R}}([c, +\infty)) = [c, +\infty), \qquad \mathrm{Lim}_{\mathbb{R}}((-\infty, c)) = \mathrm{Lim}_{\mathbb{R}}((-\infty, c]) = (-\infty, c].$$

1.1.2. *Tools for Computing* $\mathrm{Lim}_X(E)$. The process by which we computed the limit points of a given set in the Example above seems unduly complicated. The goal of this section is to introduce some tools that make the process much less cumbersome in most cases. We begin with the following questions, which the reader may have already asked themself while working the example above.

(1) Can the 'neighborhoods' in the definition of a limit point be replaced with ε-neighborhoods?
(2) Is there a relationship between the interior of a set and the set of its limit points?

The answer to the first question is 'yes'; the answer to the second is (a tiny bit) more subtle but is also 'yes'. We make each of these statements precise below.

PROPOSITION 1.4. *Let E be a subset of X. A point $x \in X$ is a limit point of E with respect to X of E if and only if every ε-neighborhood of x intersects $E\backslash\{x\}$.*

PROOF. If $x \in \mathrm{Lim}_X(E)$, then every neighborhood of x, in particular every ε-neighborhood of x, intersects $E\backslash\{x\}$. For the non-trivial direction, assume every ε-neighborhood of x intersects $E\backslash\{x\}$, and let U be a neighborhood of x in X. Then there exists $r > 0$ such that $B_X(x, r) \subset U$. Therefore $U \cap (E\backslash\{x\}) \supset B_X(x, r) \cap (E\backslash\{x\})$ is nonempty, since $B_X(x, r)$ intersects $E\backslash\{x\}$ by assumption. □

The reader might wonder now why we defined limit points in terms of neighborhoods rather than ε-neighborhoods in the first place; actually, each formulation has its advantage.

Advantages of open set definition of limit point:

- Easier to prove that a given point is *not* a limit point.
- Easier to prove things about points already known to be limit points.
- This definition is valid in a general topological space.

Advantages of ε-neighborhood formulation of a limit point:

- Easier to prove that a point is a limit point in the first place.

The reader may feel (justifiably) that we are belaboring this point. However, this theme of open sets versus open balls will come back again and again. Oftentimes, we will define some concept in terms of general open sets, then prove that a formulation in terms of open balls is equivalent. This way, we can use the advantages of each. For example, we can use the ε-neighborhood formulation to prove that a given point is a limit point of a certain set, then use the open set formulation to prove that the point in question satisfies certain properties.

PROPOSITION 1.5. *Let (X, d) be a metric space, and let E be a subset of X.*

(a) The singleton set $\{x\}$ is open in X if and only if there exists $r > 0$ such that $B_X(x, r) = \{x\}$.
(b) If $x \in \mathrm{Int}_X(E)$ but $\{x\}$ is open in X, then x is an isolated point of E with respect to X.
(c) If $x \in \mathrm{Int}_X(E)$ and $\{x\}$ is not *open in X, then $x \in \mathrm{Lim}_X(E)$.*

PROOF. (a) If $B_X(x, r) = \{x\}$ for some $r > 0$, then clearly $\{x\}$ is open in X. If $\{x\}$ is open in X, then there exists $r > 0$ such that $B_X(x, r) \subset \{x\}$, by definition of an open set. But $\{x\} \subset B_X(x, r)$ always holds. So $\{x\} = B_X(x, r)$ for this choice of $r > 0$.

(b) If $\{x\}$ is open in X, then by the previous statement, there exists $r > 0$ such that $B_X(x, r) = \{x\} \subset E$, which shows that $x \in \mathrm{Int}_X(E)$. On the other hand, $\{x\}$ is a neighborhood of x in X which

does not intersect $E\backslash\{x\}$, so $x \notin \mathrm{Lim}_X(E)$. Since $x \in E$, this means that x is an isolated point of E with respect to X.

(c) Assume that $\{x\}$ is not open in X and $x \in \mathrm{Int}_X(E)$. Choose $r > 0$ so that $B_X(x, r) \subset E$; since $\{x\}$ is not open in X we can find $y \neq x$ in $B_X(x, r)$. This y belongs to $B_X(x, r) \cap (E\backslash\{x\})$. This shows that every ε-neighborhood of x intersects $E\backslash\{x\}$, so that $x \in \mathrm{Lim}_X(E)$. □

COROLLARY 1.6. *Give $\mathbb{R}^k$ the Euclidean metric, and let E be a subset of $\mathbb{R}^k$. Then*

$$\mathrm{Int}_{\mathbb{R}^k}(E) \subset \mathrm{Lim}_{\mathbb{R}^k}(E).$$

PROOF. This statement follows immediately from the previous proposition, together with the fact that $\{x\}$ is not open in $\mathbb{R}^k$ for any $x \in \mathbb{R}^k$. □

Note that in the above Corollary, $\mathbb{R}^k$ can be replaced by any metric space in which singleton sets $\{x\}$ are never open.

REMARK 1.7. The discussion above tells us that $\mathrm{Int}_X(E)$ is 'usually' a subset of $\mathrm{Lim}_X(E)$. However, $\mathrm{Int}_X(E)$ is rarely all of $\mathrm{Lim}_X(E)$. This will be clearer when we talk about closed sets later on.

Here is another straightforward consequence of the definition:

PROPOSITION 1.8. *Let (X, d) be a metric space and assume $E \subset F \subset X$. Then*

$$\mathrm{Lim}_X(E) \subset \mathrm{Lim}_X(F).$$

PROOF. Assume $x \in \mathrm{Lim}_X(E)$, and let U be a neighborhood of x in X. Then $\emptyset \neq U \cap (E\backslash\{x\}) \subset U \cap (F\backslash\{x\})$, so $x \in \mathrm{Lim}_X(F)$. □

The following Exercise (and subsequent discussion) gives a way of computing the limit points of a finite union of sets, provided one knows the limit points of each of the sets over which the union is taken.

EXERCISE 1.1. Let E_1 and E_2 be subsets of a metric space (X, d). Prove that

$$\mathrm{Lim}_X(E_1 \cup E_2) = \mathrm{Lim}_X(E_1) \cup \mathrm{Lim}_X(E_2).$$

With the result of the above exercise in hand, an easy induction argument shows that

$$\mathrm{Lim}_X\left(\bigcup_{j=1}^{n} E_j\right) = \bigcup_{j=1}^{n} \mathrm{Lim}_X(E_j),$$

for any finite collection $\{E_j\}_{j=1}^n$ of subsets of X. However, this does *not* extend to infinite collections. Indeed,

$$\mathrm{Lim}_{\mathbb{R}}\left(\bigcup_{n=1}^{\infty}[\tfrac{1}{n}, 1]\right) = \mathrm{Lim}_{\mathbb{R}}((0, 1]) = [0, 1];$$

$$\bigcup_{n=1}^{\infty} \mathrm{Lim}_{\mathbb{R}}\left([\tfrac{1}{n}, 1]\right) = \bigcup_{n=1}^{\infty}[\tfrac{1}{n}, 1] = (0, 1].$$

Finally, whether a point x is a limit point of a given set E depends on the underlying metric space. See the Exercise below.

EXERCISE 1.2. Let (X, d) be a metric space, and assume $E \subset Y \subset X$. Prove that

$$\mathrm{Lim}_Y(E) = \mathrm{Lim}_X(E) \cap Y.$$

Hint: Use Theorem 2.13 in Chapter 3.

EXERCISE 1.3. If $(X, \mathcal{T})$ is a topological space and E is a subset of X, we say that x is a limit point of E with respect to X if every neighborhood of x in X (that is, every $U \in \mathcal{T}$ such that $U \ni x$) intersects $E\backslash\{x\}$.

(a) Suppose $\mathcal{B}$ is a basis for the topology $\mathcal{T}$ on X. Show that x is a limit point of E with respect to X if and only if every $B \in \mathcal{B}$ containing x intersects $E\backslash\{x\}$.
(b) Show that if E is any subset of $\mathbb{R}$ which is not bounded above (with respect to the usual order relation on $\mathbb{R}$), then $+\infty$ is a limit point of E with respect to $\overline{\mathbb{R}}$ (in its standard topology).

1.1.3. *Sequences and Limit Points.*

DEFINITION 1.9. Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. We say that $(x_n)_{n=1}^{\infty}$ *converges* in X to a point $x \in X$ if for every neighborhood U of x in X, there exists $N \in \mathbb{N}$ such that $n \geq N$ implies that $x_n \in U$. A sequence that converges is called *convergent*. If a sequence does not converge, we say that it *diverges*, or is *divergent*.

REMARK 1.10. The following are all standard ways to write '$(x_n)_{n=1}^{\infty}$ converges in X to $x \in X$', provided (X, d) is understood from context.

- $\lim_{n\to\infty} x_n = x$.
- $x_n \to x$ as $n \to \infty$.

If $(s_n)_{n=1}^{\infty}$ is a sequence of real numbers, one often write '$s_n \to s$ as $n \to \infty$', without even mentioning $\mathbb{R}$ or the Euclidean metric.

Whether or not a given sequence converges can depend on the metric space in which the sequence lives. However, if (X, d) is a metric space and Y is a subset of X, considered (as usual) with the same metric, then the relationship between convergence in X and convergence in Y is rather straightforward, as demonstrated by the following Exercise.

EXERCISE 1.4. Let (X, d) be a metric space, and assume $Y \subset X$. Let $(x_n)_{n=1}^{\infty}$ be a sequence in Y and let x be a point of X. Prove that the following two statements are equivalent:

(1) $x_n \to x$ in X, and $x \in Y$.
(2) $x_n \to x$ in Y.

Hint: Use Theorem 2.13 in Chapter 3.

EXERCISE 1.5. Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. Prove that the following statements are equivalent:

(1) $x_n \to x$ in X.
(2) For every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $n \geq N$ implies $x_n \in B_X(x, \varepsilon)$ (i.e. $d(x, x_n) < \varepsilon$).
(3) $d(x, x_n) \to 0$ as $n \to \infty$

EXERCISE 1.6. Let (X, d) be a metric space; let $(x_n)_{n=1}^{\infty}$ be a sequence in X, and let $(t_n)_{n=1}^{\infty}$ be a sequence of positive real numbers. Assume that $t_n \to 0$ as $n \to \infty$.

(a) Prove that if there exists $x \in X$ and $N \in \mathbb{N}$ such that $d(x_n, x) \leq t_n$ for all $n \geq N$, then $x_n \to x$ in X as $n \to \infty$.
(b) (For later use) Prove that $1/n \to 0$ as $n \to \infty$.

EXERCISE 1.7. Let $(\mathbf{x}_n)_{n=1}^{\infty}$ be a sequence in $\mathbb{R}^k$ under the Euclidean metric; for each $n \in \mathbb{N}$, denote $\mathbf{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,k})$. Prove that $(\mathbf{x}_n)_{n=1}^{\infty}$ converges to $\mathbf{x} = (x_1, \ldots, x_k)$ if and only if $(x_{n,j})_{n=1}^{\infty}$ converges to x_j for each $j \in \{1, \ldots, k\}$. (Using the square metric might be helpful here.)

The next Theorem spells out an important relationship between limits of sequences and limit points in a metric space.

THEOREM 1.11. *Let (X, d) be a metric space; let E be a subset of X. The following are equivalent:*

(1) x is a limit point of E with respect to X.
(2) Every neighborhood of x in X contains infinitely many points of E.
(3) There exists a sequence $(x_n)_{n=1}^{\infty}$ in $E\backslash\{x\}$ that converges in X to x.

PROOF. (1) $\implies$ (2) We prove the contrapositive: Let U be an open set of X that contains x, and assume that U contains only finitely many points of E; then we can write $U \cap (E\backslash\{x\}) = \{x_i\}_{i=1}^n$ for some $n \in \mathbb{N}$. Pick $r > 0$ such that $B(x,r) \subset U$, and put $\varepsilon = \min\{r, d(x,x_1), \ldots, d(x,x_n)\}$. Then $B_X(x,\varepsilon)$ is an open set of X, and

$$B_X(x,\varepsilon) \cap (E\backslash\{x\}) \subset U \cap (E\backslash\{x\}) = \{x_i\}_{i=1}^n.$$

On the other hand, $B_X(x,\varepsilon)$ contains none of the x_i, since $d(x,x_i) \geq \varepsilon$ by definition of ε. It follows that $B_X(x,\varepsilon) \cap (E\backslash\{x\})$ is empty, so x is not a limit point of E. This establishes that (1) $\implies$ (2).

(2) $\implies$ (1) follows directly from the definition of a limit point. Indeed, if U is an open set of X that contains x, and (2) holds, then U contains infinitely many points of E, therefore infinitely many points of $E\backslash\{x\}$, and certainly at least one such point, which implies that x is a limit point of E with respect to X.

(1) $\implies$ (3) Assume x is a limit point of E with respect to X. For each $n \in \mathbb{N}$, let x_n be a point of $B_X(x, \frac{1}{n}) \cap (E\backslash\{x\})$. (This is possible since x is a limit point of E.) Then the sequence $(x_n)_{n=1}^\infty$ is a sequence in $E\backslash\{x\}$; this sequence converges in X to x, since $d(x,x_n) < \frac{1}{n}$ for each $n \in \mathbb{N}$.

(3) $\implies$ (1) Assume $(x_n)_{n=1}^\infty$ is a sequence in $E\backslash\{x\}$ such that $x_n \to x$ as $n \to \infty$. Choose $\varepsilon > 0$, and then choose n large enough so that $d(x,x_n) < \varepsilon$. Then $x_n \in B_X(x,\varepsilon) \cap (E\backslash\{x\})$. Thus x is a limit point of E with respect to X. □

Note that statement (2) in the Proposition above requires that E has infinitely many points to begin with. Therefore we have the following Corollary:

COROLLARY 1.12. *Let (X,d) be a metric space, and let E be a finite subset of X. Then E has no limit points with respect to X.*

1.2. Closed Sets and the Closure.

1.2.1. *Definition of a Closed Set.*

DEFINITION 1.13. Let (X,d) be a metric space, and let F be a subset of X. We say that F is *closed* in X if $X\backslash F$ is open in X.

Important! 'Closed' is *not* the same thing as 'not open'! A set can be open, closed, neither, or both. For example,

- (a,b) is open in $\mathbb{R}$ but not closed in $\mathbb{R}$;
- $[a,b]$ is closed in $\mathbb{R}$ but not open in $\mathbb{R}$;
- $\mathbb{R}$ is both open and closed in $\mathbb{R}$;
- $(a,b]$ is neither open nor closed in $\mathbb{R}$.

PROPOSITION 1.14. *Let (X,d) be a metric space. Any intersection of closed subsets of X is a closed set set of X. Any finite union of closed subsets of X is a closed set of X.*

PROOF. Combine Proposition 2.11 in Chapter 3 with DeMorgan's Laws. □

PROPOSITION 1.15. *Let (X,d) be a metric space, and let E be a subset of X. Then E is closed in X if and only if it contains all its limit points, i.e.,* $\mathrm{Lim}_X(E) \subset E$.

PROOF. ($\implies$) Assume that E is closed, so that $X\backslash E$ is open. No point x in $X\backslash E$ can be a limit point of E, since $X\backslash E$ itself serves as a neighborhood of x that does not intersect E. In other words $X\backslash E \subset X\backslash \mathrm{Lim}_X(E)$, so that—taking complements, we conclude that $\mathrm{Lim}_X(E) \subset E$, as needed.

($\impliedby$) On the other hand, assume that $\mathrm{Lim}_X(E) \subset E$; we show that $X\backslash E$ is open in X. To this end, let x be a point of $X\backslash E$. Then $x \notin \mathrm{Lim}_X(E)$, so there is an open set U of X such that U does not intersect $E\backslash\{x\} = E$, i.e. $x \in U \subset X\backslash E$. Since U is open, it follows that $x \in \mathrm{Int}_X(U) \subset \mathrm{Int}_X(X\backslash E)$. Since x was an arbitrary point of $X\backslash E$, we conclude that every point of $X\backslash E$ is an interior point; that is, $X\backslash E$ is open. Therefore E is closed. □

EXERCISE 1.8. Let (X, d) be a metric space, and let E be a subset of X. Prove that $\mathrm{Lim}_X(E)$ is a closed set of X.

1.2.2. *The Closure of a Set.*

DEFINITION 1.16. Let (X, d) be a metric space and let E be a subset of X. The *closure* of E in X is the set E, together with the set of its limit points in X adjoined, denoted $\mathrm{Cl}_X(E)$. That is,

$$\mathrm{Cl}_X(E) = E \cup \mathrm{Lim}_X(E).$$

If (X, d) is clear from context and it is the only metric space under consideration, we may write $\overline{E}$ instead of $\mathrm{Cl}_X(E)$.

PROPOSITION 1.17. *Let (X, d) be a metric space and let E be a subset of X.*

(a) E is closed in X if and only if $E = \mathrm{Cl}_X(E)$.
(b) $\mathrm{Cl}_X(E)$ is always a closed set.
(c) If F is a closed set of X and $E \subset F$, then $\mathrm{Cl}_X(E) \subset F$.
(d) Letting $\mathcal{F}$ denote the collection of all closed sets F in X such that F contains E, we have

$$\overline{E} = \bigcap_{F \in \mathcal{F}} F.$$

In other words, $\overline{E}$ is the 'smallest' closed set that contains E.

PROOF. (a) Assume E is closed in X. Then $\mathrm{Lim}_X(E) \subset E$, so $\mathrm{Cl}_X(E) = E \cup \mathrm{Lim}_X(E) \subset E$. But $E \subset \mathrm{Cl}_X(E)$ by definition, so $E = \mathrm{Cl}_X(E)$. On the other hand, assume $E = \mathrm{Cl}_X(E)$. Then $\mathrm{Lim}_X(E) = E \cup \mathrm{Lim}_X(E) = \mathrm{Cl}_X(E) = E$, so E is closed.

(b) To show that $\mathrm{Cl}_X(E)$ is closed in X, we write

$$\mathrm{Lim}_X(E \cup \mathrm{Lim}_X(E)) = \mathrm{Lim}_X(E) \cup \mathrm{Lim}_X(\mathrm{Lim}_X(E)) \subset \mathrm{Lim}_X(E) \subset \mathrm{Cl}_X(E).$$

To get the first inclusion, we used the fact that $\mathrm{Lim}_X(E)$ is closed in X, by Exercise 1.8.

(c) Finally, let F be a closed set of X such that $E \subset F$. Then $\mathrm{Lim}_X(E) \subset \mathrm{Lim}_X(F)$, so

$$\mathrm{Cl}_X(E) = E \cup \mathrm{Lim}_X(E) \subset F \cup \mathrm{Lim}_X(F) = \mathrm{Cl}_X(F) = F.$$

(d) This follows from (b) and (c). The short proof is similar to that of Proposition 2.14. (Alternatively, one can see it as a consequence of Proposition 2.14 in Chapter 3, together with Exercise 1.9 below.) □

Note that in the course of proving the third point above, we have shown that if $E \subset F$, then $\overline{E} \subset \overline{F}$.

REMARK 1.18. A point x belongs to $\mathrm{Cl}_X(E)$ if and only if every neighborhood U of x in X intersects E. Here is a proof: If every neighborhood of x in X intersects E, then either $x \in E$ or every neighborhood intersects $E \backslash \{x\}$, in which case $x \in \mathrm{Lim}_X(E)$. Conversely, if $x \in \mathrm{Cl}_X(E)$, then either $x \in E$, in which case every neighborhood of x intersects E at x, or $x \in \mathrm{Lim}_X(E)$, in which case every neighborhood intersects the smaller set $E \backslash \{x\}$, and therefore E. As is often the case, neighborhoods may be replaced by ε-neighborhoods in this criterion.

EXERCISE 1.9. Let (X, d) be a metric space, and let E be a subset of X. Prove that

$$X \backslash \mathrm{Cl}_X(E) = \mathrm{Int}_X(X \backslash E)$$

(This can be written more concisely as $(\overline{E})^c = (E^c)^\circ$, if desired.

EXERCISE 1.10. Let (X, d) be a metric space. Let E and Y be subsets of X such that $E \subset Y$. Prove that

$$\mathrm{Cl}_Y(E) = \mathrm{Cl}_X(E) \cap Y.$$

EXERCISE 1.11. Let (X, d) be a metric space.

(a) Prove that for any collection $\mathcal{E}$ of subsets of X, we have

$$\bigcup_{E\in\mathcal{E}} \overline{E} \subset \overline{\bigcup_{E\in\mathcal{E}} E},$$

and equality holds if $\mathcal{E}$ is finite.

(b) Prove that for any collection $\mathcal{E}$ of subsets of X, we have

$$\bigcap_{E\in\mathcal{E}} \overline{E} \supset \overline{\bigcap_{E\in\mathcal{E}} E}.$$

(c) Give examples that demonstrate that equality might fail in part (a) if $\mathcal{E}$ is not finite, and equality might fail in part (b) even if $\mathcal{E}$ is finite.

1.2.3. *Density.*

DEFINITION 1.19. Let (X, d) be a metric space. A subset E of X is said to be *dense* in X if $\mathrm{Cl}_X(E) = X$.

EXERCISE 1.12. Let (X, d) be a metric space, and let E be a subset of X.

(a) Show that E is dense in X if and only if any nonempty open subset of X contains a point of E.

(b) Suppose $E \subset Y \subset X$. Prove that E is dense in Y if and only if $\mathrm{Cl}_X(E) \supset Y$.

EXERCISE 1.13. Previously, we said that a subset E of $\mathbb{R}$ was dense in $\mathbb{R}$ if for any real numbers a and b, there exists a number $c \in E$ which lies between a and b. Show that in $\mathbb{R}$, the new, more general definition of 'dense' agrees with the old one. That is, show that a subset E of $\mathbb{R}$ is dense in $\mathbb{R}$ according to the new definition if and only if it is dense according to the old one. (Hint: Use Exercise 1.12(a).)

1.3. Boundedness.

DEFINITION 1.20. Let (X, d) be a metric space. A subset E of X is called *bounded* in (X, d) if there exists $x \in X$ and $R > 0$ such that $E \subset B_{(X,d)}(x, R)$. If Z is any set and $f : Z \to X$ is a function, we say that f is *bounded* if its image $f(Z)$ is bounded in (X, d).

Note that on an ordered set equipped with a metric, 'bounded with respect to the order relation' technically means something different than 'bounded with respect to the metric.' In most cases that are useful in practice, the two notions agree.

EXERCISE 1.14. Let (X, d) be a metric space.

(a) Prove that for any $x \in X$ and $r > 0$, we have $\overline{B_X(x, r)} \subset \{y \in X : d(x, y) \leq r\}$. (Hint: Take complements and draw a picture.) Note that the inclusion $\overline{B_X(x, r)} \subset B_X(x, r + \varepsilon)$ follows for any $\varepsilon > 0$.

(b) Give an example using the discrete metric that demonstrates that equality need not hold in the inclusion $\overline{B_X(x, r)} \subset \{y \in X : d(x, y) \leq r\}$ that you proved in part (a).

(c) Prove that in $\mathbb{R}^n$ under the Euclidean metric $d(x, y) = \|x - y\|$, we have $\overline{B_{\mathbb{R}^n}(x, r)} = \{y \in \mathbb{R}^n : \|x - y\| \leq r\}$. (Again, a picture may be useful.)

(d) Using part (a), prove that if A is bounded in (X, d), then $\overline{A}$ is also bounded in (X, d).

2. Sequences in Metric Spaces

Consider the following sequence in $\mathbb{R}$:

$$(x_n)_{n=1}^{\infty} = (3, 1, 3, \tfrac{1}{2}, 3, \tfrac{1}{3}, 3, \tfrac{1}{4}, 3, \tfrac{1}{5}, 3, \tfrac{1}{6}, \ldots)$$

The sequence $(x_n)_{n=1}^{\infty}$ diverges. However, its subsequences $(3, 3, 3, 3, 3, \ldots)$ and $(\frac{1}{n})_{n=1}^{\infty}$ do converge, to 3 and 0, respectively. We would like a concept that captures this information; the limit point notion is

insufficient. Indeed, one can show that $\mathrm{Lim}_{\mathbb{R}}(\{x_n\}_{n=1}^{\infty}) = \{0\}$; the point 3 is an isolated point of the image of the sequence $(x_n)_{n=1}^{\infty}$. The notion of a subsequential limit is the concept we are looking for.

DEFINITION 2.1. Let $(x_n)_{n=1}^{\infty}$ be a sequence in a metric space (X, d), and let $(x_{n_k})_{k=1}^{\infty}$ be a subsequence of $(x_n)_{n=1}^{\infty}$. If $x_{n_k} \to x$ as $k \to \infty$, then x is called a *subsequential limit* of the original sequence $(x_n)_{n=1}^{\infty}$.

In this section, after establishing some basic properties of sequences and convergence, we will show that a subsequential limit of a sequence $(x_n)_{n=1}^{\infty}$ is either a limit point of $\{x_n\}_{n=1}^{\infty}$ or a term of the sequence that appears infinitely many times. Next, we will discuss *Cauchy sequences*, which act like convergent sequences in many ways but may or may not have a limit.

2.1. Convergent Sequences. Recall that we say that a function is bounded if its image is bounded. Since a sequence is in particular a function, this terminology applies here as well.

PROPOSITION 2.2. *(Convergent sequences are bounded.) Let $(x_n)_{n=1}^{\infty}$ be a sequence in a metric space (X, d). If $(x_n)_{n=1}^{\infty}$ converges in X, then it is bounded.*

PROOF. Assume $x_n \to x$ as $n \to \infty$. Choose $N \in \mathbb{N}$ such that $n \geq N$ implies $x_n \in B_X(x, 1)$. Then put $R = \max\{1, d(x, x_1), \ldots, d(x, x_N)\}$. Then $\{x_n\}_{n=1}^{\infty} \subset B_X(x, R)$. Thus $\{x_n\}_{n=1}^{\infty}$ is bounded. □

PROPOSITION 2.3. *Let $S = (p_n)_{n=1}^{\infty}$ be a sequence in a metric space (X, d), and assume that $p_n \to p$ as $n \to \infty$. The the following statements hold.*

- $(\mathrm{Im}\, S)' \subset \{p\}$.
- $(\mathrm{Im}\, S)' = \emptyset$ *if and only if* $p_n = p$ *for all but finitely many* $n \in \mathbb{N}$.

PROOF. Assume $q \neq p$; we show that q is not a limit point of $\mathrm{Im}\, S = \{p_n\}_{n=1}^{\infty}$. Put $r = \frac{1}{2}d(p, q)$, so that $B_X(p, r)$ and $B_X(q, r)$ are disjoint[1]. Choose N large enough so that $n \geq N$ implies that $p_n \in B_X(p, r)$. Then $B_X(q, r)$ contains at most finitely many points of $\mathrm{Im}\, S$, so q cannot be a limit point of $\mathrm{Im}\, S$.

If $(\mathrm{Im}\, S)' = \emptyset$, then (since p is not a limit point of $\mathrm{Im}\, S$) there exists a neighborhood U of p such that U contains no points of $\mathrm{Im}\, S$ other than p. On the other hand, U must contain all but finitely many of the p_n, by definition of the limit. These p_n must therefore all be equal to p.

Conversely, if $p_n = p$ for all but finitely many n, then $\mathrm{Im}\, S$ is actually a finite set; therefore it has no limit points. □

This Proposition says that for a convergent sequence $(p_n)_{n=1}^{\infty}$, there are only two possibilities for $(\mathrm{Im}\, S)'$. Either $(\mathrm{Im}\, S)'$ is a single point—the point p to which the sequence converges, or $(\mathrm{Im}\, S)'$ is empty, in which case the 'tail' of the sequence looks like $(p, p, p, p, \ldots)$.

COROLLARY 2.4. *(Limits are unique) Let $(p_n)_{n=1}^{\infty}$ be a sequence in a metric space (X, d). If $p_n \to p$ and $p_n \to q$ as $n \to \infty$, then $p = q$.*

PROOF. If $p \neq q$, then $(\mathrm{Im}\, S)' \subset \{p\} \cap \{q\} = \emptyset$. Therefore all but finitely many of the p_n must be equal to p, and all but finitely many of the p_n must be equal to q. This is impossible. □

2.2. Subsequential Limits of a Sequence. We now expand our view to include sequences that may or may not converge.

PROPOSITION 2.5. *Let $(p_n)_{n=1}^{\infty}$ be a sequence in a metric space (X, d). If $p \in (\{p_n\}_{n=1}^{\infty})'$, then there exists a subsequence $(p_{n_k})_{k=1}^{\infty}$ of (p_n) such that $p_{n_k} \to p$ as $k \to \infty$.*

[1]Proof: If $z \in B_X(x, r) \cap B_X(y, r)$, then $d(x, y) \leq d(x, z) + d(z, y) < r + r = 2r = d(x, y)$, a contradiction.

PROOF. We construct a subsequence $(p_{n_k})_{k=1}^{\infty}$ by induction. Choose $n_1 \in \mathbb{N}$ such that $p_{n_1} \in B_X(p,1)$. Assume that $n_1 < n_2 < \cdots < n_k$ are given. We claim that there exists $n_{k+1} \in \mathbb{N}$ such that $n_{k+1} > n_k$ and $p_{n_{k+1}} \in B_X(p, \frac{1}{k+1})$. Indeed, since p is a limit point of $\operatorname{Im} S$, we know that $B_X(x, \frac{1}{k+1})$ contains infinitely many points of $\operatorname{Im} S = \{p_n\}_{n=1}^{\infty}$, so it must contain infinitely many points of $\{p_n\}_{n=n_k+1}^{\infty}$. The subsequence we have constructed converges to p, since $d(p_{n_k}, p) < \frac{1}{k}$ for each $k \in \mathbb{N}$. □

The reader might wonder whether or not this proof was really necessary in light of Theorem 1.11, which guarantees under our hypotheses that there exists a sequence $(x_n)_{n=1}^{\infty}$ in $\{p_n\}_{n=1}^{\infty} \backslash \{p\}$ that converges to p. However, the order of the terms of the sequence $(x_n)_{n=1}^{\infty}$ that this Theorem provides might be completely different than the order of the terms of the sequence $(p_n)_{n=1}^{\infty}$. Therefore the sequence provided by the Theorem might not be a subsequence of $(p_n)_{n=1}^{\infty}$. This is why we paid special attention to the order of the terms when writing the proof of the Proposition.

THEOREM 2.6. *Let $S = (p_n)_{n=1}^{\infty}$ be a sequence in a metric space (X,d). Let S^* denote the set of all subsequential limits of S, and let S_∞ denote the set of all terms of S that appear infinitely many times in the sequence (i.e., $S_\infty = \{p \in X : p_n = p \text{ for infinitely many } n \in \mathbb{N}\}$). Then*

$$S^* = (\operatorname{Im} S)' \cup S_\infty,$$

and S^* is closed in X.

PROOF. We have essentially already proved the inclusion $(\operatorname{Im} S)' \cup S_\infty \subset S^*$. Indeed, Proposition 2.5 implies that $(\operatorname{Im} S)' \subset S^*$, and the inclusion $S_\infty \subset S^*$ is clear[2]. Thus $S^* \supset (\operatorname{Im} S) \cup S_\infty$.

To prove the opposite inclusion, choose $p \in S^*$. If $p \in (\operatorname{Im} S)'$, then we are done; therefore assume without loss of generality that $p \notin (\operatorname{Im} S)'$. It suffices to show, then, that $p \in S_\infty$. Since $p \in S^*$, we can find a subsequence $(p_{n_k})_{k=1}^{\infty}$ of $(p_n)_{n=1}^{\infty}$ that converges to p in X. Then $(\{p_{n_k}\}_{k=1}^{\infty})' \subset (\operatorname{Im} S)'$ (simply because $(p_{n_k})_{k=1}^{\infty}$ is a subsequence of $(p_n)_{n=1}^{\infty}$), and $(\{p_{n_k}\}_{k=1}^{\infty})' \subset \{p\}$ (by Proposition 2.5), so $(\{p_{n_k}\}_{k=1}^{\infty})' \subset (\operatorname{Im} S)' \cap \{p\}$. But $p \notin (\operatorname{Im} S)'$ by assumption, so $(\{p_{n_k}\}_{k=1}^{\infty})' = \emptyset$. Thus, by Proposition 2.3, it follows that $p_{n_k} = p$ for all but finitely many k, whence $p_n = p$ for infinitely many n, whence $p \in S_\infty$. This completes the proof of the equality $S^* = (\operatorname{Im} S)' \cup S_\infty$.

To prove that S^* is closed in X, we look at $(S^*)'$.

$$\begin{aligned}
(S^*)' &= ((\operatorname{Im} S)' \cup S_\infty)' && \text{Definition of } S^* \\
&= ((\operatorname{Im} S)')' \cup S_\infty' && \text{Exercise 1.1} \\
&\subset (\operatorname{Im} S)' && \text{Exercise 1.8, plus } S_\infty \subset \operatorname{Im} S \\
&\subset S^*.
\end{aligned}$$

This completes the proof. □

Note that as a trivial Corollary of the above Proposition, we have $S^* \subset \overline{\operatorname{Im} S}$. This fact is sometimes useful when determining S^*.

EXERCISE 2.1. Let $S = (p_n)_{n=1}^{\infty}$ be a sequence in $\mathbb{R}$ whose image is $(\mathbb{Q} \cap (0,1)) \cup \{5\}$. What are the two possibilities for S^*? Justify your answer.

PROPOSITION 2.7. *Let $S = (p_n)_{n=1}^{\infty}$ be a sequence in a metric space X, and let S^* denote the set of subsequential limits of S. The following are equivalent.*

(1) $p_n \to p$ in X as $n \to \infty$.
(2) Every subsequence of $(p_n)_{n=1}^{\infty}$ converges to p in X.

[2] In case the reader disagrees, here is a proof: If $p \in S_\infty$, then $p_n = p$ for infinitely many n. Let n_1 be the smallest integer such that $p_{n_1} = p$. Having chosen $n_1 < n_2 < \cdots < n_k$, let n_{k+1} be the smallest integer larger than n_k such that $p_{n_{k+1}} = p$. Then $(p_{n_k})_{k=1}^{\infty} = (p, p, p, p, \ldots)$ clearly converges to p in X.

(3) $S^* = \{p\}$*, and every subsequence of* $(p_n)_{n=1}^{\infty}$ *converges in* X*.*

EXERCISE 2.2. Prove Proposition 2.7.

REMARK 2.8. It is easy to find examples of sequences of (for example) real numbers which have *no* subsequential limits; the sequence $(n)_{n=1}^{\infty}$ is a simple one. However, for certain metric spaces X, the set S^* is nonempty for *every* sequence S in X. We call such metric spaces *sequentially compact*. We will study such compact metric spaces starting in Section 3. Eventually, we will see that while $\mathbb{R}$ is not sequentially compact, any closed interval $[a, b]$ in $\overline{\mathbb{R}}$ is (including the cases where a and b are infinite). We delay this discussion slightly in favor of introducing the concept of a *Cauchy sequence*, which is closely connected to both the discussion above and to a characterization of sequentially compact metric spaces.

2.3. Cauchy Sequences.

DEFINITION 2.9. Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. We say that $(x_n)_{n=1}^{\infty}$ is a *Cauchy sequence* in X (or simply *Cauchy* in X) if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $m \geq n \geq N$ implies $d(x_n, x_m) < \varepsilon$. We say that (X, d) is *complete* if every Cauchy sequence in X converges in X.

Note that the key difference between a convergent sequence and a Cauchy sequence is that there is no mention of a limit in the definition of a Cauchy sequence. However, a Cauchy sequence behaves rather like a convergent sequence, in that it 'wants' to converge somewhere. If the point to which it 'wants' to converge lies in the metric space in question, then it does converge; however, this need not be the case. Here are two illustrative examples.

EXAMPLE 2.10. Consider the sequence $S = (\frac{1}{n})_{n=1}^{\infty}$ in $\mathbb{R}$ in the Euclidean metric. Let $Y = (0, 2)$ in the same metric. Then S converges in $\mathbb{R}$, to 0, but S does not converge in Y. Indeed, if S converges to y in Y, then S also converges to y in X. But $0 \notin Y$, so $y \neq 0$. This contradicts the uniqueness of limits of a convergent sequence.

However, S *is* Cauchy in Y. Indeed, choose $\varepsilon > 0$. Choose $N \in \mathbb{N}$ such that $N > \frac{2}{\varepsilon}$. Then $m \geq n \geq N$ implies that $\left|\frac{1}{m} - \frac{1}{n}\right| \leq \frac{1}{m} + \frac{1}{n} \leq \frac{2}{N} < \varepsilon$. Thus S is Cauchy in Y.

This example also shows that $Y = (0, 2)$ is not complete (and also not sequentially compact).

Note that if a sequence $(x_n)_{n=1}^{\infty}$ in (X, d) is Cauchy in X, then it is also Cauchy in Y whenever $\{x_n\}_{n=1}^{\infty} \subset Y \subset X$. Similarly, if $(x_n)_{n=1}^{\infty}$ is a sequence in X which is Cauchy in Y, then it is also Cauchy in X. These claims follow immediately from the definition of a Cauchy sequence.

EXERCISE 2.3. Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. Prove the following statements.

(a) If $(x_n)_{n=1}^{\infty}$ converges in X, then it is Cauchy in X.
(b) If $(x_n)_{n=1}^{\infty}$ is Cauchy in X, then it is bounded in X.

In light of part (a) of this Exercise, we see another way to conclude that the sequence $S = (\frac{1}{n})_{n=1}^{\infty}$ is Cauchy in $Y = (0, 2)$ under the Euclidean metric. We know that S converges in $X = \mathbb{R}$, so it is Cauchy in $\mathbb{R}$, so it is Cauchy in Y.

EXAMPLE 2.11. Let $X = \mathbb{R}$, and let d be the Euclidean metric. Consider the sequence $S = (1, 1.4, 1.41, 1.414, \ldots)$ which converges to $\sqrt{2}$ in $\mathbb{R}$. Then, reasoning as before, we see that S converges in $\mathbb{R}$, but S does not converge in $\mathbb{Q}$ (though it is Cauchy in $\mathbb{Q}$).

The next Theorem generalizes the examples above.

THEOREM 2.12. *Let* (X, d) *be a metric space; let* Y *be a subset of* X*. The following statements hold.*

(a) If Y is complete, then Y is closed in X.
(b) If X is complete and Y is closed in X, then Y is complete.

EXERCISE 2.4. Prove Theorem 2.12. (Hints: For (a), use Theorem 1.11 and Proposition 1.15. For part (b), Exercise 1.4 is relevant.)

PROPOSITION 2.13. *Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a Cauchy sequence in X. If $(x_n)_{n=1}^{\infty}$ has a subsequence that converges in X to a point $x \in X$, then $(x_n)_{n=1}^{\infty}$ converges in X to x.*

PROOF. Choose $\varepsilon > 0$. Since (x_n) is Cauchy, we can pick $N \in \mathbb{N}$ large enough so that $m \geq n \geq N$ implies $d(x_m, x_n) < \varepsilon/2$. Next, let $(x_{n_k})_{k=1}^{\infty}$ be a subsequence of $(x_n)_{n=1}^{\infty}$ such that $x_{n_k} \to x$ in X; pick $M \in \mathbb{N}$ large enough so that $k \geq M$ implies $d(x, x_{n_k}) < \varepsilon/2$. Put $P = \max\{N, M\}$. We claim that $m \geq P$ implies $d(x, x_m) < \varepsilon$. (This will prove $x_n \to x$ in X.) In order to take advantage of both the fact that (x_n) is Cauchy and that $x_{n_k} \to x$, we pick $k \geq m$ and apply the triangle inequality:

$$d(x, x_m) \leq d(x, x_{n_k}) + d(x_{n_k}, x_m).$$

Since $n_k \geq M$, we have $d(x, x_{n_k}) < \frac{\varepsilon}{2}$; since $n_k \geq m \geq N$, we have $d(x_{n_k}, x_m) < \frac{\varepsilon}{2}$. Thus $d(x, x_m) < \varepsilon$, and we are done. □

COROLLARY 2.14. *Let (X, d) be a metric space. Let S be a Cauchy sequence in X and let S^* denote its set of subsequential limits in X. Exactly one of the following two scenarios holds.*

- *Either S converges to a point x of X, in which case $S^* = \{x\}$, or*
- *S diverges, and $S^* = \emptyset$.*

In particular, a Cauchy sequence has at most one subsequential limit.

3. Sequential Compactness

We now 'officially' introduce the notion of sequential compactness, which was previewed earlier in Remark 2.8.

DEFINITION 3.1. A metric space (X, d) is called *sequentially compact* if every sequence in X has a convergent subsequence.

In the notation of Theorem 2.6, a metric space (X, d) is sequentially compact if S^* is never empty, for any sequence S in X. As an immediate Corollary of Proposition 2.13 in the previous section, we can easily show that such spaces are *complete*:

COROLLARY 3.2. *A sequentially compact metric space is complete.*

PROOF. Let $(x_n)_{n=1}^{\infty}$ be a Cauchy sequence in a sequentially compact metric space (X, d). By sequential compactness, $(x_n)_{n=1}^{\infty}$ has a subsequence which converges in X; let x denote the subsequential limit. By Proposition 2.13, it follows that $x_n \to x$ in X. Thus X is complete. □

In fact, we can be much more precise about this relationship, as we will spell out in the next subsection.

EXERCISE 3.1. We say that a metric space (X, d) is *limit point compact* if every infinite subset of X has a limit point in X. Prove that a metric space (X, d) is limit point compact if and only if it is sequentially compact.

(The reason that 'limit point compact' is dignified with its own terminology is that it is *not* necessarily equivalent to 'sequentially compact' in a general topological space.)

3.1. Totally Bounded Metric Spaces. There is a gap in between complete metric spaces and sequentially compact ones. Basically, the issue is that completeness of (X, d) alone guarantees only that *if* a sequence has a Cauchy subsequence, then the point to which the subsequence 'wants' to converge will actually belong to X (as opposed to some possibly larger metric space), furnishing a subsequential limit. On the other hand, there may be sequences which are too 'spread out' and do not contain any Cauchy subsequences. The following Lemma gets at the main issue:

LEMMA 3.3. *Let (X, d) be a metric space, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. If there exists $\varepsilon > 0$ such that $d(x_n, x_m) \geq \varepsilon$ for all $n, m \in \mathbb{N}$, then $(x_n)_{n=1}^{\infty}$ has no Cauchy subsequence.*

PROOF. If some subsequence $(x_{n_k})_{k=1}^{\infty}$ were Cauchy, then we would have $d(x_{n_k}, x_{n_{k+1}}) < \varepsilon$ for sufficiently large k, contradicting our hypotheses. □

It is easy to come up with examples of sequences for which the hypotheses of Lemma 3.3 hold; the sequence $(n)_{n=1}^{\infty}$ in $\mathbb{R}$ that we have already mentioned has this property. Later on, we will see that $\mathbb{R}$ is a complete metric space. But $\mathbb{R}$ is not sequentially compact, since the sequence $(n)_{n=1}^{\infty}$ has no subsequential limit in $\mathbb{R}$ (though it does in $\overline{\mathbb{R}}$!). One might suspect that knowing a metric spaces is complete and *bounded* might guarantee you sequential compactness, but this is false, as the next example shows.

EXAMPLE 3.4. Let X be any infinite set, and let d be the discrete metric. Then X is bounded, since $X = B_{(X,d)}(x, 2)$ for any $x \in X$. Furthermore, X is complete: The only way for $(x_n)_{n=1}^{\infty}$ to be a Cauchy sequence is if there exists $x \in X$ such that $x_n = x$ for all sufficiently large n, in which case the sequence converges to x. However, if $(x_n)_{n=1}^{\infty}$ is any sequence in X whose terms are all distinct, then $d(x_n, x_m) = 1$ for all $n \neq m$. In particular, no subsequence of $(x_n)_{n=1}^{\infty}$ can converge (or even be Cauchy), even though X is bounded. Therefore (X, d) is not sequentially compact.

Let us propose a stronger notion of boundedness, called *total* boundedness, which eliminates the issues of the previous example. First, we need the very important definition of a *cover*.

DEFINITION 3.5. Let $\mathcal{A}$ be a collection of subsets of a set X, and let E be a subset of X. We say the collection $\mathcal{A}$ *covers* E (or is a *cover* of E) if

$$E \subset \bigcup_{A \in \mathcal{A}} A.$$

DEFINITION 3.6. Let (X, d) be a metric space. X is said to be *totally bounded* (with respect to d) if for every $\varepsilon > 0$, X can be covered by finitely many balls of radius ε, that is, if there exist finitely many points $x_1, \ldots, x_n \in X$ such that

$$X = \bigcup_{j=1}^{n} B_{(X,d)}(x_j, \varepsilon).$$

Of course, n may depend on ε, but ε is the same for each j.

The following examples distinguish between the notions of boundedness and total boundedness.

EXAMPLE 3.7. If X is any infinite set and d is the discrete metric, then X is bounded (as argued above) but not totally bounded, with respect to d. Indeed, $B(x, \frac{1}{2}) = \{x\}$ for each $x \in X$—in particular no ball of radius $\frac{1}{2}$ contains more than one point in X, so any finite union of such balls will contain only finitely many points. Such a union can never be equal to all of the infinite set X; therefore X is not totally bounded.

EXAMPLE 3.8. Let X be the set of all bounded sequences in $\mathbb{R}$; let d_u be the metric on X induced from the uniform norm on X. (That is, $d_u((x_n)_{n=1}^{\infty}, (y_n)_{n=1}^{\infty}) = \sup_{n \in \mathbb{N}} |x_n - y_n|$, c.f. Example 1.9 in Chapter 3.) Let Y be the unit ball in this metric space,

$$Y = B_{(X,d_u)}(0, 1) = \{(x_1, x_2, \ldots) : |x_j| \leq 1 \text{ for all } j \in \mathbb{N}\},$$

Then Y is manifestly bounded, being a ball itself. But it is not totally bounded. Indeed, we claim that there is no covering of Y by balls of radius $\frac{1}{2}$. To see this, for each $j \in \mathbb{N}$, let e_j be the sequence in Y with 1 as its jth term and zeros elsewhere. Then $d_u(e_i, e_j) = 1$ for $i \neq j$. Therefore $B(x, \frac{1}{2})$ contains at most one of the e_i's for any $x \in X$, so no finite collection of such balls can contain all the e_i's, and in particular no finite collection of balls of radius $\frac{1}{2}$ can cover all of Y. We conclude that Y is not totally bounded.

Incidentally, notice that the sequence $(e_1, e_2, e_3, \ldots)$ in Y satisfies the hypotheses of Lemma 3.3.

PROPOSITION 3.9. *Let (X, d) be a metric space.*

(a) If X is totally bounded, then it is bounded.
(b) If $Y \subset X$, then (Y, d) is totally bounded if and only if for any $\varepsilon > 0$, there exist $a_1, \ldots, a_J \in X$ such that $Y \subset \bigcup_{j=1}^{J} B_X(a_j, \varepsilon)$. (The point is that the a_j's do not have to lie in Y.)
(c) If X is totally bounded and $Y \subset X$, then Y is totally bounded.
(d) If $Y \subset X$ and (Y, d) is totally bounded, then $(\overline{Y}, d)$ is totally bounded.

EXERCISE 3.2. Prove Proposition 3.9.

EXERCISE 3.3. Prove that $[-R, R]^n$ is a totally bounded subset of $\mathbb{R}^n$ under the Euclidean metric. Conclude that boundedness and total boundedness are equivalent for any subset of $\mathbb{R}^n$.

EXERCISE 3.4. We say that a metric space (X, d) is *separable* if it has a countable dense subset, i.e., if there exists a countable set E such that $\overline{E} = X$. Prove that every totally bounded metric space is separable. (Hint: Use the fact that (X, d) is totally bounded—countably many times—to construct a candidate for E. Then, given $x \in X$, find a sequence in E that converges to x in X.)

3.2. A Characterization of Sequentially Compact Metric Spaces. As we have strongly hinted at in the previous subsection, the gap between 'complete' and 'sequentially compact' metric spaces is precisely total boundedness. We will prove this fact later in the present subsection. The intuition is that total boundedness of X prevents the hypotheses of Lemma 3.3 from being satisfied; it forces any sequence to have a Cauchy subsequence, since the sequence 'runs out of places to go.' Once we have a Cauchy subsequence, then completeness of X guarantees that the point where the sequence 'wants' to converge actually belongs to X. We formalize some of this intuition in the following statements.

PROPOSITION 3.10. *(X, d) is totally bounded if and only if every sequence $(x_n)_{n=1}^{\infty}$ has a Cauchy subsequence in X.*

The proof of this Proposition is rather technical, so we postpone it until after we have given the 'punchline' for this section, in the following Theorem, which we obtain as an immediate consequence of the above.

THEOREM 3.11. *Let (X, d) be a metric space. Then X is sequentially compact if and only if it is complete and totally bounded.*

PROOF. ($\Longrightarrow$) If (X, d) is sequentially compact, then it must be complete, as we have already shown (Corollary 3.2). It must also be totally bounded; otherwise, by the previous Proposition there would exist a sequence in X with no Cauchy subsequence and thus no subsequential limits. ($\Longleftarrow$) On the other hand, if it is totally bounded, then Proposition 3.10 guarantees that every sequence in X has a Cauchy subsequence; if X is also complete, then this Cauchy subsequence must converge, whence we conclude X is sequentially compact. □

The following special case of the this Theorem is useful enough to single out:

COROLLARY 3.12. *Let (X, d) be a metric space and let Y be a subset of X. If (Y, d) is sequentially compact, then Y is closed and bounded in X.*

PROOF. By the previous Theorem, Y is complete; therefore it is closed in X (Exercise 2.4. Furthermore, Y is totally bounded, therefore bounded (Proposition 3.9). □

Finally, we give the proof of the technical Proposition stated above.

PROOF OF PROPOSITION 3.10. ($\Longleftarrow$) Assume that X is *not* totally bounded, and choose $\varepsilon > 0$ such that no finite collection of ε-balls covers X. We construct a sequence $(x_n)_{n=1}^{\infty}$ as follows. Pick $x_1 \in X$ arbitrarily. Assume inductively that $x_1, \ldots, x_k$ have been chosen so that $d(x_i, x_j) \geq \varepsilon$ for all $i, j \in J_k$. Then our choice of ε guarantees that $\bigcup_{i=1}^{k} B_X(x_i, \varepsilon)$ is not all of X; we can therefore pick x_{k+1} outside this set, so that now $d(x_i, x_j) \geq \varepsilon$ for all $i, j \in J_{k+1}$. The sequence $(x_n)_{n=1}^{\infty}$ has no subsequential limits, by Lemma 3.3, so the implication is proven.

($\Longrightarrow$) Now, assume X is totally bounded. We show that every sequence in X has a Cauchy subsequence. We formulate our strategy in the following Lemma, which is designated as such for future reference.

LEMMA 3.13. *Let (X, d) be a totally bounded metric space and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. Then there is a sequence $(X_k)_{k=1}^{\infty}$ of subsets of X with the following properties:*

(i) $X_k \supset X_{k+1}$ for all $k \in \mathbb{N}$.
(ii) $\mathrm{diam} X_k \to 0$ as $k \to \infty$.
(iii) For each $k \in \mathbb{N}$, we have $x_n \in X_k$ for infinitely many $n \in \mathbb{N}$.

For any sequence $(X_k)_{k=1}^{\infty}$ satisfying these properties, there exists a corresponding subsequence $(x_{n_k})_{k=1}^{\infty}$ of $(x_n)_{n=1}^{\infty}$, such that $x_{n_k} \in X_k$ for each $k \in \mathbb{N}$. This subsequence must be Cauchy.

PROOF. Since X is totally bounded, we may choose a finite subset $\{a_{1,1}, \ldots, a_{1,M_1}\}$ of X such that $X = \bigcup_{m=1}^{M_1} B_X(a_{1,m}, 1)$. At least one of the balls $B_X(a_{1,m}, 1)$ contains infinitely many of the x_j's; let $m_1 \in \{1, \ldots, M_1\}$ be chosen so that $B_X(a_{1,m_1}, 1)$ has this property. Now, put $X_1 = B_X(a_{1,m_1}, 1)$ and repeat the process: Find a finite set $\{a_{2,1}, \ldots, a_{2,M_2}\} \subset X$ such that $X_1 = \bigcup_{m=1}^{M_2} B_{X_1}(a_{2,m}, \frac{1}{2})$. Choose $m_2 \in \{1, \ldots, M_2\}$ such that $B_{X_1}(a_{2,m_2}, \frac{1}{2})$ contains infinitely many of the x_j's, and put $X_2 = B_{X_1}(a_{2,m_2}, \frac{1}{2})$. This process can clearly be continued indefinitely to yield a sequence $(X_j)_{j=1}^{\infty}$ of subsets of X satisfying (i)–(iii).

The existence of the proposed subsequence is an immediate consequence of property (iii): Just choose n_1 such that $x_{n_1} \in X_1$, then choose $n_2 > n_1$ such that $x_{n_2} \in X_2$, and continue this process indefinitely. Finally, the fact that this subsequence is Cauchy follows from the other two properties: Since the X_k's are nested (property (i)), we have $x_{n_j} \in X_k$ for all $j \geq k$, and thus $d(x_{n_i}, x_{n_j}) \leq \mathrm{diam} X_k$ for all $i, j \geq k$. Since $\mathrm{diam} X_k \to 0$ as $k \to \infty$ (property (ii)), it follows that $(x_{n_k})_{k=1}^{\infty}$ is a Cauchy sequence. □

Since Lemma 3.13 completely takes care of the direction ($\Longrightarrow$), we have also completed the proof of Proposition 3.10. □

3.3. The Bolzano–Weierstrass Theorem and Sequentially Compact subsets of $\mathbb{R}^n$. We are almost in a position to characterize the sequentially compact subset of $\mathbb{R}^n$.

LEMMA 3.14. *Let $(I_k)_{k \in \mathbb{N}}$ be a sequence of closed intervals in $\mathbb{R}$, such that $I_k \supset I_{k+1}$ for each $k \in \mathbb{N}$. Then the following statements hold:*

(a) $A := \bigcap_{k=1}^{\infty} I_k$ is not empty.
(b) If $\mathrm{diam} I_k \to 0$ as $k \to \infty$, then A is a singleton set.

PROOF. (a) Write $I_k = [a_k, b_k]$ for each k. Since $I_k \subset I_1$, we have $b_k \leq b_1$; therefore $a_k \leq b_1$ for each $k \in \mathbb{N}$. The set of a_k is therefore nonempty and bounded above (by b_1), so $\sup a_k$ exists in $\mathbb{R}$; call it x. We claim that $x \in A$; to show this we must prove that $x \in I_\ell$ for each $\ell \in \mathbb{N}$. To this end, choose $\ell \in \mathbb{N}$. Then $x \geq a_\ell$ by definition of the supremum; to show that $x \leq b_\ell$ it suffices to show that b_ℓ is an

upper bound for the a_k's. That the latter is true follows from the inequalities $a_k \leq a_{k+\ell} \leq b_{k+\ell} \leq b_\ell$, which hold for all $k \in \mathbb{N}$.

(b) Let x and y be elements of A; we show that $x = y$. Indeed, since $x, y \in I_k$ for all k, we must have $d(x,y) \leq \text{diam} I_k$ for all k. Since $\text{diam} I_k \to 0$ as $k \to \infty$, this is only possible if $d(x,y) = 0$, i.e., if $x = y$. □

THEOREM 3.15. *The interval $[a,b]$ is a sequentially compact subset of $\mathbb{R}$ under the Euclidean metric.*

PROOF. Let $(x_n)_{n=1}^\infty$ be a sequence in $[a,b]$. at least one of the intervals $[a, \frac{a+b}{2}]$ or $[\frac{a+b}{2}, b]$ contains infinitely many of the x_j's; choose one that contains infinitely many of the x_j's and relabel it $I_1 = [a_1, b_1]$. Repeat this process: At least one of $[a_1, \frac{a_1+b_1}{2}]$ or $[\frac{a_1+b_1}{2}, b_1]$ contains infinitely many of the x_j's; choose one and relabel it $I_2 = [a_2, b_2]$. We thus obtain a sequence of closed intervals I_j with the following properties:

- $I_k \supset I_{k+1}$ for each $k \in \mathbb{N}$,
- $\text{diam} I_k = 2^{-k}(b-a) \to 0$ as $k \to \infty$,
- For each $k \in \mathbb{N}$, we have $x_n \in I_k$ for infinitely many $n \in \mathbb{N}$.

Therefore, by (the second statement in) Lemma 3.13, there exists a subsequence $(x_{n_k})_{k=1}^\infty$ such that $x_{n_k} \in I_k$ for each k. On the other hand, by Lemma 3.14(b), the set $\bigcap_{k=1}^\infty I_k$ is a singleton set $\{x\}$. Now, $d(x_{n_k}, x) \leq \text{diam} I_k$ for each $k \in \mathbb{N}$; since $\text{diam} I_k \to 0$ as $k \to \infty$ we must have $x_{n_k} \to x$ as $k \to \infty$. □

THEOREM 3.16. *A subset E of $\mathbb{R}^n$ (under the Euclidean metric) is sequentially compact if and only if it is closed and bounded.*

This Theorem is sometimes called the *Bolzano–Weierstrass Theorem*. We have already done most of the hard work of the proof.

PROOF. Note first of all that if E is sequentially compact, then it must be closed and bounded, by Corollary 3.12.

We break the other direction into two steps. First, we prove that $C = [-R,R]^n$ is sequentially compact for any $R > 0$. For this statement, it suffices to show that C is complete, since C is totally bounded (Exercise 3.3). To this end, let $(\mathbf{x}_m)_{m=1}^\infty$ be a sequence in C; for each n, denote $\mathbf{x}_m = (x_{m,1}, \ldots, x_{m,n})$. Then $(x_{n,1})_{n=1}^\infty$ is a sequence in the sequentially compact metric space $[-R,R]$; therefore it has a subsequence $(x_{m_{1,\ell},1})_{\ell=1}^\infty$ that converges in $[-R,R]$; let x_1 denote the limit. By the same reasoning, $(x_{m_{1,\ell},2})_{\ell=1}^\infty$ also has a convergent subsequence $(x_{m_{2,\ell},2})_{\ell=1}^\infty$; let x_2 denote the limit. Note that since $(m_{2,\ell})_{\ell=1}^\infty$ is a subsequence of $(m_{1,\ell})_{\ell=1}^\infty$, we still have $x_{m_{2,\ell},1} \to x_1$ as $\ell \to \infty$. Repeating this process a total of n times, we eventually obtain a subsequence $(m_{k,\ell})_{\ell=1}^\infty$ such that $(x_{m_{k,\ell},i})_{\ell=1}^\infty$ converges to some $x_i \in [-R,R]$ for each $i \in \{1, \ldots, k\}$. Thus $\mathbf{x}_{m_{k,\ell}} \to \mathbf{x} := (x_1, \ldots, x_k)$ as $\ell \to \infty$. This shows that C is sequentially compact.

Next, let E be any closed, bounded subset of $\mathbb{R}^n$. Choose $R > 0$ large enough so that $E \subset [-R,R]^n$. Then E is totally bounded, since $[-R,R]^n$ is; furthermore Theorem 2.12(b) guarantees that E is complete, since it is closed in $\mathbb{R}^n$ and therefore in the closed set $[-R,R]^n$. Having shown that E is complete and totally bounded, we may conclude that it is sequentially compact. □

COROLLARY 3.17. *$\mathbb{R}^n$ equipped with the Euclidean metric is a complete metric space.*

PROOF. Let $S = (x_m)_{m=1}^\infty$ be a Cauchy sequence in $\mathbb{R}^n$. Then S is bounded, so $K := \overline{\text{Im}\, S}$ is closed and bounded (see Exercise 1.14), therefore sequentially compact, therefore complete. Since S is a Cauchy sequence in the complete metric space K, it follows that S converges in K, therefore in $\mathbb{R}^k$. □

Note that Theorem 3.16 is not true in general metric spaces. See for example the following.

EXAMPLE 3.18. Put $X = \mathbb{R}$, under the Euclidean metric. Put $Y = \mathbb{Q}$ and $F = [\sqrt{2}, 5] \cap \mathbb{Q}$. Then F is closed and bounded in $\mathbb{Q}$, but it is not sequentially compact, since it is not closed in $\mathbb{R}$. Indeed,

$$\mathrm{Cl}_{\mathbb{R}}(F) = [\sqrt{2}, 5] \neq F.$$

Thus F is not closed in $\mathbb{R}$, hence not sequentially compact. You can also see that F is not sequentially compact by finding a Cauchy sequence in F that converges in $\mathbb{R}$ to an irrational number. This sequence will not converge in F, which will show that F is not complete and thus not sequentially compact. (This sequence will also have no subsequenial limits in F.)

4. Compactness

4.1. Basic Definitions and Examples.

DEFINITION 4.1. Let $\mathcal{A}$ be a collection of subsets of a set X, and let E be a subset of X. We say the collection $\mathcal{A}$ *covers* E (or is a *cover* of E) if

$$E \subset \bigcup_{A \in \mathcal{A}} A.$$

A *subcover* of E corresponding to the cover $\mathcal{A}$ is a subcollection $\mathcal{B} \subset \mathcal{A}$ that still covers E.

EXAMPLE 4.2. Let $X = \mathbb{R}$, $E = (0, 2) \cup \{5\}$, and $\mathcal{A} = \{(-1, 1), (0, 1), \{8\}, (0, 6)\} \cup \{(n, n+1)\}_{n=1}^{\infty}$. Then $E \subset \bigcup_{A \in \mathcal{A}} A$, so $\mathcal{A}$ is a cover of E. But some (in fact, most) of the sets of $\mathcal{A}$ were unnecessary for the purpose of covering E; if we take $\mathcal{B} = \{(-1, 1), (0, 6)\}$, then $\mathcal{B} \subset \mathcal{A}$, and $E \subset \bigcup_{B \in \mathcal{B}} B$, so $\mathcal{B}$ is a *subcover* of E associated to the cover $\mathcal{A}$. In fact, the set $\mathcal{B}$ is finite subcollection of $\mathcal{A}$; to emphasize this fact, we say that $\mathcal{B}$ is a *finite subcover* of E associated to the cover $\mathcal{A}$.

DEFINITION 4.3. Let (X, d) be a metric space. A collection $\mathcal{A}$ of subsets of X is called an *open cover* of E in X if $\mathcal{A}$ covers E and each $A \in \mathcal{A}$ is open in X. A subset K of X is called *compact* in X if every open cover of K in X has a finite subcover.

One nice property of compactness is that it behaves nicely when passing to subsets (especially in comparison with the notions of open, closed, etc.)

THEOREM 4.4. *Let (X, d) be a metric space, and let K be a subset of X. Then (K, d) is compact if and only if every covering of K by open sets in X has a finite subcover.*

The operative part of the statement above is 'open sets in X', instead of 'in K'. This Theorem allows us to talk about a 'compact subset' of a metric space. The characterization in this Theorem is sometimes taken as the *definition* of a compact set. We choose to *prove* this characterization from the stated definition stated above because the latter emphasizes that compactness is an intrinsic property of a metric space, rather than something that depends on an underlying set. Conversely, *defining* compactness of K 'with respect to' the underlying set X obscures this point. However, the convenience of being able to refer to a 'compact subset' K, and being able to take open coverings in a larger space X, does make the characterization provided by this Theorem indispensable. Note also that even if one instead defines compactness according to the hypotheses of the Theorem above, one still must prove a statement similar to the Theorem, such as 'If $K \subset Y \subset X$, then K is compact in Y if and only if it is compact in X'.

PROOF. ($\Longrightarrow$) Assume (K, d) is compact. Let $\mathcal{A}_X$ denote a covering of K by open subsets of X. Define $\mathcal{A}_K = \{A \cap K : A \in \mathcal{A}_X\}$. Then $\mathcal{A}_K$ is a covering of K by open subsets of K. Choose $A_1, \ldots, A_n$ such that $\{A_1 \cap K, \cdots, A_n \cap K\}$ is a finite subcover of K associated to the covering $\mathcal{A}_K$. Then $\{A_1, \ldots, A_n\}$ is a finite subcover of K associated to the covering $\mathcal{A}_X$.

($\Longleftarrow$) Assume that any covering of K by open subsets of X has a finite subcover. Let $\mathcal{B}_K$ be a covering of K by open subsets of K, and for each $B \in \mathcal{B}_K$, choose a subset U_B of X which is open in X and satisfies $U_B \cap K = B$. Then $\mathcal{B}_X = \{U_B : B \in \mathcal{B}_K\}$ is a cover of K by open subsets of X; let

$B_1, \ldots, B_m$ be such that $\{U_{B_1}, \ldots, U_{B_m}\}$ is a finite subcover of K associated to the covering $\mathcal{B}_X$. Then $\{B_1, \ldots, B_m\}$ is a finite subcover of K associated to the covering $\mathcal{B}_K$ of K. It follows that (K, d) is compact, as needed. □

REMARK 4.5. Because of the distinguished role played by covers of $K \subset X$ consisting of *open* subsets of the metric space (X, d), we often use the terminology 'open cover of K in X' rather than 'covering of K by open subsets of X.' If it is clear from context that the 'open' in 'open cover' refers to subsets of X rather than some subset thereof, we sometimes refer simply to an 'open cover of K.'

EXAMPLE 4.6. If (X, d) is a metric space and K is any finite subset of X, then (K, d) is compact. Indeed, if $K = \{x_1, \ldots, x_n\}$ and $\mathcal{A}$ is an open cover of K, then we can choose for each j an element A_j of $\mathcal{A}$ such that $A_j \ni x_j$. Then $K \subset \bigcup_{j=1}^n A_j$, so $\{A_j\}_{j=1}^n$ is a finite subcover of K associated to the open cover $\mathcal{A}$.

EXAMPLE 4.7. Let E denote the image of the sequence $S = (\frac{1}{n})_{n=1}^\infty$; $E = \operatorname{Im} S$. Then E (under the Euclidean metric) is not compact. Indeed, the open covering $\mathcal{A} = \{(\frac{1}{2n}, \frac{2}{n})\}_{n=1}^\infty$ of E in $\mathbb{R}$ has no finite subcover, as you can readily check. However, $K = \{0\} \cup E$ *is* compact under the Euclidean metric. (Note that K is not covered by $\mathcal{A}$.) Indeed, if $\mathcal{B}$ is an open cover of K in $\mathbb{R}$, then some $B_0 \in \mathcal{B}$ contains 0 and thus contains $\frac{1}{n}$ for all n larger than some $N \in \mathbb{N}$. For $n \leq N$, we can find $B_n \in \mathcal{B}$ containing $\frac{1}{n}$, so that $K \subset B_0 \cup B_1 \cup \cdots \cup B_N$.

4.2. A Few Basic Properties of Compact Sets in Metric Spaces. Let us establish a few basic properties of compact sets. We begin with one that is particularly intuitive:

PROPOSITION 4.8. *Let (X, d) be a compact metric space. Then (X, d) is totally bounded.*

PROOF. Choose $\varepsilon > 0$ and consider the open covering $\{B_X(x, \varepsilon)\}_{x \in X}$ of X. Since (X, d) is compact, we can extract a finite subcover $\{B_X(x_i, \varepsilon)\}_{i=1}^n$. This is a covering of X by ε-balls. Since $\varepsilon > 0$ was arbitrary, it follows that X is totally bounded. □

Note that compactness is a much more fundamental notion than total boundedness, since the latter requires the notion of a metric in order to make sense, whereas the former requires only a topology. Though we will continue to frame our discussion in the context of metric spaces, the properties of compactness in the remainder of this subsection and the next one rely only on the topology of the space under consideration.

THEOREM 4.9. *Let (X, d) be a compact metric space and let K be a closed subset of X. Then K is compact.*

PROOF. Let $\mathcal{A}$ be an open cover of K in X; denote $\mathcal{B} = \mathcal{A} \cup \{K^c\}$. Then $\mathcal{B}$ is an open cover of X and X is compact; therefore we can find a finite subcover $\mathcal{B}'$ of X associated to the open cover $\mathcal{B}$. This new collection $\mathcal{B}'$ still covers K of course (since it covers X), but it might or might not be a subcollection of $\mathcal{A}$ (depending on whether $K^c \in \mathcal{B}'$). But removing K^c from $\mathcal{B}'$ will not affect whether $\mathcal{B}'$ covers K; therefore we define $\mathcal{A}' = \mathcal{B}' \backslash \{K^c\}$. Then $\mathcal{A}'$ is a finite subcover of K associated to the open cover $\mathcal{A}$ of K in X. Therefore K is compact. □

COROLLARY 4.10. *Let (X, d) be a metric space. Assume F and K are subsets of X, with F closed and K compact. Then $F \cap K$ is compact.*

EXERCISE 4.1. Prove Corollary 4.10. (The proof is very short.)

4.3. Compact Sets and the Finite Intersection Property.

DEFINITION 4.11. Let X be a set and let $\mathcal{A}$ be a collection of subsets of X. Then $\mathcal{A}$ is said to have the *finite intersection property* if $\bigcap_{A \in \mathcal{B}} A$ is nonempty whenever $\mathcal{B}$ is a finite subset of $\mathcal{A}$.

EXAMPLE 4.12. Let $\mathcal{A} = \{A_n\}_{n=1}^{\infty}$ be a countable collection of nonempty nested sets (i.e., $A_n \supset A_{n+1}$ for each $n \in \mathbb{N}$). Then $\mathcal{A}$ has the finite intersection property. Indeed, let C be a finite subset of $\mathbb{N}$. Then C has a largest element N, and $\bigcap_{n\in C} A_n = A_N$.

THEOREM 4.13. *Let $\mathcal{A}$ be a nonempty collection of compact subsets of a metric space (X, d). Assume that $\mathcal{A}$ has the finite intersection property. Then $\bigcap_{K\in\mathcal{A}} K$ is nonempty.*

We prove the Theorem under the additional assumption that each $K \in \mathcal{A}$ is closed. As we will see in Corollary 4.17, this assumption turns out to be unnecessary, as compact subsets of a metric space are always closed.

PROOF. We argue by contradiction. Assume that $\bigcap_{K\in\mathcal{A}} K$ is empty. Then

$$\bigcup_{K\in\mathcal{A}} (X\backslash K) = X\backslash \left(\bigcap_{K\in\mathcal{A}} K \right) = X.$$

That is, the collection $\{K^c\}_{K\in\mathcal{A}}$ is an open cover for X, and thus it is an open cover for any fixed $K^* \in \mathcal{A}$. Choose such a $K^* \in \mathcal{A}$, then choose a finite subcover $\{K_1^c, \ldots, K_n^c\}$ of K^*, associated to the cover $\{K^c\}_{K\in\mathcal{A}}$ of K^*. Then

$$K^* \subset K_1^c \cup \cdots \cup K_n^c,$$

and consequently $X\backslash K^* \supset K_1 \cap \cdots \cap K_n$. But then

$$K^* \cap K_1 \cap \cdots \cap K_n \subset K^* \cap (X\backslash K^*) = \emptyset.$$

This contradicts our hypothesis since $\{K^*\} \cup \{K_j\}_{j=1}^n$ is a finite subcollection of elements of $\mathcal{A}$, and $\mathcal{A}$ has the finite intersection property. □

COROLLARY 4.14. *Let $(K_n)_{n=1}^{\infty}$ be a sequence of nonempty compact subsets of a metric space (X, d) and assume that $K_n \supset K_{n+1}$ for all $n \in \mathbb{N}$. Then*

$$\bigcap_{n=1}^{\infty} K_n \neq \emptyset.$$

PROOF. Since the K_n's are nested, the set $\{K_n\}_{n=1}^{\infty}$ has the finite intersection property (see the example above). Then by Theorem 4.13, the intersection of all of the K_n's is nonempty. □

EXAMPLE 4.15. Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}$ defined by $\mathcal{A} = \{[n, \infty)\}_{n=1}^{\infty}$. Then $\mathcal{A}$ has the finite intersection property (see the example above), but

$$\bigcap_{A\in\mathcal{A}} A = \bigcap_{n=1}^{\infty} [n, \infty) = \emptyset.$$

In light of Theorem 4.13 and Proposition 4.8, we observe that the sets $[n, \infty)$ are not bounded in $\mathbb{R}$, hence not compact in $\mathbb{R}$.

Theorem 4.13 (and Corollary 4.14) allow us to prove that compactness and sequential compactness are the same in metric spaces. These statements will also be useful in our discussion of perfect sets below. The latter will allow us to give a proof of the fact that $\mathbb{R}$ is uncountable.

4.4. Equivalence of Compactness and Sequential Compactness in a Metric Space. It turns out that compactness and sequential compactness are equivalent for metric spaces. (They are not equivalent in the setting of more general topological spaces). Here is the 'official' statement, which we will prove in the next subsection after some discussion of its consequences.

THEOREM 4.16. *Let (X, d) be a metric space. The following two statements are equivalent.*

(1) (X, d) is compact.

(2) Whenever $(F_n)_{n=1}^{\infty}$ is a nested sequence ($F_n \supset F_{n+1}$ for all $n \in \mathbb{N}$) of closed, nonempty subsets of X, we have that the intersection $\bigcap_{n=1}^{\infty} F_n$ is nonempty.
(3) (X, d) is sequentially compact.

The following two Corollaries follow immediately from the above Theorem and already-established properties of sequential compactness. It is of course possible to give self-contained proofs of these facts, without recourse to Theorem 4.16.

COROLLARY 4.17. *Let (X, d) be a metric space. If K is a compact subset of X, then K is closed in X.*

COROLLARY 4.18 (Heine-Borel Theorem). *A subset of $\mathbb{R}^n$ (under the Euclidean metric) is compact if and only if it is closed and bounded.*

EXERCISE 4.2. Give an example of a collection $\mathcal{A}$ of *bounded* subsets of $\mathbb{R}$ such that $\mathcal{A}$ has the finite intersection property, but $\bigcap_{A \in \mathcal{A}} A = \emptyset$. Hint: If $A \subset \mathbb{R}$ is bounded in $\mathbb{R}$, what else can prevent it from being compact?

4.5. Proof of Theorem 4.16. In this subsection, we give the proof of Theorem 4.16, which establishes the equivalence of compactness and sequential compactness in a metric space. To prove that sequential compactness implies compactness, we need one final Lemma.

LEMMA 4.19 (Lebesgue Number Lemma). *Let $\mathcal{A}$ be an open cover of a sequentially compact metric space (X, d). There exists $\delta > 0$ such that for any $x \in X$, $B_X(x, \delta) \subset A$ for some $A \in \mathcal{A}$.*

The number $\delta > 0$ is called the 'Lebesgue number' associated to the covering $\mathcal{A}$.

PROOF. Suppose no such $\delta > 0$ exists. Then for each $n \in \mathbb{N}$, there exists $x_n \in X$ such that $B_X(x_n, \frac{1}{n})$ is not contained in any $A \in \mathcal{A}$. Let $(x_{n_k})_{k=1}^{\infty}$ be a convergent subsequence of $(x_n)_{n=1}^{\infty}$ with limit x. Choose $A \in \mathcal{A}$ such that $x \in A$. Then choose $\varepsilon > 0$ so that $B_X(x, \varepsilon) \subset A$ (this is possible because A is open by assumption). Choose k large enough so that $d(x_{n_k}, x) < \frac{\varepsilon}{2}$ and $n_k > \frac{2}{\varepsilon}$. Then $B_X(x_{n_k}, \frac{1}{n_k}) \subset B_X(x, \varepsilon) \subset A$ (where the first inclusion is justified by the inequality $d(y, x) \leq d(y, x_{n_k}) + d(x_{n_k}, x)$, the right side of which is less than $\frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ for $y \in B_X(x_{n_k}, \frac{\varepsilon}{2})$). This contradicts the definition of x_{n_k} and thus proves the desired statement. □

THEOREM 4.16. (1) $\implies$ (2) Suppose (X, d) is compact, and let $(F_n)_{n=1}^{\infty}$ be a nested sequence of closed, nonempty subsets of X, $F_n \supset F_{n+1}$, for all $n \in \mathbb{N}$. Then each F_n is compact by Theorem 4.9, and then Corollary 4.14 implies that $\bigcap_{n=1}^{\infty} F_n$ is nonempty.

(2) $\implies$ (3) Assume (2) holds, and let $(x_n)_{n=1}^{\infty}$ be a sequence in X. Define $T_i = \{x_k\}_{k=i}^{\infty}$ (the image of the 'tail' starting at the ith term), and denote $F_i = \overline{T_i}$. Then by (2), the intersection $F := \bigcap_{i=1}^{\infty} F_i$ is nonempty; let x be an element of F. We construct a subsequence $(x_{n_k})_{k=1}^{\infty}$ of $(x_n)_{n=1}^{\infty}$ converging to x in X as follows.

For every $\varepsilon > 0$ and for every $i \in \mathbb{N}$, we have that $x \in F_i = \overline{T}_i$, so that there exists $y \in T_i$ such that $d(x, y) < \varepsilon$. Furthermore, since $T_i = \{x_k\}_{k=i}^{\infty}$, it follows that this y is equal to x_j for some $j \geq i$.

Put $n_0 = 1$. Using the above, we choose $n_1 \in \mathbb{N}$ so that $x_{n_1} \in T_1 = T_{n_0}$ and $d(x_{n_1}, x) < 1$. Then having chosen $n_0 < n_1 < \ldots < n_i$ so that $x_{n_j} \in T_{n_{j-1}}$ and $d(x_{n_j}, x) < \frac{1}{j}$ for each $j \in \{1, \ldots i\}$, we choose $n_{i+1} > n_i$ so that $x_{n_{i+1}} \in T_{n_i}$ and $d(x_{n_{i+1}}, x) < \varepsilon$. The subsequence $(x_{n_k})_{k=1}^{\infty}$ now converges to x by construction.

(3) $\implies$ (1) Assume (X, d) is sequentially compact. Let $\mathcal{A}$ be an open cover of X, and let $\delta > 0$ be the Lebesgue number associated to the cover $\mathcal{A}$. Since (X, d) is totally bounded, we can find $a_1, \ldots, a_n$ such that $X = \bigcup_{j=1}^{n} B_X(a_j, \delta)$. For each $j \in \{1, \ldots, n\}$, choose A_j such that $B_X(a_j, \delta) \subset A_j$. (This is possible by the Lebesgue number Lemma.) Then $\{A_1, \ldots, A_n\}$ is a finite subcover of X associated to the open covering $\mathcal{A}$ of X. It follows that (X, d) is compact, as needed. □

5. Perfect Sets in $\mathbb{R}^n$

DEFINITION 5.1. Let (X, d) be a metric space. A subset E of X is called *perfect* if $E = E'$.

REMARK 5.2. Closed sets are those which contain all their limit points (i.e. $E' \subset E$). Thus every perfect set is closed. But a perfect set cannot have isolated points. (If $x \in E$ is an isolated point, then $x \notin E'$, so $E \neq E'$.)

EXAMPLE 5.3. $E = [0, 1]$ is a perfect set of $\mathbb{R}$, but $F = [0, 1] \cup \{2\}$ is not, even though it is closed. The problem is that 2 is an isolated point of F.

5.1. Perfect Sets and the Uncountability of $\mathbb{R}^n$.

THEOREM 5.4. *Let P be a nonempty perfect subset of $\mathbb{R}^n$. Then P is uncountable.*

PROOF. Since P is nonempty, it contains some element p, which must be a limit point of P since P is perfect. Since finite sets have no limit points and P does have limit points, P cannot be finite. We claim that P is also not countable. To see this, we argue by contradiction. Assume that P is countable, and let $(x_1, x_2, \ldots)$ be an enumeration of its elements. Our goal is to construct a sequence $(K_i)_{i=1}^{\infty}$ of nonempty compact sets such that

- $K_{i+1} \subset K_i \subset P$ for each $i \in \mathbb{N}$ and
- For each $i \in \mathbb{N}$, we have $x_i \notin K_i$.

Assume for a moment we have done this. Then by the first two points, $K := \bigcap_{i=1}^{\infty} K_i$ is a nonempty subset of P. However, by the second point, K cannot contain any of the x_i, which means that it *is* empty. Clearly this is a contradiction. Therefore, to finish the proof, it suffices to construct compact sets K_i with the properties above.

Choose $p_1 \in P$, $p_1 \neq x_1$. Put $\varepsilon_1 = \frac{1}{2}|p_1 - x_1|$; then $\overline{B(p_1, \varepsilon_1)}$ does not contain x_1. However, since $p_1 \in P$ and P is perfect, p_1 must be a limit point of P. Therefore $B(p_1, \varepsilon_1)$ contains another point of P which is not equal to x_2; call it p_2. Put $\varepsilon_2 = \min\{\frac{1}{2}|p_2 - x_2|, \varepsilon - |p_1 - p_2|\}$. Then $B(p_2, \varepsilon_2)$ is contained in $B(p_1, \varepsilon_1)$ and $x_2 \notin \overline{B(p_2, \varepsilon_2)}$. We continue in this way, obtaining p_j, ε_j so that $\overline{B(p_j, \varepsilon_j)}$ does not contain x_j, and $B(p_{j+1}, \varepsilon_{j+1}) \subset B(p_j, \varepsilon_j)$.

Put $K_i = \overline{B(p_i, \varepsilon_i)} \cap P$ for each $i \in \mathbb{N}$. Then $\{K_i\}_{i \in \mathbb{N}}$ consists of nonempty compact sets that satisfy the two bullet points above. □

COROLLARY 5.5. *Any open or closed interval in $\mathbb{R}$ is uncountable. In particular, $\mathbb{R}$ is uncountable.*

5.2. The Cantor Set. Next, we will construct a perfect set *which contains no intervals*. This set is called the *Cantor set* after Georg Cantor. The Cantor set is an example of a fractal; its "dimension" is non-integer valued in most senses of the word dimension, and it can be put in bijective correspondence with $\mathbb{R}$.

Take $E_0 = [0, 1]$. Construct E_1 by removing the open middle third of E_0:

$$E_1 = [0, \tfrac{1}{3}] \cup [\tfrac{2}{3}, 1].$$

Construct E_2 by removing the open middle third of each of the closed intervals that make up E_1:

$$E_2 = [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}] \cup [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{8}{9}, 1].$$

We continue this process, obtaining a sequence of compact sets $(E_j)_{j=1}^{\infty}$, such that $E_j \supset E_{j+1}$ for each j, and E_j is the disjoint union of 2^j closed intervals of length 3^{-j}. In particular, each E_j is nonempty. The set

$$P = \bigcap_{j=1}^{\infty} E_j$$

is called the *Cantor set*. Note that since the E_j's are nested nonempty compact sets, $\{E_j\}$ has the finite intersection property, so P is nonempty and compact.

We claim that P contains no segment. Indeed, each interval of the form

$$I_{k,m} = \left(\frac{3k+1}{3^m}, \frac{3k+2}{3^m}\right)$$

is disjoint from E_m, therefore from P. On the other hand, any open interval contains some $I_{k,m}$. Indeed, choose any $a, b \in \mathbb{R}$ with $a < b$. We want to ensure that m is large enough so that it is possible to pick k satisfying

$$a < \frac{3k+1}{3^m} \quad \text{and} \quad \frac{3k+2}{3^m} < b.$$

These two inequalities yield

$$\frac{3^m a - 1}{3} < k < \frac{3^m b - 2}{3}.$$

In order to ensure that it is possible to pick an *integer* k satisfying both of these inequalities, we pick m large enough so that the LHS and RHS are distance greater than 1 apart. That is,

$$\frac{3^m b - 2}{3} - \frac{3^m a - 1}{3} > 1.$$

Or simplifying, $3^m(b-a) > 4$, i.e. $3^{-m} < \frac{b-a}{4}$. It follows from this construction that $I_{k,m} \in (a, b)$.

Since any interval (a, b) contains an interval of the form $I_{k,m}$, yet $I_{k,m}$ is disjoint from P, it follows that P contains no segment.

Next, we prove that P contains no isolated points, which implies that P is perfect (since we know P is closed). Choose $x \in P$, then choose $\varepsilon > 0$. Since $x \in P$, we know by definition of P that $x \in E_j$ for all $j \in \mathbb{N}$; let I_j denote the closed interval of E_j that contains x, and choose j large enough so that $I_j \subset (x-\varepsilon, x+\varepsilon) = B(x, \varepsilon)$. Let x_j be an endpoint of I_j, with $x_j \neq x$. Then $x_j \in P$ (the endpoints of the closed intervals in E_j are never removed from P), and $x_j \in I_j \subset B(x, \varepsilon)$. Thus x is a limit point of P. Since x was chosen arbitrarily, we conclude that P is perfect.

6. Connectedness

Connectedness is a crucial notion in the analysis of metric spaces, and more generally, topological spaces. However, the general definition is a bit abstract, so we preface it with a discussion of the notion of convexity, which is sort of a concrete prototype for connectedness. Unfortunately, convexity does not make sense in general metric spaces without some extra structure. Therefore we make this discussion in the setting of a real normed vector space.

6.1. Convexity in Real Normed Vector Spaces.

DEFINITION 6.1. Let V be a real vector space, and let E be a subset of V. We say that E is *convex* if whenever $a, b \in E$ and $t \in (0, 1)$, one has $(1-t)a + tb \in E$ as well.

Informally, this definition says that given any two points a and b in E, one can connect a and b with a line segment which remains in E. (Explanation: $p(t) = (1-t)a + tb$, $t \in [0, 1]$, is a parametrization of the line segment joining a and b.)

EXAMPLE 6.2. In any *normed* vector space $(V, \|\cdot\|)$, if we let $d(x, y) = \|x - y\|$, then $B_{V,d}(x, r)$ is convex. Indeed, pick $a, b \in B(x, r)$ and choose $t \in (0, 1)$. We need to show that $z = (1-t)a + tb$ is distance less than r from x. To see this, we write $x = (1-t)x + tx$, so that

$$\begin{aligned}\|x - z\| &= \|[(1-t)x + tx] - [(1-t)a + tb]\| \\ &= \|(1-t)(x-a) + t(x-b)\| \\ &\leq (1-t)\|x-a\| + t\|x-b\| < (1-t)r + tr = r.\end{aligned}$$

EXERCISE 6.1. Let $\mathcal{A}$ be a collection of convex subsets of a real vector space V. Show that $B := \bigcap_{A\in\mathcal{A}} A$ is convex.

6.2. Connectedness in Metric Spaces; Connected Subsets of $\mathbb{R}$.

DEFINITION 6.3. Let (X, d) be a metric space, and let A and B be subsets of X. A and B are said to be *separated* in X if $\mathrm{Cl}_X(A) \cap B$ and $A \cap \mathrm{Cl}_X(B)$ are both empty. That is, A and B are separated in X if they are disjoint, and neither contains a limit point of the other. A set $E \subset X$ is called *connected* if it cannot be written as a union of two nonempty sets which are separated in X. Equivalently, E is connected if and only if the following statement holds: Whenever A is nonempty and A and B are separated sets whose union is E, it follows that $A = E$ and $B = \emptyset$.

REMARK 6.4. In order for two sets A and B to be separated, they must be disjoint. But *disjointness is not enough to guarantee separatedness!* See the following Example.

EXAMPLE 6.5. If $A = (0, 1)$, $B = (1, 2)$, and $C = [1, 2)$, then A and B are separated, while A and C are not separated (even though A and C are disjoint).

EXERCISE 6.2. Let (X, d) be a metric space and let A and B be disjoint subsets of X. Prove that if A and B are both open in X, then A and B are separated.

EXERCISE 6.3. Let E be a connected subset of a metric space (X, d). Show that $\overline{E}$ is connected.

It turns out that connectedness and convexity are equivalent in $\mathbb{R}$. Furthermore, the only connected subsets of $\mathbb{R}$ are the intervals, the rays, the single point sets, and the empty set. We show the equivalence of convexity and connectedness in $\mathbb{R}$ after a Lemma about real numbers.

LEMMA 6.6. *Let $A \subset \mathbb{R}$ be nonempty and bounded above. Then $\alpha := \sup A \in \overline{A}$.*

PROOF. (By contradiction.) If $\alpha \notin \overline{A}$, then there exists $\varepsilon > 0$ so that $B(\alpha, \varepsilon)$ contains no points of A; this implies that $\alpha - \varepsilon$ is an upper bound for A, contradicting the fact that α is the *least* upper bound. □

THEOREM 6.7. *Let E be a subset of $\mathbb{R}$. Then E is connected if and only if whenever $x, y \in E$ and $x < z < y$, one has $z \in E$ as well.*

Note that the second condition in the Theorem is equivalent to saying that E is convex. (Indeed, $z = (1 - t)x + ty$ if $t = \frac{z-a}{b-a}$).

PROOF. We prove the contrapositive statement in each direction. ($\Longrightarrow$) Suppose the second condition fails, i.e. there exists $x, y \in E$ and $z \in \mathbb{R}$ such that $x < z < y$ but $z \notin E$. Put $A = (-\infty, z) \cap E$ and $B = (z, \infty) \cap E$. Then A and B are nonempty, since $x \in A$ and $y \in B$, and $A \cup B = E$. Furthermore, $\overline{A} \cap B \subset (-\infty, z] \cap (z, \infty) = \emptyset$ and similarly $A \cap \overline{B}$ is empty. Therefore A and B are separated, so E is not connected.

($\Longleftarrow$) Next, suppose that E is not connected; pick nonempty sets A and B such that $A \cup B = E$ but $\overline{A} \cap B = A \cap \overline{B} = \emptyset$. Choose $x \in A$ and $y \in B$; assume without loss of generality that $x < y$. We find an element of (x, y) that does not belong to E, proving that the second condition in the statement of the Theorem must fail.

Put $w = \sup(A \cap [x, y])$. Then by the Lemma, $w \in \overline{A \cap [x, y]} \subset \overline{A}$, so that $w \notin B$. Further, $w \neq y$ because $w \in B$. Thus $x \leq w < y$. If additionally $w \notin A$, then we can put $z = w$ and we are done, since this will imply that $z \neq x$ (so $x < z < y$) and $z \notin A \cup B = E$. Therefore we assume without loss of generality that $w \in A$.

Since $w \in A$, we know $w \notin \overline{B}$, since A and B are separated. Since $\overline{B}$ is closed, we can find an ε-neighborhood $B(w, \varepsilon) = (w - \varepsilon, w + \varepsilon)$ of z which is disjoint from $\overline{B}$. Choose $z \in (w, w + \varepsilon)$; then $z + \varepsilon \leq y$ (otherwise $y \in B(w, \varepsilon) \cap B$, a contradiction), so $z < y$. We claim that also $z \notin A$; to see this, we argue by contradiction. Suppose $z \in A$. Then since we already know $z \in [x, y)$, it follows that

$z \in A \cap [x,y]$, so $z \leq \sup(A \cap [x,y]) = w$. On the other hand, $z > w$ by construction, giving us a contradiction. We conclude that $z \notin A$, proving the auxiliary claim. It follows that $z \neq x$ (as $x \in A$), so $x < z < y$ (by combining $z \neq x$ and $z \in [x,y)$). Since also $z \notin A \cup B = E$, this finishes the proof. □

As noted above, the previous Theorem says that convexity and connectedness are equivalent in $\mathbb{R}$ (under the Euclidean metric). It is in fact true that convexity implies connectedness in $\mathbb{R}^n$ for any $n \in \mathbb{N}$. We will prove this in the next Chapter using the notion of *path connectedness*, which lies 'in between' convexity and connectedness in the context of $\mathbb{R}^n$. However, connectedness does *not* imply convexity in $\mathbb{R}^n$ for $n \geq 2$. For example, the circle $S = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ is connected, but it is not convex. To see that it is not convex, note that $a := (1,0) \in S$ and $b := (-1,0) \in S$, but $O = (0,0) \notin S$, even though $O = (1 - \frac{1}{2})a + \frac{1}{2}b$. We will soon give an easy way to show that this set is in fact connected.

THEOREM 6.8. *Let (X,d) be a metric space, and let E be a connected subset of X. If A and B are separated sets in X and $A \cup B \supset E$, then either $E \subset A$ or $E \subset B$.*

PROOF. Put $C = A \cap E$ and $D = B \cap E$. Then $E = C \cup D$, and $\overline{C} \cap D \subset \overline{A} \cap B = \emptyset$ and similarly $C \cap \overline{D} = \emptyset$. Thus C and D are separated. Since E is connected, it follows that either C or D must be empty; that is, one of C or D must contain all of E. □

EXERCISE 6.4. Let (X,d) be a metric space, and let $\mathcal{C}$ be a collection of connected subsets of X. Assume $A = \bigcap_{C \in \mathcal{C}} C$ is nonempty. Show that $B = \bigcup_{C \in \mathcal{C}} C$ is connected.

EXERCISE 6.5. Let $X = \mathbb{R}^2$. Give an example of a connected subset E of X, such that $\mathrm{Int}_X(E)$ is *not* connected. Prove both that your set E is connected and that its interior is not. (Hint: Consider the union of two convex sets joined at a point. You may assume without proof the fact that convexity implies connectedness in $\mathbb{R}^2$.)

To end this section, we give one final characterization of connectedness which is sometimes useful.

PROPOSITION 6.9. *Let (X,d) be a metric space. Then X is connected if and only if the only subsets of X which are both open and closed in X are $\emptyset$ and X.*

PROOF. Assume that X is connected and let E be a subset of X which is both open and closed in X. Then E and $X \backslash E$ are both open in X, therefore they are separated (Exercise 6.2). Since X is connected, it follows that E is either X or $\emptyset$.

On the other hand, assume that X is not connected. Then X can be written as a union of two nonempty separated sets A and B. Separatedness of A and B implies $A \cap \overline{B} = \emptyset$, so $A \subset X \backslash \overline{B}$. On the other hand, $A \cup B = X$, so $X \backslash B \subset A$. Therefore $A = X \backslash \overline{B} = X \backslash B$. It follows that A is open in X, since $\overline{B}$ is closed in X. Reasoning entirely similarly for B, we conclude that B is open in X as well. So A and B are subsets of X, neither of which is $\emptyset$ or X, and both of which are open and closed in X. □

CHAPTER 5

Functions Between Metric Spaces

1. Limits of Functions

1.1. Definition of the Limit.

DEFINITION 1.1. Let (X, d_X) and (Y, d_Y) be metric spaces, and let E be a subset of X. Let $f : E \to Y$ be a function, and let p be a limit point of E in X. We write

$$\lim_{x \to p} f(x) = q$$

to mean the following: For every neighborhood V of q in Y, there exists a neighborhood U of p in X such that $x \in U \cap E \backslash \{p\}$ implies $f(x) \in V$. In this case we say that q is the *limit* of the function $f(x)$ as x approaches p in E.

Note that another common way to write the statement $\lim_{x \to p} f(x) = q$ is '$f(x) \to q$ in Y as $x \to p$ in X', or simply '$f(x) \to q$ as $x \to p$', if X and Y are understood.

Of course, a point q as in the definition of the limit may or may not exist, depending on the function f and the point p. Therefore we sometimes say that 'the limit exists' or 'the limit does not exist', or 'f has/does not have a limit at p'. There are many other ways to express this notion; most are self-explanatory.

REMARK 1.2. The above definition is general enough to work for a topological space—just substitute $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ for (X, d_X) and (Y, d_Y), where $\mathcal{T}_X$ and $\mathcal{T}_Y$ are topologies on X and Y, respectively.

REMARK 1.3. We have already been using the notation

$$\lim_{n \to \infty} x_n = x$$

for a sequence $(x_n)_{n=1}^{\infty}$ that converges to x. This is completely consistent with the definition we have just given: $(x_n)_{n=1}^{\infty}$ is, after all, a *function* from $\mathbb{N}$ to a metric space. Here $\mathbb{N}$ is considered as a subset of $(\overline{\mathbb{R}}, \overline{d})$—or $(\overline{\mathbb{R}}, \overline{\mathcal{T}})$, if you prefer—where $\overline{d}$ and $\overline{\mathcal{T}}$ are defined in Chapter 3. Since $\mathbb{N}$ is not bounded above, $+\infty$ is a limit point of $\mathbb{N}$ in $\overline{\mathbb{R}}$, by Exercise 1.3 in Chapter 4.

You are probably accustomed to seeing an ε-δ version of Definition 1.1. In metric spaces, the two notions are equivalent. Both say that by considering x close enough to p (but not equal to p, and also in the domain of f), you can make $f(x)$ as close as you like to q. The only difference is what 'close enough' means a priori.

EXERCISE 1.1. Let (X, d_X) and (Y, d_Y) be metric spaces, and let E be a subset of X. Let $f : E \to Y$ be a function, and let p be a limit point of E in X. Prove that $f(x) \to q$ as $x \to p$ if and only if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $x \in E$ and $0 < d_X(x, p) < \delta$ imply together that $d_Y(f(x), q) < \varepsilon$.

The ε-δ formulation is often more useful in practice for proving that a given limit holds. Note that in either formulation, the function f may or may not be defined at p (i.e., p may or may not be in E). Furthermore, even if $p \in E$, it may be the case that $f(p) \neq q$.

As remarked above, limits of sequences are a special case of Definition 1.1. It turns out that in the context of metric spaces, Definition 1.1 can actually be recast entirely in terms of limits of sequences.

PROPOSITION 1.4. *Let (X, d_X) and (Y, d_Y) be metric spaces. Let E be a subset of X, and let p be a limit point of E in X. The following are equivalent:*

(1) $\lim_{x\to p} f(x) = q$.
(2) $\lim_{n\to\infty} f(x_n) = q$ *for any sequence* $(x_n)_{n=1}^{\infty}$ *in* $E\backslash\{p\}$ *that converges to* p *in* X.

Note that the assumption that p is a limit point of E guarantees that at least one sequence of the type considered in the second condition must exist.

PROOF. (1) $\implies$ (2). Assume $\lim_{x\to p} f(x) = q$ and let $(x_n)_{n=1}^{\infty}$ be a sequence in $E\backslash\{p\}$ that converges to p in X. Let V be a neighborhood of q in Y, and let U be the corresponding neighborhood of p in X such that $x \in U \cap E\backslash\{p\}$ implies $f(x) \in V$. Choose N large enough so that $n \geq N$ implies $x_n \in U$; then $n \geq N$ also implies that $f(x_n) \in V$. Thus $f(x_n) \to q$ as $n \to \infty$.

(2) $\implies$ (1). Assume (1) fails; we show (2) must as well. The failure of (1) implies that there exists a neighborhood V of q in Y such that for any neighborhood U of p in E, there exists $x \in U\backslash\{p\}$ with $f(x) \notin V$. Assume a neighborhood V is so chosen, and for each $n \in \mathbb{N}$, let x_n be a point of $B_E(p, \frac{1}{n})\backslash\{p\}$ such that $f(x_n) \notin V$. Then $d(x_n, p) < \frac{1}{n}$ for each n, so $x_n \to p$ as $n \to \infty$, but $(f(x_n))$ does not converge to q, since its image lies entirely outside the neighborhood V of q. □

COROLLARY 1.5 (Uniqueness of Limits of Functions). *Let (X, d_X) and (Y, d_Y) be metric spaces. Let E be a subset of X, and let p be a limit point of E in X. Let $f : E \to Y$ be a function. If* $\lim_{x\to p} f(x) = q$ *and* $\lim_{x\to p} f(x) = r$, *then* $q = r$.

The proof of this statement follows immediately from the sequence formulation of limits, together with the fact that limits of sequences are unique in metric spaces.

1.2. Some Limit Rules for Real-Valued Functions. The limit concept is compatible with the commutative algebra structure on $\mathbb{R}^X$ (the set of real-valued functions on the set X) when X is given a metric space structure.

THEOREM 1.6. *Let (X, d) be a metric space. Let E be a subset of X, and let $f : E \to \mathbb{R}$ and $g : E \to \mathbb{R}$ be functions. Assume that p is a limit point of E, and* $\lim_{x\to p} f(x)$, $\lim_{x\to p} g(x)$ *both exist. Then*

$$\lim_{x\to p}(f+g)(x) = \lim_{x\to p} f(x) + \lim_{x\to p} g(x),$$
$$\lim_{x\to p}(fg)(x) = \lim_{x\to p} f(x) \lim_{x\to p} g(x).$$

If additionally $\lim_{x\to p} f(x) \neq 0$, *then*

$$\lim_{x\to p}\left(\frac{1}{f}\right)(x) = \frac{1}{\lim_{x\to p} f(x)}.$$

PROOF. We prove only the statement on multiplication of limits; the others are similar. We write

$$\lim_{x\to p} f(x) = A; \qquad \lim_{x\to p} g(x) = B.$$

Choose $\varepsilon > 0$. We write

$$\begin{aligned}|(fg)(x) - AB| &= |(f(x) - A)g(x) + A(g(x) - B)| \\ &\leq |f(x) - A||g(x)| + |A||g(x) - B| \\ &\leq |f(x) - A|\big(|g(x) - B| + |B|\big) + |A||g(x) - B|.\end{aligned}$$

We want to make each of the terms on the right side less than $\frac{\varepsilon}{2}$. To accomplish this, we choose $\delta > 0$ small enough so that $|g(x) - B| < \min\{\frac{\varepsilon}{2|A|}, 1\}$ and $|f(x) - A| < \frac{\varepsilon}{2(1+|B|)}$, for all $x \in E$ such that

$0 < d(x,p) < \delta$. Then for such x, we have

$$\begin{aligned} |(fg)(x) - AB| &\leq |f(x) - A|\big(|g(x) - B| + |B|\big) + |A||g(x) - B| \\ &< \frac{\varepsilon}{2(1+|B|)} \cdot (1 + |B|) + |A| \cdot \frac{\varepsilon}{2|A|} = \varepsilon. \end{aligned}$$

□

2. Continuous Functions

2.1. Definition of Continuity.

DEFINITION 2.1. Let (X, d_X) and (Y, d_Y) be metric spaces; let $f : X \to Y$ be a function. We say that f is *continuous* at $p \in X$ if for every neighborhood V of $f(p)$, there exists a neighborhood U of p such that $x \in U$ implies $f(x) \in V$.

Once again, this is equivalent to the ε-δ formulation in metric spaces.

EXERCISE 2.1. Let (X, d_X) and (Y, d_Y) be metric spaces; let $f : X \to Y$ be a function. Prove that f is continuous at $p \in X$ if and only if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $x \in B_X(p, \delta)$ implies $f(x) \in B_Y(f(p), \varepsilon)$.

REMARK 2.2. Let (X, d_X) and (Y, d_Y) be metric spaces; let $f : X \to Y$ be a function.

- f must be defined at p in order to be continuous at p.
- If p is not a limit point of X, then f is automatically continuous at p. Indeed, let U be a neighborhood of p that does not intersect $X \backslash \{p\}$. Then p is the only point in U, so trivially, we have $f(x) = f(p) \subset V$ for all $x \in U$ and for any neighborhood V of $f(p)$.
- If p is a limit point of X, then f is continuous at p if and only if $p \in X$ and $\lim_{x \to p} f(x) = f(p)$. This should be clear after comparing the definitions of the two concepts. The only discrepancy is the lack of mention of an auxiliary set E in the definition of continuity at a point; the set E is unnecessary in the present situation because we are assuming explicitly that p is both an element of X and a limit point of X.

REMARK 2.3. We give one more way of looking at continuity at a point before proceeding. $f : X \to Y$ being continuous at $p \in X$ means that for every neighborhood V of $f(p)$, there exists a neighborhood U of p such that $f(U) \subset V$, i.e. $p \in U \subset f^{-1}(V)$. Thus, p is an interior point of $f^{-1}(V)$ for every neighborhood V of $f(p)$. This viewpoint will be useful in what follows.

DEFINITION 2.4. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is said to be *continuous* if it is continuous at every $x \in X$.

PROPOSITION 2.5 (Open Set Formulation of Continuity). *Let (X, d_X) and (Y, d_Y) be metric spaces, and let $f : X \to Y$ be a function. Then f is continuous if and only if $f^{-1}(V)$ is open in X whenever V is open in Y.*

PROOF. Assume f is continuous, and let V be an open set of Y. Choose $x \in f^{-1}(V)$; then V is a neighborhood of $f(x)$ in Y. Since f is continuous at x, it follows so x is an interior point of $f^{-1}(V)$ (see Remark 2.3). Since x was an arbitrary point of $f^{-1}(V)$, we conclude that $f^{-1}(V)$ is open in X.

On the other hand, assume that $f^{-1}(V)$ is open in X whenever V is open in Y. Pick $p \in X$, and let W be a neighborhood of $f(p)$ in Y. Then $U := f^{-1}(W)$ is open in X. Since $f(p) \in W$, we have $p \in U$, and $f(U) \subset W$ by definition of the inverse image. That is, U is a neighborhood of p in X such that $x \in U$ implies $f(x) \in W$. Thus f is continuous at x. □

As you might guess by this point, the open set formulation of continuity can be reduced (in metric spaces) to a statement about open balls in Y:

PROPOSITION 2.6. *Let $f : X \to Y$ be a function between the metric spaces (X, d_X) and (Y, d_Y). Then f is continuous if and only if for every $r > 0$ and every $y \in Y$, the set $f^{-1}(B_Y(y, r))$ is open in X.*

PROOF. If f is continuous, then clearly $f^{-1}(B_Y(y, r))$ is open in X for all $y \in Y$ and $r > 0$, as $B_Y(y, r)$ is an open set in Y. On the other hand, assume that $f^{-1}(B_Y(y, r))$ is open in X for every $y \in Y$, $r > 0$. Let V be an open set in Y and choose $x \in f^{-1}(V)$. Then $f(x)$ is an element of the open set V, so there exists $r > 0$ such that $B_Y(f(x), r) \subset V$. Thus $f^{-1}(B_Y(f(x), r)) \subset f^{-1}(V)$. Since $f^{-1}(B_Y(f(x), r))$ is open and contains x, it follows that x is an interior point of $f^{-1}(B_Y(f(x), r))$ and thus of the larger set $f^{-1}(V)$. As x was an arbitrary point of $f^{-1}(V)$, we conclude that $f^{-1}(V)$ is open in X. □

This Proposition implies in particular that if the codomain of a function is $\mathbb{R}$, then proving continuity amounts to checking the inverse images of all open intervals of $\mathbb{R}$.

REMARK 2.7. The reader might wonder why we trouble ourselves with the open set formulations of these concepts when the ε-δ version is equivalent. We give two (related) responses to this question, though the reasons below are not the only possible justifications.

- When proving that a given function between metric spaces is continuous, it is often easier to check the ε-δ criterion than the open set criterion, because there are strictly fewer things to verify. However, once we know a given function is continuous, the open set formulation gives us strictly more information about our function. Therefore, Theorems involving functions that we know are continuous to begin with are easier to prove when we have access to the open set formulation of continuity. We will see two major examples below, involving the images of compact and connected sets under a continuous function.
- The ε-δ formulation of continuity does not generalize to topological spaces, whereas the open set formulation requires essentially no modification.

EXERCISE 2.2. Assume $f : \mathbb{R} \to \mathbb{R}$ is a function satisfying $\lim_{h\to 0}[f(x + h) - f(x - h)] = 0$, for all $x \in \mathbb{R}$. Does it follow that f must be continuous? If so, give a proof; if not, give a counterexample.

EXERCISE 2.3. Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$ a function.

(a) Show that f is continuous if and only if $f^{-1}(C)$ is closed in X whenever C is closed in Y.
(b) Show that $f : X \to Y$ is continuous if and only if $f(\overline{A}) \subset \overline{f(A)}$ for every subset A of X.
(c) Consider the (continuous) function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = \frac{1}{1+x^2}$. Give an example of a subset A of $\mathbb{R}$ such that $g(\overline{A}) \neq \overline{g(A)}$.

EXERCISE 2.4. Let (X, d_X) and (Y, d_Y) be metric spaces, and let f and g be continuous functions from X to Y. Assume E is a dense subset of X.

(a) Prove that $f(E)$ is dense in $f(X)$. (Hint: Use Exercise 1.12 in Chapter 4 and Exercise 2.3 above.)
(b) Prove that if $f(x) = g(x)$ for all $x \in E$, then $f(x) = g(x)$ for all $x \in X$.

This Exercise shows, for example, that if $f : \mathbb{R} \to \mathbb{R}$ is a continuous function, and if we know what $f(x)$ is for all $x \in \mathbb{Q}$, then we can determine what $f(x)$ is for any $x \in \mathbb{R}$.

2.2. Generating Continuous Functions. We now identify a few ways to obtain continuous functions. Some of these are specific to real-valued functions; others are valid for any metric space. Most of these statements are obvious or nearly so and are presented without proof.

PROPOSITION 2.8 (Constant Functions). *Let (X, d_X) and (Y, d_Y) be metric spaces, and let $f : X \to Y$ be defined by $f(x) = c$ for some $c \in Y$. Then f is continuous.*

PROPOSITION 2.9 (Restricting the Domain or Codomain). *Let (X, d_X) and (Y, d_Y) be metric spaces, and let $f : X \to Y$ be a function.*

- *If f is continuous and E is a subset of X, then $f\big|_E : E \to Y$ is continuous.*
- *If E is a subset of X and $f\big|_E : E \to Y$ is continuous, then f is continuous at each point of $\mathrm{Int}_X(E)$.*
- *If $f(X) \subset Z \subset Y$, then the map $g : X \to Z$ obtained by restricting the codomain of f is continuous if and only if f is continuous.*

PROOF. The first and third statements are trivial to prove using the ε-δ criterion. (Purely open-set proofs are also easy but not completely trivial.) The second statement is a tiny bit subtler; therefore we prove this statement only. Choose $x \in \mathrm{Int}_X(E)$ and $\varepsilon > 0$; let $\delta_1 > 0$ be such that $d_E(x,y) < \delta_1$ implies that $d_Y(f(x), f(y)) < \varepsilon$. Choose $r > 0$ such that $B_X(x,r) \subset E$, and put $\delta_2 = \min\{\delta_1, r\}$. Then $d_X(x,y) < \delta_2$ implies that $y \in B_X(x,r) \subset E$, so $d_E(x,y) = d_X(x,y) < \delta_1$, so $d_Y(f(x), f(y)) < \varepsilon$. Thus f is continuous at x. □

The reader might wonder why it is necessary to write $\mathrm{Int}_X(E)$ in the second statement instead of just E. The problem is illustrated by the first point in the following Example.

EXAMPLE 2.10. Define $f : \mathbb{R} \to \mathbb{R}$ by the rule

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{10}$$

- Put $E = [0, \infty)$. Then $f|_E$ is a constant function, therefore continuous. But f is not continuous at 0. (In particular, f is continuous on $\mathrm{Int}_X(E)$, but not on all of E.)
- The function

$$g = f\big|_{\mathbb{R}\setminus\{0\}} = \begin{cases} 1 & \text{on } (0, +\infty) \\ 0 & \text{on } (-\infty, 0), \end{cases} \tag{11}$$

is, believe it or not, continuous (Exercise 2.5 below). Removing 0 from the domain prevents the function from 'seeing' the jump. You can make better intuitive sense of this if you consider the function

$$h = f\big|_{\mathbb{R}\setminus(0,1)} = \begin{cases} 1 & \text{on } [1, \infty) \\ 0 & \text{on } (-\infty, 0), \end{cases}$$

which shares some common features with g—but it's perhaps easier to believe that h is a function that should be called 'continuous.' One difference between these two functions is that, despite the fact that they are both continuous, h has a continuous extension to all of $\mathbb{R}$ (take $k(x) = f(x)$ on $\mathbb{R}\setminus[0,1)$ and $k(x) = x$ on $[0,1)$), whereas g does not.

EXERCISE 2.5. Consider the functions f and g defined in (10) and (11), respectively.

(a) Prove that g is continuous, using the ε-δ formulation of continuity. (Break it into cases.)
(b) Prove that g is continuous, using the open set formulation of continuity. (Using the formulation of Proposition 2.6 is okay.)
(c) Prove that f is not continuous at 0, using the ε-δ formulation of continuity at a point.
(d) Prove that f is not continuous, using the open set formulation of continuity.
(e) Prove (using whichever method you prefer) that there does not exist *any* continuous extension of g to all of $\mathbb{R}$.

PROPOSITION 2.11 (Identity and Inclusion Maps). *Let (X,d) be a metric space. The identity map $\mathrm{id}_X : X \to X$ is continuous. If $Y \subset X$, then the inclusion map $\iota : Y \to X$ is continuous.*

Note that $\iota : Y \to X$ is the same as $\mathrm{id}_X\big|_Y : Y \to X$. Also note that we assume the same metric is used in the domain and codomain. (This is in line with our usual convention, so we don't include this

in the hypotheses; however, even a function as nice as the identity map can be discontinuous if we use different metrics in the domain and codomain.)

PROPOSITION 2.12 (Composition of Continuous Functions). *Let (X, d_X), (Y, d_Y), and (Z, d_Z) be metric spaces; let $f : X \to Y$ and $g : Y \to Z$ be continuous. Then $g \circ f : X \to Z$ is continuous.*

This proof is easier using the open-set formulation of continuity.

PROOF. Let V be an open set of Z. Then $g^{-1}(V)$ is open in Y by continuity of g; thus $(g \circ f)^{-1}(V) = f^{-1}(g^{-1}(V))$ is open in X, by continuity of f. □

Note: If $f : X \to Y$ and $g : \widetilde{Y} \to Z$ are continuous functions with $Y \subset \widetilde{Y}$, then $g\big|_Y$ is continuous, so the Proposition above still guarantees continuity of $g \circ f$ (which, remember, technically means $(g\big|_Y) \circ f$ according to the definition of composition of functions in Chapter 1).

PROPOSITION 2.13 (Coordinate Functions). *Let ϕ_k denote the function $\phi_k : \mathbb{R}^n \to \mathbb{R}$ given by $\phi_k((x_1, \ldots, x_n)) = x_k$. ($\phi_k$ is called the kth coordinate function on $\mathbb{R}^n$). Then ϕ_k is continuous.*

PROOF. Choose $\varepsilon > 0$, then put $\delta = \varepsilon$. Assume $x, y \in \mathbb{R}^n$ and $|x - y| < \delta$. Write $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. Then

$$|\phi_k(x) - \phi_k(y)| = |x_k - y_k| \leq \sqrt{|(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2|} = |x - y| < \delta = \varepsilon.$$

Thus ϕ_k is continuous. □

PROPOSITION 2.14 (Sums, Products, and Quotients of Continuous Functions). *Let (X, d_X) and (Y, d_Y) be metric spaces. Let $f : X \to Y$ and $g : X \to Y$ be continuous functions. Then $f + g$ and fg are continuous, and $\frac{1}{f}$ is continuous on the smaller set $X \backslash f^{-1}(0)$ where it is defined.*

PROOF. Choose $p \in X$. If p is an isolated point of X, then $f + g$ is automatically continuous at p. Otherwise p is a limit point of X and we have

$$\lim_{x \to p}(f + g)(x) = \lim_{x \to p} f(x) + \lim_{x \to p} g(x) = f(p) + g(p) = (f + g)(p).$$

Note that the first equality above requires the existence of the two limits $\lim_{x \to p} f(x)$ and $\lim_{x \to p} g(x)$ (guaranteed by continuity), and the second equality relies on the continuity of f and g at p. Comparing the LHS and RHS, we see that $f + g$ is continuous at p. This shows that $f + g$ is continuous at every point of X and is therefore continuous. The proof for fg is entirely similar. The proof for $\frac{1}{f}$ is also entirely similar, once we note that $f\big|_{X \backslash f^{-1}(0)}$ is continuous. □

PROPOSITION 2.15 (Polynomials and Rational Functions). *Any polynomial in n variables is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$. If $r = \frac{f}{g}$ is any rational function (a quotient of the polynomials f and g) of n variables, then r is continuous on the set $\mathbb{R}^n \backslash g^{-1}(0)$.*

This Proposition follows from the fact that any polynomial or rational function in n variables can be written entirely in terms of sums, products, and quotients of coordinate functions.

PROPOSITION 2.16 ($\mathbb{R}^n$-valued functions). *Let $f_1, \ldots, f_n$ be real-valued functions on a metric space (X, d). The function $f : X \to \mathbb{R}^n$ defined by $f(x) = (f_1(x), \ldots, f_n(x))$ for all $x \in X$ is continuous if and only if each of the functions f_j is continuous.*

PROOF. Assume f is continuous. Then $f_j = \phi_j \circ f$ is also continuous. Assume conversely that each of the f_j's is continuous. Choose $\varepsilon > 0$ and $x \in X$, and let $\delta > 0$ be such that $|f_j(x) - f_j(y)| < \frac{\varepsilon}{\sqrt{n}}$ whenever $d(x, y) < \delta$, for each $j \in J_n$. Then for such y, we have

$$|f(x) - f(y)| = \sqrt{(f_1(x) - f_1(y))^2 + \cdots + (f_n(x) - f_n(y))^2} < \sqrt{\frac{\varepsilon^2}{n} + \cdots + \frac{\varepsilon^2}{n}} = \varepsilon.$$

□

2.3. Continuity and Compactness.

THEOREM 2.17 (Continuous Image of a Compact Set is Compact). *Let (X, d_X) and (Y, d_Y) be metric spaces. Let $f : X \to Y$ be a continuous function, and let E be a compact subset of X. Then $f(E)$ is compact.*

PROOF. Let $\mathcal{B}$ be an open cover of $f(E)$ in Y. We claim that $\mathcal{A} = \{f^{-1}(B)\}_{B\in\mathcal{B}}$ is an open cover of E in X. Indeed, each $f^{-1}(B)$ is open in X by continuity; furthermore, if $x \in E$, then $f(x) \in f(E)$, so $f(x) \in B_0$ for some $B_0 \in \mathcal{B}$, so $x \in f^{-1}(B_0)$. Let $\mathcal{A}' = \{f^{-1}(B_1), \ldots, f^{-1}(B_n)\}$ be a finite subcover for E associated to the open covering $\mathcal{A}$. Then

$$E \subset \bigcup_{j=1}^{n} f^{-1}(B_j)$$

implies

$$f(E) \subset f\left(\bigcup_{j=1}^{n} f^{-1}(B_j)\right) = \bigcup_{j=1}^{n} f(f^{-1}(B_j)) \subset \bigcup_{j=1}^{n} B_j.$$

Thus $\mathcal{B}' = \{B_1, \ldots, B_n\}$ is then a finite subcovering of $f(E)$ associated to the covering $\mathcal{B}$. Since $\mathcal{B}$ was an arbitrary open cover of $f(E)$, we conclude that $f(E)$ is compact. □

COROLLARY 2.18 (Extreme Value Theorem). *Let $f : X \to \mathbb{R}$ be a continuous function and let E be a compact subset of X. Then there exist $x, y \in X$ such that*

$$f(x) = \sup f(E), \qquad f(y) = \inf f(E).$$

PROOF. Since $f(E)$ is compact, it is closed and bounded in $\mathbb{R}$. Therefore $\sup f(E) \in \overline{f(E)} = f(E)$. This shows that there exists $x \in E$ such that $f(x) = \sup f(E)$. The other claim follows in the same way. □

EXERCISE 2.6. Complete the following tasks.

(a) Find a closed subset E of $\mathbb{R}$ and a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that $f(E)$ is not closed.
(b) Find a bounded subset E of $\mathbb{R}$ and a continuous function $f : E \to \mathbb{R}$ such that $f(E)$ is not bounded.
(c) Show that if E is a bounded subset of $\mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $f(E)$ is bounded.

THEOREM 2.19 (Continuity of Inverse). *Let X and Y be metric spaces, with X compact, and let $f : X \to Y$ be a continuous bijection. Then $f^{-1} : Y \to X$ is a continuous function.*

PROOF. Let C be a closed subset of X; we show that $(f^{-1})^{-1}(C) = f(C)$ is closed in Y. (This suffices by Exercise 2.3.) As C is a closed subset of the compact set X, it follows that C is compact. Therefore $f(C)$ is compact, hence closed in Y. □

Note that the utility of the Theorem above is not limited to the context of compact sets. Here is an example:

PROPOSITION 2.20. *The function $f : [0, \infty) \to \mathbb{R}$ given by $f(x) = x^{\frac{1}{n}}$ is continuous for any $n \in \mathbb{N}$.*

PROOF. Choose $x \in [0, \infty)$. Then f is continuous at x if and only if the function $\widetilde{f} : [0, x + 1] \to [0, (x + 1)^{\frac{1}{n}}]$ is continuous at x. (See Proposition 2.9.) On the other hand, $g : [0, (x + 1)^{\frac{1}{n}}] \to [0, x + 1]$ given by $g(x) = x^n$ is a continuous function on the compact set $[0, x + 1]$, whose inverse is $\widetilde{f}$. So by Theorem 2.19, it follows that $\widetilde{f}$ is continuous (in particular, at x), so that f is continuous at x. Since x was arbitrary, we conclude that f is continuous. □