COLUMBIA UNIVERSITY

Department of Mathematics

Notes in Modern Analysis

Based on the Lectures of Andre Carneiro

MATH S4061

Introduction to Modern Analysis I

Summer 2026

Student

Tyler Sotomayor

tyler.sotomayor@columbia.edu

Instructor

Andre Carneiro

arc2142@columbia.edu

This version compiled June 26, 2026[1]

[1]For latest version: https://tylersotomayor.github.io/real-anal/. Please report any errata to tyler.sotomayor@columbia.edu.

Contents

Modes of Argument

Each proof in these notes is tagged, in its caption, by its dominant *mode of argument*—the kind of reasoning it turns on. The letter A abbreviates *argument*; a proof that blends two is written, e.g., [A2+A4]. The tags are keyed below, and will eventually link to a chapter on proof-writing.

A1 ε-of-room (forcing): a nonnegative quantity below every $\varepsilon > 0$ must be 0.

A2 ε–N: convergence of a sequence—given ε, produce a threshold N.

A3 ε–δ: limits and continuity of functions.

A4 Triangle inequality, with an added-and-subtracted middle term.

A5 Supremum / infimum (least-upper-bound) argument.

A6 Open cover $\to$ finite subcover (compactness).

A7 Bisection / nested intervals closing on a point, via completeness.

A8 Subsequence extraction and diagonalisation.

A9 Contradiction, or the contrapositive.

A10 Induction.

A11 Negating a quantifier to unfold the failure of a definition.

A12 Cases, or reduction "without loss of generality."

Constr Explicit construction of the required object.

Direct A direct chain of implications, with no special device.

PART I

Lecture Notes

COURSE PRELIMINARIES

Foundations: Logic, Sets, Functions, and the Craft of Proof

MATH S4061 | Introduction to Modern Analysis I

Logical and notational prerequisites — assuming no prior experience with proofs

The only way to learn mathematics is to do mathematics.

Paul R. Halmos

Contents of these preliminaries

Overview. These preliminaries are the on-ramp to the course. They assume you have never written a proof and have never met set notation, and they build—slowly and from the ground up—everything the lectures take for granted: how to read and write mathematical statements (0.1), the quantifiers that power every definition in analysis, the language of sets, relations, and functions, and, at its centre, the *craft of proof*: how a proof is built, the handful of standard strategies (which are exactly the Modes of Argument tagged throughout this book), and how to find one when you are stuck. We close with a worked, classified problem set. Read these preliminaries actively, with pencil in hand; nothing here is meant to be watched go by.

0.1 The Language of Mathematics

Mathematics is written in declarative sentences that are either true or false, and a proof is a chain of such sentences in which each link is forced by the ones before it. Before we can build or follow such a chain, we have to be able to read and write its links *exactly*. That is the whole purpose of this first section: not to prove anything deep, but to fix the grammar—statements and the words "not", "and", "or", "if… then", and "if and only if"—so precisely that later there is never any doubt about what a theorem claims or what a proof must deliver.

Definition 0.1.1

Statement

A *statement* (or *proposition*) is a declarative sentence that is definitely true or definitely false—never both, and never neither. Its *truth value* is T (true) or F (false).

Example 0.1.2

What is and isn't a statement

"$2+2=4$" is a statement (true); "7 is even" is a statement (false). But "$x+1=3$" is *not* a statement as it stands—its truth depends on the unspecified x, so it has no fixed truth value; such a sentence is called an *open sentence*, and it becomes a statement only once x is pinned down or quantified (0.2). "This sentence is false" is not a statement either: assuming it true makes it false and vice versa, so it has no consistent truth value. Questions and commands ("Is 7 prime?", "Add 3.") are never statements.

We build complicated statements from simple ones using five *connectives*. The first three are the everyday words *not*, *and*, *or*; we record their meaning in a *truth table*, which simply lists the truth value of the compound for every combination of truth values of its parts.

Definition 0.1.3

Negation, conjunction, disjunction

Let P and Q be statements.

- the *negation* $\neg P$ ("not P") is true exactly when P is false;
- the *conjunction* $P \wedge Q$ ("P and Q") is true exactly when both are true;
- the *disjunction* $P \vee Q$ ("P or Q") is true exactly when at least one is true.

P	$\neg P$
T	F
F	T

P	Q	$P \wedge Q$	$P \vee Q$
T	T	T	T
T	F	F	T
F	T	F	T
F	F	F	F

Remark 0.1.4

"Or" is inclusive

In ordinary speech "or" often means one or the other but not both ("soup or salad"). In mathematics "or" is always *inclusive*: $P \vee Q$ is true when P is true, when Q is true, *and* when both are true—only

the last row of the table, both false, makes it false. When we genuinely mean "exactly one," we say so.

The fourth connective, *implication*, is the engine of every theorem, and it is the one beginners find most surprising. An implication does not assert that its hypothesis holds; it only promises a *link*—that whenever the hypothesis holds, the conclusion does too.

Definition 0.1.5

Implication

The *implication* $P \Rightarrow Q$ ("if P then Q", or "P implies Q") is false in exactly one situation: when P is true and Q is false. In every other case it is true.

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

Here P is the *hypothesis* (or *antecedent*) and Q the *conclusion* (or *consequent*).

Remark 0.1.6

Why a false hypothesis makes the implication true (vacuous truth)

The last two rows of the table are the ones to dwell on: when P is false, $P \Rightarrow Q$ is counted *true* no matter what Q is. This is called *vacuous truth*, and it is forced on us by what an implication promises. Read $P \Rightarrow Q$ as a promise: "whenever P holds, Q holds." The only way to *break* such a promise is to exhibit a case where P holds yet Q fails—the second row. If P never holds, the promise is never put to the test, so it cannot have been broken, and we call it true by default. A promise like "if you score 100, I will buy you dinner" is not a lie on a day you scored 60; it simply was not triggered. So "if $0 = 1$, then I am the Pope" is a true statement: its hypothesis is false, so the implication holds vacuously—it tells you nothing about the Pope. We will lean on vacuous truth constantly; for instance, "every element of the empty set is even" is true precisely because there is no element to check (0.3).

From an implication $P \Rightarrow Q$ we can form three close relatives by swapping and negating. Two of them are traps for the unwary.

Definition 0.1.7

Converse, contrapositive, inverse

Given $P \Rightarrow Q$, its

- *converse* is $Q \Rightarrow P$;
- *contrapositive* is $\neg Q \Rightarrow \neg P$;
- *inverse* is $\neg P \Rightarrow \neg Q$.

An implication and its contrapositive always have the same truth value—they are *logically equivalent*, and proving one proves the other (this is the basis of *contrapositive proof*, 0.5).

An implication and its converse are *not* equivalent: the converse may be true or false independently, and confusing the two is among the most common errors in a first proof course.

Example 0.1.8

The converse can fail

Let P be "n is divisible by 4" and Q be "n is even." Then $P \Rightarrow Q$ is true (a multiple of 4 is certainly even). Its converse $Q \Rightarrow P$—"every even number is divisible by 4"—is false ($n = 6$ is even but not a multiple of 4). Its contrapositive "if n is odd then n is not divisible by 4" is true, as it must be, being equivalent to the original. Proving $P \Rightarrow Q$ tells you nothing about whether $Q \Rightarrow P$ holds; that is a separate question.

The fifth connective ties an implication to its converse.

Definition 0.1.9

Biconditional; necessary and sufficient

The *biconditional* $P \Leftrightarrow Q$ ("P if and only if Q", abbreviated "P iff Q") is true exactly when P and Q have the same truth value. It is the conjunction of an implication and its converse: $P \Leftrightarrow Q$ means $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$. Correspondingly one says P is *sufficient* for Q (because $P \Rightarrow Q$) and P is *necessary* for Q (because $Q \Rightarrow P$, i.e. Q cannot hold without P). Proving a biconditional therefore requires proving *both* directions, usually labelled ($\Rightarrow$) and ($\Leftarrow$).

0.2 Quantifiers

In 0.1.2 we met *open sentences* such as "$x + 1 = 3$": sentences that are neither true nor false until the variable is pinned down. *Quantifiers* are the two phrases that pin a variable down by declaring *how many* of its values make the sentence true—"for all" and "there exists." They are not a side topic. Every definition in this course—convergence, continuity, the least-upper-bound property—is a short tower of stacked quantifiers, and the single most useful skill these preliminaries can give you is reading, building, and *negating* such towers fluently.

Definition 0.2.1

Predicate and domain of discourse

A *predicate* (or *open sentence*) $P(x)$ is a sentence containing a variable x that becomes a statement once x is assigned a value from a specified set D, called the *domain* (or *universe*) *of discourse*. For instance, with $P(x)$ the sentence "$x^2 = 2$" on the domain $D = \mathbb{R}$, the value $x = \sqrt{2}$ makes $P(x)$ true and $x = 1$ makes it false. A predicate may involve several variables, $P(x, y)$, each ranging over its own domain.

Definition 0.2.2

The universal and existential quantifiers

Let $P(x)$ be a predicate on a domain D.

- the *universal* statement $\forall x \in D,\ P(x)$ ("for all x in D, $P(x)$") is true exactly when $P(x)$ holds for *every* value of x in D;
- the *existential* statement $\exists x \in D,\ P(x)$ ("there exists x in D such that $P(x)$") is true exactly when $P(x)$ holds for *at least one* value of x in D.

Once quantified, x is a *bound* variable—a placeholder, so that $\forall x\, P(x)$ and $\forall t\, P(t)$ say the very same thing—and the quantified sentence is a genuine statement, with a fixed truth value that no longer depends on x. A variable left unquantified is *free*, and a sentence with a free variable is an open sentence, not a statement (0.1.2).

Example 0.2.3

Reading the quantifiers

Each line below is a complete statement; its truth value follows.

- $\forall x \in \mathbb{R},\ x^2 \geq 0$ (true: a square is never negative);
- $\exists x \in \mathbb{R},\ x^2 = 2$ (true: $x = \sqrt{2}$ is a witness);
- $\exists x \in \mathbb{Q},\ x^2 = 2$ (false: no rational number squares to 2; we prove this in 0.5);
- $\forall n \in \mathbb{N},\ n + 1 > n$ (true).

Over an *empty* domain the conventions follow vacuous truth (0.1.6): $\forall x \in \emptyset$, $P(x)$ is true (there is no x to violate it) and $\exists x \in \emptyset$, $P(x)$ is false (there is no x to witness it).

Remark 0.2.4

Proving versus disproving each quantifier

The two quantifiers call for opposite kinds of work, and recognising which you face is half of finding a proof.

- To *prove* $\forall x \in D,\ P(x)$ you must treat an *arbitrary* x: you write "fix $x \in D$" and argue for that nameless x using nothing special about it—checking examples, however many, never suffices. To *disprove* it you need only one *counterexample*: a single $x \in D$ for which $P(x)$ is false.
- To *prove* $\exists x \in D,\ P(x)$ you exhibit one *witness*: a specific x (often constructed) for which $P(x)$ holds. To *disprove* it you must show that *every* x fails, i.e. prove $\forall x \in D,\ \neg P(x)$.

Universals need an arbitrary element; existentials need a witness. This asymmetry shapes nearly every proof you will write, and we return to it in earnest in 0.5.

The most mechanical—and most valuable—fact about quantifiers is how they behave under negation. It lets you compute, with no cleverness at all, exactly what it means for a definition to *fail*.

Definition 0.2.5

Negating a quantifier

For any predicate P on a domain D,

$$\neg\big(\forall x \in D,\ P(x)\big) \ \equiv\ \exists x \in D,\ \neg P(x), \qquad \neg\big(\exists x \in D,\ P(x)\big) \ \equiv\ \forall x \in D,\ \neg P(x).$$

In words: "not every x satisfies P" means "some x fails P," and "no x satisfies P" means "every x fails P." To negate a whole string of quantifiers, sweep it left to right, turning each $\forall$ into $\exists$ and each $\exists$ into $\forall$, and finally negate the predicate at the end. This rule is the engine of mode A11 (0.5)—unfolding the failure of a definition by negating it term by term.

Example 0.2.6

Negation in action—a preview of convergence

You are not expected to follow the analysis here; watch only the mechanics. The statement "$a_n \to L$" is defined by a tower of three quantifiers,

$$\forall\, \varepsilon > 0,\ \ \exists\, N,\ \ \forall\, n \geq N,\ \ |a_n - L| < \varepsilon.$$

Negating mechanically—flip each quantifier, negate the inside—gives

$$\exists\, \varepsilon > 0,\ \ \forall\, N,\ \ \exists\, n \geq N,\ \ |a_n - L| \geq \varepsilon.$$

In words: a_n *fails* to converge to L when there is some positive ε such that, however far out you place the threshold N, some later term a_n still lies at least ε away from L. The point is that we produced this by a rule, not by insight; when the definitions get hard, this is how you will always know precisely what must be shown.

When quantifiers of different kinds are stacked, their *order* is part of the meaning; swapping a $\forall$ past an $\exists$ can turn a truth into a falsehood.

Definition 0.2.7

Nested quantifiers; order matters

Read a stacked statement left to right, as a dialogue between you and an adversary: each $\forall$ is a value handed to you by the adversary, and each $\exists$ to its right is a value *you* supply, allowed to depend on everything already named. Thus

$$\forall x \in \mathbb{R},\ \exists y \in \mathbb{R},\quad y > x$$

is *true*—handed any x, you answer $y = x + 1$—while the swap

$$\exists y \in \mathbb{R},\ \forall x \in \mathbb{R},\quad y > x$$

is *false*: now you must commit to a single y *before* any x is named, and no fixed real number exceeds every real number. Same four symbols, opposite truth values; only the order differs.

Remark 0.2.8

The dependence is the whole game

In $\forall x\, \exists y$ the witness y may depend on x; in $\exists y\, \forall x$ it may not. Every ε–N and ε–δ definition in this course is an "$\forall\exists$" statement in which the threshold N (or the radius δ) is permitted to depend on ε—and, later, the difference between ordinary and *uniform* convergence is nothing but a deliberate change in that order of dependence. Getting the order right is not pedantry; it is the mathematics.

Definition 0.2.9

Unique existence

The statement $\exists!\, x \in D,\ P(x)$ ("there exists a *unique* x in D such that $P(x)$") abbreviates

$$\exists x \in D,\ \Big(P(x)\ \wedge\ \forall y \in D,\ \big(P(y) \Rightarrow y = x\big)\Big)\,:$$

at least one x satisfies P, and any two that do are equal. Proving it therefore has two separate parts—*existence* (exhibit one x that works) and *uniqueness* (assume $P(x)$ and $P(y)$ both hold and deduce $x = y$).

0.3 Sets

Almost every object in this course—a sequence, a function, an interval, the real line itself—is, underneath, a *set*: a collection of things considered as a single whole. The language of sets is the notation in which every later definition is written, so before convergence or continuity can be stated we need to be able to name a collection, say what belongs to it, compare two collections, and combine them. None of this is deep; what it asks of you is precision, and the one new habit it demands is the pair of proof moves for comparing sets—showing one sits inside another, and showing two are equal—which we introduce here and lean on for the rest of the book.

Definition 0.3.1

Set, element, membership

A *set* is a collection of objects, considered as a single object in its own right. The objects in the collection are its *elements* (or *members*). If x is an element of the set A we write $x \in A$ ("x belongs to A"); if it is not, we write $x \notin A$. A set is determined *entirely* by which objects belong to it—nothing else about it matters, not how its elements are listed, described, or ordered, and not whether any are named twice.

That last sentence is the whole content of the idea, so it is worth saying again plainly: a set is fixed by its membership and by nothing else. Two descriptions that admit exactly the same objects describe the same set, however different they look on the page. We make that principle official in 0.3.5; first we need ways to write sets down.

Definition 0.3.2

Roster and set-builder notation

There are two standard ways to specify a set.

- *Roster (list) notation* names the elements between braces: $\{1, 2, 3\}$ is the set whose members are 1, 2, and 3. Because only membership matters, $\{1, 2, 3\}$, $\{3, 2, 1\}$, and $\{1, 1, 2, 3, 3\}$ are the *same* set—order and repetition carry no information. A one-element set such as $\{x\}$ is called a *singleton*; the braces are not optional, since $\{x\}$ is a *set* whose lone member is x, a different object from x itself. A list may trail off in a clear pattern, as in $\{0, 1, 2, 3, \dots\}$, but only when the pattern is genuinely unambiguous.
- *Set-builder notation* carves a set out of a known domain D using a predicate $P(x)$ (0.2.1):

$$\{\, x \in D : P(x) \,\}$$

  is read "the set of all x in D such that $P(x)$," and its members are exactly those $x \in D$ for which $P(x)$ is true. Thus $\{\, x \in \mathbb{R} : x^2 < 4 \,\}$ is the open interval of reals strictly between -2 and 2. The colon is sometimes written as a vertical bar, $\{x \in D \mid P(x)\}$; the two are interchangeable.

Remark 0.3.3

Set-builder is *a quantifier in disguise*

Set-builder notation is the bridge from 0.2 to this section: "$y \in \{x \in D : P(x)\}$" means nothing more nor less than "$y \in D$ and $P(y)$." Whenever you must decide whether some object lies in a set defined this way, that is the translation to reach for—membership in a set-builder set is a predicate to be checked. Always state the domain D. Writing $\{x : P(x)\}$ with no domain at all invites the trouble of 0.3.18, so in this course every set is built inside one already in hand. (Two later constructions stand apart and rightly need no ambient domain: the power set 0.3.15 gathers *all* subsets of a given set, and the union or intersection of a family 0.3.22 ranges over the family itself.)

One collection is so important, and so easy to overlook, that it gets its own name and symbol: the collection with no members at all.

Definition 0.3.4

The empty set

The *empty set* is the set with no elements; we write it $\varnothing$ (or $\{\,\}$). For every object x, the statement $x \in \varnothing$ is false. There is only *one* empty set: a set is fixed by its members (0.3.1), and any two sets with no members have exactly the same members (namely none), so they are the same set. We therefore speak of *the* empty set.

The empty set is where vacuous truth (0.1.6) earns its keep. "Every element of $\varnothing$ is even" is true—not because we checked, but because there is no element to check, so the claim cannot fail. Any sentence beginning "every element of $\varnothing\ldots$" is true, and any beginning "some element of $\varnothing\ldots$" is false. Hold onto this; it is exactly why the empty set will turn out to sit inside every set whatsoever.

Definition 0.3.5

Equality of sets (extensionality)

Two sets A and B are *equal*, written $A = B$, when they have exactly the same elements: for every object x,

$$x \in A \iff x \in B.$$

This principle—that a set is determined by its members alone—is called *extensionality*.[2] It is the only way two sets are ever shown equal, and it is the reason $\{1, 2, 3\} = \{3, 2, 1\}$.

Equality compares membership in both directions at once. Most of the time we want the one-directional comparison: every member of one set is also a member of another.

Definition 0.3.6

Subset and proper subset

A set A is a *subset* of a set B, written $A \subseteq B$, when every element of A is also an element of B:

$$A \subseteq B \quad \text{means} \quad \forall x,\ (x \in A \Rightarrow x \in B).$$

We also say A is *contained in* B, or B *contains* A. If in addition $A \neq B$—so B has at least one element outside A—we call A a *proper* subset and write $A \subsetneq B$. Comparing the

[2] The name contrasts with an *intensional* view, where the *description* would matter. Sets are extensional: only the extension, the actual collection of members, counts. "Even prime" and "{2}" describe the same set.

definition with 0.3.5 shows the link we will use constantly: $A = B$ holds exactly when $A \subseteq B$ *and* $B \subseteq A$.

Remark 0.3.7

Two proof moves to memorise now

This definition seeds the two most-used techniques in the book, both unfolded properly in 0.5; meet them here so they are familiar when the analysis starts.

- *To prove* $A \subseteq B$. The claim is a universal implication, so you prove it the way 0.2.4 prescribes: *fix* an arbitrary $x \in A$, then argue—using only that $x \in A$—that $x \in B$. The argument must work for a nameless element, never for examples. The skeleton is always "Let $x \in A$. … Hence $x \in B$, so $A \subseteq B$."
- *To prove* $A = B$. Prove $A \subseteq B$ and $B \subseteq A$ separately. This is *double inclusion*, and it is how essentially every set identity in this course is established: show each side contains the other.

Reach for these two moves automatically. The moment a problem mentions $\subseteq$ or $=$ between sets, you already know the first line of your proof.

Proposition 0.3.8

$\varnothing \subseteq A$ and $A \subseteq A$, for every set A

For every set A, the empty set is a subset of A, and A is a subset of itself.

Proof 0.3.8

Proof by Unwinding the Definition of Subset.

(1) To show $\varnothing \subseteq A$, we must show $\forall x,\ (x \in \varnothing \Rightarrow x \in A)$. Fix any x. Its hypothesis $x \in \varnothing$ is false (0.3.4), so the implication $x \in \varnothing \Rightarrow x \in A$ is true vacuously (0.1.6)—there is no x that could break it. As x was arbitrary, $\varnothing \subseteq A$.

(2) To show $A \subseteq A$, fix $x \in A$; then $x \in A$, which is what was to be shown. Hence $A \subseteq A$.

■

[Direct]

With the comparisons in place we name the sets this whole course is built on. Their careful *construction*—building each from the one before, and pinning down what makes the reals complete—is the business of the first lectures; here we only fix names and symbols.

Definition 0.3.9

The standard number sets

We write, throughout the book,

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}, \quad \mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}, \quad \mathbb{Q} = \left\{ \tfrac{p}{q} : p \in \mathbb{Z},\ q \in \mathbb{Z},\ q \neq 0 \right\},$$

the *natural numbers*, the *integers*, and the *rational numbers*. By the convention of this course $0 \in \mathbb{N}$. These nest,

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R},$$

and each inclusion is proper. How $\mathbb{N}$ is founded and induction justified (1.1), how $\mathbb{Z}$ and $\mathbb{Q}$ are constructed (1.2), why $\mathbb{Q}$ leaves gaps that force us to $\mathbb{R}$ (1.4), and how $\mathbb{R}$ is built to fill them (1.6) are all taken up later in the course; for now they are names with which to write examples.

Sets become useful once we can *combine* them. The three basic operations mirror the logical connectives of 0.1 exactly: "and" becomes intersection, "or" becomes union, "not" becomes complement. Read the definitions with that correspondence in mind and they need no memorising.

Definition 0.3.10

Union, intersection, difference, complement

Let A and B be sets, and let everything in sight be drawn from a fixed *universe* (or *universal set*) U. We define

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

the *union* and the *intersection* (the "or" here is inclusive, 0.1.4); the *difference*

$$A \setminus B = \{\, x \in U : x \in A \text{ and } x \notin B \,\}$$

("A minus B," the elements of A left after removing those in B); and, relative to the universe U, the *complement*

$$A^c = U \setminus A = \{\, x \in U : x \notin A \,\}.$$

The complement depends on the chosen U, so U must be fixed before A^c has a meaning.

Example 0.3.11

The operations on small sets

Let $A = \{1,2,3,4\}$ and $B = \{3,4,5\}$ inside the universe $U = \{1,2,3,4,5,6\}$. Then $A \cup B = \{1,2,3,4,5\}$, $A \cap B = \{3,4\}$, $A \setminus B = \{1,2\}$, $B \setminus A = \{5\}$, and $A^c = \{5,6\}$. The difference is one-sided: $A \setminus B \neq B \setminus A$ in general. For a nonexample of intersection, $\{1,2\} \cap \{3,4\} = \varnothing$—two sets can perfectly well share no element, and then their intersection is the empty set.

Definition 0.3.12

Disjoint and pairwise disjoint

Two sets A and B are *disjoint* when $A \cap B = \varnothing$—they share no element. A collection of sets is *pairwise disjoint* when any two *distinct* sets from it are disjoint. Thus $\{1,2\}$, $\{3\}$, and $\{4,5\}$ are pairwise disjoint, while $\{1,2\}$, $\{2,3\}$, $\{4\}$ are not (the first two meet in 2).

A picture fixes all of this at once. A *Venn diagram* draws the universe as a rectangle and each set as a region inside it; the operations are then the regions you shade.

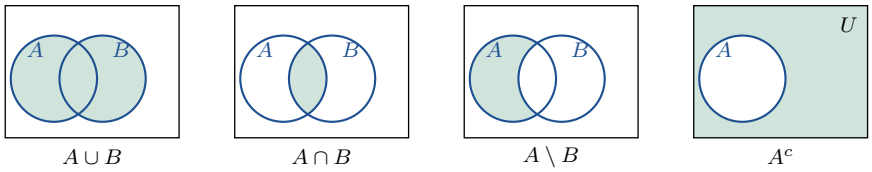


Figure 0.3.13. The four operations as shaded regions. In each panel the universe U is the rectangle and A, B are the discs; the shaded region is the named set. Union shades either disc; intersection, only the overlap; difference $A \setminus B$, the part of A outside B; complement A^c, everything in U outside A.

The correspondence between set operations and logical connectives is not a coincidence to admire and forget; it is a working tool. Every law the connectives obey in 0.1 becomes a law the operations obey, and the proofs are translations. Two families of such laws come up constantly.

Proposition 0.3.14

Distributive and De Morgan laws

For all sets A, B, C, and for A, B inside a universe U,

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

(the distributive laws), and

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

(De Morgan's laws). In words for the last pair: to be outside a union is to be outside both; to be outside an intersection is to be outside at least one. We prove the first De Morgan law below; the first distributive law is worked the same way, by double inclusion, in 0.E.18.

Proof 0.3.14

Proof by Double Inclusion of the First De Morgan Law.

(1) We prove $(A \cup B)^c = A^c \cap B^c$ by showing each side is a subset of the other (0.3.7). For the forward inclusion $\subseteq$, fix $x \in (A \cup B)^c$. Then $x \in U$ and $x \notin A \cup B$, so it is not the case that $x \in A$ or $x \in B$. The negation of an "or" is an "and" of negations, which the truth tables of 0.1.3 make plain, so $x \notin A$ and $x \notin B$. With $x \in U$ this gives $x \in A^c$ and $x \in B^c$, that is,

$$x \in A^c \cap B^c.$$

Hence $(A \cup B)^c \subseteq A^c \cap B^c$.

(2) For the reverse inclusion $\supseteq$, fix $x \in A^c \cap B^c$. Then $x \in U$ with $x \notin A$ and $x \notin B$,

so neither $x \in A$ nor $x \in B$ holds; therefore $x \notin A \cup B$. With $x \in U$ this says

$$x \in (A \cup B)^c,$$

so $A^c \cap B^c \subseteq (A \cup B)^c$.

(3) Both inclusions hold, so by double inclusion the two sets are equal.

■

[Direct]; the other three are identical translations of the truth tables of 0.1.3.

Every set so far has had numbers, or points, for elements. Nothing stops the elements *themselves* from being sets, and one such construction is indispensable: the set of all subsets of a given set.

Definition 0.3.15

Power set

The *power set* of a set S, written $\mathcal{P}(S)$, is the set whose elements are all the subsets of S:

$$\mathcal{P}(S) = \{\, A : A \subseteq S \,\}.$$

Thus "$A \in \mathcal{P}(S)$" means precisely "$A \subseteq S$." Because $\varnothing \subseteq S$ and $S \subseteq S$ for every S (0.3.8), both $\varnothing$ and S are always elements of $\mathcal{P}(S)$.

Example 0.3.16

The power set of a two-element set

For $S = \{a, b\}$ the subsets are: the empty set, the two singletons, and S itself, so

$$\mathcal{P}(\{a, b\}) = \{\ \varnothing,\ \{a\},\ \{b\},\ \{a, b\}\ \},$$

a set with *four* elements. The braces matter: $\{a\}$, the singleton, is an element of $\mathcal{P}(S)$, whereas a is not—a is an element of S, not a subset of it. Counting subsets in general: to build a subset of an n-element set you decide, independently, "in or out" for each of the n elements, which is 2 choices made n times. So a finite set with n elements has exactly 2^n subsets, and $|\mathcal{P}(S)| = 2^{|S|}$. Here $n = 2$ gives $2^2 = 4$, matching the list.

Remark 0.3.17

$\varnothing \in \mathcal{P}(S)$ versus $\varnothing \subseteq S$—a classic confusion

These two true statements look almost identical and mean different things; keeping them apart is a rite of passage. The symbol $\in$ relates an *element* to a set; the symbol $\subseteq$ relates a *set* to a set. Both

$$\varnothing \subseteq S \qquad \text{and} \qquad \varnothing \in \mathcal{P}(S)$$

are true *for every* S, and they are really the same fact wearing two hats: $\varnothing$ is a *subset* of S (0.3.8), and therefore $\varnothing$ is an *element* of the power set, because elements of $\mathcal{P}(S)$ *are* the subsets of S. But $\varnothing \in S$ is a different claim entirely, usually false: it asserts that the empty set is one of the actual members of S, and for $S = \{1, 2\}$ it plainly is not. The test is mechanical: ask whether the thing on the left is being offered as a *member* ($\in$) or as a sub-collection ($\subseteq$), and check the right object against

the matching definition. A symptom of the confusion is $\varnothing = \{\varnothing\}$; these are unequal, since the left has no elements and the right has exactly one (namely $\varnothing$).

Remark 0.3.18

Why there is no "set of all sets"

Set-builder notation feels as though it should let us form "the set of all x with property P" for *any* property P and any reach we like. It does not, and the reason is worth seeing once, because it explains the insistence in 0.3.3 on always naming a domain. Suppose we allowed an unrestricted $R = \{\, x : x \notin x \,\}$, the set of all sets that are not members of themselves. Ask whether $R \in R$. If $R \in R$, then R meets its own defining condition, so $R \notin R$; and if $R \notin R$, then R satisfies the condition, so $R \in R$. Either answer contradicts itself—this is *Russell's paradox*. The lesson is not that mathematics is broken; it is that you may not conjure a set from a property out of thin air. You carve new sets *out of sets already in hand*, with set-builder applied to a known domain D, exactly as we have done throughout. Do that, and no paradox can arise; the heavy machinery that guarantees this is the subject of axiomatic set theory, which the course will not need.

To build the plane, and later every space in the course, we need to pair elements in a way that, unlike a set, *remembers order*. The set $\{a, b\}$ forgets which came first; an *ordered pair* does not.

Definition 0.3.19

Ordered pair

An *ordered pair* (a, b) is the pairing of a *first* coordinate a with a *second* coordinate b, in that order. Its defining property is the equality rule

$$(a, b) = (c, d) \quad \Longleftrightarrow \quad a = c \ \text{ and } \ b = d\,:$$

two ordered pairs are equal exactly when they agree in the first coordinate and in the second. In particular order counts—$(1, 2) \neq (2, 1)$, although $\{1, 2\} = \{2, 1\}$—and a coordinate may repeat, as in $(7, 7)$, which no two-element set could express.

Definition 0.3.20

Cartesian product

The *Cartesian product* of sets A and B is the set of all ordered pairs with first coordinate from A and second from B:

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

The plane is the leading example: $\mathbb{R} \times \mathbb{R}$, written $\mathbb{R}^2$, is the set of all coordinate pairs (x, y) of real numbers. More generally, for finitely many sets $A_1, \ldots, A_n$ the product

$$A_1 \times \cdots \times A_n = \{\, (a_1, \ldots, a_n) : a_i \in A_i \text{ for each } i \,\}$$

collects the ordered n-tuples, with $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ (n factors) the space in which the course's metric geometry (2.2) will eventually live.

Example 0.3.21

A product is a grid; for finite sets, sizes multiply

With $A = \{1, 2, 3\}$ and $B = \{x, y\}$,

$$A \times B = \big\{(1, x), (1, y), (2, x), (2, y), (3, x), (3, y)\big\},$$

six pairs—a 3-by-2 grid of choices. In general $|A \times B| = |A| \cdot |B|$ for finite A, B, since each of the $|A|$ first coordinates may be matched with any of the $|B|$ second coordinates. Order in the product matters too: $A \times B$ and $B \times A$ are different sets here, the first holding $(1, x)$ and the second holding $(x, 1)$.

Union and intersection were defined for two sets, but analysis routinely combines *infinitely many*—one set for each natural number, or one for each real radius. We index the sets by a label drawn from an *index set* and take the union or intersection over all labels at once.

Definition 0.3.22

Indexed family; arbitrary union and intersection

An *indexed family* of sets is an assignment $\alpha \mapsto A_\alpha$ of a set A_α to each α in an *index set* I; we write it $\{A_\alpha\}_{\alpha \in I}$. Its union and intersection are

$$\bigcup_{\alpha \in I} A_\alpha = \{\, x : x \in A_\alpha \text{ for } \textit{some } \alpha \in I \,\}, \qquad \bigcap_{\alpha \in I} A_\alpha = \{\, x : x \in A_\alpha \text{ for } \textit{every } \alpha \in I \,\}.$$

The union collects everything that lands in at least one member of the family; the intersection, only what lands in all of them. The index set I may be finite, or infinite—this is exactly what lets us combine infinitely many sets in a single stroke. The quantifier inside each definition is the one to keep your eye on: "some α" for $\bigcup$, "every α" for $\bigcap$.

Example 0.3.23

An intersection over an infinite index set

Index by the positive integers $I = \{\, n \in \mathbb{N} : n \geq 1 \,\}$, and let $A_n = \big[\, 0,\ \tfrac{1}{n} \,\big]$, the closed interval from 0 to $1/n$. As n grows the intervals shrink toward the single point 0. Their union is just the largest, $\bigcup_{n \in I} A_n = [0, 1]$ (since $A_1 = [0, 1]$ contains all the others). Their intersection is

$$\bigcap_{n \in I} A_n = \{0\}.$$

Certainly $0 \in A_n$ for every n, so 0 is in the intersection. And no positive x survives: given $x > 0$, some $1/n < x$ (the Archimedean property, 1.7), so $x \notin A_n = [0, 1/n]$ for that n, hence x fails the "every α" test. The intersection of infinitely many nonempty sets can thus collapse to a single point—a phenomenon at the heart of the nested-interval arguments to come.

0.4 Relations and Functions

Sets on their own are inert collections; analysis comes alive only once we can speak of how their elements stand to one another and how one set is sent into another. "3 is less than 5," "2 divides 6," "f sends x to x^2"—each pairs up elements, and each is captured by the same single idea, a *relation*: a set of ordered pairs. From that one notion we will carve out the object this whole course turns on, the *function*, and then the two refinements of it that every later definition leans on: the way a function can be *injective* or *surjective*, and the way an *equivalence relation* sorts a set into clean disjoint pieces. We build all of it from the sets of 0.3, with the quantifier discipline of 0.2 doing the real work.

Definition 0.4.1

Ordered pair and Cartesian product

We recall two notions from 0.3.19–0.3.20, now as the substrate on which relations are built. The *ordered pair* (a, b) records two objects *in order*: $(a, b) = (c, d)$ exactly when $a = c$ and $b = d$, so $(1, 2)$ and $(2, 1)$ are different pairs even though the sets $\{1, 2\}$ and $\{2, 1\}$ are equal. Given sets A and B, their *Cartesian product* is the set of all such pairs,

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

We write A^2 for $A \times A$. Thus $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the familiar plane of coordinate pairs, and $\{1, 2\} \times \{x, y\} = \{(1, x), (1, y), (2, x), (2, y)\}$ has $2 \cdot 2 = 4$ elements.

A relation is simply a choice of *which* pairs we want to single out from this product.

Definition 0.4.2

Relation

A *relation* from A to B is a subset $R \subseteq A \times B$. We write $a \; R \; b$ (read "a is related to b") to mean $(a, b) \in R$, and $a \not R \; b$ when $(a, b) \notin R$. When $A = B$ we call R a relation *on* A. The *domain* of R is the set of first coordinates that actually occur, and the *range* the set of second coordinates,

$$\operatorname{dom} R = \{\, a \in A : \exists\, b \in B,\ a \; R \; b \,\}, \qquad \operatorname{ran} R = \{\, b \in B : \exists\, a \in A,\ a \; R \; b \,\}.$$

Example 0.4.3

Two relations you already know

The order relation "$\leq$" on $\mathbb{R}$ is the relation $R = \{(a, b) \in \mathbb{R}^2 : a \leq b\}$; writing $a \; R \; b$ is exactly writing $a \leq b$, so the familiar symbol *is* the set of pairs it picks out. Its domain and range are both all of $\mathbb{R}$. The relation "divides" on $\mathbb{N}$ is $D = \{(m, n) : \exists\, k \in \mathbb{N},\ n = mk\}$; here $2 \; D \; 6$ (take $k = 3$) but $4 \not D \; 6$. These two examples already display the structure we chase below: $\leq$ is an *order*, and—once we adjust it—a relation like "$m - n$ is even" will be an *equivalence*.

Most relations are too loose to compute with. A function is the special, disciplined kind in which every input has *exactly one* output—no input left unanswered, and none answered twice. That

single phrase, "exactly one," is the unique-existence quantifier of 0.2.9, and it is the heart of the definition.

Definition 0.4.4

Function

A *function* (or *map*) $f\colon A \to B$ is a relation $f \subseteq A \times B$ such that for every $a \in A$ there is exactly one $b \in B$ with $(a, b) \in f$:

$$\forall\, a \in A, \quad \exists!\, b \in B, \quad (a, b) \in f.$$

That unique b is written $f(a)$, the *value* of f at a. The set A is the *domain*, B the *codomain*, and the *image* (or *range*) is the set of values actually hit,

$$f(A) = \{\, f(a) : a \in A \,\} \subseteq B.$$

Two functions are equal when they have the same domain, the same codomain, and agree at every point—$f(a) = g(a)$ for all $a \in A$.

Remark 0.4.5

Well-definedness: the two ways to fail

The defining clause "exactly one b" is really two demands, and a candidate rule can break either. It must hand back *at least* one output for each input—a rule with no value at some a (such as $a \mapsto 1/a$ offered as a map on all of $\mathbb{R}$, which says nothing at $a = 0$) is not a function on that domain. And it must hand back *at most* one—a rule assigning two values to a single input is not a function at all. The square-root "rule" r on $\mathbb{R}_{\geq 0}$ that returns *both* $+\sqrt{x}$ and $-\sqrt{x}$ fails here: 4 would be related to both 2 and -2, so $(4, 2)$ and $(4, -2)$ both lie in r and the output is not determined. When a definition presents an output through some choice—"pick a representative, then…"—checking that the answer does not depend on the choice is exactly the chore called *verifying f is well defined*, and we meet it head-on with equivalence classes in 0.4.16.

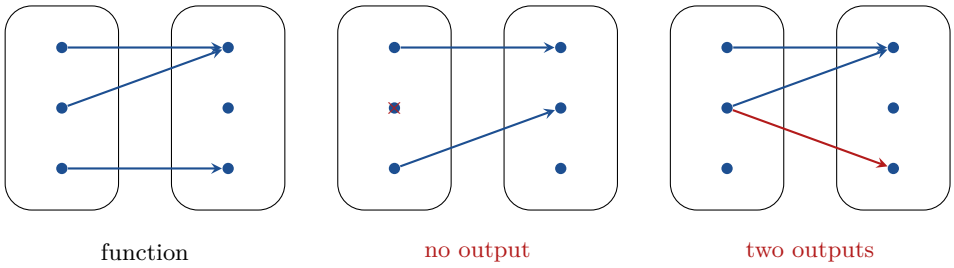


Figure 0.4.6. A FUNCTION VERSUS TWO NON-FUNCTIONS. In a function every left dot fires exactly one arrow; the middle picture leaves an input unanswered, the right picture answers one input twice.

A function moves whole *sets* around, forwards and backwards. The forward motion sends a subset of the domain to the set of its values; the backward motion—defined for *any* function, invertible or not—collects every input that lands in a target set.

Definition 0.4.7

Image and preimage of a set

Let $f\colon A \to B$, let $S \subseteq A$ and $T \subseteq B$. The *image* of S and the *preimage* (or *inverse image*) of T are

$$f(S) = \{\, f(a) : a \in S \,\} \subseteq B, \qquad f^{-1}(T) = \{\, a \in A : f(a) \in T \,\} \subseteq A.$$

The symbol $f^{-1}(T)$ is a set, read off the rule f alone; it carries *no* claim that f has an inverse function. For instance, with $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, we have $f^{-1}(\{4\}) = \{-2, 2\}$ and $f^{-1}\big((-\infty, 0)\big) = \emptyset$, even though f is nowhere invertible.

Proposition 0.4.8

Preimage respects the set operations

For any $f\colon A \to B$ and any subsets $T, U \subseteq B$,

$$f^{-1}(T \cap U) = f^{-1}(T) \cap f^{-1}(U), \qquad f^{-1}(T \cup U) = f^{-1}(T) \cup f^{-1}(U).$$

For any $S, S' \subseteq A$, the forward image is weaker: $f(S \cap S') \subseteq f(S) \cap f(S')$ always holds, but the reverse inclusion can fail.

Proof 0.4.8

Direct Proof by Membership-Chasing that the Preimage Respects Intersection.

(1) An element-of statement is proved by translating each side into its membership condition and checking the two conditions are the very same sentence. Fix $a \in A$. By the definition of preimage,

$$a \in f^{-1}(T \cap U) \iff f(a) \in T \cap U \iff \big(f(a) \in T \text{ and } f(a) \in U\big).$$

(2) The right-hand clause is, again by the definition of preimage, exactly "$a \in f^{-1}(T)$ and $a \in f^{-1}(U)$," which is to say $a \in f^{-1}(T) \cap f^{-1}(U)$. Since a was arbitrary, the two sets have the same elements and are therefore equal (0.3).

(3) For the forward image the inclusion $f(S \cap S') \subseteq f(S) \cap f(S')$ holds because any $f(a)$ with $a \in S \cap S'$ lies in both $f(S)$ and $f(S')$. The reverse fails for $f(x) = x^2$ with $S = \{1\}$, $S' = \{-1\}$: then $S \cap S' = \emptyset$ gives $f(S \cap S') = \emptyset$, yet $f(S) \cap f(S') = \{1\} \cap \{1\} = \{1\}$.

■

[Direct] the union rule is identical with "and" replaced by "or".

Two questions decide how much a function can be reversed. Does it ever *collapse* two inputs to one output? Does it *reach* every point of the codomain? A "no" to the first and a "yes" to the second are the properties *injective* and *surjective*; together they make a function perfectly reversible. Each property is a quantified sentence, and—this is the part to internalise—each

comes with a fixed recipe for how its proof must open.

Definition 0.4.9

Injective, surjective, bijective

A function $f\colon A \to B$ is

- *injective* (*one-to-one*) if distinct inputs give distinct outputs: $\forall\, x, y \in A,\ (f(x) = f(y) \Rightarrow x = y)$;
- *surjective* (*onto*) if every point of the codomain is hit: $\forall\, b \in B,\ \exists\, a \in A,\ f(a) = b$, that is, $f(A) = B$;
- *bijective* (a *one-to-one correspondence*) if it is both injective and surjective.

Remark 0.4.10

How each proof must begin

The quantifier shape of each property dictates the first line of its proof—learn these openings and you will never stare at a blank page on this kind of problem.

- To prove f *injective*, the cleanest route is the equality form of the definition: *assume* $f(x) = f(y)$ for arbitrary $x, y \in A$, and *deduce* $x = y$. (One may instead assume $x \neq y$ and derive $f(x) \neq f(y)$, but the equation form is almost always easier to manipulate.)
- To prove f *surjective*, it is a $\forall\exists$ statement (0.2.7): *fix an arbitrary* $b \in B$ and *construct a witness*—produce a specific $a \in A$, usually by solving $f(a) = b$ for a, and then verify $f(a) = b$.
- To *disprove* injectivity, exhibit one collision $x \neq y$ with $f(x) = f(y)$; to disprove surjectivity, exhibit one $b \in B$ that no input reaches.

Example 0.4.11

The same rule, four different verdicts

Whether a function is injective or surjective depends on the chosen domain and codomain, not on the formula alone. Take the rule $x \mapsto x^2$.

- $f\colon \mathbb{R} \to \mathbb{R}$ is neither: $f(-1) = f(1) = 1$ kills injectivity, and no input gives -1, killing surjectivity.
- $g\colon \mathbb{R} \to [0, \infty)$ is surjective (every $b \geq 0$ is $g(\sqrt{b})$) but still not injective.
- $h\colon [0, \infty) \to [0, \infty)$ is bijective: injective because $x^2 = y^2$ with $x, y \geq 0$ forces $x = y$, and surjective as before.

The successor map $s\colon \mathbb{N} \to \mathbb{N}$, $s(n) = n + 1$, is injective (if $n + 1 = m + 1$ then $n = m$) but not surjective: nothing maps to 0, since 0 is the least natural number. This is the first hint that an infinite set can match a *proper* subset of itself—a theme we take up under "same size" below and pursue in 2.1.

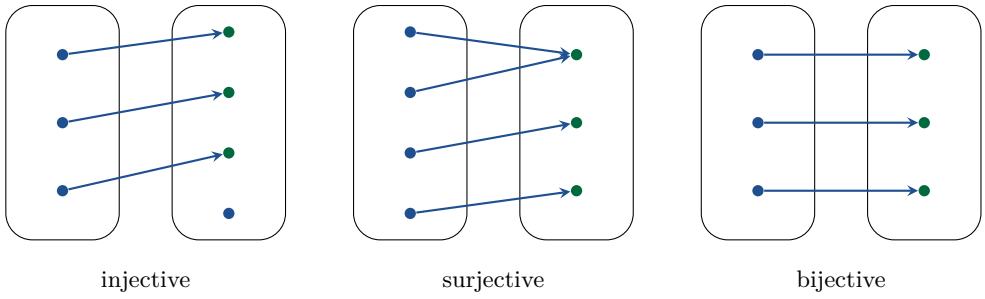


Figure 0.4.12. Injective, surjective, bijective. Injective: no two arrows share a head. Surjective: every right dot is hit. Bijective: a perfect pairing, no misses and no collisions.

Functions chain: feed the output of one into the next. This chaining, *composition*, is how complicated maps are assembled from simple ones, and it is the setting in which "reversing" a function gets its precise meaning.

Definition 0.4.13

Composition and the identity

Given $f\colon A \to B$ and $g\colon B \to C$, the *composition* $g \circ f\colon A \to C$ is the function defined by

$$(g \circ f)(a) = g\big(f(a)\big), \qquad a \in A.$$

(The codomain of f must match the domain of g, or the right-hand side is meaningless.) For each set A the *identity function* $\mathrm{id}_A\colon A \to A$ is $\mathrm{id}_A(a) = a$. Composition is *associative*—$(h \circ g) \circ f = h \circ (g \circ f)$, since both send a to $h(g(f(a)))$—and the identity is neutral: $f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f$.

Definition 0.4.14

Inverse function

A function $f\colon A \to B$ is *invertible* if there is a function $g\colon B \to A$ with

$$g \circ f = \mathrm{id}_A \qquad \text{and} \qquad f \circ g = \mathrm{id}_B,$$

that is, $g(f(a)) = a$ for all $a \in A$ and $f(g(b)) = b$ for all $b \in B$. Such a g, when it exists, is unique and is written f^{-1}.[3]

Theorem 0.4.15

A function is invertible iff it is bijective

A function $f\colon A \to B$ is invertible if and only if it is bijective.

Proof 0.4.15

Direct Proof that Invertible Equals Bijective by Constructing the Inverse.

[3] This f^{-1}, the inverse *function*, exists only when f is bijective; the preimage $f^{-1}(T)$ of 0.4.7 is a set defined for *every* f. The shared notation is standard and harmless once you know which object is meant: f^{-1} alone is a function, $f^{-1}(\,\cdot\,)$ applied to a *set* is a preimage.

(1) For the forward direction, suppose f is invertible with inverse g. To see f is injective, take $x, y \in A$ with $f(x) = f(y)$; applying g and using $g \circ f = \mathrm{id}_A$,

$$x = g\big(f(x)\big) = g\big(f(y)\big) = y.$$

To see f is surjective, fix $b \in B$ and set $a = g(b)$; then $f(a) = f(g(b)) = (f \circ g)(b) = b$, so b is hit. Hence f is bijective.

(2) For the converse, suppose f is bijective. We must *build* a function $g\colon B \to A$. Fix $b \in B$. Surjectivity gives at least one $a \in A$ with $f(a) = b$; injectivity makes it the only one, for if $f(a) = f(a') = b$ then $a = a'$. So there is exactly one such a, and defining $g(b)$ to be that a assigns one output to each input—g is a genuine function (0.4.4).

(3) This g inverts f on both sides. For $a \in A$, the unique preimage of $f(a)$ is a itself, so $g(f(a)) = a$, giving $g \circ f = \mathrm{id}_A$. For $b \in B$, $g(b)$ is by construction the input with $f(g(b)) = b$, giving $f \circ g = \mathrm{id}_B$. Thus f is invertible.

■

[Direct] cf. 0.2.9 (unique existence) and 0.4.9.

We now return to relations on a single set to isolate the kind that behaves like *equality of some feature*—"same remainder on division by n," "same length," "same direction." Such a relation does not merely compare elements; it sorts the whole set into groups of mutually-related elements, and those groups will turn out to tile the set with no gaps and no overlaps. This sorting mechanism is the engine behind the construction of the integers and the rationals (1.2) and, later, the completion that builds the reals.

Definition 0.4.16

Equivalence relation and equivalence class

A relation $\sim$ on a set A is an *equivalence relation* if it is

- *reflexive*: $\forall\, a \in A,\ a \sim a$;
- *symmetric*: $\forall\, a, b \in A,\ (a \sim b \Rightarrow b \sim a)$;
- *transitive*: $\forall\, a, b, c \in A,\ (a \sim b \text{ and } b \sim c \Rightarrow a \sim c)$.

For $a \in A$ the *equivalence class* of a is the set of everything related to it,

$$[a] = \{\, x \in A : x \sim a \,\} \subseteq A,$$

and any $x \in [a]$ is called a *representative* of that class. The set of all classes is the *quotient* $A/\!\sim\, = \{\, [a] : a \in A \,\}$.

Example 0.4.17

Congruence modulo n

Fix $n \in \mathbb{N}$ with $n \geq 1$. On $\mathbb{Z}$ declare $a \equiv b \pmod{n}$ to mean n divides $a - b$. This is an equivalence relation: it is reflexive because $n \mid 0 = a - a$; symmetric because $n \mid (a-b)$ forces $n \mid (b-a)$; transitive because if $n \mid (a-b)$ and $n \mid (b-c)$ then n divides their sum $(a-b)+(b-c) = a-c$. Its classes are the *residue classes*: for $n = 3$ they are

$$[0] = \{\dots, -3, 0, 3, 6, \dots\}, \quad [1] = \{\dots, -2, 1, 4, 7, \dots\}, \quad [2] = \{\dots, -1, 2, 5, 8, \dots\},$$

three disjoint sets whose union is all of $\mathbb{Z}$. Every integer lands in exactly one—precisely the partition phenomenon we now prove in general.

Theorem 0.4.18

Equivalence classes partition the set

Let $\sim$ be an equivalence relation on a nonempty set A. Then the distinct equivalence classes form a *partition* of A: each class is nonempty, any two classes are either identical or disjoint, and the classes together cover A. Conversely, every partition of A arises from exactly one equivalence relation, namely "$a \sim b$ iff a and b lie in the same piece."

Proof 0.4.18

Direct Proof that Equivalence Classes Partition the Set.

(1) Each class is nonempty and the classes cover A: reflexivity gives $a \sim a$, so $a \in [a]$; thus no class is empty, and every $a \in A$ sits in the class $[a]$, so the union of all classes is A.

(2) The key step is that two classes overlap only by coinciding. Suppose $[a] \cap [b] \neq \emptyset$ and pick c in the intersection, so $c \sim a$ and $c \sim b$. For any $x \in [a]$ we have $x \sim a$; chaining with $a \sim c$ (symmetry) and $c \sim b$ (transitivity, twice) gives $x \sim b$, hence $x \in [b]$. So $[a] \subseteq [b]$, and the symmetric argument gives $[b] \subseteq [a]$; therefore $[a] = [b]$. Any two classes are thus equal or disjoint, and with the previous step the classes partition A.

(3) For the converse, let $\mathcal{P}$ be a partition—a family of nonempty, pairwise-disjoint subsets of A with union A—and define $a \sim b$ to mean a, b belong to a common piece $P \in \mathcal{P}$. This is reflexive (each a lies in some piece, since the pieces cover A), symmetric (the condition is unordered in a, b), and transitive (if a, b share a piece P and b, c share a piece Q, then $b \in P \cap Q$ forces $P = Q$ by disjointness, so $a, c \in P$). Its classes are exactly the pieces of $\mathcal{P}$, and no other equivalence relation yields this partition, since the relation is forced by "same piece."

■

[Direct] the whole proof runs on the three axioms of 0.4.16.

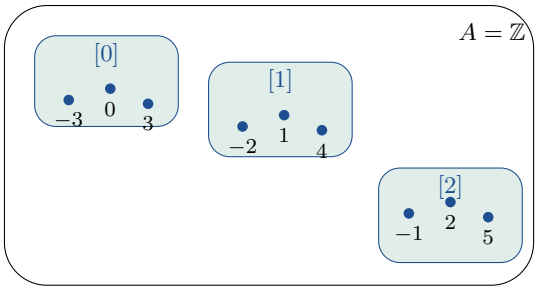


Figure 0.4.19. A PARTITION INTO EQUIVALENCE CLASSES. The relation $\sim$ carves A into non-overlapping cells, here the three residue classes $[0], [1], [2]$ of congruence modulo 3 on a patch of $\mathbb{Z}$. Every element sits in exactly one cell; the cells meet nowhere and together cover all of A.

Remark 0.4.20

Why "well defined on classes" is the recurring worry

Once a set is cut into classes, one constantly wants to define operations on the classes by working with a representative: "to add $[a]$ and $[b]$, add a and b and take $[a+b]$." The danger is that the answer might depend on *which* representatives you grabbed. Showing it does not—that $a \sim a'$ and $b \sim b'$ imply $a+b \sim a'+b'$, so $[a+b] = [a'+b']$—is the well-definedness check of 0.4.5, and it is unavoidable: skip it and the operation is simply not defined. Every quotient construction in 1.2—the integers from pairs of naturals, the rationals from pairs of integers—hinges on exactly this verification.

A bijection does something striking: it pairs the elements of two sets perfectly, one for one, with none left over on either side. That is precisely what it should mean for two sets to have *equally many* elements—and it is the only definition that survives the jump to infinite sets, where counting is impossible.

Definition 0.4.21

Equinumerous sets; finite and infinite

Two sets A and B are *equinumerous* (have the *same cardinality*), written $A \approx B$, if there exists a bijection $f\colon A \to B$. A set A is *finite* if $A \approx \{0, 1, \ldots, n-1\}$ for some $n \in \mathbb{N}$ (and then A has n elements), and *infinite* otherwise; here $n = 0$ gives the empty set $\{0, \ldots, -1\} = \emptyset$, so the empty set is finite with 0 elements, and $n = 1$ gives the one-element set $\{0\}$. Equinumerosity is itself reflexive, symmetric, and transitive—witnessed by the identity, the inverse of a bijection, and the composition of two bijections—so it behaves like an equivalence relation comparing sizes.

Remark 0.4.22

The surprise of the infinite

For *finite* sets this matches ordinary counting and a proper subset is always strictly smaller. Infinite sets break that intuition: the successor map $n \mapsto n+1$ from 0.4.11 is a bijection from $\mathbb{N}$ onto $\mathbb{N} \setminus \{0\}$, so an infinite set can be equinumerous with a proper subset of itself. This is not a paradox but the *definition* of infinite at work, and it opens the door to a genuine hierarchy of infinities—countable versus uncountable—which we develop in earnest in 2.1. We plant only the seed here.

Definition 0.4.23

Partial and total order

A relation $\leq$ on a set A is a *partial order* if it is reflexive, *antisymmetric* ($a \leq b$ and $b \leq a$ imply $a = b$), and transitive. It is a *total* (or *linear*) order if in addition any two elements are comparable: $\forall\, a, b \in A,\ (a \leq b \text{ or } b \leq a)$. The usual $\leq$ on $\mathbb{R}$ is a total order, and that comparability is exactly what makes the real line a *line*; the subset relation $\subseteq$ on the subsets of a set is a partial order that is not total, since $\{1\}$ and $\{2\}$ are incomparable. The order axioms of $\mathbb{R}$, and the supremum property that distinguishes it from $\mathbb{Q}$, are the subject of 1.3 and 1.5.

With relations and functions in hand—image and preimage, the injective/surjective/bijective trichotomy, composition and inverse, and the equivalence-class machinery—we have the vocabulary every later definition is phrased in. What remains is to learn to *wield* it: to turn these definitions into proofs. That craft is 0.5.

0.5 The Craft of Proof

Everything so far has been preparation. We fixed the language of statements (0.1), the quantifiers that turn open sentences into claims (0.2), and the sets and functions the claims are about (0.3, 0.4). Now we come to the act the whole course is built on: proving. A proof is not a different kind of writing reserved for geniuses; it is a disciplined habit of justifying every assertion, and it runs on a small, learnable stock of strategies—the very *Modes of Argument* tagged on every proof in this book. This section lays out that stock, one strategy at a time, each with a worked proof in the house style, and ends with the question every beginner actually asks: when you are staring at a blank page, how do you *find* the proof?

Concept 0.5.1

What a proof is

A *proof* of a statement P is a finite chain of statements ending in P, in which every link is one of: an axiom, a definition, a hypothesis we are allowed to assume, a result proved earlier, or a statement that follows from earlier links by a valid rule of logic (0.1). *Rigour* means exactly that no link is skipped—a careful reader, granting only the axioms and definitions, can check each step without having to guess. The standard we hold to in this course is concrete: a proof is finished when a competent reader could read it line by line and object to nothing.

Concept 0.5.2

Before you prove: name the hypotheses and the goal

The most common reason a beginner is stuck is that they have not written down, precisely, what they may *assume* and what they must *show*. Do this first, and translate each into the language of 0.1–0.2 by *unfolding the definitions*. To prove "f is injective," for instance, replace the word by its definition and you are no longer staring at a vague adjective but at a concrete target with a known shape (0.4):

$$\text{show:}\quad \forall x, y \in A,\quad f(x) = f(y) \;\Rightarrow\; x = y.$$

A universal goal tells you to fix an arbitrary x, y; the inner implication tells you to assume $f(x) = f(y)$ and drive toward $x = y$. Half of most proofs is just this honest restatement.

We now catalogue the strategies.[4] The first is the one every other strategy falls back on.

Definition 0.5.3

Direct proof

A *direct proof* of an implication $P \Rightarrow Q$ assumes P and reaches Q by a forward chain of definitions and known results, with no special device. It is the default; reach for something cleverer only when the direct route stalls.

Proof 0.5.3

Direct proof that the product of two odd integers is odd.

[4] The worked examples lean on the elementary arithmetic of the integers, taken as known throughout: even and odd numbers, divisibility ($d \mid m$ means $m = dq$ for some integer q), prime and composite numbers, and the fact that every integer is even or odd.

(1) Suppose m and n are odd. By definition of oddness there are integers j, k with $m = 2j + 1$ and $n = 2k + 1$.

(2) Multiply and regroup:

$$mn = (2j + 1)(2k + 1) = 4jk + 2j + 2k + 1 = 2(2jk + j + k) + 1.$$

(3) Since $2jk+j+k$ is an integer, mn has the form 2(integer)+1, which is the definition of an odd integer. Hence mn is odd.

■

[Direct]

Definition 0.5.4

Proof by contrapositive

A *proof by contrapositive* of $P \Rightarrow Q$ proves the equivalent statement $\neg Q \Rightarrow \neg P$ instead (0.1.7): assume $\neg Q$ and derive $\neg P$. It is the right move when "$\neg Q$" is a more concrete handle than "P"—typically when Q is itself a negation or an existence claim that is awkward to use directly.

Proof 0.5.4

Proof by contrapositive that if n^2 is even then n is even.

(1) The hypothesis "n^2 even" is hard to use directly—it tells us little about n itself. So we prove the contrapositive: *if n is odd, then n^2 is odd.*

(2) Assume n is odd, say $n = 2k + 1$. Then

$$n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1,$$

which is odd.

(3) This establishes $\neg(n \text{ even}) \Rightarrow \neg(n^2 \text{ even})$, and by 0.1.7 that is logically the same as the claim "n^2 even $\Rightarrow$ n even."

■

[A9]

Definition 0.5.5

Proof by contradiction

A *proof by contradiction* of a statement P assumes its negation $\neg P$ and derives an *absurdity*—some statement R together with its negation $\neg R$. Since a true assumption can never lead to a contradiction, $\neg P$ must be false, so P is true. It is the natural weapon against statements of the form "there is no ..." or "... is impossible."

Proof 0.5.5

Proof by contradiction that $\sqrt{2}$ is irrational.

(1) Suppose, for contradiction, that $\sqrt{2}$ is rational. Then $\sqrt{2} = p/q$ for integers p, q with $q \neq 0$, and we may take the fraction in lowest terms, so that p and q are not both even.

(2) Squaring and clearing denominators, $p^2 = 2q^2$, so p^2 is even; by 0.5.4, p is even, say $p = 2r$.

(3) Substitute: $4r^2 = 2q^2$, hence $q^2 = 2r^2$, so q^2 is even, and again by 0.5.4 q is even.

(4) But now p and q are both even, contradicting the choice of lowest terms. The assumption was impossible, so $\sqrt{2}$ is irrational.

■

[A9]; the existence gap behind 1.4.

Remark 0.5.6

Contrapositive or contradiction?

The two are cousins, and beginners often blur them. A contrapositive proof is the disciplined special case: it proves $P \Rightarrow Q$ by assuming exactly $\neg Q$ and producing exactly $\neg P$—a single clean target. A contradiction proof is freer: it assumes $\neg P$ and may derive *any* absurdity at all. Prefer the contrapositive when your goal is an implication and the negated conclusion is the better foothold; save full contradiction for statements that are not implications, like an irrationality or a nonexistence. When a "contradiction" proof of $P \Rightarrow Q$ only ever uses $\neg Q$ to build $\neg P$, it was secretly a contrapositive—write it as one.

Definition 0.5.7

Proof by cases, and "without loss of generality"

A *proof by cases* splits the hypothesis into finitely many exhaustive alternatives and proves the goal in each; since the cases cover every possibility, the goal holds always. When two cases are interchangeable—the argument for one becomes the argument for the other by renaming—we prove just one and write "*without loss of generality*" (WLOG) to record that the symmetry disposes of the rest.

Proof 0.5.7

Proof by cases that $n^2 + n$ is even for every integer n.

(1) Every integer is even or odd; take the two cases.

(2) If n is even, $n = 2k$, then $n^2 + n = 4k^2 + 2k = 2(2k^2 + k)$ is even.

(3) If n is odd, $n = 2k+1$, then $n^2+n = n(n+1) = (2k+1)(2k+2) = 2(2k+1)(k+1)$ is even.

(4) The two cases exhaust the integers, so $n^2 + n$ is even in all cases. (One could instead note $n^2 + n = n(n + 1)$ is a product of consecutive integers, one of which is even—a shorter route, but the case split is the method on display.)

■

[A12]

Concept 0.5.8

Proving and disproving quantified statements

The grammar of the goal dictates the proof's skeleton (0.2.4).

- To prove $\forall x \in D,\ P(x)$: write "*fix an arbitrary* $x \in D$" and argue for that nameless x, using nothing particular about it. To *dis*prove it, produce one *counterexample*—a single x with $P(x)$ false.
- To prove $\exists x \in D,\ P(x)$: *construct* or exhibit a specific witness x and check $P(x)$. To *dis*prove it, prove the universal $\forall x \in D,\ \neg P(x)$.

For example, $\forall n \in \mathbb{N},\ n^2 \geq n$ is proved by fixing an arbitrary n: if $n = 0$ it reads $0 \geq 0$; if $n \geq 1$ then multiplying $n \geq 1$ by $n \geq 0$ gives $n^2 \geq n$. By contrast "every prime is odd" is destroyed by the single counterexample $p = 2$.

Concept 0.5.9

Proving "if and only if," and uniqueness

A biconditional $P \Leftrightarrow Q$ is two implications (0.1.9), and a proof must supply *both*: the forward direction ($\Rightarrow$) and the backward direction ($\Leftarrow$), often by different strategies. Thus "n is even iff n^2 is even" is a one-line direct argument in one direction ($n = 2k$ gives $n^2 = 2(2k^2)$, even) and the contrapositive of 0.5.4 in the other (an odd n has an odd square); it is worked in full at 0.E.36. A *uniqueness* claim, "there is exactly one x with $P(x)$" (0.2.9), is likewise two jobs: *existence* (exhibit one) and *uniqueness* (assume $P(x)$ and $P(y)$ and deduce $x = y$).

Concept 0.5.10

Proving statements about sets

Set claims reduce to logic about membership, and two patterns cover almost everything (0.3). To prove a containment $A \subseteq B$, "*fix* $x \in A$" and show $x \in B$ (*element-chasing*). To prove an equality $A = B$, prove $A \subseteq B$ and $B \subseteq A$ separately (*double inclusion*). For instance, transitivity of inclusion—if $A \subseteq B$ and $B \subseteq C$ then $A \subseteq C$—is one element-chase: fix $x \in A$; then $x \in B$ because $A \subseteq B$, and so $x \in C$ because $B \subseteq C$.

Remark 0.5.11

Negate to discover what you must produce

When a goal is itself a negation or a failure—"f is *not* continuous," "the sequence does *not* converge," "no x satisfies P"—do not argue in fog. Negate the definition mechanically by the rule of 0.2.5, flipping every quantifier, until the negation sits on a plain predicate; what remains is a positive thing to build. This is mode A11, and in analysis it is the difference between being lost and knowing exactly which ε, or which point, you are on the hook to produce (the worked negation of continuity is 0.E.8).

Remark 0.5.12

The Modes of Argument, in one map

The strategies above are the general-purpose core, and they are the front-matter Modes of Argument by another name: Direct (0.5.3), A9 contradiction and contrapositive (0.5.4, 0.5.5), A12 cases and WLOG (0.5.7), A11 negating a quantifier (0.5.11), and A10 induction, important enough to get its own treatment next (0.6). The remaining tags name the moves special to analysis, which the lectures will exemplify in place: A1 forcing a nonnegative quantity below every ε to be zero, A2 and A3 the ε–N and ε–δ arguments, A4 the triangle inequality with an added-and-subtracted middle term, A5

supremum and infimum arguments, A6 passing from an open cover to a finite subcover, A7 bisection and nested intervals, and A8 subsequence extraction and diagonalisation. Every one is a specialisation of the habits in this section; none is magic.

Concept 0.5.13

How to find a proof when you are stuck

Knowing the strategies is not the same as knowing which to use, and the honest answer is that finding a proof is a craft sharpened by practice, not a formula. Still, when the page is blank, this checklist almost always breaks the logjam.

(1) *Understand the statement exactly.* Unfold every definition and quantifier (0.5.2); write down precisely what is assumed and what is wanted. Most "hard" problems dissolve here.

(2) *Compute small and extreme cases.* Try $n = 0, 1, 2$; try the empty set; try the trivial function. Examples reveal the mechanism and often hand you the general argument.

(3) *Work from both ends.* Push forward from the hypotheses and backward from the goal until the two chains meet in the middle.

(4) *Find a relative.* Ask which theorem mentions *both* the hypothesis and the goal; recall a solved problem of the same shape and reuse its method.

(5) *Switch strategy deliberately.* If a universal resists a direct attack, try the contrapositive or contradiction; if you need existence, try to construct the object, or to derive an absurdity from its absence.

(6) *Separate finding from writing.* Discover the idea on scratch paper, then write the proof cleanly from the top—a finished proof rarely shows the false starts that produced it.

None of this substitutes for doing the work: as the epigraph insists, the only way to learn this is to prove things yourself. The exercises that follow are where the craft is actually acquired.

0.6 Mathematical Induction

A great many statements in this course read " $P(n)$ holds for every natural number n "—a formula that sums the first n integers, an inequality that holds for all n past some point, a property shared by every member of an unending list. We cannot check infinitely many cases by hand, and 0.2.4 warned that for a universal statement no finite pile of examples will do. *Mathematical induction* is the engine built precisely for this situation: a single argument that, in two short moves, establishes $P(n)$ for all n at once. It is mode A10, and it rests on one structural fact about the naturals, which we state first.

Theorem 0.6.1

Well-ordering of the naturals

Every nonempty subset of $\mathbb{N}$ has a *least element*: if $S \subseteq \mathbb{N}$ and $S \neq \emptyset$, then there is some $m \in S$ with $m \leq s$ for every $s \in S$.

We take well-ordering as a defining feature of $\mathbb{N} = \{0, 1, 2, \dots\}$; later in the course the naturals are built instead from the successor and induction axioms (1.1), and well-ordering is then a theorem rather than a starting point.[5] What makes it powerful is that it forbids an infinite *descent*: you cannot keep stepping to a strictly smaller natural number forever, because any set of the numbers you would visit would have a bottom. That single prohibition is exactly what licenses induction.

Theorem 0.6.2

The principle of mathematical induction

Fix an integer n_0, and let $P(n)$ be a predicate defined for every integer $n \geq n_0$. Suppose

- *(base case)* $P(n_0)$ is true; and
- *(inductive step)* for every integer $k \geq n_0$, the implication $P(k) \Rightarrow P(k+1)$ is true.

Then $P(n)$ is true for every integer $n \geq n_0$.

The picture to hold in your head is a line of dominoes. The base case is your hand knocking over the first one; the inductive step is the guarantee that each domino, when it falls, topples its neighbour. Grant both and every domino falls, however long the line—not because you pushed each one, but because the two moves together force the whole cascade. The inductive step does *not* assert $P(k)$; it only promises the link from $P(k)$ to $P(k+1)$, exactly the conditional promise of 0.1.6. The assumption "$P(k)$ is true," made temporarily inside the step so as to reach $P(k+1)$, has a name worth knowing.

Concept 0.6.3

The inductive hypothesis

Inside the inductive step you fix an arbitrary $k \geq n_0$ and *assume* $P(k)$; this working assumption is the *inductive hypothesis*. Your task in the step is to use it to derive $P(k+1)$. It is not something

[5] The same statement is false for $\mathbb{Z}$ (which has no least element) and for the nonnegative rationals (the set of positive rationals has no least element—between any two there is another). Well-ordering is special to $\mathbb{N}$.

you are claiming outright—it is the antecedent of the conditional you are proving, granted to you for the length of the step and discharged the moment you reach the conclusion. A proof by induction therefore has a fixed two-part shape: verify the base case outright, then prove the conditional "if $P(k)$ then $P(k+1)$" for an arbitrary k.

Why is the principle true? Well-ordering supplies a clean answer through the back door: if induction could fail, a *least* failure would have to exist, and we show it cannot.

Proof 0.6.2

Proof that well-ordering implies the principle of induction.

(1) Suppose, for contradiction, that the base case and inductive step both hold yet $P(n)$ fails for some integer $n \geq n_0$. Collect the failures into

$$S = \{\, n \in \mathbb{Z} : \ n \geq n_0 \text{ and } P(n) \text{ is false} \,\}.$$

The shift $n \mapsto n - n_0$ carries S onto a subset of $\mathbb{N}$, and it preserves order, so a least element of S corresponds to a least element of that image and vice versa. By assumption S is nonempty, hence so is its image; well-ordering (0.6.1) gives the image a least element, and pulling back, S itself has a least element m.

(2) The base case says $P(n_0)$ is true, so $m \neq n_0$; hence $m > n_0$, and $m-1$ is an integer with $m - 1 \geq n_0$. Because m is the *least* member of S, the smaller number $m - 1$ is not in S, so $P(m-1)$ is true.

(3) Apply the inductive step with $k = m - 1$: from $P(m-1)$ true we get $P(m)$ true. But $m \in S$ means $P(m)$ is false—a contradiction.

(4) No least failure can exist, so S is empty: $P(n)$ holds for every integer $n \geq n_0$.

■

[A9] via a least counterexample.

This is one of two honest routes to the same principle. Here well-ordering is the given and induction is the theorem; when the naturals are constructed later (1.1), it is induction that is taken as an axiom and well-ordering that is derived. The two derivations run in opposite directions and neither is circular—they simply choose different members of the same family (0.6.13) as the foundation. With the principle secured, the rest of the section is practice—and induction is a craft learned only by watching it run and then running it yourself. We begin with the identity that is, by tradition, everyone's first.

Proposition 0.6.4

The sum of the first n integers

For every integer $n \geq 1$,

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

Proof 0.6.4

Proof by Induction on n of the closed form for $1 + \cdots + n$.

(1) Write $P(n)$ for the equation $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$, to be proved for all $n \geq 1$. For the base case $n = 1$ the left side is 1 and the right side is $\frac{1\cdot 2}{2} = 1$, so $P(1)$ holds.

(2) Fix an arbitrary $k \geq 1$ and assume the inductive hypothesis $P(k)$:

$$1 + 2 + \cdots + k = \frac{k(k+1)}{2}.$$

To reach $P(k+1)$, add the next term $k+1$ to both sides:

$$1+2+\cdots+k+(k+1) = \frac{k(k+1)}{2}+(k+1) = \frac{k(k+1)+2(k+1)}{2} = \frac{(k+1)(k+2)}{2}.$$

(3) The right-hand end is $\frac{(k+1)\big((k+1)+1\big)}{2}$, which is exactly the statement $P(k+1)$. Thus $P(k) \Rightarrow P(k+1)$ for every $k \geq 1$, and by 0.6.2 the identity holds for all $n \geq 1$.

■

[A10] cf. the Gauss pairing picture of 1.1.

Remark 0.6.5

What induction proves and what it hides

The proof above is airtight, yet it gives no hint of *where* the formula $\frac{n(n+1)}{2}$ came from—induction verifies a guess but never produces one. The young Gauss is said to have found it by pairing $1+2+\cdots+n$ with the same sum written backwards, $n + (n-1) + \cdots + 1$; the n columns each total $n+1$, so twice the sum is $n(n+1)$. That argument is the rectangle picture you will meet later (1.1): two copies of the staircase $1 + 2 + \cdots + n$ interlock into an $n \times (n+1)$ grid. The picture *discovers* the answer; induction then *certifies* it. Keep both in view: the explanatory argument tells you why, the induction tells you that it never fails.

The next two are run with less commentary; the shape is by now the point. Watch how each inductive step reduces the case $k+1$ to the case k by isolating a single extra term.

Proposition 0.6.6

Two further identities

For every integer $n \geq 1$,

$$\textbf{(a)}\quad 1 + 3 + 5 + \cdots + (2n-1) = n^2, \qquad \textbf{(b)}\quad 1 + 2 + 2^2 + \cdots + 2^{n-1} = 2^n - 1.$$

Proof 0.6.6

Proof by Induction of the odd-number and geometric sums.

(1) For **(a)**, let $P(n)$ be "$1 + 3 + \cdots + (2n-1) = n^2$." The base case $n = 1$ reads $1 = 1^2$, true. Assume $P(k)$, namely $1 + 3 + \cdots + (2k-1) = k^2$; the $(k+1)$st odd

number is $2(k+1)-1 = 2k+1$, so

$$1 + 3 + \cdots + (2k-1) + (2k+1) = k^2 + (2k+1) = (k+1)^2,$$

which is $P(k+1)$. Hence (a) holds for all $n \geq 1$.

(2) For **(b)**, let $Q(n)$ be "$1 + 2 + \cdots + 2^{n-1} = 2^n - 1$." The base case $n = 1$ reads $1 = 2^1 - 1$, true. Assume $Q(k)$; the next term is 2^k, so

$$(1 + 2 + \cdots + 2^{k-1}) + 2^k = (2^k - 1) + 2^k = 2 \cdot 2^k - 1 = 2^{k+1} - 1,$$

which is $Q(k+1)$. Hence (b) holds for all $n \geq 1$. The pattern in (a) is worth a second look: the running totals $1, 4, 9, 16, \ldots$ of consecutive odd numbers are exactly the perfect squares.

■

[A10]

Induction is not confined to equalities. An *inequality* that holds from some point on is proved the same way, with the inductive step doing a little estimation rather than exact algebra. The next result also shows why the base case need not be $n = 0$ or $n = 1$: the inequality is simply false for small n, so we start where it first becomes true. (The gentlest such fact is $2^n > n$, true for every $n \in \mathbb{N}$ and a clean first exercise; we prove the harder $2^n > n^2$, which needs a later starting point.)

Proposition 0.6.7

An exponential outgrows a square

For every integer $n \geq 5$, $2^n > n^2$.

Proof 0.6.7

Proof by Induction of $2^n > n^2$ for $n \geq 5$.

(1) Let $P(n)$ be "$2^n > n^2$." Small values fail—$2^4 = 16 = 4^2$ is not strict, and $2^3 = 8 < 9$; the inequality first holds at $n = 5$, where $2^5 = 32 > 25 = 5^2$. So the base case is $P(5)$.

(2) Fix $k \geq 5$ and assume $P(k)$: $2^k > k^2$. Doubling, $2^{k+1} = 2 \cdot 2^k > 2k^2$. It now suffices to show $2k^2 \geq (k+1)^2$, for then $2^{k+1} > 2k^2 \geq (k+1)^2$ and $P(k+1)$ follows.

(3) Expanding, $2k^2 \geq (k+1)^2$ is the same as $k^2 \geq 2k + 1$. Since $k \geq 5$ we have $k^2 \geq 5k = 2k + 3k > 2k + 1$, so the needed inequality holds.

(4) Chaining the two estimates gives $2^{k+1} > (k+1)^2$, which is $P(k+1)$. By 0.6.2, $2^n > n^2$ for all $n \geq 5$.

■

[A10] the base case is $n_0 = 5$, not 0.

A divisibility statement is induction in yet another costume. Here "d divides m," written $d \mid m$, means $m = d \cdot q$ for some integer q. The trick in the step is always to write the $(k+1)$st quantity in terms of the kth plus a remainder you can see is divisible.

Proposition 0.6.8

A divisibility fact

For every $n \in \mathbb{N}$, $3 \mid (n^3 - n)$.

Proof 0.6.8

Proof by Induction that $3 \mid (n^3 - n)$.

(1) Let $P(n)$ be "$3 \mid (n^3 - n)$." The base case $n = 0$ gives $0^3 - 0 = 0 = 3 \cdot 0$, divisible by 3, so $P(0)$ holds.

(2) Fix $k \geq 0$ and assume $P(k)$: there is an integer q with $k^3 - k = 3q$. Expand the next quantity and group it around $k^3 - k$:

$$(k+1)^3 - (k+1) = \big(k^3 + 3k^2 + 3k + 1\big) - (k+1) = \big(k^3 - k\big) + 3\big(k^2 + k\big).$$

(3) The first bracket is $3q$ by hypothesis and the second is $3(k^2 + k)$, so

$$(k+1)^3 - (k+1) = 3q + 3\big(k^2 + k\big) = 3\big(q + k^2 + k\big),$$

a multiple of 3. Thus $P(k+1)$ holds, and by 0.6.2 the claim holds for every $n \in \mathbb{N}$. ■

[A10]

Each proof so far reached $P(k+1)$ from its *immediate* predecessor $P(k)$. Sometimes one step back is not enough—the case n naturally breaks into pieces that are *several*, or an unknown number of, steps smaller. For these we need a sturdier form that hands us *all* earlier cases at once.

Theorem 0.6.9

Strong (complete) induction

Fix an integer n_0 and let $P(n)$ be a predicate for $n \geq n_0$. Suppose that for every integer $k \geq n_0$,

$$\Big(P(j) \text{ holds for all } j \text{ with } n_0 \leq j < k\Big) \implies P(k).$$

Then $P(n)$ holds for every integer $n \geq n_0$. (Taking $k = n_0$, there are no integers j with $n_0 \leq j < n_0$, so the antecedent is the empty conjunction—vacuously true (0.1.6)—and the implication forces $P(n_0)$ outright, so the base case is genuinely built in.)

Remark 0.6.10

When you need the strong form

The only difference from 0.6.2 is the inductive hypothesis: ordinary induction lends you $P(k)$, strong induction lends you the entire run $P(n_0), \ldots, P(k)$. Reach for the strong form whenever building case

$k+1$ sends you back to a case that is *not* the immediate predecessor—a recursion like $a_n = a_{n-1}+a_{n-2}$ that looks two steps back, or an argument that factors $k+1$ into smaller pieces of unpredictable size. The two principles are not rivals: anything provable one way is provable the other, and strong induction is itself a theorem of well-ordering by the same least-counterexample argument as 0.6.2. Use whichever makes the step honest.

The classic place where one step back fails is factoring a number into primes: knowing that k has a prime factorization tells you nothing about $k+1$, but breaking a composite $k+1 = ab$ sends you to the *smaller* factors a and b, and those are the cases strong induction makes available.

Theorem 0.6.11

Existence of prime factorizations

Every integer $n \geq 2$ is a product of (one or more) prime numbers.

Proof 0.6.11

Proof by Strong Induction of the existence of a prime factorization.

(1) Let $P(n)$ be "n is a product of primes," to be proved for all $n \geq 2$; here a single prime counts as a product of one factor. Fix $k \geq 2$ and assume the strong inductive hypothesis: $P(j)$ holds for every j with $2 \leq j < k$. We must prove $P(k)$.

(2) Split into two cases. If k is prime, then k is itself a product of one prime and $P(k)$ holds with nothing further to do.

(3) If k is not prime, then by definition $k = a \cdot b$ for some integers a, b with $2 \leq a < k$ and $2 \leq b < k$. Both a and b are smaller than k and at least 2, so the inductive hypothesis applies to each: a is a product of primes and b is a product of primes.

(4) Concatenating those two lists of prime factors expresses $k = a \cdot b$ as a product of primes, so $P(k)$ holds. In either case $P(k)$ follows from the cases below it, and by 0.6.9 every integer $n \geq 2$ has a prime factorization.

■

[A10+A12]

Remark 0.6.12

Ordinary induction would stall here

Try to run 0.6.11 with the ordinary form and you will feel the wall: to factor $k+1$ you would need a factorization of k, but there is no way to turn one into the other—k and $k+1$ share no factors. The honest reduction is to the divisors a, b of $k+1$, whose only guaranteed property is that they are *smaller*. That is precisely the gap strong induction fills, and recognising it is the whole skill: when the natural sub-problem is "some smaller case," not "the previous case," use 0.6.9.

Three principles—ordinary induction, strong induction, and well-ordering—keep reappearing, and it is worth knowing they are interchangeable.

Remark 0.6.13

Three statements of one idea

Over the naturals, well-ordering (0.6.1), the principle of induction (0.6.2), and strong induction (0.6.9) are *logically equivalent*: assume any one and you can derive the other two. We have already gone halfway—well-ordering gave us induction (0.6.2) and strong induction (0.6.10); the return trip, deriving well-ordering from induction, is set as 0.E.46. So the choice among them is one of convenience, never of strength. A proof that "descends to a smaller counterexample" (well-ordering), one that "adds one to climb" (induction), and one that "reduces to smaller pieces" (strong induction) are three dialects of the same underlying fact about $\mathbb{N}$.

Induction is delicate in one specific way, and the failure mode is worth meeting head-on, because it is invisible until you look for it: *a flawless inductive step proves nothing without a true base case*, and a step that quietly fails at one early value sinks the whole argument. The most famous illustration is a deliberately false "theorem."

Counterexample 0.6.14

All horses are the same colour

False claim. In any group of n horses, all n have the same colour. The bogus "induction" runs as follows. Base case $n = 1$: a single horse trivially matches itself. Inductive step: assume any k horses share a colour, and take a group of $k+1$ horses $h_1, \ldots, h_{k+1}$. The first k of them, $h_1, \ldots, h_k$, are a group of k, hence one colour; the last k of them, $h_2, \ldots, h_{k+1}$, are also a group of k, hence one colour. The two overlapping groups force all $k+1$ horses to the common colour—"therefore" all horses are the same colour.

FAILS: The inductive step is valid for every $k \geq 2$ but *breaks at* $k = 1$: passing from 1 horse to 2, the two groups of size $k = 1$ are $\{h_1\}$ and $\{h_2\}$, which do *not overlap*, so nothing links the colour of h_1 to that of h_2. The chain $P(1) \Rightarrow P(2)$ is the one missing link, and with it gone the dominoes past the first never fall.

Remark 0.6.15

The lesson: check the step at the very first transition

The horse argument is seductive precisely because the inductive step *looks* uniform—it seems to work "for all k"—while in truth it silently requires the two sub-groups to overlap, which needs $k \geq 2$. The argument proves $P(k) \Rightarrow P(k+1)$ for $k \geq 2$ and proves $P(1)$, but it never proves $P(1) \Rightarrow P(2)$, so the cascade is severed at the start. Two habits inoculate you. First, always verify the base case explicitly; an induction with a false or missing base proves nothing (the step "$P(k) \Rightarrow P(k+1)$ for all k" is consistent with P being false everywhere). Second, test the inductive step at its *first* instance—here $k = 1$—where hidden size assumptions surface. Most flawed inductions die at exactly that first transition.

A picture makes the mechanism plain: the base case lights the first domino, the step is the spacing that lets each fall onto the next, and a single missing tile—like the broken $P(1) \Rightarrow P(2)$ above—halts the cascade.

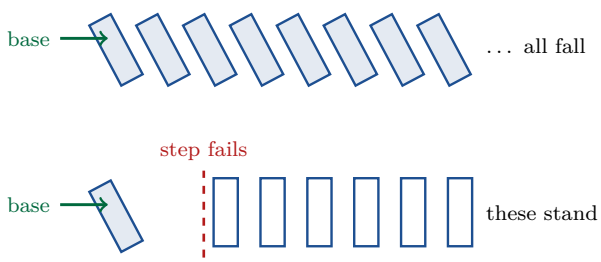


Figure 0.6.16. THE DOMINO CASCADE, AND THE GAP THAT STOPS IT. Top: base case plus uniform step topple the whole line. Bottom: the step fails at the very first transition, so every later domino stands.

With the mechanism, the two forms, and the pitfall in hand, you have mode A10 in full. The exercises that follow drill the shape on identities, inequalities, divisibility, and a recursion where only the strong form will do; treat the first few as warm-ups and the last as the real test of whether you have internalised *which* earlier case the step truly needs.

0.E Exercises

These are the foundations exercises, gathered here and grouped by topic. Each is worked in full; treat the solution as a model of how much detail a written proof is expected to carry, but try every problem yourself before reading on—reading a solution is no substitute for the struggle that makes it stick.

Statements and logic

Exercise 0.E.1 [excluded middle]

Prove that $P \vee \neg P$ is true regardless of the truth value of P. (This law is called the *law of the excluded middle*.)

Solution.

Whatever value P takes, exactly one of P and $\neg P$ is true, so their disjunction has a true disjunct and is therefore true. The truth table records both cases:

P	$\neg P$	$P \vee \neg P$
T	F	T
F	T	T

The final column is T in every row, so $P \vee \neg P$ is true for all P. A statement true under every assignment of truth values to its parts, like this one, is called a *tautology*.

■

Exercise 0.E.2 [contrapositive]

Show that an implication and its contrapositive are logically equivalent: $P \Rightarrow Q$ and $\neg Q \Rightarrow \neg P$ have the same truth value in every case.

Solution.

Compare the two final columns:

P	Q	$P \Rightarrow Q$	$\neg Q \Rightarrow \neg P$	agree?
T	T	T	T	✓
T	F	F	F	✓
F	T	T	T	✓
F	F	T	T	✓

For $\neg Q \Rightarrow \neg P$: it is false only when its hypothesis $\neg Q$ is true and its conclusion $\neg P$ is false, that is, when Q is false and P is true—exactly the one case where $P \Rightarrow Q$ is false. The two columns therefore agree in all four rows, so the statements are logically equivalent. This is what licenses proving "if P then Q" by instead assuming $\neg Q$ and deriving $\neg P$ (0.5).

■

Exercise 0.E.3 [De Morgan for connectives]

Prove that $\neg(P \wedge Q)$ and $\neg P \vee \neg Q$ are logically equivalent, and state the companion law for $\neg(P \vee Q)$.

SOLUTION.

The two are equivalent; compare the third and sixth columns.

P	Q	$\neg(P \wedge Q)$	$\neg P$	$\neg Q$	$\neg P \vee \neg Q$
T	T	F	F	F	F
T	F	T	F	T	T
F	T	T	T	F	T
F	F	T	T	T	T

They agree in every row, so $\neg(P \wedge Q) \equiv \neg P \vee \neg Q$: the negation of "both" is "at least one fails." The companion law is $\neg(P \vee Q) \equiv \neg P \wedge \neg Q$, the negation of "at least one" being "both fail." These are the logical source of the set-theoretic De Morgan laws of 0.3.14.

■

Exercise 0.E.4 [implication as "not-or"]

Show that $P \Rightarrow Q$ is logically equivalent to $\neg P \vee Q$.

SOLUTION.

Equivalent, by the truth table.

P	Q	$P \Rightarrow Q$	$\neg P \vee Q$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

The two final columns match in all four rows. The equivalence is worth keeping: an implication asserts nothing more than "the hypothesis fails or the conclusion holds," which is why a false hypothesis makes $P \Rightarrow Q$ true outright—the vacuous truth of 0.1.6 read straight off the disjunction. Negating both sides gives $\neg(P \Rightarrow Q) \equiv P \wedge \neg Q$: an implication fails exactly when its hypothesis holds and its conclusion does not.

■

Exercise 0.E.5 [converse and inverse]

Show that the converse $Q \Rightarrow P$ and the inverse $\neg P \Rightarrow \neg Q$ of an implication $P \Rightarrow Q$ are logically equivalent to each other. Give an example where the original is true yet both fail.

Solution.

The inverse is the contrapositive of the converse: contraposing $Q \Rightarrow P$ swaps and negates its parts to give $\neg P \Rightarrow \neg Q$, and a statement and its contrapositive always agree (0.1.7). So converse and inverse share a truth value in every case—which is why the warning against "assuming the converse" is the same warning against the inverse.

For the example, let P be "n is a multiple of 4" and Q be "n is even." Then $P \Rightarrow Q$ is true, the converse "every even n is a multiple of 4" is false ($n = 6$), and so its equivalent inverse "if n is not a multiple of 4 then n is odd" is false as well ($n = 6$ again). The original can hold while both converse and inverse fail.

■

Exercise 0.E.6 [necessary and sufficient]

Rewrite each as an implication and decide whether it is true: (a) "$6 \mid n$ is sufficient for n to be even"; (b) "n being even is necessary for $6 \mid n$."

Solution.

Both encode the same true implication $6 \mid n \Rightarrow 2 \mid n$. By 0.1.9, "$A$ is sufficient for B" means $A \Rightarrow B$, so (a) is "$6 \mid n \Rightarrow n$ even"; and "B is necessary for A" also means $A \Rightarrow B$ (without B, A cannot hold), so (b) is again "$6 \mid n \Rightarrow n$ even." The implication is true: if $n = 6k$ then $n = 2(3k)$ is even. The point is that "sufficient" and "necessary" name the two ends of one arrow, and reading the arrow the right way round is half of reading a theorem.

■

Quantifiers

Exercise 0.E.7 [order of quantifiers]

For the predicate $P(x, y)$ given by "$x + y = 0$" on $\mathbb{R}$, decide the truth value of each of $\forall x\, \exists y\, P(x, y)$ and $\exists y\, \forall x\, P(x, y)$, with proof.

Solution.

The first is true and the second is false.

For $\forall x\, \exists y\, (x + y = 0)$: fix an arbitrary $x \in \mathbb{R}$ and take $y = -x$; then $x + y = x + (-x) = 0$, so the inner existential holds. Since x was arbitrary, the universal holds.

For $\exists y\, \forall x\, (x + y = 0)$: suppose some fixed $y \in \mathbb{R}$ satisfied $x + y = 0$ for every x. Taking $x = 0$ forces $y = 0$; taking $x = 1$ then forces $1 + 0 = 0$, which is false. No such y exists, so

the statement is false. The two statements differ only in the order of the quantifiers, and that reversal flips the truth value—exactly the phenomenon of 0.2.7.

■

Exercise 0.E.8 [negating a definition]

Using only the mechanical rule of 0.2.5, write the negation of

$$\forall\, \varepsilon > 0,\ \ \exists\, \delta > 0,\ \ \forall x,\ \ \big(|x - a| < \delta \ \Rightarrow\ |f(x) - f(a)| < \varepsilon\big),$$

moving every negation inward until it sits on the innermost predicate.

SOLUTION.

Sweep left to right, flipping each quantifier, and finally negate the implication. The negation of an implication $A \Rightarrow B$ is $A \wedge \neg B$—it fails exactly when the hypothesis holds and the conclusion does not (0.1.5)—so $\neg(|x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon)$ becomes $|x - a| < \delta \ \wedge\ |f(x) - f(a)| \geq \varepsilon$. Hence the negation is

$$\exists\, \varepsilon > 0,\ \ \forall\, \delta > 0,\ \ \exists x,\ \ \big(|x - a| < \delta \ \wedge\ |f(x) - f(a)| \geq \varepsilon\big).$$

In words: some $\varepsilon > 0$ resists every δ—for each δ there is a point within δ of a whose image is at least ε away from $f(a)$. This is precisely what it means for f to be *discontinuous* at a, and we obtained it with no picture at all, purely by the negation rule.

■

Exercise 0.E.9 [negation and counterexample]

Write the negation of "$\forall n \in \mathbb{N},\ (n \text{ is prime} \Rightarrow n \text{ is odd})$," and decide whether the statement or its negation is true.

SOLUTION.

The negation is $\exists n \in \mathbb{N},\ (n \text{ is prime and } n \text{ is even})$, and *it* is the true one. Negating the universal turns it existential and negates the inside (0.2.5); the inner $\neg(P \Rightarrow Q)$ is $P \wedge \neg Q$ (0.1.5), here "n prime and n not odd," i.e. prime and even. The witness is $n = 2$: prime and even, so the negation holds and the original universal is false. A single even prime settles it—the counterexample pattern of 0.2.4.

■

Exercise 0.E.10 [unique existence]

Prove that there is exactly one positive real number x with $x^3 = x$.

SOLUTION.

The unique such x is 1; existence and uniqueness are separate jobs (0.2.9). Existence: $1^3 = 1$, so $x = 1$ works. Uniqueness: if $x > 0$ and $x^3 = x$, then $x^3 - x = x(x-1)(x+1) = 0$, so $x \in \{0, 1, -1\}$, of which only 1 is positive. Hence $x = 1$ is the sole positive solution, and $\exists!\, x > 0,\ x^3 = x$.

■

Exercise 0.E.11 [quantifier order]

Decide, with proof, the truth value of each (taking $n \geq 1$): (a) $\forall \varepsilon > 0,\ \exists n \in \mathbb{N},\ \frac{1}{n} < \varepsilon$; (b) $\exists n \in \mathbb{N},\ \forall \varepsilon > 0,\ \frac{1}{n} < \varepsilon$.

Solution.

Statement (a) is true and (b) is false—the same symbols in the opposite order.

For (a), fix an arbitrary $\varepsilon > 0$; by the Archimedean property (1.7) there is an integer $n \geq 1$ with $n > 1/\varepsilon$, and then $\frac{1}{n} < \varepsilon$. The witness n may depend on ε, which is the whole point (0.2.8).

For (b), one n would have to serve every $\varepsilon > 0$ at once. Given any fixed $n \geq 1$, take $\varepsilon = \frac{1}{n} > 0$; then $\frac{1}{n} < \varepsilon$ reads $\frac{1}{n} < \frac{1}{n}$, false. So no n survives and (b) fails. Reversing the quantifiers turned a truth into a falsehood (0.2.7)—the structure underneath every ε–N argument to come.

■

Exercise 0.E.12 [negating boundedness]

A sequence (a_n) is *bounded* if $\exists\, M > 0,\ \forall n,\ |a_n| \leq M$. Write, by the mechanical rule, what it means for (a_n) to be *un*bounded.

Solution.

Unbounded means $\forall\, M > 0,\ \exists\, n,\ |a_n| > M$. Negating the definition flips each quantifier and negates the inside (0.2.5): $\neg\exists M$ becomes $\forall M$, $\neg\forall n$ becomes $\exists n$, and $\neg(|a_n| \leq M)$ is $|a_n| > M$. In words: no bound M works, because for every proposed M some term overshoots it. This is the same mechanical negation that unfolds the failure of any definition built from stacked quantifiers.

■

Sets

Exercise 0.E.13 [membership vs. subset]

Decide, with a one-line reason for each, whether the following are true or false: (a) $2 \in \{1, 2, 3\}$; (b) $\{2\} \in \{1, 2, 3\}$; (c) $\{2\} \subseteq \{1, 2, 3\}$; (d) $\varnothing \in \{1, 2, 3\}$; (e) $\varnothing \subseteq \{1, 2, 3\}$; (f) $2 \subseteq \{1, 2, 3\}$.

SOLUTION.

The verdicts are: (a) true, (b) false, (c) true, (d) false, (e) true, (f) ill-formed—not a true-or-false statement at all.

For (a), 2 is one of the listed members, so $2 \in \{1, 2, 3\}$. For (b), the elements of $\{1, 2, 3\}$ are the numbers 1, 2, 3; the *set* $\{2\}$ is not among them, so $\{2\} \notin \{1, 2, 3\}$. For (c), $\{2\} \subseteq \{1, 2, 3\}$ asks whether every element of $\{2\}$ lies in $\{1, 2, 3\}$; the only element is 2, which does, so the inclusion holds. For (d), $\varnothing$ is not one of the listed members $1, 2, 3$, so $\varnothing \notin \{1, 2, 3\}$. For (e), $\varnothing \subseteq \{1, 2, 3\}$ is true by 0.3.8: the empty set is a subset of every set. For (f), "$2 \subseteq \{1, 2, 3\}$" compares a number against a set with $\subseteq$, which only relates a set to a set; 2 is not a set, so the statement is malformed—it has no truth value, and "ill-formed" is itself the answer. The whole point of the exercise is the $\in$ versus $\subseteq$ distinction of 0.3.17: $\in$ asks for membership, $\subseteq$ asks that every member be shared.

■

Exercise 0.E.14 [vacuous truth]

Prove that for every $x \in \varnothing$, one has $x^2 = 9$.

SOLUTION.

The statement is true, vacuously. Written out, the claim is the universal implication $\forall x,\ (x \in \varnothing \Rightarrow x^2 = 9)$. Fix any x. Its hypothesis $x \in \varnothing$ is false, because the empty set has no elements (0.3.4), so the implication is true by vacuous truth (0.1.6)—there is no x in $\varnothing$ that could make it fail. Since x was arbitrary, the universal statement holds. Nothing about the number 9 entered the argument: the same proof shows every element of $\varnothing$ has any property whatsoever, which is exactly why $\varnothing \subseteq A$ for all A.

■

Exercise 0.E.15 [empty set in a power set]

Let $S = \{\varnothing, \{\varnothing\}\}$. Determine $|S|$, list $\mathcal{P}(S)$ in full, and state $|\mathcal{P}(S)|$.

SOLUTION.

S has two elements, so $|S| = 2$ and $|\mathcal{P}(S)| = 2^2 = 4$.

The two elements of S are the empty set $\varnothing$ and the one-element set $\{\varnothing\}$; these are distinct, since the first has no elements and the second has exactly one (0.3.17), so $|S| = 2$. The subsets of S are the empty set, the two singletons, and S itself:

$$\mathcal{P}(S) = \Big\{\ \varnothing,\ \{\varnothing\},\ \{\{\varnothing\}\},\ \{\varnothing, \{\varnothing\}\}\ \Big\}.$$

This has 4 elements, matching $2^{|S|} = 2^2$ from 0.3.16. The first two entries deserve a glance: $\varnothing$ appears *as a subset* of S (so $\varnothing \in \mathcal{P}(S)$), while $\{\varnothing\}$ appears once as an *element* of S

and again, wrapped in braces as $\{\{\varnothing\}\}$, as a singleton subset—the nesting is genuine, not a typo.

■

Exercise 0.E.16 [$\varnothing$ as a power-set member]

Prove that $\varnothing \in \mathcal{P}(S)$ for every set S.

Solution.

The statement holds for every S. By definition of the power set, $\varnothing \in \mathcal{P}(S)$ means exactly $\varnothing \subseteq S$ (0.3.15)—the elements of $\mathcal{P}(S)$ are precisely the subsets of S. And $\varnothing \subseteq S$ holds for every set S by 0.3.8: every element of the empty set lies in S, vacuously, since there are no elements of the empty set to check (0.1.6). Therefore $\varnothing \in \mathcal{P}(S)$. The exercise is really 0.3.17 in action: the subset fact about $\varnothing$ becomes a membership fact about $\mathcal{P}(S)$.

■

Exercise 0.E.17 [absorbing/identity laws]

Let A be any set. Prove that $A \cap \varnothing = \varnothing$ and $A \cup \varnothing = A$.

Solution.

Both identities hold for every A, and each is proved by double inclusion (0.3.7).

For $A \cap \varnothing = \varnothing$: the inclusion $\varnothing \subseteq A \cap \varnothing$ is automatic from 0.3.8. Conversely, fix $x \in A \cap \varnothing$; then $x \in \varnothing$, which is impossible, so there is no such x and $A \cap \varnothing \subseteq \varnothing$ holds vacuously. Both inclusions give $A \cap \varnothing = \varnothing$.

For $A \cup \varnothing = A$: fix $x \in A \cup \varnothing$. Then $x \in A$ or $x \in \varnothing$; the second disjunct is impossible, so $x \in A$, giving $A \cup \varnothing \subseteq A$. Conversely, if $x \in A$ then $x \in A$ or $x \in \varnothing$, so $x \in A \cup \varnothing$, giving $A \subseteq A \cup \varnothing$. Both inclusions give $A \cup \varnothing = A$. (Velleman §1.4 collects these among the basic set identities.)

■

Exercise 0.E.18 [distributive law]

Prove the distributive law $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ for all sets A, B, C.

Solution.

The identity holds, and we prove it by double inclusion (0.3.7), translating the distributive law for "and" over "or" from 0.1.

For $\subseteq$, fix $x \in A \cap (B \cup C)$. Then $x \in A$, and $x \in B \cup C$, so $x \in B$ or $x \in C$. In the first case $x \in A$ and $x \in B$, so $x \in A \cap B$; in the second $x \in A$ and $x \in C$, so $x \in A \cap C$. Either

way

$$x \in (A \cap B) \cup (A \cap C),$$

so $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

For $\supseteq$, fix $x \in (A \cap B) \cup (A \cap C)$, so $x \in A \cap B$ or $x \in A \cap C$. In both cases $x \in A$, and in the first $x \in B$ while in the second $x \in C$, so in either case $x \in B$ or $x \in C$, i.e. $x \in B \cup C$. Hence $x \in A$ and $x \in B \cup C$, that is $x \in A \cap (B \cup C)$, giving $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. Both inclusions hold, so the sets are equal. The engine throughout is the logical equivalence "$P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R)$," which one reads straight off the truth tables of 0.1.3; the set proof is its faithful shadow.

■

Exercise 0.E.19 [subset and complement]

Let A, B be subsets of a universe U. Prove that $A \subseteq B$ if and only if $B^c \subseteq A^c$.

SOLUTION.

The biconditional holds; we prove both directions, and each is a contrapositive in disguise.

Assume $A \subseteq B$, and fix $x \in B^c$, so $x \in U$ and $x \notin B$. Were $x \in A$, then $A \subseteq B$ would force $x \in B$, a contradiction; so $x \notin A$, hence $x \in A^c$. As x was arbitrary, $B^c \subseteq A^c$.

Conversely assume $B^c \subseteq A^c$. Fix $x \in A$, so $x \in U$. Were $x \notin B$, then $x \in B^c$, and the hypothesis would give $x \in A^c$, i.e. $x \notin A$, contradicting $x \in A$; so $x \in B$. As x was arbitrary, $A \subseteq B$. Both directions hold, so $A \subseteq B \iff B^c \subseteq A^c$. This is the set-theoretic form of contrapositive equivalence (0.1.7): passing to complements reverses an inclusion, just as contraposition reverses an implication.

■

Exercise 0.E.20 [De Morgan for an indexed family]

Let $\{A_\alpha\}_{\alpha \in I}$ be an indexed family of subsets of a universe U, with $I \neq \varnothing$. Prove the infinite De Morgan law

$$\Big(\bigcup_{\alpha \in I} A_\alpha\Big)^c = \bigcap_{\alpha \in I} A_\alpha^c.$$

SOLUTION.

The identity holds for any index set I, finite or infinite; the proof is double inclusion (0.3.7), driven by the quantifier negation rule of 0.2.5.

For $\subseteq$, fix $x \in \big(\bigcup_\alpha A_\alpha\big)^c$, so $x \in U$ and $x \notin \bigcup_\alpha A_\alpha$. By definition of the indexed union (0.3.22), $x \in \bigcup_\alpha A_\alpha$ says "$\exists\, \alpha \in I,\ x \in A_\alpha$"; its negation is "$\forall\, \alpha \in I,\ x \notin A_\alpha$" (0.2.5). So $x \notin A_\alpha$ for every α, i.e. $x \in A_\alpha^c$ for every α, which is exactly $x \in \bigcap_\alpha A_\alpha^c$.

For $\supseteq$, fix $x \in \bigcap_\alpha A_\alpha^c$, so $x \in U$ and $x \in A_\alpha^c$ for every α, that is $x \notin A_\alpha$ for every α. Then no α puts x in A_α, so $x \notin \bigcup_\alpha A_\alpha$, hence $x \in \left(\bigcup_\alpha A_\alpha\right)^c$. Both inclusions hold, so the sets are equal. This is the two-set law of 0.3.14 with "not-or" upgraded to "not-some": negating the existential that defines a union produces the universal that defines an intersection, which is precisely 0.2.5.

■

Relations and functions

Exercise 0.E.21 [injectivity from a formula]

Let $f\colon \mathbb{R} \to \mathbb{R}$ be $f(x) = 3x - 7$. Prove that f is injective and that it is surjective, and write down its inverse.

Solution.

The map is a bijection with inverse $f^{-1}(y) = (y + 7)/3$.

For injectivity, take $x, y \in \mathbb{R}$ with $f(x) = f(y)$, that is $3x - 7 = 3y - 7$. Adding 7 and dividing by 3 gives $x = y$, which is the defining condition of 0.4.9.

For surjectivity, fix an arbitrary $b \in \mathbb{R}$ and look for a witness by solving $3x - 7 = b$; this forces $x = (b + 7)/3$, a genuine real number. Then $f((b + 7)/3) = 3 \cdot \frac{b+7}{3} - 7 = b$, so b is hit. Being injective and surjective, f is bijective, hence invertible (0.4.15). The witness just constructed is exactly the inverse rule: $f^{-1}(y) = (y + 7)/3$, and one checks $f^{-1}(f(x)) = \frac{(3x-7)+7}{3} = x$ and $f(f^{-1}(y)) = 3 \cdot \frac{y+7}{3} - 7 = y$.

■

Exercise 0.E.22 [a collision]

Let $g\colon \mathbb{R} \to \mathbb{R}$ be $g(x) = x^2 - 4x$. Decide whether g is injective and whether it is surjective, with proof.

Solution.

It is neither injective nor surjective.

To break injectivity, exhibit one collision. Completing the square, $g(x) = (x - 2)^2 - 4$, so g is symmetric about $x = 2$: $g(0) = 0$ and $g(4) = 16 - 16 = 0$, hence $g(0) = g(4)$ while $0 \neq 4$. A single such pair defeats the universal in 0.4.9.

To break surjectivity, find a target no input reaches. Since $(x - 2)^2 \geq 0$ for every real x, we have $g(x) = (x - 2)^2 - 4 \geq -4$, so the value -5 is never attained; the existential $\exists\, x,\ g(x) = -5$ is false. Either failure alone settles the question, and both hold here.

■

Exercise 0.E.23 [composition of injections]

Let $f\colon A \to B$ and $g\colon B \to C$ be injective. Prove that $g \circ f\colon A \to C$ is injective. Then show by example that $g \circ f$ can be injective while g is not.

SOLUTION.

The composition of injections is injective.

Take $x, y \in A$ with $(g \circ f)(x) = (g \circ f)(y)$, that is $g(f(x)) = g(f(y))$. Injectivity of g applied to the inputs $f(x), f(y) \in B$ gives $f(x) = f(y)$; injectivity of f then gives $x = y$. So $g \circ f$ satisfies 0.4.9.

The converse property fails: g need not be injective even when $g \circ f$ is. Take $A = \{0\}$, $B = \{0, 1\}$, $C = \{0\}$, with $f(0) = 0$ and g the constant map $g(0) = g(1) = 0$. Here g collapses 0 and 1, so it is not injective, yet $g \circ f\colon \{0\} \to \{0\}$ is trivially injective, having only one input. The injectivity of a composite constrains the *first* map applied, not the second.

■

Exercise 0.E.24 [preimage of a union]

Let $f\colon A \to B$ and let $T, U \subseteq B$. Prove $f^{-1}(T \cup U) = f^{-1}(T) \cup f^{-1}(U)$.

SOLUTION.

The two sets are equal, proved by showing each element of one lies in the other.

Fix $a \in A$. Unwinding the preimage from 0.4.7,

$$a \in f^{-1}(T \cup U) \iff f(a) \in T \cup U \iff (f(a) \in T \text{ or } f(a) \in U).$$

The final clause is, again by 0.4.7, the statement "$a \in f^{-1}(T)$ or $a \in f^{-1}(U)$," which is exactly $a \in f^{-1}(T) \cup f^{-1}(U)$. Each step is a biconditional, so the two sets have identical membership conditions and are equal. The argument is the union twin of 0.4.8; the only change is "and" becoming "or," reflecting that union is built from disjunction.

■

Exercise 0.E.25 [forward image can grow]

Let $f\colon A \to B$ and $S, S' \subseteq A$. Prove $f(S \cup S') = f(S) \cup f(S')$, and give an example where $f(S \cap S') \neq f(S) \cap f(S')$.

SOLUTION.

The image distributes over unions but not over intersections.

For the union, a point b lies in $f(S \cup S')$ exactly when $b = f(a)$ for some $a \in S \cup S'$, i.e. for

some $a \in S$ or some $a \in S'$; that is precisely $b \in f(S)$ or $b \in f(S')$, namely $b \in f(S) \cup f(S')$. Both inclusions hold, so the sets are equal.

For the intersection the equality can fail. Take $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, with $S = \{1\}$ and $S' = \{-1\}$. Then $S \cap S' = \emptyset$ gives $f(S \cap S') = \emptyset$, while $f(S) \cap f(S') = \{1\} \cap \{1\} = \{1\}$. The obstruction is exactly the collapsing that injectivity forbids: when f is injective the intersection equality is restored, since distinct inputs then keep distinct images.

■

Exercise 0.E.26 [verify an equivalence relation]

On $\mathbb{R}$ define $x \sim y$ to mean $x - y \in \mathbb{Z}$. Prove that $\sim$ is an equivalence relation and describe the equivalence class $[x]$.

Solution.

The relation is an equivalence relation, and $[x] = \{\, x + k : k \in \mathbb{Z} \,\}$.

It satisfies the three axioms of 0.4.16. Reflexive: $x - x = 0 \in \mathbb{Z}$, so $x \sim x$. Symmetric: if $x - y \in \mathbb{Z}$ then $y - x = -(x - y) \in \mathbb{Z}$, since the integers are closed under negation, so $y \sim x$. Transitive: if $x - y \in \mathbb{Z}$ and $y - z \in \mathbb{Z}$, then their sum $(x - y) + (y - z) = x - z \in \mathbb{Z}$, so $x \sim z$.

For the class, $y \in [x]$ means $x - y \in \mathbb{Z}$, that is $y = x - k$ for some integer k; as k ranges over $\mathbb{Z}$ so does $-k$, so $[x] = \{\, x + k : k \in \mathbb{Z} \,\}$, the integer translates of x. Each class meets the half-open interval $[0, 1)$ in exactly one point, the fractional part of x, so $[0, 1)$ is a complete set of class representatives and the quotient $\mathbb{R}/\!\sim$ may be pictured as a circle of circumference one (gluing 0 to 1).

■

Exercise 0.E.27 [a relation that just misses]

On $\mathbb{R}$ define $x \mathrel{R} y$ to mean $|x - y| \leq 1$. Show R is reflexive and symmetric but not transitive, so it is *not* an equivalence relation.

Solution.

The relation has the first two properties but fails the third.

Reflexive: $|x - x| = 0 \leq 1$, so $x \mathrel{R} x$ for every x. Symmetric: $|x - y| = |y - x|$, so $|x - y| \leq 1$ holds iff $|y - x| \leq 1$, giving $x \mathrel{R} y \Rightarrow y \mathrel{R} x$.

Transitivity fails, and one counterexample is enough to deny the universal in 0.4.16. Take $x = 0$, $y = 1$, $z = 2$: then $|0 - 1| = 1 \leq 1$ and $|1 - 2| = 1 \leq 1$, so $0 \mathrel{R} 1$ and $1 \mathrel{R} 2$, yet $|0 - 2| = 2 > 1$, so $0 \not\mathrel{R} 2$. Closeness within a fixed tolerance does not chain, which is why "approximately equal" is never an equivalence relation—the very obstruction that forces

analysis to work with genuine limits rather than "small enough" differences.

■

Exercise 0.E.28 [a two-sided inverse forces bijectivity]

Let $f\colon A \to B$ and suppose there is $g\colon B \to A$ with $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. Prove f is bijective *without* quoting 0.4.15, by arguing each property directly.

SOLUTION.

The function f is injective and surjective.

For injectivity, suppose $f(x) = f(y)$ with $x, y \in A$. Apply g to both sides and use $g \circ f = \mathrm{id}_A$:

$$x = g\big(f(x)\big) = g\big(f(y)\big) = y,$$

so distinct inputs cannot share an output.

For surjectivity, fix $b \in B$ and set $a = g(b) \in A$. Then $f \circ g = \mathrm{id}_B$ gives $f(a) = f(g(b)) = b$, so b has the preimage a. Holding for every b, this is the existential of 0.4.9. Hence f is bijective. The point worth keeping is that injectivity used only the left inverse and surjectivity only the right inverse—each side of the inverse equation carries exactly one of the two properties.

■

Exercise 0.E.29 [a bijection between intervals]

Construct an explicit bijection $f\colon (0,1) \to (a,b)$ for real numbers $a < b$, and prove it is one. Conclude that any two bounded open intervals are equinumerous.

SOLUTION.

The affine map $f(t) = a + (b-a)\,t$ is a bijection from $(0,1)$ onto (a,b).

First it lands in the target: for $t \in (0,1)$, multiplying the strict inequalities $0 < t < 1$ by the positive number $b - a$ and adding a gives $a < a + (b-a)t < b$, so $f(t) \in (a,b)$. It is injective because $f(s) = f(t)$ means $a + (b-a)s = a + (b-a)t$; cancelling a and dividing by $b - a \neq 0$ gives $s = t$. It is surjective because, given any $y \in (a,b)$, the value $t = (y-a)/(b-a)$ satisfies $0 < t < 1$ (divide $a < y < b$ by $b - a > 0$ after subtracting a) and $f(t) = a + (b-a) \cdot \frac{y-a}{b-a} = y$.

So f is a bijection, and by 0.4.21 the intervals $(0,1)$ and (a,b) are equinumerous. Since this holds for every choice of $a < b$, and equinumerosity is symmetric and transitive (0.4.21), any two bounded open intervals (a,b) and (c,d) are equinumerous, each being equinumerous with $(0,1)$. This finite-looking statement already foreshadows the surprises of 2.1: a comparable trick stretches a bounded interval onto the whole line.

■

The craft of proof

Exercise 0.E.30 [direct proof]

Prove that the sum of any two odd integers is even.

Solution.

Let m and n be odd, so $m = 2j + 1$ and $n = 2k + 1$ for integers j, k. Then

$$m + n = (2j + 1) + (2k + 1) = 2(j + k + 1),$$

and since $j + k + 1$ is an integer, $m + n$ is twice an integer, hence even.

■

Exercise 0.E.31 [contrapositive]

Let n be an integer. Prove that if $3n + 2$ is odd, then n is odd.

Solution.

We prove the contrapositive: if n is even, then $3n + 2$ is even. Suppose $n = 2k$. Then $3n + 2 = 6k + 2 = 2(3k + 1)$, which is even. By 0.1.7 the contrapositive is equivalent to the original implication, so "$3n + 2$ odd $\Rightarrow$ n odd" is proved. (A direct attack stalls: "$3n + 2$ odd" is awkward to use, while "n even" is an immediate handle—the cue to switch to the contrapositive.)

■

Exercise 0.E.32 [contradiction]

Prove that there is no smallest positive real number.

Solution.

Suppose, for contradiction, that there were a smallest positive real number, say $x > 0$, so that $x \leq y$ for every positive real y. Consider $x/2$. Since $x > 0$ we have $x/2 > 0$, so $x/2$ is itself a positive real; and $x/2 < x$ because $x > 0$. Thus $x/2$ is a positive real strictly smaller than x, contradicting that x was the smallest. No such smallest positive real exists.

■

Exercise 0.E.33 [contradiction]

Prove that $\sqrt{3}$ is irrational. (You may first prove the lemma: for an integer p, if 3 divides p^2 then 3 divides p.)

SOLUTION.

First the lemma, by contrapositive: if $3 \nmid p$, then p leaves remainder 1 or 2 on division by 3, i.e. $p = 3k+1$ or $p = 3k+2$. In the first case $p^2 = 9k^2 + 6k + 1 = 3(3k^2 + 2k) + 1$, and in the second $p^2 = 9k^2 + 12k + 4 = 3(3k^2 + 4k + 1) + 1$; either way p^2 leaves remainder 1, so $3 \nmid p^2$. Hence $3 \mid p^2 \Rightarrow 3 \mid p$.

Now suppose, for contradiction, that $\sqrt{3} = p/q$ with p, q integers in lowest terms, $q \neq 0$. Then $p^2 = 3q^2$, so $3 \mid p^2$, and by the lemma $3 \mid p$, say $p = 3r$. Substituting, $9r^2 = 3q^2$, so $q^2 = 3r^2$, whence $3 \mid q^2$ and, by the lemma again, $3 \mid q$. But then 3 divides both p and q, contradicting lowest terms. Therefore $\sqrt{3}$ is irrational.

■

Exercise 0.E.34 [counterexample]

Disprove: for all real numbers a and b, $(a+b)^2 = a^2 + b^2$.

SOLUTION.

The statement is false; one counterexample suffices to disprove a universal (0.5.8). Take $a = b = 1$: then $(a+b)^2 = 2^2 = 4$, while $a^2 + b^2 = 1 + 1 = 2$, and $4 \neq 2$. (The identity fails for any a, b with $ab \neq 0$, since $(a+b)^2 = a^2 + b^2 + 2ab$; a single instance is all a disproof needs.)

■

Exercise 0.E.35 [cases]

Prove that $n^3 - n$ is divisible by 6 for every integer n.

SOLUTION.

Factor $n^3 - n = (n-1)\,n\,(n+1)$, a product of three consecutive integers. Among any two consecutive integers one is even, so the product is divisible by 2; among any three consecutive integers one is divisible by 3, so the product is divisible by 3. Being divisible by both 2 and 3, and these having no common factor, the product is divisible by 6. Hence $6 \mid n^3 - n$.

■

Exercise 0.E.36 [biconditional]

Prove that an integer n is even if and only if n^2 is even.

SOLUTION.

A biconditional needs both directions (0.5.9). ($\Rightarrow$) If n is even, $n = 2k$, then $n^2 = 4k^2 = 2(2k^2)$ is even—a direct argument. ($\Leftarrow$) If n^2 is even then n is even; this is the contrapositive proof already given in 0.5.4 (an odd n has an odd square). Both directions

holding, n is even exactly when n^2 is.

■

Exercise 0.E.37 [prove or disprove]

Prove or disprove: the sum of two irrational numbers is always irrational.

Solution.

False. Take $a = \sqrt{2}$ and $b = -\sqrt{2}$; both are irrational (0.5.5), yet $a + b = 0$ is rational. The claim is a universal statement, so this single counterexample disproves it. (What *is* true is narrower: an irrational plus a *rational* is irrational—a good direct exercise on its own.)

■

Mathematical induction

Exercise 0.E.38 [summation identity]

Prove that for every integer $n \geq 1$,

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Solution.

The identity holds for all $n \geq 1$, by induction on n. For the base case $n = 1$ the left side is $1^2 = 1$ and the right side is $\frac{1\cdot 2\cdot 3}{6} = 1$, so the two agree.

Fix $k \geq 1$ and assume $1^2 + \cdots + k^2 = \frac{k(k+1)(2k+1)}{6}$. Adding the next term $(k+1)^2$ and putting everything over 6,

$$\begin{aligned}\sum_{i=1}^{k+1} i^2 &= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\ &= \frac{(k+1)\big[k(2k+1) + 6(k+1)\big]}{6} = \frac{(k+1)\big(2k^2+7k+6\big)}{6}.\end{aligned}$$

The quadratic factors as $2k^2 + 7k + 6 = (k+2)(2k+3)$, so the right side becomes $\frac{(k+1)(k+2)(2k+3)}{6}$, which is the formula with $k+1$ in place of n. The inductive step holds, and by 0.6.2 the identity is valid for all $n \geq 1$.

■

Exercise 0.E.39 [telescoping sum]

Prove that for every integer $n \geq 1$,

$$\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \cdots + \frac{1}{n(n+1)} = \frac{n}{n+1}.$$

SOLUTION.

The identity holds for all $n \geq 1$. For the base case $n = 1$ the left side is $\frac{1}{1 \cdot 2} = \frac{1}{2}$ and the right side is $\frac{1}{2}$, which agree.

Assume the identity for a fixed $k \geq 1$. Adding the term $\frac{1}{(k+1)(k+2)}$ to both sides,

$$\begin{aligned} \sum_{i=1}^{k+1} \frac{1}{i(i+1)} &= \frac{k}{k+1} + \frac{1}{(k+1)(k+2)} = \frac{k(k+2)+1}{(k+1)(k+2)} \\ &= \frac{(k+1)^2}{(k+1)(k+2)} = \frac{k+1}{k+2}, \end{aligned}$$

which is the formula at $k+1$. By 0.6.2 it holds for every $n \geq 1$. (The name "telescoping" fits: since $\frac{1}{i(i+1)} = \frac{1}{i} - \frac{1}{i+1}$, the sum collapses to $1 - \frac{1}{n+1} = \frac{n}{n+1}$, which is the quicker route once you see it.)

■

Exercise 0.E.40 [geometric sum]

Let r be a real number with $r \neq 1$. Prove that for every integer $n \geq 0$,

$$1 + r + r^2 + \cdots + r^n = \frac{r^{n+1} - 1}{r - 1}.$$

SOLUTION.

The formula holds for all $n \geq 0$, by induction on n. For the base case $n = 0$ the left side is the single term 1 and the right side is $\frac{r^1 - 1}{r - 1} = \frac{r-1}{r-1} = 1$ (well defined since $r \neq 1$), so they agree.

Fix $k \geq 0$ and assume $1 + r + \cdots + r^k = \frac{r^{k+1} - 1}{r - 1}$. Adding r^{k+1},

$$\begin{aligned} 1 + r + \cdots + r^k + r^{k+1} &= \frac{r^{k+1} - 1}{r - 1} + r^{k+1} \\ &= \frac{r^{k+1} - 1 + r^{k+1}(r-1)}{r - 1} = \frac{r^{k+2} - 1}{r - 1}, \end{aligned}$$

after cancelling $r^{k+1} - r^{k+1} = 0$ in the numerator. This is the formula at $k+1$, so by 0.6.2 it holds for all $n \geq 0$. Taking $r = 2$ recovers 0.6.6(b).

■

Exercise 0.E.41 [Bernoulli's inequality]

Let x be a real number with $x \geq -1$. Prove *Bernoulli's inequality*: for every $n \in \mathbb{N}$,

$$(1+x)^n \geq 1 + nx.$$

Solution.

The inequality holds for all $n \in \mathbb{N}$ and every fixed $x \geq -1$. For the base case $n = 0$ both sides equal 1 (since $(1+x)^0 = 1$ and $1 + 0 \cdot x = 1$), so $\geq$ holds.

Fix $k \geq 0$ and assume $(1+x)^k \geq 1 + kx$. The hypothesis $x \geq -1$ makes $1 + x \geq 0$, so multiplying the inductive hypothesis by the nonnegative factor $1 + x$ preserves the inequality:

$$(1+x)^{k+1} = (1+x)^k(1+x) \;\geq\; (1+kx)(1+x) = 1 + (k+1)x + kx^2.$$

Since $kx^2 \geq 0$, the right side is at least $1 + (k+1)x$, hence $(1+x)^{k+1} \geq 1 + (k+1)x$, which is the claim at $k+1$. By 0.6.2 the inequality holds for all $n \in \mathbb{N}$. The step used $1 + x \geq 0$ in an essential way; without $x \geq -1$ the factor could be negative and reverse the inequality.

■

Exercise 0.E.42 [divisibility]

Prove that for every $n \in \mathbb{N}$, $6 \mid (n^3 - n)$. (This sharpens 0.6.8.)

Solution.

The claim holds for all $n \in \mathbb{N}$. For the base case $n = 0$, $0^3 - 0 = 0 = 6 \cdot 0$, divisible by 6.

Fix $k \geq 0$ and assume $6 \mid (k^3 - k)$, say $k^3 - k = 6q$ for some integer q. As in 0.6.8,

$$(k+1)^3 - (k+1) = (k^3 - k) + 3(k^2 + k) = 6q + 3k(k+1).$$

Among the two consecutive integers k and $k+1$ one is even, so their product $k(k+1)$ is even, say $k(k+1) = 2m$; then $3k(k+1) = 6m$. Hence

$$(k+1)^3 - (k+1) = 6q + 6m = 6(q+m),$$

a multiple of 6, which is the claim at $k+1$. By 0.6.2 it holds for every $n \in \mathbb{N}$. The extra factor of 2 over 0.6.8 comes exactly from the even member of the consecutive pair $k, k+1$.

■

Exercise 0.E.43 [strong induction (Fibonacci)]

Define the Fibonacci numbers by $F_1 = 1$, $F_2 = 1$, and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. Prove that $F_n < 2^n$ for every $n \geq 1$.

SOLUTION.

The bound $F_n < 2^n$ holds for all $n \geq 1$, by strong induction (0.6.9); the recursion reaches *two* steps back, so the strong form is the natural tool. Because the recursion only kicks in at $n \geq 3$, we verify the two starting values directly: $F_1 = 1 < 2 = 2^1$ and $F_2 = 1 < 4 = 2^2$.

Fix $k \geq 2$ and assume $F_j < 2^j$ for every j with $1 \leq j \leq k$. Then $k+1 \geq 3$, so the recursion gives $F_{k+1} = F_k + F_{k-1}$, and both k and $k-1$ lie in the range covered by the hypothesis. Hence

$$F_{k+1} = F_k + F_{k-1} < 2^k + 2^{k-1} = 2^{k-1}(2+1) = 3 \cdot 2^{k-1} < 4 \cdot 2^{k-1} = 2^{k+1}.$$

Thus $F_{k+1} < 2^{k+1}$, and by 0.6.9 the bound holds for all $n \geq 1$. Ordinary induction would stall: F_{k+1} depends on F_{k-1} as well as F_k, and only the strong hypothesis supplies both at once.

■

Exercise 0.E.44 [strong induction (coins)]

Prove that every integer amount of $n \geq 8$ cents can be paid exactly using only 3-cent and 5-cent coins.

SOLUTION.

Every $n \geq 8$ is so payable, by strong induction (0.6.9). Three base cases are needed because the step will reach three back: $8 = 3 + 5$, $9 = 3 + 3 + 3$, and $10 = 5 + 5$ are each payable.

Fix $k \geq 10$ and assume every amount j with $8 \leq j \leq k$ is payable. Consider $k + 1$. Since $k + 1 \geq 11$, the smaller amount $(k + 1) - 3 = k - 2$ satisfies $k - 2 \geq 8$, so by the inductive hypothesis $k - 2$ is payable with 3- and 5-cent coins. Adding one more 3-cent coin pays $k + 1$. Hence every $n \geq 8$ is payable; the three base cases plus this step cover all of them by 0.6.9.

The reduction is to the case three smaller, not the immediate predecessor, which is exactly why the strong form and three base cases are the honest setup. (The amount 7 is *not* payable, so 8 is the true threshold.)

■

Exercise 0.E.45 [find the flaw]

The following purports to prove that every natural number is equal to 0. Find the exact step that fails. *"Claim: for all $n \in \mathbb{N}$, $n = 0$. Strong induction. Base case $n = 0$: $0 = 0$. Step: fix*

k and assume $j = 0$ for all j with $0 \le j \le k$. Write $k+1 = a+b$ with $a, b \in \mathbb{N}$ and $a, b \le k$. By the hypothesis $a = 0$ and $b = 0$, so $k+1 = a+b = 0$. Hence $n = 0$ for all n."

Solution.

The fatal line is the claim "write $k+1 = a+b$ with $a, b \le k$," which is impossible at the very first transition $k = 0$. When $k = 0$ we are proving $P(1)$, i.e. $1 = 0$, and we would need $1 = a+b$ with both $a, b \le 0$; the only option is $a = b = 0$, giving $a+b = 0 \neq 1$. No valid decomposition exists, so the inductive step is never established for $k = 0$, and the chain breaks immediately after the base case exactly as in the horse fallacy (0.6.14).

For $k \ge 1$ a decomposition $k+1 = a+b$ with $a, b \le k$ does exist (e.g. $a = 1$, $b = k$), so the argument *looks* uniform—but a step that holds for all $k \ge 1$ together with a base case at $k = 0$ proves nothing past $P(0)$, since the link $P(0) \Rightarrow P(1)$ is precisely the one missing. The moral of 0.6.15 applies: test the step at its first instance, where the hidden requirement (here, a decomposition into smaller summands) can fail.

■

Exercise 0.E.46 [equivalence of the principles]

Assume the principle of mathematical induction (0.6.2) and use it to *prove* well-ordering (0.6.1): every nonempty $S \subseteq \mathbb{N}$ has a least element. (Together with the converse proved in 0.6.2, this shows the two are equivalent.)

Solution.

We prove the contrapositive: a subset $S \subseteq \mathbb{N}$ with *no* least element must be empty. Suppose S has no least element. Let $P(n)$ be the statement "no natural number $m \le n$ belongs to S," and prove $P(n)$ for all $n \in \mathbb{N}$ by ordinary induction.

For the base case $n = 0$: if $0 \in S$ then 0 would be a least element of S (nothing in $\mathbb{N}$ is below it), contradicting our supposition; so $0 \notin S$, and $P(0)$ holds. For the step, fix k and assume $P(k)$, so no $m \le k$ lies in S. If $k+1$ belonged to S, then since all of $0, 1, \ldots, k$ are absent from S, $k+1$ would be the least element of S—again a contradiction. Hence $k+1 \notin S$, and combined with $P(k)$ this gives $P(k+1)$.

By 0.6.2, $P(n)$ holds for every n, meaning no natural number lies in S; that is, $S = \emptyset$. Contrapositively, every nonempty $S \subseteq \mathbb{N}$ has a least element. Since 0.6.2 derived induction from well-ordering, the two principles—and, by 0.6.10, strong induction—are logically equivalent, as recorded in 0.6.13.

■

LECTURE 1

From the Naturals to the Reals: Induction, Ordered Fields, and Completeness

MATH S4061 | Introduction to Modern Analysis I

Tuesday, 26 May 2026

God made the integers; all else is the work of man.

Leopold Kronecker

Contents of this lecture

Overview. The course opens with a deceptively simple question—is $0.999\ldots = 1$?—whose honest answer requires building the real numbers from the ground up. We begin with the natural numbers and induction, construct $\mathbb{Z}$ and $\mathbb{Q}$, and isolate the field and order axioms they satisfy. The rationals then fall short for analysis: $\sqrt{2}$ is missing, and the sets bounding it have neither a largest nor a smallest element. Repairing that gap is the *least-upper-bound property*, which characterises $\mathbb{R}$; from it follow the Archimedean property and the density of $\mathbb{Q}$—and, at last, the resolution of $0.999\ldots = 1$. The companion text is Rudin, Chapter 1, especially the ordered-field and least-upper-bound material R 1.5–R 1.20.

1.1 The Natural Numbers and Induction

Definition 1.1.1

The natural numbers

The *natural numbers* are $\mathbb{N} = \{0, 1, 2, 3, \dots\}$[6], generated from 0 by the *successor function*

$$S : \mathbb{N} \to \mathbb{N}, \qquad n \mapsto n + 1.$$

Axiom 1.1.2

Induction

If $A \subseteq \mathbb{N}$ satisfies $0 \in A$ and $n \in A \Rightarrow n + 1 \in A$, then $A = \mathbb{N}$. Rudin treats induction as prerequisite proof background rather than a numbered analysis result.

Theorem 1.1.3

The principle of mathematical induction

Let $P(n)$ be a statement for each $n \in \mathbb{N}$. If $P(0)$ holds, and $P(n) \Rightarrow P(n + 1)$ for every n, then $P(n)$ holds for all $n \in \mathbb{N}$. This is the proof principle behind later uses of induction; Rudin assumes it as background.

Proof 1.1.3

Proof of the Induction Principle from the Axiom.

(1) Let $A = \{n \in \mathbb{N} : P(n) \text{ holds}\}$.

(2) Since $P(0)$ holds, $0 \in A$; and $P(n) \Rightarrow P(n + 1)$ says $n \in A \Rightarrow n + 1 \in A$.

(3) By the induction axiom (1.1.2), $A = \mathbb{N}$; that is, $P(n)$ holds for every n.

■

[Direct] uses 1.1.2.

Example 1.1.4

Gauss's sum

For every integer $n \geq 1$, $\quad 1 + 2 + \cdots + n = \dfrac{n(n+1)}{2}$. This is included as induction practice rather than as a Rudin-numbered item.

Proof 1.1.4

Proof by Induction of Gauss's Formula.

(1) For $n = 1$ the formula reads $1 = \dfrac{1 \cdot 2}{2}$, which holds.

[6] We take $0 \in \mathbb{N}$; Rudin instead starts the positive integers at 1. The convention is harmless—only the base of an induction shifts (Gauss's sum, 1.1.4, begins at $n = 1$).

(2) Assume $1+\cdots+n=\dfrac{n(n+1)}{2}$. Adding $n+1$ to both sides,

$$1+\cdots+n+(n+1)=\frac{n(n+1)}{2}+(n+1)=\frac{(n+1)(n+2)}{2},$$

which is the formula for $n+1$. The result follows by induction (1.1.3). ∎

[A10] from the base $n=1$ (apply 1.1.3 to the shifted statement).

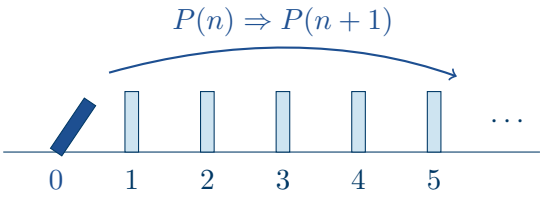


Figure 1.1.5. INDUCTION AS TOPPLING DOMINOES. The base case $P(0)$ topples the first domino; the inductive step $P(n)\Rightarrow P(n+1)$ passes the fall to the next, so every $P(n)$ falls in turn—the conclusion of 1.1.2 and 1.1.3.

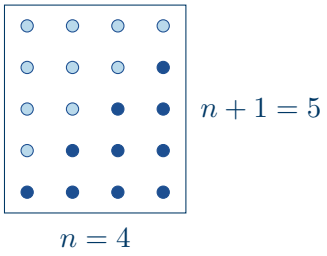


Figure 1.1.6. GAUSS'S SUM AS HALF A RECTANGLE. Two copies of the staircase $1+2+3+4$ interlock into a 4×5 grid, so $2(1+2+3+4)=4\cdot 5$; in general $1+\cdots+n=\frac{n(n+1)}{2}$ (1.1.4).

1.2 The Integers and the Rationals

Construction 1.2.1

The integers

The *integers* are signed naturals with the two zeros glued together:

$$\mathbb{Z}:=(\{+,-\}\times\mathbb{N})/\!\sim,\qquad (+,0)\sim(-,0).$$

Writing $+n,-n$ for the classes recovers $\mathbb{Z}=\{\dots,-2,-1,0,1,2,\dots\}$. Rudin treats $\mathbb{Z}$ as background and begins the analysis notation with the rational field in R 1.4.

Construction 1.2.2

The rationals

The *rationals* are equivalence classes of integer pairs with nonzero second coordinate:

$$\mathbb{Q} := \{(p,q) \in \mathbb{Z} \times \mathbb{Z} : q \neq 0\}/\sim, \qquad (p,q) \sim (r,s) \iff ps = rq.$$

The class of (p,q) is written $\dfrac{p}{q}$. Rudin records the rational-number notation and convention in R 1.4.

Remark 1.2.3

One number, many fractions

A rational *is* the whole class, e.g. $\left[\frac{1}{2}\right] = \left\{\frac{1}{2}, \frac{2}{4}, \frac{-1}{-2}, \frac{17}{34}, \dots\right\}$. In practice one names it by its *preferred representative*: the irreducible fraction $\dfrac{a}{b}$ with $\gcd(a,b) = 1$ and the sign carried in the numerator. This is the equivalence-class bookkeeping behind the notation in R 1.4.

1.3 Fields and Ordered Fields

Definition 1.3.1

Field

A *field* is a set F with two operations $+ : F \times F \to F$ and $\cdot : F \times F \to F$ satisfying eleven axioms[7]—making F an abelian group under $+$, making $F \setminus \{0\}$ an abelian group under $\cdot$, with multiplication distributing over addition.

Definition 1.3.2

Order

An *order* on a set S is a relation $<$ such that (cf. R 1.5)

(1) (*trichotomy*) for $x, y \in S$, exactly one of $x < y$, $y < x$, $x = y$ holds;

(2) (*transitivity*) $x < y$ and $y < z$ imply $x < z$.

A set carrying an order is *totally* (or *linearly*) ordered. Rudin calls this an ordered set in R 1.6.

Definition 1.3.3

Ordered field

An *ordered field* is a field F with an order $<$ compatible with the operations (R 1.17):

(1) $y < z \Rightarrow x + y < x + z$ for all x;

(2) $x > 0$ and $y > 0 \Rightarrow xy > 0$.

Remark 1.3.4

The rationals are an ordered field

[7] Five for addition (A1–A5), five for multiplication (M1–M5), and the distributive law (D); the full statements are R 1.12.

$\mathbb{Q}$ is an ordered field. The rest of the lecture asks what it *lacks*—and why that makes it unfit for analysis. Compare Rudin's ordered-field definition R 1.17 and the real-field existence theorem R 1.19.

1.4 Why the Rationals Are Not Enough

Proposition 1.4.1

No rational squares to 2

There is no $x \in \mathbb{Q}$ with $x^2 = 2$ (R 1.1).

Proof 1.4.1

Proof by Contradiction of the Irrationality of $\sqrt{2}$.

(1) Suppose $x = \dfrac{p}{q} \in \mathbb{Q}$ is irreducible with $x^2 = 2$. Then $p^2 = 2q^2$, so p^2 is even, hence p is even (an odd integer has an odd square); write $p = 2k$.

(2) Substituting, $4k^2 = 2q^2$, so $q^2 = 2k^2$ is even, hence q is even.

(3) Then $2 \mid \gcd(p, q)$, contradicting that $\dfrac{p}{q}$ is irreducible. So no such x exists.

■

[A9] cf. R 1.1; uses "n^2 even $\Rightarrow$ n even".

Proposition 1.4.2

The rationals have a gap

Let $A = \{p \in \mathbb{Q} : p > 0,\ p^2 < 2\}$ and $B = \{p \in \mathbb{Q} : p > 0,\ p^2 > 2\}$. Then A has no largest element and B has no smallest element.

Proof 1.4.2

Proof by an Explicit Construction of the Gap.

(1) Given $p > 0$, set
$$q = p - \frac{p^2 - 2}{p + 2} = \frac{2p + 2}{p + 2} > 0.$$

(2) A computation gives
$$q^2 - 2 = \frac{2\,(p^2 - 2)}{(p + 2)^2},$$
so $q^2 - 2$ has the same sign as $p^2 - 2$: if $p \in A$ then $q \in A$, and if $p \in B$ then $q \in B$.

(3) Also $q - p = -\dfrac{p^2 - 2}{p + 2}$, so $q > p$ when $p \in A$ and $q < p$ when $p \in B$.

(4) Hence every $p \in A$ is exceeded by a larger $q \in A$, so A has no largest element; and every $p \in B$ is undercut by a smaller $q \in B$, so B has no smallest element.

■

[Constr] cf. R 1.1.

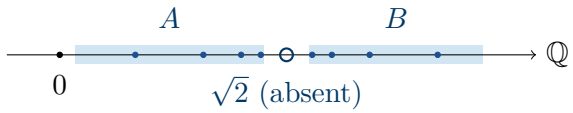


Figure 1.4.3. The gap at $\sqrt{2}$. The positive rationals A ($p^2 < 2$) climb toward $\sqrt{2}$ with no largest member; B ($p^2 > 2$) descend toward it with no smallest; in $\mathbb{Q}$ the meeting point $\sqrt{2}$ is simply absent.

1.5 Bounds and the Supremum

Definition 1.5.1

Upper bound

Let S be an ordered set and $E \subseteq S$. An element $\beta \in S$ is an *upper bound* for E if $x \leq \beta$ for every $x \in E$ (R 1.7).

Definition 1.5.2

Least upper bound (supremum)

β is the *least upper bound* of E if it is an upper bound for E and no smaller element is: $\alpha < \beta$ implies α is not an upper bound. We write $\beta = \sup E = \operatorname{lub} E$ (R 1.8).

Definition 1.5.3

Lower bound and infimum

Dually, $\beta \in S$ is a *lower bound* for E if $\beta \leq x$ for every $x \in E$, and the *greatest lower bound* (or *infimum*) if no larger element is a lower bound: $\beta < \alpha$ implies α is not a lower bound. We write $\beta = \inf E = \operatorname{glb} E$ (cf. R 1.8). Since $\inf E = -\sup(-E)$, the least-upper-bound property (1.6.1) carries a matching greatest-lower-bound property.

Remark 1.5.4

The ε-characterisation of the supremum

The supremum, when it exists, is *unique* (by trichotomy, 1.3.2) and *need not* belong to E.[8] The least-upper-bound condition has a working reformulation that is the tool behind nearly every supremum argument in the course:

$$\alpha = \sup E \iff \alpha \text{ is an upper bound of } E, \text{ and for every } \varepsilon > 0 \text{ there is } x \in E \text{ with } x > \alpha - \varepsilon.$$

The first clause says α is an upper bound; the second says nothing smaller is—any proposed reduction $\alpha - \varepsilon$ is already overshot by a member of E. Dually,

$$\alpha = \inf E \iff \alpha \text{ is a lower bound of } E, \text{ and for every } \varepsilon > 0 \text{ there is } x \in E \text{ with } x < \alpha + \varepsilon.$$

This rephrasing (cf. R 1.8) is what one actually reaches for: to use a supremum, feed it an ε and extract the witness x.

[8] For $E = \{x \in \mathbb{Q} : 0 < x^2 < 2\}$ the supremum is $\sqrt{2} \notin E$ (indeed $\notin \mathbb{Q}$); for $E = \{1 - \frac{1}{n} : n \geq 1\}$ it is $1 \notin E$. When $\sup E$ *does* lie in E it is the *maximum* of E.

Proof 1.5.4

Proof of the ε-Characterisation.

(1) ($\Rightarrow$) Let $\alpha = \sup E$. Then α is an upper bound. Given $\varepsilon > 0$, the smaller number $\alpha - \varepsilon$ is *not* an upper bound (else α would not be the *least* one), so some $x \in E$ exceeds it: $x > \alpha - \varepsilon$.

(2) ($\Leftarrow$) Suppose α is an upper bound with the stated ε-property, and let $\gamma < \alpha$. Put $\varepsilon = \alpha - \gamma > 0$; the property yields $x \in E$ with $x > \alpha - \varepsilon = \gamma$, so γ is not an upper bound. Hence no $\gamma < \alpha$ bounds E, and $\alpha = \sup E$.

■

[Direct] uses 1.5.2; the inf case is dual.

1.6 The Real Number System

Theorem 1.6.1

The real number system

There is an ordered field $\mathbb{R} \supseteq \mathbb{Q}$ in which every nonempty subset that is bounded above has a least upper bound. It is unique up to isomorphism.[9] This is Rudin's existence theorem for the real field (R 1.19).

Remark 1.6.2

Completeness and density

The property in 1.6.1 is *completeness*: $\mathbb{R}$ is complete iff it has the least-upper-bound property. A first consequence is that $\mathbb{Q}$ is *dense* in $\mathbb{R}$—between any two reals lies a rational (1.7.2; cf. R 1.19–R 1.20).

Remark 1.6.3

Three ways to build $\mathbb{R}$

- *Accept it*—take 1.6.1 on faith and learn to use sup.
- *Dedekind cuts* (1.6.4)—realise each real as a downward-closed set of rationals.
- *Completion*—the general procedure that completes any metric space (the abstract construction is deferred). Here it realises $\mathbb{R}$ via *Cauchy sequences of rationals*: a sequence in $\mathbb{Q}$ that "ought to converge" *is* declared a real number. The missing $\sqrt{2}$ reappears as, for instance, the recursion

$$q_0 = 1, \qquad q_{n+1} = \frac{2q_n + 2}{q_n + 2} \qquad (\text{terms } 1,\ 1.4,\ 1.41,\ \dots),$$

whose map $p \mapsto \dfrac{2p+2}{p+2}$ is exactly the gap-construction of 1.4.2: it fixes the value $\sqrt{2}$ and drives every starting rational toward it, so $q_n \to \sqrt{2}$. Many different rational sequences crowd toward the same target, so one declares any two that "converge together" *equivalent*, and a real number is such an equivalence class.

[9] An isomorphism of ordered fields is a bijection preserving $+$, $\cdot$, and $<$; "unique up to isomorphism" means any two complete ordered fields are indistinguishable as such. Two were built in 1872: Dedekind's cuts (*Stetigkeit und irrationale Zahlen*) and Cantor's equivalence classes of Cauchy sequences of rationals.

Rudin states the existence theorem abstractly in R 1.19; these are model-building viewpoints.

Definition 1.6.4

Dedekind cut

A *cut* is a set $\alpha \subseteq \mathbb{Q}$ such that

(1) $\alpha \neq \varnothing$ and $\alpha \neq \mathbb{Q}$;

(2) α is downward closed: $r \in \alpha$ and $s < r$ imply $s \in \alpha$;

(3) α has no largest element.

The cuts, suitably ordered, *are* the real numbers; this is one construction behind Rudin's real-field theorem R 1.19.

Example 1.6.5

$\sqrt{2}$ *as a cut*

The real number $\sqrt{2}$ is the cut

$$\alpha = \{q \in \mathbb{Q} : q < 0\} \ \cup \ \{q \in \mathbb{Q} : q^2 < 2\}.$$

This is the set A of 1.4.2 together with the non-positive rationals; the former "gap" is now a bona fide element (cf. R 1.1, R 1.19).

1.7 The Archimedean Property, Density, and Decimals

Theorem 1.7.1

Archimedean property

If $x, y \in \mathbb{R}$ and $x > 0$, then there is a positive integer n with $nx > y$.[10] This is the Archimedean part of R 1.20.

Proof 1.7.1

Proof by Contradiction of the Archimedean Property.

(1) Suppose not: the set $A = \{nx : n \in \mathbb{N}\}$ is bounded above. By the least-upper-bound property (1.6.1) it has a supremum $\alpha = \sup A$.

(2) Since $x > 0$, $\alpha - x < \alpha$, so $\alpha - x$ is not an upper bound of A: some $m \in \mathbb{N}$ has $mx > \alpha - x$.

(3) Then $(m+1)x > \alpha$, yet $(m+1)x \in A$—contradicting $\alpha = \sup A$. So A is not bounded above, and in particular some $nx > y$.

■

[A5+A9] cf. R 1.20; via the supremum; uses 1.6.1.

[10] Named by Otto Stolz in the 1880s after Archimedes, who postulated it in *On the Sphere and the Cylinder*; the idea is older still—Euclid records it (*Elements* V, Def. 4) and credits Eudoxus.

Theorem 1.7.2

Density of $\mathbb{Q}$ in $\mathbb{R}$

If $x, y \in \mathbb{R}$ and $x < y$, then there is a rational q with $x < q < y$. This is the density part of R 1.20.

Proof 1.7.2

Proof by the Archimedean Property of the Density of $\mathbb{Q}$.

(1) Since $y - x > 0$, the Archimedean property (1.7.1) gives $n \in \mathbb{N}$ with $n(y - x) > 1$, i.e. $ny - nx > 1$.

(2) Since $ny - nx > 1$, the interval (nx, ny) contains an integer m (§1.A.1).

(3) Dividing by $n > 0$, $x < \dfrac{m}{n} < y$, so $q = \dfrac{m}{n}$ works.

■

[Direct] cf. R 1.20; uses 1.7.1.

Construction 1.7.3

Real numbers as decimals

Every *nonnegative* real is the supremum of its finite decimal truncations (a negative real is then named by its sign):

$$1.4112\ldots \longmapsto \sup\{\,1,\ 1.4,\ 1.41,\ 1.411,\ \ldots\}.$$

This uses the least-upper-bound and Archimedean machinery in R 1.19–R 1.20.

Example 1.7.4

$0.999\ldots = 1$

This settles the opening question. Put $x = 1 - 0.999\ldots$; then $0 \le x < 10^{-k}$ for every $k \in \mathbb{N}$. By the Archimedean property (1.7.1)[11] no positive real lies below every 10^{-k}, so $x = 0$—that is, $0.999\ldots = 1$. More generally every terminating decimal equals its repeating-9 twin ($2.5 = 2.4999\ldots$, and so on), so

$$\mathbb{R} = \{\text{decimal expansions}\}/\!\sim,$$

the quotient that identifies each terminating decimal with that twin (cf. R 1.20).

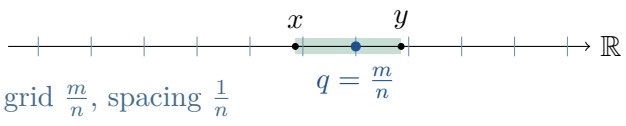


Figure 1.7.5. DENSITY OF $\mathbb{Q}$ IN $\mathbb{R}$. Choosing n with $1/n < y - x$ makes the grid of multiples m/n finer than the gap, so some m/n lands strictly between x and y (1.7.2).

[11] If $x > 0$, the Archimedean property gives n with $nx > 1$; taking k with $10^k \ge n$ then yields $10^k x > 1$, i.e. $x > 10^{-k}$.

Appendix

Supplementary material—additional proofs, context, and other treatments; not part of the lecture proper.

Supplement 1.A.1

An interval longer than 1 *contains an integer*

If $s, t \in \mathbb{R}$ with $t - s > 1$, then there is an integer m with $s < m < t$. (Used silently in the density proof, 1.7.2.)

Proof 1.A.1

Direct Proof from the Archimedean Property.

(1) The integers exceeding s are nonempty (by the Archimedean property) and bounded below by s, so they have a least element m; thus $m - 1 \leq s < m$.

(2) Hence $m \leq s + 1 < t$, and therefore $s < m < t$.

■

[Direct] uses 1.7.1; cf. R 1.20.

Supplement 1.A.2

Every nonnegative real has a decimal expansion

Construction 1.7.3 names each nonnegative real as the supremum of its truncations; here we build those truncations digit by digit and confirm the supremum is the number we began with. Given $x \geq 0$, there are an integer $a_0 \geq 0$ and digits $d_1, d_2, \cdots \in \{0, \ldots, 9\}$ whose truncations

$$t_k \;=\; a_0 + \sum_{i=1}^{k} \frac{d_i}{10^i} \qquad \text{satisfy} \qquad 0 \leq x - t_k < 10^{-k},$$

so that $x = \sup_k t_k$. The rule below always takes the *largest* admissible digit, so a number with a terminating expansion receives it ($1 = 1.000\ldots$), and the repeating-9 twin (1.7.4) is the single expansion this rule never produces. The key order input is R 1.20.

Proof 1.A.2

Direct Proof by Greedy Digit Selection.

(1) For the integer part, let a_0 be the greatest integer with $a_0 \leq x$ — it exists by the least-integer argument of 1.A.1, applied to the integers exceeding x, and $a_0 \geq 0$ since $x \geq 0$. Then $t_0 = a_0$ obeys $0 \leq x - t_0 < 1$.

(2) For the digits, suppose t_{k-1} has been built with $0 \leq x - t_{k-1} < 10^{-(k-1)}$. Let d_k be the largest digit in $\{0, \ldots, 9\}$ with $t_{k-1} + d_k\, 10^{-k} \leq x$; the digit 0 qualifies, and $d_k \leq 9$ because $x < t_{k-1} + 10 \cdot 10^{-k} = t_{k-1} + 10^{-(k-1)}$. Put $t_k = t_{k-1} + d_k\, 10^{-k}$. Then $t_k \leq x$, while maximality of d_k gives $x < t_k + 10^{-k}$; hence $0 \leq x - t_k < 10^{-k}$ and the invariant persists.

(3) For the supremum: every $t_k \leq x$, so x is an upper bound. If $\beta < x$, then $x - \beta > 10^{-k}$ for some k (Archimedean property, 1.7.1), so $t_k > x - 10^{-k} > \beta$; thus no

$\beta < x$ bounds the truncations, and $x = \sup_k t_k$.

■

[Direct] uses 1.7.1, 1.A.1.

LECTURE 2

Counting the Infinite, and the Geometry of Distance: Countability and Metric Spaces

MATH S4061 | Introduction to Modern Analysis I

Thursday, 28 May 2026

No one shall expel us from the paradise that Cantor has created.

David Hilbert

Contents of this lecture

Overview. Lecture 1 built $\mathbb{R}$ as the complete ordered field and drew from completeness the Archimedean property (1.7.1) and the density of $\mathbb{Q}$ (1.7.2); the same property gives infima as readily as suprema (1.5.3). Two new questions open here. First, *how large* is $\mathbb{R}$? Cantor's notion of countability lets us compare infinite sets, and his diagonal argument shows $\mathbb{R}$ is strictly larger than $\mathbb{Q}$—uncountable, with $\mathbb{Q}$ a vanishingly "thin" subset. Second, what is the right language for *distance*? Metric spaces distil the Euclidean norm into three axioms, and the open sets they generate are the gateway to topology, where the rest of the course lives. The companion text is Rudin, Chapter 2, especially R 2.1–R 2.19.

2.1 Counting the Infinite: Countable and Uncountable Sets

Definition 2.1.1

Injective, surjective, bijective

Let $f : X \to Y$ be a function (cf. R 2.1–R 2.3).

- f is *injective* if $f(x) = f(x') \Rightarrow x = x'$ (distinct inputs give distinct outputs);
- f is *surjective* if every $y \in Y$ equals $f(x)$ for some $x \in X$;
- f is *bijective* if it is both. A bijection has a two-sided inverse $f^{-1} : Y \to X$.

Definition 2.1.2

Equinumerous sets; finite, countable, uncountable

Sets X and Y are *equinumerous*, written $X \sim Y$, if there is a bijection $f : X \to Y$ (R 2.3). With $J_0 = \emptyset$ and $J_n = \{1, \ldots, n\}$, a set X is

- *finite* if $X \sim J_n$ for some n, and *infinite* otherwise;
- *countable* if $X \sim \mathbb{N}$;
- *uncountable* if it is infinite and $X \not\sim \mathbb{N}$.[12]

This packages Rudin's finite/countable terminology R 2.4–R 2.7.

Proposition 2.1.3

Equinumerosity is an equivalence relation

The relation $\sim$ is reflexive, symmetric, and transitive. This is the basic relation behind R 2.3.

Proof 2.1.3

Direct Proof that Equinumerosity is an Equivalence Relation.

(1) For reflexivity, the identity $\mathrm{id}_X : X \to X$, $x \mapsto x$, is a bijection, so $X \sim X$.

(2) For symmetry, if $X \sim Y$ via a bijection f, then $f^{-1} : Y \to X$ is a bijection, so $Y \sim X$.

(3) For transitivity, if $X \sim Y$ via f and $Y \sim Z$ via g, then $g \circ f : X \to Z$ is a bijection (its inverse is $f^{-1} \circ g^{-1}$), so $X \sim Z$.

■

[Direct] bijections invert and compose, with $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Example 2.1.4

$\mathbb{Z}$ is countable

[12] The definition can feel strange—how can $\mathbb{N}$ be the "same size" as a set that looks bigger or smaller? The shift is to measure size by *pairing* rather than counting: two sets match if their elements can be lined up in perfect one-to-one correspondence with none left over. For finite sets this is just counting; for infinite sets it is the only thing that still works, and it makes you give up the finite habit that a part must be smaller than the whole—$n \mapsto 2n$ already pairs all of $\mathbb{N}$ with only the even numbers.

The identity shows $\mathbb{N}$ is countable. For $\mathbb{Z}$, interleave the signs:

$$f : \mathbb{N} \to \mathbb{Z}, \qquad f(n) = \begin{cases} n/2 & n \text{ even}, \\ -(n+1)/2 & n \text{ odd}, \end{cases}$$

a bijection that lists $\mathbb{Z}$ as $0, -1, 1, -2, 2, -3, 3, \ldots$ (cf. R 2.12–R 2.13).

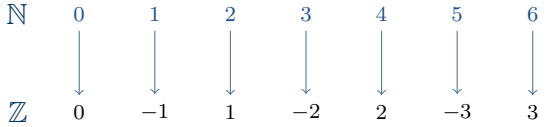


Figure 2.1.5. $\mathbb{N}$ counts $\mathbb{Z}$. The bijection of 2.1.4 walks outward from 0 in alternating steps, reaching every integer exactly once.

Definition 2.1.6

Sequences and the listing criterion

A *sequence* in a set X is a function $f : \mathbb{N} \to X$, written $\{x_n\}_{n\in\mathbb{N}}$ with $x_n = f(n)$. Thus X is countable precisely when some sequence lists every element of X exactly once. Rudin's sequence definition is R 2.7.

Example 2.1.7

$\mathbb{Q}$ is countable

Identifying a fraction with its pair of integers exhibits $\mathbb{Q}$ inside $\mathbb{Z} \times \mathbb{Z}$, which a diagonal sweep lists as a single sequence.

Proof 2.1.7

Proof by Diagonal Enumeration that $\mathbb{Q}$ is Countable.

(1) Arrange the pairs (p, q) with $q \neq 0$ on a grid and traverse them along successive finite diagonals (each diagonal $|p| + |q| = \text{const}$ holds only finitely many pairs).

(2) This runs through every pair; deleting pairs with a common factor and any repeats leaves a sequence that lists each rational exactly once.

(3) Hence $\mathbb{Q} \sim \mathbb{N}$ (2.1.6).

■

[Constr] cf. R 2.12–R 2.13.

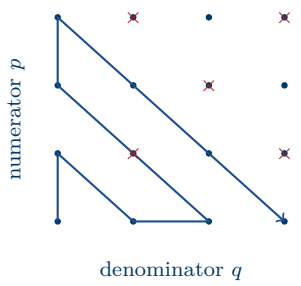


Figure 2.1.8. LISTING THE RATIONALS. Sweeping the lattice of pairs (p, q) along finite diagonals threads them into one sequence; pairs with a common factor (e.g. $\frac{2}{2}, \frac{2}{4}$) are skipped (2.1.7).

Proposition 2.1.9

An infinite subset of a countable set is countable

If X is countable and $A \subseteq X$ is infinite, then A is countable (R 2.8).

Proof 2.1.9

Proof by Recursive Selection of Indices that the Subset is Countable.

(1) Write $X = \{x_n\}_{n\in\mathbb{N}}$, every element listed once (2.1.6).

(2) Choose indices recursively: let n_0 be the least index with $x_{n_0} \in A$; having chosen $n_0 < \cdots < n_k$, let n_{k+1} be the least index $> n_k$ with $x_{n_{k+1}} \in A$. Such an index exists because A, being infinite, cannot lie inside the finite set $\{x_0, \ldots, x_{n_k}\}$, so some $x_m \in A$ has $m > n_k$.

(3) The map $k \mapsto x_{n_k}$ is strictly increasing, hence injective, and reaches every element of A (each $a = x_m \in A$ is selected no later than stage m); so it is a bijection $\mathbb{N} \to A$, and $A \sim \mathbb{N}$.

■

[A10] cf. R 2.8; the recursion is an induction, justified at each step by A being infinite.

Theorem 2.1.10

Cantor: the binary sequences are uncountable

The set $2^{\mathbb{N}} = \{\, f : \mathbb{N} \to \{0, 1\} \,\}$ of infinite binary sequences—equivalently, the subsets of $\mathbb{N}$—is uncountable.[13]

Proof 2.1.10

Proof by Cantor's Diagonal Argument that the Binary Sequences are Uncountable.

(1) Suppose $2^{\mathbb{N}}$ were countable, listed as $a_1, a_2, a_3, \ldots$, each $a_k = (a_k^1, a_k^2, a_k^3, \ldots)$ a binary sequence.

(2) Define $\bar{b} \in 2^{\mathbb{N}}$ by flipping the diagonal: $\bar{b}_n = 1 - a_n^n$.

(3) For each n, $\bar{b}$ differs from a_n in its nth entry ($\bar{b}_n \neq a_n^n$), so $\bar{b} \neq a_n$ for every n.

(4) Thus $\bar{b}$ appears nowhere in the list—contradicting that the list was all of $2^{\mathbb{N}}$. Hence $2^{\mathbb{N}}$ is uncountable.

■

[A8+A9] diagonalisation, by contradiction; cf. R 2.14.

[13] The diagonal proof below can look like a sleight of hand. Against *any* proposed list of binary sequences it produces one the list left out, so no list can hold them all—and you read that missing sequence straight off the list: make it disagree with the kth listed sequence in its kth entry, and it differs from the first (in entry 1), the second (entry 2), and every one after. The move is worth keeping on its own: to escape an enumeration, differ from item k in coordinate k.

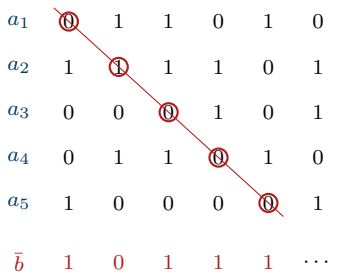


Figure 2.1.11. Cantor's diagonal. Flipping the boxed diagonal entries yields a sequence $\bar{b}$ differing from each listed a_k in place k, so $\bar{b}$ is absent from the list (2.1.10).

Theorem 2.1.12

The interval $[0, 1]$ is uncountable

The interval $[0, 1] \subseteq \mathbb{R}$ is uncountable.

Proof 2.1.12

Proof by Reduction to Binary Sequences that $[0, 1]$ *is Uncountable.*

(1) Since $[0, 1] \supseteq [0, 1)$, it suffices to show $[0, 1)$ is uncountable. Send each $x \in [0, 1)$ to the binary expansion $(b_n) \in 2^{\mathbb{N}}$ that does *not* end in all 1s; on $[0, 1)$ this representative exists and is unique, so the map is injective, with image exactly the sequences not ending in all 1s.

(2) If $[0, 1)$ were countable, that image would be countable, so $2^{\mathbb{N}}$—the image together with the countably many all-1-tail sequences—would be countable too (R 2.12), contradicting 2.1.10.

(3) Hence $[0, 1)$, and therefore $[0, 1]$, is uncountable.

■

[Direct] uses 2.1.10; barring an all-1 tail makes binary expansions unique ($0.0111\ldots_2 = 0.1000\ldots_2$), and the barred eventually-all-1 sequences are countable.

Corollary 2.1.13

$\mathbb{R}$ is uncountable

Were $\mathbb{R}$ countable, its infinite subset $[0, 1]$ would be countable too (2.1.9); but it is not (2.1.12). So $\mathbb{R}$ is uncountable.[14] Compare Rudin's diagonal proof for binary sequences R 2.14 and later perfect-set result R 2.43.

Remark 2.1.14

$\mathbb{Q}$ *is thin in* $\mathbb{R}$

[14] "Uncountable" says something stronger than "a bigger infinity": *no list can ever catch all the reals*—hand over any enumeration and the diagonal argument hands back a real it missed. To gauge the size of the gap: the numbers you can pin down as a root of a polynomial with integer coefficients (the *algebraic* numbers) form only a countable set, so *almost every* real lies beyond them—*transcendental*—even though such numbers are hard to point to.

Thus $\mathbb{Q}$ is countable yet dense (1.7.2), while $\mathbb{R}$ is uncountable: the rationals are a vanishingly small part of the line. Made precise by measure, $\mathbb{Q}$—indeed any countable set—has Lebesgue length zero, so a real drawn from any distribution with a density is irrational with probability 1. Compare the countability results in R 2.12–R 2.13 and the diagonal uncountability result R 2.14; the measure language is later perspective.

Theorem 2.1.15

A countable union of countable sets is countable

If $A_1, A_2, A_3, \dots$ are countable sets, then $\bigcup_{n=1}^{\infty} A_n$ is countable (R 2.12).[15]

Proof 2.1.15

Proof by Diagonal Enumeration of the Array that the Union is Countable.

(1) List each set as a sequence (2.1.6), say $A_n = \{x_{n,1}, x_{n,2}, x_{n,3}, \dots\}$, and arrange all the elements in the doubly-indexed array

$$\begin{array}{cccc} x_{1,1} & x_{1,2} & x_{1,3} & \cdots \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \end{array}$$

whose nth row exhausts A_n.

(2) Enumerate the array along its successive anti-diagonals—the entries $x_{n,m}$ with $n + m = k$, taken for $k = 2, 3, 4, \dots$:

$$x_{1,1}; \quad x_{1,2}, x_{2,1}; \quad x_{1,3}, x_{2,2}, x_{3,1}; \quad \dots$$

Each diagonal is finite, so this is a single sequence, and it visits every entry of the array, hence every element of $\bigcup_n A_n$.

(3) This sequence may repeat an element (the sets need not be disjoint). If the union is finite there is nothing to prove; otherwise discard repetitions, leaving an infinite subsequence that lists $\bigcup_n A_n$ without repetition (2.1.9). Hence $\bigcup_{n=1}^{\infty} A_n$ is countable.

■

[A8] the array threaded along finite diagonals, repetitions pruned by 2.1.9; cf. R 2.12.

[15] The trick is to defeat *two* indices at once. The elements come labelled by a row n (which set) and a column m (which element of that set), so they fill a square array—and a square array is "two-dimensional," seemingly too big to list. The diagonal sweep is the resolution: each anti-diagonal $n + m = k$ is *finite*, so walking the diagonals one after another threads the whole grid onto a single line. The same diagonal that listed the pairs for $\mathbb{Q}$ (2.1.7) is doing the work here.

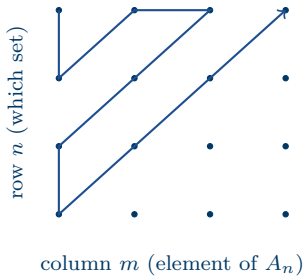


Figure 2.1.16. Threading the array. Sweeping the array $(x_{n,m})$ along finite anti-diagonals $n+m=k$ threads the whole double-indexed family onto one sequence (2.1.15).

Corollary 2.1.17

A finite product of countable sets is countable

If $A_1, \dots, A_k$ are countable sets, then the product $A_1 \times \cdots \times A_k$ is countable (R 2.13). In particular $\mathbb{Z} \times \mathbb{Z}$ is countable, giving a second route to $\mathbb{Q}$ countable (2.1.7).

Proof 2.1.17

Proof by Induction on the Number of Factors that the Product is Countable.

(1) The case $k = 1$ is immediate. We induct on k, taking $k = 2$ as the base recorded by the array.

(2) *Base* $k = 2$. Write $A_1 \times A_2 = \bigcup_{a \in A_1} (\{a\} \times A_2)$. Each slice $\{a\} \times A_2$ is in bijection with the countable set A_2, hence countable, and the union runs over the countable index set A_1. So $A_1 \times A_2$ is a countable union of countable sets, countable by 2.1.15.

(3) *Step.* Suppose $A_1 \times \cdots \times A_{k-1}$ is countable. Decompose

$$A_1 \times \cdots \times A_k = \bigcup_{a \in A_k} ((A_1 \times \cdots \times A_{k-1}) \times \{a\}),$$

again a countable union (over A_k) of countable sets, so it is countable by 2.1.15.

(4) By induction $A_1 \times \cdots \times A_k$ is countable for every k. Taking $A_1 = A_2 = \mathbb{Z}$ shows $\mathbb{Z} \times \mathbb{Z}$ is countable; since $\mathbb{Q}$ embeds in it as an infinite subset (a fraction in lowest terms is a pair (p, q), $q > 0$), $\mathbb{Q}$ is countable by 2.1.9.

■

[A10] base case $k = 2$ via 2.1.15; cf. R 2.13.

2.2 Metric Spaces

The Euclidean distance on $\mathbb{R}^n$, $d(x, y) = \|x - y\| = \left(\sum_{i=1}^n |x_i - y_i|^2\right)^{1/2}$, built from the inner product $\langle x, y \rangle = \sum_i x_i y_i$, is the model. A metric space keeps only the features of d that analysis actually uses.

Definition 2.2.1

Metric space

A *metric space* (R 2.15)[16] is a set X with a *metric* (or distance) $d : X \times X \to \mathbb{R}^{\geq 0}$ satisfying, for all $x, y, z \in X$:

1. $d(x, y) \geq 0$, with $d(x, y) = 0 \iff x = y$;
2. $d(x, y) = d(y, x)$ (symmetry);
3. $d(x, z) \leq d(x, y) + d(y, z)$ (the triangle inequality).

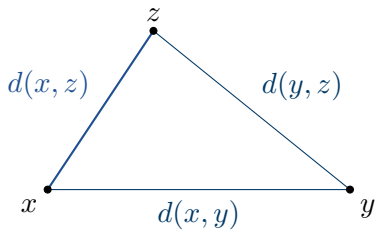


Figure 2.2.2. THE TRIANGLE INEQUALITY. The direct distance $d(x, z)$ never exceeds the detour $d(x, y) + d(y, z)$ through y—axiom (3) of 2.2.1.

Example 2.2.3

The standard metrics

$(\mathbb{R}, |\cdot|)$ with $d(x, y) = |x - y|$ is a metric space, as are $(\mathbb{Z}, |\cdot|)$, $(\mathbb{N}, |\cdot|)$, and $(\mathbb{Q}, |\cdot|)$. More generally any subset Y of a metric space (X, d) is itself a metric space under the restricted distance $d|_{Y \times Y}$. The Euclidean space $(\mathbb{R}^n, d_2)$, $d_2(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^2\right)^{1/2}$, is the prototype. These are Rudin's basic metric examples (cf. R 2.16).

Example 2.2.4

The p-metrics and the discrete metric

On $\mathbb{R}^n$, for $p \geq 1$,[17]

$$d_p(x, y) = \Big(\sum_{i=1}^n |x_i - y_i|^p\Big)^{1/p}, \qquad d_\infty(x, y) = \max_i |x_i - y_i|.$$

On any set X, the *discrete metric*—$d(x, y) = 0$ if $x = y$ and 1 otherwise—is a metric. The triangle inequality for d_p is Minkowski's inequality, taken here as a standard fact (cf. R 2.16).

Example 2.2.5

Metrics on a function space

[16] Read the three axioms not as arbitrary rules but as the *least* a notion of "distance" must obey for analysis to run on it. (1) distances are honest numbers that vanish only when two points genuinely coincide, so d can tell points apart; (2) distance doesn't care which direction you travel; (3) the triangle inequality says a detour is never shorter than going straight—the one axiom with real force, and the workhorse of almost every proof to come (already 2.3.5). That is the whole point of abstracting: keep only these, and every notion built from them works on *any* such space at once—Euclidean space, the function spaces of 2.2.5, and more.

[17] The board writes $p > 1$; in fact d_p is a metric for every $p \geq 1$, with d_1 and d_∞ the boundary cases.

Let $C = C[0,1]$ be the continuous functions $f : [0,1] \to \mathbb{R}$.[18] Both

$$d_\infty(f,g) = \sup_{[0,1]} |f-g| \qquad \text{and} \qquad d_1(f,g) = \int_0^1 |f(x)-g(x)|\,dx$$

are metrics on C. Here d_∞ is finite because a continuous function on $[0,1]$ is bounded, and $d_1(f,g) = 0$ forces $f = g$ since $|f-g|$ is continuous. As metric examples, compare R 2.16.

2.3 Open Sets and Open Balls

Basic *topology* studies a space through its *open subsets*[19]—concepts that can be phrased using open sets alone. Everything starts with the open ball.

Definition 2.3.1

Open ball

In a metric space (X,d), the *open ball* of radius $r > 0$ about $p \in X$ is

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

Rudin calls this a neighbourhood of p in R 2.18.

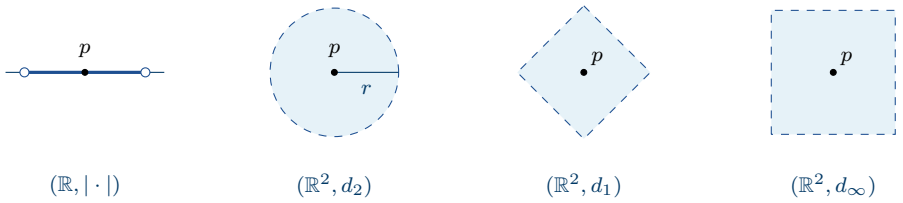


Figure 2.3.2. An open ball takes the shape of its metric. Same centre, same radius: an interval in $(\mathbb{R}, |\cdot|)$, a round disk under d_2, a diamond under d_1, a square under d_∞—each the set $\{x : d(x,p) < r\}$ of 2.3.1.

Definition 2.3.3

Interior point; open set

Let (X,d) be a metric space and $E \subseteq X$. A point $p \in E$ is an *interior point* of E if $B_r(p) \subseteq E$ for some $r > 0$, and E is *open in* X if every point of E is interior (R 2.18). Conversely, p fails to be interior exactly when $B_r(p) \not\subseteq E$ for *every* $r > 0$; so E is *not* open as soon as one of its points is not interior.

[18] The leap to make here is to stop picturing a function as a *graph* and start treating it as a single *point* in a space whose points happen to be functions. Then "two functions are close" just means their distance is small—and which metric you pick decides what "close" means: under d_∞ two functions are close only if they stay near each other *everywhere at once* (uniformly), while under d_1 they may disagree sharply on a thin set so long as the *average* gap is small. Same functions, two genuinely different geometries.

[19] The picture to carry: a set is *open* when none of its points sits on the "edge"—every point has a whole little ball of room around it that is still inside the set, so from any point you can step a little in every direction without leaving. That "room to move" is what makes open sets the right language for analysis: continuity, convergence, and compactness can all be expressed through open sets alone, never mentioning the particular distance that produced them (indeed the d_1, d_2, d_∞ of 2.2.4 produce the very same open sets on $\mathbb{R}^n$).

Remark 2.3.4

Openness is relative

Openness is meaningful only *relative to the ambient space*: one says "E is open in X," never merely "E is open." The same set can be open in one metric space and not in another. This anticipates Rudin's relative-openness theorem R 2.30.

Proposition 2.3.5

Every open ball is open

In any metric space (X, d), the ball $B_r(p)$ is open in X, for every $p \in X$ and $r > 0$.

Proof 2.3.5

Proof by the Triangle Inequality that Every Open Ball is Open.

(1) Fix $q \in B_r(p)$; then $d(p,q) < r$, so $r_q := r - d(p,q) > 0$.

(2) Take any $x \in B_{r_q}(q)$. By the triangle inequality,

$$d(x,p) \le d(x,q) + d(q,p) < r_q + d(q,p) = r.$$

(3) Hence $x \in B_r(p)$, so $B_{r_q}(q) \subseteq B_r(p)$. Every $q \in B_r(p)$ is therefore interior, and $B_r(p)$ is open.

■

[Direct] the template for openness proofs; uses 2.2.1; cf. R 2.19.

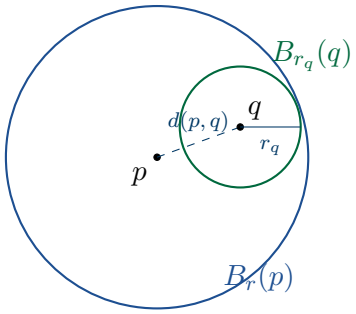


Figure 2.3.6. WHY A BALL IS OPEN. For $q \in B_r(p)$, the radius $r_q = r - d(p,q)$ keeps $B_{r_q}(q)$ inside $B_r(p)$—the triangle inequality at work (proof of 2.3.5).

Remark 2.3.7

A recipe for proving openness

To show E is open in X: (1) take an arbitrary $p \in E$; (2) produce a radius $r_p > 0$ (guided by the geometry); (3) verify $B_{r_p}(p) \subseteq E$. The proof of 2.3.5 is the model, and this is the working form of R 2.18–R 2.19.

LECTURE 3

The Topology of Metric Spaces: Open and Closed Sets, and Compactness

MATH S4061 | Introduction to Modern Analysis I

Tuesday, 2 June 2026

> *Topology is precisely the mathematical discipline that allows the passage from local to global.*
>
> René Thom

Contents of this lecture

Overview. Lecture 2 introduced metric spaces and the open ball (2.3.1). This lecture builds the *topology* they generate: which sets are open—closed under arbitrary unions and *finite* intersections—then the limit (cluster) points that define *closed* sets, the clean duality E open $\iff$ E^c closed, how openness is *relative* to a subspace, and finally *compactness*: the open-cover/finite-subcover property that, unlike open or closed, is *intrinsic* to a set. One method recurs throughout: to show a set is open, write it as a union of balls centred at its own points. The companion text is Rudin, Chapter 2, especially R 2.18–R 2.33.

3.1 The Algebra of Open Sets

Definition 3.1.1

Indexed unions and intersections

Let A be an index set and $V_\alpha \subseteq X$ for each $\alpha \in A$. The *union* and *intersection* are

$$\bigcup_{\alpha\in A} V_\alpha = \{x : x \in V_\alpha \text{ for some } \alpha\}, \qquad \bigcap_{\alpha\in A} V_\alpha = \{x : x \in V_\alpha \text{ for every } \alpha\};$$

in particular $p \in \bigcap_\alpha V_\alpha \iff p \in V_\alpha$ for all α. This is the indexed-family notation of R 2.9.

Proposition 3.1.2

Unions and finite intersections of open sets

Let (X, d) be a metric space with $V_\alpha \subseteq X$ open for every $\alpha \in A$.

1. $\bigcup_{\alpha\in A} V_\alpha$ is open (any union of open sets is open);
2. $V_{\alpha_1} \cap \cdots \cap V_{\alpha_n}$ is open (any *finite* intersection is open).[20]

Proof 3.1.2

Direct Proof that Unions and Finite Intersections of Open Sets Are Open.

(1) For the union, let $p \in \bigcup_\alpha V_\alpha$, say $p \in V_\beta$. As V_β is open, some $B_{r_p}(p) \subseteq V_\beta \subseteq \bigcup_\alpha V_\alpha$; so p is interior and the union is open.

(2) For a finite intersection, let $p \in V_{\alpha_1} \cap \cdots \cap V_{\alpha_n}$. Each V_{α_i} open gives $r_i > 0$ with $B_{r_i}(p) \subseteq V_{\alpha_i}$. Put $r = \min(r_1, \ldots, r_n) > 0$; then $B_r(p) \subseteq V_{\alpha_i}$ for every i, so $B_r(p) \subseteq \bigcap_i V_{\alpha_i}$. The minimum of *finitely* many positive numbers is positive—this is where finiteness is used.

■

[Direct] uses 2.3.3; finiteness enters in (2); cf. R 2.24.

Example 3.1.3

Infinite intersections can fail

The bound $r = \min r_i$ becomes an infimum that may vanish. In $(\mathbb{R}, |\cdot|)$ the open intervals $I_n = (-\frac{1}{n}, \frac{1}{n})$ satisfy

$$\bigcap_{n\geq 1} I_n = \{0\},$$

which is not open in $\mathbb{R}$. So "finite" in 3.1.2(2) cannot be dropped (cf. R 2.24).

Proposition 3.1.4

X and $\emptyset$ are open

In any metric space (X, d), both X and $\emptyset$ are open in X (cf. R 2.24).

[20] The asymmetry is the whole point. Openness gives every point a ball of room; for a *union* a point needs that room from only the one piece it already lies in, so any number of open sets stays open. An *intersection* demands room from *all* the pieces at once—and finitely many positive radii have a positive minimum, while infinitely many can shrink to zero and leave no room at all (3.1.3). Finiteness is exactly what keeps that minimum positive.

Proof 3.1.4

Direct Proof that X and $\emptyset$ Are Open, the Second Case Vacuously.

(1) For any $p \in X$ and any $r > 0$, $B_r(p) \subseteq X$, so every point of X is interior: X is open.

(2) $\emptyset$ has no points, so "every point is interior" holds vacuously: $\emptyset$ is open.

■

[Direct]

Proposition 3.1.5

Open sets are exactly unions of open balls

A set $V \subseteq X$ is open if and only if it is a union of open balls. This is the union-of-neighbourhoods viewpoint behind R 2.19 and R 2.24.

Proof 3.1.5

Constructive Proof that Open Sets Are Exactly Unions of Open Balls.

(1) If V is open, each $p \in V$ has an $r_p > 0$ with $B_{r_p}(p) \subseteq V$. Every ball lies in V, and each p lies in its own ball, so

$$V = \bigcup_{p \in V} B_{r_p}(p),$$

a union of open balls.

(2) Conversely any union of open balls is open: each ball is open (2.3.5), and a union of open sets is open (3.1.2).

■

[Constr] balls are open by 2.3.5; the converse uses 3.1.2.

3.2 Limit Points and Closed Sets

Definition 3.2.1

Cluster (limit) point

Let (X, d) be a metric space and $E \subseteq X$. A point $p \in X$ is a *cluster point*[21] of E if every open ball $B_r(p)$ contains a point of E other than p: for all $r > 0$ there is $q \in E$ with $q \in B_r(p)$ and $q \neq p$. Consequently a finite set has no cluster points.

Definition 3.2.2

Isolated point

[21] Also called a *limit point* (Rudin's term) or *accumulation point*. Because every ball holds a point of $E \setminus \{p\}$, it in fact holds infinitely many points of E: a ball with only finitely many would have a least positive distance from p among them, so a smaller ball meets $E \setminus \{p\}$ nowhere (cf. R 2.20).

A point $p \in E$ is an *isolated point* of E if some ball meets E only at p: there is $r > 0$ with $B_r(p) \cap E = \{p\}$ (R 2.18).

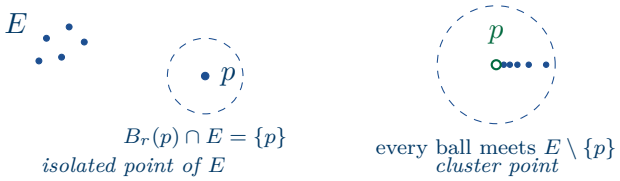


Figure 3.2.3. ISOLATED VERSUS CLUSTER POINTS. Left: p is *isolated* — a ball meets E only at p. Right: p is a *cluster point* — every ball, however small, meets E away from p (3.2.1, 3.2.2).

Definition 3.2.4

Closed set

E is *closed in* X if it contains all of its cluster points: whenever $p \in X$ is a cluster point of E, $p \in E$.[22] To prove E is *not* closed, exhibit one cluster point of E lying outside E (R 2.18).

Example 3.2.5

Cluster points and closedness

In $(\mathbb{R}, |\cdot|)$:

- $E = \{\frac{1}{n} : n \geq 1\}$ has 0 as its only cluster point[23] (every $B_r(0)$ contains some $\frac{1}{n}$, by the Archimedean property, 1.7.1); since $0 \notin E$, E is *not* closed, whereas $\{0\} \cup \{\frac{1}{n}\}$ *is*.
- $\mathbb{Z}$ is closed because it has *no* cluster points: each integer is isolated, and any $p \notin \mathbb{Z}$ has $\operatorname{dist}(p, \mathbb{Z}) > 0$, so a small ball about p misses $\mathbb{Z}$. Vacuously, $\mathbb{Z}$ holds all its cluster points.
- Every real is a cluster point of $\mathbb{Q}$ (by density, 1.7.2); $\mathbb{Q}$ is far from closed.

Compare Rudin's limit-point and closed-set definitions in R 2.18 and the nearby limit-point facts in R 2.20.

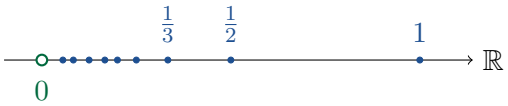


Figure 3.2.6. THE CLUSTER POINT OF $\{1/n\}$. The points $1/n$ bunch arbitrarily close to 0, so every ball about 0 meets $E = \{1/n\}$; yet $0 \notin E$, so E is not closed (3.2.5).

Remark 3.2.7

Open, closed, neither, both

[22] "Closed" is not "not open"; it means closed *under taking limits*. Read it as a sealing condition: anything E gets arbitrarily close to—a cluster point—already lies in E, so no limiting process can leak out of the set. That is what it means for a closed set to have no missing edges, and it is why closed sets are exactly those containing the limits of their own convergent sequences.

[23] Each $\frac{1}{n}$ is isolated, and any $p \notin \{0\} \cup E$ has a neighbourhood missing E; so 0 is the sole cluster point.

"Closed" is relative: say "E is closed *in* X." And open and closed are not opposites—in $(\mathbb{R}, |\cdot|)$ a set may be open, closed, neither, or both (here $p < q$). The interval (p, q) is open; $[p, q]$ is closed; $(p, q]$ is neither; $\mathbb{Q}$ is neither; and $\emptyset$ and X are *both* ("clopen"). Compare the complement duality and set algebra in R 2.23–R 2.24.

Theorem 3.2.8

E open $\iff$ E^c closed

For $E \subseteq X$ with complement $E^c = \{p \in X : p \notin E\}$: E is open in X if and only if E^c is closed in X.[24]

Proof 3.2.8

Proof via Cluster Points that E Is Open $\iff$ E^c Is Closed.

(1) ($\Rightarrow$) Let E be open and $p \in E$. Some $B_r(p) \subseteq E$, so $B_r(p) \cap E^c = \emptyset$ and p is *not* a cluster point of E^c. Thus no point of E is a cluster point of E^c: every cluster point of E^c lies in E^c, i.e. E^c is closed.

(2) ($\Leftarrow$) Let E^c be closed and $p \in E$. Then $p \notin E^c$; since E^c holds all its cluster points, p is not one, so some $B_r(p) \cap E^c = \emptyset$, i.e. $B_r(p) \subseteq E$. Hence p is interior and E is open.

■

[Direct] cf. R 2.23; recall $(E^c)^c = E$.

Lemma 3.2.9

De Morgan

For any family $\{E_\alpha\}_{\alpha \in A}$ of subsets of X,

$$\Big(\bigcup_\alpha E_\alpha\Big)^c = \bigcap_\alpha E_\alpha^c, \qquad \Big(\bigcap_\alpha E_\alpha\Big)^c = \bigcup_\alpha E_\alpha^c.$$

Proof 3.2.9

Proof of De Morgan's Laws by Element-Chasing.

(1) For the first identity,

$$p \in \Big(\bigcup_\alpha E_\alpha\Big)^c \iff p \notin \bigcup_\alpha E_\alpha \iff p \notin E_\alpha \ \forall\alpha \iff p \in E_\alpha^c \ \forall\alpha \iff p \in \bigcap_\alpha E_\alpha^c.$$

(2) The second follows by replacing each E_α with E_α^c and taking complements.

■

[Direct] cf. R 2.22.

Proposition 3.2.10

Intersections and finite unions of closed sets

[24] Treat this less as a fact than as a tool: it lets any statement about open sets be reread as one about closed sets, and back, just by complementing. That is how 3.2.10 falls out of 3.1.2 for free—complement, apply the open version, complement again—so the closed-set theorems never need a fresh proof.

Let $F_\alpha \subseteq X$ be closed in X for every $\alpha \in A$.

1. $\bigcap_{\alpha\in A} F_\alpha$ is closed (any intersection of closed sets is closed);
2. $F_{\alpha_1} \cup \cdots \cup F_{\alpha_n}$ is closed (any *finite* union is closed).

Proof 3.2.10

Proof by Dualising the Open Case that Intersections and Finite Unions of Closed Sets Are Closed.

(1) For the intersection, de Morgan gives $\left(\bigcap_\alpha F_\alpha\right)^c = \bigcup_\alpha F_\alpha^c$. Each F_α^c is open (3.2.8), and a union of open sets is open (3.1.2); so the complement is open and $\bigcap_\alpha F_\alpha$ is closed (3.2.8).

(2) For a finite union, $\left(F_{\alpha_1} \cup \cdots \cup F_{\alpha_n}\right)^c = F_{\alpha_1}^c \cap \cdots \cap F_{\alpha_n}^c$ is a *finite* intersection of open sets, hence open (3.1.2); so the union is closed. (Infinite unions can fail, dually to 3.1.3.)

■

[Direct] via 3.2.8, 3.2.9, 3.1.2; cf. R 2.24. A proof straight from the cluster-point definition (the board's exercise) also works.

3.3 The Subspace Topology

Proposition 3.3.1

Relative openness

Let (X, d) be a metric space, $Y \subseteq X$ a subspace, and $E \subseteq Y$. Then E is open in Y if and only if $E = Y \cap G$ for some G open in X. (Here the ball in Y is $B_r^Y(p) = \{x \in Y : d(x,p) < r\} = B_r(p) \cap Y$.)

Proof 3.3.1

Proof of Relative Openness by Carving Y-balls from X-balls.

(1) ($\Leftarrow$) Suppose $E = Y \cap G$, G open in X. For $p \in E$ we have $p \in G$, so some $B_{r_p}(p) \subseteq G$; intersecting with Y, $B_{r_p}^Y(p) = B_{r_p}(p) \cap Y \subseteq G \cap Y = E$. Thus p is interior in Y, and E is open in Y.

(2) ($\Rightarrow$) Suppose E is open in Y: each $p \in E$ has $r_p > 0$ with $B_{r_p}^Y(p) \subseteq E$. Put $G = \bigcup_{p\in E} B_{r_p}(p)$, open in X as a union of X-balls (3.1.5). Then

$$G \cap Y = \bigcup_{p\in E} \left(B_{r_p}(p) \cap Y\right) = \bigcup_{p\in E} B_{r_p}^Y(p) = E.$$

■

[Direct] cf. R 2.30.

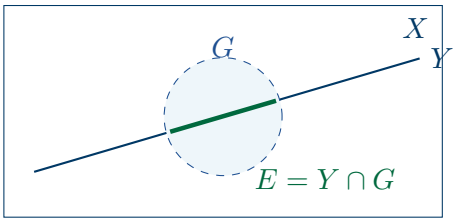


Figure 3.3.2. Open in a subspace. E is open in Y exactly when $E = Y \cap G$ for some G open in the whole space X: an X-ball G meets the subspace Y in the relative set E (3.3.1).

Remark 3.3.3

Openness by balls at its own points

The recurring method, used in 3.1.5 and 3.3.1: to prove a set is open, write it as a union of open balls centred at its own points (cf. R 2.19, R 2.30).

3.4 Compactness

Definition 3.4.1

Open cover; compact set

Let (X, d) be a metric space and $K \subseteq X$. An *open cover* of K is a family $\{V_\alpha\}$ of sets open in X with $K \subseteq \bigcup_\alpha V_\alpha$. The set K is *compact* if every open cover of K admits a *finite subcover*: indices $\alpha_1, \dots, \alpha_n$ with $K \subseteq V_{\alpha_1} \cup \cdots \cup V_{\alpha_n}$.[25] This combines Rudin's open-cover definition R 2.31 and compactness definition R 2.32.

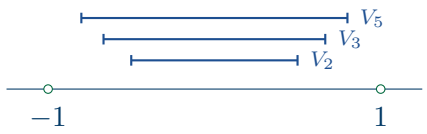
Example 3.4.2

Two non-compact sets

In $(\mathbb{R}, |\cdot|)$:

- $\mathbb{R}$ is not compact: the cover $V_n = (-n, n)$ $(n \geq 1)$ has $\bigcup_n V_n = \mathbb{R}$, but any finite subfamily lies in V_N with $N = \max n_i$, missing all $|x| \geq N$.
- $(-1, 1)$ is not compact: $V_n = (-1 + \frac{1}{n}, 1 - \frac{1}{n})$ $(n \geq 2)$ cover $(-1, 1)$, yet any finite subfamily is contained in the largest V_N, missing points near ± 1.

These are direct open-cover failures for Rudin's compactness definition R 2.32; compare the Heine–Borel theorem R 2.41.



[25] The naive candidate "closed and bounded" depends on the ambient space and the metric; the open-cover definition is *intrinsic* (3.4.4) and survives passage to continuous images, which is why it is the right notion.

Figure 3.4.3. AN OPEN COVER WITH NO FINITE SUBCOVER. The intervals $V_n = (-1 + \frac{1}{n}, 1 - \frac{1}{n})$ cover $(-1, 1)$, but each stops short of ± 1; a finite subfamily reaches only the largest V_N — so $(-1, 1)$ is not compact (3.4.2).

Proposition 3.4.4

Compactness is intrinsic

Let $K \subseteq Y \subseteq X$, where (X, d) is a metric space and Y carries the subspace metric $d|_{Y \times Y}$. Then K is compact in Y if and only if K is compact in X. So one may say simply "K is compact," with no ambient space.

Proof 3.4.4

Proof that Compactness Is Intrinsic by Trading Y-covers for X-covers.

(1) ($\Leftarrow$) Assume K compact in X. Let $\{V_\alpha\}$ be a Y-open cover of K. By 3.3.1 each $V_\alpha = G_\alpha \cap Y$ with G_α open in X, and $\{G_\alpha\}$ covers K. By X-compactness, $K \subseteq G_{\alpha_1} \cup \cdots \cup G_{\alpha_n}$; intersecting with Y (and using $K \subseteq Y$), $K \subseteq V_{\alpha_1} \cup \cdots \cup V_{\alpha_n}$.

(2) ($\Rightarrow$) Assume K compact in Y. Let $\{G_\alpha\}$ be an X-open cover of K. Set $V_\alpha = G_\alpha \cap Y$, open in Y (3.3.1) and covering K. By Y-compactness, $K \subseteq V_{\alpha_1} \cup \cdots \cup V_{\alpha_n} \subseteq G_{\alpha_1} \cup \cdots \cup G_{\alpha_n}$.

■

[Direct] uses 3.3.1; cf. R 2.33.

LECTURE 4

Compactness: Heine–Borel and Bolzano–Weierstrass

MATH S4061 | Introduction to Modern Analysis I

Thursday, 4 June 2026

Mathematics is the science of the infinite.

Hermann Weyl

Contents of this lecture

Overview. Lecture 3 defined compactness—every open cover has a finite subcover (3.4.1)—and showed it is intrinsic (3.4.4). Here we draw out its consequences: a compact set is closed and bounded, a closed subset of a compact set is compact, and nested nonempty compact sets meet. Specialising to $\mathbb{R}^k$ through nested intervals and k-cells yields the two pillars of the subject—*Heine–Borel* (in $\mathbb{R}^k$, compact = closed and bounded) and *Bolzano–Weierstrass* (every bounded infinite set has a cluster point). A short topological coda—interior, closure, connectedness—closes the chapter.

4.1 Compactness, Closedness, and Boundedness

Theorem 4.1.1

A compact set is closed and bounded

A compact subset K of a metric space (X, d) is closed and bounded (4.3.4).[26]

Proof 4.1.1

Direct Proof that the Complement Is Open.

(1) If $K = \emptyset$ it is closed and bounded vacuously, so assume $K \neq \emptyset$. Fix $p \notin K$. For each $q \in K$ set $r_q = \frac{1}{2}d(p, q) > 0$; the triangle inequality makes $B_{r_q}(p) \cap B_{r_q}(q) = \emptyset$.

(2) The balls $\{B_{r_q}(q)\}_{q \in K}$ cover K, so by compactness finitely many do: $K \subseteq B_{r_{q_1}}(q_1) \cup \cdots \cup B_{r_{q_n}}(q_n)$.

(3) Put $r = \min(r_{q_1}, \ldots, r_{q_n}) > 0$. Then $B_r(p)$ is disjoint from every $B_{r_{q_i}}(q_i)$, hence from K. So p is interior to K^c; as $p \notin K$ was arbitrary, K^c is open and K is closed. ∎

[Direct] ball separation; cf. R 2.34.

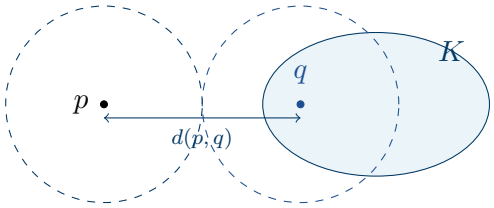


Figure 4.1.2. SEPARATING A POINT FROM A COMPACT SET. For $p \notin K$ and $q \in K$, the balls of radius $r_q = \frac{1}{2}d(p, q)$ about p and q are disjoint; as q ranges over K they cover it, and a finite subcover walls p off — so K^c is open (proof of 4.1.1).

Proposition 4.1.3

A closed subset of a compact set is compact

If K is compact and $F \subseteq K$ is closed in K, then F is compact.

Proof 4.1.3

Direct Proof that a Closed Subset of a Compact Set Is Compact.

(1) Let $\{V_\alpha\}$ be an open cover of F. Since F is closed, F^c is open in K, and $\{V_\alpha\} \cup \{F^c\}$ is an open cover of K.

(2) By compactness, finitely many cover K: $K \subseteq V_{\alpha_1} \cup \cdots \cup V_{\alpha_n} \cup F^c$.

(3) Since $F \cap F^c = \emptyset$, dropping F^c still covers F: $F \subseteq V_{\alpha_1} \cup \cdots \cup V_{\alpha_n}$. So every open

[26] Bounded: fix any p; the balls $B_n(p)$ $(n \geq 1)$ cover K, so finitely many do, and $K \subseteq B_N(p)$ for the largest N.

cover of F has a finite subcover, and F is compact.

■

[Direct] cf. R 2.35.

4.2 The Finite-Intersection Property and Nested Sets

Proposition 4.2.1

The finite-intersection property

Let $\{K_\alpha\}$ be a family of nonempty compact subsets of X such that every finite subfamily has nonempty intersection. Then $\bigcap_\alpha K_\alpha \neq \emptyset$.[27]

Proof 4.2.1

Proof by Contradiction of the Finite-Intersection Property.

(1) Suppose $\bigcap_\alpha K_\alpha = \emptyset$ and fix one K_1. Then $K_1 \cap \bigcap_{\alpha\neq 1} K_\alpha = \emptyset$, i.e. $K_1 \subseteq \left(\bigcap_{\alpha\neq 1} K_\alpha\right)^c = \bigcup_{\alpha\neq 1} K_\alpha^c$.

(2) Each K_α is compact, hence closed (4.1.1), so each K_α^c is open: $\{K_\alpha^c\}_{\alpha\neq 1}$ is an open cover of K_1.

(3) By compactness of K_1, finitely many cover it: $K_1 \subseteq K_{\alpha_1}^c \cup \cdots \cup K_{\alpha_n}^c$, i.e. $K_1 \cap K_{\alpha_1} \cap \cdots \cap K_{\alpha_n} = \emptyset$ — a finite subfamily with empty intersection, contradicting the hypothesis.

■

[A9] uses 4.1.1; cf. R 2.36.

Corollary 4.2.2

Nested nonempty compact sets meet

A nested sequence $K_1 \supseteq K_2 \supseteq \cdots$ of nonempty compact sets has $\bigcap_n K_n \neq \emptyset$.[28] This is the countable nested compact corollary to Rudin's finite-intersection theorem (R 2.36).

Proposition 4.2.3

Nested closed intervals in $\mathbb{R}$

A nested sequence of nonempty closed intervals $I_n = [a_n, b_n]$ $(I_{n+1} \subseteq I_n)$ has $\bigcap_n I_n = [\sup_n a_n,\ \inf_n b_n] \neq \emptyset$.

[27] Read this as the open-cover definition turned inside out. Complementing the K_α trades "no finite subfamily covers" for "no finite subfamily has empty intersection," so a family of compact sets in which every finite batch meets cannot have empty total intersection. The moral to keep: with compactness, sharing every *finite* subfamily is enough to force one common point—nothing can leak away to infinity.

[28] A descending chain of nonempty compact sets cannot shrink away to nothing. This is the shape of almost every existence proof that follows: trap your target inside nested compact sets whose diameters tend to 0, and the single point they all share is the object you were chasing—the move behind k-cells being compact (4.3.1) and behind finding limits of sequences. Compactness is exactly what stops the chain from closing on the empty set.

Proof 4.2.3

Proof by a Supremum Argument that Nested Closed Intervals Meet.

(1) Nesting gives $a_1 \le a_2 \le \cdots \le b_2 \le b_1$, so every b_m is an upper bound for $\{a_n\}$. By the least-upper-bound property (1.6.1), $\alpha = \sup_n a_n$ exists and $a_n \le \alpha \le b_n$ for all n; thus $\alpha \in \bigcap_n I_n$.

(2) Moreover $x \in \bigcap_n I_n \iff a_n \le x \le b_n \ \forall n \iff \sup_n a_n \le x \le \inf_n b_n$, so $\bigcap_n I_n = [\sup_n a_n, \inf_n b_n]$.

■

[A5] uses 1.6.1; cf. R 2.38.

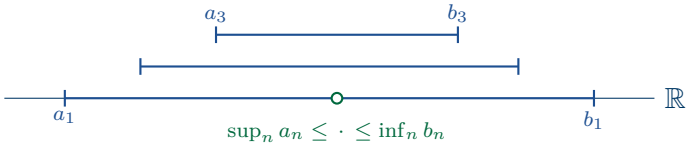


Figure 4.2.4. NESTED CLOSED INTERVALS. A descending chain $[a_1, b_1] \supseteq [a_2, b_2] \supseteq \cdots$ shares every point of $[\sup_n a_n, \inf_n b_n]$ — in particular it is never empty (4.2.3).

Definition 4.2.5

k-cell

A *k-cell* (R 2.17) is a product of closed intervals

$$I = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_k, b_k] \subseteq \mathbb{R}^k.$$

Proposition 4.2.6

Nested k-cells meet

A nested sequence of nonempty k-cells $I_1 \supseteq I_2 \supseteq \cdots$ has $\bigcap_n I_n \neq \emptyset$.

Proof 4.2.6

Proof Coordinatewise that Nested k-Cells Meet.

(1) Write $I_n = [a_1^{(n)}, b_1^{(n)}] \times \cdots \times [a_k^{(n)}, b_k^{(n)}]$. For each fixed coordinate i the intervals $[a_i^{(n)}, b_i^{(n)}]$ are nested, so by 4.2.3 some $\alpha_i \in \bigcap_n [a_i^{(n)}, b_i^{(n)}]$.

(2) Then $\alpha = (\alpha_1, \ldots, \alpha_k) \in \bigcap_n I_n$.

■

[A5] applies 4.2.3 in each coordinate; cf. R 2.39.

4.3 Heine–Borel and Bolzano–Weierstrass

Theorem 4.3.1

Every k-cell is compact

Every k-cell $I \subseteq \mathbb{R}^k$ is compact.[29]

Proof 4.3.1

Proof by Repeated Bisection that the k-cell Is Compact.

(1) Let $\delta = \left(\sum_{i=1}^{k}(b_i - a_i)^2\right)^{1/2}$, so $d(x, y) \leq \delta$ for all $x, y \in I$. Suppose I is *not* compact: some open cover $\{G_\alpha\}$ has no finite subcover.

(2) Bisecting each edge at its midpoint splits I into 2^k sub-cells. If each had a finite subcover, so would I; hence some sub-cell I_1 has none. Iterating gives nested cells $I \supseteq I_1 \supseteq I_2 \supseteq \cdots$, each without a finite subcover, with $\operatorname{diam} I_n \leq 2^{-n}\delta$ (each bisection halves every edge, hence the diagonal).

(3) By 4.2.6 some $x \in \bigcap_n I_n$. As $\{G_\alpha\}$ covers I, $x \in G_{\alpha_0}$ for some α_0, and G_{α_0} open gives $r > 0$ with $B_r(x) \subseteq G_{\alpha_0}$. Choose N with $2^{-N}\delta < r$ (Archimedean, 1.7.1); then $I_N \subseteq B_r(x) \subseteq G_{\alpha_0}$ — so I_N is covered by the single G_{α_0}, contradicting that it has no finite subcover.

■

[A7] bisection and nested cells (4.2.6); cf. R 2.40.

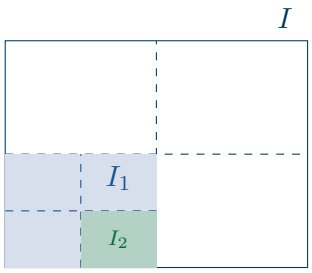


Figure 4.3.2. Bisecting a k-cell. A k-cell without a finite subcover has a 2^k-bisection inheriting the defect; iterating gives nested cells $I \supset I_1 \supset I_2$ of diameter $\to 0$, trapping a point that a single cover element must contain (proof of 4.3.1).

Theorem 4.3.3

Bolzano–Weierstrass

Every infinite subset E of a compact metric space[30] K has a cluster point in K.

Proof 4.3.3

Proof by Contradiction that E Would Be Finite.

[29] Why is a solid box compact when an open or unbounded set is not? Bisect it into 2^k smaller boxes: if no finite part of a cover handled the whole box, some sub-box inherits the same defect, and repeating traps a single point whose own little neighbourhood—sitting inside *one* set of the cover—then swallows an entire tiny box, the contradiction. The trap works because halving sends the diameter to 0 across finitely many coordinates; that is just what breaks down in infinitely many dimensions, and why Heine–Borel (4.3.5) is a fact about $\mathbb{R}^k$ and not metric spaces at large.

[30] Rudin reserves the name *Bolzano–Weierstrass* (R 2.42) for the bounded-$\mathbb{R}^k$ case; the general compact statement here is R 2.37, whence the $\mathbb{R}^k$ case follows by enclosing a bounded set in a k-cell (4.3.1).

(1) Suppose no point of K is a cluster point of E. Then each $q \in K$ has $r_q > 0$ with $B_{r_q}(q) \cap E \subseteq \{q\}$.

(2) The balls $\{B_{r_q}(q)\}_{q \in K}$ cover K; by compactness finitely many do. Each holds at most one point of E, so E is finite — contradicting that E is infinite.

■

[A9] cf. R 2.37.

Definition 4.3.4

Bounded set

$E \subseteq X$ is *bounded* (R 2.18) if $E \subseteq B_r(p)$ for some $p \in X$ and $r > 0$.

Theorem 4.3.5

Heine–Borel

For $E \subseteq \mathbb{R}^k$ the following are equivalent:[31]

1. E is closed and bounded;
2. E is compact;
3. every infinite subset of E has a cluster point in E.

Proof 4.3.5

Proof of the Heine–Borel Theorem by Cycling (a)⇒(b)⇒(c)⇒(a).

(1) (a)⇒(b). Bounded E lies in some k-cell I; I is compact (4.3.1) and $E \subseteq I$ is closed, so E is compact (4.1.3).

(2) (b)⇒(c). This is Bolzano–Weierstrass (4.3.3).

(3) (c)⇒(a), contrapositive. If E is *not closed*, a cluster point $q \notin E$ exists; choose $p_n \in E$ with $\|p_n - q\| < \frac{1}{n}$. The set $\{p_n\}$ is infinite (a finite range would repeat some value $p \in E$ infinitely, forcing $q = p \in E$), and its only possible cluster point is $q \notin E$: any $y \in E$ has $\|p_n - y\| \geq \|q - y\| - \frac{1}{n} \geq \frac{1}{2}\|q - y\|$ eventually. If E is *not bounded*, choose $p_n \in E$ with $\|p_n\| > n$; then $\{p_n\}$ is infinite (the norms are unbounded) with no cluster point, since any $B_1(y)$ holds only finitely many p_n. Either way (c) fails.

■

[A9+A4] uses 4.1.3, 4.3.1, 4.3.3; cf. R 2.41.

4.4 Interior, Closure, and Connectedness

Definition 4.4.1

Interior and closure

[31] The equivalence (a) $\iff$ (b) is *special to* $\mathbb{R}^k$: in a general metric space a compact set is closed and bounded (4.1.1), but closed-and-bounded need not be compact.

For $E \subseteq X$, the *interior* $\mathring{E}$ is the set of interior points of E (2.3.3); it is open, and E is open iff $E = \mathring{E}$. The *closure* (R 2.27) is

$$\overline{E} = E \cup \{\text{cluster points of } E\} \text{ (3.2.1)};$$

it is closed — indeed the smallest closed set containing E.[32]

Definition 4.4.2

Connected and disconnected

A metric space (X, d) is *connected* (R 2.45) if its only subsets that are both open and closed ("clopen") are $\emptyset$ and X. Equivalently, X is *disconnected* when $X = A \cup B$ for nonempty, disjoint, open sets A, B.[33]

Theorem 4.4.3

Connected subsets of $\mathbb{R}$ are exactly the intervals

A subset $E \subseteq \mathbb{R}$ is connected (4.4.2) if and only if it is an *interval* (R 2.47): whenever $x, y \in E$ and $x < z < y$, we have $z \in E$.[34]

Proof 4.4.3

Proof That a Subset of $\mathbb{R}$ *Is Connected Iff It Is an Interval.*

(1) ($\Rightarrow$), contrapositive. Suppose E is *not* an interval: there are $x, y \in E$ and $z \notin E$ with $x < z < y$. In the subspace E put

$$U = (-\infty, z) \cap E, \qquad V = (z, \infty) \cap E.$$

Both are open in E (intersections of E with open subsets of $\mathbb{R}$), nonempty ($x \in U$, $y \in V$), and disjoint. Since $z \notin E$, we have $U \cup V = E$. So $E = U \cup V$ is a disconnection, and E is not connected (4.4.2).

(2) ($\Leftarrow$). Suppose E has the interval property but is *not* connected, say $E = A \cup B$ with A, B nonempty, open in E, and disjoint. Pick $x \in A$ and $y \in B$; relabel so that $x < y$, and set

$$z = \sup\,(A \cap [x, y]),$$

which exists by the least-upper-bound property (1.6.1) since $A \cap [x, y]$ is nonempty

[32] If $C \supseteq E$ is closed then C contains every cluster point of E, so $C \supseteq \overline{E}$; and $\overline{E}$ is itself closed, so it is the least such set.

[33] The picture is just "one piece." A clopen set other than $\emptyset$ or X would split X into two parts, each open, with neither approaching the other—a clean break. Requiring the only clopen sets to be $\emptyset$ and X says precisely that no such split exists, so the space hangs together in a single piece.

[34] The two directions split the labour cleanly. "Not an interval" means E has a visible gap at some z, and the two open rays on either side of z are the separation you want—no metric subtlety needed. The reverse direction is the only place the completeness of $\mathbb{R}$ enters: a disconnection of an interval would have to have a boundary point $z = \sup A$, and the least-upper-bound property is exactly what produces that z and traps it between the two halves. This is the fact our Intermediate Value Theorem (7.5.5) rests on—a continuous image of an interval is connected, hence again an interval.

(x lies in it) and bounded above by y. As $x \le z \le y$ and E has the interval property, $z \in E$, so $z \in A$ or $z \in B$.

(3) If $z \in A$, then $z \neq y$ (as $y \in B$), so $z < y$. Since A is open in E, some $(z-\varepsilon, z+\varepsilon) \cap E \subseteq A$; points just above z then lie in $A \cap [x, y]$, contradicting $z = \sup(A \cap [x, y])$.

(4) If $z \in B$, then $z \neq x$ (as $x \in A$), so $x < z$. Since B is open in E, some $(z - \varepsilon, z + \varepsilon) \cap E \subseteq B$; in particular $(z - \varepsilon, z) \cap E \subseteq B$ is disjoint from A, so no point of $A \cap [x, y]$ exceeds $z - \varepsilon$ — again contradicting $z = \sup(A \cap [x, y])$. Either way we reach a contradiction, so E is connected.

■

[A4+A5] uses 1.6.1; cf. R 2.47.

Definition 4.4.4

Dense and perfect

Let $E \subseteq X$. We call E *dense* in X (R 2.18) if $\overline{E} = X$ — equivalently, if every nonempty open subset of X meets E.[35] We call E *perfect* (R 2.43) if E is closed and every point of E is a cluster point of E — that is, $E = E'$, where E' is the derived set of cluster points.

Example 4.4.5

$\mathbb{Q}$ is dense in $\mathbb{R}$

$\mathbb{Q}$ is dense in $\mathbb{R}$: by the density of the rationals (1.7.2), every open interval $(a, b) \subseteq \mathbb{R}$ contains a rational. Hence every real number is a cluster point of $\mathbb{Q}$, and $\overline{\mathbb{Q}} = \mathbb{R}$.

[35] Density says E leaves no room: every point of X is either in E or arbitrarily well approximated by it, so you can never carve out an open patch of X that avoids E entirely. The two phrasings are the same fact seen from the closure (4.4.1) and from its complement.

LECTURE 5

Sequences: Convergence, Cauchy Completeness, and $\mathbb{R}^k$

MATH S4061 | Introduction to Modern Analysis I

Tuesday, 9 June 2026

> *A mathematician who is not also something of a poet will never be a complete mathematician.*
>
> Karl Weierstrass

Contents of this lecture

Overview. A sequence converges to p when it eventually lies in every ball about p. Two things must hold: the terms must settle, *and* a limit must exist in the space—the second can fail. We develop the calculus of limits (the open-set view, uniqueness, boundedness, the algebra of limits in $\mathbb{R}/\mathbb{C}$, coordinatewise convergence in $\mathbb{R}^k$), then *subsequences*—every sequence in a compact space has a convergent one—and finally *Cauchy sequences* and *completeness*: a Cauchy sequence is one whose terms bunch together, every convergent sequence is Cauchy, and a space is *complete* when the converse holds. Compact spaces and $\mathbb{R}^k$ are complete.

5.1 Convergence in a Metric Space

Definition 5.1.1

Convergent sequence

A sequence $\{p_n\}$ in (X, d) (2.1.6)[36] *converges* (R 3.1) to $p \in X$ if for every $\varepsilon > 0$ there is an N with $d(p_n, p) < \varepsilon$ for all $n \geq N$. We write $p_n \to p$ or $\lim_{n\to\infty} p_n = p$, and call p the *limit*.[37]

Remark 5.1.2

Two things must happen

Convergence asks both that the terms settle *and* that a limit exist *in* X. The latter can fail: in $X = (0, \infty)$ with the usual metric, $p_n = \frac{1}{n}$ has terms that settle, yet diverges—its would-be limit 0 is not in X (cf. R 3.1, R 3.12).

Proposition 5.1.3

Basic properties of limits

Let $\{p_n\}$ be a sequence in (X, d).

1. $p_n \to p \iff$ every open set $U \ni p$ contains p_n for all but finitely many n;
2. limits are unique: if $p_n \to p$ and $p_n \to p'$ then $p = p'$;
3. if $\{p_n\}$ converges then it is bounded;
4. if p is a cluster point of $E \subseteq X$, some sequence in $E \setminus \{p\}$ converges to p.

Proof 5.1.3

Direct Proof of the Four Properties.

(1) For the open-set criterion, if $p_n \to p$ and $U \ni p$ is open, some $B_r(p) \subseteq U$; past the N for $\varepsilon = r$, $p_n \in B_r(p) \subseteq U$. Conversely, $B_\varepsilon(p)$ is open and contains p, so it holds p_n for all but finitely many n; past those, $d(p_n, p) < \varepsilon$.

(2) For uniqueness, given $\varepsilon > 0$, $d(p, p') \leq d(p, p_n) + d(p_n, p') < 2\varepsilon$ for large n. As this holds for *every* $\varepsilon > 0$, $d(p, p') = 0$, so $p = p'$.

(3) For boundedness, take N with $d(p_n, p) < 1$ for $n \geq N$ ($\varepsilon = 1$). With $r = 1 + \max\{0, d(p_1, p), \ldots, d(p_{N-1}, p)\}$, every $p_n \in B_r(p)$, so $\{p_n\}$ is bounded.

(4) For the cluster-point claim, pick for each n a point $p_n \in B_{1/n}(p) \cap E$ with $p_n \neq p$ (possible since p is a cluster point). Then $d(p_n, p) < \frac{1}{n}$, so $p_n \to p$.

■

[Direct] cf. R 3.2. Uniqueness needs no Archimedean property: $d(p, p') < 2\varepsilon$ for all $\varepsilon > 0$ already forces $d(p, p') = 0$.

[36] We index from $n = 1$ here (Rudin's convention), so $\frac{1}{n}$ is always defined; the sequence definition 2.1.6 permits $n = 0$, which changes nothing.

[37] Read the two quantifiers as a challenge and a response: you name a tolerance ε—a ball $B_\varepsilon(p)$, however small—and the sequence answers with a stage N past which *every* later term sits inside that ball. Convergence is the promise that any challenge can be met: each ball about p, no matter how tight, eventually traps the whole tail. The order is what bites—ε comes first, and N may depend on it.

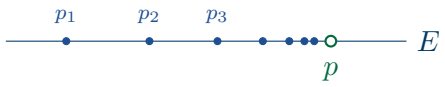


Figure 5.1.4. A cluster point is a limit of points of E. Choosing $p_n \in B_{1/n}(p) \cap E$ with $p_n \neq p$ produces a sequence in E converging to the cluster point p (5.1.3(d)).

5.2 The Algebra of Limits and Convergence in $\mathbb{R}^k$

Proposition 5.2.1

The algebra of limits

Let $s_n \to s$ and $t_n \to t$ in $\mathbb{R}$ (or $\mathbb{C}$). Then $s_n + t_n \to s + t$; $cs_n \to cs$ for any constant c; $s_n t_n \to st$; and $1/s_n \to 1/s$ provided $s_n, s \neq 0$.

Proof 5.2.1

Proof of the Algebra of Limits by Adding and Subtracting Cross Terms.

(1) For the sum, $|(s_n + t_n) - (s + t)| \leq |s_n - s| + |t_n - t| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ once both terms are $< \frac{\varepsilon}{2}$.

(2) For a scalar $c \neq 0$, $|cs_n - cs| = |c|\,|s_n - s| < |c| \cdot \frac{\varepsilon}{|c|} = \varepsilon$; the case $c = 0$ is trivial.

(3) For the product, writing $s_n t_n - st = (s_n - s)(t_n - t) + s(t_n - t) + t(s_n - s)$, the last two terms $\to 0$ by the scalar case, and $|(s_n - s)(t_n - t)| < \varepsilon^2$ eventually; so $s_n t_n \to st$.

(4) For the reciprocal, since $s_n \to s \neq 0$, eventually $|s_n| > \frac{1}{2}|s|$, so

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| = \frac{|s_n - s|}{|s_n|\,|s|} < \frac{2}{|s|^2}\,|s_n - s| \to 0.$$

■

[A4] cf. R 3.3.

Proposition 5.2.2

Coordinatewise convergence in $\mathbb{R}^k$

For $x_n = (\alpha_{1,n}, \ldots, \alpha_{k,n})$ and $x = (\alpha_1, \ldots, \alpha_k)$ in $\mathbb{R}^k$,[38]

$$x_n \to x \iff \alpha_{j,n} \to \alpha_j \text{ for every } j = 1, \ldots, k.$$

[38] This collapses convergence in $\mathbb{R}^k$ to k separate questions about real sequences: a vector sequence converges exactly when each coordinate does, to the matching coordinate of the limit. The reason is that the norm both dominates each single coordinate gap and is dominated by their sum, so "the whole vector is close" and "every coordinate is close" come to the same thing. In practice you rarely handle $\|x_n - x\|$ directly—you check the coordinates.

Proof 5.2.2

Proof of Coordinatewise Convergence by Comparing the Norm with Its Coordinates.

(1) ($\Rightarrow$) Each $|\alpha_{j,n} - \alpha_j| \le \|x_n - x\| = \left(\sum_i |\alpha_{i,n} - \alpha_i|^2\right)^{1/2} \to 0$.

(2) ($\Leftarrow$) Given $\varepsilon > 0$, choose N_j with $|\alpha_{j,n} - \alpha_j| < \varepsilon/\sqrt{k}$ for $n \ge N_j$; with $N = \max_j N_j$,

$$\|x_n - x\| = \Big(\sum_{j=1}^{k} |\alpha_{j,n} - \alpha_j|^2\Big)^{1/2} < \Big(k \cdot \tfrac{\varepsilon^2}{k}\Big)^{1/2} = \varepsilon.$$

■

[A4] cf. R 3.4.

The same argument works for any p-norm $\|\cdot\|_p$ or the sup norm; under the discrete metric, $p_n \to p$ iff $p_n = p$ for all but finitely many n.

5.3 Subsequences

Definition 5.3.1

Subsequence

Given a sequence $\{p_n\}$ and indices $n_1 < n_2 < n_3 < \cdots$, the sequence $\{p_{n_k}\}$ is a *subsequence* of $\{p_n\}$ (R 3.5). For instance $a_n = (-1)^n$ diverges, yet its even and odd subsequences converge, to $+1$ and -1.

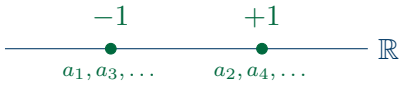


Figure 5.3.2. A DIVERGENT SEQUENCE WITH CONVERGENT SUBSEQUENCES. $a_n = (-1)^n$ never settles, but its even terms sit at $+1$ and its odd terms at -1: $a_{2k} \to 1$ and $a_{2k+1} \to -1$ (5.3.1).

Theorem 5.3.3

Convergent subsequences

1. In a compact metric space, every sequence has a convergent subsequence.
2. Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.[39]

Proof 5.3.3

Proof of the Convergent-Subsequence Theorem by Extraction at a Cluster Point.

[39] A sequence can refuse to settle and still be forced to settle *somewhere*: if its terms cannot escape—trapped in a compact space, or merely bounded in $\mathbb{R}^k$—then infinitely many of them pile up near some point, and reading those off gives a convergent subsequence. The hypothesis does all the work; drop boundedness and $p_n = n$ has no convergent subsequence at all. This is the sequential face of compactness.

(1) For (a), if $\{p_n\}$ takes only finitely many values, one value recurs infinitely often, giving a constant—hence convergent—subsequence. Otherwise the set $\{p_n\}$ is infinite, so by Bolzano–Weierstrass (4.3.3) it has a cluster point p. Choose $n_1 < n_2 < \cdots$ with $d(p_{n_k}, p) < \frac{1}{k}$ (each ball $B_{1/k}(p)$ meets the infinite value-set in terms of arbitrarily large index); then $p_{n_k} \to p$.

(2) For (b), a bounded sequence in $\mathbb{R}^k$ lies in a k-cell, which is compact (4.3.1); now apply (a).

■

[A8] uses 4.3.3, 4.3.1; cf. R 3.6.

5.4 Cauchy Sequences and Completeness

Definition 5.4.1

Cauchy sequence

A sequence $\{p_n\}$ in (X, d) is a *Cauchy sequence* (R 3.8) if for every $\varepsilon > 0$ there is an N with $d(p_m, p_n) < \varepsilon$ for all $m, n \geq N$. No limit is named.[40]

Proposition 5.4.2

Convergent sequences are Cauchy

Every convergent sequence is Cauchy.

Proof 5.4.2

Direct Proof that Convergent Sequences Are Cauchy by Halving ε.

(1) If $p_n \to p$, given $\varepsilon > 0$ take N with $d(p_n, p) < \frac{\varepsilon}{2}$ for $n \geq N$. Then for $m, n \geq N$, $d(p_m, p_n) \leq d(p_m, p) + d(p, p_n) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$.

■

[A4] cf. R 3.11.

Definition 5.4.3

Complete metric space

(X, d) is *complete* (R 3.12) if every Cauchy sequence in X converges.[41] Thus $\mathbb{R}$, $\mathbb{R}^k$, $\mathbb{C}$, and every compact space are complete (5.5.3), while $\mathbb{Q}$ and $(0, 1)$ are not—a sequence of rationals approaching $\sqrt{2}$ is Cauchy but has no rational limit. A closed subset of a complete space is itself complete (a Cauchy sequence in it converges in the ambient space, and the limit lies in the subset since it is closed).

[40] Set this beside convergence (5.1.1): there a single target p is named and the terms must approach *it*; here only the terms appear, and they must approach *each other*. That makes "Cauchy" an internal test—checkable knowing nothing about where the sequence might be heading, or whether it heads anywhere at all. The interest of everything that follows is the gap between bunching together and actually landing.

[41] Read completeness as a promise about the *space*, not the sequence: it says the space has no holes, so any sequence whose terms bunch up (5.4.1) finds an actual limit waiting inside. The Cauchy condition is the sequence trying to converge; completeness is the space letting it. Where the promise fails—as in $\mathbb{Q}$ or $(0, 1)$ just below—the terms can bunch arbitrarily tightly around a point that simply is not there.

Theorem 5.4.4

Cauchy completion

Every metric space (X, d) embeds isometrically as a dense subspace of a complete metric space $(\widehat{X}, \widehat{d}\,)$, essentially unique, where $\widehat{X}$ is the set of Cauchy sequences in X modulo $\{p_n\} \sim \{q_n\} \iff d(p_n, q_n) \to 0$. (Adjoining the suprema of bounded sets to $\mathbb{Q}$ to build $\mathbb{R}$, Lecture 1, is this construction in disguise.)

Proof 5.4.4

Construction of the Completion by Equivalence Classes of Cauchy Sequences.

↪ Rudin, Exercise 3.24; each $p \in X$ maps to the class of the constant sequence $[p, p, \dots]$.

5.5 Completeness of Compact Spaces and of $\mathbb{R}^k$

Definition 5.5.1

Diameter and the tail of a sequence

The *diameter* of a nonempty $E \subseteq X$ is $\operatorname{diam} E = \sup_{p,q \in E} d(p, q)$ (R 3.9) (finite iff E is bounded). The *tail* of $\{p_n\}$ at N is $E_N = \{p_N, p_{N+1}, \dots\}$. Then $\{p_n\}$ is Cauchy $\iff$ $\operatorname{diam} E_N \to 0$.[42]

Proposition 5.5.2

Diameter and nested compacts

Let (X, d) be a metric space.

1. $\operatorname{diam} E = \operatorname{diam} \overline{E}$ for every nonempty $E \subseteq X$;
2. a nested sequence $K_1 \supseteq K_2 \supseteq \cdots$ of nonempty compact sets with $\operatorname{diam} K_n \to 0$ has $\bigcap_n K_n$ a single point.

Proof 5.5.2

Proof of the Diameter and Nested-Compact Claims by Approximation and Contradiction.

(1) For (1), $E \subseteq \overline{E}$ gives $\operatorname{diam} E \le \operatorname{diam} \overline{E}$. For $p, q \in \overline{E}$ and $\varepsilon > 0$, pick $p', q' \in E$ with $d(p, p'), d(q, q') \le \varepsilon$; then $d(p, q) \le d(p, p') + d(p', q') + d(q', q) \le \operatorname{diam} E + 2\varepsilon$. Taking the supremum over p, q and letting $\varepsilon \to 0$, $\operatorname{diam} \overline{E} \le \operatorname{diam} E$.

(2) For (2), $K = \bigcap_n K_n \neq \emptyset$ (4.2.2). If K held two distinct points then $\operatorname{diam} K > 0$; but $K \subseteq K_n$ gives $\operatorname{diam} K \le \operatorname{diam} K_n \to 0$, a contradiction. So K is a single point. ∎

[A4+A9] uses 4.2.2; cf. R 3.10.

Theorem 5.5.3

Compact spaces and $\mathbb{R}^k$ are complete

A compact metric space is complete; and $\mathbb{R}^k$ (with the usual metric) is complete.

[42] $\operatorname{diam} E_N \le \varepsilon$ says exactly that $d(p_m, p_n) \le \varepsilon$ for all $m, n \ge N$.

Proof 5.5.3

Proof of Completeness by Trapping the Limit in Shrinking Tails.

(1) For the compact case, let $\{p_n\}$ be Cauchy in compact X. The tail-closures $\overline{E_N}$ are closed in X, hence compact (4.1.3), are nested, and $\operatorname{diam} \overline{E_N} = \operatorname{diam} E_N \to 0$ (Cauchy, 5.5.2(1)). By 5.5.2(2), $\bigcap_N \overline{E_N} = \{p\}$, and then $p_n \to p$: given $\varepsilon > 0$, $\operatorname{diam} \overline{E_N} \le \varepsilon$ for large N forces $d(p_n, p) \le \varepsilon$ there.

(2) For $\mathbb{R}^k$, a Cauchy sequence is bounded—take N with $\|p_m - p_n\| < 1$ for $m, n \ge N$, so the tail lies in $B_1(p_N)$ while the finitely many earlier terms are bounded—hence its closure is closed and bounded, so compact (Heine–Borel, 4.3.5); a Cauchy sequence in that compact set converges by the first part. So every Cauchy sequence in $\mathbb{R}^k$ converges.

■

[A7] uses 4.1.3, 4.3.5, 5.5.2; cf. R 3.11. Here $\overline{E_N}$ is compact as a closed subset of the compact X (in $\mathbb{R}^k$, closed-and-bounded).

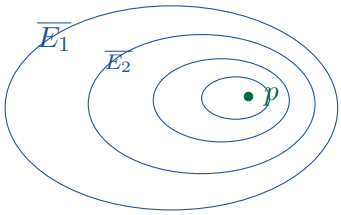


Figure 5.5.4. Trapping the limit in shrinking tails. For a Cauchy sequence the tail-closures $\overline{E_1} \supseteq \overline{E_2} \supseteq \cdots$ are nested compacts with $\operatorname{diam} \overline{E_N} \to 0$; their single common point is the limit (proof of 5.5.3).

LECTURE 6

Monotone Convergence, the Squeeze Theorem, and the Upper and Lower Limits

MATH S4061 | Introduction to Modern Analysis I

Thursday, 11 June 2026

> *Of the small there is no smallest, but always a smaller; and of the large there is always a larger.*
>
> Anaxagoras, fragment B3 (5th c. bce)

Contents of this lecture

Overview. Picking up from completeness, this lecture assembles the basic machinery for limits of real sequences. Monotonicity plus boundedness already forces convergence; the comparison and squeeze principles let limits be read off from neighbours; and for sequences that oscillate, the upper and lower limits lim sup and lim inf record the band the sequence is eventually trapped in, coinciding exactly when the sequence converges. After recording four limits that recur throughout the subject, we open the theory of series: convergence through partial sums, the Cauchy criterion, and the first divergence test. The companion text is Rudin, Chapter 3, especially R 3.12–R 3.40.

6.1 Review of Completeness and Cauchy Sequences

Definition 6.1.1

Complete metric space

A metric space (X, d) is *complete* (R 3.12) if every Cauchy sequence in X converges to a point of X.

Example 6.1.2

Complete and incomplete spaces

- Compact metric spaces are complete (R 3.11).
- $\mathbb{R}^k$ is complete (R 3.11).
- $\mathbb{Q}$ and $(0, 1)$ are *not* complete (cf. R 3.12).

Remark 6.1.3

Testing convergence without the limit

In a complete space — in particular in $\mathbb{R}^k$ — it is enough to check whether a sequence is Cauchy; there is no need to know the limit in advance. In an incomplete space such as $\mathbb{Q}$ this fails: a Cauchy sequence may have no limit (6.1.2; cf. R 3.11–R 3.12).

6.2 Monotone Sequences

Definition 6.2.1

Monotonic sequences

A sequence $\{s_n\}$ is monotonic in Rudin's sense (R 3.13) as follows:

- *monotonic increasing* if $s_n \leq s_{n+1}$ for all n;
- *monotonic decreasing* if $s_n \geq s_{n+1}$ for all n.

"Monotonic" means either increasing or decreasing.

Theorem 6.2.2

Monotone convergence

If $\{s_n\} \subseteq \mathbb{R}$ is monotonic and bounded, then $\{s_n\}$ converges (R 3.14).[43]

Proof 6.2.2

Proof by a Supremum Argument of the Monotone Convergence Theorem.

(1) Treat the increasing case (the decreasing case is analogous). Since $\{s_n\}$ is bounded above, the least-upper-bound property of $\mathbb{R}$ provides

$$s := \sup_n s_n.$$

[43] This is the first place completeness earns its keep by *manufacturing* a limit rather than asking you to guess one. A monotone sequence can head only one way, and boundedness blocks the exit, so the terms have nowhere to go but to pile up against their least upper bound—and that bound, $\sup_n s_n$, *is* the limit. Read the two hypotheses as a division of labour: monotone forbids oscillation, bounded forbids escape to infinity.

(2) Let $\varepsilon > 0$. Because $s - \varepsilon < s$, the number $s - \varepsilon$ is not an upper bound of $\{s_n\}$, so there is N_ε with

$$s_{N_\varepsilon} > s - \varepsilon.$$

(3) As the sequence is increasing and s is an upper bound, for every $n \geq N_\varepsilon$

$$s - \varepsilon < s_{N_\varepsilon} \leq s_n \leq s,$$

hence $|s_n - s| < \varepsilon$. Therefore $s_n \to s$.

■

[A5] cf. R 3.14; the decreasing case follows by applying the increasing case to $\{-s_n\}$.

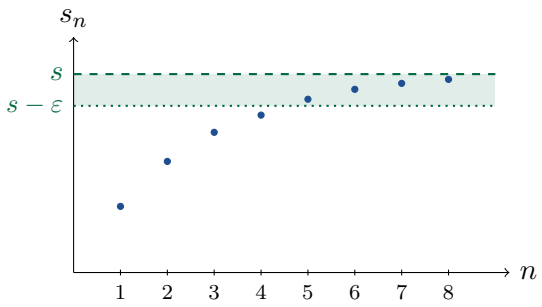


Figure 6.2.3. CLIMBING TO THE SUPREMUM. An increasing sequence bounded above closes in on $s = \sup_n s_n$: the terms rise but never cross the dashed bound, and from some N_ε on they all lie in the band $(s - \varepsilon,\, s]$.

6.3 Comparison and Squeeze

Proposition 6.3.1

Order is preserved in the limit

Let $\{s_n\}$ and $\{t_n\}$ be convergent real sequences with $s_n \leq t_n$ for all n. Then

$$\lim_{n\to\infty} s_n \;\leq\; \lim_{n\to\infty} t_n.$$

This is the ordinary-limit form of Rudin's comparison theorem for upper and lower limits (cf. R 3.19).

Idea. Suppose instead that $s := \lim s_n$ and $t := \lim t_n$ satisfy $s > t$. Separate s and t by disjoint ε-balls, with $\varepsilon < \frac{1}{2}|s - t|$, so that $B_\varepsilon(t)$ and $B_\varepsilon(s)$ sit on opposite sides of the midpoint $\frac{s+t}{2}$.

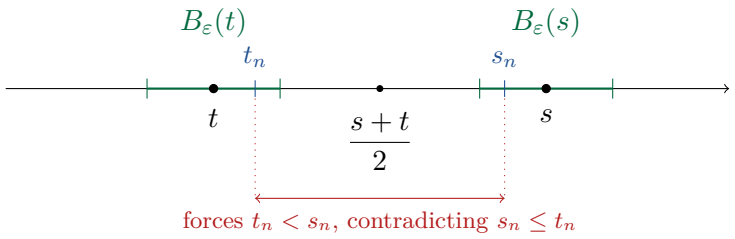


Figure 6.3.2. Separating the limits. With $\varepsilon < \frac{1}{2}|s-t|$, the balls $B_\varepsilon(t)$ and $B_\varepsilon(s)$ fall on opposite sides of the midpoint $\frac{s+t}{2}$, which is impossible once $s_n \le t_n$.

Proof 6.3.1

Proof by Contradiction that Order Passes to the Limit.

(1) Write $s = \lim s_n$, $t = \lim t_n$, and suppose for contradiction $s > t$. Since $s - t > 0$, choose

$$0 < \varepsilon < \tfrac{1}{2}|s-t| = \tfrac{1}{2}(s-t).$$

(2) By convergence there are N_1, N_2 with $|s_n - s| < \varepsilon$ for $n \ge N_1$ and $|t_n - t| < \varepsilon$ for $n \ge N_2$. With $N = \max\{N_1, N_2\}$, for every $n \ge N$

$$s_n > s - \varepsilon \qquad \text{and} \qquad t_n < t + \varepsilon.$$

(3) The choice of ε puts these on opposite sides of the midpoint:

$$t + \varepsilon < t + \tfrac{1}{2}(s-t) = \frac{s+t}{2} = s - \tfrac{1}{2}(s-t) < s - \varepsilon.$$

(4) Chaining the inequalities, for every $n \ge N$,

$$t_n < t + \varepsilon < \frac{s+t}{2} < s - \varepsilon < s_n,$$

so $t_n < s_n$, contradicting $s_n \le t_n$. Hence $s \le t$, i.e. $\lim s_n \le \lim t_n$.

■

[A9]

Proposition 6.3.3

Squeeze theorem

Suppose $r_n \le s_n \le t_n$ for all n, where $r_n \to \ell$ and $t_n \to \ell$. Then $s_n \to \ell$. This is the comparison principle used in Rudin's special-sequence computations (cf. R 3.20).

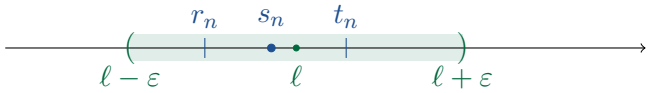


Figure 6.3.4. THE SQUEEZE. Once r_n and t_n lie in $(\ell-\varepsilon, \ell+\varepsilon)$, the trapped term s_n must lie there too.

Proof 6.3.3

Proof by an ε–N Estimate of the Squeeze Theorem.

(1) Let $\varepsilon > 0$. Since $r_n \to \ell$ and $t_n \to \ell$, there is N with, for all $n \geq N$,

$$|r_n - \ell| < \varepsilon \qquad \text{and} \qquad |t_n - \ell| < \varepsilon,$$

so $r_n, t_n \in (\ell-\varepsilon, \ell+\varepsilon)$.

(2) Using $r_n \leq s_n \leq t_n$, for every $n \geq N$

$$\ell - \varepsilon < r_n \leq s_n \leq t_n < \ell + \varepsilon,$$

hence $|s_n - \ell| < \varepsilon$. As $\varepsilon > 0$ was arbitrary, $s_n \to \ell$.

■

[A2]

6.4 The Upper and Lower Limits

Convergence $s_n \to \ell$ says that for every $\varepsilon > 0$ there is N with $s_n \in (\ell-\varepsilon, \ell+\varepsilon)$ for $n \geq N$. For a bounded sequence that does *not* converge, no single ℓ works. Instead there are two values $\ell^- \leq \ell^+$ such that the sequence is eventually trapped in any ε-enlargement of the band $[\ell^-, \ell^+]$:

$$\forall\, \varepsilon > 0 \;\; \exists\, N \;\; \text{s.t.} \;\; s_n \in (\ell^- - \varepsilon,\ \ell^+ + \varepsilon) \quad \text{for } n \geq N.$$

We write $\ell^- = \liminf s_n$ and $\ell^+ = \limsup s_n$.

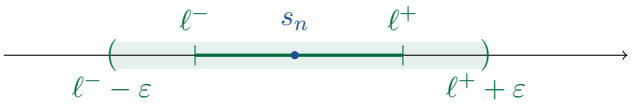


Figure 6.4.1. THE TRAPPING BAND. For a bounded sequence, every ε-enlargement of $[\ell^-, \ell^+]$ eventually contains all the terms.

The two coincide exactly when the sequence converges:

$$s_n \to \ell \iff \liminf s_n = \limsup s_n = \ell.$$

Example 6.4.2

An oscillating sequence

For $s_n = (-1)^n \dfrac{n}{n-1}$ (with $n \geq 2$), the even terms approach $+1$ from above and the odd terms approach -1 from below. Hence $\liminf s_n = -1$ and $\limsup s_n = +1$, and the sequence does not converge, since these differ (cf. R 3.18).

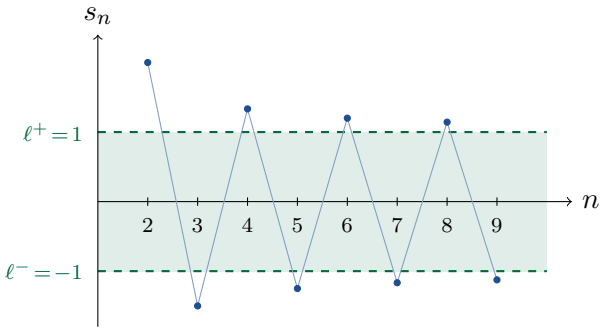


Figure 6.4.3. Two limit points. The oscillating sequence $s_n = (-1)^n \frac{n}{n-1}$ has no limit: the even terms fall toward $\ell^+ = \limsup s_n = 1$ from above and the odd terms rise toward $\ell^- = \liminf s_n = -1$ from below. Every ε-enlargement of the shaded band $[\ell^-, \ell^+]$ eventually traps the whole tail.

Definition 6.4.4

Divergence to $\pm\infty$

For a real sequence $\{t_n\}$ (R 3.15),

- $t_n \to +\infty$ means: for every $M > 0$ there is N with $t_n > M$ for all $n \geq N$;
- $t_n \to -\infty$ means: for every $M > 0$ there is N with $t_n < -M$ for all $n \geq N$.

Such sequences diverge, but in a specific, controlled way.

Definition 6.4.5

Limit inferior and limit superior

The *upper* and *lower limits* of a real sequence $\{s_n\}$ are defined by the tail extrema

$$\limsup s_n = \lim_{n\to\infty} \Big(\sup_{k \geq n} s_k \Big), \qquad \liminf s_n = \lim_{n\to\infty} \Big(\inf_{k \geq n} s_k \Big).$$

The tails are nested, $\{s_k : k \geq n+1\} \subseteq \{s_k : k \geq n\}$, so $\sup_{k\geq n} s_k$ is non-increasing in n (and $\inf_{k\geq n} s_k$ non-decreasing); hence both limits always exist in the extended reals $[-\infty, +\infty]$.[44] Two characterizations follow.

(1) Subsequential limits. If $\{s_n\}$ is bounded and E is its set of subsequential limits, then

$$\liminf s_n = \min E, \qquad \limsup s_n = \max E.$$

[44] Picture it this way: $\sup_{k\geq n} s_k$ is the ceiling over the tail from n onward, and discarding more early terms can only lower a supremum, so these ceilings march downward to a limit—the lim sup. It is the highest value the sequence keeps crowding back up to, while lim inf is the matching floor. A convergent sequence has ceiling and floor equal; a bounded oscillating one holds them apart, pinning the tail between two honest numbers (6.4.2).

(2) Eventual trapping (finite $\liminf, \limsup$). For every $\varepsilon > 0$ there is N_ε with

$$s_n \in \big(\liminf s_n - \varepsilon,\ \limsup s_n + \varepsilon\big) \qquad \text{for } n \geq N_\varepsilon,$$

and $[\liminf s_n,\ \limsup s_n]$ is the smallest closed interval with this property.

This tail-extrema presentation is equivalent to Rudin's subsequential-limit definition (R 3.16) and characterization (R 3.17).

Remark 6.4.6

The extreme limits are attained

For a *bounded* sequence, E is closed (6.4.8), bounded, and nonempty, so it contains its own extrema: $\inf E \in E$ and $\sup E \in E$. Hence $\liminf s_n$ and $\limsup s_n$ are themselves subsequential limits of $\{s_n\}$. (For an unbounded sequence this holds in the extended reals, with $\pm\infty$ admitted as subsequential limits.) Compare R 3.17.

Remark 6.4.7

Existence

If $\{s_n\}$ is bounded, then E is bounded and nonempty, so $\liminf s_n$ and $\limsup s_n$ exist as finite reals. If $\{s_n\}$ is unbounded, then E may include $+\infty$ and/or $-\infty$, and $\liminf s_n, \limsup s_n$ may equal $\pm\infty$ (cf. R 3.16–R 3.17).

Proposition 6.4.8

The set of subsequential limits is closed

Let $\{p_n\}$ be a sequence in a metric space (X, d), and let $E^* \subseteq X$ be its set of subsequential limits. Then E^* is a closed subset of X (R 3.7). In particular, for a bounded real sequence the set E of 6.4.5–6.4.6 is closed, justifying that it contains its extrema.

Proof 6.4.8

Proof by Diagonal Extraction that E^* is Closed.

(1) Let q be a cluster point of E^*; we construct a subsequence of $\{p_n\}$ converging to q, which shows $q \in E^*$ and hence that E^* is closed. Set $\delta = d(q, p_{n_1})$ where $p_{n_1} \neq q$ (if $\{p_n\}$ is eventually constant at q, then $q \in E^*$ trivially).

(2) Suppose $n_1, \ldots, n_{k-1}$ have been chosen. Since q is a cluster point of E^*, pick $x \in E^*$ with $d(x, q) < 2^{-k}\delta$. As $x \in E^*$, some subsequence of $\{p_n\}$ converges to x, so we may choose $n_k > n_{k-1}$ with $d(p_{n_k}, x) < 2^{-k}\delta$.

(3) By the triangle inequality,

$$d(p_{n_k}, q) \leq d(p_{n_k}, x) + d(x, q) < 2^{-k}\delta + 2^{-k}\delta = 2^{1-k}\delta \xrightarrow[k\to\infty]{} 0.$$

Hence $p_{n_k} \to q$, so $q \in E^*$. Therefore E^* contains all its cluster points and is closed.

■

[A8] cf. R 3.7; build a subsequence converging to a given cluster point of E^*.

6.5 Some Standard Limits

Proposition 6.5.1

Four standard limits

(a) If $|x| < 1$, then $x^n \to 0$.

(b) If $p > 0$, then $\dfrac{1}{n^p} \to 0$.

(c) If $p > 1$, then $\sqrt[n]{p} \to 1$.

(d) $\sqrt[n]{n} \to 1$.

These are the recurring special limits collected in R 3.20(a)–(c),(e).

Proof 6.5.1.a

Proof by the Binomial Inequality of $x^n \to 0$.

(1) If $x = 0$ then $x^n = 0 \to 0$, so assume $0 < |x| < 1$. Then $1/|x| > 1$, so $|x| = \dfrac{1}{1+p}$ for a unique $p > 0$ (the map $p \mapsto \frac{1}{1+p}$ is a bijection $(0,\infty) \to (0,1)$).

(2) By the binomial theorem, keeping the nonnegative $k = 0, 1$ terms,

$$(1+p)^n = \sum_{k=0}^{n} \binom{n}{k} p^k \;\geq\; 1 + np.$$

(3) Hence

$$|x|^n = \frac{1}{(1+p)^n} \leq \frac{1}{1+np} \xrightarrow[n\to\infty]{} 0,$$

and since $|x^n| = |x|^n \to 0$, we conclude $x^n \to 0$. ∎

[A2]

Proof 6.5.1.b

Proof by an ε*–*N *Estimate of* $1/n^p \to 0$.

(1) Let $\varepsilon > 0$ and choose N with $N > (1/\varepsilon)^{1/p}$, equivalently $\dfrac{1}{N^p} < \varepsilon$.

(2) Then for every $n \geq N$,

$$\frac{1}{n^p} \leq \frac{1}{N^p} < \varepsilon,$$

so $\dfrac{1}{n^p} \to 0$. ∎

[A2]

Proof 6.5.1.c

Proof by the Squeeze Theorem of $\sqrt[n]{p} \to 1$.

(1) Assume $p > 1$ and set $x_n = \sqrt[n]{p} - 1 \geq 0$. By the binomial inequality from (a),

$$1 + nx_n \leq (1 + x_n)^n = p,$$

so $0 \leq x_n \leq \dfrac{p-1}{n}$.

(2) Since $\dfrac{p-1}{n} \to 0$, the squeeze theorem (6.3.3) gives $x_n \to 0$, i.e. $\sqrt[n]{p} = 1 + x_n \to 1$. ■

[A2] uses 6.3.3.

Proof 6.5.1.d

Proof by the Squeeze Theorem of $\sqrt[n]{n} \to 1$.

(1) Set $x_n = \sqrt[n]{n} - 1 \geq 0$, so $n = (1 + x_n)^n$. The linear term is too weak, so keep the $k = 2$ term:

$$n = (1 + x_n)^n \geq \binom{n}{2} x_n^2 = \frac{n(n-1)}{2}\, x_n^2 \qquad (n \geq 2).$$

(2) Rearranging, $x_n^2 \leq \dfrac{2}{n-1}$, hence $0 \leq x_n \leq \sqrt{\dfrac{2}{n-1}}$.

(3) Since $\sqrt{\frac{2}{n-1}} \to 0$, the squeeze theorem (6.3.3) gives $x_n \to 0$, i.e. $\sqrt[n]{n} = 1 + x_n \to 1$. ■

[A2] keep the $k = 2$ binomial term; uses 6.3.3.

Remark 6.5.2

The case $0 < p < 1$ *in (c)*

For $0 < p < 1$, apply part (c) to $\sigma = 1/p > 1$: since $\sqrt[n]{\sigma} \to 1$, also $\sqrt[n]{p} = 1/\sqrt[n]{\sigma} \to 1$ (and $\sqrt[n]{1} = 1$). Hence $\sqrt[n]{p} \to 1$ for every $p > 0$; the full $p > 0$ statement is R 3.20(b).

Proposition 6.5.3

Polynomial versus geometric growth

If $p > 0$ and $\alpha \in \mathbb{R}$, then

$$\lim_{n \to \infty} \frac{n^\alpha}{(1+p)^n} = 0.$$

Powers of n grow more slowly than any fixed exponential base $1 + p > 1$; this is R 3.20(d).[45]

Proof 6.5.3

Proof by the Binomial Inequality and Squeeze of $n^\alpha/(1+p)^n \to 0$.

[45] The single linear or quadratic binomial term used for (a) and (d) above is no longer enough: against n^α one must keep a binomial term of degree higher than α. Keeping the k-th term with $k > \alpha$ buries the numerator's n^α under an n^k in the denominator, and the squeeze finishes the job.

(1) Choose an integer $k > \alpha$ with $k > 0$. For $n > 2k$ the binomial theorem, keeping only the nonnegative k-th term, gives

$$(1+p)^n \geq \binom{n}{k} p^k = \frac{n(n-1)\cdots(n-k+1)}{k!}\, p^k > \Big(\frac{n}{2}\Big)^k \frac{p^k}{k!},$$

since each of the k factors $n, n-1, \ldots, n-k+1$ exceeds $n/2$ once $n > 2k$.

(2) Inverting and multiplying by $n^\alpha > 0$,

$$0 < \frac{n^\alpha}{(1+p)^n} < \frac{2^k k!}{p^k}\, n^{\alpha-k} \qquad (n > 2k).$$

(3) Since $\alpha - k < 0$, write $n^{\alpha-k} = 1/n^{k-\alpha}$ with $k - \alpha > 0$; by 6.5.1(b) this tends to 0, so the right-hand bound tends to 0. The squeeze theorem (6.3.3) then gives $\dfrac{n^\alpha}{(1+p)^n} \to 0$.

■

[A2] cf. R 3.20(d); keep the k-th binomial term with $k > \alpha$; uses 6.3.3 and 6.5.1(b).

6.6 Introduction to Series

(Not covered on the midterm.)

As a field, $\mathbb{R}$ hands us directly only the polynomials $1, x, x^2, \ldots$ and, as their quotients, the rational functions $P(x)/Q(x)$. What about e^x, $\sin x$, $\cos x$? These are reached not by field operations but by infinite sums — for instance

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

To make sense of such a sum we must first say what an infinite sum *is*.

Definition 6.6.1

Series and partial sums

Given a sequence $\{a_n\}$ in $\mathbb{R}$, its *k-th partial sum* is

$$s_k = a_1 + \cdots + a_k = \sum_{i=1}^{k} a_i \qquad (s_0 = 0, \text{ the empty sum}).$$

If $s_k \to s$, the *series* $\sum a_n$ *converges*, and we write

$$\sum_{i=1}^{\infty} a_i = \lim_{k\to\infty} s_k = s.$$

A series may instead be indexed from any starting point—often $n = 0$, as for $\sum c_n x^n$—with

the partial sums adjusted accordingly.[46] This is Rudin's series convention (R 3.21) with the real-valued specialization used in lecture.

Theorem 6.6.2

Cauchy criterion for series

A series $\sum a_n$ converges if and only if for every $\varepsilon > 0$ there is N such that

$$\Big|\sum_{k=m}^{n} a_k\Big| < \varepsilon \qquad \text{for all } n \geq m \geq N.$$

This is the Cauchy criterion for series (R 3.22).

Proof 6.6.2

Proof by Completeness of $\mathbb{R}$ of the Cauchy Criterion.

(1) For $n \geq m \geq 1$ one has $s_n - s_{m-1} = \sum_{k=m}^{n} a_k$ (with $s_0 = 0$), so the displayed condition says precisely that the partial sums $\{s_k\}$ form a Cauchy sequence.

(2) Since $\mathbb{R}$ is complete (6.1.1), $\{s_k\}$ converges if and only if it is Cauchy. Hence $\sum a_n$ converges if and only if the criterion holds.

■

[Direct] cf. R 3.22; the partial sums converge iff they are Cauchy, and $\mathbb{R}$ is complete.

Corollary 6.6.3

Vanishing of the terms

If $\sum a_n$ converges, then $a_n \to 0$ (R 3.23).[47]

Proof 6.6.3

Proof by the Cauchy Criterion with $m = n$.

(1) Take $m = n$ in the criterion (6.6.2): for every $\varepsilon > 0$ there is N with

$$|a_n| = \Big|\sum_{k=n}^{n} a_k\Big| < \varepsilon \qquad \text{for all } n \geq N.$$

(2) Therefore $a_n \to 0$.

■

[Direct] cf. R 3.23.

So $a_n \to 0$ is *necessary* for convergence. Is it *sufficient*? No.

Proposition 6.6.4

[46] A series introduces no genuinely new operation; there is no act of adding infinitely many numbers at once. The symbol $\sum a_n$ is shorthand for a *sequence*—the partial sums s_k—together with its limit, so every question about a series is a question about that sequence, and every tool for sequences (monotonicity, the Cauchy criterion, the algebra of limits) carries straight over.

[47] Use this in one direction only. It says convergence *forces* the terms to die out, so if $a_n \not\to 0$ the series cannot converge—a quick divergence test. It does not say the reverse: terms dying out is not enough, as the harmonic series shows next (6.6.4). The terms must not merely vanish but vanish fast enough to be summable.

The harmonic series diverges

Although $\dfrac{1}{n} \to 0$, the harmonic series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. This is the $p = 1$ case of Rudin's p-series theorem (R 3.28).

Proof 6.6.4

Proof by Dyadic Grouping of the Divergence of the Harmonic Series.

(1) Round each term down to the next power of $\frac{1}{2}$: replace $\frac{1}{3}$ by $\frac{1}{4}$; then $\frac{1}{5}, \frac{1}{6}, \frac{1}{7}$ by $\frac{1}{8}$; and so on. If s'_k denotes the resulting partial sums, then $s_k \geq s'_k$ for every k.

(2) The rounded series groups into dyadic blocks of equal weight:

$$1 + \tfrac{1}{2} + \left(\tfrac{1}{4} + \tfrac{1}{4}\right) + \left(\tfrac{1}{8} + \tfrac{1}{8} + \tfrac{1}{8} + \tfrac{1}{8}\right) + \cdots = 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \cdots,$$

each block summing to $\frac{1}{2}$; hence $s'_{2^m} = 1 + \frac{m}{2}$.

(3) Therefore $s_{2^m} \geq s'_{2^m} = 1 + \frac{m}{2} \to \infty$, so the partial sums $\{s_n\}$ have a subsequence diverging to $+\infty$ and cannot converge. Thus $\sum \frac{1}{n}$ diverges.

■

[Direct] Cauchy condensation; cf. R 3.27–R 3.28.

Proposition 6.6.5

The p-series

The series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges for $p > 1$. Rudin proves the full convergence/divergence dichotomy in R 3.28.

Proof 6.6.5

Proof of the Convergence of the p-Series.

↪ Stated in lecture as a fact, for use as a comparison series; cf. R 3.28.

Proposition 6.6.6

A nonnegative series converges when bounded

A series $\sum a_n$ of nonnegative terms ($a_n \geq 0$) converges if and only if its partial sums are bounded (R 3.24).

Proof 6.6.6

Proof by Monotone Convergence for a Series of Nonnegative Terms.

(1) Since $a_{n+1} \geq 0$, the partial sums satisfy $s_{n+1} = s_n + a_{n+1} \geq s_n$, so $\{s_n\}$ is increasing.

(2) If the partial sums are bounded, then $\{s_n\}$ is increasing and bounded, hence converges by the monotone convergence theorem (6.2.2); that is, $\sum a_n$ converges. Conversely, a convergent sequence is bounded.

■

[A5] cf. R 3.24; the partial sums increase; apply 6.2.2.

Proposition 6.6.7

The geometric series

For $|x| < 1$,

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

This is the convergent half of Rudin's geometric-series theorem (R 3.26).

Proof 6.6.7

Proof by the Partial-Sum Formula of the Geometric Series.

(1) For $x \neq 1$ the partial sums have the closed form

$$s_k = 1 + x + \cdots + x^k = \frac{1 - x^{k+1}}{1-x}.$$

(2) Since $|x| < 1$ gives $|x|^{k+1} \to 0$, we have $x^{k+1} \to 0$ and hence

$$s_k \xrightarrow[k\to\infty]{} \frac{1}{1-x},$$

so $\displaystyle\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$. ∎

[Direct] cf. R 3.26; uses $|x|^k \to 0$ for $|x| < 1$ (6.5.1).

Theorem 6.6.8

Comparison test for series

(a) If $|a_n| \le c_n$ for all $n \ge N_0$ and $\sum c_n$ converges, then $\sum a_n$ converges.

(b) If $a_n \ge d_n \ge 0$ for all n and $\sum d_n$ diverges, then $\sum a_n$ diverges.

This is Rudin's comparison test (R 3.25).[48]

Proof 6.6.8

Proof by the Cauchy Criterion of the Comparison Test.

(1) For (a), let $\varepsilon > 0$. Since $\sum c_n$ converges, the Cauchy criterion (6.6.2) gives some $N \ge N_0$ with $\sum_{k=m}^{n} c_k < \varepsilon$ for all $n \ge m \ge N$. Then for such n, m,

$$\Big|\sum_{k=m}^{n} a_k\Big| \le \sum_{k=m}^{n} |a_k| \le \sum_{k=m}^{n} c_k < \varepsilon,$$

so $\sum a_n$ meets the Cauchy criterion and therefore converges.

[48] Convergence of a series is really a question of how fast the terms shrink, and comparison is the basic way to settle it: outrun a series known to converge and you converge too; be outrun by one known to diverge and you diverge. Most of the later tests are this one in disguise—they simply pick a convenient series, usually geometric, to measure against.

(2) For (b), suppose instead $\sum a_n$ converged. Since $0 \le d_n \le a_n = |a_n|$, part (a) (with $c_n = a_n$) would force $\sum d_n$ to converge — contradicting that it diverges. Hence $\sum a_n$ diverges.

■

[A4] cf. R 3.25; bound the tail of $\sum a_n$ by that of $\sum c_n$.

The natural series to compare against is the geometric one. Comparing $\sum a_n$ to a geometric $\sum b_n$: if the ratio b_{n+1}/b_n is constant we are led to the *ratio test*, and if the root $\sqrt[n]{b_n}$ is constant, to the *root test*.

Theorem 6.6.9

Root test

Given $\sum a_n$, set $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. Then

- if $\alpha < 1$, $\sum a_n$ converges;
- if $\alpha > 1$, $\sum a_n$ diverges;
- if $\alpha = 1$, the test is inconclusive.

This is Rudin's root test (R 3.33).

Proof 6.6.9

Proof by Comparison with a Geometric Series of the Root Test.

(1) If $\alpha < 1$, choose β with $\alpha < \beta < 1$. Then $\sqrt[n]{|a_n|} < \beta$ for all large n, so $|a_n| < \beta^n$; as $\sum \beta^n$ converges, the comparison test (6.6.8) gives convergence.

(2) If $\alpha > 1$, then $\sqrt[n]{|a_n|} > 1$ for infinitely many n, so $|a_n| > 1$ infinitely often and $a_n \not\to 0$; by 6.6.3 the series diverges. For $\alpha = 1$ the test decides nothing: both $\sum \frac{1}{n}$ (divergent) and $\sum \frac{1}{n^2}$ (convergent) have $\alpha = 1$.

■

[A1] cf. R 3.33; uses the comparison test (6.6.8).

Theorem 6.6.10

Ratio test

Suppose $a_n \neq 0$ for all large n. Then $\sum a_n$ converges if $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$, and diverges if there is N with $\left|\frac{a_{n+1}}{a_n}\right| \ge 1$ for all $n \ge N$. This is Rudin's ratio test (R 3.34).

Proof 6.6.10

Proof by Comparison with a Geometric Series of the Ratio Test.

(1) If $\limsup |a_{n+1}/a_n| < 1$, choose β in between; then $|a_{n+1}/a_n| < \beta$ for $n \ge N$, so $|a_n| \le |a_N|\,\beta^{\,n-N}$ for $n \ge N$. Comparison with the geometric series (6.6.8) gives convergence.

(2) If $|a_{n+1}/a_n| \ge 1$ for all $n \ge N$, then $|a_n| \ge |a_N| > 0$ for $n \ge N$, so $a_n \not\to 0$ and the series diverges (6.6.3).

■

[A1] cf. R 3.34; uses the comparison test (6.6.8).

Definition 6.6.11

Power series and radius of convergence

A *power series* is a function defined by

$$f(x) = \sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots .$$

Fix $x \neq 0$ and apply the root test (6.6.9)[49] to $\sum c_n x^n$. With $\alpha := \limsup_{n\to\infty} \sqrt[n]{|c_n|}$,

$$\beta := \limsup_{n\to\infty} \sqrt[n]{|c_n x^n|} = |x|\,\alpha,$$

so the series converges when $\beta = |x|\,\alpha < 1$ and diverges when $|x|\,\alpha > 1$. Hence it converges for $|x| < R$, where

$$R := \frac{1}{\alpha} \qquad \big(R := +\infty \text{ if } \alpha = 0, \qquad R := 0 \text{ if } \alpha = +\infty\big).$$

R is the *radius of convergence*: the series converges on $(-R, R)$ and diverges for $|x| > R$; at $x = 0$ it is just $f(0) = c_0$, and the endpoints $|x| = R$ require separate analysis. This is Rudin's power-series setup and radius formula (R 3.38–R 3.39), restricted here to real x.

Example 6.6.12

The exponential series

For $f(x) = \sum_{n=0}^{\infty} \dfrac{x^n}{n!}$ the ratios have a limit, so they compute α (the valid case for the ratio shortcut):

$$\lim_{n\to\infty} \frac{1/(n+1)!}{1/n!} = \lim_{n\to\infty} \frac{1}{n+1} = 0 = \alpha,$$

so $R = 1/\alpha = +\infty$: the series converges for every x. This function is $f(x) = e^x$. Compare Rudin's examples following the radius-of-convergence theorem (R 3.40).

Remark 6.6.13

Analytic $\Rightarrow C^\infty$, but not conversely

A function defined by a power series on $(-R, R)$ is called *analytic* there. Such a function may be differentiated termwise infinitely often, and each derivative is again given by a power series with the same radius of convergence—so an analytic function is automatically C^∞, i.e. smooth. The converse, however, fails: there are functions that are infinitely differentiable everywhere yet are *not* analytic, because their Taylor series, though convergent, does not return the function. Being analytic is thus strictly stronger than being smooth. A standard such example—e^{-1/x^2} extended by $f(0) = 0$, whose derivatives at 0 all vanish—appears later (8.7.2). The termwise differentiation of power series will be proved when we treat uniform convergence.

[49] The root test is used here because it is strictly more general than the ratio test; for the precise relationship see the appendix, §6.A.1.

6.7 Absolute Convergence and Rearrangements

Adding two convergent series is harmless: $(\sum a_n) + (\sum b_n) = \sum(a_n + b_n)$. Multiplying or *reordering* the terms of a series is far more delicate and demands a stronger hypothesis than mere convergence. The right hypothesis is that the series of absolute values converge.

Definition 6.7.1

Absolute convergence

A series $\sum a_n$ *converges absolutely* if $\sum |a_n|$ converges (R 3.45). A series that converges but does not converge absolutely is called *conditionally convergent*.

Proposition 6.7.2

Absolute convergence implies convergence

If $\sum a_n$ converges absolutely, then $\sum a_n$ converges (R 3.45).[50]

Proof 6.7.2

Proof by the Cauchy Criterion that Absolute Convergence Implies Convergence.

(1) Let $\varepsilon > 0$. Since $\sum |a_n|$ converges, the Cauchy criterion (6.6.2) provides N with $\sum_{k=m}^{n} |a_k| < \varepsilon$ for all $n \geq m \geq N$. Then for such n, m,

$$\Big|\sum_{k=m}^{n} a_k\Big| \leq \sum_{k=m}^{n} |a_k| < \varepsilon.$$

(2) Thus $\sum a_n$ satisfies the Cauchy criterion (6.6.2), and therefore converges.

■

[A4] cf. R 3.45; the triangle inequality reduces the tail of $\sum a_n$ to that of $\sum |a_n|$; uses 6.6.2.

Definition 6.7.3

Rearrangement

Let $\{k_n\}$ be a sequence of positive integers in which every positive integer appears exactly once (a bijection of $\mathbb{N}$). Put $a'_n = a_{k_n}$. The series $\sum a'_n$ is called a *rearrangement* of $\sum a_n$ (R 3.52); it has exactly the same terms, reordered.

Theorem 6.7.4

Riemann's rearrangement theorem

Suppose $\sum a_n$ converges but does not converge absolutely. Let $-\infty \leq \alpha \leq \beta \leq +\infty$ be given. Then there is a rearrangement $\sum a'_n$, with partial sums s'_n, such that

$$\liminf_{n\to\infty} s'_n = \alpha \qquad \text{and} \qquad \limsup_{n\to\infty} s'_n = \beta.$$

[50] The converse is false. The alternating harmonic series $\sum (-1)^{n+1}/n$ converges, yet the series of absolute values is the harmonic series (6.6.4), which diverges. Absolute convergence is the genuinely stronger property, and—as the next theorem shows—the only one robust enough to survive reordering.

In particular, a conditionally convergent series can be rearranged to converge to *any* prescribed sum (take $\alpha = \beta$), or to diverge to $\pm\infty$. This is R 3.54.[51]

Proof 6.7.4

Proof Sketch by Alternately Overshooting β and Undershooting α.

(1) **The two parts each diverge.** Write $p_n = \max\{a_n, 0\}$ and $q_n = \max\{-a_n, 0\}$ for the positive and negative parts, so $a_n = p_n - q_n$ and $|a_n| = p_n + q_n$. If both $\sum p_n$ and $\sum q_n$ converged, then $\sum |a_n| = \sum(p_n + q_n)$ would converge, contradicting non-absolute convergence; if exactly one converged, then $\sum a_n = \sum(p_n - q_n)$ would diverge, contradicting convergence. Hence *both* $\sum p_n$ and $\sum q_n$ diverge to $+\infty$. Also $a_n \to 0$ (6.6.3), so $p_n \to 0$ and $q_n \to 0$.

(2) **Alternately overshoot and undershoot.** Take the positive terms in order and add just enough of them to push the partial sum strictly above β; this is possible because $\sum p_n = +\infty$. Then take negative terms in order and add just enough to pull the partial sum strictly below α; this is possible because $\sum q_n = +\infty$. Repeat indefinitely. Every term is used exactly once (each part is exhausted in order), so the result is a genuine rearrangement.

(3) **The overshoots vanish.** At each switch the partial sum crosses β (resp. α) by no more than the last term added, and since $p_n, q_n \to 0$ these overshoots tend to 0. Therefore the partial sums s'_n cluster down to α and up to β with no other limit points, giving $\liminf s'_n = \alpha$ and $\limsup s'_n = \beta$. When $\alpha = \beta$ (finite) this forces $s'_n \to \alpha$; when $\alpha = \beta = \pm\infty$ the same scheme drives $s'_n \to \pm\infty$.

■

[A9] sketch, following R 3.54; positive and negative parts each diverge while the terms vanish.

[51] This is the cautionary tale of the subject: for a merely conditionally convergent series the value of the sum is an artifact of the *order* of summation, not of the terms alone. Absolute convergence (6.7.1) is precisely what immunizes a series against this—every rearrangement of an absolutely convergent series converges, and to the same sum (R 3.55).

Appendix

Supplementary material—additional proofs, context, and other treatments; not part of the lecture proper.

Supplement 6.A.1

The root test subsumes the ratio test

For any sequence $\{c_n\}$ with $c_n \neq 0$,

$$\liminf_n \left|\tfrac{c_{n+1}}{c_n}\right| \;\le\; \liminf_n \sqrt[n]{|c_n|} \;\le\; \limsup_n \sqrt[n]{|c_n|} \;\le\; \limsup_n \left|\tfrac{c_{n+1}}{c_n}\right|$$

(R 3.37). Hence whenever the ratio test settles convergence the root test settles it too, but not conversely; and *if* $\lim_n |c_{n+1}/c_n| = L$ exists then the root quantity tends to L as well, so the ratio test gives the same radius $R = 1/L$ (the converse fails — the radii can agree without the ratio limit existing). This is why 6.6.11 fixes R through the root quantity $\alpha = \limsup_n \sqrt[n]{|c_n|}$.

LECTURE 7

Function Limits, Continuity, and the Topology of Continuous Maps

MATH S4061 | Introduction to Modern Analysis I

Thursday, 18 June 2026

Natura non facit saltus.

Gottfried Wilhelm Leibniz, New Essays IV.16

Contents of this lecture

Overview. This lecture begins Chapter 4 by replacing the sequential limit $p_n \to p$ with the function limit $\lim_{x \to p} f(x) = q$ in metric spaces. The ε–δ definition is tied back to sequences, then specialized to continuity at a point and continuity on a set. With the open-set characterization in hand, continuity becomes a topological operation: preimages of open sets are open, compositions are continuous, compact sets have compact images, connected sets have connected images, and the classical extreme-value and intermediate-value theorems become consequences. The lecture closes with homeomorphisms and the compactness theorem that upgrades pointwise continuity to uniform continuity. The companion text is Rudin, Chapter 4, especially R 4.1–R 4.26.

7.1 Function Limits

Theorem 7.1.1

Extreme values on a closed interval

If $f : [a,b] \to \mathbb{R}$ is continuous, then f attains a maximum and a minimum: there exist $x, y \in [a,b]$ such that

$$f(x) \le f(z) \le f(y) \qquad \text{for every } z \in [a,b].$$

This is the interval case of Rudin's extreme-value theorem (R 4.16); the proof is deferred until compactness of images is available in 7.5.2.

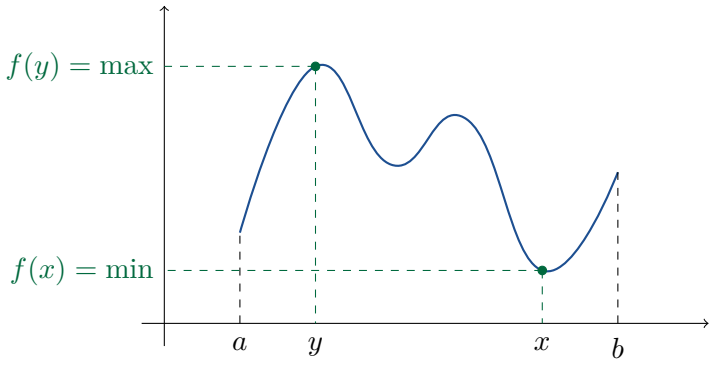


Figure 7.1.2. The calculus picture of extrema. A continuous real-valued function on a closed interval cannot miss its highest and lowest values. The rigorous proof will come from compactness of the image (7.5.2).

Remark 7.1.3

Sequences as functions

For a sequence $\{p_n\}$ in a metric space X, writing $p_n \to p$ means $\lim_{n\to\infty} p_n = p$ in the ε–N sense. A sequence is simply a function

$$f : \mathbb{N} \to X, \qquad n \mapsto p_n,$$

and asks what happens as n gets large (cf. R 2.7, R 3.1).

Definition 7.1.4

Function limit in metric spaces

Let (X, d_X) and (Y, d_Y) be metric spaces, let $E \subseteq X$, let p be a cluster point of E, and let $f : E \to Y$. We say that the *limit of f as $x \to p$ is q* (R 4.1) if for every $\varepsilon > 0$ there exists $\delta_{p,\varepsilon} > 0$ such that

$$0 < d_X(x,p) < \delta_{p,\varepsilon} \quad \Longrightarrow \quad d_Y\big(f(x), q\big) < \varepsilon.$$

We write

$$f(x) \to q \quad \text{as } x \to p, \qquad \text{or} \qquad \lim_{x \to p} f(x) = q.$$

The condition $0 < d_X(x,p)$ forces $x \neq p$; therefore $f(p)$ need not be defined, and even when it is defined its value does not affect this limit.[52]

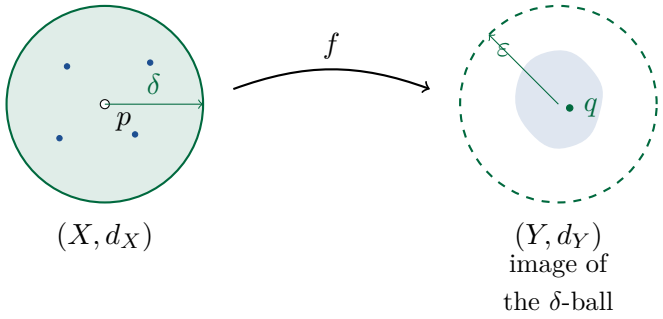


Figure 7.1.5. THE ε–δ PICTURE. The punctured δ-ball around p in the domain is carried inside the ε-ball around q in the target.

Proposition 7.1.6

Sequential criterion for function limits

Let (X, d_X) and (Y, d_Y) be metric spaces, let $E \subseteq X$, let $f : E \to Y$, and let p be a cluster point of E. Then

$$\lim_{x \to p} f(x) = q \quad \Longleftrightarrow \quad f(x_n) \to q \text{ for every sequence } \{x_n\} \subseteq E \setminus \{p\} \text{ with } x_n \to p.$$

This is Rudin's sequential characterization of function limits (R 4.2).

Proof 7.1.6

Proof by the ε–δ Definition and Its Contrapositive.

(1) Suppose $\lim_{x \to p} f(x) = q$ and let $\{x_n\} \subseteq E \setminus \{p\}$ satisfy $x_n \to p$. Given $\varepsilon > 0$, choose $\delta > 0$ such that

$$0 < d_X(x,p) < \delta \quad \Longrightarrow \quad d_Y(f(x), q) < \varepsilon.$$

Since $x_n \to p$, for all large n one has $d_X(x_n, p) < \delta$, and the punctured condition is automatic because $x_n \neq p$. Hence $d_Y(f(x_n), q) < \varepsilon$ eventually, so $f(x_n) \to q$.

(2) Conversely, prove the contrapositive. If the function limit fails, then for some $\varepsilon_0 > 0$ every $\delta > 0$ fails. Taking $\delta = 1/n$, choose $x_n \in E$ with

$$0 < d_X(x_n, p) < \frac{1}{n} \qquad \text{but} \qquad d_Y(f(x_n), q) \geq \varepsilon_0.$$

[52] A limit is a statement about the *approach* to p, never about the value at p. The punctured condition $0 < d_X(x,p)$ deliberately looks away from $x = p$, so $f(p)$ may be undefined, or defined "wrongly," without changing $\lim_{x \to p} f$. The gap between the value the approach predicts and the value actually assigned at p is exactly what a removable discontinuity (7.3.7) records.

Then $x_n \to p$, while $f(x_n)$ does not converge to q. This contradicts the sequential condition, so the function limit holds.

■

[A2+A11] cf. R 4.2; use $\delta_n = 1/n$ in the reverse direction.

The logic is: if A is the function-limit statement and B is the sequential statement, prove $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$.

Proposition 7.1.7

Arithmetic of function limits

Let $E \subseteq X$, let p be a cluster point of E, and let $f, g : E \to \mathbb{R}$ satisfy $\lim_{x\to p} f(x) = A$ and $\lim_{x\to p} g(x) = B$. Then

$$\lim_{x\to p}(f+g)(x) = A + B, \qquad \lim_{x\to p}(fg)(x) = AB, \qquad \lim_{x\to p}\frac{f}{g}(x) = \frac{A}{B} \quad (B \neq 0).$$

The same statements hold coordinatewise for $\mathbb{R}^k$-valued f, g, and for their inner product $f \cdot g$. This is the function-limit form of the algebraic limit laws (R 4.4); because limits of functions reduce to limits of sequences, the Chapter 3 laws carry over verbatim.

Proof 7.1.7

Proof by the Sequential Criterion and the Chapter 3 Limit Laws.

(1) Let $\{x_n\} \subseteq E \setminus \{p\}$ be any sequence with $x_n \to p$. By the sequential criterion 7.1.6, the hypotheses give

$$f(x_n) \to A, \qquad g(x_n) \to B.$$

(2) The algebraic limit laws for sequences (R 3.3) then apply termwise:

$$f(x_n) + g(x_n) \to A + B, \qquad f(x_n)\, g(x_n) \to AB, \qquad \frac{f(x_n)}{g(x_n)} \to \frac{A}{B} \ \ (B \neq 0),$$

where in the quotient case $B \neq 0$ keeps the terms eventually defined and bounded away from 0.

(3) Since these hold for *every* such sequence $x_n \to p$, the reverse direction of 7.1.6 converts each one back into the corresponding function limit. This proves the three identities.

(4) For $\mathbb{R}^k$-valued f, g the same argument runs in each coordinate, so the vector limits hold coordinatewise; and since $f \cdot g = \sum_{i=1}^{k} f_i\, g_i$ is built from the real products and sums just treated, $\lim_{x\to p}(f \cdot g)(x) = A \cdot B$ as well.

■

[A2] cf. R 4.4; uses 7.1.6 and the sequence limit laws R 3.3.

7.2 Continuity and Composition

Definition 7.2.1

Continuity at a point

Let $f : E \to Y$ and let $p \in E$. The function f is *continuous at* p (R 4.5) if

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

Equivalently, for every $\varepsilon > 0$ there exists $\delta_{\varepsilon,p} > 0$ such that

$$d_X(x,p) < \delta_{\varepsilon,p} \quad \Longrightarrow \quad d_Y\big(f(x), f(p)\big) < \varepsilon.$$

For continuity the ball around p is not punctured: the point $x = p$ is allowed, and the limit value is required to be exactly $f(p)$ (cf. R 4.6 when p is a cluster point).

Definition 7.2.2

Continuity on a set

A function $f : E \to Y$ is *continuous on* E if it is continuous at every $p \in E$ (R 4.5).

Remark 7.2.3

Isolated points

If p is an isolated point of E, then f is continuous at p vacuously: for a sufficiently small ball, there are no points of E near p except p itself (cf. R 4.5).

Proposition 7.2.4

Composition of continuous functions

If $f : X \to Y$ is continuous at x and $g : Y \to Z$ is continuous at $f(x)$, then $g \circ f : X \to Z$ is continuous at x. Consequently, the composition of continuous functions is continuous (R 4.7).

Proof 7.2.4

Proof by an ε–η–δ Chase.

(1) Let $\varepsilon > 0$. Since g is continuous at $f(x)$, there exists $\eta > 0$ such that

$$d_Y(y, f(x)) < \eta \quad \Longrightarrow \quad d_Z\big(g(y), g(f(x))\big) < \varepsilon.$$

(2) Since f is continuous at x, there exists $\delta > 0$ such that

$$d_X(u,x) < \delta \quad \Longrightarrow \quad d_Y\big(f(u), f(x)\big) < \eta.$$

(3) Combining the two implications, $d_X(u,x) < \delta$ implies

$$d_Z\big((g \circ f)(u), (g \circ f)(x)\big) < \varepsilon,$$

so $g \circ f$ is continuous at x. ∎

[A3] cf. R 4.7.

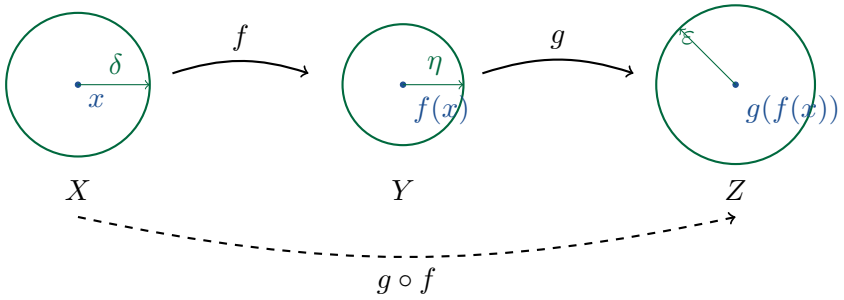


Figure 7.2.5. The composition chase. Continuity of g supplies the η-ball in Y; continuity of f supplies the δ-ball in X whose image lands inside it.

Remark 7.2.6

Sequential proof of composition

The sequential criterion gives the same result at a glance:

$$x_n \to x \quad \Longrightarrow \quad f(x_n) \to f(x) \quad \Longrightarrow \quad g(f(x_n)) \to g(f(x)).$$

This is the sequential version of R 4.7, using R 4.2.

7.3 Examples and Discontinuities

Example 7.3.1

Basic continuous real functions

The constant function $f(x) \equiv c$ is continuous: given $\varepsilon > 0$, every value of f already lies in $(c-\varepsilon, c+\varepsilon)$, so any $\delta > 0$ works. The identity $f(x) = x$ is continuous by taking $\delta = \varepsilon$. Building from constants and the identity through the algebra of limits and composition, every polynomial is continuous, and every rational function $p(x)/q(x)$ is continuous where $q(x) \neq 0$ (cf. R 4.9–R 4.11).

Proposition 7.3.2

The reciprocal is continuous on $(0, \infty)$

The function $f(t) = 1/t$ is continuous on $(0, \infty)$, but the required δ depends on both the point x and ε. This is a concrete ε–δ proof of the reciprocal case covered abstractly by R 4.9.

Proof 7.3.2

Direct ε–δ Proof for the Reciprocal.

(1) Fix $x > 0$ and $\varepsilon > 0$. If $|t - x| < \delta$, then

$$\left|\frac{1}{t} - \frac{1}{x}\right| = \frac{|t-x|}{tx}.$$

(2) First keep t away from 0. If $\delta \le x/2$, then $t > x/2$, so $tx > x^2/2$ and

$$\left|\frac{1}{t} - \frac{1}{x}\right| < \frac{\delta}{x^2/2} = \frac{2\delta}{x^2}.$$

(3) Therefore it is enough to require $\delta < \varepsilon x^2/2$. One may take

$$\delta(x, \varepsilon) = \min\left(\frac{x}{2}, \frac{\varepsilon x^2}{2}\right).$$

This proves continuity at x. The factor x^2 shows why δ shrinks as $x \to 0$. ■

[A3] cf. R 4.9; the point-dependence is the point of the example.

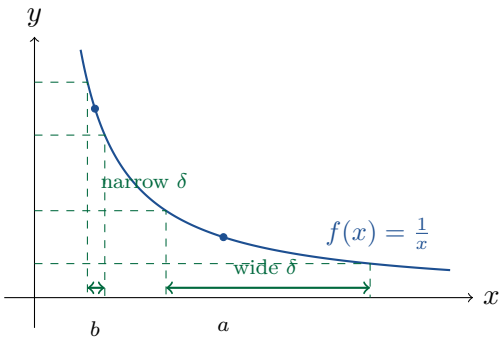


Figure 7.3.3. POINT-DEPENDENT δ FOR $1/x$. The same ε-band in the target pulls back to a narrow input interval where the graph is steep and a wider interval where the graph is flat.

Remark 7.3.4

Target and domain roles

The ε always measures closeness in the target Y; the δ always measures closeness in the domain X. In 7.3.2 the roles do not swap. Only the size of δ changes with the base point (cf. R 4.5–R 4.8).

Remark 7.3.5

Power series and familiar functions

Power series define continuous functions inside their radius of convergence. In particular,

$$e^x = \sum_{n \ge 0} \frac{x^n}{n!}$$

is continuous, and similarly $\sin x$, $\cos x$, $\tan x$, and the other familiar functions are continuous on their natural domains. Compare the radius-of-convergence discussion in 6.6.11 and Rudin's power-series radius theorem R 3.39.

Example 7.3.6

Standard discontinuities

The floor function $\lfloor x \rfloor$ is discontinuous at each integer. The piecewise function

$$f(x) = \begin{cases} e^x, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

is discontinuous at 0, since $\lim_{x \to 0^+} f(x) = 1$ while $\lim_{x \to 0^-} f(x) = 0$. Compare Rudin's catalogue of discontinuity examples in R 4.27.

Definition 7.3.7

Type I and Type II discontinuities

For a real function discontinuous at x_0, the one-sided limits classify the discontinuity.

- Type I: both one-sided limits exist, but either they differ (a jump) or they agree without matching $f(x_0)$ (a removable discontinuity).
- Type II: at least one one-sided limit fails to exist. This is often called an essential discontinuity.

This matches Rudin's first-kind/second-kind terminology (R 4.25–R 4.26).

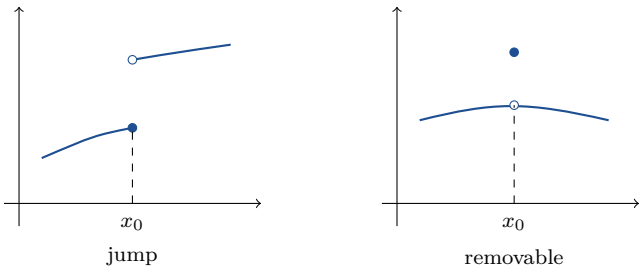


Figure 7.3.8. Type I discontinuities. A jump has different one-sided limits; a removable discontinuity has a hole at the limiting value and the function value placed elsewhere.

Example 7.3.9

The essential discontinuity of $\sin(1/x)$

The function

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

has no limit at 0. Near 0 the graph oscillates infinitely often between -1 and 1; for different sequences $x_n \to 0$, the values $f(x_n)$ can approach different limits (cf. R 4.27).

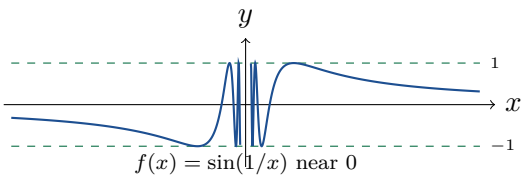


Figure 7.3.10. A TYPE II DISCONTINUITY. The graph of $\sin(1/x)$ oscillates indefinitely near 0, so no single limiting value can govern all approaches to 0.

Proposition 7.3.11

Projections and the norm are continuous

The coordinate projections and the norm on $\mathbb{R}^k$ are continuous; indeed each is Lipschitz, so $\delta = \varepsilon$ works at every point. Explicitly, for the i-th projection $\varphi_i : \mathbb{R}^k \to \mathbb{R}$, $\varphi_i(x_1, \dots, x_k) = x_i$, and for the norm $\|\cdot\| : \mathbb{R}^k \to \mathbb{R}$,

$$|\varphi_i(x) - \varphi_i(y)| \le \|x - y\|, \qquad \big|\, \|x\| - \|y\| \,\big| \le \|x - y\|.$$

These are the building blocks behind the continuity of polynomials and rational functions on $\mathbb{R}^k$ (R 4.10).

Proof 7.3.11

Proof by Two Lipschitz Estimates.

(1) For the projection, the i-th coordinate difference is dominated by the full Euclidean distance:

$$|\varphi_i(x) - \varphi_i(y)| = |x_i - y_i| \le \Big(\sum_{j=1}^{k} |x_j - y_j|^2\Big)^{1/2} = \|x - y\|.$$

Hence $\|x - y\| < \varepsilon$ forces $|\varphi_i(x) - \varphi_i(y)| < \varepsilon$, so $\delta = \varepsilon$ proves continuity of φ_i at every point.

(2) For the norm, the reverse triangle inequality gives

$$\big|\, \|x\| - \|y\| \,\big| \le \|x - y\|,$$

so again $\|x - y\| < \varepsilon$ forces $\big|\, \|x\| - \|y\| \,\big| < \varepsilon$, and $\delta = \varepsilon$ proves continuity of $\|\cdot\|$. ∎

[A3] cf. R 4.10; the second bound is the reverse triangle inequality.

Remark 7.3.12

Coordinatewise reduction of vector-valued continuity

A map $f : X \to \mathbb{R}^k$ is continuous if and only if each of its coordinate functions $f_i : X \to \mathbb{R}$ is continuous (R 4.10). One direction is immediate from 7.3.11: if f is continuous then $f_i = \varphi_i \circ f$ is a composition of continuous maps (7.2.4), hence continuous. Conversely, if every f_i is continuous then $\|f(x) - f(p)\|^2 = \sum_{i=1}^{k} |f_i(x) - f_i(p)|^2 \to 0$ as $x \to p$, so f is continuous. This reduces vector-valued continuity to the real-valued case.

7.4 Open-Set Characterizations

Proposition 7.4.1

Sequential characterization of continuity

A function $f : E \to Y$ is continuous at $p \in E$ if and only if for every sequence $\{x_n\} \subseteq E$ with $x_n \to p$, one has $f(x_n) \to f(p)$. This specializes Rudin's sequential limit criterion to continuity (cf. R 4.2, R 4.6).

Proof 7.4.1

Proof by Specializing the Limit Criterion.

(1) If f is continuous at p, apply 7.1.6 with $q = f(p)$; the non-punctured case causes no issue, since a sequence equal to p along a subsequence already has image equal to $f(p)$ along that subsequence.

(2) Conversely, if continuity fails at p, then for some $\varepsilon_0 > 0$ every $\delta > 0$ fails. With $\delta = 1/n$, choose $x_n \in E$ such that

$$d_X(x_n, p) < \frac{1}{n} \qquad \text{but} \qquad d_Y(f(x_n), f(p)) \geq \varepsilon_0.$$

Then $x_n \to p$, but $f(x_n)$ does not converge to $f(p)$.

■

[A2+A11] cf. R 4.2; the reverse direction again uses $\delta_n = 1/n$.

Proposition 7.4.2

Continuity at a point via neighbourhoods

The function $f : X \to Y$ is continuous at p if and only if for every open set $V \subseteq Y$ containing $f(p)$, there is an open set $U \subseteq X$ containing p such that

$$f(U) \subseteq V \qquad \text{equivalently} \qquad U \subseteq f^{-1}(V).$$

This is the pointwise neighbourhood form underlying Rudin's global open-set characterization (R 4.8).

Proof 7.4.2

Proof by Translating Balls into Open Neighbourhoods.

(1) If f is continuous at p and $V \subseteq Y$ is open with $f(p) \in V$, choose $\varepsilon > 0$ such that $B_\varepsilon(f(p)) \subseteq V$. Continuity gives $\delta > 0$ such that

$$f(B_\delta(p)) \subseteq B_\varepsilon(f(p)) \subseteq V.$$

Taking $U = B_\delta(p)$ gives the required neighbourhood.

(2) Conversely, take $V = B_\varepsilon(f(p))$. The neighbourhood condition gives an open $U \ni p$ with $f(U) \subseteq V$. Since U is open, some $B_\delta(p) \subseteq U$, and hence

$$f(B_\delta(p)) \subseteq B_\varepsilon(f(p)).$$

This is exactly the ε–δ condition for continuity at p.

■

[A3]

Proposition 7.4.3

Global open-set characterization of continuity

A function $f : X \to Y$ is continuous on X if and only if $f^{-1}(V)$ is open in X for every open set $V \subseteq Y$.[53] This is Rudin's open-set characterization of continuity (R 4.8).

Proof 7.4.3

Proof by Pulling Back Open Balls.

(1) Suppose f is continuous and let $V \subseteq Y$ be open. To show $f^{-1}(V)$ is open, take $x \in f^{-1}(V)$. Since V is open, there is $\varepsilon > 0$ with $B_\varepsilon(f(x)) \subseteq V$. Continuity at x gives $\delta > 0$ with

$$f(B_\delta(x)) \subseteq B_\varepsilon(f(x)) \subseteq V.$$

Hence $B_\delta(x) \subseteq f^{-1}(V)$, so every point of $f^{-1}(V)$ is interior.

(2) Conversely, fix $x \in X$ and $\varepsilon > 0$. The ball $B_\varepsilon(f(x))$ is open in Y, so $f^{-1}(B_\varepsilon(f(x)))$ is open in X by hypothesis. Since x lies in this preimage, there is $\delta > 0$ with

$$B_\delta(x) \subseteq f^{-1}(B_\varepsilon(f(x))).$$

Thus $f(B_\delta(x)) \subseteq B_\varepsilon(f(x))$, so f is continuous at x. Since x was arbitrary, f is continuous on X.

■

[Direct] cf. R 4.8.

Proposition 7.4.4

Composition via preimages

The composition of continuous functions is continuous. This is a second proof of R 4.7, using R 4.8.

Proof 7.4.4

Proof by the Open-Set Characterization.

(1) Let $X \xrightarrow{f} Y \xrightarrow{g} Z$ with f and g continuous, and let $V \subseteq Z$ be open.

(2) Since g is continuous, $g^{-1}(V)$ is open in Y. Since f is continuous, $f^{-1}(g^{-1}(V))$ is open in X.

[53] This is what continuity *really* is; the ε–δ definition is a local symptom of it. Continuity means the map never tears the space open—pull any open target set back and the domain set is still open. It has to be the *preimage*, not the image: preimages respect unions, intersections, and complements exactly, so they carry the open-set bookkeeping that images scramble. Read this way, continuity of compositions and the compact- and connected-image theorems drop out almost in one line (7.4.4, 7.5.1, 7.5.4).

(3) But preimage commutes with composition:

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

Thus the preimage under $g \circ f$ of every open set is open, so $g \circ f$ is continuous. ■

[Direct] a second proof of 7.2.4; cf. R 4.8.

This is the clean topological proof of composition: it works because preimages, not images, commute exactly with composition.

Remark 7.4.5

Pugh's moral

The sequential criterion is often the easiest way to check continuity at a glance: to test continuity at p, push every sequence $x_n \to p$ through the function and see whether $f(x_n) \to f(p)$ (cf. R 4.2).

7.5 Continuous Images of Compact and Connected Sets

Theorem 7.5.1

Continuous image of a compact set

If $f : X \to Y$ is continuous and X is compact, then the image

$$f(X) = \{y \in Y : \exists\, x \in X \text{ with } f(x) = y\}$$

is compact in Y (R 4.14).[54]

Proof 7.5.1

Proof by Pulling Back an Open Cover.

(1) Let $\{V_\alpha\}$ be an open cover of $f(X)$, with each V_α open in Y. Since f is continuous, each $f^{-1}(V_\alpha)$ is open in X.

(2) These preimages cover X:

$$X = \bigcup_\alpha f^{-1}(V_\alpha).$$

Indeed, if $x \in X$, then $f(x) \in f(X)$, so $f(x) \in V_\alpha$ for some α, and hence $x \in f^{-1}(V_\alpha)$.

[54] Compactness and continuity are made for each other. Compactness speaks in open covers; continuity turns a cover of the image into a cover of the domain by pulling each open set back (7.4.3), where a finite subcover is guaranteed—and pushing forward gives a finite subcover of the image. The most useful corollary drops out immediately: a continuous real function on a compact set attains its maximum and minimum (7.5.2).

(3) By compactness of X, choose finitely many indices $\alpha_1, \dots, \alpha_n$ such that

$$X = f^{-1}(V_{\alpha_1}) \cup \cdots \cup f^{-1}(V_{\alpha_n}).$$

Applying f, the sets $V_{\alpha_1}, \dots, V_{\alpha_n}$ cover $f(X)$. Thus every open cover of $f(X)$ has a finite subcover, so $f(X)$ is compact.

■

[A6] cf. R 4.14; uses 7.4.3.

Corollary 7.5.2

Extreme value theorem

If $f : [a, b] \to \mathbb{R}$ is continuous, then it attains a maximum and a minimum: there exist $m, M \in [a, b]$ such that

$$f(m) \le f(x) \le f(M) \qquad \text{for all } x \in [a, b].$$

This is Rudin's extreme-value theorem for real-valued continuous functions on compact domains (R 4.16).

Proof 7.5.2

Proof by Compactness of the Image.

(1) The interval $[a, b]$ is compact, so $f([a, b])$ is compact by 7.5.1. By Heine–Borel in $\mathbb{R}$, it is closed and bounded.

(2) Since $f([a, b])$ is bounded, the real numbers

$$\alpha = \inf f([a, b]), \qquad \beta = \sup f([a, b])$$

exist.

(3) Since $f([a, b])$ is closed, it contains its infimum and supremum. Thus $\alpha, \beta \in f([a, b])$, so there are $m, M \in [a, b]$ with

$$f(m) = \alpha, \qquad f(M) = \beta.$$

Therefore $f(m) \le f(x) \le f(M)$ for every $x \in [a, b]$.

■

[A6+A5] cf. R 4.16; uses 7.5.1 and Heine–Borel.

Definition 7.5.3

Connected metric space

A metric space (X, d) is *connected* if it cannot be written as $X = A \cup B$ where

$$A \neq \emptyset, \qquad B \neq \emptyset, \qquad A \cap B = \emptyset, \qquad A, B \text{ are open.}$$

Such a decomposition is called a *separation*. Equivalently, the only clopen subsets of X are $\emptyset$ and X itself.[55] In $\mathbb{R}$, the connected subsets are exactly the intervals (cf. R 2.45, R 2.47).

Theorem 7.5.4

Continuous image of a connected set

If X is connected and $f : X \to Y$ is continuous, then $f(X)$ is connected (R 4.22).

Proof 7.5.4

Proof by Pulling Back a Separation.

(1) Suppose for contradiction that $f(X) = A \cup B$ is a separation of $f(X)$: the sets A and B are nonempty, disjoint, and open in the subspace $f(X)$.

(2) Then $f^{-1}(A)$ and $f^{-1}(B)$ cover X, are disjoint, and are open in X by continuity. They are also nonempty: because $A, B \subseteq f(X)$ are nonempty, each contains some value $f(x)$.

(3) Thus $X = f^{-1}(A) \cup f^{-1}(B)$ is a separation of X, contradicting connectedness. Therefore $f(X)$ is connected.

■

[A9] cf. R 4.22; uses 7.4.3.

Corollary 7.5.5

Intermediate value theorem

Let $f : [a, b] \to \mathbb{R}$ be continuous and assume, without loss of generality, that $f(a) \leq f(b)$. Then for every $y \in [f(a), f(b)]$ there exists $x \in [a, b]$ with $f(x) = y$. This is the intermediate-value theorem (R 4.23).

Proof 7.5.5

Proof by Connectedness of the Image.

(1) The interval $[a, b]$ is connected, so $f([a, b])$ is connected by 7.5.4. Since connected subsets of $\mathbb{R}$ are intervals, $f([a, b])$ is an interval.

(2) This interval contains $f(a)$ and $f(b)$, and therefore contains every $y \in [f(a), f(b)]$. Hence for each such y there is $x \in [a, b]$ with $f(x) = y$.

■

[Direct] cf. R 4.23; compactness gives closed and bounded; connectedness gives an interval.

The structural mirror is useful: continuity preserves compactness, giving maxima and minima; continuity preserves connectedness, giving intermediate values.

[55] Read "connected" as "all one piece": you cannot break X into two nonempty open parts with nothing bridging them. The clopen rephrasing is the test that gets used—a set that is both open and closed has cleanly detached from its complement, so any such set beyond $\emptyset$ and X *is* a break. On the line this is just the statement that the connected sets are the intervals: anything with a gap splits at the gap.

7.6 Homeomorphisms

Two metric spaces are considered topologically the same when there is a bijection between them that preserves the open-set structure in both directions.

Definition 7.6.1

Homeomorphism

A *homeomorphism* is a continuous bijection $f : X \to Y$ whose inverse $f^{-1} : Y \to X$ is also continuous. If such an f exists, then X and Y are *homeomorphic*. Rudin uses the compact-domain theorem for continuous bijections in R 4.17.

Theorem 7.6.2

A continuous bijection from a compact domain is a homeomorphism

Let $f : X \to Y$ be continuous and bijective. If X is compact, then $f^{-1} : Y \to X$ is continuous, so f is a homeomorphism (R 4.17).

Proof 7.6.2

Proof by Showing the Inverse Pulls Back Opens to Opens.

(1) To prove f^{-1} is continuous, it is enough to show that for every open $V \subseteq X$, the preimage of V under f^{-1} is open in Y. Since f is a bijection,

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

(2) Let $V \subseteq X$ be open. Then V^c is closed in X, and since X is compact, V^c is compact.

(3) By 7.5.1, $f(V^c)$ is compact in Y, hence closed. Because f is a bijection,

$$f(V) = \big(f(V^c)\big)^c.$$

Therefore $f(V)$ is open in Y. Thus f^{-1} is continuous. ∎

[A6] cf. R 4.17; compactness makes continuous images of closed sets closed.

Counterexample 7.6.3

A continuous bijection need not be a homeomorphism

Let

$$\varphi : [0, 2\pi) \to S^1 \subseteq \mathbb{R}^2, \qquad \varphi(t) = (\cos t, \sin t).$$

This map is a continuous bijection onto the unit circle, but $[0, 2\pi)$ is not compact and φ^{-1} is not continuous: points on S^1 approaching $(1, 0)$ from the fourth quadrant have arguments tending to 2π, while $\varphi^{-1}(1, 0) = 0$ (cf. R 4.21).

FAILS: the claim that every continuous bijection is automatically a homeomorphism.

Remark 7.6.4

The compactness engine

The proof of 7.6.2 says that f^{-1} is continuous exactly when f carries open sets to open sets; equivalently, it is enough here that f carry closed sets to closed sets. On a compact domain this is automatic: closed subsets of X are compact, continuous images of compact sets are compact, and compact subsets of a metric space are closed (cf. R 4.17).

Remark 7.6.5

Recognising homeomorphic spaces

The basic homeomorphism questions are: given metric spaces M and N, are they homeomorphic, and what are the continuous maps $M \to N$? Typical comparison examples include the unit circle, a trefoil knot, the perimeter of a square, the 2-torus, $[0,1]$, and the unit disc. Compare the compact-domain homeomorphism theorem R 4.17 and circle example R 4.21.

7.7 Uniform Continuity

Definition 7.7.1

Uniform continuity

A function $f : X \to Y$ is *uniformly continuous* on X (R 4.18) if for every $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$d_X(p,q) < \delta_\varepsilon \quad \Longrightarrow \quad d_Y(f(p), f(q)) < \varepsilon.$$

The difference from ordinary continuity is that δ depends only on ε, not on the base point.[56]

Theorem 7.7.2

Compactness gives uniform continuity

If X is compact and $f : X \to Y$ is continuous, then f is uniformly continuous (R 4.19).

Proof 7.7.2

Proof by a Finite Half-Radius Cover.

(1) Let $\varepsilon > 0$. Since f is continuous at each $p \in X$, choose $\delta(p) > 0$ such that

$$d_X(p,q) < \delta(p) \quad \Longrightarrow \quad d_Y(f(p), f(q)) < \frac{\varepsilon}{2}.$$

(2) Define the half-radius ball

$$J(p) := B^X_{\delta(p)/2}(p).$$

The collection $\{J(p)\}_{p \in X}$ is an open cover of X, so compactness gives $p_1, \dots, p_N$ with

$$X \subseteq J(p_1) \cup \cdots \cup J(p_N).$$

[56] The whole content sits in the order of the quantifiers: a single δ must now work *everywhere at once*, whereas ordinary continuity may choose δ after the point and let it shrink from place to place. Picture sliding a fixed window of width δ along the graph—uniform continuity says one width holds the output swing under ε at every location. The reciprocal $1/x$ fails this near 0, where the graph steepens without bound and the usable δ collapses (7.7.3).

(3) Set
$$\delta = \frac{1}{2}\min(\delta(p_1), \dots, \delta(p_N)) > 0.$$
This δ depends only on ε.

(4) Take $x, y \in X$ with $d_X(x, y) < \delta$. Since the $J(p_i)$ cover X, choose i with $x \in J(p_i)$. Then
$$d_X(p_i, x) < \frac{1}{2}\delta(p_i),$$
so
$$d_Y(f(p_i), f(x)) < \frac{\varepsilon}{2}.$$

(5) Also,
$$d_X(y, p_i) \le d_X(y, x) + d_X(x, p_i) < \delta + \frac{1}{2}\delta(p_i) \le \delta(p_i),$$
so
$$d_Y(f(y), f(p_i)) < \frac{\varepsilon}{2}.$$

(6) By the triangle inequality in Y,
$$d_Y(f(y), f(x)) \le d_Y(f(y), f(p_i)) + d_Y(f(p_i), f(x)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$
Thus f is uniformly continuous.

■

[A6+A4] cf. R 4.19; the half-radius leaves room for nearby points.

Counterexample 7.7.3

The reciprocal is not uniformly continuous on $(0, \infty)$

The function $f(x) = 1/x$ is continuous on $(0, \infty)$, but not uniformly continuous there. In the explicit continuity estimate from 7.3.2,
$$\delta(x, \varepsilon) = \min\left(\frac{x}{2}, \frac{\varepsilon x^2}{2}\right) \to 0 \qquad \text{as } x \to 0.$$
Thus no single δ works for all $x > 0$. The domain $(0, \infty)$ is not compact, so 7.7.2 does not apply. On a compact interval $[a, b]$ with $a > 0$, the same function is uniformly continuous. Compare Rudin's noncompact-domain counterexamples in R 4.20.

FAILS: the claim that ordinary continuity on a noncompact domain implies uniform continuity.

Remark 7.7.4

Three compactness-for-free results

Compactness upgrades continuity in three recurring ways: a continuous image of a compact set is compact (7.5.1); a continuous bijection from a compact domain is a homeomorphism (7.6.2); and a continuous map from a compact domain is uniformly continuous (7.7.2). These are R 4.14, R 4.17, and R 4.19.

7.8 Monotonic Functions

For a real function f on an interval, the one-sided limits at an interior point x are written

$$f(x^+) := \lim_{t \to x^+} f(t), \qquad f(x^-) := \lim_{t \to x^-} f(t),$$

when these limits exist; equivalently $f(x^+) = \lim_n f(t_n)$ for every $t_n \to x$ with $t_n > x$, and $f(x^-) = \lim_n f(t_n)$ for every $t_n \to x$ with $t_n < x$ (cf. 7.3.7). The two-sided limit $\lim_{t \to x} f(t)$ exists exactly when $f(x^-)$ and $f(x^+)$ exist and agree. For *monotone* functions these one-sided limits always exist, which forces every discontinuity to be a jump and limits how many there can be (R 4.29–R 4.30).

Definition 7.8.1

Monotonic function

A real function $f : (a, b) \to \mathbb{R}$ is *monotonically increasing* if

$$x < y \quad \Longrightarrow \quad f(x) \le f(y),$$

and *monotonically decreasing* if

$$x < y \quad \Longrightarrow \quad f(x) \ge f(y).$$

A function that is one or the other is called *monotone* (R 4.28). Statements proved for increasing f transfer to decreasing f by replacing f with $-f$.

Proposition 7.8.2

One-sided limits of a monotone function

Let $f : (a, b) \to \mathbb{R}$ be monotonically increasing. Then $f(x^-)$ and $f(x^+)$ exist at every $x \in (a, b)$, and

$$f(x^-) = \sup_{a<t<x} f(t) \;\le\; f(x) \;\le\; \inf_{x<t<b} f(t) = f(x^+).$$

This is Rudin's existence theorem for one-sided limits of monotone functions (R 4.29).

Proof 7.8.2

Proof by a Supremum Argument.

(1) Consider the set $A = \{f(t) : a < t < x\}$. Since f is increasing, $f(t) \le f(x)$ for every $t < x$, so A is bounded above by $f(x)$. Hence

$$\alpha := \sup A$$

exists and $\alpha \le f(x)$.

(2) We show $f(x^-) = \alpha$. Let $\varepsilon > 0$. Since $\alpha - \varepsilon$ is not an upper bound of A, there is

$\delta > 0$ with $a < x - \delta < x$ and

$$\alpha - \varepsilon < f(x - \delta) \leq \alpha.$$

(3) For any $t \in (x - \delta, x)$ monotonicity gives $f(x - \delta) \leq f(t) \leq \alpha$, hence

$$\alpha - \varepsilon < f(t) \leq \alpha, \qquad \text{i.e.} \qquad |f(t) - \alpha| < \varepsilon.$$

Thus $f(x^-) = \alpha = \sup_{a<t<x} f(t)$.

(4) The argument for the right-hand limit is symmetric: $\{f(t) : x < t < b\}$ is bounded below by $f(x)$, its infimum β satisfies $\beta \geq f(x)$, and the same estimate gives $f(x^+) = \beta = \inf_{x<t<b} f(t)$. Combining the two yields $f(x^-) \leq f(x) \leq f(x^+)$.

■

[A5] cf. R 4.29; monotonicity turns the sup into the left-hand limit.

Corollary 7.8.3

Monotone functions have only first-kind discontinuities

If $f : (a, b) \to \mathbb{R}$ is monotone, then every discontinuity of f is of the first kind: both one-sided limits exist, and a discontinuity at x occurs precisely when $f(x^-) \neq f(x^+)$, a jump (R 4.29).

Proof 7.8.3

Proof by Existence of Both One-Sided Limits.

(1) By 7.8.2 (applied to f, or to $-f$ if f is decreasing), the limits $f(x^-)$ and $f(x^+)$ exist at every $x \in (a, b)$. By 7.3.7 this rules out a discontinuity of the second kind.

(2) If $f(x^-) = f(x^+)$ then, since $f(x^-) \leq f(x) \leq f(x^+)$, the common value equals $f(x)$ and f is continuous at x. Hence a discontinuity forces $f(x^-) \neq f(x^+)$, which is a jump.

■

[Direct] cf. R 4.29; immediate from 7.8.2.

Theorem 7.8.4

A monotone function has at most countably many discontinuities

If $f : (a, b) \to \mathbb{R}$ is monotonically increasing, then the set of points at which f is discontinuous is at most countable. The same holds for any monotone f (R 4.30).

Proof 7.8.4

Proof by Injecting a Rational into Each Jump.

(1) Let D be the set of discontinuities. By 7.8.3, each $x \in D$ is a jump, so $f(x^-) < f(x^+)$ and the open interval $(f(x^-), f(x^+))$ is nonempty. By density of the rationals, choose a rational $r(x) \in (f(x^-), f(x^+))$.

(2) The assignment $x \mapsto r(x)$ is injective. Indeed, if $x_1 < x_2$ are both in D, then for

any t with $x_1 < t < x_2$ monotonicity gives

$$f(x_1^+) = \inf_{s>x_1} f(s) \le f(t) \le \sup_{s<x_2} f(s) = f(x_2^-).$$

Hence the jump interval $\big(f(x_1^-), f(x_1^+)\big)$ lies entirely to the left of $\big(f(x_2^-), f(x_2^+)\big)$; the two are disjoint, so $r(x_1) < r(x_2)$.

(3) Thus $r : D \to \mathbb{Q}$ is an injection into a countable set, so D is at most countable. ■

[A5] cf. R 4.30; disjoint jump intervals inject into $\mathbb{Q}$.

Remark 7.8.5

Sharpness: a prescribed countable discontinuity set

The bound "at most countably many" is sharp: for *any* countable set $E = \{x_1, x_2, \dots\} \subseteq (a, b)$ there is a monotone function discontinuous exactly on E. Fix positive reals $c_n > 0$ with $\sum_n c_n$ convergent and set

$$f(x) = \sum_{x_n < x} c_n.$$

Then f is monotonically increasing, with jump $f(x_n^+) - f(x_n^-) = c_n > 0$ at each x_n (so f is discontinuous at every point of E), while f is continuous at every point of E^{c}. The convergence of $\sum c_n$ keeps f finite and bounded. Compare Rudin's construction of a function with prescribed discontinuities (R 4.31).

LECTURE 8

Derivatives, Taylor's Theorem, and Analyticity

MATH S4061 | Introduction to Modern Analysis I

Tuesday, 23 June 2026

And what are these fluxions? The velocities of evanescent increments.

George Berkeley, The Analyst (1734)

Contents of this lecture

Overview. This lecture begins the differential part of the course. The derivative is first defined by the difference quotient, then immediately reinterpreted as the best first-order linear approximation. That viewpoint makes the usual rules of differentiation bookkeeping in powers of the increment, and it also explains the multivariable definition. The lecture then turns the derivative into a global tool through Fermat's interior-extremum theorem, Rolle's theorem, the Mean Value Theorem, monotonicity, and Darboux's theorem. It closes with Taylor's theorem and the distinction between smooth and analytic functions, first over $\mathbb{R}$ and then over $\mathbb{C}$. The companion text is Rudin, Chapter 5, with the closest matches marked by the locators R 5.1–R 5.16.

8.1 Derivatives as First-Order Approximations

Definition 8.1.1

Derivative at a point

Let $f : (a, b) \to \mathbb{R}$ and let $x_0 \in (a, b)$. The function f is *differentiable at* x_0 (R 5.1) if the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

exists. This limit is the *derivative* of f at x_0, denoted $f'(x_0)$.[57]

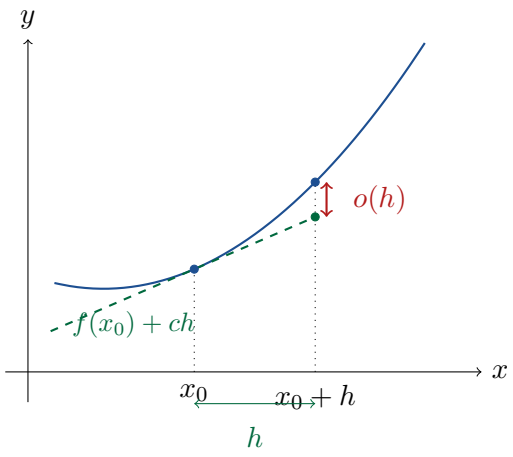


Figure 8.1.2. Best linear approximation. Near x_0, differentiability says the graph of f is shadowed by an affine line whose error is smaller than the displacement.

Proposition 8.1.3

Differentiability as a first-order expansion

For $f : (a, b) \to \mathbb{R}$ and $x_0 \in (a, b)$, the following are equivalent:

$$f \text{ is differentiable at } x_0 \qquad \Longleftrightarrow \qquad f(x_0 + h) = f(x_0) + ch + o(h)$$

for some $c \in \mathbb{R}$, where $o(h)$ means an error $E(h)$ satisfying

$$\frac{E(h)}{h} \longrightarrow 0 \qquad (h \to 0).$$

In that case $c = f'(x_0)$. This is Rudin's derivative definition R 5.1 written in the increment variable.

[57] The difference quotient is how you *compute* a derivative; the picture behind it is what to keep. Zoom in on the graph at x_0 and it looks more and more like a straight line, and $f'(x_0)$ is that line's slope—the limit of the chord slopes $(f(x) - f(x_0))/(x - x_0)$ as the far point slides in. 8.1.3 says this outright: to be differentiable is to be matched by *some* straight line with an error smaller than the step h. It is that reading, not the quotient, that survives into $\mathbb{R}^n$ (8.1.4), where one cannot divide by a vector.

Proof 8.1.3

Proof by Rearranging the Difference Quotient.

(1) Suppose f is differentiable at x_0. Define

$$E(h) = f(x_0 + h) - f(x_0) - f'(x_0)h.$$

Then the desired expansion holds with $c = f'(x_0)$, and for $h \neq 0$,

$$\frac{E(h)}{h} = \frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) \longrightarrow 0.$$

(2) Conversely, if $f(x_0 + h) = f(x_0) + ch + E(h)$ and $E(h)/h \to 0$, then

$$\frac{f(x_0 + h) - f(x_0)}{h} = c + \frac{E(h)}{h} \longrightarrow c.$$

Hence the difference quotient converges and $f'(x_0) = c$.

■

[Direct] increment form of R 5.1.

Definition 8.1.4

Differentiability from $\mathbb{R}^n$ to $\mathbb{R}^m$

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and let $x_0 \in \mathbb{R}^n$. The function F is *differentiable at* x_0 if there exists a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$F(x_0 + h) = F(x_0) + A(h) + E(h), \qquad \frac{\|E(h)\|}{\|h\|} \longrightarrow 0 \quad (h \to 0).$$

The linear map A is unique; in the standard bases it is represented by the Jacobian matrix. Compare Rudin's one-parameter vector-valued derivative in R 5.16; the present definition replaces one real increment by a vector increment.

Remark 8.1.5

Why the linear map replaces division

In one real variable, the quotient

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

is available because nonzero real increments have inverses. A vector increment $h \in \mathbb{R}^n$ has no comparable division operation, so the higher-dimensional definition subtracts a candidate linear map and asks that the remaining vector error be smaller than $\|h\|$. Complex differentiability is a special two-dimensional exception: identifying $\mathbb{R}^2$ with $\mathbb{C}$ restores division by a nonzero increment, but imposes a much stronger condition than ordinary real differentiability. Compare the vector-valued one-parameter derivative in R 5.16. More basically still, the difference quotient $[f(x) - f(x_0)]/(x - x_0)$ already presumes that the domain and target of f are one and the same complete field—for instance $\mathbb{R}$ or $\mathbb{C}$—so that subtraction and division are both available.

8.2 Rules for Derivatives

Theorem 8.2.1

Differentiability implies continuity

If f is differentiable at x_0, then f is continuous at x_0 (R 5.2).

Proof 8.2.1

Proof from the First-Order Expansion.

(1) By 8.1.3,
$$f(x_0+h) = f(x_0) + f'(x_0)h + E(h), \qquad \frac{E(h)}{h} \to 0.$$

(2) Since $E(h) = h(E(h)/h)$, we have $E(h) \to 0$. Therefore
$$f(x_0+h) - f(x_0) = f'(x_0)h + E(h) \longrightarrow 0,$$
which is continuity at x_0.

■

[Direct] cf. R 5.2.

Corollary 8.2.2

Discontinuity obstructs differentiability

If f is not continuous at x_0, then f is not differentiable at x_0. This is the contrapositive of Rudin's differentiability-implies-continuity theorem R 5.2.

Proof 8.2.2

Proof by Contrapositive.

(1) The contrapositive of 8.2.1 says precisely that differentiability at x_0 forces continuity at x_0.

(2) Therefore a failure of continuity at x_0 rules out differentiability there.

■

[A9]

Proposition 8.2.3

Algebra rules for derivatives

Suppose f and g are differentiable at x_0. Then
$$(f \pm g)'(x_0) = f'(x_0) \pm g'(x_0), \qquad (cf)'(x_0) = c\,f'(x_0),$$
$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$
If $g(x_0) \neq 0$, then f/g is differentiable at x_0 and
$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

These are Rudin's algebra rules for derivatives (R 5.3).

Proof 8.2.3

Proof by First-Order Bookkeeping.

(1) Write

$$f(x_0+h) = f(x_0) + f'(x_0)h + E_f(h), \qquad g(x_0+h) = g(x_0) + g'(x_0)h + E_g(h),$$

with $E_f(h)/h \to 0$ and $E_g(h)/h \to 0$.

(2) Adding the expansions gives

$$(f+g)(x_0+h) = (f+g)(x_0) + \big(f'(x_0)+g'(x_0)\big)h + \big(E_f(h)+E_g(h)\big),$$

and the last error is $o(h)$ because

$$\frac{E_f(h)+E_g(h)}{h} = \frac{E_f(h)}{h} + \frac{E_g(h)}{h} \longrightarrow 0.$$

The difference and scalar-multiple rules are the same calculation.

(3) Multiplying the expansions and collecting the constant and first-order terms gives

$$\begin{aligned}(fg)(x_0+h) = {} & f(x_0)g(x_0) + f'(x_0)g(x_0)h + f(x_0)g'(x_0)h \\ & + E_f(h)\big(g(x_0)+g'(x_0)h+E_g(h)\big) \\ & + E_g(h)\big(f(x_0)+f'(x_0)h\big) + f'(x_0)g'(x_0)h^2.\end{aligned}$$

After division by h, the last line tends to 0 because the error ratios tend to 0 and the bracketed factors remain bounded. Hence the coefficient of h is the derivative of fg.

(4) If $g(x_0) \neq 0$, then by 8.2.1 the values $g(x_0+h)$ stay nonzero for all sufficiently small h. Thus

$$\begin{aligned}\left(\frac{1}{g}\right)'(x_0) &= \lim_{h\to 0}\frac{1}{h}\left(\frac{1}{g(x_0+h)} - \frac{1}{g(x_0)}\right) \\ &= -\lim_{h\to 0}\frac{g(x_0+h)-g(x_0)}{h}\cdot\frac{1}{g(x_0+h)g(x_0)} = -\frac{g'(x_0)}{g(x_0)^2}.\end{aligned}$$

Applying the product rule to $f\cdot(1/g)$ gives the quotient rule.

■

[Direct] cf. R 5.3.

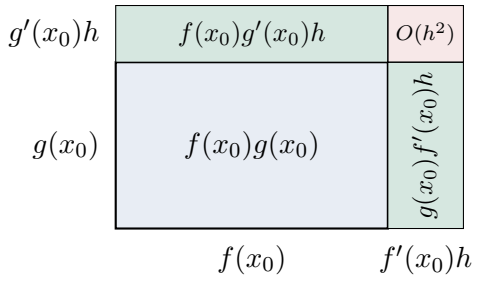


Figure 8.2.4. The product rule. The two first-order strips contribute to the derivative of the area; the corner is second-order.

Proposition 8.2.5

Chain rule

Suppose f is differentiable at x_0 and g is differentiable at $f(x_0)$. Then $g \circ f$ is differentiable at x_0 and

$$(g \circ f)'(x_0) = g'(f(x_0))f'(x_0).$$

This is the chain rule (R 5.5).

Proof 8.2.5

Proof by Composing First-Order Errors.

(1) Let

$$k(h) = f(x_0 + h) - f(x_0).$$

Differentiability of f gives $k(h) = f'(x_0)h + E_1(h)$, with $E_1(h)/h \to 0$; in particular $k(h) \to 0$.

(2) Differentiability of g at $f(x_0)$ means there is an error $E_2(u)$ such that

$$g(f(x_0) + u) = g(f(x_0)) + g'(f(x_0))u + E_2(u), \qquad \frac{E_2(u)}{u} \to 0.$$

To avoid dividing by a possibly zero displacement, define

$$H(u) = \begin{cases} E_2(u)/u, & u \neq 0, \\ 0, & u = 0. \end{cases}$$

Then $H(u) \to 0$ as $u \to 0$ and $E_2(u) = H(u)u$ for every u.

(3) Substituting $u = k(h)$ gives

$$\begin{aligned} g(f(x_0 + h)) &= g(f(x_0)) + g'(f(x_0))k(h) + H(k(h))k(h) \\ &= g(f(x_0)) + g'(f(x_0))f'(x_0)h + g'(f(x_0))E_1(h) + H(k(h))k(h). \end{aligned}$$

(4) The remainder divided by h tends to 0:

$$g'(f(x_0))\frac{E_1(h)}{h} + H(k(h))\frac{k(h)}{h} \longrightarrow 0 + 0 \cdot f'(x_0) = 0.$$

Therefore the coefficient of h is the derivative of $g \circ f$ at x_0. ■

[Direct] cf. R 5.5.

Proposition 8.2.6

Power rule

For every integer $n \geq 0$,

$$\frac{d}{dx}x^n = nx^{n-1},$$

where the case $n = 0$ is read separately as the derivative of the constant function 1, namely 0. R 5.4

Proof 8.2.6

Proof by Induction Using the Product Rule.

(1) For $n = 0$, the function $x^0 = 1$ is constant, so its derivative is 0.

(2) For $n = 1$, the difference quotient of x is 1.

(3) Assume $(x^n)' = nx^{n-1}$. Since $x^{n+1} = x \cdot x^n$, the product rule gives

$$(x^{n+1})' = (x)'x^n + x(x^n)' = x^n + nx^n = (n+1)x^n.$$

The formula follows by induction. ■

[A10]

8.3 Oscillation Examples

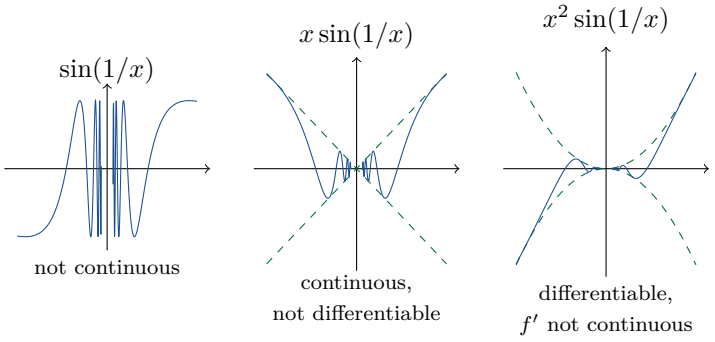


Figure 8.3.1. Three oscillatory functions. Multiplying by powers of x progressively tames the oscillation of $\sin(1/x)$ near 0.

Example 8.3.2

$\sin(1/x)$ is not continuous at 0

Define $f(0) = 0$ and $f(x) = \sin(1/x)$ for $x \neq 0$. Along sequences approaching 0, the values can approach different limits; for example one may choose points where $\sin(1/x) = 1$ and others where $\sin(1/x) = -1$. Thus f is not continuous at 0, and by 8.2.2 it is not differentiable there (cf. R 5.6).

Example 8.3.3

$x\sin(1/x)$ is continuous but not differentiable at 0

Let $f(0) = 0$ and $f(x) = x\sin(1/x)$ for $x \neq 0$. Since

$$|x\sin(1/x)| \leq |x|,$$

the squeeze theorem gives $f(x) \to 0 = f(0)$. But the difference quotient at 0 is

$$\frac{f(h) - f(0)}{h} = \sin(1/h),$$

which has no limit. Hence f is continuous at 0 but not differentiable there (R 5.6).

Example 8.3.4

$x^2\sin(1/x)$ is differentiable but not C^1 at 0

Let $f(0) = 0$ and $f(x) = x^2\sin(1/x)$ for $x \neq 0$. Then

$$\frac{f(h) - f(0)}{h} = h\sin(1/h) \longrightarrow 0,$$

so $f'(0) = 0$. For $x \neq 0$, the product and chain rules give

$$f'(x) = 2x\sin(1/x) - \cos(1/x).$$

The first term tends to 0, but $\cos(1/x)$ has no limit as $x \to 0$. Thus f is differentiable at 0, yet f' is not continuous at 0 (R 5.6).

Remark 8.3.5

The hierarchy already separates

The three examples show that differentiability is stronger than continuity, and continuous differentiability is stronger than differentiability. Each extra factor of x narrows the envelope around the oscillation and buys one more degree of regularity at the origin (cf. R 5.6).

8.4 Extrema and the Mean Value Theorem

Definition 8.4.1

Local extrema

Let $f : E \to \mathbb{R}$ and let $p \in E$. The function f has a *local maximum* at p if there exists $\delta > 0$ such that

$$q \in E, \quad |q - p| < \delta \quad \Longrightarrow \quad f(q) \leq f(p).$$

A *local minimum* is defined by reversing the inequality (R 5.7).

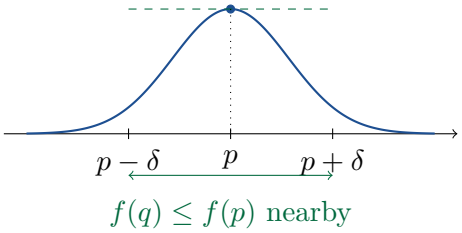


Figure 8.4.2. A LOCAL MAXIMUM. On a sufficiently small window around p, the value $f(p)$ dominates nearby values of the function.

Theorem 8.4.3

Interior extremum theorem

Let $f : (a, b) \to \mathbb{R}$ be differentiable. If $x_0 \in (a, b)$ is a local maximum or a local minimum, then

$$f'(x_0) = 0.$$

This is Fermat's interior-extremum theorem (R 5.8).

Proof 8.4.3

Proof by One-Sided Difference Quotients.

(1) Assume x_0 is a local maximum. For x sufficiently close to x_0,

$$f(x) - f(x_0) \leq 0.$$

(2) If $x < x_0$, then $x - x_0 < 0$, so

$$\frac{f(x) - f(x_0)}{x - x_0} \geq 0.$$

Passing to the left-hand limit gives $f'(x_0) \geq 0$.

(3) If $x > x_0$, then $x - x_0 > 0$, so

$$\frac{f(x) - f(x_0)}{x - x_0} \leq 0.$$

Passing to the right-hand limit gives $f'(x_0) \leq 0$.

(4) The two one-sided limits are equal because $f'(x_0)$ exists. Therefore

$$0 \leq f'(x_0) \leq 0,$$

and $f'(x_0) = 0$. The minimum case is identical with the inequalities reversed. ■

[A12] cf. R 5.8.

Remark 8.4.4

Why the hypotheses matter

The point must be interior because the proof uses both sides. At an endpoint only one one-sided inequality is available. The converse also fails: $f(x) = x^3$ has $f'(0) = 0$, but 0 is not a local extremum (cf. R 5.8).

Proposition 8.4.5

Rolle's theorem

Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there exists $c \in (a, b)$ such that

$$f'(c) = 0.$$

This is the Rolle special case behind Rudin's generalized mean-value theorem (R 5.9).

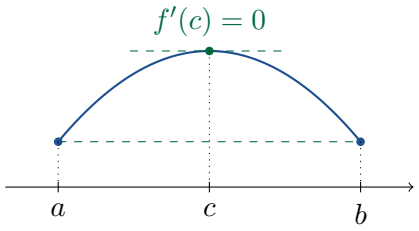


Figure 8.4.6. Rolle's theorem. Equal endpoint heights force a horizontal tangent at some interior point, unless the function is constant.

Proof 8.4.5

Proof by the Extreme-Value Theorem and Interior Extrema.

(1) Since $[a, b]$ is compact and f is continuous, f attains a maximum and a minimum on $[a, b]$ by the extreme-value theorem from Lecture 7.

(2) If either extreme is attained at an interior point $c \in (a, b)$, then 8.4.3 gives $f'(c) = 0$.

(3) If no extreme is attained in the interior, then both the maximum and minimum occur at endpoints. Since $f(a) = f(b)$, the maximum and minimum values coincide. Hence f is constant on $[a, b]$, and any $c \in (a, b)$ satisfies $f'(c) = 0$. ■

[A6+A12] Rolle step behind R 5.9.

Corollary 8.4.7

Mean Value Theorem

Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that

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

This is the Mean Value Theorem (R 5.10).[58]

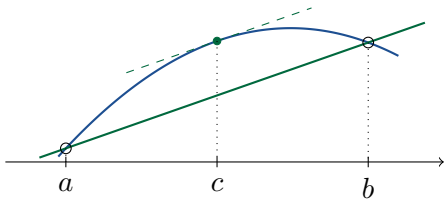


Figure 8.4.8. MEAN VALUE THEOREM. Some interior tangent is parallel to the chord joining the endpoints.

Proof 8.4.7

Proof by Tilting the Chord Flat.

(1) Let L be the secant line

$$L(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a),$$

and define $h = f - L$.

(2) The function h is continuous on $[a, b]$ and differentiable on (a, b). Also

$$h(a) = 0, \qquad h(b) = 0.$$

(3) By Rolle's theorem, there exists $c \in (a, b)$ with $h'(c) = 0$. Since

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

the conclusion follows.

■

[Constr] cf. R 5.10.

[58] This theorem looks modest and carries most of the chapter. Read it as a bridge from *local* to *global*: it turns information about f' at single points into control over how much f can change across a whole interval. Nearly every later result passes through it—the monotonicity test 8.4.9 is the first (the *sign* of f' alone decides whether f rises or falls), and "$f' \equiv 0 \Rightarrow f$ constant" is the same move. The figure above is the whole content: the average slope over $[a, b]$, the chord, is genuinely attained as a tangent slope at some interior c.

Corollary 8.4.9

Monotonicity test

Let f be differentiable on (a, b).

1. If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is nondecreasing.
2. If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant.
3. If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is nonincreasing.

These are the standard monotonicity consequences of the Mean Value Theorem (R 5.11).

Proof 8.4.9

Proof by Applying the Mean Value Theorem on Subintervals.

(1) Fix $x_1 < x_2$ in (a, b). By 8.4.7, there exists $\xi \in (x_1, x_2)$ such that

$$f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1).$$

(2) Since $x_2 - x_1 > 0$, the sign of $f(x_2) - f(x_1)$ is the sign of $f'(\xi)$. The three conclusions follow immediately from the three sign hypotheses.

■

[Direct] cf. R 5.11.

8.5 Darboux's Theorem

Proposition 8.5.1

Darboux's theorem

Let f be differentiable on an interval containing $a < b$. If λ lies strictly between $f'(a)$ and $f'(b)$, then there exists $c \in (a, b)$ such that

$$f'(c) = \lambda.$$

Thus derivatives have the intermediate value property, even though derivatives need not be continuous (R 5.12).

Proof 8.5.1

Proof by Tilting and Taking an Interior Extremum.

(1) Assume $f'(a) < \lambda < f'(b)$; the reverse case is obtained by changing signs. Define

$$g(x) = f(x) - \lambda x.$$

Then $g'(a) < 0$ and $g'(b) > 0$.

(2) Since g is continuous on $[a, b]$, it attains a minimum at some point $c \in [a, b]$.

(3) The condition $g'(a) < 0$ means that just to the right of a the values of g are smaller than $g(a)$, so the minimum is not attained at a. Similarly, $g'(b) > 0$ implies that just to the left of b the values of g are smaller than $g(b)$, so the minimum is not attained at b.

(4) Hence the minimum is attained at an interior point $c \in (a, b)$. By 8.4.3, $g'(c) = 0$, and therefore

$$f'(c) - \lambda = 0.$$

■

[A12+Constr] cf. R 5.12.

Remark 8.5.2

What Darboux rules out

Darboux's theorem forbids derivatives from having jump discontinuities, because a jump would skip the intermediate values. The theorem does not say that derivatives are continuous: 8.3.4 gives a derivative with an oscillatory discontinuity at 0 (cf. R 5.12).

8.6 Taylor Polynomials and Taylor's Theorem

Definition 8.6.1

Taylor polynomial

Let $n \geq 1$, and suppose f has derivatives through order $n-1$ at α. The *Taylor polynomial of degree at most $n-1$ centered at α* is

$$P_{n-1}(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(x-\alpha)^k, \qquad f^{(0)} = f.$$

Rudin introduces higher derivatives in R 5.14 and this polynomial in R 5.15.

Proposition 8.6.2

Matching and uniqueness

The Taylor polynomial P_{n-1} is the unique polynomial of degree at most $n-1$ satisfying

$$P_{n-1}^{(k)}(\alpha) = f^{(k)}(\alpha), \qquad k = 0, 1, \ldots, n-1.$$

This is the matching property built into Rudin's Taylor polynomial in R 5.15.

Proof 8.6.2

Proof in the Shifted Polynomial Basis.

(1) Write a general degree-at-most $n-1$ polynomial in the shifted basis:

$$Q(x) = \sum_{k=0}^{n-1} c_k (x-\alpha)^k.$$

(2) Differentiating k times and evaluating at α leaves only the k-th shifted monomial:

$$Q^{(k)}(\alpha) = k!\, c_k.$$

(3) The matching conditions therefore force

$$c_k = \frac{f^{(k)}(\alpha)}{k!},$$

exactly the coefficients of P_{n-1}. Hence the matching polynomial exists and is unique.

■

[Direct]

Theorem 8.6.3

Taylor's theorem with Lagrange remainder

Let $n \geq 1$. Suppose $f : [a, b] \to \mathbb{R}$ has $f^{(n-1)}$ continuous on $[a, b]$ and $f^{(n)}$ exists on (a, b). If $\alpha, \beta \in [a, b]$ are distinct and P_{n-1} is the Taylor polynomial centered at α, then there is a point ξ strictly between α and β such that

$$f(\beta) = P_{n-1}(\beta) + \frac{f^{(n)}(\xi)}{n!}(\beta - \alpha)^n.$$

This is Rudin's Taylor theorem with Lagrange remainder (R 5.15).[59]

Proof 8.6.3

Proof by Iterated Rolle's Theorem.

(1) Assume $\alpha < \beta$; the other order is identical on the interval between them. Choose M by

$$M = \frac{f(\beta) - P_{n-1}(\beta)}{(\beta - \alpha)^n},$$

so that $f(\beta) = P_{n-1}(\beta) + M(\beta - \alpha)^n$.

(2) Define

$$g(x) = f(x) - P_{n-1}(x) - M(x - \alpha)^n.$$

By the matching property in 8.6.2,

$$g(\alpha) = g'(\alpha) = \cdots = g^{(n-1)}(\alpha) = 0,$$

[59] This extends the lecture's opening idea. In 8.1.3 a derivative let us stand in for f near α by a *line*—a degree-one polynomial—with error $o(h)$; here f is stood in for by a degree-$(n-1)$ *polynomial*, the one matching f and its first $n-1$ derivatives at α (8.6.2), and the remainder names exactly what that polynomial misses. More matched derivatives, a closer fit near α. From here the remainder is the whole question: whether it shrinks as the degree grows is what decides whether the polynomials rebuild f (8.7.1).

and by the choice of M, $g(\beta) = 0$.

(3) Apply Rolle's theorem repeatedly. First, $g(\alpha) = g(\beta) = 0$ gives a point $c_1 \in (\alpha, \beta)$ with $g'(c_1) = 0$. Then $g'(\alpha) = g'(c_1) = 0$ gives a point $c_2 \in (\alpha, c_1)$ with $g''(c_2) = 0$. Continuing n times gives a point $\xi \in (\alpha, \beta)$ with

$$g^{(n)}(\xi) = 0.$$

(4) Since $P_{n-1}^{(n)} = 0$ and $\frac{d^n}{dx^n} M(x-\alpha)^n = n!M$,

$$0 = g^{(n)}(\xi) = f^{(n)}(\xi) - n!M.$$

Thus $M = f^{(n)}(\xi)/n!$, and substituting this into the definition of M gives the stated formula.

■

[A6+Constr] cf. R 5.15.

Remark 8.6.4

The qualitative little-o form

The Lagrange formula gives the quantitative error. If, in addition, $f^{(n)}$ is locally bounded near α–for instance if $f^{(n)}$ is continuous near α–then

$$f(t) - P_{n-1}(t) = O\big((t-\alpha)^n\big),$$

and hence

$$\frac{f(t) - P_{n-1}(t)}{(t-\alpha)^{n-1}} \longrightarrow 0.$$

The boundedness condition is doing real work here; the existence of $f^{(n)}$ alone does not by itself provide local boundedness. This is the qualitative consequence of the Lagrange remainder in R 5.15.

8.7 Real and Complex Analyticity

Definition 8.7.1

Real analyticity at a point

A smooth real function f is *real analytic at* α if its Taylor series at α converges to f on some neighbourhood of α:

$$f(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\alpha)}{k!}(t-\alpha)^k$$

for all t sufficiently close to α. This goes beyond Rudin's finite Taylor theorem R 5.15: one must also control the limiting behaviour of the remainders as the degree tends to infinity.

Counterexample 8.7.2

A smooth function need not be analytic

Define

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

This function is C^∞, and all derivatives at 0 vanish.[60] Its Taylor series at 0 is therefore identically 0, but $f(x) > 0$ for $x \neq 0$. Hence smoothness does not imply real analyticity. This is beyond Rudin's finite Taylor theorem R 5.15.

FAILS: the claim that knowing all derivatives at one point determines a smooth real function.

Theorem 8.7.3

Weierstrass Function

There exists a continuous function $W : \mathbb{R} \to \mathbb{R}$ that is differentiable at no point. This is an advanced separation result beyond Rudin Chapter 5.

Proof 8.7.3

Quoted Advanced Result.

↪ This theorem is stated as part of the regularity landscape; its proof is beyond the lecture.

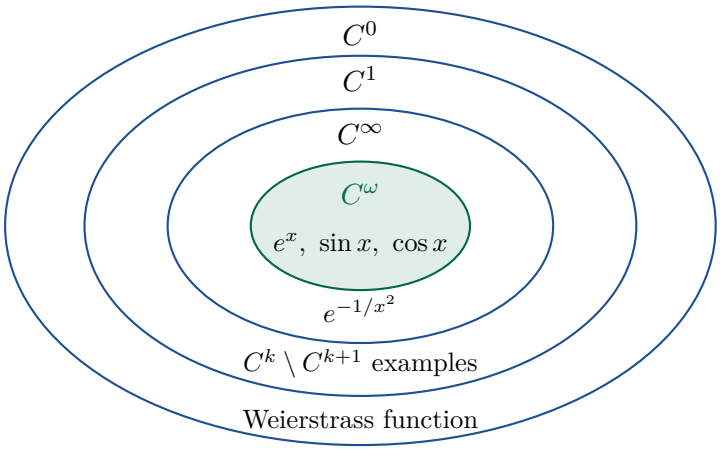


Figure 8.7.4. The real regularity tower. For real functions the inclusions between regularity classes are strict; each rung contains functions not lying on the next rung.

Remark 8.7.5

Derivative growth and Taylor convergence

Taylor convergence is controlled by the growth of the coefficients $f^{(r)}(\alpha)/r!$. If these coefficients grow too quickly, the formal Taylor series can have radius of convergence 0. A useful sufficient condition in

[60] It is tempting to assume that a function and *all* its derivatives at one point must determine it. Over $\mathbb{R}$ they do not, and this is the example to remember. Every derivative of f at 0 is zero, so its Taylor series at 0 is the zero series—which rebuilds the zero function, not f. The series converges perfectly well; it just converges to the *wrong* function. So for a smooth real function the Taylor series is a *hope*, not a promise; "analytic" (8.7.1) is the word for the functions that keep it, and most smooth ones do not.

the other direction is uniform control of derivatives on a neighbourhood: if the Lagrange remainders satisfy a bound forcing

$$\left|\frac{f^{(n)}(\xi)}{n!}(t-\alpha)^n\right| \longrightarrow 0,$$

then the Taylor polynomials converge back to f near α. Compare the finite Lagrange-remainder control in R 5.15.

Theorem 8.7.6

Holomorphic functions are analytic

Let $U \subseteq \mathbb{C}$ be open. If $f : U \to \mathbb{C}$ is complex differentiable at every point of U, then f is complex analytic: near each point of U, it is represented by a convergent power series. This complex-analysis theorem is beyond Rudin Chapter 5.[61]

Proof 8.7.6

Quoted Theorem from Complex Analysis.

↪ The result is recorded to contrast real and complex differentiability; its proof belongs to complex analysis.

Remark 8.7.7

The real tragedy and the complex collapse

Over $\mathbb{R}$, the classes $C^1, C^2, \ldots, C^\infty, C^\omega$ do not collapse. Over $\mathbb{C}$, the hypothesis is not ordinary real differentiability of a map $\mathbb{R}^2 \to \mathbb{R}^2$; it is complex differentiability on an open set, a much more rigid condition. Under that holomorphic hypothesis, one derivative already forces local power-series representation. This is the contrast with Rudin's finite Taylor theorem R 5.15 and beyond the real-variable scope of Chapter 5.

8.8 The Cauchy Mean Value Theorem and L'Hôpital's Rule

Theorem 8.8.1

Cauchy Mean Value Theorem

Let $f, g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $t \in (a, b)$ such that

$$[f(b) - f(a)]g'(t) = [g(b) - g(a)]f'(t).$$

If in addition g' is nonzero on (a, b) and $g(b) \neq g(a)$, this is the symmetric ratio form

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(t)}{g'(t)}.$$

[61] Why is $\mathbb{C}$ so much more rigid than $\mathbb{R}$? Complex differentiability demands far more than it appears to. The quotient $(f(z) - f(z_0))/(z - z_0)$ must approach the *same* limit no matter which direction z comes in from in the plane—a heavy constraint with no counterpart on the line, where a point has only a left and a right. Imposed at every point of U, that one requirement already forces a convergent power series, and the strict real tower $C^1 \supsetneq C^2 \supsetneq \cdots \supsetneq C^\omega$ of 8.7.4 collapses to a single floor.

Taking $g(x) = x$ recovers the ordinary Mean Value Theorem 8.4.7; this is the generalized mean value theorem behind Rudin (R 5.9).[62]

Proof 8.8.1

Proof by Rolle Applied to a Combined Function.

(1) Define $h : [a, b] \to \mathbb{R}$ by

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

As a linear combination of f and g, the function h is continuous on $[a, b]$ and differentiable on (a, b).

(2) Evaluate at the endpoints. Both reduce to the same value:

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

Hence $h(a) = h(b)$.

(3) By Rolle's theorem 8.4.5, there exists $t \in (a, b)$ with $h'(t) = 0$. Since

$$h'(x) = [f(b) - f(a)]g'(x) - [g(b) - g(a)]f'(x),$$

this gives $[f(b) - f(a)]g'(t) = [g(b) - g(a)]f'(t)$, which is the claim. When g' is nonzero on (a, b), the Mean Value Theorem applied to g forces $g(b) \neq g(a)$, and dividing by $[g(b) - g(a)]g'(t) \neq 0$ produces the ratio form.

■

[A6+A12] Rolle step 8.4.5; cf. R 5.9.

Theorem 8.8.2

L'Hôpital's Rule, the 0/0 case

Let $-\infty \leq a < b \leq +\infty$, and suppose f and g are differentiable on (a, b) with $g'(x) \neq 0$ for all $x \in (a, b)$. Assume

$$f(x) \to 0 \quad \text{and} \quad g(x) \to 0 \qquad (x \to a^+),$$

and that, for some $A \in \mathbb{R}$,

$$\frac{f'(x)}{g'(x)} \longrightarrow A \qquad (x \to a^+).$$

[62] The ordinary Mean Value Theorem compares f against the clock x; the Cauchy version compares f against *any* second travelling quantity g, so it reads slopes "per unit g" rather than "per unit x." That is exactly the leverage L'Hôpital's rule 8.8.2 needs to compare two functions that are both vanishing at the same point: one cannot use the clock, because the destination is $0/0$, so one races f directly against g.

Then

$$\frac{f(x)}{g(x)} \longrightarrow A \qquad (x \to a^+).$$

This is L'Hôpital's rule for the indeterminate form $0/0$ (R 5.13).[63]

Proof 8.8.2

Proof by the Cauchy Mean Value Theorem and Pinching.

(1) We treat the case $A \neq \pm\infty$. Fix any real number $r > A$. Since $f'(x)/g'(x) \to A$ as $x \to a^+$, there exists $c \in (a, b)$ such that

$$\frac{f'(x)}{g'(x)} < r \qquad \text{for all } x \in (a, c).$$

(2) Let $x, y \in (a, c)$ with $x < y$. The Cauchy Mean Value Theorem 8.8.1, applied to f and g on $[x, y]$, supplies a point $t \in (x, y) \subseteq (a, c)$ with

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(t)}{g'(t)} < r.$$

(The denominator $g(x) - g(y)$ is nonzero because g' does not vanish on (a, b), by the Mean Value Theorem.)

(3) Hold $y \in (a, c)$ fixed and let $x \to a^+$. Then $f(x) \to 0$ and $g(x) \to 0$, so the left-hand side tends to $\dfrac{-f(y)}{-g(y)} = \dfrac{f(y)}{g(y)}$, and the strict inequality passes to the limit as

$$\frac{f(y)}{g(y)} \leq r \qquad \text{for all } y \in (a, c).$$

(4) The number $r > A$ was arbitrary, so $\limsup_{y \to a^+} f(y)/g(y) \leq A$. A symmetric argument with any $q < A$ produces some $d \in (a, b)$ with

$$\frac{f(y)}{g(y)} \geq q \qquad \text{for all } y \in (a, d),$$

whence $\liminf_{y \to a^+} f(y)/g(y) \geq A$. The two bounds squeeze the ratio, giving $f(y)/g(y) \to A$ as $y \to a^+$.

■

[Constr] uses 8.8.1; cf. R 5.13.

8.9 Differentiation of Vector-Valued Functions

Definition 8.9.1

[63] The rule is not "differentiate top and bottom and hope." What licenses it is the Cauchy Mean Value Theorem 8.8.1: on any short interval the genuine ratio f/g equals a ratio of *derivatives* f'/g' at some interior point, and as the interval shrinks toward a that interior point is trapped near a too, where f'/g' is already close to A. The pinching by $r > A$ and $q < A$ below is just the ε–free way of saying "close to A from both sides."

Differentiable vector-valued function

Let $f : [a, b] \to \mathbb{R}^k$, written in components $f = (f_1, \dots, f_k)$ with each $f_i : [a, b] \to \mathbb{R}$. The function f is *differentiable at* x if the limit

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \in \mathbb{R}^k$$

exists. Here the numerator $f(x + h) - f(x)$ is a vector and the increment h is a scalar, so the quotient is a vector divided by a number. This is Rudin's one-parameter vector-valued derivative (R 5.16).[64]

Proposition 8.9.2

Differentiation is componentwise

A function $f = (f_1, \dots, f_k) : [a, b] \to \mathbb{R}^k$ is differentiable at x if and only if each component f_i is differentiable at x, in which case

$$f'(x) = (f_1'(x), \dots, f_k'(x)).$$

In particular, a complex-valued function $f = f_1 + if_2 : [a, b] \to \mathbb{C}$ is differentiable exactly when f_1 and f_2 are, with $f'(x) = f_1'(x) + if_2'(x)$. This is Rudin's componentwise differentiation (R 5.16).

Proof 8.9.2

Proof by Reading the Quotient Componentwise.

(1) For $h \neq 0$ the difference quotient is the vector whose i-th coordinate is the scalar difference quotient of f_i:

$$\frac{f(x+h) - f(x)}{h} = \left(\frac{f_1(x+h) - f_1(x)}{h}, \dots, \frac{f_k(x+h) - f_k(x)}{h} \right).$$

(2) A sequence of vectors in $\mathbb{R}^k$ converges if and only if each coordinate sequence converges, and the limit is taken coordinatewise. Hence the vector limit defining $f'(x)$ exists precisely when every $f_i'(x)$ exists, and then $f'(x) = (f_1'(x), \dots, f_k'(x))$.

(3) Identifying $\mathbb{C}$ with $\mathbb{R}^2$ through $f_1 + if_2 \leftrightarrow (f_1, f_2)$ specialises the statement to the complex case, giving $f' = f_1' + if_2'$.

■

[Direct] cf. R 5.16.

Remark 8.9.3

Which rules survive, and which break

[64] This is a genuinely different object from the multivariable derivative 8.1.4, even though both involve $\mathbb{R}^k$. Here the *input* is still one real number, so one may legitimately divide the vector displacement by the scalar step h; the derivative is itself a vector (a velocity). In 8.1.4 the input is a vector, division is impossible, and the derivative is a linear map. Same target, opposite side.

The sum, scalar-multiple, and product rules carry over to vector-valued functions. In particular, if $f, g : [a, b] \to \mathbb{R}^k$ are differentiable then the scalar-valued dot product $f \cdot g$ is differentiable with

$$(f \cdot g)' = f' \cdot g + f \cdot g'.$$

Two things *fail*, however. There is no quotient rule, simply because $\mathbb{R}^k$ has no division for $k \geq 2$. More importantly, the equality form of the Mean Value Theorem breaks: there need be no single c with $f(b) - f(a) = f'(c)(b - a)$. The standard witness is $f(x) = (\cos x, \sin x)$ on $[0, 2\pi]$, for which $f(2\pi) - f(0) = (0, 0)$, yet

$$\|f'(x)\| = \|(-\sin x, \cos x)\| = 1 \neq 0 \quad \text{for every } x,$$

so no interior point can make $f'(c)\,(2\pi - 0)$ vanish (cf. R 5.18).

Proposition 8.9.4

Mean Value Inequality for vector-valued functions

Let $f : [a, b] \to \mathbb{R}^k$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $x \in (a, b)$ such that

$$\|f(b) - f(a)\| \leq (b - a)\,\|f'(x)\|.$$

This inequality replaces the (failed) equality form of the Mean Value Theorem in the vector-valued setting (R 5.19).

Proof 8.9.4

Proof by Projecting Onto the Displacement and Cauchy–Schwarz.

(1) Set $z = f(b) - f(a) \in \mathbb{R}^k$ and define the real-valued function $\varphi : [a, b] \to \mathbb{R}$ by $\varphi(t) = z \cdot f(t)$, the projection of f onto the fixed displacement vector z. Then φ is continuous on $[a, b]$ and differentiable on (a, b) with $\varphi'(t) = z \cdot f'(t)$.

(2) By the scalar Mean Value Theorem 8.4.7, there exists $x \in (a, b)$ with

$$\varphi(b) - \varphi(a) = \varphi'(x)(b - a) = z \cdot f'(x)\,(b - a).$$

The left-hand side is $z \cdot f(b) - z \cdot f(a) = z \cdot \big(f(b) - f(a)\big) = z \cdot z = \|z\|^2$.

(3) By the Cauchy–Schwarz inequality, $z \cdot f'(x) \leq \|z\|\,\|f'(x)\|$, so

$$\|z\|^2 = z \cdot f'(x)\,(b - a) \leq \|z\|\,\|f'(x)\|\,(b - a).$$

(4) If $z \neq 0$, divide by $\|z\| > 0$ to obtain $\|z\| \leq (b - a)\|f'(x)\|$, which is the claim. If $z = 0$ the left-hand side $\|f(b) - f(a)\|$ is zero and the inequality holds trivially for any $x \in (a, b)$.

■

[Constr] uses 8.4.7; cf. R 5.19.

Remark 8.9.5

The physical reading

If $f(t)$ is the position of a particle at time t, then $f'(t)$ is its velocity vector and $\|f'(t)\|$ its speed. The inequality 8.9.4 then says that the straight-line distance between the endpoints, $\|f(b) - f(a)\|$, is at most the maximum speed times the elapsed time—distance travelled cannot exceed speed multiplied by duration, exactly as expected (cf. R 5.19).

Appendix

Supplementary material—additional proofs, context, and other treatments; not part of the lecture proper.

The first variation of the energy functional. This worked example was presented in lecture as an extended illustration of the abstract derivative 8.1.3; it is enrichment, and is *not* examinable. It carries the best-linear-approximation definition into a setting where the "point" is an entire curve, so the increment is itself a curve and there is nothing to divide by—exactly the situation 8.1.5 anticipated. The first variation yields only a *necessary* condition (a critical curve); a separate Cauchy–Schwarz bound (8.A.9) then shows that curve is the *global* minimiser.

Setup. Fix $p, q \in \mathbb{R}^k$ and let

$$\mathcal{C} = \{\, \gamma : [0,1] \to \mathbb{R}^k \mid \gamma(0) = p,\ \gamma(1) = q,\ \gamma \in C^1 \,\}$$

be the C^1 curves joining them: $\gamma(t)$ is the point reached at time t, the endpoint conditions pin p and q, and $\gamma'(t) \in \mathbb{R}^k$ is the velocity. A single "point" of $\mathcal{C}$ is a whole curve, so $\mathcal{C}$ is infinite-dimensional; it is an *affine*, not a vector, space (adding two curves through p, q lands at $2p, 2q$), which is why the perturbation directions form the separate set $\mathcal{V}$ of 8.A.2.

Definition 8.A.1

The energy functional

The *energy* of a curve $\gamma \in \mathcal{C}$ is

$$E(\gamma) = \int_0^1 \|\gamma'(t)\|^2 \, dt, \qquad \|\gamma'(t)\|^2 = \gamma'(t) \cdot \gamma'(t) = \sum_{j=1}^{k} \big(\gamma_j'(t)\big)^2,$$

the integral of the squared speed, so $E : \mathcal{C} \to \mathbb{R}$ sends a curve to a single real number. We ask which $\gamma \in \mathcal{C}$ has least energy. No naive difference quotient $\big(E(\gamma + h) - E(\gamma)\big)/h$ is available—"h" would be a curve, and one cannot divide by a curve—so we use the abstract derivative 8.1.3, which asks only for the best linear approximation, never for a quotient.

Definition 8.A.2

Admissible variations

To probe E near γ we nudge it in a direction drawn from

$$\mathcal{V} = \{\, \eta : [0,1] \to \mathbb{R}^k \mid \eta \in C^1,\ \eta(0) = \eta(1) = 0 \,\},$$

scaled by a real parameter h: the nudged curve $\gamma + h\eta$ moves each point $\gamma(t)$ by $h\,\eta(t)$. The endpoint conditions $\eta(0) = \eta(1) = 0$ keep it admissible, since

$$(\gamma + h\eta)(0) = p + h \cdot 0 = p, \qquad (\gamma + h\eta)(1) = q + h \cdot 0 = q$$

for *every* h; so $\mathcal{V}$ plays the role of the tangent space to $\mathcal{C}$ at γ.

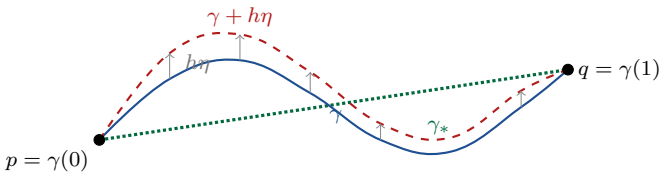


Figure 8.A.3. A curve and an admissible variation. The variation η is a vector field along γ vanishing at the endpoints, so every perturbed curve $\gamma + h\eta$ still joins p to q (no arrows at the ends). The dotted line is the energy-minimising straight curve γ_* of 8.A.8.

Step 1: reduce to one real variable. For fixed γ and admissible η, set

$$g(h) := E(\gamma + h\eta) = \int_0^1 \|\gamma'(t) + h\,\eta'(t)\|^2\,dt, \qquad g : \mathbb{R} \to \mathbb{R},$$

using $(\gamma + h\eta)' = \gamma' + h\eta'$. This is an ordinary real function of h, which one-variable calculus can attack.

Proposition 8.A.4

The energy along a variation is quadratic in h

For fixed γ and admissible η,

$$g(h) = E(\gamma + h\eta) = E(\gamma) + 2Bh + Ch^2, \qquad B = \int_0^1 \gamma' \cdot \eta'\,dt, \quad C = \int_0^1 \|\eta'\|^2\,dt.$$

Against the abstract expansion $\varphi(x_0 + h) = \varphi(x_0) + A(h) + o(h)$ of 8.1.3: the constant $E(\gamma)$ is $\varphi(x_0)$, the linear term $2Bh$ is the derivative $A(h)$, and the quadratic Ch^2 is the error—genuinely $o(h)$, since $|Ch^2|/|h| = |C|\,|h| \to 0$.

Proof 8.A.4

Direct Proof by Expanding and Integrating.

(1) Expand the integrand with $\|v\|^2 = v \cdot v$ and bilinearity:

$$\|\gamma' + h\eta'\|^2 = (\gamma' + h\eta') \cdot (\gamma' + h\eta') = \|\gamma'\|^2 + 2h\,(\gamma' \cdot \eta') + h^2\|\eta'\|^2,$$

the two cross terms coinciding because the dot product is symmetric.

(2) Integrate over $[0, 1]$ and pull the constants $2h, h^2$ outside (linearity of $\int$):

$$E(\gamma + h\eta) = \int_0^1 \|\gamma'\|^2\,dt + 2h\int_0^1 \gamma' \cdot \eta'\,dt + h^2\int_0^1 \|\eta'\|^2\,dt = E(\gamma) + 2Bh + Ch^2.$$

(3) For the remainder, $|Ch^2|/|h| = |C|\,|h| \to 0$ as $h \to 0$, so $Ch^2 = o(h)$.

■

[Direct] bilinearity and symmetry of $\cdot$; linearity of $\int$.

Definition 8.A.5

First variation

The *first variation* of E at γ in the direction η is the coefficient of h in 8.A.4,[65]

$$\delta E(\gamma)[\eta] := \frac{d}{dh}\Big|_{h=0} E(\gamma + h\eta) = g'(0) = 2B = 2\int_0^1 \gamma'(t)\cdot\eta'(t)\,dt,$$

the directional (Gâteaux) derivative of E at γ. (Differentiating $g(h) = E(\gamma) + 2Bh + Ch^2$ gives $g'(h) = 2B + 2Ch$, so $g'(0) = 2B$.)

Step 2: integration by parts. Assume now $\gamma \in C^2$,[66] so that γ'' is continuous, and rewrite $\int_0^1 \gamma'\cdot\eta'\,dt$ to move the derivative off the variation η and onto the curve γ.

Proposition 8.A.6

The first variation after integration by parts

For $\gamma \in C^2$ and admissible η,

$$\delta E(\gamma)[\eta] = 2\int_0^1 \gamma'\cdot\eta'\,dt = -2\int_0^1 \gamma''(t)\cdot\eta(t)\,dt.$$

Proof 8.A.6

Direct Proof by the Product Rule and the Fundamental Theorem.

(1) The product rule for $t \mapsto \gamma'(t)\cdot\eta(t) = \sum_j \gamma_j'\eta_j$ gives

$$\tfrac{d}{dt}(\gamma'\cdot\eta) = \gamma''\cdot\eta + \gamma'\cdot\eta', \qquad \text{so} \qquad \gamma'\cdot\eta' = \tfrac{d}{dt}(\gamma'\cdot\eta) - \gamma''\cdot\eta.$$

(2) Integrate over $[0,1]$; the exact-derivative term becomes a boundary value by the Fundamental Theorem of Calculus:

$$\int_0^1 \gamma'\cdot\eta'\,dt = \big[\gamma'(t)\cdot\eta(t)\big]_0^1 - \int_0^1 \gamma''\cdot\eta\,dt = \gamma'(1)\cdot\eta(1) - \gamma'(0)\cdot\eta(0) - \int_0^1 \gamma''\cdot\eta\,dt.$$

(3) The endpoint conditions $\eta(0) = \eta(1) = 0$ (8.A.2) kill the boundary term, leaving $\int_0^1 \gamma'\cdot\eta'\,dt = -\int_0^1 \gamma''\cdot\eta\,dt$. Multiplying by 2 gives the claim. ∎

[Direct] uses 8.A.2; the endpoint conditions kill the boundary term.

Step 3: stationarity and the fundamental lemma. If γ minimises E, then for every admissible η the function $g(h) = E(\gamma + h\eta)$ has a minimum at $h = 0$, so $g'(0) = \delta E(\gamma)[\eta] = 0$;

[65] On the board the first variation is written *with* a factor of h, as $E'(\gamma)\cdot\eta = 2h\int_0^1 \gamma'\cdot\eta'\,dt$, straight from the $2Bh$ term; we define it instead as the coefficient of h (a derivative should not still contain h). Both give the same stationarity condition—setting either to zero, the nonzero scalar divides out. Some texts write ε for the variation parameter in place of h.

[66] The board worked in C^1. Taking $\gamma \in C^2$ is what makes the integration by parts valid as written; with only $\gamma \in C^1$ the same conclusion $\gamma'' \equiv 0$ follows from the du Bois-Reymond lemma, at the cost of a longer argument. We follow the lecture's C^2 route.

by 8.A.6 this says $\int_0^1 \gamma'' \cdot \eta \, dt = 0$ for all admissible η. One integral vanishing does not force its integrand to vanish—but vanishing against *every* test direction does:

Lemma 8.A.7

Fundamental lemma of the calculus of variations

Let $f : [0,1] \to \mathbb{R}^k$ be continuous. If $\int_0^1 f(t) \cdot \eta(t) \, dt = 0$ for every $\eta \in C^1([0,1], \mathbb{R}^k)$ with $\eta(0) = \eta(1) = 0$, then $f \equiv 0$ on $[0,1]$.

Proof 8.A.7

Proof by Contradiction with a Bump Test Field.

(1) It suffices that each component $f_j \equiv 0$: taking $\eta = \phi \, e_j$ for a scalar ϕ and the standard basis vector e_j reduces the hypothesis to $\int_0^1 f_j \, \phi \, dt = 0$.

(2) Suppose $f_j(t_0) \neq 0$ for some $t_0 \in (0,1)$; say $f_j(t_0) > 0$. By continuity, $f_j > 0$ throughout an open interval $(a,b) \subseteq (0,1)$ containing t_0.

(3) Take the bump $\phi(t) = (t-a)^2(b-t)^2$ on (a,b) and $\phi = 0$ elsewhere. Then $\phi \in C^1([0,1])$ (value and first derivative vanish at a, b), $\phi(0) = \phi(1) = 0$, and $\phi > 0$ on (a,b), so

$$\int_0^1 f_j \, \phi \, dt = \int_a^b \underbrace{f_j}_{>0} \underbrace{\phi}_{>0} \, dt > 0,$$

contradicting the hypothesis. Hence $f_j \equiv 0$ for each j, and $f \equiv 0$.

■

[A9+Constr] reduce to one component, then localise with an explicit bump.

Proposition 8.A.8

The unique critical curve is the straight line

If $\gamma \in C^2$ satisfies $\delta E(\gamma)[\eta] = 0$ for all admissible η, then $\gamma'' \equiv 0$, equivalently

$$\gamma_*(t) = p + t(q-p) = (1-t)\, p + t\, q, \qquad t \in [0,1],$$

the constant-speed straight line from p to q, with energy $E(\gamma_*) = \|q-p\|^2$.

Proof 8.A.8

Constructive Proof by Integrating Twice.

(1) Stationarity and 8.A.6 give $\int_0^1 \gamma'' \cdot \eta \, dt = 0$ for all admissible η; since γ'' is continuous, 8.A.7 forces $\gamma'' \equiv 0$ (the Euler–Lagrange equation).

(2) Integrate twice: $\gamma'' = 0$ gives $\gamma'(t) = a$, then $\gamma(t) = a\, t + b$, with $a, b \in \mathbb{R}^k$.

(3) The endpoints give $\gamma(0) = b = p$ and $\gamma(1) = a + b = q$, so $a = q - p$ and $\gamma_*(t) = p + t(q-p)$, with velocity $\gamma_*'(t) = q - p$ and energy $E(\gamma_*) = \int_0^1 \|q-p\|^2 \, dt = \|q-p\|^2$.

■

[Constr] stationarity + 8.A.7 give $\gamma'' \equiv 0$; the endpoints fix the constants.

Theorem 8.A.9

The straight line is the global minimiser

For every $\gamma \in \mathcal{C}$,

$$E(\gamma) \;\geq\; \|q-p\|^2,$$

with equality if and only if $\gamma = \gamma_*$. Hence $\gamma_*(t) = p + t(q-p)$ is the unique minimiser of E over $\mathcal{C}$.

Proof 8.A.9

Direct Proof by the Cauchy–Schwarz Inequality.

(1) By the Fundamental Theorem of Calculus, componentwise, $q - p = \gamma(1) - \gamma(0) = \int_0^1 \gamma'(t)\,dt$.

(2) Apply the scalar Cauchy–Schwarz inequality $\left(\int_0^1 u\,dt\right)^2 \leq \int_0^1 u^2\,dt$ (the constant 1 in the second slot) to each component $u = \gamma_j'$, then sum:

$$\|q-p\|^2 = \sum_{j=1}^{k}\Big(\int_0^1 \gamma_j'\,dt\Big)^2 \leq \sum_{j=1}^{k}\int_0^1 (\gamma_j')^2\,dt = \int_0^1 \|\gamma'\|^2\,dt = E(\gamma).$$

(3) The straight line meets the bound, $E(\gamma_*) = \|q-p\|^2$, so it is a global minimiser. Equality in Cauchy–Schwarz forces each γ_j' constant, hence $\gamma' \equiv q - p$ and $\gamma = \gamma_*$; the minimiser is unique.

■

[Direct] a lower bound holding for every admissible curve; cf. 8.A.8.

Remark 8.A.10

The general Euler–Lagrange pattern

This is the first instance of a general pattern. For a Lagrangian $L = L(t, x, v)$, the functional $I[\gamma] = \int_0^1 L\big(t, \gamma, \gamma'\big)\,dt$ has first variation given by the same argument—first variation, then integration by parts—producing the *Euler–Lagrange equation*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial v}\right) - \frac{\partial L}{\partial x} = 0.$$

For the energy $L(t, x, v) = \|v\|^2$ one has $\partial L/\partial v = 2v$ and $\partial L/\partial x = 0$, so the equation reads $\frac{d}{dt}(2\gamma') = 0$, i.e. $\gamma'' = 0$—exactly 8.A.8.

Remark 8.A.11

Energy versus length

Energy is not length, but Cauchy–Schwarz links them. With $\text{Length}(\gamma) = \int_0^1 \|\gamma'\|\,dt$, treating the integrand as $\|\gamma'\| \cdot 1$,

$$\text{Length}(\gamma) = \int_0^1 \|\gamma'\| \cdot 1\,dt \leq \Big(\int_0^1 \|\gamma'\|^2\,dt\Big)^{1/2}\Big(\int_0^1 1\,dt\Big)^{1/2} = E(\gamma)^{1/2},$$

so $\text{Length}(\gamma)^2 \leq E(\gamma)$, with equality iff $\|\gamma'\|$ is constant. With $\text{Length}(\gamma) \geq \|q-p\|$ (triangle in-

equality for integrals) this chains to $E(\gamma) \geq \text{Length}(\gamma)^2 \geq \|q - p\|^2$. Length is reparametrisation-invariant but energy is not: minimising energy both shortens the curve *and* selects the constant-speed parametrisation—which is why energy, not length, is the convenient functional for geodesics.

LECTURE 9

The Riemann Integral: Construction, Integrability, and the Fundamental Theorem

MATH S4061 | Introduction to Modern Analysis I

Thursday, 25 June 2026

> *I take the bills and coins out of my pocket and give them to the creditor in the order I find them . . . This is the Riemann integral. But I can proceed differently: . . . I order the bills and coins according to identical values and . . . pay the several heaps one after the other . . . This is my integral.*
>
> Henri Lebesgue, letter to Paul Montel

Contents of this lecture

Overview. The calculus slogan is "area under the curve," but here the integral comes first: area is *defined by* the integral, not the reverse. We build the Riemann integral for any *bounded* f on a closed bounded interval—continuity is not required—by trapping the curve between under- and over-estimating staircases, and we call f integrable exactly when the two estimates can be made to agree. Refining a partition squeezes the sums together, and that single fact yields the working test for integrability: $U - L$ can be made smaller than any ε. The test then delivers three classes of integrable functions—continuous, monotone, and boundedly discontinuous at finitely many points—and points past Ch. 6 to the full Lebesgue criterion, which the Cantor set shows is genuinely sharper. We close by recording the algebraic properties of the integral and the two halves of the Fundamental Theorem, whose proofs open the next lecture. The companion text is Rudin, Chapter 6, read throughout with $\alpha(x) = x$, so $\int f\,d\alpha$ becomes $\int f\,dx$ (R 6.1–R 6.21).

9.1 Upper and Lower Integrals

In the calculus picture one partitions $[a,b]$ into equal pieces, samples f at a point x_i of each, and reads $\int_a^b f(x)\,dx$ as the limit of the Riemann sums $R_n(f) = \sum_{i=1}^{n} f(x_i)\,\Delta x_i$. We will be more general: continuity is *not* required, and many discontinuous functions still turn out to be integrable. We ask only that

$$f : [a,b] \to \mathbb{R} \quad \text{be a } \textit{bounded} \text{ function}, \qquad M := \sup_{[a,b]} |f(x)| < \infty,$$

on a closed and bounded interval.[67]

Definition 9.1.1

Partition of $[a,b]$

A *partition* of $[a,b]$ is a finite set $P = \{x_0, x_1, \dots, x_n\} \subseteq [a,b]$ with

$$a = x_0 \le x_1 \le x_2 \le \cdots \le x_n = b,$$

and we write $\Delta x_i = x_i - x_{i-1}$ for the length of the i-th subinterval (R 6.1). Consecutive points may coincide.

Definition 9.1.2

Upper and lower sums

Given a partition P of $[a,b]$, set

$$M_i = \sup_{[x_{i-1},x_i]} f, \qquad m_i = \inf_{[x_{i-1},x_i]} f,$$

the supremum and infimum of f over the i-th subinterval. The *upper* and *lower Riemann sums* of f over P are

$$U(P,f) = \sum_{i=1}^{n} M_i\,\Delta x_i, \qquad L(P,f) = \sum_{i=1}^{n} m_i\,\Delta x_i$$

(R 6.1). Since $m_i \le M_i$ on each subinterval, $L(P,f) \le U(P,f)$ for every P.

[67] Rudin builds the more general *Riemann–Stieltjes* integral $\int f\,d\alpha$ against an integrator α (R 6.1–R 6.2). Throughout this lecture we take $\alpha(x) = x$, which collapses $\int f\,d\alpha$ to the ordinary $\int f(x)\,dx$; every weight $\Delta\alpha_i$ below is just $\Delta x_i = x_i - x_{i-1}$. Boundedness is what makes the suprema and infima of 9.1.2 real numbers; both the closedness and the boundedness of $[a,b]$ matter.

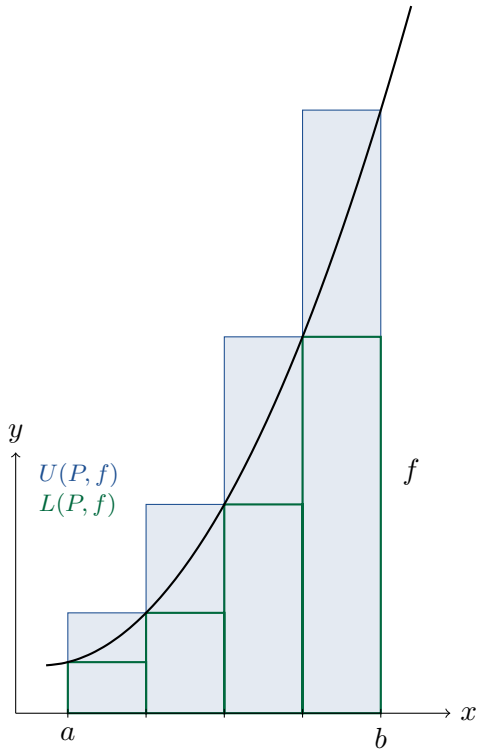


Figure 9.1.3. UPPER AND LOWER SUMS. The upper sum $U(P,f)$ is the area of the staircase whose steps reach the suprema M_i (tops above the curve); the lower sum $L(P,f)$ is the area of the staircase whose steps reach the infima m_i (tops below). The graph of f is caught between them, so $L(P,f) \le \int_a^b f \le U(P,f)$ whenever the integral exists.

Definition 9.1.4

Upper and lower integrals; Riemann integrable

The *upper* and *lower integrals* of f over $[a,b]$ are

$$\overline{\int_a^b} f = \inf_P U(P,f), \qquad \underline{\int_a^b} f = \sup_P L(P,f),$$

the infimum and supremum taken over all partitions P of $[a,b]$. We say f is *Riemann integrable*, written $f \in \mathcal{R}$, when these agree,

$$\underline{\int_a^b} f = \overline{\int_a^b} f,$$

and then their common value is the *integral* $\displaystyle\int_a^b f$ (R 6.2). Here $\mathcal{R} = \mathcal{R}[a,b]$ denotes the class of Riemann-integrable functions on $[a,b]$. If instead $\underline{\int} f \neq \overline{\int} f$, then $f \notin \mathcal{R}$.

Counterexample 9.1.5

The Dirichlet function

Let $f : [a, b] \to \mathbb{R}$ be the indicator of the irrationals,

$$f(x) = \begin{cases} 1, & x \notin \mathbb{Q}, \\ 0, & x \in \mathbb{Q}, \end{cases}$$

which is discontinuous at every point. Every subinterval $[x_{i-1}, x_i]$ contains both rationals and irrationals, so for any partition P,

$$M_i = 1, \qquad m_i = 0,$$

whence

$$U(P, f) = \sum_i M_i \, \Delta x_i = b - a = \overline{\int_a^b} f, \qquad L(P, f) = 0 = \underline{\int_a^b} f.$$

Since $b - a \neq 0$, the upper and lower integrals differ, so $f \notin \mathcal{R}$.[68]

FAILS: the claim that every bounded function is Riemann integrable.

9.2 Refinement and the Integrability Criterion

Definition 9.2.1

Refinement and common refinement

A partition P^* *refines* P if $P^* \supseteq P$ as sets of points; P^* then has all of P's division points and possibly more. Partitions are only *partially* ordered: for two partitions P_1, P_2 it may happen that $P_1 \not\subseteq P_2$ and $P_2 \not\subseteq P_1$. Their *common refinement* is the union $P_1 \cup P_2$, which refines both:

$$P_1 \cup P_2 \supseteq P_1, \qquad P_1 \cup P_2 \supseteq P_2$$

(R 6.3). The *mesh* of P is $|P| = \max_i \Delta x_i$, the width of its widest subinterval.[69]

Proposition 9.2.2

Refinement inequalities

If P^* refines P, then

$$L(P, f) \leq L(P^*, f) \qquad \text{and} \qquad U(P^*, f) \leq U(P, f)$$

(R 6.4). Refining a partition can only raise the lower sum and lower the upper sum.

Idea. Refining pushes the two staircases toward the curve, so the sums move together. It is enough to add a single point and iterate: passing to the smaller subintervals can only raise an

[68] The orientation is immaterial: putting $f = 1$ on $\mathbb{Q}$ instead swaps the names but leaves every $M_i = 1$, every $m_i = 0$, and the same gap $U - L = b - a$. The function reappears in 9.3.8, where the Lebesgue integral—grouping by value rather than by location—recovers a value Riemann cannot.

[69] Morally one wants $\int_a^b f = \lim_{|P| \to 0} R_n(f)$, the limit of the Riemann sums as the mesh shrinks. Refinement makes this precise without mentioning a limit: rather than driving $|P| \to 0$, we compare a partition with its refinements and show the upper and lower sums close in on each other (9.2.2).

infimum and lower a supremum.

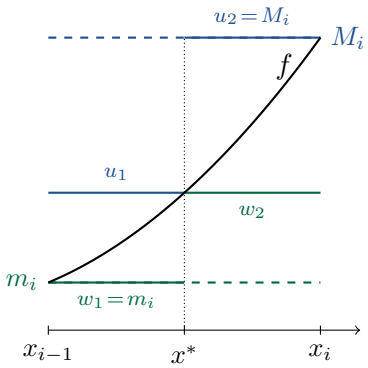


Figure 9.2.3. ADDING ONE POINT TO A PARTITION. Inserting x^* splits $[x_{i-1}, x_i]$ into two pieces. The infimum over the whole, m_i, is at most the infima w_1, w_2 over the halves (inf over a smaller set rises), so the lower sum grows; the supremum over the whole, M_i, is at least the suprema u_1, u_2 (sup over a smaller set falls), so the upper sum shrinks.

Proof 9.2.2

Proof by Induction on the Number of Added Points.

(1) It suffices to treat $P^* = P \cup \{x^*\}$, one extra point: a general refinement is reached by adding the finitely many new points one at a time, and chaining the resulting inequalities gives the full statement.

(2) Say x^* falls in the i-th subinterval, $x_{i-1} \le x^* \le x_i$, splitting it into $[x_{i-1}, x^*]$ and $[x^*, x_i]$. Write

$$w_1 = \inf_{[x_{i-1}, x^*]} f, \quad w_2 = \inf_{[x^*, x_i]} f, \qquad u_1 = \sup_{[x_{i-1}, x^*]} f, \quad u_2 = \sup_{[x^*, x_i]} f.$$

(3) The infimum over a set is at most the infimum over any subset, and dually for suprema, so

$$m_i \le w_1, \qquad m_i \le w_2, \qquad M_i \ge u_1, \qquad M_i \ge u_2.$$

(4) All other subintervals are untouched, so on passing from P to P^* only the i-th term changes. Using $\Delta x_i = (x^* - x_{i-1}) + (x_i - x^*)$,

$$L(P^*, f) - L(P, f) = w_1(x^* - x_{i-1}) + w_2(x_i - x^*) - m_i\,\Delta x_i \ge 0,$$

by Step (3), and likewise

$$U(P, f) - U(P^*, f) = M_i\,\Delta x_i - u_1(x^* - x_{i-1}) - u_2(x_i - x^*) \ge 0.$$

Hence $L(P,f) \le L(P^*,f)$ and $U(P^*,f) \le U(P,f)$.

■

[A10] cf. R 6.4; reduce to a single inserted point.

Corollary 9.2.4

Every lower sum is below every upper sum

For any two partitions P_1, P_2 of $[a,b]$,

$$L(P_1,f) \le U(P_2,f)$$

(R 6.5).

Proof 9.2.4

Proof by Passing Through the Common Refinement.

(1) Let $P^* = P_1 \cup P_2$ be the common refinement, which refines both P_1 and P_2 (9.2.1). Then
$$L(P_1,f) \;\le\; L(P_1 \cup P_2,f) \;\le\; U(P_1 \cup P_2,f) \;\le\; U(P_2,f).$$

(2) The first inequality is 9.2.2 applied to P_1 (lower sums rise under refinement), the middle is $L \le U$ for the single partition P^* (9.1.2), and the last is 9.2.2 applied to P_2 (upper sums fall under refinement). Reading the ends gives the claim.

■

[Direct] cf. R 6.5.

Corollary 9.2.5

Lower integral does not exceed upper integral

For every bounded f on $[a,b]$,

$$\underline{\int_a^b} f \;\le\; \overline{\int_a^b} f$$

(R 6.5).

Proof 9.2.5

Proof by a Supremum-then-Infimum Argument.

(1) Fix P_2. By 9.2.4, $L(P_1,f) \le U(P_2,f)$ for every P_1, so $U(P_2,f)$ is an upper bound for the lower sums; taking the supremum over P_1,
$$\underline{\int_a^b} f = \sup_{P_1} L(P_1,f) \;\le\; U(P_2,f).$$

(2) This now holds for every P_2, so $\underline{\int} f$ is a lower bound for the upper sums; taking the infimum over P_2,
$$\underline{\int_a^b} f \;\le\; \inf_{P_2} U(P_2,f) = \overline{\int_a^b} f.$$

■

[A5] cf. R 6.5; take sup then inf in 9.2.4.

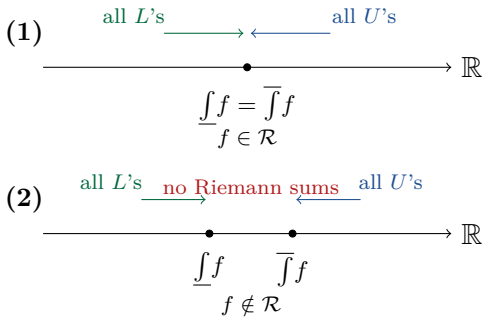


Figure 9.2.6. THE TWO PICTURES. All lower sums lie to the left, all upper sums to the right, and $\underline{\int} f \le \overline{\int} f$ separates them. **(1)** The two families meet at a single point: $\underline{\int} f = \overline{\int} f$ and $f \in \mathcal{R}$. **(2)** A gap remains, spanned by no Riemann sum: $\underline{\int} f < \overline{\int} f$ and $f \notin \mathcal{R}$.

Theorem 9.2.7

The integrability criterion

A bounded function f is Riemann integrable on $[a, b]$ if and only if for every $\varepsilon > 0$ there is a partition P with

$$U(P, f) - L(P, f) < \varepsilon$$

(R 6.6).

Idea. Closeness of the two sums means closeness of the two integrals. If U and L can be brought within ε, the upper and lower integrals—trapped between them—are within ε too, for every ε, hence equal.

Proof 9.2.7

Proof of the Integrability Criterion.

(1) ($\Rightarrow$) Suppose $f \in \mathcal{R}$, so $\underline{\int} f = \overline{\int} f = \int_a^b f$. Let $\varepsilon > 0$. By the definitions of the infimum and supremum (9.1.4) there are partitions P_1, P_2 with

$$U(P_1, f) < \int_a^b f + \tfrac{\varepsilon}{2}, \qquad L(P_2, f) > \int_a^b f - \tfrac{\varepsilon}{2}.$$

Put $P = P_1 \cup P_2$. By 9.2.2, $U(P, f) \le U(P_1, f)$ and $L(P, f) \ge L(P_2, f)$, so

$$U(P, f) - L(P, f) \;\le\; U(P_1, f) - L(P_2, f) \;<\; \varepsilon.$$

(2) ($\Leftarrow$) Suppose that for every $\varepsilon > 0$ some P has $U(P,f) - L(P,f) < \varepsilon$. Since always

$$L(P,f) \;\le\; \underline{\int_a^b} f \;\le\; \overline{\int_a^b} f \;\le\; U(P,f),$$

the gap satisfies $0 \le \overline{\int} f - \underline{\int} f \le U(P,f) - L(P,f) < \varepsilon$. As $\varepsilon > 0$ was arbitrary, $\overline{\int} f - \underline{\int} f = 0$, i.e. $f \in \mathcal{R}$.

■

[A1] cf. R 6.6; ($\Leftarrow$) is the two-pictures argument of 9.2.6.

Proposition 9.2.8

A working tool

Suppose $f \in \mathcal{R}$ on $[a,b]$, let $\varepsilon > 0$, and let P be a partition with $U(P,f) - L(P,f) < \varepsilon$. Then:

(a) the same inequality holds for every refinement $P^* \supseteq P$;

(b) for any sample points $s_i, t_i \in [x_{i-1}, x_i]$, $\displaystyle\sum_{i=1}^{n} |f(s_i) - f(t_i)|\,\Delta x_i < \varepsilon$;

(c) for any sample points $t_i \in [x_{i-1}, x_i]$, $\displaystyle\Big|\sum_{i=1}^{n} f(t_i)\,\Delta x_i - \int_a^b f\Big| < \varepsilon$.

This is the technical tool drawn on in the proofs below (R 6.7).

Proof 9.2.8

Proof by Sandwiching Between the Infima and Suprema.

(1) For (a), refining only shrinks U and grows L (9.2.2), so $U(P^*,f) - L(P^*,f) \le U(P,f) - L(P,f) < \varepsilon$.

(2) For (b), on each subinterval $f(s_i), f(t_i) \in [m_i, M_i]$, hence $|f(s_i) - f(t_i)| \le M_i - m_i$. Summing,

$$\sum_i |f(s_i) - f(t_i)|\,\Delta x_i \;\le\; \sum_i (M_i - m_i)\,\Delta x_i \;=\; U(P,f) - L(P,f) \;<\; \varepsilon.$$

(3) For (c), since $t_i \in [x_{i-1}, x_i]$ we have $m_i \le f(t_i) \le M_i$, so both $\sum_i f(t_i)\,\Delta x_i$ and $\int_a^b f$ lie in the interval $[L(P,f), U(P,f)]$, of length $< \varepsilon$. Two numbers in an interval of length $< \varepsilon$ differ by $< \varepsilon$.

■

[Direct] cf. R 6.7.

9.3 Classes of Integrable Functions

Theorem 9.3.1

Continuous functions are integrable

If $f : [a,b] \to \mathbb{R}$ is continuous, then $f \in \mathcal{R}$ (R 6.8).

Idea. The key step is *uniform* continuity: on the compact interval $[a,b]$ a single δ controls the oscillation everywhere at once, so a fine enough partition makes every $M_i - m_i$ uniformly small.

Proof 9.3.1

Proof by Uniform Continuity on a Compact Interval.

(1) Since $[a,b]$ is compact, f is *uniformly continuous* (7.7.2). Given $\varepsilon > 0$ there is $\delta = \delta(\varepsilon) > 0$ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \frac{\varepsilon}{b-a}.$$

(2) Choose a partition P with $\Delta x_i < \delta$ for all i—for instance n equal subintervals with $(b-a)/n < \delta$. On each subinterval f attains its supremum and infimum (a continuous function on a compact set attains its extrema), so $M_i = f(\xi_i)$ and $m_i = f(\eta_i)$ for some $\xi_i, \eta_i \in [x_{i-1}, x_i]$ with $|\xi_i - \eta_i| < \delta$; hence

$$M_i - m_i < \frac{\varepsilon}{b-a} \qquad \text{for all } i.$$

(3) Therefore

$$U(P,f) - L(P,f) = \sum_{i=1}^{n} (M_i - m_i)\,\Delta x_i < \frac{\varepsilon}{b-a} \sum_{i=1}^{n} \Delta x_i = \frac{\varepsilon}{b-a}\,(b-a) = \varepsilon,$$

using $\sum_i \Delta x_i = b - a$. By the integrability criterion (9.2.7), $f \in \mathcal{R}$. ∎

[A6+A1] cf. R 6.8; uniform continuity is 7.7.2.

Theorem 9.3.2

Monotone functions are integrable

If f is monotone on $[a,b]$, then $f \in \mathcal{R}$ (R 6.9). Such an f may have at most countably many jump discontinuities.

Idea. For a monotone f the differences $M_i - m_i$ telescope: on equal subintervals the total oscillation is just $f(b) - f(a)$, spread over n pieces, so it vanishes as $n \to \infty$.

Proof 9.3.2

Proof by Telescoping on an Equal Partition.

(1) Treat f monotone increasing. Let $\varepsilon > 0$ and take the equal partition with

$$\Delta x_i = \frac{b-a}{n}, \qquad n > \frac{(b-a)\big(f(b) - f(a)\big)}{\varepsilon}.$$

(2) On $[x_{i-1}, x_i]$ monotonicity gives $m_i = f(x_{i-1})$ and $M_i = f(x_i)$. Pulling the con-

stant $\Delta x_i = (b-a)/n$ out of the sum, the differences telescope:

$$U(P,f) - L(P,f) = \sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))\Delta x_i = \frac{b-a}{n} \sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))$$
$$= \frac{b-a}{n}(f(b) - f(a)).$$

(3) By the choice of n this is $< \varepsilon$. The criterion (9.2.7) gives $f \in \mathcal{R}$.

■

[Direct] cf. R 6.9; the decreasing case follows by applying the increasing case to $-f$.

Theorem 9.3.3

Finitely many discontinuities still permit integrability

If $f : [a,b] \to \mathbb{R}$ is bounded and has only *finitely many* discontinuities, then $f \in \mathcal{R}$ (R 6.10).

Idea. Finitely many points can be covered by intervals of arbitrarily small total length. Box the bad points into tiny intervals where the bounded f can misbehave by at most $2M$ each; on the compact remainder f is uniformly continuous, so the earlier argument applies. The two contributions to $U - L$ are then each $O(\varepsilon)$.

Proof 9.3.3

Constructive Proof by Covering the Discontinuities.

(1) Let $\varepsilon > 0$ and set $M := \sup_{[a,b]} |f| < \infty$, so $-M \le f \le M$. Let $E \subseteq [a,b]$ be the (finite) set of discontinuities. Cover E by finitely many disjoint closed intervals $[u_j, v_j]$, $j = 1, \dots, m$, such that

(1) each point of $E \cap (a,b)$ is interior to some $[u_j, v_j]$; (2) $\sum_j (v_j - u_j) < \varepsilon$.

(2) The set

$$K := [a,b] \setminus \bigcup_j (u_j, v_j)$$

is closed in $[a,b]$ and hence compact, and by construction f is continuous on K. Therefore f is uniformly continuous on K (7.7.2): there is $\delta > 0$ with $x, y \in K$, $|x-y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon$.

(3) Form a partition P such that: (i) every $u_j, v_j \in P$; (ii) no point of (u_j, v_j) lies in P; (iii) every subinterval not of the form $[u_j, v_j]$ has length $< \delta$. Split the upper-minus-lower sum according to the two flavours of subinterval:

$$U(P,f) - L(P,f) = \underbrace{\sum_{\text{covered}} (M_i - m_i)\,\Delta x_i}_{\text{the } [u_j, v_j]} + \underbrace{\sum_{\text{other}} (M_i - m_i)\,\Delta x_i}_{\text{length } < \delta}.$$

(4) On a covered subinterval $M_i - m_i \le 2M$, and these subintervals have total length $< \varepsilon$, so the first sum is $\le 2M\varepsilon$. On the others, the subinterval lies in K, where f is uniformly continuous, so $M_i - m_i \le \varepsilon$ by Step (2); the second sum is then $\le \varepsilon \sum_i \Delta x_i \le (b-a)\varepsilon$. Hence

$$U(P,f) - L(P,f) \;\le\; 2M\varepsilon + (b-a)\varepsilon.$$

The right side is an arbitrary multiple of ε, so $U - L$ can be made smaller than any prescribed positive number; by the criterion (9.2.7), $f \in \mathcal{R}$.

■

[Constr] cf. R 6.10; a two-flavour partition splits $U - L$ into a covered part and a uniformly continuous part.

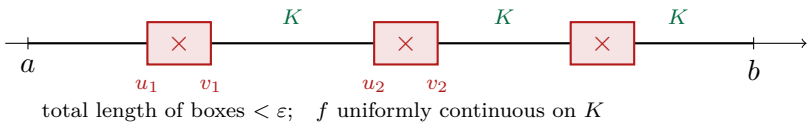


Figure 9.3.4. COVERING THE DISCONTINUITIES. The discontinuities (marked ×) are boxed inside small closed intervals $[u_j, v_j]$ of total length $< \varepsilon$. On the compact remainder $K = [a,b] \setminus \bigcup_j (u_j, v_j)$ the function is continuous, hence uniformly continuous, so a partition fine away from the boxes controls $U - L$.

Theorem 9.3.5

The Lebesgue criterion

Let $f : [a,b] \to \mathbb{R}$ be bounded. Then $f \in \mathcal{R}$ if and only if the set $E \subseteq [a,b]$ of discontinuities of f can be covered by countably many intervals of total length $< \varepsilon$ for every $\varepsilon > 0$—that is, E has *length (measure) zero*. This is the full integrability criterion; the lecture proved only the finite case (9.3.3).[70]

Proof 9.3.5

The General Lebesgue Criterion.

↪ Stated as the complete characterization of $\mathcal{R}$; its proof, and the precise notion of measure zero, lie beyond Ch. 6 (Rudin Ch. 11). The lecture established only the finite-discontinuity case (9.3.3).

Example 9.3.6

The Cantor set: uncountable, yet length zero

[70] "Length zero" is the intuitive face of *Lebesgue measure zero*: a set E is null if for every $\varepsilon > 0$ it fits inside a countable union of intervals whose lengths total less than ε. The finite cover of 9.3.3 is the simplest instance. The notion—and this characterization of $\mathcal{R}$—is made precise in the measure theory of Ch. 11 (R 11.33).

Build the middle-thirds set by deleting open middle thirds: $C_0 = [0,1]$; C_1 removes the middle third, leaving two intervals; C_2 removes the middle third of each, leaving four; and so on. The *Cantor set* is the intersection

$$C = \bigcap_{n \geq 0} C_n.$$

It is *uncountable*: a point of C is determined by an infinite sequence of left/right choices, one at each stage, so by the binary-choice argument of 2.1.10 there are uncountably many. Yet C has length zero,[71] since the deleted thirds exhaust the unit length,

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \frac{8}{81} + \cdots = \frac{1}{3} \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^k = 1, \qquad \text{so} \qquad 1 - \frac{1}{3} - \frac{2}{9} - \frac{4}{27} - \cdots = 0.$$

By the Lebesgue criterion (9.3.5), a bounded function discontinuous *exactly* on C is still Riemann integrable—the sharp contrast with the Dirichlet function (9.1.5), which is discontinuous on all of $[a,b]$, a set of full length $b-a$ and far from null.

Figure 9.3.7. The Cantor middle-thirds construction. Each stage deletes the open middle third of every surviving interval: C_0 is one interval, C_1 two, C_2 four. The total deleted length is $\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = 1$, so $C = \bigcap_n C_n$ has length zero—while remaining uncountable.

Remark 9.3.8

Toward the Lebesgue integral

Where Riemann sums by *location*, Lebesgue sums by *value*. For the Dirichlet function this recovers an integral Riemann cannot assign:

$$\int_{[a,b]} f = \int_{[a,b]\cap\mathbb{Q}} f + \int_{[a,b]\setminus\mathbb{Q}} f = 0 \cdot \ell\big([a,b]\cap\mathbb{Q}\big) + 1 \cdot \ell\big([a,b]\setminus\mathbb{Q}\big) = b - a,$$

since the rationals have length 0 and the irrationals length $b-a$. Riemann cannot do this, because $\{\mathbb{Q},\ \mathbb{R}\setminus\mathbb{Q}\}$ is "not a partition" of $[a,b]$ in his sense—it splits by value, not by intervals. This is the move the epigraph describes, and it belongs to Ch. 11.

9.4 Properties and the Fundamental Theorem

The remaining results were stated in lecture with their proofs *deferred to Lecture 10*; we record them here as stated facts so the lecture is complete and the Fundamental Theorem is in place. Each proof is supplied next time.

[71] The geometric-series computation is the intuitive face of a rigorous fact: C has *Lebesgue measure zero*, by the countable additivity of measure (Ch. 11). That machinery is not needed here—the elementary length bound suffices—but it is what makes "length zero" precise.

Theorem 9.4.1

Composition with a continuous function

If $f \in \mathcal{R}$ on $[a,b]$ with $m \le f \le M$, and φ is continuous on the closed interval $[m,M]$, then $\varphi \circ f \in \mathcal{R}$ (R 6.11).

Proof 9.4.1

Composition Preserves Integrability.

↪ Proof deferred to Lecture 10.

Proposition 9.4.2

Properties of the integral

For $f, g \in \mathcal{R}$ on $[a,b]$:

(a) *Linearity:* $\int_a^b (f+g) = \int_a^b f + \int_a^b g$ and $\int_a^b cf = c\int_a^b f$ for $c \in \mathbb{R}$.

(b) *Monotonicity:* if $f \le g$ on $[a,b]$ then $\int_a^b f \le \int_a^b g$; and if $|f(x)| \le M$ then $\left|\int_a^b f\right| \le M(b-a)$.

(c) *Products:* $fg \in \mathcal{R}$.

(d) *Absolute value:* $|f| \in \mathcal{R}$, and $\left|\int_a^b f\right| \le \int_a^b |f|$.

(e) *Additivity:* for $c \in [a,b]$, $\int_a^b f = \int_a^c f + \int_c^b f$.

These are Rudin's basic properties of the integral (R 6.12–R 6.13).

Proof 9.4.2

Properties of the Integral.

↪ Proofs deferred to Lecture 10.

Theorem 9.4.3

Fundamental Theorem of Calculus, I

Let $f \in \mathcal{R}$ on $[a,b]$. Then the "antiderivative"

$$F(x) = \int_a^x f(t)\,dt$$

is uniformly continuous on $[a,b]$; and if f is continuous at a point $t_0 \in [a,b]$, then F is differentiable at t_0 with

$$F'(t_0) = f(t_0)$$

(R 6.20).

Proof 9.4.3

The Fundamental Theorem, I.

↪ Proof deferred to Lecture 10.

Theorem 9.4.4

Fundamental Theorem of Calculus, II

Let $f \in \mathcal{R}$ on $[a,b]$, and suppose there is $F : [a,b] \to \mathbb{R}$ with $F' = f$ on $[a,b]$. Then

$$\int_a^b f = F(b) - F(a)$$

(R 6.21).

Proof 9.4.4

The Fundamental Theorem, II.

↪ Proof deferred to Lecture 10.

Remark 9.4.5

Integration smooths, differentiation roughens

A closing thought from lecture. Integration makes functions *better*: it carries a merely integrable f to a continuous F, and a continuous f to a differentiable one (9.4.3). Differentiation runs the other way, making functions *worse*—a differentiable function can have a discontinuous derivative. The two operations undo each other, which is exactly the content of the Fundamental Theorem (9.4.3, 9.4.4).

PART II

Problem Sets

Problem Set Conventions

MATH S4061 | Introduction to Modern Analysis I

Problem Sets in This Part

Note to the reader. Problems from Walter Rudin are taken from *Principles of Mathematical Analysis*, 3rd ed. (McGraw-Hill, 1976); the remaining problems come from course homework, practice material, and exams. A **green** left rule marks an **assigned** problem, and a **blue** left rule marks an **unassigned** problem included for completeness. Each solution is followed by a gray Reflections block; these reflections are the student's study notes and are *not* part of the official submitted solution.

HOMEWORK 1

The Real Numbers

MATH S4061 | Introduction to Modern Analysis I

Rudin Ch. 1–2, Induction, an Equivalence Relation, an Ordered-Field Obstruction

Problems in This Set

Problem A1 Rudin Ch. 1, Ex. 1 — A Rational Plus an Irrational

Problem A2 Rudin Ch. 1, Ex. 2 — No Rational Squares to 12

Problem A3 Rudin Ch. 1, Ex. 3 — Multiplicative Cancellation Laws

Problem A4 Rudin Ch. 1, Ex. 4 — A Lower Bound Is at Most an Upper Bound

Problem A5 Rudin Ch. 1, Ex. 5 — $\inf A = -\sup(-A)$

Problem A6 Rudin Ch. 1, Ex. 6 — Real Powers from Rational Powers

Problem A7 Rudin Ch. 1, Ex. 7 — Logarithms from Completeness

Problem A8 Rudin Ch. 1, Ex. 8 — $\mathbb{C}$ Admits No Order

Problem A9 Rudin Ch. 1, Ex. 9 — Lexicographic Order on $\mathbb{C}$

Problem A10 Rudin Ch. 1, Ex. 10 — Complex Square Roots

Problem A11 Rudin Ch. 1, Ex. 11 — Polar Decomposition

Problem A12 Rudin Ch. 1, Ex. 12 — The Finite Triangle Inequality

Problem A13 Rudin Ch. 1, Ex. 13 — The Reverse Triangle Inequality

Problem A14 Rudin Ch. 1, Ex. 14 — A Unit-Circle Identity

Problem A15 Rudin Ch. 1, Ex. 15 — Equality in Schwarz

Problem A16 Rudin Ch. 1, Ex. 16 — Intersecting Two Spheres

Problem A17 Rudin Ch. 1, Ex. 17 — The Parallelogram Identity

Problem A18 Rudin Ch. 1, Ex. 18 — Orthogonal Vectors

Problem A19 Rudin Ch. 1, Ex. 19 — An Apollonius Sphere

Problem A20 Rudin Ch. 1, Ex. 20 — Cuts Without the No-Largest Condition

Problem B Induction: $3 \mid n^3 + 2n$

Problem C Induction: $2^n > n^2$ for $n > 4$

Problem D The Defining Relation of $\mathbb{Q}$ Is an Equivalence Relation

Problem E $ab = c \Rightarrow a \le \sqrt{c}$ or $b \le \sqrt{c}$

Problem F Rudin Ch. 2, Ex. 2 — The Algebraic Numbers Are Countable

Problem A1 *Rudin, Chapter 1, Exercise 1*

If r is rational ($r \neq 0$) and x is irrational, prove that $r + x$ and rx are irrational.

SOLUTION.

Recall that $\mathbb{Q}$ is a field: it is closed under subtraction and multiplication, and every nonzero rational has a rational inverse.

Suppose, toward a contradiction, that $r + x$ is rational. Since r is rational and $\mathbb{Q}$ is closed under subtraction, $x = (r + x) - r \in \mathbb{Q}$, contradicting that x is irrational. Hence $r + x$ is irrational.

Suppose instead that rx is rational. Since $r \neq 0$, its inverse r^{-1} is rational, so $x = (rx)\, r^{-1} \in \mathbb{Q}$, again a contradiction. Hence rx is irrational. The hypothesis $r \neq 0$ is used exactly once, to form r^{-1}; without it the product claim is false, since $0 \cdot x = 0$.

■

REFLECTIONS.

The word that sets the strategy is *irrational*. It is a negative property—it says a number is *not* in $\mathbb{Q}$—and a negative conclusion gives a direct proof nothing to push on, so the reflex is contradiction (0.5.5): assume the conclusion fails and manufacture a rational number that has no right to exist. The only other ingredient on offer is that r is a nonzero rational, which is exactly the raw material such a contradiction runs on, because $\mathbb{Q}$ is a field (1.3.1; the quotable "$\mathbb{Q}$ is an ordered field"): closed under the operations, with an inverse for every nonzero element.

So the plan writes itself—assume rationality and undo the rational part with a field operation:

(1) if $r + x \in \mathbb{Q}$, then $x = (r + x) - r$ is rational by closure under subtraction, and the contradiction is immediate;

(2) if $rx \in \mathbb{Q}$, then $x = (rx)\, r^{-1}$ is rational—and here $r \neq 0$ earns its keep, since r^{-1} exists exactly when $r \neq 0$.

Seeing where each hypothesis is spent shows the statement is sharp in both directions: drop "r rational" and it collapses ($\sqrt{2} + (-\sqrt{2}) = 0$, $\sqrt{2} \cdot \sqrt{2} = 2$); drop $r \neq 0$ and the product fails, since $rx = 0$ for every x.

The lasting lesson is how to handle "irrational" at all. The irrationals have no closure of their own, so you rarely argue with them head-on: assume rationality, isolate the known irrational using only operations that keep you inside $\mathbb{Q}$, and let field closure spring the trap.

Problem A2 *Rudin, Chapter 1, Exercise 2*

Prove that there is no rational number whose square is 12.

Solution.

Suppose, toward a contradiction, that $p \in \mathbb{Q}$ and $p^2 = 12$. Write $p = m/n$ in lowest terms, with $m, n \in \mathbb{Z}$, $n \neq 0$, and no prime dividing both m and n. Then

$$m^2 = 12n^2 = 3 \cdot 4n^2.$$

Thus $3 \mid m^2$, hence $3 \mid m$ by Euclid's lemma. Write $m = 3k$. Substituting back gives

$$9k^2 = 12n^2, \qquad \text{so} \qquad 3k^2 = 4n^2.$$

The right-hand side is divisible by 3, so $3 \mid n^2$, hence $3 \mid n$. This contradicts that m/n was in lowest terms. Therefore no rational number has square 12.

■

Reflections.

This is the same descent obstruction as the standard proof that $\sqrt{2}$ is irrational (1.4.1). A square equal to 12 would force the prime 3 into the numerator, and then the same equation forces it into the denominator. The contradiction is not about the size of the number; it is about unique prime divisibility being incompatible with a fraction already reduced to lowest terms.

Problem A3 *Rudin, Chapter 1, Exercise 3*

Prove Rudin's Proposition 1.15: in any field, if $x \neq 0$ then multiplication by x can be cancelled, the multiplicative identity is unique, inverses are unique, and $1/(1/x) = x$.

Solution.

Let F be a field and let $x \in F$ with $x \neq 0$.

For cancellation, suppose $xy = xz$. Multiplying both sides by $1/x$ and using associativity, commutativity, and the identity law gives

$$y = 1y = \big((1/x)x\big)y = (1/x)(xy) = (1/x)(xz) = \big((1/x)x\big)z = 1z = z.$$

This proves: if $x \neq 0$ and $xy = xz$, then $y = z$.

If $xy = x$, then $xy = x \cdot 1$, so cancellation gives $y = 1$. If $xy = 1$, then also $x(1/x) = 1$, so cancellation gives $y = 1/x$. Finally, since $x(1/x) = 1$ and multiplication is commutative, $(1/x)x = 1$; by uniqueness of the inverse of $1/x$, we have

$$\frac{1}{1/x} = x.$$

■

REFLECTIONS.

The proof is the multiplicative twin of Rudin's additive Proposition 1.14. The condition $x \neq 0$ is the only new issue: it lets us multiply by $1/x$, and after that the field axioms reduce every clause to the same cancellation step. Once cancellation is proved, the remaining parts are just special cases: cancel x from $xy = x \cdot 1$, cancel it from $xy = x(1/x)$, and read the last statement as uniqueness of the inverse of $1/x$.

Problem A4 *Rudin, Chapter 1, Exercise 4*

Let E be a nonempty subset of an ordered set; suppose α is a lower bound of E and β is an upper bound of E. Prove that $\alpha \leq \beta$.

SOLUTION.

Because E is nonempty, fix some $t \in E$. Since α is a lower bound of E we have $\alpha \leq t$, and since β is an upper bound of E we have $t \leq \beta$. Chaining these by transitivity of the order, $\alpha \leq t \leq \beta$, gives $\alpha \leq \beta$.

■

REFLECTIONS.

The tell here is that the hypotheses never relate α and β to each other: α is tied only to the elements of E, as a lower bound, and β only to the elements of E, as an upper bound (1.5.1). When two quantities share no direct link but each speaks to a common set, compare them *through* a member of that set. That single idea also picks out the hypothesis that does the work—$E \neq \varnothing$, which is what lets you draw a witness $t \in E$; the rest is reading $\alpha \leq t$ and $t \leq \beta$ off the two conditions and chaining by transitivity (1.3.2).

Nonemptiness is the crux, not a technicality, and you see that by asking what breaks without it. Over the empty set both bound conditions hold vacuously (0.1.6), so *every* pair would qualify, including $\alpha > \beta$—in $\mathbb{R}$ both 1 and 0 vacuously bound $\varnothing$. Arguing by contradiction instead ($\beta < \alpha$ gives $\beta < \alpha \leq t \leq \beta$) spends the same witness and the same nonemptiness, good evidence that the format is cosmetic and the real ingredient is the element of E.

Carry the principle: to compare two things each pinned only to a common set, route the comparison through a witness drawn from the set—and confirm the set is nonempty before you reach in.

Problem A5 *Rudin, Chapter 1, Exercise 5*

Let A be a nonempty set of real numbers which is bounded below. Let $-A$ be the set of all numbers $-x$, where $x \in A$. Prove that $\inf A = -\sup(-A)$.

Solution.

Because A is bounded below it has a lower bound m; write $-A = \{-x : x \in A\}$.

The set $-A$ is nonempty and bounded above: nonempty because A is, and bounded above because negating $m \leq x$ (which reverses the order) gives $-x \leq -m$ for every $x \in A$, so $-m$ is an upper bound. By the least-upper-bound property the number $\beta := \sup(-A)$ exists in $\mathbb{R}$, and we show $\inf A = -\beta$ by checking that $-\beta$ is the greatest lower bound of A.

First, $-\beta$ is a lower bound: for each $x \in A$ we have $-x \in -A$, so $-x \leq \beta$, and negating gives $-\beta \leq x$. Second, it is the greatest one: if γ is any lower bound of A, then $-x \leq -\gamma$ for every $x \in A$, so $-\gamma$ is an upper bound of $-A$; since β is the *least* upper bound, $\beta \leq -\gamma$, and negating gives $\gamma \leq -\beta$.

Thus $-\beta$ is a lower bound of A at least as large as every other, so $\inf A = -\beta = -\sup(-A)$.

■

Reflections.

Two words in the statement fix the whole approach. The first is *infimum*: the only tool in our notes that conjures a supremum or infimum out of a bare set is the least-upper-bound property (1.6.1), so the moment a problem asks you to *produce* an infimum, that property is your destination. The second is the set $-A$, handed to you unbidden—ask why the statement bothers to name the reflection of A through 0. Because negation reverses order (1.3.3), it ought to trade lower bounds for upper bounds and inf for sup, and our completeness tool speaks only of suprema: the statement is quietly telling you to cross to the supremum side, prove the thing there, and cross back. And "prove $\inf A = \cdots$" asks you to *certify* a number as the infimum, whose only certificate is the definition (1.5.3)—a greatest lower bound, two conditions, not one.

Believe it before proving it. With $A = [2, \infty)$ we get $\inf A = 2$, while $-A = (-\infty, -2]$ has supremum -2 and $-(-2) = 2$. Reflecting A across the origin turned its bottom edge into the top edge of $-A$; that one picture is the whole content.

Now watch what each step rests on:

(1) naming a lower bound m is legitimate only because A is bounded below, and passing from $m \leq x$ to $-x \leq -m$ is the order-reversal of 1.3.3, which makes $-m$ an upper bound of $-A$;

(2) with $-A$ nonempty (from $A \neq \varnothing$) and bounded above, the least-upper-bound property (1.6.1; List of Facts) produces $\beta = \sup(-A)$—the one and only use of completeness;

(3) reading the definition of supremum (1.5.2) one way, β bounds $-A$ above, so $-x \leq \beta$

and hence $-\beta \le x$: the easy half, that $-\beta$ is *a* lower bound;

(4) the crux is the other half—for any lower bound γ, the number $-\gamma$ is an upper bound of $-A$, and because β is the *least*, $\beta \le -\gamma$, i.e. $\gamma \le -\beta$. Everything hinges on "least"; drop it and $-\beta$ is merely *a* lower bound, not the infimum.

Two checks finish honestly. Both clauses of "greatest lower bound" (1.5.3) are present, and stopping after the first is the usual way this is half-solved and lost; and each hypothesis is load-bearing—without $A \neq \varnothing$ the set $-A$ is empty with no supremum, and without "bounded below" there is no m and $\inf A$ need not exist. Nothing is circular: we never assumed the infimum existed, we built it from a supremum.

That last fact is the lesson. It is exactly how the notes secure the greatest-lower-bound property with no second axiom—a set bounded below has an infimum because its reflection has a supremum, which is why infimum sits beside supremum at 1.5.3. Keep the technique: when a claim about infima resists you, reflect the set through 0 into a statement about suprema, settle it there, and reflect back. Negation is an order-reversing mirror that swaps inf with sup.

Problem A6 *Rudin, Chapter 1, Exercise 6*

Fix $b > 1$.

(a) If m, n, p, q are integers, $n > 0$, $q > 0$, and $m/n = p/q$, prove that

$$(b^m)^{1/n} = (b^p)^{1/q}.$$

Hence it makes sense to define $b^r = (b^m)^{1/n}$ for $r = m/n$.

(b) Prove that $b^{r+s} = b^r b^s$ if r and s are rational.

(c) If x is real, define $B(x)$ to be the set of all numbers b^t, where t is rational and $t \le x$. Prove that $b^r = \sup B(r)$ when r is rational. Hence it makes sense to define

$$b^x = \sup B(x)$$

for every real x.

(d) Prove that $b^{x+y} = b^x b^y$ for all real x and y.

SOLUTION.

For (a), write $r = m/n = p/q$. Then $mq = np$. Both numbers in question are positive, and their (nq)th powers agree:

$$\left((b^m)^{1/n}\right)^{nq} = b^{mq} = b^{np} = \left((b^p)^{1/q}\right)^{nq}.$$

Uniqueness of positive roots gives $(b^m)^{1/n} = (b^p)^{1/q}$, so b^r is well defined.

For (b), write $r = m/n$ and $s = p/q$. By (a), use the common denominator nq:

$$b^r = (b^{mq})^{1/(nq)}, \qquad b^s = (b^{np})^{1/(nq)}.$$

The root product rule gives

$$b^r b^s = (b^{mq+np})^{1/(nq)} = b^{(mq+np)/(nq)} = b^{r+s}.$$

For (c), first note that rational powers are increasing: if $t < r$, then

$$b^r = b^t b^{r-t} > b^t,$$

because $r - t > 0$ and $b^{r-t} > 1$. Hence every element of $B(r)$ is at most b^r, and $b^r \in B(r)$. Thus b^r is actually the maximum of $B(r)$, so $b^r = \sup B(r)$. For arbitrary real x, the set $B(x)$ is nonempty and bounded above: choose rationals below and above x using density of $\mathbb{Q}$ in $\mathbb{R}$ (1.7.2). Therefore completeness gives $\sup B(x)$, and the definition $b^x = \sup B(x)$ makes sense.

For (d), use rational approximation from below. If $r < x$ and $s < y$ are rational, then

$$b^r b^s = b^{r+s} \le b^{x+y},$$

so taking suprema first over r and then over s gives $b^x b^y \le b^{x+y}$. Conversely, first suppose $t < x + y$ is rational. Choose a rational r with $t - y < r < x$; then $s = t - r$ is rational and $s < y$, so

$$b^t = b^r b^s \le b^x b^y.$$

If $t = x + y$ is rational, the same argument applied to $t - 1/n$ gives $b^{t-1/n} \le b^x b^y$ for every large n. If $b^t > b^x b^y$, put $c = b^t/(b^x b^y) > 1$; the elementary estimate $b^{1/n} \to 1$ gives some n with $b^{1/n} < c$, and then $b^{t-1/n} = b^t/b^{1/n} > b^x b^y$, a contradiction. Thus $b^t \le b^x b^y$ for every rational $t \le x + y$.

Taking the supremum over all such t gives $b^{x+y} \le b^x b^y$. The two inequalities prove $b^{x+y} = b^x b^y$.

■

Reflections.

The exercise builds real exponents by the same completeness pattern used throughout Lecture 1: define the object first on $\mathbb{Q}$, show the rational definition is independent of the chosen fraction, then fill the real gaps with a supremum. Part (d) is the only subtle step. It says the algebraic law survives passage from rationals to real suprema, and the density

of $\mathbb{Q}$ is what lets the rational approximants approach x, y, and $x+y$ from below.

Problem A7 *Rudin, Chapter 1, Exercise 7*

Fix $b > 1$, $y > 0$, and prove that there is a unique real x such that $b^x = y$ by completing Rudin's outline.

SOLUTION.

For every positive integer n,

$$b^n - 1 = (b-1)(1 + b + \cdots + b^{n-1}) \geq n(b-1),$$

which proves (a). Applying this to $b^{1/n}$ gives

$$b - 1 = (b^{1/n})^n - 1 \geq n(b^{1/n} - 1),$$

which is (b). If $t > 1$ and $n > (b-1)/(t-1)$, then

$$b^{1/n} - 1 \leq \frac{b-1}{n} < t - 1,$$

so $b^{1/n} < t$, proving (c).

For (d), suppose $b^w < y$. Put $t = yb^{-w} > 1$. By (c), $b^{1/n} < t$ for all sufficiently large n, and hence

$$b^{w+1/n} = b^w b^{1/n} < b^w t = y.$$

For (e), suppose $b^w > y$. Put $t = b^w/y > 1$. Again $b^{1/n} < t$ for all sufficiently large n, so

$$b^{w-1/n} = b^w b^{-1/n} > y.$$

Let

$$A = \{w \in \mathbb{R} : b^w < y\}.$$

The set A is nonempty and bounded above. Indeed, by (a) and the Archimedean property, positive integer powers of b eventually exceed both y and $1/y$, so some negative integer belongs to A and some positive integer bounds A above. Let $x = \sup A$.

If $b^x < y$, part (d) gives $b^{x+1/n} < y$ for some n, so $x + 1/n \in A$, contradicting that x is an upper bound. If $b^x > y$, part (e) gives $b^{x-1/n} > y$ for some n; monotonicity then makes $x - 1/n$ an upper bound of A, contradicting the definition of x as the least upper bound. Hence $b^x = y$.

Finally, if $x_1 < x_2$, then $x_2 - x_1 > 0$, so $b^{x_2 - x_1} > 1$ and

$$b^{x_2} = b^{x_1} b^{x_2 - x_1} > b^{x_1}.$$

Thus b^x is strictly increasing, and at most one x can satisfy $b^x = y$.

■

Reflections.

This is the logarithm as a supremum argument. The set $A = \{w : b^w < y\}$ records all exponents that are still too small. Parts (a)–(e) prove there is no jump at the boundary: if the supremum exponent were still below y, one could move a little to the right; if it were above y, one could move the upper bound a little to the left. Completeness supplies the boundary point, and monotonicity supplies uniqueness.

Problem A8 *Rudin, Chapter 1, Exercise 8*

Prove that no order can be defined in the complex field that turns it into an ordered field. *Hint:* -1 is a square.

Solution.

Suppose, toward a contradiction, that some total order makes $\mathbb{C}$ an ordered field.

First, every nonzero square is positive. Let $w \neq 0$; by trichotomy $w > 0$ or $w < 0$. If $w > 0$ then $w^2 = w \cdot w > 0$, a product of positives. If $w < 0$ then $-w > 0$, so $(-w)(-w) > 0$; and $(-w)(-w) = w^2$, so again $w^2 > 0$.

Apply this with $w = 1$ to get $1 = 1^2 > 0$, and adding -1 to both sides gives $-1 < 0$. Apply it with $w = i \neq 0$ to get $i^2 > 0$; but $i^2 = -1$, so $-1 > 0$. Then -1 is both positive and negative, contradicting trichotomy. Hence no order turns $\mathbb{C}$ into an ordered field.

■

Reflections.

The hint—-1 is a square—is the proof in disguise, and hearing what such a hint says is the real exercise. In any ordered field the order is rigid: from the axioms (1.3.3) a product of positives is positive, so two facts get pinned down that ordinarily live apart—every nonzero square is positive, and -1 is negative. They collide only if some single number is at once -1 and a square, and the hint supplies exactly that: $-1 = i^2$. Hence the natural shape is contradiction (0.5.5).

The argument has two moving parts:

(1) the lemma "nonzero squares are positive," got by splitting on the sign of w (a clean two-case argument, 0.5.7) and using product-positivity; it gives $1 = 1^2 > 0$, hence $-1 < 0$;

(2) the collision: apply the lemma to $i \neq 0$ to force $i^2 = -1 > 0$, against $-1 < 0$, breaking trichotomy.

Notice the lemma uses only trichotomy and product-positivity, while the collision uses only

that i is a nonzero square root of -1.

The reason to remember the problem is that nothing in it is special to $\mathbb{C}$: in *any* ordered field $1 > 0$ forces $-1 < 0$ while squares stay ≥ 0, so a field becomes unorderable the instant -1 acquires a square root. "Squares are nonnegative" is the master constraint an order imposes; to show a field admits none, exhibit a quantity the order would call both positive and negative.

Problem A9 *Rudin, Chapter 1, Exercise 9*

Suppose $z = a + bi$ and $w = c + di$. Define $z < w$ if $a < c$, and also if $a = c$ but $b < d$. Prove that this turns the set of complex numbers into an ordered set. Does this ordered set have the least-upper-bound property?

SOLUTION.

The relation is lexicographic order. For any two complex numbers $a + bi$ and $c + di$, trichotomy for real numbers gives exactly one of $a < c$, $a = c$, or $c < a$. If $a \neq c$, this decides exactly one of $z < w$ or $w < z$; if $a = c$, trichotomy for b and d decides exactly one of $b < d$, $b = d$, or $d < b$. Thus exactly one of $z < w$, $z = w$, or $w < z$ holds.

Transitivity follows by cases. If $z < w$ and $w < u$, compare first the real parts. A strict increase in the real part persists through the chain; if the real parts are equal throughout, transitivity of the imaginary parts gives the result. Hence this is an order on $\mathbb{C}$.

It does not have the least-upper-bound property. Let

$$E = \{a + bi : a < 0\}.$$

This set is nonempty and is bounded above: every number $0 + ti$ is an upper bound. But there is no least upper bound, because if $0 + ti$ is an upper bound, then $0 + (t - 1)i$ is a smaller upper bound.

■

REFLECTIONS.

Lexicographic order makes $\mathbb{C}$ an ordered set by reading complex numbers like dictionary entries: first compare real parts, and use imaginary parts only to break ties. The failure of completeness comes from the vertical line over real part 0; all of it bounds the left half-plane, and there is no lowest point on that vertical line.

Problem A10 *Rudin, Chapter 1, Exercise 10*

Suppose $z = a + bi$, $w = u + iv$, and

$$a = \left(\frac{|w| + u}{2}\right)^{1/2}, \qquad b = \left(\frac{|w| - u}{2}\right)^{1/2}.$$

Prove that $z^2 = w$ if $v \geq 0$ and that $\overline{z}^2 = w$ if $v \leq 0$. Conclude that every complex number, with one exception, has two complex square roots.

Solution.

The displayed quantities are well defined because $|w| \geq |u|$. By construction,

$$a^2 - b^2 = u.$$

Also,

$$4a^2b^2 = (|w| + u)(|w| - u) = |w|^2 - u^2 = v^2,$$

so $2ab = |v|$. Therefore

$$z^2 = (a + bi)^2 = (a^2 - b^2) + 2abi = u + i|v|.$$

If $v \geq 0$, this is $u + iv = w$. If $v \leq 0$, then

$$\overline{z}^2 = (a - bi)^2 = (a^2 - b^2) - 2abi = u - i|v| = u + iv = w.$$

The exception is $w = 0$, whose only square root is 0. If $w \neq 0$, the construction gives one square root, and its negative gives a second; a quadratic equation over a field has at most two roots, so these are exactly the two square roots.

■

Reflections.

The formula is just the real and imaginary parts of $(a + bi)^2 = u + iv$ solved backward: $a^2 - b^2 = u$ and $2ab = v$. The absolute value decides the sign of the imaginary part. Once one square root of a nonzero complex number is found, the other is its negative, and there cannot be a third.

Problem A11 *Rudin, Chapter 1, Exercise 11*

If z is a complex number, prove that there exists an $r \geq 0$ and a complex number w with $|w| = 1$ such that $z = rw$. Are w and r always uniquely determined by z?

Solution.

If $z \neq 0$, take

$$r = |z|, \qquad w = \frac{z}{|z|}.$$

Then $r > 0$, $|w| = |z|/|z| = 1$, and $rw = z$.

If $z = 0$, take $r = 0$ and any complex number w with $|w| = 1$, for instance $w = 1$; then $z = rw$.

The number r is always unique, since $z = rw$ and $|w| = 1$ imply $|z| = r$. When $z \neq 0$,

$w = z/r$ is also unique. When $z = 0$, $r = 0$ is forced but w can be any unit complex number, so w is not unique.

■

REFLECTIONS.

This is the polar decomposition stripped of angles. The modulus supplies the size r, and division by the size leaves only direction. The only place direction is undefined is the origin, which explains exactly why w fails to be unique only when $z = 0$.

Problem A12 *Rudin, Chapter 1, Exercise 12*

If $z_1, \dots, z_n$ are complex, prove that

$$|z_1 + \cdots + z_n| \le |z_1| + \cdots + |z_n|.$$

SOLUTION.

We prove the claim by induction on n. For $n = 1$ it is equality, and for $n = 2$ it is Rudin's triangle inequality for complex numbers. Suppose the result holds for n terms. Then

$$\begin{aligned} |z_1 + \cdots + z_n + z_{n+1}| &\le |z_1 + \cdots + z_n| + |z_{n+1}| \\ &\le |z_1| + \cdots + |z_n| + |z_{n+1}|. \end{aligned}$$

Thus the inequality holds for every positive integer n.

■

REFLECTIONS.

The two-term triangle inequality is the whole engine. Induction lets the same inequality absorb one new summand at a time, turning a local estimate into a finite one.

Problem A13 *Rudin, Chapter 1, Exercise 13*

If x and y are complex, prove that

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

SOLUTION.

By the triangle inequality,

$$|x| = |(x - y) + y| \le |x - y| + |y|,$$

so $|x| - |y| \le |x - y|$. Swapping x and y gives

$$|y| - |x| \le |x - y|.$$

Together these two inequalities say

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

■

Reflections.

This is the reverse triangle inequality. The direct triangle inequality bounds a whole by its pieces; rearranging it says that changing a complex number by $x - y$ can change its length by no more than $|x - y|$.

Problem A14 *Rudin, Chapter 1, Exercise 14*

If z is a complex number such that $|z| = 1$, compute

$$|1 + z|^2 + |1 - z|^2.$$

Solution.

Since $|z| = 1$, we have $z\overline{z} = 1$. Then

$$|1 + z|^2 = (1 + z)(1 + \overline{z}) = 2 + z + \overline{z}$$

and

$$|1 - z|^2 = (1 - z)(1 - \overline{z}) = 2 - z - \overline{z}.$$

Adding gives

$$|1 + z|^2 + |1 - z|^2 = 4.$$

■

Reflections.

The mixed terms cancel. Geometrically, this is the parallelogram identity in the special case of the two vectors 1 and z on the unit circle: the two squared diagonal lengths add to 4.

Problem A15 *Rudin, Chapter 1, Exercise 15*

Under what conditions does equality hold in the Schwarz inequality?

Solution.

Let

$$A = \sum_{j=1}^{n} |a_j|^2, \qquad B = \sum_{j=1}^{n} |b_j|^2, \qquad C = \sum_{j=1}^{n} a_j \overline{b_j}.$$

In Rudin's proof, if $B > 0$ then

$$\sum_{j=1}^{n} |Ba_j - Cb_j|^2 = B(AB - |C|^2).$$

Equality in Schwarz means $AB = |C|^2$, so the sum of squares on the left is 0. Hence

$$Ba_j = Cb_j \qquad (1 \le j \le n),$$

which says $a_j = (C/B)b_j$ for every j.

If $B = 0$, then every $b_j = 0$ and equality is automatic. Therefore equality holds exactly when one of the two vectors is zero or when there is a complex scalar λ such that

$$a_j = \lambda b_j \qquad (1 \le j \le n).$$

■

REFLECTIONS.

Schwarz becomes an equality precisely when the "error vector" in Rudin's proof vanishes. That is the algebraic version of linear dependence: one vector is a scalar multiple of the other. If one vector is zero, the same description is understood as the degenerate equality case.

Problem A16 *Rudin, Chapter 1, Exercise 16*

Suppose $k \ge 3$, $x, y \in \mathbb{R}^k$, $|x - y| = d > 0$, and $r > 0$. Prove:

(a) If $2r > d$, there are infinitely many $z \in \mathbb{R}^k$ such that $|z - x| = |z - y| = r$.

(b) If $2r = d$, there is exactly one such z.

(c) If $2r < d$, there is no such z.

How must these statements be modified if k is 2 or 1?

SOLUTION.

Let $m = (x+y)/2$ and let $u = (y-x)/d$. Then $|u| = 1$. A point z satisfies $|z-x| = |z-y|$ exactly when $z - m$ is orthogonal to u; geometrically, it lies in the perpendicular bisector hyperplane of the segment from x to y. For such z,

$$|z - x|^2 = |z - m|^2 + \left(\frac{d}{2}\right)^2.$$

Thus the equations $|z-x| = |z-y| = r$ are equivalent to

$$z - m \perp u, \qquad |z-m|^2 = r^2 - \frac{d^2}{4}.$$

If $2r > d$, the right-hand side is positive. In dimension $k \geq 3$, the perpendicular subspace has dimension $k-1 \geq 2$, so its sphere of that positive radius contains infinitely many points. If $2r = d$, the radius is 0, so $z = m$ is the unique point. If $2r < d$, the required squared radius is negative, so no point exists.

For $k = 2$, the perpendicular subspace is one-dimensional: if $2r > d$ there are exactly two points, if $2r = d$ one point, and if $2r < d$ none. For $k = 1$, the perpendicular subspace is zero-dimensional: there is one point when $2r = d$ and no point otherwise.

■

Reflections.

The problem is the intersection of two equal-radius spheres centered at x and y. The midpoint separates the geometry from the algebra: first move to the perpendicular bisector, then ask for a sphere of radius $\sqrt{r^2 - d^2/4}$ inside that perpendicular space. The dimension of that perpendicular space is why the answer changes for $k = 2$ and $k = 1$.

Problem A17 *Rudin, Chapter 1, Exercise 17*

Prove that

$$|x+y|^2 + |x-y|^2 = 2|x|^2 + 2|y|^2$$

if $x, y \in \mathbb{R}^k$. Interpret this geometrically, as a statement about parallelograms.

Solution.

Using the inner product,

$$|x+y|^2 = (x+y)\cdot(x+y) = |x|^2 + 2x\cdot y + |y|^2$$

and

$$|x-y|^2 = (x-y)\cdot(x-y) = |x|^2 - 2x\cdot y + |y|^2.$$

Adding cancels the cross terms and gives

$$|x+y|^2 + |x-y|^2 = 2|x|^2 + 2|y|^2.$$

Geometrically, x and y span a parallelogram whose diagonals are $x+y$ and $x-y$; the sum of the squares of the diagonals equals the sum of the squares of the four sides.

■

Reflections.

The identity is algebraic cancellation of the cross terms $2x \cdot y$ and $-2x \cdot y$. Its geometry is that the two diagonals of a parallelogram encode the same total squared length as the four sides.

Problem A18 *Rudin, Chapter 1, Exercise 18*

If $k \geq 2$ and $x \in \mathbb{R}^k$, prove that there exists $y \in \mathbb{R}^k$ such that $y \neq 0$ but $x \cdot y = 0$. Is this also true if $k = 1$?

SOLUTION.

If $x = 0$, take any nonzero y. If $x \neq 0$, choose an index j with $x_j \neq 0$ and choose $\ell \neq j$. Define y by

$$y_j = -x_\ell, \qquad y_\ell = x_j, \qquad y_i = 0 \quad (i \neq j, \ell).$$

Then $y \neq 0$, since $y_\ell = x_j \neq 0$, and

$$x \cdot y = x_j(-x_\ell) + x_\ell x_j = 0.$$

For $k = 1$ this is not true for every x: if $x \neq 0$, then $x \cdot y = xy = 0$ forces $y = 0$.

■

REFLECTIONS.

In dimension at least two there is room to turn one nonzero coordinate into a perpendicular direction. In one dimension there is no sideways direction, so only the zero vector is orthogonal to a nonzero vector.

Problem A19 *Rudin, Chapter 1, Exercise 19*

Suppose $a, b \in \mathbb{R}^k$. Find $c \in \mathbb{R}^k$ and $r > 0$ such that

$$|x - a| = 2|x - b|$$

if and only if $|x - c| = r$. Rudin gives the solution $3c = 4b - a$, $3r = 2|b - a|$.

SOLUTION.

Assume first that $a \neq b$, so the displayed r is positive. Squaring the equation gives

$$|x - a|^2 = 4|x - b|^2.$$

Expanding with the inner product,

$$|x|^2 - 2a \cdot x + |a|^2 = 4|x|^2 - 8b \cdot x + 4|b|^2.$$

Move all terms to one side and complete the square with

$$c = \frac{4b - a}{3}.$$

A direct simplification gives

$$|x - a|^2 - 4|x - b|^2 = -3|x - c|^2 + \frac{4}{3}|b - a|^2.$$

Thus $|x - a| = 2|x - b|$ is equivalent to

$$|x - c|^2 = \frac{4}{9}|b - a|^2,$$

or $|x - c| = r$, where

$$r = \frac{2}{3}|b - a|.$$

If $a = b$, the original equation reduces to $|x - a| = 2|x - a|$, hence $x = a$; this is the degenerate case of the same formula with $r = 0$.

■

Reflections.

This is an Apollonius sphere: the points whose distance from a is twice their distance from b form a sphere with center shifted past b away from a. The computation is just completing the square after expanding the two squared distances.

Problem A20 *Rudin, Chapter 1, Exercise 20*

With reference to Rudin's Appendix, suppose property (III) were omitted from the definition of a cut. Keep the same definitions of order and addition. Show that the resulting ordered set has the least-upper-bound property, that addition satisfies axioms (A1) to (A4) with a slightly different zero-element, but that (A5) fails.

Solution.

Call the modified objects *weak cuts*: nonempty proper subsets of $\mathbb{Q}$ that are downward closed, but may have a largest element.

The least-upper-bound proof from Rudin's Appendix still works. If $\mathcal{A}$ is a nonempty family of weak cuts bounded above by a weak cut β, then

$$\gamma = \bigcup_{\alpha \in \mathcal{A}} \alpha$$

is nonempty, proper because $\gamma \subseteq \beta$, and downward closed. Hence γ is a weak cut. It is plainly an upper bound of $\mathcal{A}$, and any smaller weak cut fails to contain some element of some $\alpha \in \mathcal{A}$, so it is not an upper bound. Thus $\gamma = \sup \mathcal{A}$.

Addition is still defined by

$$\alpha + \beta = \{r + s : r \in \alpha,\ s \in \beta\}.$$

The sum of two weak cuts is again a weak cut, and commutativity and associativity come from the corresponding laws in $\mathbb{Q}$. The additive identity is now

$$0^\dagger = \{q \in \mathbb{Q} : q \leq 0\},$$

not Rudin's original open cut $\{q : q < 0\}$. Indeed, if α is downward closed, then $\alpha + 0^\dagger \subseteq \alpha$, while $p = p + 0 \in \alpha + 0^\dagger$ for every $p \in \alpha$. Thus axioms (A1)–(A4) hold.

Axiom (A5) fails. Let

$$\alpha = \{q \in \mathbb{Q} : q < 0\}.$$

If some weak cut β satisfied $\alpha + \beta = 0^\dagger$, then $0 \in \alpha + \beta$, so there would be $r < 0$ and $s \in \beta$ with $r + s = 0$. Thus $s = -r > 0$. But then $-s/2 \in \alpha$, so

$$(-s/2) + s = s/2 > 0$$

would belong to $\alpha + \beta$, contradicting $\alpha + \beta = 0^\dagger$. Hence this α has no additive inverse.

■

REFLECTIONS.

The no-largest-element condition is exactly what separates open cuts from rational endpoint cuts. If it is removed, completeness by unions survives and addition still has an identity, but the identity becomes the closed cut at 0. The old open zero-cut then has no additive inverse, so the field structure breaks at (A5).

Problem B *Induction: Divisibility by* 3

Prove by induction that $n^3 + 2n$ is divisible by 3 for every integer $n \geq 0$.

SOLUTION.

For $n \geq 0$ let $P(n)$ be the statement $3 \mid n^3 + 2n$. The base case $n = 0$ holds, since $0^3 + 2 \cdot 0 = 0 = 3 \cdot 0$.

Assume $P(n)$, say $n^3 + 2n = 3k$ with $k \in \mathbb{Z}$. Then

$$\begin{aligned}(n+1)^3 + 2(n+1) &= (n^3 + 3n^2 + 3n + 1) + (2n + 2) \\ &= (n^3 + 2n) + 3n^2 + 3n + 3 \\ &= 3k + 3\,(n^2 + n + 1) = 3\,(k + n^2 + n + 1),\end{aligned}$$

and $k + n^2 + n + 1 \in \mathbb{Z}$, so $3 \mid (n+1)^3 + 2(n+1)$, i.e. $P(n+1)$. By induction $P(n)$ holds

for all $n \geq 0$.

■

Reflections.

"Divisible by 3 for every n" is indexed by the naturals with successive cases linked by a small algebraic change—the signature of induction (1.1.3, 0.6.2). But induction pays off only if the step *reuses* the previous case, so the genuine task is algebraic: write $f(n+1) = (n+1)^3 + 2(n+1)$ as $f(n)$ plus a remainder you can see is a multiple of 3.

Doing that exposes the increment $3n^2 + 3n + 3 = 3(n^2 + n + 1)$, and now the hypothesis in usable form ($f(n) = 3k$ for an actual integer k, not the vague "3 divides it") adds to a multiple of 3 and keeps $f(n+1)$ divisible. That increment is the entire engine.

Seeing the same statement without induction explains *why* it is true, not just *that* it is: $n^3 + 2n = (n-1)\,n\,(n+1) + 3n$, and among three consecutive integers one is a multiple of 3. (The notes prove the cousin $3 \mid n^3 - n$ the identical way at 0.6.8—worth reading side by side.)

The habit to carry: to prove $p \mid f(n)$ by induction, never stare at $f(n+1)$ whole—compute the increment $f(n+1) - f(n)$, check it is a multiple of p, and divisibility is inherited from one case to the next.

Problem C *Induction: An Exponential Inequality*

Prove by induction that $2^n > n^2$ for every integer $n > 4$.

Solution.

Induct on n, starting at $n = 5$, the least integer greater than 4. The base case holds: $2^5 = 32 > 25 = 5^2$.

Assume $2^n > n^2$ for some $n \geq 5$. Doubling (multiplication by $2 > 0$ preserves the inequality), $2^{n+1} = 2 \cdot 2^n > 2n^2$. Moreover $2n^2 \geq (n+1)^2$ for $n \geq 5$, since

$$2n^2 - (n+1)^2 = n^2 - 2n - 1 = (n-1)^2 - 2 \geq (5-1)^2 - 2 = 14 > 0.$$

Combining, $2^{n+1} > 2n^2 \geq (n+1)^2$, which is the statement at $n+1$. By induction $2^n > n^2$ for every $n > 4$.

■

Reflections.

Two features flag the method. "For every integer $n > 4$" is an inductive shape (1.1.3), and the oddly specific cutoff $n > 4$ is a clue in its own right: it says the inequality is *false* below 5 and the induction must begin exactly where the statement first holds for good. It does—$n = 2, 3, 4$ give $4 = 4$, $8 < 9$, $16 = 16$, and only from $n = 5$ onward is it true, the same result the notes establish at 0.6.7.

The step's idea is that an exponential, once doubled, overshoots a polynomial: $2^{n+1} = 2 \cdot 2^n > 2n^2$, and $2n^2$ already clears $(n+1)^2$. So the inductive step reduces to a purely polynomial inequality, $2n^2 \geq (n+1)^2$, i.e. $(n-1)^2 \geq 2$—which you must check across the *whole* range $n \geq 5$, not merely at the base, or the chain could snap somewhere up the line. The hypothesis carries the exponential growth; a side inequality, valid throughout, bridges $2n^2$ to $(n+1)^2$.

For any exponential-beats-polynomial claim the recipe is the same: double the hypothesis, reduce the step to a polynomial inequality, confirm it on the entire tail, and anchor the base case where the statement first becomes permanently true.

Problem D *An Equivalence Relation on the Rationals*

To define $\mathbb{Q}$, we declared $\frac{p}{q} \sim \frac{r}{s} \iff ps = rq$ in $\mathbb{Z}$. Show this relation is reflexive, symmetric, and transitive — an equivalence relation.

SOLUTION.

The relation acts on symbols $\frac{p}{q}$ with $p, q \in \mathbb{Z}$ and $q \neq 0$, by $\frac{p}{q} \sim \frac{r}{s} \iff ps = rq$.

Reflexivity is $pq = pq$, which is exactly $\frac{p}{q} \sim \frac{p}{q}$. Symmetry is immediate: $ps = rq$ reversed is $rq = ps$, i.e. $\frac{r}{s} \sim \frac{p}{q}$.

For transitivity, suppose $\frac{p}{q} \sim \frac{r}{s}$ and $\frac{r}{s} \sim \frac{m}{n}$, so $ps = rq$ and $rn = ms$. Multiplying the first by n and the second by q, $psn = rqn$ and $rnq = msq$, and since $rqn = rnq$ the right-hand sides agree, giving $psn = msq$, that is $s(pn) = s(mq)$. As $s \neq 0$ and $\mathbb{Z}$ has no zero divisors, we cancel s to get $pn = mq$, i.e. $\frac{p}{q} \sim \frac{m}{n}$. Being reflexive, symmetric, and transitive, $\sim$ is an equivalence relation.

■

REFLECTIONS.

On its face this asks you to check the three clauses of an equivalence relation (0.4.16), but the deeper clue is *which* relation: the very rule our construction of $\mathbb{Q}$ rests on (1.2.2), deciding when two fraction symbols name the same rational. So it is not a drill but the well-definedness check that makes $\mathbb{Q}$ a legitimate object—and the reason one rational has many names (1.2.3).

That reframing locates the content. Reflexivity and symmetry fall straight out of equality of integers ($pq = pq$; and $ps = rq$ reversed is $rq = ps$), so everything real lives in transitivity, where the shared symbol $\frac{r}{s}$ must be eliminated. From $ps = rq$ and $rn = ms$ you cannot chain directly, so you manufacture a common factor—multiply to reach $s(pn) = s(mq)$—and cancel s.

That cancellation is the whole game, and it is legal only because $s \neq 0$ and $\mathbb{Z}$ is an integral domain; it is exactly the step that consumes the nonzero-denominator convention. Allow

$s = 0$ and transitivity can fail, which is precisely why $\mathbb{Q}$ is built from fractions with nonzero denominators (1.2.2).

The lesson generalises: an equation-defined relation is typically shown transitive by combining its two defining equations into one that shares a factor and then cancelling, and such cancellation is licensed in an integral domain exactly when the shared factor is nonzero.

Problem E *An Inequality with Square Roots*

Prove that if real numbers a, b, c satisfy $ab = c$ then either $a \le \sqrt{c}$ or $b \le \sqrt{c}$.

Solution.

The symbol $\sqrt{c}$ is defined only for $c \ge 0$, so that is the domain of the statement (for $c < 0$ it is not well-posed, and $ab = c$ alone permits $c < 0$, e.g. $a = 1$, $b = -1$). Write $\sqrt{c} \ge 0$ for the nonnegative root, so $(\sqrt{c})^2 = c$.

Suppose, toward a contradiction, that $ab = c$ yet $a > \sqrt{c}$ and $b > \sqrt{c}$—the negation of the conclusion. Since $\sqrt{c} \ge 0$, both a and b are positive. Multiplying $a > \sqrt{c}$ by $b > 0$ gives $ab > \sqrt{c}\,b$, and multiplying $b > \sqrt{c}$ by $\sqrt{c} \ge 0$ gives $\sqrt{c}\,b \ge (\sqrt{c})^2 = c$. Hence $ab > c$, contradicting $ab = c$. Therefore $a \le \sqrt{c}$ or $b \le \sqrt{c}$.

■

Reflections.

The conclusion is a disjunction, and disjunctions resist a head-on attack because you cannot say in advance which alternative holds; the "or" is itself the clue to negate. Assuming *both* $a > \sqrt{c}$ and $b > \sqrt{c}$ collapses the two unknowns into one usable conjunction—the proving-disjunctions and negation strategy of 0.5.8 and 0.5.11, on the connective laws of 0.1.3.

A quieter clue comes first: a radical $\sqrt{c}$ appears, defined only for $c \ge 0$, so that is the genuine domain of the problem rather than an extra hypothesis ($ab = c$ by itself allows $c < 0$). Pinning the domain hands you $\sqrt{c} \ge 0$ and $(\sqrt{c})^2 = c$, the facts the argument leans on.

Now the negation does the work: $a, b > \sqrt{c} \ge 0$ makes both positive, which is what licenses multiplying the two inequalities in an ordered field (1.3.3), yielding $ab > (\sqrt{c})^2 = c$ against $ab = c$. The picture to keep is a threshold—$\sqrt{c}$ is a gate, and if a product equals c its two factors cannot both stand strictly above the gate. The $\le$ cannot be tightened to $<$, as $a = b = \sqrt{c}$ shows, so equality must be allowed.

Two reusable moves fall out: to prove "P or Q," assume $\neg P$ and $\neg Q$ together and hunt for a contradiction; and to weigh a product against a threshold T, compare each factor against $\sqrt{T}$.

Problem F *Rudin, Chapter 2, Exercise 2*

A complex number z is *algebraic* if there are integers $a_0, \dots, a_n$, not all zero, with $a_0 z^n + a_1 z^{n-1} + \cdots + a_n = 0$. Prove that the set of all algebraic numbers is countable. *Hint:* for each N there are finitely many equations with $n + |a_0| + \cdots + |a_n| = N$.

SOLUTION.

To each integer polynomial $a_0 z^n + \cdots + a_n$ (coefficients not all zero) assign the *height*

$$N = n + |a_0| + |a_1| + \cdots + |a_n| \geq 1.$$

Fix $N \geq 1$. A height-N polynomial has $n \leq N$ and $|a_i| \leq N$, so there are at most $N+1$ choices for n and at most $(2N+1)^{n+1}$ for the coefficients; hence only finitely many integer polynomials have height N. Each is nonzero of degree $\leq N$, so it has at most N roots in $\mathbb{C}$. Therefore the set A_N of roots of all height-N polynomials is a finite union of finite sets, hence finite.

Every algebraic number is a root of some integer polynomial, of some height $N \geq 1$, so it lies in A_N; conversely each element of every A_N is algebraic. Thus $\mathbb{A} = \bigcup_{N=1}^{\infty} A_N$ is a countable union of finite sets, so it is at most countable. Finally $\mathbb{A}$ is infinite, since every integer k is a root of $z - k$; an infinite, at-most-countable set is countable, so $\mathbb{A}$ is countable.

■

REFLECTIONS.

"Prove this set is countable" has, in our notes, essentially one playbook: present the set as a countable union of finite or countable pieces—the quotable "a countable union of countable sets is countable", with "a subset of a countable set is finite or countable" (2.1.9) in support. The art is choosing the slicing, and the hint points straight at it: organise the polynomials by an integer *height* $N = n + \sum |a_i|$. Why a height is the right device is the part to understand—it must be *finite-to-one*, only finitely many objects sharing each value, and this one is, since N simultaneously caps the degree ($n \leq N$) and every coefficient ($|a_i| \leq N$).

From there the steps are short, each closed by a specific fact:

(1) finitely many polynomials per height, and a nonzero degree-n polynomial has at most n roots, so each height contributes only finitely many algebraic numbers;

(2) assembling over all heights gives $\bigcup_N A_N$, a countable union of finite sets, hence at most countable;

(3) "at most countable" upgrades to "countable" only once you note the set is infinite—every integer k is a root of $z - k$, and $\mathbb{Z}$ is countable (2.1.4).

> That last step is the easy one to forget. And the transferable idea is the height itself: anything assembled from finite integer data—polynomials, finite words over a finite alphabet, finite sequences—can be sliced into finite layers by a well-chosen integer "size" with finite fibres, and then "countable union of countable sets" finishes the job.

HOMEWORK 2

The Topology of Metric Spaces

MATH S4061 | Introduction to Modern Analysis I

Rudin Chapter 2 — Limit Points, Compactness, and Perfect, Connected, and Separable Sets

Problems in This Set

Problem 1 *Rudin, Chapter 2, Exercise 1*

Prove that the empty set is a subset of every set.

Solution.

Let A be any set. The claim $\varnothing \subseteq A$ unfolds, by the definition of subset, into the implication

$$x \in \varnothing \quad \Longrightarrow \quad x \in A,$$

asserted for every x. Its hypothesis $x \in \varnothing$ is false for every x, since $\varnothing$ has no elements. An implication with a false hypothesis is true, so the implication holds for every x, with no instance left to check. Hence $\varnothing \subseteq A$, and since A was arbitrary, the empty set is a subset of every set.

■

Reflections.

The word that fixes the strategy is *subset*. A claim "$\varnothing \subseteq A$" is not a claim about a single object but a universal implication—"for every x, if $x \in \varnothing$ then $x \in A$"—which is exactly how a subset statement is built (0.3.6), and the standard way to prove one is to take an arbitrary element of the left-hand set and show it lies in the right (0.5.10). The second clue is the set on the left: it is $\varnothing$, the one set with no elements at all (0.3.4). The reflex those two clues should trigger is to look at the *hypothesis* of the implication rather than its conclusion, because that hypothesis is the thing the empty set makes impossible.

It helps to believe it first on a small case. Asking whether $\varnothing \subseteq \{1, 2\}$ means asking whether every element of $\varnothing$ lies in $\{1, 2\}$; there is no element of $\varnothing$ to inspect, so there is nothing that could go wrong, and the answer is yes by default. The same picture works for any target set, which is already the whole theorem.

The argument is short, and each move rests on the one before:

(1) fix an arbitrary set A and unfold $\varnothing \subseteq A$ into the implication "$x \in \varnothing \Rightarrow x \in A$" for every x, the only thing a subset claim means (0.3.6);

(2) read the hypothesis $x \in \varnothing$: it is false for every x, because the empty set has no elements (0.3.4);

(3) conclude the implication is true, since an implication with a false hypothesis is true—this is vacuous truth (0.1.6, on the definition of implication 0.1.5), and it is the single load-bearing step;

(4) since A was named with no special property, the conclusion holds for every set.

The rigor check is to see that there is genuinely no gap rather than a missing case. The proof never produces an element of $\varnothing$—it cannot, and that is the point; it argues that the implication to be verified has no false instance, so it is true with nothing to verify. Drop the feature that $\varnothing$ is empty and the reasoning collapses at once: if the left-hand set had even one element outside A, the implication would have a true hypothesis and a false conclusion, and the subset claim would fail. So emptiness is exactly the hypothesis doing the work, which is why the result records a basic property of $\varnothing$ alone (0.3.8).

The transferable move is how to handle any universal statement quantified over a set that might be empty: it is automatically true there, because a "for all" over nothing has no counterexample. This is the same vacuous-truth reflex that lets a proof safely open "take an arbitrary element of E" even when E could turn out to be empty (0.3.7)—the empty case needs no separate argument, it is handled for free.

Problem 2 *Rudin, Chapter 2, Exercise 2*

A complex number z is said to be *algebraic* if there are integers $a_0, \dots, a_n$, not all zero, such that

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0.$$

Prove that the set of all algebraic numbers is countable. *Hint:* For every positive integer N there are only finitely many equations with

$$n + |a_0| + |a_1| + \cdots + |a_n| = N.$$

SOLUTION.

To each integer polynomial $a_0 z^n + a_1 z^{n-1} + \cdots + a_n$ with coefficients not all zero, assign the *height*

$$N = n + |a_0| + |a_1| + \cdots + |a_n| \geq 1.$$

Fix a positive integer N. A polynomial of height N has degree $n \leq N$, and every coefficient satisfies $|a_i| \leq N$; so there are at most $N+1$ choices for n and at most $(2N+1)^{n+1}$ choices for the coefficients. Hence only finitely many integer polynomials have height N. Each such polynomial is nonzero of degree $\leq N$, so it has at most N complex roots. Let A_N be the set of all roots of all height-N polynomials; being a finite union of finite sets, A_N is finite.

Every algebraic number is a root of some integer polynomial, which has some height $N \geq 1$, so it lies in the corresponding A_N; conversely every element of every A_N is algebraic.

Writing $\mathbb{A}$ for the set of algebraic numbers,

$$\mathbb{A} = \bigcup_{N=1}^{\infty} A_N,$$

a countable union of finite sets, which is at most countable. Finally $\mathbb{A}$ is infinite, since each integer k is algebraic as the root of $z - k$, and $\mathbb{Z}$ is countably infinite (2.1.4). An infinite, at-most-countable set is countable, so $\mathbb{A}$ is countable.

■

Reflections.

The phrase *prove that this set is countable* fixes the strategy before anything else: in our notes the only practical way to certify countability is to display the set as a countable union of finite or countable pieces—the quotable "a countable union of countable sets is countable", supported by "a subset of a countable set is finite or countable" (2.1.9). What that pushes the work onto is the *slicing*: into which finite or countable pieces should the algebraic numbers be cut? Trying to list the roots directly is hopeless, because there is no natural order on $\mathbb{C}$ and no first algebraic number to start from.

The hint names the slicing and is worth taking apart, since it teaches the whole technique. An algebraic number is not a free-standing object; it is manufactured by a polynomial with integer coefficients, and a polynomial is finite integer data—a degree together with a finite list of coefficients. Whenever objects are built from finite integer data, the move is to attach to that data a single integer *size* that is finite-to-one, meaning only finitely many objects share each value. The height $N = n + \sum_i |a_i|$ is exactly such a size, because fixing N simultaneously caps the degree ($n \leq N$) and every coefficient ($|a_i| \leq N$); that simultaneous cap is the reason only finitely many equations have a given height, and it is the one thing to verify rather than assert.

Believe it on a small case first. Height $N = 2$ forces $n + \sum |a_i| = 2$ with the coefficients not all zero; the polynomials are $\pm z$, ± 2 (degree 0), and a few low-degree pieces, a finite list one can write out by hand, and each contributes only its handful of roots. Stacking these finite layers $N = 1, 2, 3, \ldots$ builds the whole set from the bottom up.

The proof then runs in three short moves, each closed by a named result:

(1) each height gives only finitely many polynomials, and a nonzero polynomial of degree n has at most n complex roots, so the height-N roots form a finite set A_N—this is where the hint is spent, and the root bound is the only fact from outside topology that the argument needs;

(2) gathering over all heights writes $\mathbb{A} = \bigcup_{N \geq 1} A_N$, a countable union of finite sets, hence at most countable by the countable-union fact;

(3) "at most countable" becomes "countable" only after observing the set is infinite, and the cheapest witness is that every integer k is the root of $z - k$, so $\mathbb{A} \supseteq \mathbb{Z}$ is infinite (2.1.4).

The rigor check is mostly that third step, which is the easy one to drop: a countable union of finite sets is only guaranteed to be *at most* countable, so without an explicit infinite subset one has proved less than "countable." The other hidden hypothesis is that the polynomials are taken nonzero ("not all zero"); the zero polynomial vanishes everywhere and would drag all of $\mathbb{C}$ into a single layer, destroying the finiteness of A_N—which is exactly why the definition excludes it.

The transferable move is the height itself. Anything assembled from finite integer data—polynomials, finite words over a finite alphabet, finite sequences of integers—can be sorted into finite layers by a well-chosen integer size with finite fibres, after which the countable-union fact finishes the count. This same template returns in Problem 3, where the countability proved here is played against the uncountability of $\mathbb{R}$ ("$\mathbb{R}$ is uncountable") to force the existence of transcendental reals.

Problem 3 *Rudin, Chapter 2, Exercise 3*

Prove that there exist real numbers which are not algebraic.

SOLUTION.

A complex number is *algebraic* when it is a root of some nonzero integer polynomial; write $\mathbb{A}$ for the set of all algebraic numbers and $\mathbb{A}_{\mathbb{R}} = \mathbb{A} \cap \mathbb{R}$ for the real ones.

The set $\mathbb{A}$ is countable: this is Problem F (Rudin, Chapter 2, Exercise 2), which sorts the integer polynomials by the height $N = n + |a_0| + \cdots + |a_n|$, finds only finitely many of each height and at most N roots apiece, and assembles $\mathbb{A}$ as a countable union of finite sets. Hence $\mathbb{A}_{\mathbb{R}}$, a subset of $\mathbb{A}$, is finite or countable (subset of a countable set; 2.1.9).

Suppose, toward a contradiction, that every real number were algebraic. Then $\mathbb{R} = \mathbb{A}_{\mathbb{R}}$ would be at most countable. But $\mathbb{R}$ is uncountable ($\mathbb{R}$ is uncountable; 2.1.13), and an uncountable set cannot equal an at-most-countable one. This contradiction shows some real number lies outside $\mathbb{A}$, i.e. there exist real numbers that are not algebraic.

■

REFLECTIONS.

The statement asks only that a non-algebraic real *exist*, naming no candidate and demanding no formula—and a pure existence claim with no constructive thread is the classic signal for a counting argument by contradiction (0.5.5): rather than exhibit a transcendental number, show that the algebraic ones are too few to fill $\mathbb{R}$. The two halves of that strategy are already on the page. "Algebraic" was the subject of the previous problem, where it was shown to be a *countable* property, and "real numbers" carries the one fact

about $\mathbb{R}$ that beats countability head-on—its uncountability ($\mathbb{R}$ is uncountable, 2.1.13). A countable family inside an uncountable set cannot exhaust it, so something must be left over.

Believe it by sizes before proving it. The algebraic numbers can be listed $\alpha_1, \alpha_2, \alpha_3, \ldots$ (every rational, $\sqrt{2}$, the golden ratio, and so on), while the reals provably cannot be put in any such list; a list cannot contain everything off a longer-than-any- list line, so reals outside the list survive. Those survivors are exactly the transcendental numbers.

The argument is three short moves, each leaning on a named prior result:

(1) $\mathbb{A}$ is countable—this is not reproved here but imported wholesale from Problem F, whose height argument writes $\mathbb{A}$ as a countable union of finite root-sets;

(2) the real algebraic numbers $\mathbb{A}_{\mathbb{R}} = \mathbb{A} \cap \mathbb{R}$ form a subset of a countable set, hence are at most countable (subset of a countable set, 2.1.9);

(3) assuming every real were algebraic makes $\mathbb{R} = \mathbb{A}_{\mathbb{R}}$ at most countable, which collides with the uncountability of $\mathbb{R}$ (2.1.13)—the contradiction that forces a non-algebraic real.

The reasoning is honest because the two ingredients are genuinely independent and neither is assumed of the conclusion: countability of $\mathbb{A}$ comes from organising polynomials, uncountability of $\mathbb{R}$ from Cantor's diagonal (2.1.12), and nothing here presumes a transcendental number in advance—it is produced, not posited. Every hypothesis earns its place: drop the countability of $\mathbb{A}$ and the comparison is empty, drop the uncountability of $\mathbb{R}$ and a countable line could be all-algebraic (as the rational line nearly is). Restricting to $\mathbb{A}_{\mathbb{R}}$ rather than all of $\mathbb{A}$ matters too, since the comparison must happen inside $\mathbb{R}$.

The move to keep is the cardinality existence proof: to show an object with some property exists, prove the objects *lacking* it are at most countable and then drop them into an uncountable ambient set—the overflow is your witness. It is the same accounting that makes $\mathbb{Q}$ thin in $\mathbb{R}$ (2.1.14); here it conjures transcendentals without ever building one, which is why exhibiting a specific transcendental (Liouville, e, π) is a strictly harder and separate task.

Problem 4 *Rudin, Chapter 2, Exercise 4*

Is the set of all irrational real numbers countable?

Solution.

No. The set $\mathbb{R} \setminus \mathbb{Q}$ of irrational reals is uncountable.

The argument rests on two facts: $\mathbb{Q}$ is countable (List of Facts; 2.1.7) and $\mathbb{R}$ is uncountable (List of Facts; 2.1.13). Suppose, toward a contradiction, that $\mathbb{R} \setminus \mathbb{Q}$ were countable. Every

real is rational or irrational, so

$$\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$$

would be a union of two countable sets, hence countable ("a countable union of countable sets is countable"). That contradicts the uncountability of $\mathbb{R}$, so $\mathbb{R} \setminus \mathbb{Q}$ is uncountable.

It remains to see why $\mathbb{R}$ is uncountable, the one input above worth reproving. It is enough to show the subset $(0, 1)$ cannot be listed, since a countable $\mathbb{R}$ would force the subset $(0, 1)$ to be at most countable ("a subset of a countable set is finite or countable"). Suppose $(0, 1)$ were listed as $x_1, x_2, x_3, \ldots$, each written in decimal as $x_n = 0.d_{n1}d_{n2}d_{n3}\cdots$. Build $y = 0.e_1e_2e_3\cdots$ along the diagonal by

$$e_n = \begin{cases} 4 & \text{if } d_{nn} \neq 4, \\ 7 & \text{if } d_{nn} = 4. \end{cases}$$

Every digit of y is a 4 or a 7, so y avoids the ambiguous tails $0.\cdots 999\cdots$ and $0.\cdots 000\cdots$ and thus has a single decimal expansion. That expansion differs from x_n in the nth digit, so $y \neq x_n$ for every n; yet $y \in (0, 1)$, so the list omits a point of $(0, 1)$. The list was supposed to be exhaustive, a contradiction. Hence $(0, 1)$ is uncountable, and so is $\mathbb{R} \supseteq (0, 1)$.

■

REFLECTIONS.

The question asks whether the irrationals can be threaded onto a single list $a_1, a_2, a_3, \ldots$ the way $\mathbb{Z}$ can—whether there are only countably many of them. The word *countable* names the lecture object (2.1.2), and the honest first reaction is that listing the irrationals *directly* is hopeless: they carry no first element and no successor rule, nothing to enumerate along. That obstruction is itself the clue. When a set resists a head-on count, the move is to place it inside a set you understand and compare sizes, and the irrationals come pre-packaged as the part of $\mathbb{R}$ left over from $\mathbb{Q}$.

Believe the splitting before proving it. Each real is rational or irrational and never both, so $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ partitions the line into the listable part and the part in question. If the leftover were also listable, interleaving the two lists would enumerate all of $\mathbb{R}$—and that is exactly what cannot happen.

The proof is a short contradiction (0.5.5) resting on two quotable facts and one diagonal core:

(1) assume $\mathbb{R} \setminus \mathbb{Q}$ is countable; since $\mathbb{Q}$ is countable as well (List of Facts; 2.1.7), the union $\mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ is a union of two countable sets, hence countable by "a countable union of countable sets is countable";

(2) that union is all of $\mathbb{R}$, so $\mathbb{R}$ would be countable, against "$\mathbb{R}$ is uncountable" (2.1.13)—

the collision that closes the contradiction;

(3) the only input not to be taken on faith is the uncountability of $\mathbb{R}$, secured by Cantor's diagonal argument on $(0,1)$ (2.1.12), the decimal cousin of the binary statement "the 0–1 sequences are uncountable" (2.1.10).

The diagonal step is the one to actually understand. Given any candidate list of $(0,1)$ written as decimals, read down the diagonal and change each digit; the number you build disagrees with the nth entry in the nth place, so it lies in $(0,1)$ yet sits on no row—the list was incomplete. The lone subtlety is that a few reals wear two decimal coats ($0.4999\cdots = 0.5000\cdots$), which would let a "new" number secretly equal a listed one; restricting the manufactured digits to 4 and 7 dodges both ambiguous tails, so the constructed y has a unique expansion and the disagreement is real. Drop that precaution and the argument genuinely has a gap, which is why it is built in rather than waved past.

The rigor turns on the asymmetry between the two facts: countability is preserved by the union of *two* pieces, but only because finitely (indeed countably) many countable sets stay countable; were $\mathbb{R}\setminus\mathbb{Q}$ countable, $\mathbb{R}$ would inherit that, and the diagonal forbids it. Nothing is circular—the uncountability of $\mathbb{R}$ is proved independently of any claim about the irrationals.

The transferable move is the whole point: to size a set that defies direct enumeration, write the ambient set as that set together with a countable remainder, and let a known uncountability do the work. The same accounting shows the transcendentals are uncountable (the algebraic numbers being only countable, as in Problem F); a countable exception cannot dent an uncountable whole, so the strange numbers are the overwhelming majority, even when not one is easy to name.

Problem 5 *Rudin, Chapter 2, Exercise 5*

Construct a bounded set of real numbers with exactly three limit points.

Solution.

Glue together three shifted copies of the sequence $\frac{1}{n}$, aimed at the separated targets 0, 2, 4. With n ranging over the positive integers, set

$$E = \left\{\tfrac{1}{n}\right\} \cup \left\{2+\tfrac{1}{n}\right\} \cup \left\{4+\tfrac{1}{n}\right\}.$$

The verdict: E is bounded, and its set of cluster points is exactly $\{0,2,4\}$.

The set is bounded. From $0 < \frac{1}{n} \le 1$, $2 < 2+\frac{1}{n} \le 3$, and $4 < 4+\frac{1}{n} \le 5$, every element lies in $[0,5]$, so $E \subseteq [0,5]$ is bounded (4.3.4).

Each of 0, 2, 4 is a cluster point. Fix $c \in \{0,2,4\}$ and any radius $r > 0$. By the Archimedean property (Archimedean) there is an n with $\frac{1}{n} < r$; the point $c+\frac{1}{n}$ lies in E,

differs from c, and satisfies $|(c+\frac{1}{n})-c| = \frac{1}{n} < r$. Every ball about c thus contains a point of E other than c, which is the definition of a cluster point (3.2.1).

No other point is a cluster point. Take $p \notin \{0, 2, 4\}$ and set

$$\delta = \min\{\, |p|,\ |p-2|,\ |p-4| \,\} > 0,$$

the distance from p to the nearest target. Each of the three families converges to its target, so for all but finitely many n the points $\frac{1}{n}$, $2+\frac{1}{n}$, $4+\frac{1}{n}$ lie within $\delta/2$ of 0, 2, 4 respectively, hence outside the ball $B(p, \delta/2)$. That ball therefore meets E in only finitely many points, and a cluster point would force infinitely many points of E into every one of its balls (a cluster point is approached infinitely often). So p is not a cluster point.

Thus E is bounded with cluster-point set precisely $\{0, 2, 4\}$.

■

REFLECTIONS.

The phrase that fixes the strategy is *limit point*—a cluster point, a place a set piles up so that every ball around it catches another member of the set (3.2.1). The request is constructive: build one set that stays inside a bounded interval, accumulates at three spots, and accumulates nowhere else. So the question is really "what is the cheapest way to manufacture a single accumulation point, and how do I place three of them without their debris colliding?"

The cheapest accumulation point is a convergent sequence that never reaches its limit. The numbers $\frac{1}{n}$ crowd toward 0 without ever hitting it, so 0 is a cluster point of $\{\frac{1}{n}\}$ and—because the terms march away to the right—its only one (3.2.5). Believe that picture first: a string of dots bunching up against a wall they never touch. To get three cluster points, take three such strings and aim them at 0, 2, 4; keeping the targets two apart, and the whole set inside $[0, 5]$, both bounds the set and stops the three families from interfering with one another.

The two halves of the proof are unequal in difficulty, and reading which is which is the lesson:

(1) that 0, 2, 4 *are* cluster points is the easy half—it falls straight out of the convergences, with the Archimedean property (Archimedean) supplying, for any radius r, an index n with $\frac{1}{n} < r$, so the point $c + \frac{1}{n} \in E$ sits in the ball about c and differs from c;

(2) that there are *no others* is the real work, and the right tool is that a cluster point pulls infinitely many points of the set into every ball around it (a cluster point is approached infinitely often); so to acquit a candidate $p \notin \{0, 2, 4\}$ it is enough to trap it in one ball holding only finitely many points of E, which the separation gap

$\delta = \min\{|p|, |p-2|, |p-4|\}$ provides, since all but finitely many points of each family have already gathered next to their target.

The rigor turns on that gap being strictly positive, which is exactly why the targets had to be kept apart: $\delta > 0$ holds precisely because p avoids all three targets, and it is what guarantees only a finite remnant of each family can intrude. Boundedness is independent and never in doubt, since $E \subseteq [0, 5]$ outright (4.3.4). The two directions together—"these three are" and "nothing else is"—are both needed; stopping after the first would only show $\{0, 2, 4\}$ is contained in the cluster set, not equal to it.

The move generalises into a template. For a bounded set with exactly k cluster points, union k copies of a sequence converging to 0, each shifted onto a distinct, well-separated target; the targets become the cluster points, and the infinitely-many-points criterion (a cluster point is approached infinitely often) keeps any stray accumulation from appearing. The same idea, pushed to countably many targets that themselves accumulate, builds far richer cluster sets—it is the engine behind Problem 13.

Problem 6 *Rudin, Chapter 2, Exercise 6*

Let E' be the set of all limit points of a set E. Prove that E' is closed. Prove that E and $\overline{E}$ have the same limit points. (Recall that $\overline{E} = E \cup E'$.) Do E and E' always have the same limit points?

Solution.

Work in a metric space (X, d), and write $B_r(p) = \{x : d(p, x) < r\}$ for the open ball. A point p is a limit point of a set S when every $B_r(p)$ contains a point of S other than p (3.2.1); S is closed when it contains all of its limit points (3.2.4).

The set E' is closed. Let p be a limit point of E' and fix any radius $r > 0$; the goal is to produce a point of E inside $B_r(p)$ different from p. Since p is a limit point of E', there is a point $q \in E'$ with $q \neq p$ and $d(p, q) < r$. Put $s = \min\{\, r - d(p, q),\, d(p, q) \,\} > 0$. Since q is itself a limit point of E, there is a point $x \in E$ with $x \neq q$ and $d(q, x) < s$. Two estimates now finish it. By the triangle inequality

$$d(p, x) \leq d(p, q) + d(q, x) < d(p, q) + \big(r - d(p, q)\big) = r,$$

so $x \in B_r(p)$; and again by the triangle inequality, in the reverse form,

$$d(p, x) \geq d(p, q) - d(q, x) > d(p, q) - d(p, q) = 0,$$

so $x \neq p$. Every ball about p thus contains a point of E other than p, so $p \in E'$. Hence E' contains all its limit points and is closed.

The sets E and $\overline{E}$ have the same limit points. Since $E \subseteq \overline{E}$, any ball that meets E in a

point other than p also meets $\overline{E}$ there, so every limit point of E is a limit point of $\overline{E}$. For the reverse, let p be a limit point of $\overline{E}$ and fix $r > 0$. Choose $q \in \overline{E}$ with $q \neq p$ and $d(p, q) < r$. By $\overline{E} = E \cup E'$ there are two cases. If $q \in E$, then $B_r(p)$ already contains the point $q \in E$ with $q \neq p$. If instead $q \in E'$, then q is a limit point of E, and the same two-step estimate as above—taking $x \in E$ with $x \neq q$ and $d(q, x) < \min\{\, r - d(p, q),\, d(p, q) \,\}$—yields a point $x \in E$ with $x \neq p$ and $d(p, x) < r$. In either case $B_r(p)$ contains a point of E other than p, so p is a limit point of E. Therefore E and $\overline{E}$ have exactly the same limit points.

The sets E and E' need not have the same limit points. In $\mathbb{R}$ take

$$E = \left\{ \tfrac{1}{n} : n \in \mathbb{Z}^+ \right\}.$$

Its only limit point is 0, so $E' = \{0\}$. Then 0 is a limit point of E but not of E': the one-point set $E' = \{0\}$ has no limit point at all, since the ball $B_1(0)$ contains no point of E' different from 0. So E has a limit point that E' does not, and the two sets do not always share their limit points.

■

REFLECTIONS.

The whole problem is built from one definition—a limit point of S is a point every ball around which catches a member of S other than itself (3.2.1)—and the recurring move is to chain two of these statements together. "Prove E' is closed" is a request to show E' contains its own limit points (3.2.4), and the only data you start with is that p is close to points q that are in turn close to points of E; passing the closeness along, in two hops $p \rightsquigarrow q \rightsquigarrow x$, is the entire idea. The hard part is purely bookkeeping on the radii, and the quotable fact that a limit point attracts *infinitely many* members (cluster points pull in infinitely many points) is what guarantees each hop has somewhere to land even after you shrink the tolerance.

A picture makes the radius juggling believable before the inequalities are written. Imagine p with a ball of radius r; pick $q \in E'$ strictly inside, at distance $d(p, q) < r$. Two things could spoil the landing point $x \in E$ near q: it could be pushed out past radius r, or it could land exactly on p. Shrinking the search radius around q to $s = \min\{r - d(p, q),\, d(p, q)\}$ closes off both: $r - d(p, q)$ is the slack left before the outer wall, and $d(p, q)$ is the room between q and p, so an x within s of q stays inside the r-ball yet cannot reach p.

The proof then runs in clean steps, each resting on the one before:

(1) from $p \in (E')'$ and the definition of a limit point (3.2.1), extract $q \in E'$ with $q \neq p$ and $d(p, q) < r$—the first hop;

(2) set the guarded radius $s = \min\{r - d(p, q),\, d(p, q)\}$, positive because $d(p, q)$ lies strictly between 0 and r, and use $q \in E'$ to extract $x \in E$ with $x \neq q$ and $d(q, x) < s$—the

second hop;

(3) the forward triangle inequality $d(p,x) \le d(p,q) + d(q,x) < r$ keeps x inside $B_r(p)$, and the reverse form $d(p,x) \ge d(p,q) - d(q,x) > 0$ keeps x off p, so $B_r(p)$ holds a point of E unequal to p, which is exactly $p \in E'$ (3.2.4).

The equality of limit points for E and $\overline{E}$ reuses the very same machine. One inclusion is free from monotonicity—a bigger set has at least as many limit points—and the other splits on $\overline{E} = E \cup E'$ (4.4.1): a nearby point of $\overline{E}$ is either already in E, and then there is nothing to do, or it is a limit point of E, and then the two-hop estimate of step (3) manufactures the needed point of E. This is also why E' being closed is not a throwaway first part—it is the engine that makes $\overline{E} = E \cup E'$ closed at all ($E \cup E'$ is closed), since adjoining a closed set of limit points to E cannot create new ones to escape through.

Two checks keep the argument honest. The guard s must be *strictly* positive, which is why both inequalities $0 < d(p,q) < r$ are needed—the first from $q \neq p$, the second from $d(p,q) < r$—and dropping either one lets x collapse onto p or spill outside the ball. And the case split on $\overline{E} = E \cup E'$ is genuinely two cases: skipping the easy $q \in E$ branch would leave a gap, not a shortcut. There is no circularity, since closedness of E' is established first and only then fed into the closure statement.

The final question is the one that separates the pattern from its limits. Closing a set never adds limit points, but *replacing* a set by its limit points can lose them: $E = \{1/n\}$ piles up at 0, so $0 \in E'$, yet $E' = \{0\}$ is a lone point with no pile-up of its own. The takeaway is that $E \mapsto \overline{E}$ is idempotent—E and $\overline{E}$ carry exactly the same limit points, so $\overline{\overline{E}} = \overline{E}$—while $E \mapsto E'$ is a genuine collapse that can strip a set down to isolated points: here it sends $\{1/n\}$ to $\{0\}$ and then to $\varnothing$, so $E'' = \varnothing \neq E'$. Passing to $\overline{E}$ is the safe closure operation and passing to E' is not.

Problem 7 *Rudin, Chapter 2, Exercise 7*

Let $A_1, A_2, A_3, \ldots$ be subsets of a metric space.

(a) If $B_n = \bigcup_{i=1}^{n} A_i$, prove that $\overline{B}_n = \bigcup_{i=1}^{n} \overline{A}_i$, for $n = 1, 2, 3, \ldots$.

(b) If $B = \bigcup_{i=1}^{\infty} A_i$, prove that $\overline{B} \supset \bigcup_{i=1}^{\infty} \overline{A}_i$.

Show, by an example, that this inclusion can be proper.

Solution.

Throughout, the closure $\overline{E} = E \cup E'$ is the smallest closed set containing E, and it is monotone: if $A \subseteq B$, then $\overline{A}$ is contained in the closed set $\overline{B}$, so $\overline{A} \subseteq \overline{B}$.

For (a), fix n. Each $A_i \subseteq B_n$ for $i \le n$, so monotonicity gives $\overline{A}_i \subseteq \overline{B}_n$, and hence $\bigcup_{i=1}^{n} \overline{A}_i \subseteq \overline{B}_n$. For the reverse inclusion, each $\overline{A}_i$ is closed (4.4.1), so $\bigcup_{i=1}^{n} \overline{A}_i$ is a *finite* union of closed sets and is therefore closed (finite unions of closed sets are closed; 3.2.10).

It contains $B_n = \bigcup_{i=1}^{n} A_i$, since $A_i \subseteq \overline{A}_i$ for each i. As $\overline{B}_n$ is the smallest closed set containing B_n, it follows that $\overline{B}_n \subseteq \bigcup_{i=1}^{n} \overline{A}_i$. The two inclusions give

$$\overline{B}_n = \bigcup_{i=1}^{n} \overline{A}_i.$$

For (b), every $A_i \subseteq B$, so monotonicity gives $\overline{A}_i \subseteq \overline{B}$ for each i, and therefore

$$\bigcup_{i=1}^{\infty} \overline{A}_i \subseteq \overline{B}.$$

The inclusion in (b) can be proper. Take $A_i = \{1/i\} \subseteq \mathbb{R}$ for $i = 1, 2, 3, \ldots$. Each singleton is closed (a finite set has no cluster points, so it contains them all vacuously, 3.2.4), so $\overline{A}_i = A_i$ and

$$\bigcup_{i=1}^{\infty} \overline{A}_i = \left\{ \tfrac{1}{i} : i \geq 1 \right\}.$$

The point 0 is a cluster point of $B = \bigcup_{i=1}^{\infty} A_i = \{1/i : i \geq 1\}$: given $\varepsilon > 0$, the Archimedean property (Archimedean property) supplies an integer $i > 1/\varepsilon$, and then $0 < 1/i < \varepsilon$, so the ball $B_\varepsilon(0)$ contains the point $1/i \neq 0$ of B. Hence $0 \in \overline{B}$, while $0 \notin \bigcup_{i=1}^{\infty} \overline{A}_i$, and the inclusion is strict.

■

REFLECTIONS.

The question is whether one of the two operations in sight commutes with the other: does closing a union equal the union of the closures? The word *closure* points at the one structural fact that drives the whole problem, that $\overline{E} = E \cup E'$ is the *smallest* closed set containing E (4.4.1). Once closure is read that way—as a minimal closed hull rather than "E with its limit points glued on"—the easy half falls out of monotonicity and the hard half turns into a question about whether a union of closed sets stays closed, which is exactly where "finite" versus "infinite" will decide the answer.

Monotonicity is worth isolating because it is used in every part and is itself just the minimality of the hull: if $A \subseteq B$, then $\overline{B}$ is a closed set that already contains A, so the smallest such set $\overline{A}$ cannot be larger. That single observation gives $\bigcup \overline{A}_i \subseteq \overline{\bigcup A_i}$ for free, in both the finite and the infinite case—each piece A_i sits inside the union, so its hull sits inside the union's hull. The content of the problem is never this direction; it is the reverse one in (a), and its failure in (b).

The reverse inclusion in (a) rests on one load-bearing step, which is where you should slow down:

(1) each $\overline{A}_i$ is closed (4.4.1), and a *finite* union of closed sets is again closed (finite unions

of closed sets are closed; 3.2.10), so $\bigcup_{i=1}^{n} \overline{A}_i$ is itself a closed set;

(2) this closed set contains B_n, because $A_i \subseteq \overline{A}_i$ pushes the whole union $B_n = \bigcup_{i \le n} A_i$ inside it;

(3) $\overline{B}_n$ is by definition the *smallest* closed set containing B_n (4.4.1), so it can only be smaller than the closed set produced in (1), giving $\overline{B}_n \subseteq \bigcup_{i=1}^{n} \overline{A}_i$ and, with monotonicity, equality.

The entire argument is the minimality of the hull played against a closed set you have built by hand; nothing is circular, since the built set is closed for an independent reason.

Step (1) is exactly where the infinite case breaks, and seeing why is the point of the counterexample. An *infinite* union of closed sets need not be closed—the same phenomenon as an infinite intersection of open sets failing to be open (3.1.3)—so for the infinite union there is no closed set built from the $\overline{A}_i$ to play against, and $\overline{B}$ is free to contain points lying in no single $\overline{A}_i$. The choice $A_i = \{1/i\}$ is the cleanest witness: each set is a lone closed point, equal to its own closure, yet the points $1/i$ together pile up at 0. That 0 is a cluster point of B is the Archimedean property in disguise ($\inf\{1/n\} = 0$)—every ball about 0 swallows a tail of the $1/i$ (a cluster point is approached by infinitely many points of the set)—so $0 \in \overline{B}$, while 0 belongs to none of the $\overline{A}_i = \{1/i\}$.

The transferable reading is that closure commutes with *finite* unions and only finite ones, and the gap is always carried by cluster points that no single piece A_i can reach—points produced only by drawing terms from infinitely many of the sets at once. Whenever a topological identity holds for finite unions but you suspect it fails in the limit, look for the operation's stability theorem (here, finite unions of closed sets are closed, finite unions of closed sets are closed) and probe exactly the hypothesis "finite": the counterexample lives at the limit point the finite version was quietly forbidding.

Problem 8 *Rudin, Chapter 2, Exercise 8*

Is every point of every open set $E \subset \mathbb{R}^2$ a limit point of E? Answer the same question for closed sets in $\mathbb{R}^2$.

Solution.

Yes for open sets; no for closed sets.

Let $E \subseteq \mathbb{R}^2$ be open and fix $p \in E$. Openness makes p an interior point, so some ball $B_\varepsilon(p) \subseteq E$ with $\varepsilon > 0$ (2.3.3). To see p is a limit point, take any radius $s > 0$, set $\rho = \min(\varepsilon, s) > 0$, and let $q = p + (\frac{\rho}{2}, 0)$. Then $|q - p| = \frac{\rho}{2} < \rho \le \varepsilon$, so $q \in B_\varepsilon(p) \subseteq E$; and $\frac{\rho}{2} < \rho \le s$, so $q \in B_s(p)$; and $q \neq p$. Every ball about p thus contains a point of E other than p, which is exactly the definition of a limit point (3.2.1). Hence every point of an open set in $\mathbb{R}^2$ is a limit point of it.

For closed sets the answer is no. Fix any $p \in \mathbb{R}^2$ and take $E = \{p\}$. This E is closed: being finite it has no limit points—a limit point would force infinitely many points of E into each of its neighbourhoods (a cluster point captures infinitely many points of the set, 3.2.1)—so E vacuously contains all of them and is closed (3.2.4). Yet p is *not* a limit point of E: the ball $B_1(p)$ holds no point of E different from p, so p is an isolated point of E (3.2.2). Thus p is a point of a closed set that fails to be a limit point of it.

■

REFLECTIONS.

The two answers split because "open" and "closed" constrain genuinely different things, and the way to read the problem is to translate each into the language of limit points. A limit point is a *crowded* point: every ball around it, no matter how small, still catches another point of the set (3.2.1). So the real question is whether merely belonging to the set forces that crowding—and the equivalent notion, an *isolated* point (3.2.2), a member with a ball around it meeting the set in itself alone, is precisely the witness that crowding can fail. The contrast the problem is fishing for is whether open or closed sets are allowed isolated points.

Believe the open case through the picture first: a point of an open set sits at the centre of a whole disk lying inside the set (that is what 2.3.3 means), and in the plane no disk is a single point—slide a hair in any direction and you are still inside it. So neighbours of p inside the set are unavoidable. The closed case is believed through the cleanest counterexample, a lone point $\{p\}$: there is nothing nearby to crowd it, and closedness has no complaint, since a finite set has no limit points to omit.

The open half runs in three short moves:

(1) openness gives a ball $B_\varepsilon(p) \subseteq E$, because each point of an open set is interior to it (2.3.3)—this is the only hypothesis used, and it is the whole supply of nearby points;

(2) to test a limit point you must beat an *arbitrary* radius s, so you shrink to $\rho = \min(\varepsilon, s)$ to stay inside both the witnessing ball and the test ball at once;

(3) the explicit neighbour $q = p + (\frac{\rho}{2}, 0)$ lands in E and in $B_s(p)$ and differs from p, meeting the definition (3.2.1) on the nose.

Move (2) is where rigour lives: a limit point must be checked against *every* radius, not one convenient one, and taking the minimum is the standard way to honour that "for all s" without case work.

The closed half is a single counterexample, and its honesty rests on why $\{p\}$ is closed: a finite set has no limit points at all, since a limit point would drag infinitely many members into every neighbourhood (the infinitely-many-points criterion, 3.2.1), so the "contains all

its limit points" condition (3.2.4) holds vacuously. One example is enough to answer a "does this always hold?" question in the negative, and one is all that is needed.

Each verdict turns on isolated points, and that is the transferable line to keep (3.2.7): an open set in $\mathbb{R}^k$ has none—nothing about $\mathbb{R}^2$ was special, every $\mathbb{R}^k$ behaves the same—so every one of its points is a limit point; a closed set may consist entirely of isolated points. When a problem wants a point of a set that is *not* a limit point, reach for a closed set with isolated points: a finite set, or a discrete lattice like $\mathbb{Z}^2$.

Problem 9 *Rudin, Chapter 2, Exercise 9*

Let E° denote the set of all interior points of a set E. (E° is called the *interior* of E.)

(a) Prove that E° is always open.

(b) Prove that E is open if and only if $E^\circ = E$.

(c) If $G \subset E$ and G is open, prove that $G \subset E^\circ$.

(d) Prove that the complement of E° is the closure of the complement of E.

(e) Do E and $\overline{E}$ always have the same interiors?

(f) Do E and E° always have the same closures?

Solution.

Throughout, E is a subset of a metric space X and every complement is taken in X. A point p is *interior* to E when some ball $N_r(p) \subseteq E$, and E° is the set of such points (2.3.3); by definition $E^\circ \subseteq E$.

(a) Let $p \in E^\circ$, so $N_r(p) \subseteq E$ for some $r > 0$. The ball $N_r(p)$ is itself open (2.3.5; the open balls are open), so each $q \in N_r(p)$ has a ball $N_s(q) \subseteq N_r(p) \subseteq E$, making q an interior point of E. Thus $N_r(p) \subseteq E^\circ$, so p is interior to E°. Every point of E° is interior to it, hence E° is open.

(b) If $E^\circ = E$, then E is open by (a). Conversely, if E is open, every $p \in E$ has a ball $N_r(p) \subseteq E$, so $p \in E^\circ$; with the standing inclusion $E^\circ \subseteq E$ this gives $E^\circ = E$.

(c) Let $p \in G$. Since G is open, some ball $N_r(p) \subseteq G$, and $G \subseteq E$ gives $N_r(p) \subseteq E$; thus $p \in E^\circ$. Hence $G \subseteq E^\circ$.

(d) We prove $(E^\circ)^c = \overline{E^c}$ by showing the two sides have the same points. A point p fails to lie in E° exactly when *no* ball about p is contained in E, that is, when *every* ball about p contains a point of E^c. On the other side, $\overline{E^c} = E^c \cup (E^c)'$ (4.4.1), so $p \in \overline{E^c}$ means $p \in E^c$ or p is a cluster point of E^c (3.2.1)—and either way every ball about p meets E^c (if $p \in E^c$, the point p itself is in the ball). Conversely, if every ball about p meets E^c: should $p \notin E^c$, each such meeting point differs from p, making p a cluster point of E^c, so $p \in \overline{E^c}$; and if $p \in E^c$ then $p \in \overline{E^c}$ outright. The two conditions coincide, so $(E^\circ)^c = \overline{E^c}$.

(e) No. In $\mathbb{R}$ take $E = (-1,0)\cup(0,1)$. Then E is open, so $E^\circ = E$ by (b); but $\overline{E} = [-1,1]$, whose interior is $(-1,1)$. Since $0 \in (-1,1) \setminus E$, the interiors of E and $\overline{E}$ differ.

(f) No. In $\mathbb{R}$ take $E = \{0\}$. A one-point set has no cluster point, so it is closed (3.2.4) and $\overline{E} = \{0\}$; but 0 is not interior to E, so $E^\circ = \varnothing$ and $\overline{E^\circ} = \varnothing$. The closures of E and E° differ.

■

REFLECTIONS.

The single word that organises the whole problem is *interior*. An interior point is one with room to spare—a ball entirely inside E (2.3.3)—and that "ball inside" clause is the only machinery the first three parts ever touch. Recognising this is what tells you that (a), (b), (c) are not three separate tasks but three readings of one definition: (a) says E° is open, (c) says it swallows every open subset of E, and together they identify E° as the *largest* open set inside E, the inside-out twin of the closure (4.4.1), which is the smallest closed set containing E. Part (b) is then the fixed-point statement: a set is open exactly when it already equals its own core.

Believe that picture on an example before proving it. For $E = [0,1]$ the interior is the open core $(0,1)$ and the closure is $[0,1]$; the endpoints are interior to neither but cluster points of both. Interior carves *off* the boundary, closure glues it *on*, and part (d) is the precise statement that these two operations are mirror images through complementation.

Each step rests on something named:

(1) for (a), the leap from "p has a ball in E" to "every *nearby* point does too" is powered by the fact that a ball is open (2.3.5; open balls are open): inside $N_r(p)$ each q gets its own sub-ball, so $N_r(p) \subseteq E^\circ$ and openness propagates from the one point to the whole ball;

(2) (b) is a biconditional, so it is two implications (0.5.9)—one direction is just (a), the other reads "E open" through the definition of interior, and the standing inclusion $E^\circ \subseteq E$ supplies the half you never have to argue;

(3) (c) chains two containments, $N_r(p) \subseteq G$ from openness of G and $G \subseteq E$ from hypothesis, so the witnessing ball lands in E and certifies $p \in E^\circ$ (0.3.6);

(4) (d) is the crux, an equality of sets proved by matching membership conditions (0.5.10): negating "some ball lies in E" yields "every ball meets E^c," and that is exactly the neighbourhood test for the closure $E^c \cup (E^c)'$ (4.4.1, 3.2.1). The lone subtlety is the point p itself—when $p \in E^c$ a ball meets E^c at p without p being a cluster point, which is why the closure is $E^c \cup (E^c)'$ and not $(E^c)'$ alone.

The rigour check is to confirm each direction of (d) is reversible and that the two coun-

terexamples genuinely separate the operations. They do, in opposite directions: closing can *grow* the interior, as $(-1,0)\cup(0,1)$ (its own interior) closes to $[-1,1]$ whose interior $(-1,1)$ has swallowed the missing 0; while taking the interior can *erase* a set, as $\{0\}$ (closed, equal to its closure) has empty interior whose closure is empty. So E, E°, and $\overline{E}$ are three honestly different sets, and neither rounding operation is stable under the other.

The move to carry away is duality by complement: part (d) says $(E^\circ)^c = \overline{E^c}$, so every theorem about interiors converts—swap $E \leftrightarrow E^c$ and interior$\leftrightarrow$ closure—into a theorem about closures, and you prove only one of each pair. It is the same bookkeeping that turns "open iff complement closed" (open $\iff$ complement closed, 3.2.8) into a two-for-one, and the reason interior and closure always travel together.

Problem 10 *Rudin, Chapter 2, Exercise 10*

Let X be an infinite set. For $p \in X$ and $q \in X$, define

$$d(p,q) = \begin{cases} 1 & (\text{if } p \neq q) \\ 0 & (\text{if } p = q). \end{cases}$$

Prove that this is a metric. Which subsets of the resulting metric space are open? Which are closed? Which are compact?

Solution.

The map d is a metric. Every subset of X is both open and closed. The compact subsets are exactly the finite ones.

First we check that d is a metric. From the definition $d(p,q) \geq 0$, with $d(p,q) = 0$ precisely when $p = q$, and $d(p,q) = d(q,p)$ since "$p \neq q$" is symmetric in p and q. For the triangle inequality, fix $p, q, r \in X$. If $p = q$ the left side $d(p,q) = 0 \leq d(p,r) + d(r,q)$. If $p \neq q$, then r cannot equal both p and q, so at least one of $d(p,r), d(r,q)$ equals 1, giving $d(p,q) = 1 \leq d(p,r) + d(r,q)$. Hence all four axioms hold and d is a metric (2.2.1); this is the discrete metric of 2.2.4.

Everything topological is read off the open balls, so we compute them next. For any $p \in X$ and radius $0 < \varepsilon \leq 1$,

$$B_\varepsilon(p) = \{q \in X : d(p,q) < \varepsilon\} = \{p\},$$

since the only point within distance 1 of p is p itself; for $\varepsilon > 1$ the ball is all of X. Everything below rests on the first case.

Every subset is open. Let $E \subseteq X$ and $p \in E$. The ball $B_{1/2}(p) = \{p\} \subseteq E$ witnesses p as an interior point, so E is open (2.3.3). (Equivalently, each singleton $\{p\} = B_{1/2}(p)$ is open by "open balls are open", and $E = \bigcup_{p\in E}\{p\}$ is a union of open sets, hence open by the union rule.)

Every subset is closed as well. Each subset E has open complement $X \setminus E$, so E is closed (3.2.8; "open $\iff$ complement closed"). Every subset is therefore clopen.

Finally, the compact subsets are exactly the finite ones. A finite set $\{x_1, \ldots, x_m\}$ is compact: given any open cover, choose for each x_j one cover member containing it, and these m members form a finite subcover (3.4.1). Conversely, suppose $E \subseteq X$ is infinite. The family $\{\, B_{1/2}(p) : p \in E \,\} = \{\, \{p\} : p \in E \,\}$ is an open cover of E in which each member contains exactly one point of E. Any finite subfamily covers only finitely many points, so it omits a point of the infinite set E; no finite subcover exists, and E is not compact. Thus a subset is compact iff it is finite.

■

REFLECTIONS.

The statement hands you an explicit two-valued formula for d rather than a named space, so the first job is to recognise the object: distances are only 0 or 1, which is the discrete metric of our notes (2.2.4). The word *metric* then points at the four axioms (2.2.1) as the thing to verify, and the three follow-up questions—open, closed, compact—each name a definition to apply (2.3.3, 3.2.4, 3.4.1). The strategic clue is that in a metric space every notion of nearness is read off the open balls, so before answering anything you compute the balls; here they are so coarse that one calculation settles all three questions at once.

Believe the coarseness first. Picture the points of X as scattered so that any two distinct ones are a full unit apart—a cloud of isolated specks, no two ever closer. A ball of radius $\frac{1}{2}$ then reaches nobody but its own centre, and a ball of radius 2 swallows the entire space; there is no intermediate scale. That single picture is the whole problem.

The argument runs in four moves, each resting on the one before it or on a cited definition:

(1) the metric axioms (2.2.1) reduce to one real check, the triangle inequality, and that is settled by a counting remark: a third point r cannot coincide with two distinct points p, q at once, so one of the legs $d(p, r), d(r, q)$ is already 1 and dominates the unit left side;

(2) the crux computation $B_{1/2}(p) = \{p\}$ comes straight from the definition of the ball (2.3.1), because no point other than p lies within distance 1;

(3) openness of every E is then immediate—each point $p \in E$ has the witnessing ball $B_{1/2}(p) = \{p\} \subseteq E$, so p is interior (2.3.3); equivalently every singleton is an open ball (open balls are open) and E is their union (unions of open sets are open);

(4) closedness needs no new work: with every set open, every set has open complement, so every set is closed (3.2.8; open $\iff$ complement closed). The same ball calculation gives the direct reading too—no point is a cluster point of anything, since a ball around

p contains no point of a set other than p itself (3.2.1; cluster points are approached infinitely often), so E contains its empty set of cluster points vacuously.

Compactness is the lone notion the coarse metric does *not* trivialise, and seeing why sharpens the whole space. The verdict "compact = finite" is proved in both directions and both halves are load-bearing: finiteness alone yields a finite subcover, while infiniteness is exploited by the cover of singletons, where every member is forced to do the work of a single point so that nothing finite can suffice. This is also where the infinitude of X in the hypothesis finally matters—it guarantees X itself is a closed and bounded (every distance is ≤ 1) set that fails to be compact.

That failure is the reason to keep the example. Heine–Borel (4.3.5; "compact $\iff$ closed and bounded in $\mathbb{R}^k$") makes "closed and bounded" equal to "compact" only inside $\mathbb{R}^k$; the discrete metric on an infinite set is the clean witness that the equivalence is special to Euclidean space, since here closed-and-bounded sets are everywhere yet only finite sets are compact. So when you need a metric in which some Euclidean intuition breaks—a bounded set with no convergent subsequence, an infinite set with no cluster point, a space where every set is clopen and hence totally disconnected (4.4.2)—the discrete metric is the first place to reach. The transferable move is the one that opened the solution: in any metric space, compute the open balls before answering a single topological question, because in a metric space every property of nearness is encoded in them.

Problem 11 *Rudin, Chapter 2, Exercise 11*

For $x \in \mathbb{R}^1$ and $y \in \mathbb{R}^1$, define

$$d_1(x,y) = (x-y)^2, \quad d_2(x,y) = \sqrt{|x-y|}, \qquad d_3(x,y) = |x^2 - y^2|,$$
$$d_4(x,y) = |x - 2y|, \quad d_5(x,y) = \frac{|x-y|}{1+|x-y|}.$$

Determine, for each of these, whether it is a metric or not.

Solution.

The metrics are d_2 and d_5; the functions d_1, d_3, d_4 each fail one axiom of a metric (2.2.1). A metric on $\mathbb{R}$ must be nonnegative and vanish exactly on the diagonal $x = y$, be symmetric, and satisfy the triangle inequality.

The function d_1 fails the triangle inequality: with $x = 0$, $z = 2$, and the intermediate point $y = 1$,

$$d_1(0,2) = 4 > 2 = 1 + 1 = d_1(0,1) + d_1(1,2).$$

The function d_3 fails definiteness, since $d_3(1,-1) = |1^2 - (-1)^2| = 0$ although $1 \neq -1$; distinct points sit at distance 0. The function d_4 fails already at $d(x,x) = 0$, since $d_4(1,1) = |1 - 2 \cdot 1| = 1 \neq 0$. One broken axiom is enough to disqualify each.

For d_2, nonnegativity, the vanishing $d_2(x,x) = \sqrt{0} = 0$ together with $d_2(x,y) > 0$ when $x \neq y$, and symmetry are inherited directly from $|x-y|$, since the square root is nonnegative, vanishes only at 0, and ignores order. For the triangle inequality, fix $x, y, z \in \mathbb{R}$ and start from the ordinary one, $|x-z| \leq |x-y| + |y-z|$. The square root is increasing, so

$$d_2(x,z) = \sqrt{|x-z|} \leq \sqrt{|x-y| + |y-z|} \leq \sqrt{|x-y|} + \sqrt{|y-z|} = d_2(x,y) + d_2(y,z),$$

the last step being $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, which holds because squaring the right-hand side gives $a + b + 2\sqrt{ab} \geq a + b$. Thus d_2 is a metric.

For d_5, write $f(t) = \dfrac{t}{1+t}$ for $t \geq 0$, so that $d_5(x,y) = f(|x-y|)$. Again nonnegativity, definiteness ($f(t) = 0$ iff $t = 0$), and symmetry come from $|x-y|$. The map f is increasing on $[0, \infty)$, since $f(t) = 1 - \frac{1}{1+t}$ grows as t does, and it is subadditive: for $a, b \geq 0$,

$$f(a+b) = \frac{a+b}{1+a+b} = \frac{a}{1+a+b} + \frac{b}{1+a+b} \leq \frac{a}{1+a} + \frac{b}{1+b} = f(a) + f(b),$$

because $1 + a + b \geq 1 + a$ and $1 + a + b \geq 1 + b$ make each fraction only larger when the denominator shrinks. Combining monotonicity with the ordinary triangle inequality,

$$d_5(x,z) = f(|x-z|) \leq f\big(|x-y| + |y-z|\big) \leq f(|x-y|) + f(|y-z|) = d_5(x,y) + d_5(y,z).$$

Hence d_5 is a metric.

■

REFLECTIONS.

The single word governing every part is *metric*: each candidate is to be tested against the four clauses of the definition (2.2.1)—nonnegativity, vanishing exactly on the diagonal, symmetry, and the triangle inequality—and the verdict is whichever clause first holds or breaks. Because the task splits cleanly into "produce a counterexample" for the failures and "verify all four clauses" for the successes, this is a proof by cases (0.5.7) with two different burdens, and reading the burden correctly is the whole skill. To *refute* "d is a metric" you negate a universally quantified statement (0.2.5), so a single triple x, y, z suffices; to *confirm* it you owe all four clauses for all inputs.

That asymmetry sets the order of work: hunt first for a cheap failure, and only do the labour of verification on what survives. The cheapest failures to scan for are a nonzero $d(x,x)$, two distinct points at distance 0, and a three-point violation of the triangle inequality—which is exactly how d_4, d_3, and d_1 fall:

(1) d_4 dies at the first clause, since $d_4(1,1) = |1-2| = 1 \neq 0$: the formula $|x-2y|$ treats its two slots unequally, so a point is not even at distance 0 from itself (it also fails symmetry, but one broken axiom is all the negation needs);

(2) d_3 dies at definiteness, since $x \mapsto x^2$ collapses 1 and -1 to the same value, giving $d_3(1,-1) = 0$ with $1 \neq -1$—a metric must *separate* distinct points, and any non-injective inner transformation forfeits that;

(3) d_1 satisfies the first three clauses but breaks the triangle inequality, because squaring inflates a long step out of proportion to two short ones: $d_1(0,2) = 4$ exceeds $d_1(0,1) + d_1(1,2) = 2$, the standard "too convex" failure.

The two survivors share a structure worth naming: each is $g(|x-y|)$ for a function g that fixes the ordinary distance on $\mathbb{R}$. The first three clauses then transfer for free—$g(0) = 0$ keeps the diagonal, $g(t) > 0$ for $t > 0$ keeps separation, and $|x-y|$ is already symmetric. Only the triangle inequality can be lost, and it is preserved precisely when g is increasing and subadditive ($g(a+b) \leq g(a)+g(b)$): feed the ordinary inequality $|x-z| \leq |x-y|+|y-z|$ through the increasing g, then split with subadditivity, exactly the two-line chain the solution runs for both. For $g = \sqrt{\cdot}$ subadditivity is $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$; for $g(t) = t/(1+t)$ it is the bounded-distance trick that recurs throughout the subject, since this g caps every distance below 1 yet still defines the same notion of nearness as $|x-y|$.

The cases are exhaustive—five functions, each assigned exactly one fate, with at least one explicit axiom failing in every rejection and all four verified in every acceptance—and nothing is circular, as the verifications lean only on the ordinary metric on $\mathbb{R}$ (2.2.3) and on elementary inequalities. The transferable move is this dichotomy in how you read "is it a metric?": a refutation is one well-chosen triple, while a confirmation is four clauses earned in full, and the fastest path is to chase the cheap counterexample before committing to the verification. The deeper lesson, the one the survivors teach, is that an increasing subadditive reshaping of a metric is again a metric—the source of $t/(1+t)$ as the standard way to make any metric bounded without disturbing its topology.

Problem 12 *Rudin, Chapter 2, Exercise 12*

Let $K \subset \mathbb{R}^1$ consist of 0 and the numbers $1/n$, for $n = 1, 2, 3, \ldots$. Prove that K is compact directly from the definition (without using the Heine–Borel theorem).

Solution.

Write $K = \{0\} \cup \{1/n : n \geq 1\}$, and let $\{G_\alpha\}$ be an arbitrary open cover of K; we must extract a finite subcover.

The point 0 lies in some member G_{α_0} of the cover. Since G_{α_0} is open and 0 is one of its points, 0 is an interior point (2.3.3), so some ball $(-r, r) \subseteq G_{\alpha_0}$ with $r > 0$. By the Archimedean property (1.7.1) there is an integer $N \geq 1$ with $1/N < r$. For every $n \geq N$ then $0 < 1/n \leq 1/N < r$, so $1/n \in (-r, r) \subseteq G_{\alpha_0}$; and $0 \in G_{\alpha_0}$ as well. Thus the single set G_{α_0} already contains the entire tail $\{0\} \cup \{1/n : n \geq N\}$.

What remains uncovered lies in the finite set $\{1, 1/2, \ldots, 1/(N-1)\}$. For each such point

choose one member of the cover containing it, say $G_{\alpha_1}, \ldots, G_{\alpha_{N-1}}$. The finite collection

$$G_{\alpha_0},\, G_{\alpha_1},\, \ldots,\, G_{\alpha_{N-1}}$$

then covers all of K: the tail by G_{α_0}, each leftover by its chosen set. Every open cover of K admits a finite subcover, so K is compact by definition (3.4.1).

■

REFLECTIONS.

The phrase "directly from the definition (without using Heine–Borel)" is an explicit instruction about which tool to reach for. The shortcut would be to remark that $K \subseteq [0,1]$ is closed and bounded and quote Heine–Borel (4.3.5); the problem forbids exactly that, leaving only the raw definition of compactness—every open cover has a finite subcover (3.4.1). So the task is fixed in advance: start from an *arbitrary* open cover $\{G_\alpha\}$ and produce a finite subcollection that still covers, with the obstacle that K is infinite while a subcover may use only finitely many sets.

The second clue is the shape of K itself: a sequence $1/n$ together with the one point 0 it approaches. That 0 is the lone accumulation point, and reading it as a limit is the whole idea, since $1/n \to 0$ means every neighbourhood of 0 swallows all but finitely many terms (the tail lands in every ball about the limit, 5.1.3). A cover set that catches 0 is forced, by openness, to contain a whole interval around 0, and that interval drags in the entire tail for free. Picture it on the line: pin an interval at 0 and only the few largest terms $1, 1/2, \ldots$ poke out to its right; everything else is already inside.

That picture is the proof, step by step:

(1) some G_{α_0} contains 0 because $\{G_\alpha\}$ covers K and $0 \in K$, and openness makes 0 an interior point (2.3.3), so a ball $(-r, r) \subseteq G_{\alpha_0}$;

(2) the Archimedean property (1.7.1; $\inf\{1/n\} = 0$) supplies an N with $1/N < r$, and then $1/n \le 1/N < r$ for all $n \ge N$, so this one set G_{α_0} holds the entire infinite tail $\{0\} \cup \{1/n : n \ge N\}$;

(3) only the finitely many points $1, 1/2, \ldots, 1/(N-1)$ are left, and finitely many points are harmless—assign one cover member to each;

(4) G_{α_0} together with those finitely many sets is a finite subcover, which is the definition (3.4.1) satisfied.

Each hypothesis earns its place, and dropping the point 0 shows why. The bare set $\{1/n : n \ge 1\}$ is *not* compact: cover it by tiny disjoint intervals, one snug around each $1/n$, and no finite subfamily reaches the points near 0 (3.4.2). What rescues K is precisely that 0 has been included, so a single cover set is compelled to absorb the tail; without it nothing

forces that absorption. The Archimedean step is equally load-bearing—it is what turns "$1/n$ gets small" into a concrete cutoff N that splits K cleanly into a covered tail and a finite head.

The transferable move is the head–tail split for compactness from scratch: find the one point that catches infinitely many elements, let the open set around it inhale the whole tail (this is where $x_n \to p$ does the work), and mop up the finite remainder one set at a time. It is the model argument for "a convergent sequence together with its limit is compact," and it is exactly the engine behind the more elaborate construction in Problem 13.

Problem 13 *Rudin, Chapter 2, Exercise 13*

Construct a compact set of real numbers whose limit points form a countable set.

Solution.

Push the construction of Problem 5 one level deeper: instead of three target points, take the countably many targets $1, \frac{1}{2}, \frac{1}{3}, \dots$, attach a sequence accumulating at each, and add the point 0 at which the targets themselves accumulate. With n, k ranging over the positive integers, set

$$K = \{0\} \cup \left\{\tfrac{1}{n} : n \geq 1\right\} \cup \left\{\tfrac{1}{n} + \tfrac{1}{nk} : n, k \geq 1\right\}, \qquad L = \{0\} \cup \left\{\tfrac{1}{n} : n \geq 1\right\}.$$

The verdict: K is compact, and its set of cluster points is exactly L, which is countable.

The set is bounded. For every $n, k \geq 1$ we have $0 < \frac{1}{n} \leq 1$ and $0 < \frac{1}{n} + \frac{1}{nk} \leq 1 + 1 = 2$, the maximum occurring at $n = k = 1$; so $K \subseteq [0, 2]$ is bounded (4.3.4).

Every point of L is a cluster point of K. Fix n and any radius $r > 0$. By the Archimedean property (Archimedean) there is a k with $\frac{1}{nk} < r$; the point $\frac{1}{n} + \frac{1}{nk}$ lies in K, differs from $\frac{1}{n}$, and sits within r of $\frac{1}{n}$, so $\frac{1}{n}$ is a cluster point. For 0: given $r > 0$, choose n with $\frac{1}{n} < r$; then $\frac{1}{n} \in K$, $\frac{1}{n} \neq 0$, and $\frac{1}{n}$ is within r of 0, so 0 is a cluster point as well (3.2.1).

No point outside L is a cluster point. The key is that L is closed: its only cluster point is 0, since $\frac{1}{n} \to 0$ while the terms march away to the right (3.2.5), and $0 \in L$. Take $p \notin L$; in particular $p \neq 0$. If $p < 0$ then the ball $B(p, |p|)$ misses $K \subseteq [0, 2]$ entirely, so p is trivially not a cluster point; assume henceforth $p > 0$. Because $p \notin L$ and L is closed, p is not a cluster point of L, so some ball $B(p, \rho)$ with $\rho > 0$ is disjoint from L; shrinking ρ, arrange in addition that $\rho < \frac{p}{2}$, so that $B(p, \rho) \subseteq (\frac{p}{2}, p + \rho)$ stays a fixed distance $\frac{p}{2} > 0$ to the right of 0. I claim $B(p, \rho)$ contains only finitely many points of K. Each remaining family $\{\frac{1}{n} + \frac{1}{nk} : k \geq 1\}$ converges to $\frac{1}{n}$ as $k \to \infty$, and $\frac{1}{n} + \frac{1}{nk} \leq \frac{2}{n}$, so once $\frac{2}{n} \leq \frac{p}{2}$ the *entire* family for that n lies inside $(0, \frac{p}{2}]$, hence to the left of $B(p, \rho)$ and disjoint from it. Only the finitely many indices n with $\frac{2}{n} > \frac{p}{2}$ survive; for each such n the sequence $\frac{1}{n} + \frac{1}{nk} \to \frac{1}{n} \notin B(p, \rho)$, so only finitely many of its terms can lie in $B(p, \rho)$. A finite union of finite sets is finite, so $B(p, \rho) \cap K$ is finite, and a cluster point would force infinitely

many points of K into every ball around it (a cluster point is approached infinitely often). Hence p is not a cluster point of K.

The cluster points of K are thus exactly the points of $L = \{0\} \cup \{\frac{1}{n}\}$, a subset of the countable set $\mathbb{Q}$, hence countable (a subset of a countable set is countable). Finally K contains all its cluster points, so K is closed; being closed and bounded in $\mathbb{R}$, it is compact by Heine–Borel (4.3.5; closed and bounded $\Rightarrow$ compact).

■

REFLECTIONS.

Two demands sit in the statement and they pull against each other, which is the first thing to read. *Compact* in $\mathbb{R}$ means closed and bounded (Heine–Borel, 4.3.5), and "closed" insists the set already *owns* every one of its cluster points (3.2.5); yet the problem wants the cluster set to be *countably infinite*—an honest, unbounded-in-cardinality supply of accumulation points. So the design problem is to engineer infinitely many places where the set piles up (3.2.1), keep them all inside a bounded interval, and then make sure not a single accumulation point has been left out, or closedness fails and compactness with it.

The cheapest accumulation point, as in Problem 5, is a sequence converging to a limit it never reaches—$\frac{1}{n}$ bunching against 0. There the trick was three such strings aimed at three separated targets. The leap here is to let the *targets* be a convergent sequence in their own right: aim one little string at each of $1, \frac{1}{2}, \frac{1}{3}, \ldots$, by hanging $\frac{1}{n} + \frac{1}{nk}$ $(k \to \infty)$ off the point $\frac{1}{n}$. Believe the picture before proving anything: clusters of dots bunch against each $\frac{1}{n}$, and the $\frac{1}{n}$ themselves bunch against 0—accumulation at two scales, a fractal-flavoured staircase. That second scale is exactly why 0 must be thrown in by hand; it is a cluster point of the targets, so it is a cluster point of K, and an unowned cluster point is precisely what would make K non-closed.

The proof then has four moves, each resting on the one before or on a quotable fact:

(1) boundedness is read straight off the formula—$\frac{1}{n} + \frac{1}{nk} \leq 2$ with equality at $n = k = 1$, so $K \subseteq [0, 2]$ (4.3.4)—and it never comes into doubt again;

(2) that every $\frac{1}{n}$ and 0 *is* a cluster point is the easy direction, falling out of the two convergences, with the Archimedean property (Archimedean) producing, for any radius, an index that lands a genuine point of K inside the ball;

(3) that *nothing else* is a cluster point is the real labour, and the lever is that a cluster point drags infinitely many points of K into every ball around it (a cluster point is approached infinitely often); so a candidate $p \notin L$ is acquitted the moment one ball around it is shown to meet K only finitely often—which works because L is closed, giving a ball about p that avoids L and 0, after which the bound $\frac{1}{n} + \frac{1}{nk} \leq \frac{2}{n}$ sweeps all but finitely many whole families down near 0 and out of the ball;

(4) with the cluster set pinned to $L = \{0\} \cup \{\frac{1}{n}\} \subseteq \mathbb{Q}$, countability is immediate (a subset of a countable set is countable), and $K \supseteq L$ means K owns its cluster points, so K is closed and therefore compact (4.3.5).

The rigor hinges on two places where the construction could quietly leak. First, the secondary limit 0: omit it and K has a cluster point it does not contain, so K fails to be closed and Heine–Borel never fires—this is the single most common way the problem is botched. Second, the uniform bound $\frac{1}{n} + \frac{1}{nk} \leq \frac{2}{n}$ is what collapses an *entire* family down toward 0 at once, clear of a fixed ball about p; without controlling the whole family for large n one cannot reduce to finitely many indices, and the "finitely many points" argument stalls. Both halves of "the cluster set is exactly L" are needed: showing L is contained would not stop some rogue point from also being a cluster point, and showing nothing outside L qualifies would not confirm the $\frac{1}{n}$ actually do.

The transferable move is the engineering recipe and its closure tax. To prescribe the cluster set of a compact set, choose your target points, hang a convergent sequence on each to force accumulation exactly there, and then—this is the part the bookkeeping demands—include every cluster point of the targets themselves, so that the finished set owns all its accumulation points and is closed (adding the cluster points closes a set). Countably many targets accumulating at a single new point give a countable cluster set; the same idea iterated builds the Cantor–Bendixson towers that classify how complicated a closed set's accumulation structure can be.

Problem 14 *Rudin, Chapter 2, Exercise 14*

Give an example of an open cover of the segment $(0, 1)$ which has no finite subcover.

Solution.

For each integer $n \geq 2$ put

$$G_n = \left(\tfrac{1}{n}, 1\right),$$

an open interval, and consider the family $\{G_n : n \geq 2\}$.

This is an open cover of $(0, 1)$. Given $x \in (0, 1)$, the Archimedean property supplies an integer n with $n > 1/x$; then $\frac{1}{n} < x < 1$, so $x \in G_n$. Hence $(0, 1) \subseteq \bigcup_{n\geq 2} G_n$, and since each $G_n \subseteq (0, 1)$ the union is exactly $(0, 1)$.

No finite subcollection covers $(0, 1)$. Suppose $G_{n_1}, \ldots, G_{n_k}$ are finitely many of these sets, and let $N = \max\{n_1, \ldots, n_k\}$. The intervals are nested, $G_n \subseteq G_N$ whenever $n \leq N$, so

$$\bigcup_{j=1}^{k} G_{n_j} = G_N = \left(\tfrac{1}{N}, 1\right).$$

This set misses the point $\frac{1}{N} \in (0, 1)$ (in fact every point of $(0, \frac{1}{N}]$), so it does not cover

$(0,1)$. Thus the open cover $\{G_n\}$ admits no finite subcover.

■

REFLECTIONS.

The phrase "open cover ... with no finite subcover" is the literal negation of the definition of compactness (3.4.1): a set is compact when *every* open cover has a finite subcover, so to certify a set is not compact you produce one cover that defeats every finite selection. The task is therefore not to study $(0,1)$ in the abstract but to engineer a single bad cover, and the shape of $(0,1)$—an open interval missing its endpoints—tells you where to aim. The trouble must sit at a missing endpoint, since that is the only thing distinguishing $(0,1)$ from the compact $[0,1]$; the notes flag exactly this in their two standard non-compact examples (3.4.2).

Picture it before proving it. Slide a window $(\frac{1}{n},1)$ leftward as n grows: each window swallows more of the interval, and any fixed $x>0$ is eventually inside one. But the left edges $\frac{1}{n}$ march toward 0 without ever arriving, so no *single* window—and hence no finite batch of them, which reaches only as far left as the smallest edge—can reach all the way down to 0. The cover chases the absent endpoint 0 and never catches it.

The argument has two halves, and each rests on something concrete:

(1) the family covers $(0,1)$ because every $x>0$ has some $\frac{1}{n}<x$, which is precisely the Archimedean property ($\inf\{1/n\}=0$); this is the only place the ambient field does any work;

(2) no finite subfamily covers, because finitely many indices have a largest one N, the intervals are nested so their union collapses to the single largest interval $(\frac{1}{N},1)$, and that interval visibly omits $\frac{1}{N}$.

The rigor turns entirely on the word *finite* in step (2): an infinite subfamily—say all of them—does cover, so the obstruction is not the cover itself but the impossibility of choosing finitely many edges that reach 0. Take a finite set of indices, extract its maximum, and the nesting hands you a single offending interval; no maximum exists for the full infinite family, which is exactly why the whole cover succeeds where every finite part fails. That contrast is the entire content. (One can also confirm non-compactness through Heine–Borel: $(0,1)$ is bounded but not closed, since 0 is a cluster point lying outside it, and a compact subset of $\mathbb{R}^k$ is closed and bounded would force closedness. But the exercise asks for the explicit cover, which exhibits the failure directly rather than appealing to a theorem.)

The transferable move is the recipe for "show X is not compact": find the point the space is missing and build a cover whose members each stop just short of it, indexed so that any finite subcollection has a last member and therefore still falls short. The same template handles $(0,1/n)$ shrinking to a missing 0 or $[n,\infty)$ running off to infinity, the two failures

> the notes single out (3.4.2); a finite subfamily always has an extremal index, and that is what you turn against the cover.

Problem 15 *Rudin, Chapter 2, Exercise 15*

Show that Theorem 2.36 and its Corollary become false (in $\mathbb{R}^1$, for example) if the word "compact" is replaced by "closed" or by "bounded."

Solution.

Rudin's Theorem 2.36 is our finite-intersection property (4.2.1): a family of compact sets whose finite subfamilies all have nonempty intersection has nonempty total intersection. Its Corollary is our nested statement (4.2.2, the quotable nested nonempty compact sets meet): a decreasing chain $K_1 \supseteq K_2 \supseteq \cdots$ of nonempty compact sets has $\bigcap_n K_n \neq \varnothing$. Each fails over $\mathbb{R}^1$ once "compact" is weakened to "closed" or to "bounded," and a single nested family settles both the theorem and its corollary at once, since a nested chain automatically has the finite-intersection property.

Weaken "compact" to "closed" first. For $n \in \mathbb{N}$ with $n \geq 1$ put

$$F_n = [n, \infty).$$

Each F_n is nonempty and closed (its complement $(-\infty, n)$ is open, 3.2.8), and the chain is nested: $F_1 \supseteq F_2 \supseteq \cdots$. Any finite subfamily intersects in its smallest member,

$$F_{n_1} \cap \cdots \cap F_{n_k} = [N, \infty) \neq \varnothing, \qquad N = \max\{n_1, \ldots, n_k\},$$

so the family has the finite-intersection property. Yet $\bigcap_{n \geq 1} F_n = \varnothing$: a point in the intersection would be a real number $x \geq n$ for every n, which the Archimedean property forbids ($\mathbb{R}$ is Archimedean, 1.7.1). Thus closed, nonempty, nested sets with the finite-intersection property can have empty intersection.

Now weaken "compact" to "bounded" instead. For $n \in \mathbb{N}$ with $n \geq 1$ put

$$B_n = \left(0, \tfrac{1}{n}\right).$$

Each B_n is nonempty and bounded (contained in $[0, 1]$, 4.3.4), and the chain is nested: $B_1 \supseteq B_2 \supseteq \cdots$. Any finite subfamily intersects in its smallest member,

$$B_{n_1} \cap \cdots \cap B_{n_k} = \left(0, \tfrac{1}{N}\right) \neq \varnothing, \qquad N = \max\{n_1, \ldots, n_k\},$$

so again the finite-intersection property holds. Yet $\bigcap_{n \geq 1} B_n = \varnothing$: a point in the intersection would be a real x with $0 < x < \frac{1}{n}$ for every n, contradicting that $\inf\{1/n\} = 0$, the Archimedean property once more ($\mathbb{R}$ is Archimedean, 1.7.1). Thus bounded, nonempty, nested sets with the finite-intersection property can have empty intersection.

In both families the sets are nested, so each is simultaneously a counterexample to Theorem 2.36 (via the finite-intersection property) and to its nested Corollary. What each family drops is one half of "closed and bounded," which by Heine–Borel (compact in $\mathbb{R}^k$ means closed and bounded, 4.3.5) is exactly compactness: $\{[n, \infty)\}$ keeps closed but discards bounded, and $\{(0, 1/n)\}$ keeps bounded but discards closed.

■

REFLECTIONS.

The instruction is unusually direct—"show the theorem becomes false if you weaken the hypothesis"—so the reading is not which tool to invoke but which hypothesis to attack. Theorem 2.36, our finite-intersection property (4.2.1), and its Corollary, the nested statement (4.2.2, the quotable nested nonempty compact sets meet), both turn a local promise (every finite subfamily meets) into a global one (the whole family meets), and the only hypothesis paying for that leap is compactness. So the task is to find a family that satisfies everything *except* compactness and still fails to meet—one family per weakening the problem names, "closed" and "bounded."

The right picture comes from asking what compactness buys here. In $\mathbb{R}^1$ Heine–Borel (closed and bounded, 4.3.5) factors compactness into two independent conditions, so to break the theorem you keep one and sabotage the other, and there are exactly two ways a nested chain of nonempty sets can drain to nothing: it can escape to infinity, or it can shrink onto a point it never reaches. Escaping to infinity, $[n, \infty)$, stays closed but is unbounded; shrinking to a missing endpoint, $(0, 1/n)$, stays bounded but is not closed. Each keeps exactly the half the other discards.

Watching the closed family run shows where each requirement is spent:

(1) the sets $F_n = [n, \infty)$ are closed because each complement $(-\infty, n)$ is open (3.2.8), and they are nested by construction, $F_1 \supseteq F_2 \supseteq \cdots$;

(2) nestedness hands over the finite-intersection property for free—a finite subfamily meets in its smallest member $[N, \infty)$, which is nonempty—so the hypothesis of Theorem 2.36 is genuinely satisfied and not quietly skipped;

(3) the intersection is empty because no real number exceeds every natural number, which is the Archimedean property ($\mathbb{R}$ is Archimedean, 1.7.1)—the single fact that powers the collapse, and the thing compactness would have prevented by forbidding the escape to infinity.

The bounded family $B_n = (0, 1/n)$ runs identically, with the same Archimedean fact appearing as $\inf\{1/n\} = 0$: the chain shrinks toward 0, but 0 was deleted from every B_n, so nothing survives.

The rigour to check is that the finite-intersection property really holds—it does, and nestedness is why, which is also the economy of the example: a nested chain is automatically a finite-intersection family, so one family disproves both Theorem 2.36 and its nested Corollary, no second construction needed. It is worth seeing that neither half alone is enough and each is load-bearing: drop bounded and $[n, \infty)$ kills it; drop closed and $(0, 1/n)$ kills it; restore both and Heine–Borel makes the sets compact, at which point 4.2.2 applies and the intersection must be nonempty. There is no circularity—the counterexamples never assume the conclusion, they exhibit it failing.

The transferable move is the diagnostic habit behind "show this becomes false." When a theorem bundles two conditions into one named hypothesis, a sharp counterexample keeps one condition and breaks the other, isolating which half does the work; here Heine–Borel is the bundle, and the two escapes a nested real chain can make—off to infinity, or onto a deleted limit—are precisely the two ways compactness can be missing while the finite-intersection property survives.

Problem 16 *Rudin, Chapter 2, Exercise 16*

Regard $\mathbb{Q}$, the set of all rational numbers, as a metric space, with $d(p, q) = |p - q|$. Let E be the set of all $p \in \mathbb{Q}$ such that $2 < p^2 < 3$. Show that E is closed and bounded in $\mathbb{Q}$, but that E is not compact. Is E open in $\mathbb{Q}$?

Solution.

E is bounded and is both closed and open in $\mathbb{Q}$, yet it is not compact; so the answer to the last question is yes.

No rational squares to 2 or to 3 (1.4.1; the same parity/divisibility argument settles 3), so for $p \in \mathbb{Q}$ the condition $2 < p^2 < 3$ is exactly $\sqrt{2} < |p| < \sqrt{3}$. Writing

$$U = (-\sqrt{3}, -\sqrt{2}) \cup (\sqrt{2}, \sqrt{3}), \qquad C = [-\sqrt{3}, -\sqrt{2}] \cup [\sqrt{2}, \sqrt{3}],$$

both subsets of $\mathbb{R}$, we therefore have $E = \mathbb{Q} \cap U$ *and* $E = \mathbb{Q} \cap C$: since $\pm\sqrt{2}, \pm\sqrt{3} \notin \mathbb{Q}$, adjoining or deleting those four endpoints changes nothing inside $\mathbb{Q}$.

The set E is bounded, since every $p \in E$ has $p^2 < 3 < 4$, so $|p| < 2$; thus $E \subseteq B_2(0)$ in $\mathbb{Q}$.

It is open in $\mathbb{Q}$, being the trace $E = \mathbb{Q} \cap U$ of the set U, which is open in $\mathbb{R}$; the subspace criterion that the open sets of $\mathbb{Q}$ are exactly the intersections of $\mathbb{Q}$ with open sets of $\mathbb{R}$ then applies (3.3.1).

It is also closed in $\mathbb{Q}$. Since C is closed in $\mathbb{R}$, its complement $\mathbb{R} \setminus C$ is open (3.2.8), so the trace $\mathbb{Q} \setminus E = \mathbb{Q} \cap (\mathbb{R} \setminus C)$ is open in $\mathbb{Q}$ by the same subspace criterion, and its complement E is therefore closed in $\mathbb{Q}$ (3.2.8).

The set E is not compact. For each integer $n \geq 1$ set

$$G_n = \left\{ q \in \mathbb{Q} : \ q^2 < 3 - \tfrac{1}{n} \right\} = \mathbb{Q} \cap (-\sqrt{3 - \tfrac{1}{n}}, \ \sqrt{3 - \tfrac{1}{n}}),$$

open in $\mathbb{Q}$ as the trace of an open interval. These cover E: if $q \in E$ then $3 - q^2 > 0$, so by the Archimedean property (1.7.1; List of Facts) $\frac{1}{n} < 3 - q^2$ for some n, whence $q \in G_n$. The G_n increase with n, so any finite subfamily lies inside a single G_N, the one of largest index. But G_N omits every $q \in E$ with $\sqrt{3 - \frac{1}{N}} \leq |q| < \sqrt{3}$, and such a rational exists: since $\mathbb{Q}$ is dense in $\mathbb{R}$ (1.7.2; List of Facts), the interval $(\sqrt{3 - \frac{1}{N}}, \sqrt{3})$ contains some $q \in \mathbb{Q}$, and then $2 < q^2 < 3$ puts $q \in E \setminus G_N$. No finite subfamily covers E, so E is not compact (3.4.1).

■

REFLECTIONS.

The phrase that should stop you is "regard $\mathbb{Q}$ *as a metric space*": the problem is quietly warning that every topological word—open, closed, compact, bounded—is to be read inside $\mathbb{Q}$, not inside $\mathbb{R}$. In $\mathbb{R}^k$ those words are tightly linked by Heine–Borel, which equates compactness with closed-and-bounded (4.3.5; List of Facts); seeing a set that is closed and bounded yet *not* compact is the standard signal that the ambient space has been changed, and that the link Heine–Borel forges is a property of $\mathbb{R}^k$ rather than of sets in the abstract. So the real content is to take each property relative to $\mathbb{Q}$ and watch what survives.

Believe it through one picture before proving anything. In $\mathbb{R}$ the set lives in two open intervals with the irrational endpoints $\pm\sqrt{2}, \pm\sqrt{3}$; passing to $\mathbb{Q}$ deletes those endpoints from the universe entirely. A set whose only boundary points have been removed from the space has no boundary left to own or to miss, so it comes out both open and closed—and a sequence in E creeping up toward $\sqrt{3}$ is trying to converge to a point that simply is not there, which is exactly how compactness will die.

The cleanest route is to stop fighting the metric pointwise and use the subspace description. The engine is that E is squeezed between irrational thresholds, so it equals the trace on $\mathbb{Q}$ of *both* the open set U and the closed set C that differ only at those absent endpoints:

(1) because no rational squares to 2 or 3 (1.4.1), the rational solutions of $2 < p^2 < 3$ are the same whether the endpoints are included or not, giving the two readings $E = \mathbb{Q} \cap U = \mathbb{Q} \cap C$ on which everything else rests;

(2) boundedness is immediate from $p^2 < 3 < 4$, so $|p| < 2$ for all $p \in E$, with no reference to the space at all;

(3) openness in $\mathbb{Q}$ is the trace of the open U, by the subspace criterion that the open sets of $\mathbb{Q}$ are exactly the intersections $\mathbb{Q} \cap V$ with V open in $\mathbb{R}$ (3.3.1);

(4) closedness in $\mathbb{Q}$ follows by taking complements: C closed makes $\mathbb{R} \setminus C$ open (3.2.8), its trace $\mathbb{Q} \setminus E$ is open in $\mathbb{Q}$ by the same criterion, and so E, its complement, is closed (List of Facts);

(5) noncompactness is exhibited by hand: the increasing cover $G_n = \{q \in \mathbb{Q} : q^2 < 3 - \frac{1}{n}\}$ catches each point of E once $\frac{1}{n}$ drops below $3 - q^2$, which the Archimedean property guarantees (1.7.1), yet a finite subfamily reduces to the largest G_N, and density of $\mathbb{Q}$ (1.7.2) plants a rational in $(\sqrt{3 - \frac{1}{N}}, \sqrt{3})$ that G_N leaves uncovered, so no finite subfamily suffices (3.4.1).

The argument is honest about where each space-dependent fact is spent. Openness and closedness both ride entirely on the endpoints being irrational—were $\sqrt{3}$ rational, C and U would no longer have the same trace and E would lose one of the two properties. The noncompactness needs both the Archimedean property (every point is eventually covered) and density (the finite stage always leaves a rational stranded near the missing $\sqrt{3}$); drop density and one could not produce the witness $q \in E \setminus G_N$. There is no circularity: the cover is built explicitly and the failure is shown directly from the definition of compactness, not from any converse of Heine–Borel.

The move to carry away is the subspace dictionary itself—a set is open (closed) in $\mathbb{Q}$ exactly when it is the trace of an open (closed) set of $\mathbb{R}$, so questions about the induced metric become questions about $\mathbb{R}$ that you already know how to answer. And the cautionary half is the lasting one: "closed and bounded" is only as strong as the space it lives in. Heine–Borel (4.3.5) is a theorem about $\mathbb{R}^k$; over $\mathbb{Q}$ a set can be closed, bounded, and still riddled with missing limits, and those holes—here the irrational $\pm\sqrt{2}, \pm\sqrt{3}$—are precisely what a finite subcover cannot corner.

Problem 17 *Rudin, Chapter 2, Exercise 17*

Let E be the set of all $x \in [0, 1]$ whose decimal expansion contains only the digits 4 and 7. Is E countable? Is E dense in $[0, 1]$? Is E compact? Is E perfect?

Solution.

E is uncountable, not dense in $[0, 1]$, compact, and perfect.

Each point of E is determined by a choice of one digit from $\{4, 7\}$ in every decimal place, so write

$$x = 0.a_1a_2a_3\cdots = \sum_{n=1}^{\infty} a_n\, 10^{-n}, \qquad a_n \in \{4, 7\}.$$

Distinct digit sequences give distinct points: if (a_n) and (b_n) first differ at place m, with $a_m > b_m$ say, then $a_m - b_m = 3$, while the tails contribute at most $\sum_{n>m} 3 \cdot 10^{-n} = 10^{-m}/3 < 3 \cdot 10^{-m}$ to either number, so the two values cannot coincide. Hence the map $(a_n) \mapsto x$ is a bijection of $\{4, 7\}^{\mathbb{N}}$ onto E. The set of all such sequences is uncountable

(2.1.10, relabelling $\{4,7\}$ as $\{0,1\}$), so E is uncountable.

E is not dense in $[0,1]$. Every $x \in E$ has first digit 4 or 7, so $x \geq 0.4\overline{4} = 4/9$; thus $E \cap (0, 4/9) = \varnothing$. A point like $1/9 = 0.111\cdots$ has a neighbourhood, say $(0, 4/9)$, that meets no point of E, so $\overline{E} \neq [0,1]$.

E is closed. For $m \geq 1$ and each finite block $(a_1, \ldots, a_m) \in \{4,7\}^m$ form the length-10^{-m} closed interval

$$I_{(a_1,\ldots,a_m)} = \Big[\, s,\ s + 10^{-m} \Big], \qquad s = \sum_{j=1}^{m} a_j \, 10^{-j},$$

and let $F_m = \bigcup_{(a_1,\ldots,a_m)\in\{4,7\}^m} I_{(a_1,\ldots,a_m)}$ be the union of the 2^m of them. Each F_m is a finite union of closed sets, hence closed (3.2.10). I claim

$$E = \bigcap_{m=1}^{\infty} F_m.$$

If $x \in E$ then for every m its first m digits lie in $\{4,7\}$, so x lies in the corresponding interval and $x \in F_m$. Conversely suppose $x \in \bigcap_m F_m$; for each m the interval of F_m containing x pins the first m decimal digits of x to values in $\{4,7\}$, and these choices are consistent as m grows, so x has a decimal expansion using only 4 and 7 and $x \in E$. Being an intersection of closed sets, E is closed (3.2.10).

E is compact. It is closed, and $E \subseteq [0,1]$ is bounded, so by Heine–Borel a closed bounded subset of $\mathbb{R}$ is compact (4.3.5).

E is perfect: it is closed, and every point is a limit point of E. Fix $x = 0.a_1a_2\cdots \in E$ and any radius $\varepsilon > 0$. Choose m with $3 \cdot 10^{-m} < \varepsilon$, and let $x_m \in E$ be the point obtained by switching the mth digit ($4 \leftrightarrow 7$) and keeping the rest. Then $x_m \in E$, $x_m \neq x$, and $|x_m - x| = 3 \cdot 10^{-m} < \varepsilon$, so every ball about x contains a point of E other than x. Thus x is a cluster point of E (3.2.1). A closed set all of whose points are cluster points is perfect, so E is perfect.

■

REFLECTIONS.

The statement is a checklist—countable, dense, compact, perfect—and the way to read it is to notice what the *description* of E controls: each point is a free choice of one of *two* symbols in *every* decimal place. "Two choices in each of countably many slots" is the exact signature of the binary sequences, whose uncountability is the headline theorem of the counting lecture (2.1.10); so the moment you see "only the digits 4 and 7" you should expect E to be in bijection with $\{0,1\}^{\mathbb{N}}$ and hence uncountable. The remaining three questions are topological, and the word that unlocks all of them is the one hiding in "compact" and "perfect": both are built on *closed* (3.2.4), so the real engine of the problem is a single structural fact—that E is closed—reused three times.

Believe the picture first. Pinning the first digit to $\{4,7\}$ confines x to two short intervals inside $[0.4, 0.8]$; pinning the first two digits splits each into two shorter ones, and so on. What you are watching is a base-ten Cantor set: at stage m you keep 2^m intervals of length 10^{-m} and discard everything between them, and E is what survives forever. That image makes every verdict visible—it is a dust of uncountably many points, it sits entirely above $4/9$ so it cannot be dense, and it has no isolated specks because each surviving interval always contains two of the next stage's.

The four verdicts then come out in order, each leaning on a named result:

(1) the digit-to-point map is a bijection onto E, the one delicate point being that distinct sequences give distinct numbers—a digit gap of 3 at the first disagreement outweighs the entire tail—so E inherits uncountability from $\{4,7\}^{\mathbb{N}}$ (2.1.10);

(2) density fails for a concrete reason, not by vague smallness: every point of E is at least $0.4\overline{4} = 4/9$, so the whole interval $(0, 4/9)$ is missed and points like $1/9$ have a neighbourhood disjoint from E, whence $\overline{E} \neq [0,1]$;

(3) closedness is the crux, and the move is to write E as the nested intersection $\bigcap_m F_m$ of the stage-m unions of closed intervals—an intersection of closed sets is closed (3.2.10), and the equality holds because lying in every F_m forces every finite initial digit-block into $\{4,7\}$, which is exactly membership in E;

(4) compactness is then immediate from Heine–Borel, since E is closed and bounded in $\mathbb{R}$ (4.3.5), and perfectness adds only the limit-point check: flip a far-out digit to produce a distinct point of E within $3 \cdot 10^{-m}$, so every point is a cluster point (3.2.1).

Two checks keep the argument honest. "Not dense" is not the same as "small"—E is uncountable yet fails density, because density is about *reaching every neighbourhood*, not about cardinality; the gap below $4/9$ is what kills it. And perfect demands *both* halves: closed and no isolated points. Closedness alone is not perfect (a single point is closed), and "every point a limit point" alone is not perfect either (the open interval $(0,1)$ qualifies but is not closed); E passes both, the digit-flip supplying nearby companions and the intersection presentation supplying closedness, so no step is circular.

The transferable move is the digit-restriction construction itself. Whenever a set is carved out by forcing each entry of an infinite expansion into a fixed finite menu, read off its properties from the menu: two-or-more choices per slot make it uncountable, a forbidden leading value makes it miss an interval, and the stage-by-stage nested-interval picture makes it a closed—and in fact perfect—Cantor-type set. The very same recipe, with zeros wedged between the free digits, builds a perfect set of irrationals in Problem 18.

Problem 18 *Rudin, Chapter 2, Exercise 18*

Is there a nonempty perfect set in $\mathbb{R}^1$ which contains no rational number?

SOLUTION.

Yes. The verdict is built from a single device: take the digit-restriction trick of Problem 17, which already yields a perfect set, and space the free digits out so that ever-longer blocks of forced zeros sit between them—those zero blocks are what bar the rationals. Write $T_n = \frac{n(n+1)}{2}$ for the n-th triangular number, so $T_1, T_2, T_3, \cdots = 1, 3, 6, 10, \ldots$, and put

$$P = \Big\{ \textstyle\sum_{n=1}^{\infty} a_n \, 10^{-T_n} \;:\; a_n \in \{4, 7\} \text{ for every } n \Big\}.$$

A point of P is the real number whose decimal expansion carries a 4 or a 7 in each triangular place T_n and a 0 in every other place. The series converges absolutely (its terms are dominated by $7 \cdot 10^{-T_n}$), so P is a well-defined nonempty subset of $[0, 1]$; we show it is perfect—closed with no isolated point—and disjoint from $\mathbb{Q}$.

P is closed. For each $m \geq 1$ and each choice $(a_1, \ldots, a_m) \in \{4, 7\}^m$ of the first m free digits, the points of P whose leading free digits are exactly $a_1, \ldots, a_m$ all share the partial sum $s = \sum_{n=1}^{m} a_n 10^{-T_n}$, and the unwritten tail $\sum_{n>m} a_n 10^{-T_n}$ is a nonnegative number at most $\sum_{k>T_m} 9 \cdot 10^{-k} = 10^{-T_m}$; so each such point lies in the closed interval $[\, s, \; s + 10^{-T_m} \,]$. Let C_m be the union of these 2^m closed intervals, one per choice of $(a_1, \ldots, a_m)$. Each C_m is a finite union of closed sets, hence closed (3.2.10; a finite union of closed sets is closed). Now $P = \bigcap_{m \geq 1} C_m$: every point of P lies in every C_m by the bound just given, while a real number in all the C_m has, for each m, its first m free digits drawn from $\{4, 7\}$ and zeros elsewhere, so all its free digits are in $\{4, 7\}$ and it belongs to P. An arbitrary intersection of closed sets is closed (3.2.10; any intersection of closed sets is closed), so P is closed.

Every point of P is a cluster point of P. Fix $x = \sum_{n \geq 1} a_n 10^{-T_n} \in P$ and any radius $r > 0$. Choose m with $3 \cdot 10^{-T_m} < r$, and let x_m be the point of P obtained by flipping the single digit a_m ($4 \leftrightarrow 7$) and leaving every other a_n unchanged. Then $x_m \in P$, $x_m \neq x$, and

$$|x_m - x| = |a_m - a_m'| \, 10^{-T_m} = 3 \cdot 10^{-T_m} < r,$$

so the ball $(x - r, x + r)$ contains a point of P other than x. Thus x is a cluster point of P (3.2.1). With P closed and every point a cluster point, P is perfect; and $P \neq \varnothing$ (for instance $\sum_n 4 \cdot 10^{-T_n} \in P$).

No point of P is rational. A rational has an eventually periodic decimal expansion: dividing by a denominator q leaves only q possible remainders, so the remainders—and hence the quotient digits— must eventually cycle. Suppose $x \in P$ were rational, with its expansion eventually periodic of period ℓ from some place onward. The gaps between consecutive triangular places grow without bound, $T_{n+1} - T_n = n + 1 \to \infty$, so beyond any point of

the expansion there is a run of more than ℓ consecutive zeros (the stretch strictly between two far-apart free places). A periodic block that contains ℓ consecutive zeros is the all-zero block, forcing the expansion to be eventually 0; but P has a nonzero digit 4 or 7 in every triangular place T_n, arbitrarily far out. This contradiction shows x is irrational. Hence P is a nonempty perfect set containing no rational number.

■

Reflections.

The word *perfect* is the whole instruction set. In our vocabulary it unpacks into two demands that must be met together: the set is closed—it owns every one of its cluster points (3.2.4)—and it has no isolated point, so every point is itself a cluster point (3.2.1, 3.2.2). The extra clause "contains no rational" is a number-theoretic constraint laid on top, and the tension between the two is what makes the problem more than a definition-chase: rationals are *dense* ($\mathbb{Q}$ is dense in $\mathbb{R}$, 1.7.2), so any interval is riddled with them, which is exactly why a perfect set with room to spread—had it contained an interval—could never dodge them. The set must be perfect yet nowhere interval-like, a Cantor-type dust.

That points straight at the construction from Problem 17, where restricting decimal digits to $\{4, 7\}$ produced a perfect set with empty interior. Believe that picture first: prescribing a digit at each place and giving it two free choices builds a set that, at every point, has a near-twin got by flipping one far-out digit—so no point stands alone—while leaving gaps no interval can fit through. The single new idea here is to thin the free places out, parking them at the triangular positions $T_n = \frac{n(n+1)}{2}$ so that the runs of forced zeros between them lengthen without bound. Those growing zero runs are not decoration; they are the engine that defeats rationality.

The proof has three independent obligations, each closed by a named result or by the step before it:

(1) closedness comes from exhibiting P as an intersection $\bigcap_m C_m$, where C_m is the finite union of the 2^m closed "decimal intervals" pinned down by the first m free digits; a finite union of closed sets is closed and so is an arbitrary intersection of them (3.2.10; intersections of closed sets are closed), and the bookkeeping that makes $P = \bigcap_m C_m$ exact is that membership in every C_m is the same as having all free digits in $\{4, 7\}$;

(2) that every point is a cluster point is the digit-flip move: flipping the m-th free digit perturbs the value by exactly $3 \cdot 10^{-T_m}$, which the freedom in m drives below any radius r, so every ball around x catches a different point of P—the definition of a cluster point (3.2.1), and equivalently the failure of x to be isolated (3.2.2);

(3) irrationality rests on the one fact about rationals worth quoting here—their decimals are eventually periodic, because long division by q admits only q remainders—played

against the widening zero gaps: past any place there sits a zero run longer than the alleged period, a period meeting a full zero run is all zeros, yet P keeps planting a 4 or 7 arbitrarily far out, and the two cannot coexist.

The rigor turns on where the triangular spacing earns its keep, and it is worth checking by asking what a *constant* spacing would cost. Problem 17's set, with a free digit in every place, is perfect too, but it is stuffed with rationals—$0.4444\cdots = \frac{4}{9}$ for one—because a fixed periodic digit pattern is itself a periodic decimal. Only by letting the gaps grow, $T_{n+1} - T_n = n + 1 \to \infty$, do we guarantee a zero run beating *any* prescribed period, which is precisely the lever the eventual-periodicity test needs; drop the growth and the argument collapses. Closedness and the no-isolated-point property, by contrast, never used the spacing at all—they survive verbatim from Problem 17—so the design cleanly separates "perfect" (handled by the digit menu) from "no rationals" (handled by the spacing). Nothing is circular: closedness is built from the closed-set algebra, the cluster condition from an explicit nearby point, and irrationality from a property of $\mathbb{Q}$ alone.

The transferable lesson is how decimal (or base-b) digit menus turn topological and arithmetic wishes into independent dials. A two-symbol alphabet at chosen places forces perfection and uncountability (two-valued digit sequences are uncountable, 2.1.10); spreading those places out destroys eventual periodicity and so banishes the rationals; thinning them further still controls measure and dimension. When a problem asks for a set that is large in one sense (perfect, uncountable) yet evasive in another (rational-free, interval-free), reach for a Cantor-style digit construction and tune the placement of the free digits to whatever the second condition forbids.

Problem 19 *Rudin, Chapter 2, Exercise 19*

Call two subsets A, B of a metric space X *separated* if $A \cap \overline{B} = \varnothing$ and $\overline{A} \cap B = \varnothing$ — neither meets the closure of the other.

(a) If A and B are disjoint closed sets in some metric space X, prove that they are separated.

(b) Prove the same for disjoint open sets.

(c) Fix $p \in X$, $\delta > 0$, define A to be the set of all $q \in X$ for which $d(p, q) < \delta$, define B similarly, with $>$ in place of $<$. Prove that A and B are separated.

(d) Prove that every connected metric space with at least two points is uncountable. *Hint:* Use (c).

SOLUTION.

Throughout, $\overline{E} = E \cup E'$ is the closure (4.4.1), so a point lies in $\overline{E}$ exactly when it is in E or is a cluster point of E — equivalently, when every ball about it meets E (3.2.1).

(a) Let A, B be disjoint and closed. A closed set equals its own closure (3.2.4), so $\overline{A} = A$

and $\overline{B} = B$. Then $A \cap \overline{B} = A \cap B = \varnothing$ and $\overline{A} \cap B = A \cap B = \varnothing$, so A and B are separated.

(b) Let A, B be disjoint and open. Take any $x \in A$. Since A is open, some ball $N_r(x) \subseteq A$, and $A \cap B = \varnothing$ forces $N_r(x) \cap B = \varnothing$; a ball about x thus misses B, so x is neither in B nor a cluster point of B, i.e. $x \notin \overline{B}$. As $x \in A$ was arbitrary, $A \cap \overline{B} = \varnothing$. Interchanging the roles of A and B (both are open) gives $\overline{A} \cap B = \varnothing$. So A and B are separated.

(c) With

$$A = \{q \in X : d(p,q) < \delta\}, \qquad B = \{q \in X : d(p,q) > \delta\},$$

fix $q \in A$ and set $r = \delta - d(p,q) > 0$. If $d(q,y) < r$, the triangle inequality (2.2.1) gives

$$d(p,y) \leq d(p,q) + d(q,y) < d(p,q) + r = \delta,$$

so $N_r(q) \subseteq A$ and meets B nowhere; hence $q \notin \overline{B}$, and $A \cap \overline{B} = \varnothing$. Symmetrically, fix $q \in B$ and set $r = d(p,q) - \delta > 0$. If $d(q,y) < r$, the triangle inequality gives

$$d(p,y) \geq d(p,q) - d(q,y) > d(p,q) - r = \delta,$$

so $N_r(q) \subseteq B$ misses A, whence $q \notin \overline{A}$ and $\overline{A} \cap B = \varnothing$. Thus A and B are separated.

(d) Let X be connected with two distinct points p, q, and put $D = d(p,q) > 0$. For each $t \in (0, D)$ set

$$A_t = \{x \in X : d(p,x) < t\}, \qquad B_t = \{x \in X : d(p,x) > t\}.$$

The set $A_t = N_t(p)$ is an open ball, hence open (2.3.5; open balls are open); and B_t is open by the argument of (c), since each of its points carries a ball staying in B_t. The two are disjoint, and both are nonempty: $p \in A_t$ (as $0 < t$), while $q \in B_t$ (as $d(p,q) = D > t$).

I claim some $x \in X$ has $d(p,x) = t$. If not, then no point lands on the sphere $d(p,x) = t$, so every point of X falls into A_t or B_t, giving $X = A_t \cup B_t$ as a union of two nonempty, disjoint, open sets — which is precisely a disconnection of X (4.4.2), against the hypothesis that X is connected. So a point x_t with $d(p, x_t) = t$ exists.

The assignment $t \mapsto x_t$ is injective: $d(p, x_t) = t$ recovers t from x_t, so distinct parameters give distinct points. Hence $(0, D)$ injects into X. The interval $(0, D)$ is uncountable — the map $s \mapsto Ds$ carries $(0,1)$ bijectively onto it, and $(0,1)$ is uncountable ($\mathbb{R}$ is uncountable; 2.1.12). An injection of an uncountable set into X forbids X from being countable, so X is uncountable.

■

Reflections.

The load-bearing word is *separated*, and the statement is generous enough to hand you its meaning: A and B are separated when $A \cap \overline{B} = \varnothing$ and $\overline{A} \cap B = \varnothing$. That this is strictly stronger than disjoint is the thing to register first — it forbids each set even from *touching* the other's closure, so the entire problem is about the closure $\overline{E} = E \cup E'$ (4.4.1) and its neighbourhood test: $x \in \overline{E}$ exactly when every ball about x meets E (3.2.1). Once you read "separated" as "a ball walls each point of one set off from the other," parts (a)–(c) become three settings in which such a wall is free, and the closure is the only machinery any of them touches.

The cleanest picture is part (c): the open ball $d(p,x) < \delta$ and its far exterior $d(p,x) > \delta$ sit on opposite sides of the sphere $d(p,x) = \delta$, and the sphere itself is the buffer that keeps either set off the other's closure. The shell of thickness given by "how far q is from the threshold" is exactly the radius that fits a ball staying on q's own side.

Each part rests on a single named fact, applied to closures:

(1) closed disjoint sets are separated because a closed set is its own closure (3.2.4), so $A \cap \overline{B}$ and $\overline{A} \cap B$ collapse to the given $A \cap B = \varnothing$ — nothing to prove beyond reading off the definition;

(2) disjoint open sets are separated because openness gives each $x \in A$ a ball $N_r(x) \subseteq A$, which is disjoint from B, so x flunks the neighbourhood test for $\overline{B}$ and $A \cap \overline{B} = \varnothing$; the same with A, B swapped finishes it;

(3) the two sides of a metric sphere are separated by a quantitative version of the same move: from $q \in A$, the slack $r = \delta - d(p,q)$ feeds the triangle inequality (2.2.1) to put $N_r(q)$ wholly inside A, hence off $\overline{B}$; the mirror estimate $d(p,y) \geq d(p,q) - d(q,y)$ handles $q \in B$;

(4) connectedness turns (c) into a counting bound: for each threshold $t \in (0, D)$ the open ball A_t (open by open balls are open, 2.3.5) and the open exterior B_t are nonempty and disjoint, so if no point sat on the sphere $d(p,x) = t$ they would disconnect X (4.4.2); connectedness therefore forces a witness x_t at every radius, and $t \mapsto x_t$ injects the uncountable interval $(0, D)$ into X.

The rigor of (d) turns on two checkpoints. First, B_t must genuinely be open, and that is not a freebie — it is exactly part (c) reused, which is why the problem stages (c) before (d). Second, the injection has to be honest: $d(p, x_t) = t$ recovers t, so distinct radii give distinct points, and $(0, D)$ is uncountable as an affine copy of $(0, 1)$ ($\mathbb{R}$ is uncountable; 2.1.12). Both hypotheses in (d) are load-bearing: drop "at least two points" and $D = 0$ leaves no interval to inject; drop "connected" and the conclusion fails outright, since $\mathbb{Q}$ as a metric space is countable yet has at least two points — it is its *disconnectedness*, a missing irrational threshold splitting it at every t, that lets it stay small.

The move to carry away is that connectedness is a no-gaps condition you exploit by manufacturing a candidate gap and daring the space to have it. Any clopen splitting is forbidden, so each nonempty disjoint pair of open sets you can build must fail to cover X, and the points it misses are forced to exist; sweeping the threshold across an interval converts "the space has no gaps" into "the space has at least a continuum of points." The same template — present a would-be disconnection, let connectedness veto it, and harvest the witness — is the standard way the clopen definition (4.4.2) does real work.

Problem 20 *Rudin, Chapter 2, Exercise 20*

Are closures and interiors of connected sets always connected? (Look at subsets of $\mathbb{R}^2$.)

Solution.

Closures of connected sets are always connected; interiors need not be.

For the closure, let E be connected and suppose, toward a contradiction, that $\overline{E} = A \cup B$ with A, B nonempty and *separated*, meaning $A \cap \overline{B} = \varnothing$ and $\overline{A} \cap B = \varnothing$. Intersecting with E gives $E = (E \cap A) \cup (E \cap B)$, and these two pieces are separated, since a subset of a separated set stays separated: $E \cap A \subseteq A$ and $E \cap B \subseteq B$, so $(E \cap A) \cap \overline{E \cap B} \subseteq A \cap \overline{B} = \varnothing$ and likewise on the other side. Because E is connected, one piece is empty; say $E \cap A = \varnothing$, so $E \subseteq B$ and hence $\overline{E} \subseteq \overline{B}$. But $A \subseteq \overline{E}$, so $A \subseteq \overline{B}$, forcing $A = A \cap \overline{B} = \varnothing$ against A nonempty. The case $E \cap B = \varnothing$ is identical. Thus no such separation exists and $\overline{E}$ is connected.

For interiors, one counterexample suffices. In $\mathbb{R}^2$ take the two closed unit disks tangent at the origin,

$$E = \{(x, y) : (x + 1)^2 + y^2 \leq 1\} \cup \{(x, y) : (x - 1)^2 + y^2 \leq 1\}.$$

Each disk is convex, hence connected, and the two disks share the point $(0, 0)$, so their union E is connected (a union of two connected sets with a common point cannot be separated, the same argument as for the closure). Its interior, however, is the union of the two *open* disks,

$$E^\circ = \{(x, y) : (x + 1)^2 + y^2 < 1\} \cup \{(x, y) : (x - 1)^2 + y^2 < 1\},$$

because the tangency point $(0, 0)$ lies on both bounding circles and so is not interior to either disk. These two open disks are nonempty, disjoint, and open, hence separated: a disjoint open set cannot meet the closure of another, since each of its points has a neighbourhood (the set itself) missing the other (this is exactly Problem 19(b)). Therefore E° is disconnected, and interiors of connected sets need not be connected.

■

REFLECTIONS.

The phrase *connected set* fixes the entire toolkit. Connectedness is defined negatively—a set is connected when it admits no splitting into two nonempty separated pieces (4.4.2)—and a negative definition is a standing invitation to argue by contradiction (0.5.5): assume the separation you are forbidden and break it. The other two words, *closure* and *interior*, name the operations $\overline{E} = E \cup E'$ and E° from 4.4.1, and the bracketed steer "look at subsets of $\mathbb{R}^2$" is itself a clue: in $\mathbb{R}$ the connected sets are exactly the intervals and both operations keep you among intervals, so any honest counterexample must live in a space with more room—the plane.

Before proving anything, picture why the two halves should differ. Adding boundary points to a connected blob cannot tear it apart—the new points cling to the set they came from—so the closure should stay in one piece. But carving out the boundary can delete the very point that fused two lobes together: two disks kissing at a single point form one connected set, yet erasing that shared point to pass to the interior leaves two disjoint open disks drifting free. That single picture is the whole content of the problem.

The closure half runs through four linked steps:

(1) assume a forbidden separation $\overline{E} = A \cup B$ with A, B nonempty separated, which is the negation of "$\overline{E}$ connected" (4.4.2) and the only foothold a negative hypothesis offers;

(2) restrict it to E, writing $E = (E \cap A) \cup (E \cap B)$, and check the pieces stay separated because a subset of a separated set is separated—shrinking a set only shrinks its closure, so the empty intersections survive;

(3) invoke the connectedness of E itself to kill one piece, say $E \cap A = \varnothing$, which pins $E \subseteq B$ and therefore $\overline{E} \subseteq \overline{B}$ since closure is monotone (4.4.1);

(4) spring the trap: $A \subseteq \overline{E} \subseteq \overline{B}$ collides with the separation condition $A \cap \overline{B} = \varnothing$, forcing the nonempty A to be empty.

The hinge is step (2)'s monotonicity together with the defining asymmetry of separation—separated is strictly more than disjoint, demanding $A \cap \overline{B} = \varnothing$, and it is precisely that closure appearing in the condition that the argument exploits.

The interior half is a single well-chosen example, and its rigor lives in two checks. First, E really is connected: two convex (hence connected) disks meeting at $(0, 0)$ join into one piece, again because two connected sets sharing a point admit no separation. Second, E° really is disconnected: the tangency point sits on both circles, so it is interior to neither disk, and the interior collapses to two disjoint open disks; disjoint open sets are separated, which is Problem 19(b), so the splitting is genuine. The example is sharp—move the disks apart and E disconnects; overlap them and E° reconnects; tangency is the knife's

edge where the closure stays whole but the interior does not. And nothing is circular: the closure direction never assumed what it proved, building the contradiction only from monotonicity of closure and the meaning of separation.

The transferable move is to treat "is this operation connectedness-preserving?" as two opposite questions. Adding closure points cannot create a separation, so closure (and in fact any set squeezed between E and $\overline{E}$) inherits connectedness; deleting points to form an interior can sever a one-point bridge, so it does not. Whenever an operation only *enlarges* toward the closure, suspect permanence and chase a contradiction through monotonicity; whenever it can *remove* a load-bearing point, suspect failure and hunt in $\mathbb{R}^2$ for the tangency that the operation cuts.

Problem 21 *Rudin, Chapter 2, Exercise 21*

Let A and B be separated subsets of some $\mathbb{R}^k$, suppose $\mathbf{a} \in A$, $\mathbf{b} \in B$, and define

$$\mathbf{p}(t) = (1-t)\mathbf{a} + t\mathbf{b}$$

for $t \in \mathbb{R}^1$. Put $A_0 = \mathbf{p}^{-1}(A)$, $B_0 = \mathbf{p}^{-1}(B)$. [Thus $t \in A_0$ if and only if $\mathbf{p}(t) \in A$.]

1. Prove that A_0 and B_0 are separated subsets of $\mathbb{R}^1$.
2. Prove that there exists $t_0 \in (0,1)$ such that $\mathbf{p}(t_0) \notin A \cup B$.
3. Prove that every convex subset of $\mathbb{R}^k$ is connected.

Solution.

Two sets S, T are *separated* when neither meets the other's closure: $\overline{S} \cap T = \varnothing$ and $S \cap \overline{T} = \varnothing$. The single fact that drives every part is that $\mathbf{p}$ moves distances by a fixed factor: for all $s, t \in \mathbb{R}$,

$$|\mathbf{p}(s) - \mathbf{p}(t)| = |(s-t)(\mathbf{b}-\mathbf{a})| = |s-t|\,|\mathbf{b}-\mathbf{a}|,$$

so $t_n \to t$ in $\mathbb{R}$ forces $\mathbf{p}(t_n) \to \mathbf{p}(t)$ in $\mathbb{R}^k$.

(a) The claim reduces to one transfer: if $t \in \overline{A_0}$ then $\mathbf{p}(t) \in \overline{A}$. Take $t \in \overline{A_0} = A_0 \cup A_0'$. If $t \in A_0$ then $\mathbf{p}(t) \in A \subseteq \overline{A}$. If t is a cluster point of A_0, some sequence (t_n) in A_0 has $t_n \to t$ (a cluster point is a sequential limit); then $\mathbf{p}(t_n) \in A$ for every n and $\mathbf{p}(t_n) \to \mathbf{p}(t)$, so $\mathbf{p}(t)$ is a cluster point of A (or already lies in A), hence $\mathbf{p}(t) \in \overline{A}$. The same argument gives $t \in \overline{B_0} \Rightarrow \mathbf{p}(t) \in \overline{B}$.

Now suppose $t \in \overline{A_0} \cap B_0$. Then $\mathbf{p}(t) \in \overline{A}$ by the transfer, and $\mathbf{p}(t) \in B$ since $t \in B_0$, so $\mathbf{p}(t) \in \overline{A} \cap B = \varnothing$ as A, B are separated — impossible. Hence $\overline{A_0} \cap B_0 = \varnothing$, and interchanging the roles of A and B gives $A_0 \cap \overline{B_0} = \varnothing$. Thus A_0 and B_0 are separated.

(b) Endpoints land where they should: $\mathbf{p}(0) = \mathbf{a} \in A$ and $\mathbf{p}(1) = \mathbf{b} \in B$, so $0 \in A_0$ and $1 \in B_0$. Suppose, toward a contradiction, that $\mathbf{p}(t) \in A \cup B$ for every $t \in [0,1]$. Writing $P = [0,1] \cap A_0$ and $Q = [0,1] \cap B_0$, this assumption says $[0,1] = P \cup Q$; moreover $0 \in P$ and $1 \in Q$, so both are nonempty, they are disjoint, and they are separated, being subsets of the separated sets A_0, B_0 from part (a). I show no such split of $[0,1]$ exists.

Let $c = \sup P$. The set P is nonempty (it contains 0) and bounded above by 1, so $c \in [0,1]$ exists by the least-upper-bound property (the supremum property), and being a supremum, $c \in \overline{P}$. Separation forbids $c \in Q$ (else $c \in \overline{P} \cap Q = \varnothing$), so $c \in P$; in particular $c < 1$, since $1 \in Q$. Every $t \in (c,1]$ exceeds $\sup P$, hence lies outside P, hence in Q; letting $t \downarrow c$ shows $c \in \overline{Q}$. Then $c \in P \cap \overline{Q} = \varnothing$, again impossible. This contradiction shows some $t_0 \in [0,1]$ has $\mathbf{p}(t_0) \notin A \cup B$; and since $\mathbf{p}(0), \mathbf{p}(1)$ do lie in $A \cup B$, that $t_0 \in (0,1)$.

(c) Let $C \subseteq \mathbb{R}^k$ be convex, meaning $(1-t)\mathbf{a} + t\mathbf{b} \in C$ whenever $\mathbf{a}, \mathbf{b} \in C$ and $0 \le t \le 1$. Suppose, toward a contradiction, that C is disconnected: $C = A \cup B$ with A, B nonempty, separated (4.4.2). Pick $\mathbf{a} \in A$, $\mathbf{b} \in B$. By convexity $\mathbf{p}(t) \in C = A \cup B$ for all $t \in [0,1]$. But part (b), applied to these separated A, B, produces a $t_0 \in (0,1)$ with $\mathbf{p}(t_0) \notin A \cup B = C$ — contradicting $\mathbf{p}(t_0) \in C$. Hence no such split exists, and C is connected.

■

REFLECTIONS.

The word that fixes the whole problem is *separated*, and reading it correctly is the exercise. "Disjoint" would only forbid a shared point; "separated" is the stronger relation $\overline{A} \cap B = A \cap \overline{B} = \varnothing$ that our notes use to define disconnectedness (4.4.2), and a relation phrased through closures is one you attack with closures. So the part-(a) target, "A_0, B_0 are separated," is really the claim that a closure on the line ($\overline{A_0}$) maps into the matching closure upstairs ($\overline{A}$) — which is exactly what the affine map $\mathbf{p}$, being distance-scaling, delivers. The second clue is the explicit formula $\mathbf{p}(t) = (1-t)\mathbf{a} + t\mathbf{b}$: a parametrised segment turns a geometric statement in $\mathbb{R}^k$ into a question about the interval $[0,1]$ in $\mathbb{R}^1$, the place where "connected" is easiest to settle. The third is the word *convex* in part (c) — convexity is precisely the promise that this segment never leaves C, which is what lets the line problem report back to the set.

Believe it on a picture first. With $A = [0,1)$ and $B = (1,2]$ on the line, the gap at the coordinate 1 belongs to neither set, yet any segment running from a point of A to a point of B must pass through it; so the segment leaves $A \cup B$ at some interior parameter. Convexity forbids the gap from lying outside the set, so a convex set cannot be cut this way.

The argument runs in three linked stages:

(1) part (a) rests on the distance identity $|\mathbf{p}(s) - \mathbf{p}(t)| = |s-t|\,|\mathbf{b} - \mathbf{a}|$, which makes $\mathbf{p}$ carry convergent sequences to convergent sequences; reading $\overline{A_0}$ as A_0 together with

its cluster points (3.2.1, closure = set + cluster points) and pulling a sequence out of any cluster point (cluster points are sequential limits), the image sequence lands in A and converges, so $\mathbf{p}$ sends $\overline{A_0}$ into $\overline{A}$ — and a point of $\overline{A_0} \cap B_0$ would put $\mathbf{p}(t)$ in $\overline{A} \cap B$, which separation has emptied;

(2) part (b) is the connectedness of $[0,1]$ in disguise: a split $[0,1] = P \cup Q$ into nonempty separated pieces cannot exist, and the supremum $c = \sup P$ of the piece holding 0 exhibits the impossibility — $c \in \overline{P}$ keeps c out of Q, yet the values just above c are forced into Q so $c \in \overline{Q}$ keeps it out of P, leaving c nowhere to go. This is the one place completeness enters, through the least-upper-bound property (1.6.1, the supremum property); the assumption that $\mathbf{p}$ stays inside $A \cup B$ is exactly what manufactures the forbidden split, so denying it yields the escape value t_0, and the endpoints pin t_0 to the open interval;

(3) part (c) is pure assembly: convexity holds the segment inside C, so any separation $C = A \cup B$ would feed parts (a) and (b) a segment they force to step outside $A \cup B = C$ — a contradiction (0.5.5), so C admits no separation and is connected (4.4.2).

Each hypothesis earns its place, and dropping one breaks the chain. Without *separated* (only disjoint), part (a)'s closure transfer collapses — $A = [0,1)$, $B = [1,2]$ are disjoint but $1 \in \overline{A} \cap B$, and the segment never escapes $A \cup B$, so the conclusion of (b) genuinely fails. Without *convex* in (c), the set $\{|\mathbf{x}| < 1\} \cup \{|\mathbf{x}| > 2\}$ in $\mathbb{R}^k$ splits into separated pieces and is disconnected, because the segment between the two shells does leave the set. Nothing is circular: I never assumed $[0,1]$ connected as a quoted theorem but built its connectedness from the supremum, the same axiom the notes lean on throughout.

The move to keep is the transport principle. A distance-controlled map carries closures forward, hence separations backward, so connectedness of a complicated set can be tested along its segments by exporting the whole question to the one space where connectedness is transparent — the interval, where the least-upper-bound property settles it. That is why "convex ⇒ connected" sits at the end of the chapter (4.4.2): it is the first payoff of having both completeness and the topology of $\mathbb{R}^k$ in hand at once.

Problem 22 *Rudin, Chapter 2, Exercise 22*

A metric space is called *separable* if it contains a countable dense subset. Show that $\mathbb{R}^k$ is separable. *Hint:* consider the set of points which have only rational coordinates.

SOLUTION.

The set $D = \mathbb{Q}^k = \{(q_1, \ldots, q_k) : q_j \in \mathbb{Q}\}$ is a countable dense subset of $\mathbb{R}^k$, so $\mathbb{R}^k$ is separable.

D is countable. It is the k-fold product of the countable set $\mathbb{Q}$, and a finite product of

countable sets is countable (List of Facts; $\mathbb{Q}$ countable by 2.1.7).

D is dense, meaning every point of $\mathbb{R}^k$ lies in D or is a cluster point of D; equivalently, every open ball meets D. Fix $x = (x_1, \dots, x_k) \in \mathbb{R}^k$ and $\varepsilon > 0$. Because $\mathbb{Q}$ is dense in $\mathbb{R}$ (List of Facts; 1.7.2), for each coordinate choose $q_j \in \mathbb{Q}$ with $|q_j - x_j| < \varepsilon/\sqrt{k}$. Then $q = (q_1, \dots, q_k) \in D$ and

$$|q - x| = \Big(\sum_{j=1}^{k} (q_j - x_j)^2\Big)^{1/2} < \Big(\sum_{j=1}^{k} \frac{\varepsilon^2}{k}\Big)^{1/2} = \Big(k \cdot \frac{\varepsilon^2}{k}\Big)^{1/2} = \varepsilon,$$

so $q \in B_\varepsilon(x)$. Hence every ball about every point of $\mathbb{R}^k$ contains a point of D, which is exactly density (4.4.1). With D countable and dense, $\mathbb{R}^k$ is separable.

■

REFLECTIONS.

The definition is handed to you, so the reading is of the word *separable* itself: it asks for a countable set that approximates every point arbitrarily well, and "approximates every point" is the density condition—closure equal to the whole space (4.4.1)—while "countable" is the counting condition. The hint names the candidate, the rational grid $\mathbb{Q}^k$, and the only real question is why it is the right one. It is right because it is built coordinatewise from $\mathbb{Q}$, which is both countable (2.1.7) and dense in $\mathbb{R}$ (1.7.2); the two properties demanded of D are inherited from those two properties of $\mathbb{Q}$, one for counting and one for approximation.

Picture $k = 2$: $\mathbb{Q}^2$ is the lattice of points with both coordinates rational, a countable web that nonetheless threads arbitrarily close to every point of the plane. Believe that the web is fine enough to enter any disc, and the proof is just making "fine enough" quantitative.

The verification splits cleanly along the definition:

(1) countability is bookkeeping—$\mathbb{Q}^k$ is a product of k copies of the countable set $\mathbb{Q}$, and a finite product of countable sets is countable (the product fact, resting on 2.1.7); the finiteness of k is what keeps the product countable;

(2) density is where the geometry lives—to land a rational point inside the ball $B_\varepsilon(x)$ (2.3.1), approximate each of the k coordinates of x by a rational within $\varepsilon/\sqrt{k}$, which density of $\mathbb{Q}$ in $\mathbb{R}$ supplies ($\mathbb{Q}$ dense; 1.7.2);

(3) the per-coordinate tolerance $\varepsilon/\sqrt{k}$ is then forced by the Euclidean norm: k squared errors each below ε^2/k sum to less than ε^2, so the distance is below ε and the rational point meets the ball, which is the definition of density (4.4.1, via cluster points 3.2.1).

The estimate is honest because the $\sqrt{k}$ is not a fudge but exactly the factor the norm contributes from k coordinates—split the tolerance any more coarsely and the sum of squares could exceed ε^2. The finiteness of k is load-bearing twice over: it is what makes

the product countable, and it is what lets the single split $\varepsilon/\sqrt{k}$ work for every coordinate at once. In an infinite-dimensional space neither step survives unchanged, which is why separability there is a genuine question rather than a formality.

The transferable move is the recipe for separability of a product: take the product of countable dense sets in the factors, and split the tolerance across finitely many coordinates so the norm reassembles them under ε. The same rational grid is the workhorse behind the next exercise—a countable dense set is precisely what lets you build a countable base by centering balls at its points (Problem 23).

Problem 23 *Rudin, Chapter 2, Exercise 23*

A collection $\{V_\alpha\}$ of open subsets of X is said to be a *base* for X if the following is true: for every $x \in X$ and every open set $G \subset X$ such that $x \in G$, we have $x \in V_\alpha \subset G$ for some α. In other words, every open set in X is the union of a subcollection of $\{V_\alpha\}$.
Prove that every separable metric space has a *countable* base. *Hint:* take all neighborhoods with rational radius and center in some countable dense subset of X.

Solution.

Let X be separable, fix a countable dense subset $D = \{x_1, x_2, \dots\}$, and set

$$\mathcal{B} = \{\, B_r(x_i) : x_i \in D,\ r \in \mathbb{Q},\ r > 0 \,\}.$$

Each member is an open ball, hence open, so $\mathcal{B}$ is a collection of open sets; the claim is that it is a countable base.

It is countable. The map $(x_i, r) \mapsto B_r(x_i)$ sends $D \times (\mathbb{Q} \cap (0, \infty))$ onto $\mathcal{B}$, and that product of two countable sets is countable, so $\mathcal{B}$ is at most countable.

It is a base. Let $G \subseteq X$ be open and $x \in G$. Because G is open, x is an interior point, so there is $R > 0$ with $B_R(x) \subseteq G$. Density of D gives a center $x_i \in D$ with $d(x, x_i) < R/4$, and density of $\mathbb{Q}$ in $\mathbb{R}$ gives a rational radius r with

$$d(x, x_i) < r < \tfrac{R}{2}.$$

Then $x \in B_r(x_i)$, since $d(x, x_i) < r$. Moreover $B_r(x_i) \subseteq B_R(x)$: for any $y \in B_r(x_i)$ the triangle inequality gives

$$d(y, x) \le d(y, x_i) + d(x_i, x) < r + \tfrac{R}{4} < \tfrac{R}{2} + \tfrac{R}{4} = \tfrac{3R}{4} < R,$$

so $y \in B_R(x) \subseteq G$. Thus $x \in B_r(x_i) \subseteq G$ with $B_r(x_i) \in \mathcal{B}$.

Since each $x \in G$ sits in some member of $\mathcal{B}$ contained in G, the set G is the union of those members; equivalently every open G is a union of a subcollection of $\mathcal{B}$. Hence $\mathcal{B}$ is

a countable base for X.

■

REFLECTIONS.

The defining clue is the word *base*: a supply of open sets out of which every open set can be rebuilt by taking unions, so that to describe an open G around a point x it is enough to slot one member of the supply between x and G. Built into the very definition of open set (2.3.3) is that $x \in G$ open means some ball $B_R(x) \subseteq G$, so the open balls already form a base—and the real demand here is the adjective *countable*. The other clue is the hypothesis *separable*, which by definition hands you a countable dense subset D (Problem 22 is where $\mathbb{R}^k$ earns that label); "dense" is the promise that points of D sit arbitrarily close to every x, which is exactly what lets a ball centered in D still catch x.

Two independent infinities threaten countability—which centers, and which radii—and the cure is to discretize both. Restrict the centers to the countable set D and the radii to the positive rationals, so the catalogue $\mathcal{B}$ is indexed by a product of two countable sets and is therefore countable ($\mathbb{Q}$ is countable, with a finite product of countable sets is countable). Before proving that this thinned-out catalogue still works, picture why it should: x may be irrational and the enclosing radius R may be irrational, yet a rational-radius ball pinned to a nearby dense center can be slid into the gap, because density leaves room on both sides.

The fitting argument is a two-sided ε estimate, and each move rests on something named:

(1) from $x \in G$ open, the definition of interior point (2.3.3) produces a radius $R > 0$ with $B_R(x) \subseteq G$—the room inside G that everything else must fit within;

(2) denseness of D supplies a center $x_i \in D$ within $R/4$ of x, so the base ball can be anchored at a legitimate (countable) center near x rather than at x itself;

(3) density of $\mathbb{Q}$ in $\mathbb{R}$ ($\mathbb{Q}$ is dense, 1.7.2) supplies a rational radius r with $d(x, x_i) < r < R/2$: the left inequality forces $x \in B_r(x_i)$, so the ball still catches x, and the right inequality reserves the slack the next step spends;

(4) the triangle inequality bounds any $y \in B_r(x_i)$ by $d(y, x) < r + R/4 < 3R/4 < R$, placing $B_r(x_i) \subseteq B_R(x) \subseteq G$, so the chosen ball lies wholly inside G;

(5) with one member of $\mathcal{B}$ wedged between each x and G, the open set G is the union of those members, which is what "base" demands.

The rigor turns on the margins, and it is worth seeing each one carry its weight. Centering at x_i within $R/4$ and capping r below $R/2$ leaves $d(y, x) < R/4 + R/2 < R$ with room to spare; had the radius been allowed up to R the containment $B_r(x_i) \subseteq B_R(x)$ could fail. The lower bound $d(x, x_i) < r$ is equally load-bearing: drop it and the ball might miss x,

so it would not witness $x \in B_r(x_i) \subseteq G$. Both halves of "base"—containing the point and sitting inside the open set—are checked, and nothing is circular, since the catalogue is built first and only afterward shown to suffice.

The transferable move is the double discretization: when a continuum of choices threatens countability, replace each continuous parameter by a countable dense set of values and let a margin-splitting estimate absorb the error. This is precisely the route from separable to "second countable," and the countable base it yields is the workhorse behind later reductions of open covers and sequential arguments.

Problem 24 *Rudin, Chapter 2, Exercise 24*

Let X be a metric space in which every infinite subset has a limit point. Prove that X is separable. *Hint:* Fix $\delta > 0$, and pick $x_1 \in X$. Having chosen $x_1, \dots, x_j \in X$, choose $x_{j+1} \in X$, if possible, so that $d(x_i, x_{j+1}) \geq \delta$ for $i = 1, \dots, j$. Show that this process must stop after a finite number of steps, and that X can therefore be covered by finitely many neighborhoods of radius δ. Take $\delta = 1/n$ $(n = 1, 2, 3, \dots)$, and consider the centers of the corresponding neighborhoods.

Solution.

Call a finite set δ-separated if any two of its points are at distance $\geq \delta$. The whole argument rests on one claim: under the hypothesis, for each $\delta > 0$ there is no infinite δ-separated set, so X admits a finite cover by balls of radius δ.

Fix $\delta > 0$ and run the greedy process: pick $x_1 \in X$, and having chosen $x_1, \dots, x_j$, pick x_{j+1} with $d(x_i, x_{j+1}) \geq \delta$ for all $i \leq j$ whenever such a point exists. The points so chosen form a δ-separated set, and the process must terminate. If it did not, it would produce an infinite δ-separated set $E = \{x_1, x_2, \dots\}$; by hypothesis E has a limit point $p \in X$, and every ball about p contains infinitely many points of E, in particular two distinct points x_i, x_k in the ball $B_{\delta/2}(p)$. The triangle inequality then gives $d(x_i, x_k) \leq d(x_i, p) + d(p, x_k) < \delta/2 + \delta/2 = \delta$, contradicting δ-separation. So the process halts, say after $x_1, \dots, x_m$.

Once it halts, the balls $B_\delta(x_1), \dots, B_\delta(x_m)$ cover X: if some $x \in X$ lay in none of them, then $d(x_i, x) \geq \delta$ for every $i \leq m$, so x would be a legitimate next choice and the process would not have stopped. Thus for each $\delta > 0$ a *finite* set of centers gives a cover of X by δ-balls.

Apply this with $\delta = 1/n$ for every $n \geq 1$: choose a finite set $S_n = \{x_1^{(n)}, \dots, x_{m_n}^{(n)}\} \subseteq X$ with $X = \bigcup_{j=1}^{m_n} B_{1/n}(x_j^{(n)})$, and put $S = \bigcup_{n=1}^{\infty} S_n$. Being a countable union of finite sets, S is at most countable.

Finally S is dense: given $x \in X$ and $\varepsilon > 0$, pick n with $1/n < \varepsilon$; the radius-$1/n$ balls centered at S_n cover X, so $x \in B_{1/n}(s)$ for some $s \in S_n \subseteq S$, whence $d(x, s) < 1/n < \varepsilon$. Every ball about every point of X therefore meets S, so S is a countable dense subset and X is separable.

■

REFLECTIONS.

The lone hypothesis—"every infinite subset has a limit point"—is the entire engine, and the word that unlocks it is *limit point*. A cluster point is not merely approached; every ball around it swallows infinitely many points of the set (3.2.1; the quotable "every ball about a limit point contains infinitely many points"). That "infinitely many in an arbitrarily small ball" is the lever: it says points cannot stay spread out forever, which is exactly the obstruction the hint's construction is built to expose. Read the goal the same way—"separable" means a countable dense subset, and density is a statement about closures (4.4.1): a set whose closure is all of X, i.e. one that comes within every ε of every point. So the target is a countable set of points fine enough to approximate all of X.

Believe it on a familiar space first. In $\mathbb{R}$ the rationals are dense and countable; the way you build such a set by hand is to lay down a grid at each scale—points $1/n$ apart—and let the meshes shrink. The construction here is that picture made intrinsic: at scale δ you greedily scatter points that stay δ apart, and the hypothesis guarantees you run out of room after finitely many.

The proof is a short chain, each link resting on the one before:

(1) the greedy process at scale δ produces points pairwise at distance $\geq \delta$; if it never stopped it would yield an *infinite* such set, and the hypothesis hands that set a limit point p;

(2) the cluster property (3.2.1, infinitely many in every ball) crowds infinitely many of the points into $B_{\delta/2}(p)$, so two distinct ones sit in that half-radius ball, and the triangle inequality squeezes them to distance $< \delta$—against δ-separation, so the process must halt;

(3) once it halts at $x_1, \ldots, x_m$, those centers' δ-balls cover X, since any uncovered point would itself have been a legal next pick—so finitely many δ-balls cover X (2.3.1, the definition of the ball $B_\delta(x_i) = \{x : d(x_i, x) < \delta\}$);

(4) running this at every scale $\delta = 1/n$ gives finite center-sets S_n, and $S = \bigcup_n S_n$ is a countable union of finite sets, hence at most countable (a countable union of countable sets is countable; 2.1.9);

(5) S is dense: for $x \in X$ and $\varepsilon > 0$, take $1/n < \varepsilon$, and the cover at scale $1/n$ puts some $s \in S_n$ within $1/n < \varepsilon$ of x.

The rigor turns on two points worth checking. The radius $\delta/2$ in step (2) is not cosmetic: the ball must be small enough that *any* two of its points are closer than δ, and only a radius strictly below $\delta/2$—which $\delta/2$ achieves through the strict triangle inequality on distinct

points—forces the contradiction. And the hypothesis is genuinely load-bearing: drop it and the conclusion fails, as an uncountable set with the discrete metric shows—every pair sits at distance 1, no infinite subset has a limit point, and no countable set can be dense because every ball of radius $\frac{1}{2}$ holds a single point. There is no circularity: the limit point in step (1) is supplied by the hypothesis, never assumed of S.

The transferable move is the greedy maximal separated set, the metric-space cousin of a packing argument: to manufacture a countable approximating skeleton, pack in points kept a fixed distance apart, let a compactness-flavored hypothesis cap each packing at finite size, and shrink the spacing to 0 along $1/n$. This is the same machine that drives the next problems in the set—Problem 25 reaches a countable base from compactness by covering at every scale $1/n$, and Problem 26 turns this very separability back into compactness—so the $1/n$-net is worth keeping in hand.

Problem 25 *Rudin, Chapter 2, Exercise 25*

Prove that every compact metric space K has a countable base, and that K is therefore separable. *Hint:* For every positive integer n, there are finitely many neighbourhoods of radius $1/n$ whose union covers K.

Solution.

A *base* for K is a collection $\mathcal{B}$ of open sets such that every open $G \subseteq K$ is a union of members of $\mathcal{B}$; equivalently, for every open G and every $x \in G$ there is some $B \in \mathcal{B}$ with $x \in B \subseteq G$. We produce a countable such collection, and a countable dense set, simultaneously.

Fix $n \in \mathbb{N}$ with $n \geq 1$. The open balls $\{B_{1/n}(x) : x \in K\}$ cover K, since each x lies in its own ball $B_{1/n}(x)$, and each is open (2.3.5; open balls are open). By compactness this cover has a finite subcover (3.4.1): there are centres $x_{n,1}, \ldots, x_{n,m_n} \in K$ with

$$K = \bigcup_{j=1}^{m_n} B_{1/n}(x_{n,j}).$$

Collect the balls over all scales and all the centres they generate,

$$\mathcal{B} = \{\, B_{1/n}(x_{n,j}) : n \geq 1,\ 1 \leq j \leq m_n \,\}, \qquad D = \{\, x_{n,j} : n \geq 1,\ 1 \leq j \leq m_n \,\}.$$

Each is a countable union of finite sets, hence countable (a countable union of countable sets is countable).

The collection $\mathcal{B}$ is a base. Let $G \subseteq K$ be open and $x \in G$; by openness choose $\varepsilon > 0$ with $B_\varepsilon(x) \subseteq G$ (2.3.3). Pick n with $2/n < \varepsilon$. The radius-$1/n$ balls cover K, so $x \in B_{1/n}(x_{n,j})$

for some j. For every y in that ball the triangle inequality gives

$$d(y,x) \le d(y,x_{n,j}) + d(x_{n,j},x) < \tfrac{1}{n} + \tfrac{1}{n} = \tfrac{2}{n} < \varepsilon,$$

so $B_{1/n}(x_{n,j}) \subseteq B_\varepsilon(x) \subseteq G$. Thus $x \in B_{1/n}(x_{n,j}) \subseteq G$ with $B_{1/n}(x_{n,j}) \in \mathcal{B}$, and $\mathcal{B}$ is a countable base.

Finally D is dense: given $x \in K$ and $r > 0$, choose n with $1/n < r$; the radius-$1/n$ balls cover K, so $x \in B_{1/n}(x_{n,j})$ for some j, that is $d(x,x_{n,j}) < 1/n < r$. Every ball about every point of K therefore meets the countable set D, so K is separable.

■

REFLECTIONS.

The words that set the strategy are *compact* and *base*. Compactness is a statement about covers—every open cover admits a finite subcover (3.4.1)—so the way to mine it is to hand it a cover and pocket the finite list it returns. A *base* is a fixed supply of open sets out of which every open set is rebuilt by unions (this is exactly the notion of Problem 23), and "countable base" is the demand that this supply be countable. The hint names the cover to feed in: at each scale $1/n$, the balls of radius $1/n$. The reason that single family of radii is enough is the part to internalise—the radii $1/n$ shrink to 0, so once you can land inside any prescribed ball $B_\varepsilon(x)$, some scale $1/n$ is fine enough to fit a covering ball between x and $B_\varepsilon(x)$.

Believe it before proving it. Picture K tiled, for each n, by finitely many balls of radius $1/n$; as n grows the tiles get finer, and the centres pile up densely. Any open neighbourhood $B_\varepsilon(x)$, no matter how small, is eventually coarser than some tiling: pick n so that a radius-$1/n$ tile, being half the width of $B_\varepsilon(x)$, slides inside it. The countable heap of all these tiles is the base, and their centres are the dense set.

The construction runs through four linked steps, each resting on the one before:

(1) at each scale n the balls $\{B_{1/n}(x) : x \in K\}$ form an open cover—open because balls are open (2.3.5; open balls are open), a cover because every point sits in its own ball—and compactness extracts a finite subcover $B_{1/n}(x_{n,1}), \dots, B_{1/n}(x_{n,m_n})$ (3.4.1); this is the only place compactness is spent;

(2) gathering these finite lists over all n gives the family $\mathcal{B}$ and the centre set D, each a countable union of finite sets, hence countable (a countable union of countable sets is countable);

(3) $\mathcal{B}$ is a base: inside a given $B_\varepsilon(x) \subseteq G$, choose n with $2/n < \varepsilon$, take the covering ball $B_{1/n}(x_{n,j})$ that contains x, and the triangle inequality keeps that whole ball inside $B_\varepsilon(x)$, since two radii $1/n$ sum to less than ε—this is the two-ball estimate, and the

slack $2/n < \varepsilon$ is exactly what makes it close;

(4) D is dense by the same covering balls read at scale $1/n < r$: the centre nearest x at that scale lies within $1/n < r$, so every ball about x meets D.

The argument is honest about where each hypothesis lives. Compactness is load-bearing and used once, in step (1): without it the scale-$1/n$ cover need not reduce to a finite list, the centre set D can be uncountable, and the base falls apart—an uncountable discrete space (every ball of radius $1/n < 1$ is a single point) is the standard space with no countable base and no countable dense set. The choice of radius is also genuine, not cosmetic: a covering ball of radius $1/n$ with $1/n < \varepsilon$ would only place its *centre* near x, but its far side could poke outside $B_\varepsilon(x)$; demanding $2/n < \varepsilon$ buys the margin the triangle inequality needs. Nothing is circular—the base is built from balls produced by compactness, never assumed—and the two conclusions share one engine, so separability costs no extra work once the centres are in hand.

The transferable move is to read "compact" as "hand it a cover, keep the finite subcover," and to realise that running this at every scale $1/n$ converts the qualitative finiteness into a single countable structure that approximates all open sets at once. This is the compact end of a chain the homework builds deliberately: separable yields a countable base by Problem 23, and here compactness yields the countable dense set outright, so compact $\Rightarrow$ separable $\Rightarrow$ countable base, with compactness the strongest link that drives the whole thing.

Problem 26 *Rudin, Chapter 2, Exercise 26*

Let X be a metric space in which every infinite subset has a limit point. Prove that X is compact. *Hint:* By Exercises 23 and 24, X has a countable base. It follows that every open cover of X has a *countable* subcover $\{G_n\}$, $n = 1, 2, 3, \ldots$. If no finite subcollection of $\{G_n\}$ covers X, then the complement F_n of $G_1 \cup \cdots \cup G_n$ is nonempty for each n, but $\bigcap F_n$ is empty. If E is a set which contains a point from each F_n, consider a limit point of E, and obtain a contradiction.

Solution.

Call a metric space *limit-point compact* if every infinite subset has a cluster point; this is the standing hypothesis on X. The goal is genuine compactness: every open cover admits a finite subcover.

First reduce arbitrary covers to countable ones. By Problem 24 the hypothesis makes X separable, and by Problem 23 a separable metric space has a countable base $\{V_k\}_{k\geq 1}$. Let $\{U_\alpha\}$ be any open cover of X. For each index k such that some U_α contains V_k, fix one such cover member and call it $U_{\alpha(k)}$. The chosen sets cover X: given $x \in X$, pick α with $x \in U_\alpha$; since $\{V_k\}$ is a base there is a V_k with $x \in V_k \subseteq U_\alpha$, and the member $U_{\alpha(k)}$ chosen for that k then contains x. Indexing by a subset of the positive integers, $\{U_{\alpha(k)}\}$ is countable, so

every open cover of X has a countable subcover. Enumerate one as $G_1, G_2, G_3, \ldots$.

It now suffices to show this countable cover has a finite subcover. Suppose, toward a contradiction, that it does not. For each $n \geq 1$ set

$$F_n = X \setminus (G_1 \cup \cdots \cup G_n).$$

Each F_n is the complement of a finite union of open sets, hence closed (3.2.8), and each is nonempty, since no finite collection $G_1, \ldots, G_n$ covers X. The sets descend, $F_1 \supseteq F_2 \supseteq \cdots$, and

$$\bigcap_{n=1}^{\infty} F_n = X \setminus \bigcup_{n=1}^{\infty} G_n = \varnothing,$$

because the G_n cover X.

Choose a point $x_n \in F_n$ for each n and put $E = \{x_n : n \geq 1\}$. The set E is infinite. Were it finite, some single point p would equal x_n for infinitely many n; as the F_n are nested, p would then lie in F_n for arbitrarily large n, hence in every F_n, contradicting $\bigcap_n F_n = \varnothing$.

Being infinite, E has a cluster point $x \in X$ by hypothesis. Fix N with $x \in G_N$, which exists because the G_n cover X. Since G_N is open, it is a neighborhood of x, and a ball around a cluster point of E contains infinitely many points of E (3.2.1; List of Facts). In particular G_N contains some x_m with $m \geq N$. But $x_m \in F_m \subseteq F_N = X \setminus (G_1 \cup \cdots \cup G_N)$, so $x_m \notin G_N$—a contradiction.

Therefore the countable cover $\{G_n\}$ has a finite subcover. Since every open cover of X reduces to such a countable cover, every open cover of X has a finite subcover, and X is compact (3.4.1).

■

REFLECTIONS.

The phrase "every infinite subset has a limit point" is the diagnostic clue: it is the conclusion of the Bolzano–Weierstrass theorem, "an infinite subset of a compact space has a cluster point" (4.3.3). The problem hands you that consequence as a hypothesis and asks you to recover compactness, so it is asking whether the implication reverses. In $\mathbb{R}^k$ the answer is yes by Heine–Borel, but here X is an arbitrary metric space, so you cannot lean on boundedness or cells; you must build a finite subcover out of nothing but the limit-point property itself.

The hint is really a map of the whole proof, and reading it teaches the strategy. "X has a countable base" points back through Problem 24 (the same hypothesis yields separability) and Problem 23 (separable metric spaces are second countable); that base is what shrinks any open cover to a countable one. The rest of the hint is the signature of a nested-closed-sets argument—nonempty descending closed sets F_n with empty intersection, one point

chosen from each, and a cluster point used to spring the trap. The cluster point is where the lone hypothesis finally gets spent.

Believe the shape before checking it. The F_n are the parts of X still uncovered after n of the G_i; if no finite stage ever covers everything, each F_n has a leftover point, and stringing those leftovers together produces an infinite set E marching out toward the part of X that the cover is failing to reach. A cluster point of E is a point the cover *does* reach—and that is the collision.

Each step rests on a named support:

(1) the countable base from Problem 23 lets you pick, for each basic set that fits inside some cover member, a single such member; these cover X exactly because a base rebuilds every open set, so an arbitrary cover collapses to a countable subcover $\{G_n\}$;

(2) assuming no finite subcover, the sets $F_n = X \setminus (G_1 \cup \cdots \cup G_n)$ are nonempty (the failure hypothesis), closed as complements of finite unions of open sets (3.2.8), nested, and with empty intersection precisely because the G_n cover X;

(3) a transversal E choosing one $x_n \in F_n$ is infinite, since a repeated value would, by nesting, sit in every F_n and violate $\bigcap_n F_n = \varnothing$;

(4) the standing hypothesis gives E a cluster point x, which lies in some G_N; openness makes G_N a neighborhood, and a neighborhood of a cluster point holds infinitely many points of E (3.2.1; List of Facts), so it catches some x_m with $m \geq N$ even though $x_m \in F_m \subseteq F_N$ was built to avoid G_N.

The rigor turns on two places where a weaker reading would fail. The contradiction needs x_m with an index $m \geq N$, not merely *some* $x_m \in G_N$; that is exactly why "infinitely many points of E" is the right tool rather than "some point"—finitely many points could all carry indices below N and slip through. And the chosen E must be genuinely infinite, or it would have no cluster point to exploit; the nesting of the F_n is what guarantees infinitude, and it is the same nesting that makes $x_m \in F_m \subseteq F_N$ land where the contradiction needs it. The limit-point hypothesis is load-bearing once: drop it and the construction still produces nested nonempty closed F_n with empty intersection, which is precisely the picture of a non-compact space such as $\mathbb{Q}$ or an open interval, so without it nothing forces a finite subcover.

The transferable move is the conversion itself. To turn a sequential or limit-point property into the covering definition of compactness, first buy a countable base so that "cover" means "countable cover," then run the nested-closed-sets contradiction: the tails of an unending cover leave a descending chain of nonempty closed sets, a transversal through them is infinite, and its cluster point must lie in a cover member that the transversal was

constructed to avoid. It is the metric-space twin of the finite-intersection characterization at 4.2.1, read in reverse—there nested compact sets are forced to meet, here a would-be-non-compact space is forced to meet itself and contradict.

Problem 27 *Rudin, Chapter 2, Exercise 27*

Define a point p in a metric space X to be a *condensation point* of a set $E \subset X$ if every neighborhood of p contains uncountably many points of E.
Suppose $E \subset \mathbb{R}^k$, E is uncountable, and let P be the set of all condensation points of E. Prove that P is perfect and that at most countably many points of E are not in P. In other words, show that $P^c \cap E$ is at most countable. *Hint:* Let $\{V_n\}$ be a countable base of $\mathbb{R}^k$, let W be the union of those V_n for which $E \cap V_n$ is at most countable, and show that $P = W^c$.

SOLUTION.

First fix a countable base for $\mathbb{R}^k$: let

$$\{V_n\}_{n\in\mathbb{N}} = \{\, B(q,r) : q \in \mathbb{Q}^k,\ r \in \mathbb{Q},\ r > 0 \,\},$$

the family of open balls with rational centre and rational radius. This is a countable family, being indexed by $\mathbb{Q}^k \times \mathbb{Q}$, a finite product of countable sets. It is a *base*: if U is open and $x \in U$, choose $\rho > 0$ with $B(x,\rho) \subset U$, then a rational point q with $d(x,q) < \rho/2$ and a rational r with $d(x,q) < r < \rho/2$; then $x \in B(q,r)$, and for any $y \in B(q,r)$ one has $d(x,y) \le d(x,q) + d(q,y) < \rho/2 + \rho/2 = \rho$, so $B(q,r) \subset B(x,\rho) \subset U$. Thus every open set is a union of members of $\{V_n\}$.

Let $W = \bigcup\{\, V_n : E \cap V_n \text{ is at most countable} \,\}$, the union of the "small" basic sets, and we show $P = W^c$.

If $p \in W$, then $p \in V_n$ for some n with $E \cap V_n$ at most countable. Since V_n is open it is a neighborhood of p meeting E in at most countably many points, so p is not a condensation point; that is, $p \notin P$. Hence $W \subset P^c$.

If $p \notin P$, some neighborhood of p contains at most countably many points of E; shrinking, there is $\rho > 0$ with $E \cap B(p,\rho)$ at most countable. By the base property there is a basic set V_n with $p \in V_n \subset B(p,\rho)$, and then $E \cap V_n \subset E \cap B(p,\rho)$ is at most countable, so V_n is one of the sets defining W and $p \in V_n \subset W$. Hence $P^c \subset W$. Combining the two inclusions, $P = W^c$.

Now W is a union of open sets, hence open, so $P = W^c$ is closed. Moreover

$$E \cap W = \bigcup\{\, E \cap V_n : E \cap V_n \text{ at most countable} \,\}$$

is a union of at most countably many at-most-countable sets (there are only countably

many basic sets in all), hence at most countable. Since

$$P^c \cap E = W \cap E$$

this says exactly that at most countably many points of E lie outside P.

It remains to see that every point of P is a limit point of P. Fix $p \in P$ and any $\rho > 0$. As p is a condensation point, $E \cap B(p, \rho)$ is uncountable. Removing the at-most-countable set $E \cap W$ cannot exhaust an uncountable set, so

$$E \cap B(p, \rho) \cap W^c = E \cap B(p, \rho) \cap P$$

is uncountable; in particular it contains a point of P other than p. Thus every ball about p contains a point of P distinct from p, so p is a limit point of P. Being closed with no isolated points, P is perfect.

■

Reflections.

The statement defines its own vocabulary, and the first job is to translate "condensation point" into a familiar shape. It is the cluster-point definition (3.2.1) with "infinitely many" upgraded to "uncountably many"—every neighbourhood of p meets E in an uncountable set. The conclusion asks for two things at once, "P is perfect" and "$P^c \cap E$ is at most countable," and the second clause is the louder clue: a countability bound on a topologically defined set is the signature of a *countable base*. The reason a base helps is that it converts the uncountably many neighbourhoods hiding in the definition into a countable checklist, and the hint hands you exactly this device.

Believe the mechanism on a small picture before proving it. Take $E = [0, 1] \cup \{2\}$ in $\mathbb{R}$. Every point of $[0, 1]$ is a condensation point, since any interval around it catches a whole subinterval, which is uncountable; but the isolated point 2 is not, as the neighbourhood $(1.5, 2.5)$ meets E only in $\{2\}$. So $P = [0, 1]$, which is perfect, and the single stray point 2 is the entire countable remainder $P^c \cap E$. The proof simply makes "stray points are rare" precise by collecting them inside basic neighbourhoods that are individually thin.

The argument runs in stages, each resting on a named result:

(1) manufacture the countable base from rational-centre, rational-radius balls; it is countable because $\mathbb{Q}^k \times \mathbb{Q}$ is a finite product of countable sets ($\mathbb{Q}$ is countable, finite products of countable sets are countable), and it is a base because $\mathbb{Q}$ is dense ($\mathbb{Q}$ dense in $\mathbb{R}$, 1.7.2) lets you slip such a ball between any point and any open neighbourhood, the same balls-are-open fact behind open balls are open (2.3.5) and the description of open sets as unions of balls (3.1.5);

(2) with W the union of the basic sets meeting E countably, identify $P = W^c$: a point in W sits in a thin basic neighbourhood, so it fails the condensation condition, while a point failing condensation has *some* thin neighbourhood and the base supplies a thin basic set inside it;

(3) read off that P is closed because W, a union of open sets, is open, so its complement is closed—the open/closed duality open iff complement closed (3.2.8, 3.2.4);

(4) bound $P^c \cap E = W \cap E$: it is the union of the countably many sets $E \cap V_n$ that defined W, each at most countable, so a countable union of countable sets is at most countable (countable union of countable sets, subsets of countable sets);

(5) prove every $p \in P$ is a limit point of P: each ball $B(p, \rho)$ meets E uncountably, and deleting the at-most-countable $E \cap W$ leaves an uncountable—hence nonempty, and in fact infinite—set of points of $E \cap P$ near p, forcing a point of P other than p into the ball.

The one inequality doing all the work in step (5) is that an uncountable set minus a countable set is still uncountable; this is where "E uncountable" is finally spent, and dropping it breaks everything (P would be empty). The other hypotheses are equally load-bearing: without the countable base, step (4) has no reason to produce a countable total—an uncountable union of thin sets need not be thin—which is precisely why the construction lives in $\mathbb{R}^k$, where a countable base exists. The argument is not circular: P is built from W, and W from the fixed base, with the condensation property only read off afterward, never assumed.

The transferable move is the one the hint enforces: when a property is quantified over *all* neighbourhoods, replace "all neighbourhoods" by "all basic neighbourhoods," and if the base is countable an uncountable quantifier collapses to a countable bookkeeping problem. The same trick—small basic sets union to a controllably small set—is what makes second-countable spaces so well-behaved, and it is exactly the strengthening of the cluster-point idea of 3.2.1 from "infinitely many" to "uncountably many."

Problem 28 *Rudin, Chapter 2, Exercise 28*

Prove that every closed set in a separable metric space is the union of a (possibly empty) perfect set and a set which is at most countable. (*Corollary:* Every countable closed set in $\mathbb{R}^k$ has isolated points.) *Hint:* Use Exercise 27.

SOLUTION.

Call a point p a *condensation point* of a set E if every neighborhood of p contains uncountably many points of E; call a closed set *perfect* if every one of its points is a cluster point of it. Let F be a closed subset of a separable metric space X, let P be the set of

condensation points of F, and write $F \setminus P$ for the rest. The claim is that P is perfect (possibly empty), that $F \setminus P$ is at most countable, and that $F = P \cup (F \setminus P)$.

By Problem 23 a separable metric space has a countable base $\{V_n\}_{n\geq 1}$. Let W be the union of those basic sets V_n for which $F \cap V_n$ is at most countable. The first task, exactly as in Problem 27, is to identify the leftover set with W:

$$P^c = W, \qquad \text{equivalently} \qquad P = W^c.$$

If $p \in W$, then $p \in V_n$ for a basic set with $F \cap V_n$ at most countable, so p has a neighborhood meeting F in at most countably many points and is not a condensation point. Conversely, if $p \notin P$, some neighborhood U of p has $F \cap U$ at most countable; since $\{V_n\}$ is a base there is a V_n with $p \in V_n \subseteq U$, whence $F \cap V_n \subseteq F \cap U$ is at most countable and V_n is one of the sets forming W, so $p \in W$.

P is closed, being the complement of the open set W (3.2.8; List of Facts). The leftover $F \setminus P = F \cap W$ is at most countable: it is contained in the union of the sets $F \cap V_n$ used to build W, a countable union of at-most-countable sets, hence at most countable (List of Facts, with "a subset of a countable set is finite or countable", 2.1.9).

The perfect part lies inside F. If $p \notin F$, then because F is closed its complement $X \setminus F$ is open (3.2.8) and is a neighborhood of p disjoint from F, so p meets F in no points at all and is not a condensation point. Thus $P \subseteq F$, and here the closedness of F is spent.

P is perfect. It is closed, so only the cluster-point condition remains. Fix $p \in P$ and any neighborhood U of p. Then $U \cap F$ is uncountable, since p is a condensation point. Removing the at most countable set $F \setminus P$ leaves

$$U \cap P = (U \cap F) \setminus (F \setminus P)$$

still uncountable, hence infinite, hence containing some point of P other than p. Every neighborhood of p therefore contains a point of P distinct from p, so p is a cluster point of P (3.2.1). As $p \in P$ was arbitrary, every point of P is a cluster point of P, and P is perfect. Combining,

$$F = P \cup (F \setminus P),$$

the union of a (possibly empty) perfect set and an at most countable set.

For the corollary, let $F \subseteq \mathbb{R}^k$ be nonempty, countable, and closed, and suppose toward a contradiction that F has no isolated points. Then every point of F is a cluster point of F, and since F is closed it is therefore a nonempty perfect subset of $\mathbb{R}^k$. But a nonempty perfect subset of $\mathbb{R}^k$ is uncountable. To see this, suppose such a set P were the countable list $\{p_1, p_2, \dots\}$. Pick a closed ball $\overline{B}_1$ centered at a point of P with $p_1 \notin \overline{B}_1$ and $\overline{B}_1 \cap P \neq \varnothing$;

this is possible because p_2 (say) is a cluster point of P, so points of P accumulate near it and a small enough ball excludes the single point p_1. Inductively, given a closed ball $\overline{B}_n$ with $\overline{B}_n \cap P \neq \varnothing$, its center is a cluster point of P, so $\overline{B}_n$ contains a point q of P with $q \neq p_{n+1}$; choose a closed ball $\overline{B}_{n+1} \subseteq \overline{B}_n$ centered at q, of radius small enough that $p_{n+1} \notin \overline{B}_{n+1}$, with $\overline{B}_{n+1} \cap P \neq \varnothing$. The sets $K_n = \overline{B}_n \cap P$ are nonempty; each is closed and bounded in $\mathbb{R}^k$, hence compact (4.3.5; List of Facts), and they are nested, $K_1 \supseteq K_2 \supseteq \cdots$. By the nested-compact-sets property their intersection is nonempty (4.2.2; List of Facts), so some $q^* \in \bigcap_n K_n \subseteq P$. But $q^* = p_m$ for some m, while $q^* \in K_m \subseteq \overline{B}_m$ was built to exclude p_m—a contradiction. Hence P is uncountable, contradicting that F is countable. Therefore a nonempty countable closed set in $\mathbb{R}^k$ must have an isolated point.

■

REFLECTIONS.

The clue is the single word *perfect*, paired with the instruction to "use Exercise 27." A perfect set is a closed set all of whose points are cluster points (3.2.1, 3.2.4); to split an arbitrary closed F into a perfect piece and a small remainder you need a canonical rule for deciding which points of F are "robustly surrounded" by F and which are not. Problem 27 built exactly that rule—the condensation points, the points where F stays *uncountably* large in every neighborhood—and showed they form a closed set P whose complement captures all but at most countably many points of E. Reading "perfect set + countable set" as "condensation points + the rest" is the whole move; the present problem is Problem 27 re-run on a closed set, with one extra observation that closedness contributes.

The hidden hinge is the word *separable*. Without a countable base there is no way to assemble the small leftover, because the bound on $F \setminus P$ comes from there being only countably many basic sets to fail on. Problem 23 supplies the base $\{V_n\}$, and the union W of the basic sets that meet F countably is the engine: it is open by construction, so its complement P is closed (3.2.8; List of Facts), and $F \cap W$ is swept up inside a countable union of countable sets (List of Facts).

Believe the shape on a familiar set before checking it. Take $F = \{0\} \cup \{1/n : n \geq 1\}$ in $\mathbb{R}$, a countable closed set. No point is a condensation point, so $P = \varnothing$ and $F = \varnothing \cup F$ is the whole of it as the countable remainder—and the corollary is visible too, since 1 is an isolated point. For an uncountable example, $F = [0, 1]$ has $P = [0, 1]$ and the remainder empty. The split always sends the "thick" part to P and the "thin" part to the countable scrap.

Each step rests on a named support:

(1) the countable base from Problem 23 lets you define W as the union of the basic sets meeting F in at most countably many points; the base property gives $P = W^c$, exactly as in Problem 27, because every neighborhood with countable F-trace contains a basic

such neighborhood;

(2) P is closed as the complement of the open W (3.2.8; List of Facts), and $F \setminus P = F \cap W$ is at most countable, lying in a countable union of at-most-countable sets (List of Facts, subset of a countable set, 2.1.9);

(3) closedness of F forces $P \subseteq F$: a point outside the closed F has the open complement $X \setminus F$ (3.2.8) as a F-free neighborhood, so it cannot be a condensation point— this is the lone place "F closed" is used, and it is what keeps the perfect piece honestly inside F;

(4) P is perfect because near any $p \in P$ the neighborhood meets F uncountably, and subtracting the at most countable $F \setminus P$ still leaves uncountably many points of P, in particular one $\neq p$, so p is a cluster point of P (3.2.1).

The rigor turns on the arithmetic of cardinalities in step (4): *uncountable minus at-most-countable is uncountable*. That is why Problem 27 insists on *condensation* points rather than mere cluster points—an ordinary cluster point only guarantees *infinitely* many nearby points of F (List of Facts, 3.2.1), and "infinite minus countable" can be empty, so the left-over could swallow every nearby point and break the cluster-point check. The uncountable local mass is exactly the surplus that survives the subtraction. Closedness of F is load-bearing in step (3): drop it and a condensation point of F could sit on the boundary outside F, so P would not lie in F and the decomposition $F = P \cup (F \setminus P)$ would fail. Nothing is circular: P is defined from F and the base, never assumed perfect in advance.

The corollary is the payoff and shows why the ambient space matters. A nonempty countable closed set in $\mathbb{R}^k$ cannot be perfect, because a nonempty perfect subset of $\mathbb{R}^k$ is uncountable—the classical fact proved here by trapping a point of P inside a nested chain of compact pieces $K_n = \overline{B}_n \cap P$, each nonempty (the centers are cluster points), each closed and bounded hence compact (4.3.5; List of Facts), with nonempty intersection (4.2.2; List of Facts) engineered to dodge the n-th listed point. So a nonempty countable closed set, having $P = \varnothing$, equals its countable remainder, and a set all of whose points were cluster points would be perfect and so uncountable—hence some point must be isolated.

The transferable idea is the cleavage itself, the Cantor–Bendixson move: to peel a closed set into a "thick" core and a "thin" shell, define the core by a local-largeness condition stable under shrinking neighborhoods (here, uncountably many nearby points), prove the core closed and the shell small with a countable base, and then exploit that the largeness condition survives removing the shell. The same template recovers the perfect kernel of any closed set and, in $\mathbb{R}^k$, separates the countable closed sets (all isolated debris, no kernel) from the uncountable ones.

Problem 29 *Rudin, Chapter 2, Exercise 29*

Prove that every open set in $\mathbb{R}^1$ is the union of an at most countable collection of disjoint segments. *Hint:* Use Exercise 22.

SOLUTION.

Let $E \subseteq \mathbb{R}$ be open. For each $x \in E$ define

$$I_x = \bigcup \{\, J : \ J \text{ is an open interval with } x \in J \subseteq E \,\}.$$

Because E is open, at least one such J exists, so I_x is nonempty and contains x.

First, each I_x is itself an open interval contained in E. It is open as a union of open sets, and it lies in E since every J in the union does. To see it is an interval, take $u < v$ in I_x and w with $u < w < v$; say $u \in J_1$ and $v \in J_2$, where J_1, J_2 are intervals from the union, both containing x. If $w \leq x$ then $u < w \leq x$ forces $w \in J_1$ (an interval containing both u and x); if $w \geq x$ then $x \leq w < v$ forces $w \in J_2$. Either way $w \in I_x$, so I_x is an interval. By construction it is the *largest* open interval with $x \in I_x \subseteq E$ (an infinite endpoint is allowed).

Second, any two of these intervals are equal or disjoint. If $I_x \cap I_y \neq \varnothing$, then $I_x \cup I_y$ is again an open interval contained in E (two intervals that meet have an interval union), and it contains x, so maximality of I_x gives $I_x \cup I_y \subseteq I_x$, i.e. $I_y \subseteq I_x$; symmetrically $I_x \subseteq I_y$, hence $I_x = I_y$. The distinct sets among the I_x are therefore pairwise disjoint, and since each $x \in E$ lies in its own I_x, their union is exactly E.

It remains to count them. List the rationals as a sequence $r_1, r_2, r_3, \ldots$, which is possible because $\mathbb{Q}$ is countable (2.1.7; List of Facts). Each of these intervals is a nonempty open interval, so it contains a rational by density of $\mathbb{Q}$ in $\mathbb{R}$ (1.7.2; List of Facts); assign to each distinct interval I the rational r_I of least index lying in I. Distinct intervals are disjoint, so they receive distinct rationals, and $I \mapsto r_I$ is an injection from the collection of intervals into $\mathbb{Q}$. A set that injects into the countable set $\mathbb{Q}$ is finite or countable (a subset of a countable set is finite or countable). Thus E is the union of an at most countable collection of pairwise disjoint segments.

■

REFLECTIONS.

The phrase "disjoint segments" is the whole map. On the line a connected open set is precisely an open interval, so "write E as disjoint open intervals" is the same request as "split E into its connected components" (4.4.2)—and that single reframing tells you to manufacture, around each point, the biggest interval-shaped chunk of E that holds it. The second clue is "at most countable," the standard countability verdict, which forces the question of *how* to count an a-priori-unknown family; and the hint, "use Exercise 22,"

points at the device, since Exercise 22 is exactly where the density of $\mathbb{Q}$ is harnessed to count—there to find a countable dense $\mathbb{Q}^k$, here to tag each interval with a rational.

Believe the statement on a picture before proving it: $E = (0,1) \cup (2,5) \cup (7,\infty)$ is already three disjoint segments, and any open E you draw breaks into such pieces with gaps (points not in E) between them. The content is that this always happens and that the pieces are never more than countably many.

The proof has three movements, each resting on the one before or on a named result:

(1) around each $x \in E$ form I_x, the union of *all* open intervals trapped between x and E; it is open as a union of opens (any union of open sets is open), it is an interval because all the pieces share the anchor x (splitting on whether the in-between point w sits left or right of x, a clean two-case check, 0.5.7), and by construction it is the maximal such interval—this is the component of x;

(2) maximality forces a dichotomy: if I_x and I_y overlap, their union is a single larger interval inside E containing x, which maximality cannot allow unless $I_x = I_y$, so distinct components are disjoint and, as each point lies in its own, they cover E;

(3) to count, send each distinct interval to the least-indexed rational it contains; density of $\mathbb{Q}$ ($\mathbb{Q}$ is dense in $\mathbb{R}$) guarantees there is one, and disjointness makes the assignment injective, so the family injects into the countable $\mathbb{Q}$ and is at most countable (a subset of a countable set is finite or countable).

Each hypothesis earns its place. Openness of E is what lets step (1) start—without it a point of E need sit in no interval inside E, and I_x could be empty or a single point, not a segment. The maximality built in step (1) is the sole engine of the disjointness in step (2): drop it and overlapping intervals would double-count the same stretch of E, so the family would neither be disjoint nor inject into $\mathbb{Q}$. And the counting in step (3) leans on disjointness, not on the intervals being long; an interval of any positive length still catches a rational, so even a swarm of tiny components stays countable. The argument is not circular: the components are constructed from E outward, and only after they are pinned down are they counted.

The transferable move is the rational-tagging trick, and it is worth abstracting past intervals: any family of pairwise disjoint nonempty open subsets of $\mathbb{R}$ (and of $\mathbb{R}^k$ too, using $\mathbb{Q}^k$ as in Exercise 22) is at most countable, because each open set swallows a point of the countable dense set and disjointness keeps those points distinct. Whenever a problem hands you a disjoint family of fat sets and asks "how many?", reach for a countable dense set and let disjointness do the injecting.

Problem 30 *Rudin, Chapter 2, Exercise 30*

Imitate the proof of Theorem 2.43 to obtain the following result: if $\mathbb{R}^k = \bigcup_{n=1}^{\infty} F_n$, where each F_n is a closed subset of $\mathbb{R}^k$, then at least one F_n has a nonempty interior. *Equivalent statement:* if G_n is a dense open subset of $\mathbb{R}^k$ for $n = 1, 2, 3, \ldots$, then $\bigcap_{n=1}^{\infty} G_n$ is not empty (in fact, it is dense in $\mathbb{R}^k$). *(This is a special case of Baire's theorem.)*

SOLUTION.

The two formulations are complementary, and it is the dense-open form that carries the construction; the closed-set form follows by taking complements. So let $G_1, G_2, \ldots$ be dense open subsets of $\mathbb{R}^k$.

It suffices to show that $\bigcap_{n\geq 1} G_n$ meets every nonempty open set, which is exactly density (4.4.1) and in particular forces nonemptiness. Fix a nonempty open $U \subseteq \mathbb{R}^k$. The plan is to manufacture a nested sequence of closed balls, each one trapped inside both the previous ball and the next G_n, and then catch a common point with the nested-compact-set property.

Since G_1 is dense, $U \cap G_1$ is nonempty, and since G_1 is open it is open; pick a point in it, and because it is open choose a radius small enough that the *closed* ball about that point still lies inside, with radius less than 1:

$$K_1 = \overline{B}(x_1, r_1) \subseteq U \cap G_1, \qquad 0 < r_1 < 1.$$

(One can always shrink: if $B(p, \rho) \subseteq W$ with W open, then $\overline{B}(p, \rho/2) \subseteq B(p, \rho) \subseteq W$.)

Suppose $K_{n-1} = \overline{B}(x_{n-1}, r_{n-1})$ has been chosen. The open ball $B(x_{n-1}, r_{n-1})$ is nonempty and open (2.3.1), and G_n is dense and open, so $B(x_{n-1}, r_{n-1}) \cap G_n$ is nonempty and open. Shrinking a closed ball into it as before, choose

$$K_n = \overline{B}(x_n, r_n) \subseteq B(x_{n-1}, r_{n-1}) \cap G_n, \qquad 0 < r_n < \tfrac{1}{n}.$$

Each K_n is closed and bounded in $\mathbb{R}^k$, hence compact by Heine–Borel (4.3.5; List of Facts), and each is nonempty. Because $K_n \subseteq B(x_{n-1}, r_{n-1}) \subseteq K_{n-1}$, the balls are nested:

$$K_1 \supseteq K_2 \supseteq K_3 \supseteq \cdots.$$

A nested sequence of nonempty compact sets has nonempty intersection (4.2.2; List of Facts), so there is a point $x \in \bigcap_{n\geq 1} K_n$. From $K_1 \subseteq U$ we get $x \in U$, and from $K_n \subseteq G_n$ for every n we get $x \in G_n$ for every n; hence

$$x \in U \cap \bigcap_{n=1}^{\infty} G_n.$$

Thus $\bigcap_{n\geq 1} G_n$ meets every nonempty open U, so it is dense in $\mathbb{R}^k$, and being dense it is nonempty.

Now the closed-set form. Suppose $\mathbb{R}^k = \bigcup_{n\geq 1} F_n$ with each F_n closed, and suppose toward a contradiction that every F_n has empty interior. Put $G_n = \mathbb{R}^k \setminus F_n$. Each G_n is open because F_n is closed (3.2.8; List of Facts), and G_n is dense: a point p with some ball $B(p,\rho) \subseteq F_n$ would be an interior point of F_n, against the assumption, so every ball about every point meets G_n. By the result just proved $\bigcap_{n\geq 1} G_n \neq \varnothing$. But

$$\bigcap_{n=1}^{\infty} G_n = \mathbb{R}^k \setminus \bigcup_{n=1}^{\infty} F_n = \mathbb{R}^k \setminus \mathbb{R}^k = \varnothing,$$

a contradiction. Hence at least one F_n has nonempty interior.

■

Reflections.

The instruction "imitate the proof of Theorem 2.43" is the loudest clue in the problem: that theorem (perfect sets in $\mathbb{R}^k$ are uncountable) runs on a single engine, a shrinking tower of closed balls pinned down at the end by the fact that nested nonempty compact sets meet (4.2.2; List of Facts). So before reading another word you should expect to build $K_1 \supseteq K_2 \supseteq \cdots$ and harvest a point from the intersection. The only question is what each ball must dodge, and the two hypotheses answer it—"open" is what lets you fit a ball inside at all, and "dense" is what guarantees the next G_n reaches into whatever ball you have so far.

The other half of reading the problem is recognising that "F_n closed with empty interior" and "$G_n = F_n^c$ dense and open" are the same statement seen from opposite sides. Complementation turns closed into open (3.2.8; List of Facts), turns the union $\bigcup F_n$ into the intersection $\bigcap G_n$, and—this is the part worth pausing on—turns "F_n has empty interior" into "G_n is dense," because a point fails to be in the closure of F_n^c exactly when some ball around it avoids F_n^c, i.e. sits inside F_n as an interior point (4.4.1). Once you see this dictionary, you prove whichever form is easier and read the other off for free.

The construction itself is a short loop, each step leaning on the one before:

(1) density of G_1 makes $U \cap G_1$ nonempty and openness keeps it open, so it contains a point with a whole ball around it; passing to a closed ball of half the radius gives $K_1 \subseteq U \cap G_1$, the base of the tower;

(2) at stage n the *open* ball $B(x_{n-1}, r_{n-1})$ is a legitimate nonempty open target (2.3.1), so density of G_n lets it reach in and openness lets you again shrink to a closed ball $K_n \subseteq B(x_{n-1}, r_{n-1}) \cap G_n$—and one shrinks into the open ball rather than the closed ball K_{n-1} precisely because the shrinking step needs an *open* set to seat a closed ball

inside, $B(x_{n-1}, r_{n-1}) \cap G_n$ being open while $K_{n-1} \cap G_n$ need not be; nesting $K_n \subseteq K_{n-1}$ then comes for free from $B(x_{n-1}, r_{n-1}) \subseteq \overline{B}(x_{n-1}, r_{n-1}) = K_{n-1}$;

(3) each K_n is closed and bounded, hence compact in $\mathbb{R}^k$ by Heine–Borel (4.3.5; List of Facts)—this is where "$\mathbb{R}^k$" is spent, since a closed ball need not be compact in a general metric space;

(4) nested nonempty compacts meet (4.2.2; List of Facts), producing x; then $K_1 \subseteq U$ and $K_n \subseteq G_n$ place x in $U \cap \bigcap_n G_n$, so the intersection meets the arbitrary open set U and is therefore dense (4.4.1).

It is worth checking that the radii $r_n < 1/n$ are decoration here, not load: nested nonempty compact sets already meet without any control on size, so the shrinking radii buy nothing for this problem. They are kept only to mirror Theorem 2.43, where the radii must vanish so that the intersection is a single point and the branching produces distinct limits. The genuinely load-bearing hypotheses are the three the construction actually consumes—openness (to fit a ball), density (to reach the next set), and the completeness of $\mathbb{R}^k$ packaged as Heine–Borel and the nested-compact property. Drop completeness and the theorem fails: $\mathbb{Q}$ is a countable union of its (closed, empty-interior) singletons, so the closed-set form is false there, and the matching dense open sets $\mathbb{Q} \setminus \{q\}$ have empty intersection.

The transferable move is the one the hint is really teaching: when you must produce a point lying in infinitely many sets at once, do not seek it directly—trap it. Nest a tower of compact sets so that membership in the n-th set is forced by the n-th ball, and let the nested-compact-set property (4.2.2) hand you the common point. This is the skeleton of every Baire-category argument, and it is the same skeleton our notes use to extract an uncountable perfect set.

HOMEWORK 3

Sequences, Series, and Continuity

MATH S4061 | Introduction to Modern Analysis I

Rudin Chapters 3 and 4 — Sequences, Series, and Continuity

Problems in This Set

Rudin Chapter 3 – Numerical Sequences and Series

Problem 3.1 Rudin Ch. 3, Ex. 1

Problem 3.2 Rudin Ch. 3, Ex. 2

Problem 3.3 Rudin Ch. 3, Ex. 3

Problem 3.4 Rudin Ch. 3, Ex. 4

Problem 3.5 Rudin Ch. 3, Ex. 5

Problem 3.6 Rudin Ch. 3, Ex. 6

Problem 3.7 Rudin Ch. 3, Ex. 7

Problem 3.8 Rudin Ch. 3, Ex. 8

Problem 3.9 Rudin Ch. 3, Ex. 9

Problem 3.10 Rudin Ch. 3, Ex. 10

Problem 3.11 Rudin Ch. 3, Ex. 11

Problem 3.12 Rudin Ch. 3, Ex. 12

Problem 3.13 Rudin Ch. 3, Ex. 13

Problem 3.14 Rudin Ch. 3, Ex. 14

Problem 3.15 Rudin Ch. 3, Ex. 15

Problem 3.16 Rudin Ch. 3, Ex. 16

Problem 3.17 Rudin Ch. 3, Ex. 17

Problem 3.18 Rudin Ch. 3, Ex. 18

Problem 3.19 Rudin Ch. 3, Ex. 19

Problem 3.20 Rudin Ch. 3, Ex. 20

Problem 3.21 Rudin Ch. 3, Ex. 21

Problem 3.22 Rudin Ch. 3, Ex. 22

Problem 3.23 Rudin Ch. 3, Ex. 23

Problem 3.24 Rudin Ch. 3, Ex. 24

Problem 3.25 Rudin Ch. 3, Ex. 25

Rudin Chapter 4 – Continuity

Rudin Chapter 3 – Numerical Sequences and Series

Problem 3.1 *Rudin, Chapter 3, Exercise 1*

Prove that convergence of $\{s_n\}$ implies convergence of $\{|s_n|\}$. Is the converse true?

SOLUTION.

The sequence $\{s_n\}$ lives in $\mathbb{R}$ (or $\mathbb{C}$), with $|\cdot|$ the absolute value (modulus); a sequence $x_n \to x$ when for every $\varepsilon > 0$ there is an N with $|x_n - x| < \varepsilon$ for all $n \geq N$ (5.1.1).

Suppose $s_n \to s$. The reverse triangle inequality gives, for every n,

$$\big|\,|s_n| - |s|\,\big| \le |s_n - s|.$$

Fix $\varepsilon > 0$. By convergence there is an N with $|s_n - s| < \varepsilon$ for all $n \ge N$, and then $\big|\,|s_n| - |s|\,\big| < \varepsilon$ for all $n \ge N$. Hence $|s_n| \to |s|$, so $\{|s_n|\}$ converges.

The converse is false. Take $s_n = (-1)^n$ in $\mathbb{R}$. Then $|s_n| = 1$ for every n, so $\{|s_n|\}$ is the constant sequence 1 and converges to 1. But $\{s_n\}$ diverges: it has a subsequence $s_{2n} = 1 \to 1$ and a subsequence $s_{2n+1} = -1 \to -1$, and a convergent sequence has a single limit (at most one limit; 5.1.3), so two subsequences with different limits forbid convergence. Thus $\{|s_n|\}$ can converge while $\{s_n\}$ does not.

■

Reflections.

The statement is an implication between two convergence claims, so the word doing the work is *convergence* itself, and the lecture object it names is the ε–N definition of a limit (5.1.1). Reading the hypothesis and the conclusion side by side shows what the proof must supply: the hypothesis hands you control of $|s_n - s|$, and the conclusion asks for control of $\big|\,|s_n| - |s|\,\big|$, so the entire task is to bound the second quantity by the first. The clue that such a bound exists is the absolute-value bars wrapped around the terms—whenever a problem replaces s_n by $|s_n|$, the reverse triangle inequality $\big|\,|a| - |b|\,\big| \le |a - b|$ is the tool that compares the two, because it says taking absolute values can only shrink distances, never enlarge them.

It helps to see why that inequality should hold before leaning on it. The ordinary triangle inequality $|a| \le |a-b| + |b|$ rearranges to $|a| - |b| \le |a-b|$, and swapping the roles of a and b gives $|b| - |a| \le |a - b|$; together these two say $\big|\,|a| - |b|\,\big| \le |a - b|$. So the map $x \mapsto |x|$ is 1-Lipschitz, and a 1-Lipschitz map cannot pull a convergent sequence apart.

With that in hand the forward direction is a single clean estimate:

(1) the reverse triangle inequality bounds $\big|\,|s_n| - |s|\,\big|$ by $|s_n - s|$ for every n, turning a question about $\{|s_n|\}$ into the question about $\{s_n\}$ the hypothesis already answers;

(2) given $\varepsilon > 0$, convergence $s_n \to s$ supplies an N past which $|s_n - s| < \varepsilon$ (5.1.1), and chaining the two inequalities makes $\big|\,|s_n| - |s|\,\big| < \varepsilon$ for all $n \ge N$, which is exactly $|s_n| \to |s|$.

The same N that works for $\{s_n\}$ works verbatim for $\{|s_n|\}$, since the bound holds termwise; no new tolerance has to be manufactured. Phrased through the tail, every ball about s eventually swallows the s_n (all but finitely many terms), and the Lipschitz bound carries that tail straight to a ball about $|s|$.

The converse question is answered by a counterexample, and the reason one is even worth hunting for is that the forward proof used *both* the size and the sign of s_n only through $|s_n - s|$; the modulus $|s_n|$ throws the sign away, and information thrown away cannot in general be recovered. The cleanest witness is $s_n = (-1)^n$: its moduli are all 1, a constant convergent sequence, while the signs oscillate forever. The rigor of "$\{s_n\}$ diverges" rests on uniqueness of limits (a sequence has at most one limit; 5.1.3)—two subsequences converging to 1 and to -1 cannot both be tails of one convergent sequence. So the implication is strictly one-directional: passing to $|s_n|$ is a genuine loss, not an equivalence.

The transferable move is to recognise that any 1-Lipschitz (or even continuous) operation applied termwise preserves convergence, the absolute value being the first instance; the matching lesson is that such an operation can collapse distinct inputs, so the converse fails exactly when the operation is not injective on the limit's neighbourhood. The borderline case is worth keeping in mind: if $s_n \to 0$ then $|s_n| \to 0$, and here the converse *does* hold, since $|s_n| \to 0$ forces $s_n \to 0$ by the very same termwise bound $|s_n - 0| = \big|\,|s_n| - 0\,\big|$—the sign cannot hide near 0.

Problem 3.2 *Rudin, Chapter 3, Exercise 2*

Calculate $\lim_{n\to\infty}\left(\sqrt{n^2+n}-n\right)$.

SOLUTION.

The limit is $\frac{1}{2}$. For $n \geq 1$ write $a_n = \sqrt{n^2+n} - n$. Rationalising the difference by its conjugate,

$$a_n = (\sqrt{n^2+n} - n)\cdot\frac{\sqrt{n^2+n}+n}{\sqrt{n^2+n}+n} = \frac{(n^2+n)-n^2}{\sqrt{n^2+n}+n} = \frac{n}{\sqrt{n^2+n}+n},$$

the denominator being positive, so no division is illegitimate. Dividing numerator and denominator by $n > 0$,

$$a_n = \frac{1}{\sqrt{1+\frac{1}{n}}+1}.$$

As $n \to \infty$ the standard limit $\frac{1}{n} \to 0$ (6.5.1; Archimedean) gives $1+\frac{1}{n} \to 1$, and the square root is continuous at 1, so $\sqrt{1+\frac{1}{n}} \to 1$; by the algebra of limits (5.2.1; limits respect the algebraic operations) the denominator tends to $1+1 = 2 \neq 0$, and the quotient tends to $\frac{1}{1+1} = \frac{1}{2}$.

To keep the appeal to the square root self-contained, the same value follows from a squeeze. Since $\left(n+\frac{1}{2}\right)^2 = n^2+n+\frac{1}{4} > n^2+n > n^2$, taking positive roots gives $n < \sqrt{n^2+n} < n+\frac{1}{2}$, hence $2n < \sqrt{n^2+n}+n < 2n+\frac{1}{2}$. Inverting and multiplying by $n > 0$,

$$\frac{n}{2n+\frac{1}{2}} < a_n < \frac{n}{2n} = \frac{1}{2}, \qquad \text{that is} \qquad \frac{1}{2+\frac{1}{2n}} < a_n < \frac{1}{2}.$$

The left bound is $\dfrac{1}{2+1/(2n)} \to \dfrac{1}{2+0} = \frac{1}{2}$ by the algebra of limits, and the right bound is the constant $\frac{1}{2}$; both endpoints converge to $\frac{1}{2}$, so by the squeeze theorem (6.3.3) $a_n \to \frac{1}{2}$. Either way,

$$\lim_{n\to\infty}\left(\sqrt{n^2+n}-n\right) = \frac{1}{2}.$$

■

Reflections.

The expression $\sqrt{n^2+n}-n$ is a difference of two quantities that each march off to $+\infty$, so the symbol $\infty-\infty$ it suggests is indeterminate—the limit laws (5.2.1; limits respect the algebraic operations) say nothing about a difference of two divergent sequences, and reading that is the first move. The clue is the lone square root standing against a polynomial: a root minus its near-polynomial is the textbook cue to multiply by the conjugate, because $(\sqrt{A}-B)(\sqrt{A}+B) = A-B^2$ erases the root and turns a fragile difference into a quotient the algebra of limits can actually touch.

It is worth believing the answer before proving it. The root $\sqrt{n^2+n}$ sits between n and $n+\frac{1}{2}$—it is the geometric-mean-like object just above n—so the gap $\sqrt{n^2+n}-n$ hovers just under $\frac{1}{2}$ and ought to settle there; the perturbation $+n$ under the root contributes exactly half a unit in the limit, not a full one, because the root dilutes it.

The computation then runs in a few controlled steps, each resting on a named result:

(1) multiply by the conjugate $\sqrt{n^2+n}+n$ over itself, legitimate since that factor is strictly positive, collapsing the numerator to $(n^2+n)-n^2 = n$ and producing the quotient $a_n = \dfrac{n}{\sqrt{n^2+n}+n}$;

(2) divide top and bottom by n to expose $a_n = \dfrac{1}{\sqrt{1+1/n}+1}$, the form in which the $n\to\infty$ behaviour is visible;

(3) send $\frac{1}{n}\to 0$, the standard limit (6.5.1) that is the Archimedean property in disguise (Archimedean), so the denominator tends to $1+1=2$, a nonzero value, and the quotient to $\frac{1}{2}$ by the algebra of limits (5.2.1);

(4) to avoid leaning on continuity of the root, bracket $n < \sqrt{n^2+n} < n+\frac{1}{2}$ from $(n+\frac{1}{2})^2 > n^2+n > n^2$, which pins a_n between $\dfrac{1}{2+1/(2n)}$ and $\frac{1}{2}$, both tending to $\frac{1}{2}$, so the squeeze theorem (6.3.3) delivers the same value.

The two routes are not redundant: they isolate exactly the one fact that must be supplied from outside the algebra of limits. The quotient form needs $\sqrt{1+1/n}\to 1$, which is continuity of the square root at 1; the squeeze form replaces that with the bare algebraic

inequality $(n+\frac{1}{2})^2 > n^2+n$ and the order facts behind a squeeze (6.3.1), so it assumes nothing about roots beyond their monotonicity. Both need the denominator's limit to be nonzero—the one hypothesis of the quotient law (5.2.1) that is genuinely load-bearing here; were it 0 the algebra of limits would not apply and the squeeze would be the only honest path. Nothing is circular, since $\frac{1}{n} \to 0$ is proved independently of this limit.

The transferable move is the conjugate trick for any $\infty - \infty$ built from a root: rationalise to convert the indeterminate difference into a determinate quotient, then strip the dominant growth (here a factor of n) so the limit laws can finish. The recurring discipline is to refuse the limit laws on a difference of divergent sequences and instead reshape the expression until every sub-limit it contains exists—only then is the algebra of limits (limits respect the algebraic operations) licensed.

Problem 3.3 *Rudin, Chapter 3, Exercise 3*

If $s_1 = \sqrt{2}$, and

$$s_{n+1} = \sqrt{2+\sqrt{s_n}} \qquad (n = 1, 2, 3, \ldots),$$

prove that $\{s_n\}$ converges, and that $s_n < 2$ for $n = 1, 2, 3, \ldots$.

SOLUTION.

Every term is a square root of a positive number, so $s_n > 0$ for all n, and the recursion is the single increasing map $f(x) = \sqrt{2+\sqrt{x}}$ applied repeatedly: $s_{n+1} = f(s_n)$. The plan is to show $\{s_n\}$ is increasing and bounded above by 2, after which monotone convergence forces a limit.

First, $s_n < 2$ for every n, by induction on n. The base case is $s_1 = \sqrt{2} < 2$. Assume $s_n < 2$. Then $\sqrt{s_n} < \sqrt{2} < 2$, so $2 + \sqrt{s_n} < 4$, and taking square roots (which preserves the order on nonnegative reals) gives

$$s_{n+1} = \sqrt{2+\sqrt{s_n}} < \sqrt{4} = 2.$$

By induction $s_n < 2$ for all $n = 1, 2, 3, \ldots$.

Next, $\{s_n\}$ is increasing, i.e. $s_n < s_{n+1}$ for every n, again by induction. For the base case compare the squares: $s_1^2 = (\sqrt{2})^2 = 2$, while $s_2^2 = 2 + \sqrt{\sqrt{2}} > 2$ since $\sqrt{\sqrt{2}} > 0$. Both s_1 and s_2 are positive, so $s_2^2 > s_1^2$ gives $s_2 > s_1$. For the step, assume $s_n < s_{n+1}$. The map f is increasing on $[0, \infty)$: it is the composition $x \mapsto \sqrt{x} \mapsto 2 + \sqrt{x} \mapsto \sqrt{2+\sqrt{x}}$ of order-preserving steps on the nonnegative reals. Applying f to $s_n < s_{n+1}$ therefore preserves the inequality,

$$s_{n+1} = f(s_n) < f(s_{n+1}) = s_{n+2},$$

which is the statement at $n+1$. By induction $s_n < s_{n+1}$ for all n.

Thus $\{s_n\}$ is a monotonically increasing sequence of real numbers bounded above (by

2). By monotone convergence (a bounded monotonic real sequence converges; 6.2.2) the sequence converges to a limit L, with $s_n \leq L \leq 2$ for every n. In particular $\{s_n\}$ converges and $s_n < 2$ for $n = 1, 2, 3, \ldots$.

■

Reflections.

The statement hands you a sequence not by a formula for s_n but by a rule that builds each term from the one before, $s_{n+1} = \sqrt{2 + \sqrt{s_n}}$—a recursively defined sequence. There is no closed form to take a limit of, so the convergence test that needs no formula and no advance knowledge of the limit is the one to reach for: monotone convergence (6.2.2), which certifies a limit from two qualitative facts, monotonicity and boundedness (6.2.1). That the problem also asks for the explicit ceiling $s_n < 2$ is the second clue: it is not a side request but the very bound the convergence proof will run on, so the two halves of the task are really one.

Believe the picture before proving it. Plotting the first terms, $s_1 \approx 1.414$, then 1.786, 1.827, 1.831, …, the values climb but slow down and crowd against a ceiling a little below 1.832, never crossing 2. A sequence that rises yet is penned beneath a fixed wall has nowhere to go but toward a limit—this is exactly the intuition monotone convergence makes rigorous.

The argument splits into three moves, each leaning on induction (0.6.2) or the convergence theorem:

(1) the ceiling $s_n < 2$ comes by induction: $s_1 = \sqrt{2} < 2$, and if $s_n < 2$ then $\sqrt{s_n} < \sqrt{2} < 2$ forces $2 + \sqrt{s_n} < 4$ and hence $s_{n+1} < \sqrt{4} = 2$—the bound reproduces itself because 2 is a fixed point of the estimate;

(2) monotonicity $s_n < s_{n+1}$ comes by a second induction, but the cleanest engine is that the generating map $f(x) = \sqrt{2 + \sqrt{x}}$ is *increasing*, being a composition of order-preserving steps; once $s_2 > s_1$ is checked by hand (squaring turns it into $2 + \sqrt{\sqrt{2}} > 2$), feeding $s_n < s_{n+1}$ through the increasing f yields $s_{n+1} < s_{n+2}$ for free;

(3) with the sequence increasing and bounded above by 2, monotone convergence (bounded monotonic real sequence; 6.2.2) delivers a limit L, and order-preservation in the limit (6.3.1) keeps $L \leq 2$.

The rigor rests on the two inductions being genuinely independent and each self-propagating. The bound in (1) is what makes the increasing step in (2) more than wishful—a sequence can rise without bound, and "increasing" alone never gives convergence—while monotonicity is what rules out the bounded sequence merely oscillating below 2 without settling. Drop either hypothesis and the theorem has nothing to stand on: that is the

content of 6.2.2, and it is why exhibiting both is the whole proof. The increasing-map argument for (2) deserves a second look, since the alternative—comparing s_{n+1} and s_n by squaring at every step—drowns in nested radicals; reading the recursion as "apply one fixed increasing function" converts monotonicity of the sequence into the single, checkable fact that f preserves order, and then induction propagates one inequality the length of the chain.

This problem also quietly answers "what is the limit?" even though it was not asked. Passing to the limit in $s_{n+1} = f(s_n)$, the limit L must satisfy $L = \sqrt{2+\sqrt{L}}$: the left side $s_{n+1} \to L$, while the right side $f(s_n) \to f(L)$ because f is continuous, and a sequence has one limit, so $L = f(L)$; thus L is the fixed point of f, the wall the terms approached. The transferable move is the template for any sequence given by $s_{n+1} = f(s_n)$: guess the trapping bound (often the fixed point of the estimate), prove it by induction, prove monotonicity—most painlessly by showing f is increasing and checking the first comparison—and let monotone convergence finish. The recursion never needs to be unwound; its limit is read off afterward by solving $L = f(L)$.

Problem 3.4 *Rudin, Chapter 3, Exercise 4*

Find the upper and lower limits of the sequence $\{s_n\}$ defined by

$$s_1 = 0; \qquad s_{2m} = \frac{s_{2m-1}}{2}; \qquad s_{2m+1} = \frac{1}{2} + s_{2m}.$$

SOLUTION.

The verdict: $\liminf_{n\to\infty} s_n = \frac{1}{2}$ and $\limsup_{n\to\infty} s_n = 1$.

Compute the first terms to read the pattern. From $s_1 = 0$ the recursion gives $s_2 = 0$, $s_3 = \frac{1}{2}$, $s_4 = \frac{1}{4}$, $s_5 = \frac{3}{4}$, $s_6 = \frac{3}{8}$, $s_7 = \frac{7}{8}$, $s_8 = \frac{7}{16}$, $s_9 = \frac{15}{16}$, which suggests two interleaved patterns. We prove, for every $m \geq 1$,

$$s_{2m+1} = 1 - \frac{1}{2^m}, \qquad s_{2m} = \frac{1}{2} - \frac{1}{2^m}.$$

Both formulas follow by induction on m. For $m = 1$: $s_2 = s_1/2 = 0 = \frac{1}{2} - \frac{1}{2}$, and $s_3 = \frac{1}{2} + s_2 = \frac{1}{2} = 1 - \frac{1}{2}$, so the base case holds. Assume both hold at some $m \geq 1$. The even recursion at $m+1$ reads $s_{2(m+1)} = s_{2m+1}/2$, and using $s_{2m+1} = 1 - 2^{-m}$,

$$s_{2m+2} = \frac{1 - 2^{-m}}{2} = \frac{1}{2} - \frac{1}{2^{m+1}}.$$

The odd recursion then gives

$$s_{2(m+1)+1} = \frac{1}{2} + s_{2m+2} = \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{2^{m+1}}\right) = 1 - \frac{1}{2^{m+1}},$$

which is the pair of formulas at $m+1$. By induction they hold for all $m \geq 1$.

These closed forms split $\{s_n\}$ into its even- and odd-indexed parts. Since $2^{-m} \to 0$ (Archimedean property; the standard limit 6.5.1), the even subsequence $\{s_{2m}\}_{m\geq 1}$ increases to $\frac{1}{2}$ and the odd subsequence $\{s_{2m+1}\}_{m\geq 1}$ increases to 1; each is bounded and monotone, so each converges (monotone convergence; 6.2.2). Thus $\frac{1}{2}$ and 1 are subsequential limits of $\{s_n\}$ (5.3.1).

No other value is a subsequential limit. Suppose $\{s_{n_k}\}$ converges to some ℓ. Among its indices n_k, either infinitely many are even or infinitely many are odd, so $\{s_{n_k}\}$ has a further subsequence lying entirely inside the even family $\{s_{2m}\}$ or entirely inside the odd family $\{s_{2m+1}\}$. A subsequence of a convergent sequence has the same limit—if a sequence converges to a point, every neighborhood of that point contains all but finitely many terms, hence all but finitely many terms of any subsequence (tail trapping; 5.1.3). So that further subsequence converges to $\frac{1}{2}$ or to 1; but it is also a subsequence of $\{s_{n_k}\}$, which converges to ℓ, so by the same principle its limit is ℓ. Hence $\ell \in \{\frac{1}{2}, 1\}$.

The set of subsequential limits is therefore exactly $E = \{\frac{1}{2}, 1\}$. The sequence is bounded (every term lies in $[0,1]$), so the upper and lower limits are the largest and smallest members of E (6.4.5):

$$\liminf_{n\to\infty} s_n = \min E = \frac{1}{2}, \qquad \limsup_{n\to\infty} s_n = \max E = 1.$$

■

Reflections.

The phrase *upper and lower limits* names the lecture object outright: the lim sup and lim inf, defined through the tail extrema but most usable through the characterization that, for a bounded sequence, they are the largest and smallest of the *subsequential* limits (6.4.5). So the task is not to evaluate a single limit—there is none, the terms oscillate forever—but to find every value the sequence clusters at and then read off the top and bottom of that set. The recursion is the second clue: it defines s_{2m} and s_{2m+1} by separate rules, so the sequence is two interleaved threads, and "even index versus odd index" is the parity split that turns one oscillating sequence into two well-behaved subsequences (5.3.1).

Believe it first by listing terms. The even-indexed values $0, \frac{1}{4}, \frac{3}{8}, \frac{7}{16}, \ldots$ climb toward $\frac{1}{2}$, while the odd-indexed values $\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \ldots$ climb toward 1; the sequence is forever ratcheting up to 1 on odd steps and getting halved back down toward $\frac{1}{2}$ on even steps. Two accumulation heights, $\frac{1}{2}$ and 1, are visible before any formula is written.

The proof then proceeds in four movements, each resting on a named result:

(1) guess and verify closed forms $s_{2m+1} = 1 - 2^{-m}$ and $s_{2m} = \frac{1}{2} - 2^{-m}$ by induction on

m (6.2.1 for the monotone reading; the induction itself is 0.6.2), the even and odd recursions feeding each other so the two formulas advance together in one step;

(2) read off that each subsequence is bounded and increasing, hence convergent by monotone convergence (a bounded monotone sequence converges; 6.2.2), with limits $\frac{1}{2}$ and 1 since $2^{-m} \to 0$ (6.5.1);

(3) rule out any third cluster value: an arbitrary convergent subsequence must reuse infinitely many even *or* infinitely many odd indices, so it has a sub-subsequence trapped in one thread, and a subsequence inherits its parent's limit—every neighborhood of the limit holds all but finitely many terms, so it holds all but finitely many of any subsequence (tail trapping; 5.1.3)—pinning the limit to $\frac{1}{2}$ or 1;

(4) with the set of subsequential limits now exactly $E = \{\frac{1}{2}, 1\}$ and the sequence bounded, invoke the characterization $\liminf = \min E$, $\limsup = \max E$ (6.4.5).

Step (3) is the one that is easy to skip and the one that makes the answer honest. Exhibiting two subsequential limits proves only $\frac{1}{2}, 1 \in E$, hence $\liminf \leq \frac{1}{2}$ and $\limsup \geq 1$; it does *not* show these are the extremes, because a stray subsequence drawing terms from both threads could in principle settle somewhere else. The parity dichotomy closes that door: every subsequence is eventually dominated by one thread, so E can hold nothing beyond $\frac{1}{2}$ and 1. Without this upper bound on E one has computed two cluster points, not the upper and lower limits. The boundedness of $\{s_n\}$ is also load-bearing—it is the hypothesis under which $\limsup$ and $\liminf$ equal $\max E$ and $\min E$ rather than $\pm\infty$ (6.4.7), and here it is free since every term lands in $[0, 1]$.

The transferable move is to treat a sequence defined by alternating rules as a braid of subsequences indexed by the residue class: split on the index modulo the period, settle each thread on its own (monotone convergence does it here), and then assemble the upper and lower limits as the max and min of the threads' limits—valid precisely because every subsequence is eventually swallowed by one thread. The extremes are attained as actual subsequential limits (6.4.6), which is why the verdict is a clean $\frac{1}{2}$ and 1 rather than an unapproached bound.

Problem 3.5 *Rudin, Chapter 3, Exercise 5*

For any two real sequences $\{a_n\}$, $\{b_n\}$, prove that

$$\limsup_{n\to\infty}(a_n + b_n) \leq \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n,$$

provided the sum on the right is not of the form $\infty - \infty$.

SOLUTION.

Read lim sup through its definition as a limit of tail suprema (6.4.5): for a real sequence $\{c_n\}$ write

$$C_n = \sup_{k \geq n} c_k \in (-\infty, +\infty],$$

the supremum of a nonempty set of reals, so C_n is never $-\infty$. The tails are nested, so C_n is non-increasing in n, and $\limsup_n c_n = \lim_n C_n = \inf_n C_n$ in the extended reals. Apply this to the three sequences in play, setting

$$A_n = \sup_{k \geq n} a_k, \qquad B_n = \sup_{k \geq n} b_k, \qquad S_n = \sup_{k \geq n} (a_k + b_k),$$

each lying in $(-\infty, +\infty]$, with $\alpha := \limsup a_n = \lim_n A_n$, $\beta := \limsup b_n = \lim_n B_n$, and $\sigma := \limsup(a_n + b_n) = \lim_n S_n$.

The inequality holds at every finite stage. Fix n. For each index $k \geq n$ one has $a_k \leq A_n$ and $b_k \leq B_n$, so $a_k + b_k \leq A_n + B_n$. Because $A_n, B_n > -\infty$, the right-hand side is a well-defined element of $(-\infty, +\infty]$ and is an upper bound for $\{a_k + b_k : k \geq n\}$; taking the supremum over $k \geq n$ gives

$$S_n \leq A_n + B_n \qquad \text{for every } n.$$

Now pass to the limit. As $n \to \infty$, $A_n \downarrow \alpha$ and $B_n \downarrow \beta$ with $\alpha, \beta \in [-\infty, +\infty]$. The hypothesis excludes $\alpha + \beta = \infty - \infty$, so the sum $\alpha + \beta$ is defined in the extended reals and $A_n + B_n \to \alpha + \beta$ (the limit of a sum is the sum of the limits whenever the latter is not of the form $\infty - \infty$). Letting $n \to \infty$ in $S_n \leq A_n + B_n$ and using that the order $\leq$ is preserved under limits (6.3.1) yields

$$\sigma = \lim_n S_n \leq \lim_n (A_n + B_n) = \alpha + \beta,$$

that is, $\limsup(a_n + b_n) \leq \limsup a_n + \limsup b_n$.

■

Reflections.

The single word that fixes the strategy is *limsup*, and reading the statement well means deciding *which* of its faces to use. Our notes carry two: the upper limit is the largest subsequential limit (6.4.6), and it is the limit of the tail suprema $\sup_{k \geq n} c_k$ (6.4.5). A subsequential-limit attack would have to extract a subsequence of $a_n + b_n$ converging to $\limsup(a_n + b_n)$ and then control a and b *along the same indices*—awkward, because the subsequence chasing $\limsup(a_n + b_n)$ need not chase either $\limsup a_n$ or $\limsup b_n$. The tail-supremum face has no such coupling: a supremum over a block of indices already dominates every term in the block at once, which is exactly the kind of crude, simultaneous bound an inequality between three separate lim sups wants.

Believe the inequality, and its slack, on a small case first. Take $a_n = (-1)^n$ and $b_n = (-1)^{n+1}$, so each oscillates between ± 1 with $\limsup a_n = \limsup b_n = 1$; the right side is 2. But $a_n + b_n = 0$ for every n, so $\limsup(a_n + b_n) = 0$. The gap $0 < 2$ is not a defect—it is the whole phenomenon: each sequence peaks at $+1$, but on *opposite* parities, so their peaks never coincide and the sum never sees a 2. The right-hand side optimizes a and b independently; the left-hand side is forced to pick a common index. That picture is why one expects $\leq$ and not $=$.

The proof is short once the tail-supremum reading is in hand, and each move rests on the one before:

(1) rewrite all three lim sups as limits of tail suprema A_n, B_n, S_n (6.4.5), and note each tail supremum, being a sup of reals, lies in $(-\infty, +\infty]$ and decreases as the tail shrinks—this nestedness is what makes $\lim_n$ exist as $\inf_n$;

(2) at a *fixed* n, every $k \geq n$ satisfies $a_k \leq A_n$ and $b_k \leq B_n$, hence $a_k + b_k \leq A_n + B_n$; this is the only genuine idea, the subadditivity of a supremum, and it is legitimate because $A_n, B_n > -\infty$ keeps the bound $A_n + B_n$ from ever being an undefined $\infty - \infty$;

(3) taking the supremum over $k \geq n$ promotes the term bound to $S_n \leq A_n + B_n$, true for every n;

(4) let $n \to \infty$: order survives the limit (6.3.1), each tail supremum being non-increasing so that its limit is its infimum in the extended reals—which is how the upper limit is defined and why it always exists (6.4.5, 6.4.7), no boundedness assumed— and $A_n + B_n \to \alpha + \beta$ because the limit of a sum is the sum of the limits in the extended reals—valid precisely when the sum is not $\infty - \infty$, which is where the lone hypothesis is spent.

The rigor turns entirely on extended-real bookkeeping, and there are two places to be careful. First, a tail supremum is never $-\infty$ (it is the sup of a nonempty set of real numbers), so it can only be a real or $+\infty$; this is what guarantees the finite-stage sum $A_n + B_n$ is always defined, even though $\alpha + \beta$ might not be. Second, the hypothesis "not $\infty - \infty$" is load-bearing exactly at the passage to the limit: drop it and one could have $\limsup a_n = +\infty$, $\limsup b_n = -\infty$, where the right side is meaningless while the left may be anything—e.g. $a_n = n$, $b_n = -n$ gives a right side $\infty - \infty$ and a left side 0, so the statement could not even be parsed, let alone proved. When the right side *is* defined and equals $+\infty$ the claim is vacuous, and the only substantive case is finite α, β, which the same chain covers without separate treatment.

The move to carry away is that an inequality between lim sups of *different* sequences is almost always cleanest through the tail-supremum definition, never the subsequential-limit one: a supremum bounds an entire block of indices simultaneously, so a per-term

inequality survives both the block supremum and the limit untouched, and one never has to align subsequences. The same template, run with inf in place of sup and the inequalities reversed, gives the companion $\liminf(a_n+b_n) \geq \liminf a_n+\liminf b_n$; and the strict-versus-equal question raised by the ± 1 example—when do the peaks line up?—is the upper-limit bookkeeping that drives the oscillating sequence of Problem 3.4.

Problem 3.6 *Rudin, Chapter 3, Exercise 6*

Investigate the behavior (convergence or divergence) of $\sum a_n$ if

1. $a_n = \sqrt{n+1} - \sqrt{n}$;
2. $a_n = \dfrac{\sqrt{n+1} - \sqrt{n}}{n}$;
3. $a_n = (\sqrt[n]{n} - 1)^n$;
4. $a_n = \dfrac{1}{1+z^n}$, for complex values of z.

Solution.

Throughout, the rationalisation $\sqrt{n+1} - \sqrt{n} = \dfrac{1}{\sqrt{n+1}+\sqrt{n}}$, obtained by multiplying by the conjugate, converts each difference of roots into a positive fraction whose size is transparent.

For (a), the series *diverges*. Rationalising,

$$a_n = \frac{1}{\sqrt{n+1}+\sqrt{n}} \geq \frac{1}{2\sqrt{n+1}} > 0,$$

so by comparison (6.6.8) it suffices that $\sum 1/\sqrt{n+1}$ diverge, which it does: it is the p-series with $p = \frac{1}{2} \leq 1$ (6.6.5). One can also see it outright, since the partial sums telescope: $\sum_{n=0}^{N}(\sqrt{n+1} - \sqrt{n}) = \sqrt{N+1} \to \infty$.

For (b), the series *converges* (the terms begin at $n = 1$, where a_n is defined). Rationalising the numerator,

$$a_n = \frac{\sqrt{n+1} - \sqrt{n}}{n} = \frac{1}{n(\sqrt{n+1}+\sqrt{n})} \leq \frac{1}{n \cdot \sqrt{n}} = \frac{1}{n^{3/2}},$$

since $\sqrt{n+1} + \sqrt{n} > \sqrt{n}$. The bounding series $\sum 1/n^{3/2}$ is the p-series with $p = \frac{3}{2} > 1$, hence convergent (6.6.5); as $0 \leq a_n \leq n^{-3/2}$, comparison (6.6.8) gives convergence.

For (c), the series *converges*, by the root test. With $a_n = (\sqrt[n]{n} - 1)^n \geq 0$,

$$\sqrt[n]{a_n} = \sqrt[n]{n} - 1.$$

The standard limit $\sqrt[n]{n} \to 1$ (6.5.1) gives $\sqrt[n]{a_n} \to 0$, so $\limsup_n \sqrt[n]{a_n} = 0 < 1$, and the

root test (6.6.9) yields convergence.

For (d) the answer depends on $|z|$: the series *converges* when $|z| > 1$ and *diverges* when $|z| \leq 1$ (excluding any z with $z^n = -1$ for some n, where a term is undefined). The dividing line is whether the terms tend to 0, the necessary condition for convergence (6.6.3).

- If $|z| < 1$, then $z^n \to 0$, so $a_n = \dfrac{1}{1+z^n} \to 1 \neq 0$; the terms do not vanish, so the series diverges.
- If $|z| = 1$, then $|z^n| = 1$, so $|1 + z^n| \leq 1 + |z^n| = 2$ and hence $|a_n| = \dfrac{1}{|1+z^n|} \geq \frac{1}{2}$ whenever $1 + z^n \neq 0$; again the terms fail to tend to 0, and the series diverges.
- If $|z| > 1$, then $|z^n| = |z|^n \to \infty$, and the reverse triangle inequality gives $|1 + z^n| \geq |z|^n - 1 > 0$ for all large n, so

$$|a_n| = \frac{1}{|1+z^n|} \leq \frac{1}{|z|^n - 1}.$$

  Taking nth roots, $\left(|z|^n - 1\right)^{1/n} = |z|\,(1 - |z|^{-n})^{1/n} \to |z|$, so $\limsup_n \sqrt[n]{|a_n|} \leq 1/|z| < 1$; the root test (6.6.9) makes $\sum |a_n|$ converge, and absolute convergence forces convergence.

Thus $\sum a_n$ converges precisely for $|z| > 1$.

■

REFLECTIONS.

Every part asks the same question—does $\sum a_n$ converge?—and the word *investigate* is the instruction to first reach a verdict and only then justify it, so each part opens by deciding what to prove. The toolkit is fixed and small: a candidate series is weighed against one we already understand, either by the comparison test (6.6.8) against a p-series (6.6.5) or a geometric series (6.6.7), by the root test (6.6.9) when the nth term is built from an nth power, or—when nothing converges—by the cheapest disqualifier of all, that the terms must tend to 0 (6.6.3). Reading which clue each a_n carries is the whole skill the problem trains.

Parts (a) and (b) both wear a difference of square roots, and the reflex a radical difference should trigger is to multiply by the conjugate: $\sqrt{n+1} - \sqrt{n} = 1/(\sqrt{n+1} + \sqrt{n})$ turns a small difference of large numbers into a transparent positive fraction. Once rationalised, the size is plain. In (a) the term is $\asymp 1/(2\sqrt{n})$, a p-series with $p = \frac{1}{2}$, just on the divergent side of the boundary $p = 1$; in (b) the extra factor of n in the denominator pushes the exponent to $p = \frac{3}{2}$, just on the convergent side. The two parts are deliberately paired to sit on opposite shores of the same divide, and the lesson is that $\sum n^{-p}$ converges exactly when $p > 1$ (6.6.5)—the single fact both verdicts rest on.

It is worth believing (a) without any test at all, because its structure is a gift: the partial sum $\sum_{n=0}^{N}(\sqrt{n+1}-\sqrt{n})$ telescopes to $\sqrt{N+1}$, which marches off to infinity. A telescoping sum exposes the partial sums directly (6.6.1), and seeing $\sqrt{N+1}\to\infty$ is more convincing than any comparison.

Part (c) advertises its method in its shape: an nth term that is itself an nth power, $(\sqrt[n]{n}-1)^n$, is precisely the form the root test was built for (6.6.9), since taking $\sqrt[n]{\cdot}$ peels the power away cleanly. What remains, $\sqrt[n]{n}-1$, is governed by the standard limit $\sqrt[n]{n}\to 1$ (6.5.1); the base creeps down to 1, so the whole power collapses geometrically, and $\limsup\sqrt[n]{a_n}=0<1$ closes it. Trying the ratio test here would mean wrestling with $(\sqrt[n+1]{n+1}-1)^{n+1}/(\sqrt[n]{n}-1)^n$, a far uglier object—the nth power is the signpost pointing at the root test, not the ratio test.

Part (d) is the one that rewards reading the statement slowly: "for complex values of z" is an invitation to split on $|z|$, because the entire behaviour of z^n is dictated by its modulus—it shrinks to 0, stays on the unit circle, or blows up, according to whether $|z|$ is below, on, or above 1. The argument tracks those three regimes:

(1) for $|z|<1$ the powers $z^n\to 0$ (5.1.1), so $a_n\to 1\neq 0$ and the term test (6.6.3) kills convergence before any comparison is needed;

(2) for $|z|=1$ the powers ride the unit circle, so $|1+z^n|\le 2$ forces $|a_n|\ge\frac{1}{2}$, and the terms again fail to vanish—the term test does the work a second time, now via a lower bound rather than a limit;

(3) for $|z|>1$ the powers escape to infinity, the reverse triangle inequality gives $|1+z^n|\ge |z|^n-1$, and $|a_n|\le 1/(|z|^n-1)$ behaves like the convergent geometric tail $|z|^{-n}$; the root test (6.6.9), using $(|z|^n-1)^{1/n}\to|z|$, certifies $\limsup\sqrt[n]{|a_n|}\le 1/|z|<1$ and hence absolute convergence.

Two points keep (d) honest. The cases $|z|<1$ and $|z|=1$ are genuinely different arguments, not one with a shared limit: on the circle z^n need not converge at all (take $z=-1$, where z^n alternates and some terms are even undefined), so one cannot say $a_n\to$ anything—only that $|a_n|$ stays bounded below, which is all the term test needs. And the comparison in case (3) must run on the *moduli*; $\sum a_n$ is a complex series, so convergence is established through $\sum|a_n|$, the absolute-convergence detour that lets a real positive test settle a complex series. Across all four parts no step is circular, since each verdict is reached by comparison to a series whose fate is known independently.

The transferable instinct is to classify before you compute: a difference of roots wants its conjugate and a p-series comparison; an nth power of an nth term wants the root test; a parameter z entering only through powers wants a split on $|z|$; and whenever a comparison or root estimate would be laborious, first check the free necessary condition that the terms

tend to 0 (6.6.3)—it disqualifies a divergent series in one line, as it does twice in part (d).

Problem 3.7 *Rudin, Chapter 3, Exercise 7*

Prove that the convergence of $\sum a_n$ implies the convergence of

$$\sum \frac{\sqrt{a_n}}{n},$$

if $a_n \geq 0$.

SOLUTION.

The terms run over $n \geq 1$, so that $1/n$ is defined; every term $\sqrt{a_n}/n$ is ≥ 0, since $a_n \geq 0$.

The one inequality the proof needs is the bound between geometric and arithmetic means of two nonnegative reals: for $x, y \geq 0$,

$$\sqrt{xy} \leq \tfrac{1}{2}(x+y),$$

which follows from $\left(\sqrt{x}-\sqrt{y}\right)^2 \geq 0$, i.e. $x - 2\sqrt{xy} + y \geq 0$. Apply it with $x = a_n$ and $y = 1/n^2$, both ≥ 0:

$$\frac{\sqrt{a_n}}{n} = \sqrt{a_n \cdot \frac{1}{n^2}} \leq \frac{1}{2}\Big(a_n + \frac{1}{n^2}\Big).$$

The right-hand side is the term of a convergent series. The series $\sum a_n$ converges by hypothesis, and the p-series $\sum 1/n^2$ converges, being a p-series with $p = 2 > 1$ (6.6.5). Convergent series add term by term (5.2.1; algebra of limits applied to the partial sums), so

$$\sum_{n\geq 1} \frac{1}{2}\Big(a_n + \frac{1}{n^2}\Big) = \frac{1}{2}\sum_{n\geq 1} a_n + \frac{1}{2}\sum_{n\geq 1}\frac{1}{n^2}$$

converges.

Now the comparison test finishes it. The estimate $0 \leq \sqrt{a_n}/n \leq \frac{1}{2}(a_n + 1/n^2)$ holds for every $n \geq 1$ with a convergent majorant, so $\sum \sqrt{a_n}/n$ converges by comparison (6.6.8).

■

REFLECTIONS.

Two features of the statement set the whole approach. First, the hypothesis $a_n \geq 0$ together with $1/n > 0$ makes every term $\sqrt{a_n}/n$ nonnegative, and a nonnegative series is the friendly case: its partial sums increase, so it converges exactly when they stay bounded (6.6.6), and the entire arsenal of comparison-type tests is available (6.6.8). Second, the term $\sqrt{a_n}/n$ is a *product* of two pieces—$\sqrt{a_n}$, which the hypothesis controls only through $\sum a_n$, and $1/n$, which we know converges when squared. The reflex when an unknown quantity is multiplied against a known summable one is to split the product so each piece feeds a series we already understand, and the clean tool for splitting a product of nonnegatives

into a *sum* is the arithmetic–geometric mean inequality.

Believe it first by trusting the shape of the estimate. The expression $\sqrt{a_n}/n$ is the geometric mean $\sqrt{a_n \cdot 1/n^2}$ of a_n and $1/n^2$; the AM–GM bound trades that geometric mean for the average $\frac{1}{2}(a_n + 1/n^2)$, which is exactly a blend of the two series whose convergence we can certify. The square-root, which is what makes the term awkward to bound directly, is precisely what AM–GM is built to remove.

The argument runs in three steps, each resting on a named result:

(1) the pointwise bound $\sqrt{a_n}/n \le \frac{1}{2}(a_n + 1/n^2)$ comes from $(\sqrt{a_n} - 1/n)^2 \ge 0$, valid because $a_n \ge 0$ lets $\sqrt{a_n}$ exist—this is where the sign hypothesis is spent;

(2) the majorant converges, since $\sum a_n$ converges by hypothesis and $\sum 1/n^2$ is a p-series with $p = 2 > 1$ (6.6.5), and convergent series combine linearly term by term (5.2.1; the algebra of limits on the partial sums);

(3) the comparison test (6.6.8), whose hypothesis is exactly $0 \le b_n \le c_n$ with $\sum c_n$ convergent, transfers convergence from the majorant to $\sum \sqrt{a_n}/n$.

Each hypothesis of the comparison test is genuinely load-bearing, and dropping the sign assumption shows why. The test needs the terms to be nonnegative and dominated by a convergent series; the lower bound $0 \le \sqrt{a_n}/n$ is what rules out the cancellation that would otherwise let a bounded-above series still diverge. The hypothesis $a_n \ge 0$ is used twice over—once so that $\sqrt{a_n}$ is even defined, and once so that the comparison applies—and it is not cosmetic: were the a_n allowed to be negative, $\sqrt{a_n}$ would be meaningless and the statement would not parse. The factor $1/n$ is doing real work too; pairing a_n with $1/n^2$ rather than some larger weight is what keeps the auxiliary p-series convergent, and a heavier weight such as $1/\sqrt{n}$ (whose square $1/n$ gives the divergent harmonic series) would break step (2).

The transferable move is the splitting itself: to tame a product $\sqrt{a_n} \cdot w_n$ of a crudely-controlled factor against a known weight, use AM–GM to dominate it by $\frac{1}{2}(a_n + w_n^2)$ and sum the two halves separately. It is the discrete shadow of the Cauchy–Schwarz reflex, and it works precisely because $\sum w_n^2$—here the p-series $\sum 1/n^2$ (6.6.5)—converges, so the price of removing the square root is a series we have already paid for.

Problem 3.8 *Rudin, Chapter 3, Exercise 8*

If $\sum a_n$ converges, and if $\{b_n\}$ is monotonic and bounded, prove that $\sum a_n b_n$ converges.

Solution.

Write $A_n = \sum_{k=0}^{n} a_k$ for the partial sums of $\sum a_n$, with the convention $A_{-1} = 0$. Since $\sum a_n$ converges, the sequence $\{A_n\}$ converges, hence is bounded (a convergent sequence is bounded): there is M with $|A_n| \le M$ for all n.

The sequence $\{b_n\}$ is monotonic and bounded, so it converges, say $b_n \to b$ (a bounded monotonic sequence converges; 6.2.2). Setting $c_n = b_n - b$, the sequence $\{c_n\}$ is monotonic and $c_n \to 0$. Splitting the terms,

$$a_n b_n = a_n c_n + b\, a_n,$$

the series $\sum b\, a_n = b \sum a_n$ converges (a convergent series scaled by a constant), so it suffices to prove that $\sum a_n c_n$ converges. We verify the Cauchy criterion for it (6.6.2).

The tool is summation by parts. For $0 \le p \le q$, using $a_k = A_k - A_{k-1}$ and reindexing,

$$\sum_{k=p}^{q} a_k c_k = \sum_{k=p}^{q} (A_k - A_{k-1}) c_k = A_q c_q - A_{p-1} c_p + \sum_{k=p}^{q-1} A_k\,(c_k - c_{k+1}).$$

Bound each piece by $|A_k| \le M$. The boundary terms give $|A_q c_q| \le M|c_q|$ and $|A_{p-1} c_p| \le M|c_p|$. For the sum, $\{c_n\}$ is monotonic, so all the differences $c_k - c_{k+1}$ carry the same sign; hence

$$\sum_{k=p}^{q-1} |c_k - c_{k+1}| = \Big|\sum_{k=p}^{q-1} (c_k - c_{k+1})\Big| = |c_p - c_q| \le |c_p| + |c_q|,$$

the middle equality being a telescoping sum. Combining,

$$\begin{aligned} \Big|\sum_{k=p}^{q} a_k c_k\Big| &\le M|c_q| + M|c_p| + M \sum_{k=p}^{q-1} |c_k - c_{k+1}| \\ &\le M|c_q| + M|c_p| + M\big(|c_p| + |c_q|\big) = 2M\big(|c_p| + |c_q|\big). \end{aligned}$$

Now $c_n \to 0$, so given $\varepsilon > 0$ choose N with $|c_n| < \dfrac{\varepsilon}{4M+1}$ for all $n \ge N$ (the $+1$ guards the case $M = 0$). For $q \ge p \ge N$,

$$\Big|\sum_{k=p}^{q} a_k c_k\Big| \le 2M\big(|c_p| + |c_q|\big) < 2M \cdot \frac{2\varepsilon}{4M+1} = \frac{4M}{4M+1}\,\varepsilon < \varepsilon.$$

This is the Cauchy criterion for $\sum a_n c_n$, so that series converges (6.6.2). Adding back the convergent series $b \sum a_n$ gives that $\sum a_n b_n = \sum a_n c_n + b \sum a_n$ converges.

■

REFLECTIONS.

The hypotheses are oddly mismatched, and that mismatch is the whole clue. About $\sum a_n$ we are told only that it *converges*—nothing about its terms or its sign—while about $\{b_n\}$ we are handed two strong shape conditions, *monotonic* and *bounded*. When a weak control on one factor is paired with rigid monotonicity on the other, the pairing names a single technique: summation by parts, the discrete analogue of integration by parts, which trades

the unknown product $a_n b_n$ for partial sums A_n (which we *do* control) multiplied against the *differences* $b_k - b_{k+1}$ (which monotonicity makes well-behaved). The two words "monotonic and bounded" together are also the exact hypothesis of monotone convergence (a bounded monotonic sequence converges; 6.2.2), so $\{b_n\}$ has a limit b to peel off.

A reader should believe the shape of the answer before the inequalities. The convergence of $\sum a_n$ pins the partial sums A_n inside a bounded window (a convergent sequence is bounded); a monotonic $\{c_n\}$ sliding down to 0 supplies weights whose *total* variation is finite—in fact just $|c_p - c_q|$, because a one-signed difference telescopes. Bounded sums against finite-variation weights ought to converge, and that intuition is exactly what the algebra below makes precise. A concrete instance fitting the hypotheses: take the convergent alternating series $\sum a_n = \sum(-1)^n/(n+1)$ and weight it by $b_n = n/(n+1)$, which increases monotonically to 1 and stays bounded; the theorem promises $\sum a_n b_n$ converges, and it does. The degenerate weight $b_n \equiv 1$ recovers the bare convergence of $\sum a_n$, so the content is precisely that a monotone bounded reweighting cannot break a convergent series.

The proof runs in a few linked moves, each resting on the one before:

(1) convergence of $\sum a_n$ makes $\{A_n\}$ a convergent, hence bounded, sequence (convergent $\Rightarrow$ bounded), giving the uniform bound $|A_n| \le M$ that every later estimate leans on;

(2) "monotonic and bounded" lets monotone convergence (6.2.2) extract the limit b and split $b_n = b + c_n$ with $c_n \to 0$ monotonic; the constant part $b\sum a_n$ converges by the algebra of series (limits respect the algebraic operations; 5.2.1), reducing everything to the single series $\sum a_n c_n$;

(3) summation by parts rewrites $\sum_{k=p}^{q} a_k c_k$ via $a_k = A_k - A_{k-1}$ as two boundary terms plus $\sum A_k(c_k - c_{k+1})$—this is the step that swaps the uncontrolled a_k for the controlled A_k;

(4) monotonicity of $\{c_n\}$ forces the differences $c_k - c_{k+1}$ to share one sign, so their absolute values telescope to $|c_p - c_q|$, and the whole block is bounded by $2M(|c_p| + |c_q|)$;

(5) $c_n \to 0$ drives that bound below any ε for p, q large, which is precisely the Cauchy criterion for series (6.6.2), and completeness of $\mathbb{R}$ turns "Cauchy" into "convergent" (Cauchy sequences in $\mathbb{R}$ converge).

Each hypothesis is load-bearing, and dropping any one breaks the result. Without convergence of $\sum a_n$ there is no bound M, and the boundary term $A_q c_q$ can run off even as $c_q \to 0$ (take $a_n = 1$, $b_n = 1$: $\sum a_n b_n$ diverges). Without monotonicity of $\{b_n\}$ the telescoping collapses—the differences may alternate in sign and their absolute sum can grow without bound, so the estimate $\sum |c_k - c_{k+1}| = |c_p - c_q|$ fails; a bounded but wildly oscillating

$\{b_n\}$ can destroy convergence even when $\sum a_n$ converges. Boundedness of $\{b_n\}$ is what manufactures the limit b; a monotonic but unbounded $\{b_n\}$ (say $b_n = n$) has $c_n \to \pm\infty$ and the argument has nothing to drive to zero. The reasoning is not circular: the bound on A_n and the limit b are established independently before summation by parts is invoked, and the Cauchy criterion is verified directly, never assuming the conclusion.

The transferable move is summation by parts itself, and the principle it encodes: to control a series whose terms are a product, do not estimate the product head-on—resum so that the well-behaved factor appears as a bounded partial sum and the rigid factor appears only through its differences, whose total variation a monotone sequence keeps finite. This is the engine behind both Dirichlet's test (bounded partial sums against a monotone null sequence) and Abel's test (a convergent series against a monotone bounded sequence), and the present problem is exactly Abel's test. The same trade—swap a difference operator onto the tame factor and a summation operator onto the partial sums—is what integration by parts does for integrals, and recognising the discrete twin is the lasting lesson.

Problem 3.9 *Rudin, Chapter 3, Exercise 9*

Find the radius of convergence of each of the following power series:

(a) $\sum n^3 z^n$,

(b) $\sum \dfrac{2^n}{n!} z^n$,

(c) $\sum \dfrac{2^n}{n^2} z^n$,

(d) $\sum \dfrac{n^3}{3^n} z^n$.

SOLUTION.

For a power series $\sum c_n z^n$ the radius of convergence is $R = 1/\alpha$, where $\alpha = \limsup_{n\to\infty} |c_n|^{1/n}$, with the conventions $R = +\infty$ when $\alpha = 0$ and $R = 0$ when $\alpha = +\infty$ (6.6.11). The single standard limit

$$n^{1/n} \longrightarrow 1 \qquad (6.5.1)$$

settles all four cases, since each $|c_n|^{1/n}$ converges—so its lim sup is its ordinary limit (6.4.5). The verdicts are $R = 1,\ +\infty,\ \frac{1}{2},\ 3$.

For (a), $c_n = n^3$, so $|c_n|^{1/n} = \left(n^{1/n}\right)^3 \to 1^3 = 1$ by the algebra of limits (limits respect products). Hence $\alpha = 1$ and $R = 1$.

For (b), $c_n = 2^n/n!$. Here the ratio is cleaner than the root: with all terms positive,

$$\frac{|c_{n+1}|}{|c_n|} = \frac{2^{n+1}/(n+1)!}{2^n/n!} = \frac{2}{n+1} \longrightarrow 0.$$

Since $|c_{n+1}|/|c_n| \to 0$, the same value bounds $\limsup |c_n|^{1/n}$ from above (6.A.1, the inequality behind the fact that the root test subsumes the ratio test), so $\alpha = 0$ and $R = +\infty$: the series converges for every $z \in \mathbb{C}$.

For (c), $c_n = 2^n/n^2$, so

$$|c_n|^{1/n} = \frac{2}{(n^2)^{1/n}} = \frac{2}{(n^{1/n})^2} \longrightarrow \frac{2}{1^2} = 2.$$

Hence $\alpha = 2$ and $R = \frac{1}{2}$.

For (d), $c_n = n^3/3^n$, so

$$|c_n|^{1/n} = \frac{(n^{1/n})^3}{3} \longrightarrow \frac{1}{3}.$$

Hence $\alpha = \frac{1}{3}$ and $R = 3$.

■

Reflections.

The phrase *radius of convergence* names exactly one object in the notes and hands you its formula at the same time: for $\sum c_n z^n$ the radius is $R = 1/\alpha$ with $\alpha = \limsup_n |c_n|^{1/n}$ (6.6.11), the Cauchy–Hadamard recipe. So a problem that asks for four radii is really asking for four computations of $\limsup |c_n|^{1/n}$; the conceptual work is done before any algebra, and what remains is to extract an nth root and take a limit. The reason the root appears at all is that $\sum c_n z^n$ converges precisely where the root test (6.6.9) on the terms $c_n z^n$ returns a value below 1, and $\limsup |c_n z^n|^{1/n} = |z|\,\alpha < 1$ is the inequality $|z| < 1/\alpha$ in disguise.

A single fact does almost all the lifting, and recognising it is the point of the exercise: every coefficient here is a power of n divided by an exponential, and $(n^p)^{1/n} = (n^{1/n})^p \to 1$ because $n^{1/n} \to 1$ (6.5.1). Believe that limit first on (a): the coefficients n^3 grow without bound, yet their nth roots $(n^{1/n})^3$ slide down to 1, so a polynomial factor, however large its degree, is invisible to the root test—it contributes a factor tending to 1 and cannot move the radius. That is why (a) has $R = 1$ exactly as the bare geometric series $\sum z^n$ does.

The four computations then run on the same two moves:

(1) take the nth root and split it across the product or quotient, which is legitimate because limits respect the algebraic operations (the algebra of limits, 5.2.1); the polynomial part becomes a power of $n^{1/n} \to 1$ and disappears, leaving only the exponential base—1 in (a), 2 in (c), $\frac{1}{3}$ in (d)—as the value of α;

(2) read off $R = 1/\alpha$, with $\alpha = 0 \Rightarrow R = +\infty$ and $\alpha = +\infty \Rightarrow R = 0$ the two boundary conventions (6.6.11).

Part (b) is the one place a root is awkward—$(n!)^{1/n}$ is not a standard limit in our notes—so the ratio test is the right instrument: $|c_{n+1}|/|c_n| = 2/(n+1) \to 0$. The factorial crushes the geometric growth 2^n, the ratio tends to 0, and a ratio that tends to a limit forces the root lim sup to the same limit (6.A.1); hence $\alpha = 0$ and the series is entire, $R = +\infty$. This is the same mechanism that makes the exponential series converge everywhere (6.6.12).

Two points keep the argument rigorous. First, the formula is stated with a lim sup, not a plain limit, because a general coefficient sequence need not have $|c_n|^{1/n}$ converging; here every one of the four does converge, so the lim sup collapses to the ordinary limit (6.4.5) and no upper envelope is needed—worth saying explicitly rather than silently treating lim sup as lim. Second, the radius records only the open disk $|z| < R$ where convergence is absolute; on the circle $|z| = R$ the formula is silent, and behaviour there genuinely varies—$\sum z^n$, $\sum z^n/n$, and $\sum z^n/n^2$ all have $R = 1$ yet differ completely on $|z| = 1$—so a complete answer names R and stops, without claiming anything on the boundary.

The transferable reading is that the root test is the engine of the radius, and within it a polynomial coefficient is weightless: R is governed entirely by the exponential rate, the base b in a coefficient $\sim n^p b^n$ giving $\alpha = b$ and $R = 1/b$, while factorials send α to 0 and the radius to infinity. When the root is clean, compute $\limsup |c_n|^{1/n}$ directly; when a factorial or a stubborn product appears, switch to the ratio $|c_{n+1}/c_n|$, which agrees with the root whenever it converges (6.A.1).

Problem 3.10 *Rudin, Chapter 3, Exercise 10*

Suppose that the coefficients of the power series $\sum a_n z^n$ are integers, infinitely many of which are distinct from zero. Prove that the radius of convergence is at most 1.

SOLUTION.

The radius of convergence of $\sum a_n z^n$ is given by the Cauchy–Hadamard formula (6.6.11)

$$R = \frac{1}{\limsup_{n\to\infty} |a_n|^{1/n}},$$

with the conventions $1/0 = +\infty$ and $1/\infty = 0$. To prove $R \le 1$ it is enough to prove $\limsup_{n\to\infty} |a_n|^{1/n} \ge 1$.

Let $S = \{n : a_n \neq 0\}$ be the set of indices of nonzero coefficients. By hypothesis S is infinite, so its elements form a strictly increasing sequence $n_1 < n_2 < n_3 < \cdots$. For each $n_k \in S$ the coefficient a_{n_k} is a *nonzero integer*, hence $|a_{n_k}| \ge 1$, and therefore

$$|a_{n_k}|^{1/n_k} \ge 1^{1/n_k} = 1 \qquad \text{for every } k.$$

Thus the subsequence $\left(|a_{n_k}|^{1/n_k}\right)_k$ of the sequence $\left(|a_n|^{1/n}\right)_n$ is bounded below by 1. Being a sequence in the (extended) reals, it has at least one subsequential limit L, and since each

of its terms is ≥ 1 and $[1, +\infty]$ is closed, that limit satisfies $L \geq 1$. The limit superior of $\left(|a_n|^{1/n}\right)_n$ is the largest of all its subsequential limits, and L is one of them (6.4.5, 6.4.6); hence

$$\limsup_{n\to\infty} |a_n|^{1/n} \;\geq\; L \;\geq\; 1.$$

Consequently

$$R = \frac{1}{\limsup_{n\to\infty} |a_n|^{1/n}} \;\leq\; \frac{1}{1} = 1,$$

so the radius of convergence is at most 1.

■

Reflections.

The phrase *radius of convergence* fixes the entire approach: the one tool in our notes that turns a sequence of coefficients into a radius is the Cauchy–Hadamard formula $R = 1/\limsup |a_n|^{1/n}$ (6.6.11). Reading the goal $R \leq 1$ through that formula flips it into a statement about the denominator—$R \leq 1$ is the same as $\limsup |a_n|^{1/n} \geq 1$—and a lower bound on a lim sup is the cheapest kind of statement to establish, because the limit superior only has to be large *somewhere*, along a single subsequence, not everywhere.

That is where the second clue does its work. The hypothesis is oddly specific: the coefficients are not merely real but *integers*, and *infinitely many* are nonzero. Each word earns its place. A nonzero integer cannot be small—it has absolute value at least 1—so on the indices where $a_n \neq 0$ the quantity $|a_n|^{1/n}$ is pinned at or above 1; and "infinitely many" guarantees those indices form a genuine subsequence rather than a finite scatter that a lim sup could ignore. The integrality supplies the floor and the infinitude supplies the subsequence, and the limit superior is precisely the device built to hear an infinitely-recurring floor.

It helps to believe the mechanism on a degenerate case first. Take $a_n = 1$ for all n: every $|a_n|^{1/n} = 1$, so $\limsup = 1$ and $R = 1$, the geometric series $\sum z^n$ converging exactly on $|z| < 1$ (6.6.7). Padding with zeros, as in $\sum z^{n^2}$ where the nonzero coefficients are 1, changes nothing essential: the nonzero terms still force the lim sup to be 1, and the radius stays 1. The bound $R \leq 1$ is therefore sharp, attained by the simplest integer series there is.

The argument is short, and each move rests on the one before:

(1) rewrite the target through Cauchy–Hadamard (6.6.11): $R \leq 1$ is equivalent to $\limsup |a_n|^{1/n} \geq 1$, so the radius is never touched again and the whole problem becomes a statement about the coefficient sequence;

(2) collect the nonzero indices $S = \{n : a_n \neq 0\}$, infinite by hypothesis, and list them as a strictly increasing subsequence $n_1 < n_2 < \cdots$—this is the step that spends "infinitely

many," since only an infinite S yields a true subsequence;

(3) on that subsequence $|a_{n_k}| \geq 1$ because a nonzero integer has absolute value at least 1, and raising a number ≥ 1 to the positive power $1/n_k$ keeps it ≥ 1—this is where "integers" is spent, and it is the only arithmetic in the proof;

(4) read off the lim sup: the subsequence sitting at or above 1 has a cluster value, itself ≥ 1 because $[1, +\infty]$ is closed, and the limit superior, as the largest subsequential limit (6.4.5, 6.4.6), dominates that cluster value, hence $\limsup |a_n|^{1/n} \geq 1$ and $R \leq 1$.

The rigor turns on using lim sup rather than lim in step (4), and on seeing that both hypotheses are load-bearing. The ordinary limit of $|a_n|^{1/n}$ need not exist—between the nonzero indices the coefficients may be 0, where $|a_n|^{1/n} = 0$, so the sequence can oscillate between 0 and values ≥ 1 forever. The limit superior is exactly the right instrument because it ignores the dips to 0 and reports the recurring height ≥ 1; replacing it by a plain limit would be circular, assuming a convergence the problem does not give. Drop "integers" and the floor vanishes—coefficients like $a_n = 1/n^n$ are nonzero yet have $|a_n|^{1/n} = 1/n \to 0$, giving $R = \infty$, so the conclusion is simply false. Drop "infinitely many nonzero" and the surviving terms are a finite set the lim sup discards; a polynomial such as $1 + z$ has all but finitely many coefficients zero and radius $R = \infty$. Each hypothesis blocks exactly one of these escapes.

The transferable move is to read a radius-of-convergence question as a question about the limit superior of the coefficients, and then to control that lim sup along a well-chosen subsequence rather than the whole sequence. A lower bound on lim sup needs only one recurring floor; an upper bound needs the floor to hold eventually everywhere. Here a discreteness constraint—nonzero integers cannot be small—manufactures the floor, the same discreteness that makes $\mathbb{Z}$ a closed, spread-out set and that powers many "integers cannot crowd" arguments. Whenever you must keep $|a_n|^{1/n}$ from collapsing to 0, look for a reason the nonzero a_n stay bounded away from 0, and let the lim sup hear it.

Problem 3.11 *Rudin, Chapter 3, Exercise 11*

Suppose $a_n > 0$, $s_n = a_1 + \cdots + a_n$, and $\sum a_n$ diverges.

1. Prove that $\sum \dfrac{a_n}{1 + a_n}$ diverges.
2. Prove that
$$\frac{a_{N+1}}{s_{N+1}} + \cdots + \frac{a_{N+k}}{s_{N+k}} \geq 1 - \frac{s_N}{s_{N+k}}$$
and deduce that $\sum \dfrac{a_n}{s_n}$ diverges.
3. Prove that
$$\frac{a_n}{s_n^2} \leq \frac{1}{s_{n-1}} - \frac{1}{s_n}$$

and deduce that $\sum \dfrac{a_n}{s_n^2}$ converges.

4. What can be said about

$$\sum \frac{a_n}{1+na_n} \quad \text{and} \quad \sum \frac{a_n}{1+n^2a_n}?$$

Solution.

Throughout, the terms are positive, so each partial sum s_n is strictly increasing; and because $\sum a_n$ diverges, the increasing sequence (s_n) is unbounded, that is $s_n \to +\infty$ (6.4.4). In particular $1/s_n \to 0$ (Archimedean).

(a) The series $\sum \frac{a_n}{1+a_n}$ diverges. Suppose, toward a contradiction, that it converges. Then its terms vanish (6.6.3), so $\frac{a_n}{1+a_n} \to 0$; solving $b_n = \frac{a_n}{1+a_n}$ for $a_n = \frac{b_n}{1-b_n}$ shows $a_n \to 0$ as well. Hence $a_n < 1$ for all n beyond some N, and there $1 + a_n < 2$, so

$$\frac{a_n}{1+a_n} > \frac{a_n}{2} \qquad (n \geq N).$$

Both series have positive terms, so by the comparison test (6.6.8) the convergence of $\sum \frac{a_n}{1+a_n}$ forces the convergence of $\sum \frac{a_n}{2}$, hence of $\sum a_n$. This contradicts the hypothesis that $\sum a_n$ diverges, so $\sum \frac{a_n}{1+a_n}$ diverges.

(b) Fix N and $k \geq 1$. For each index n with $N+1 \leq n \leq N+k$ the monotonicity $s_n \leq s_{N+k}$ gives $\frac{a_n}{s_n} \geq \frac{a_n}{s_{N+k}}$. Summing over these k indices,

$$\sum_{j=1}^{k} \frac{a_{N+j}}{s_{N+j}} \;\geq\; \frac{1}{s_{N+k}} \sum_{j=1}^{k} a_{N+j} \;=\; \frac{s_{N+k} - s_N}{s_{N+k}} \;=\; 1 - \frac{s_N}{s_{N+k}},$$

where $\sum_{j=1}^{k} a_{N+j} = s_{N+k} - s_N$ is the definition of the partial sums. This is the claimed inequality.

To deduce divergence, fix N and let $k \to \infty$. Since $s_{N+k} \to +\infty$ while s_N is fixed, $\frac{s_N}{s_{N+k}} \to 0$, so the right-hand side tends to 1; in particular there is a k with $1 - \frac{s_N}{s_{N+k}} > \frac{1}{2}$, hence

$$\frac{a_{N+1}}{s_{N+1}} + \cdots + \frac{a_{N+k}}{s_{N+k}} > \tfrac{1}{2}.$$

As this holds for *every* starting index N, the partial sums of $\sum \frac{a_n}{s_n}$ are not Cauchy: no tail block can be made uniformly small. By the Cauchy criterion for series (6.6.2), $\sum \frac{a_n}{s_n}$ diverges.

(c) For $n \geq 2$, the difference of reciprocals telescopes a single term:

$$\frac{1}{s_{n-1}} - \frac{1}{s_n} = \frac{s_n - s_{n-1}}{s_{n-1}\, s_n} = \frac{a_n}{s_{n-1}\, s_n}.$$

Since $0 < s_{n-1} < s_n$, we have $s_{n-1}s_n < s_n^2$, so $\frac{a_n}{s_{n-1}s_n} \geq \frac{a_n}{s_n^2}$, which is the claimed inequality

$$\frac{a_n}{s_n^2} \leq \frac{1}{s_{n-1}} - \frac{1}{s_n} \qquad (n \geq 2).$$

The right-hand side is a telescoping series with nonnegative terms, whose partial sums are

$$\sum_{n=2}^{m}\Big(\frac{1}{s_{n-1}} - \frac{1}{s_n}\Big) = \frac{1}{s_1} - \frac{1}{s_m} < \frac{1}{s_1} = \frac{1}{a_1},$$

bounded above by $1/a_1$. By comparison (6.6.8) the partial sums of the nonnegative series $\sum_{n\geq 2} \frac{a_n}{s_n^2}$ are likewise bounded, and a nonnegative series with bounded partial sums converges (6.6.6). Adjoining the single term a_1/s_1^2 does not affect convergence, so $\sum \frac{a_n}{s_n^2}$ converges. (The bound in fact gives $\sum_{n\geq 2} \frac{a_n}{s_n^2} \leq 1/a_1$.)

(d) The second series *always* converges; the first one can do either, so nothing can be said about it in general.

For $\sum \frac{a_n}{1+n^2a_n}$, drop the 1 in the denominator: for $n \geq 1$,

$$\frac{a_n}{1+n^2a_n} < \frac{a_n}{n^2a_n} = \frac{1}{n^2}.$$

The series $\sum 1/n^2$ converges (a p-series with $p = 2$, 6.6.5), so by comparison (6.6.8) the nonnegative series $\sum \frac{a_n}{1+n^2a_n}$ converges, whatever the divergent $\sum a_n$ may be.

For $\sum \frac{a_n}{1+na_n}$ no verdict is possible, as two divergent choices of $\sum a_n$ show.

- Taking $a_n = 1$ gives the divergent $\sum 1$, and $\frac{a_n}{1+na_n} = \frac{1}{1+n}$, so $\sum \frac{a_n}{1+na_n} = \sum \frac{1}{1+n}$ diverges (the harmonic series, 6.6.4).
- Taking
$$a_n = \begin{cases} 1, & n = k^2 \text{ for some integer } k \geq 1, \\ 1/n^2, & \text{otherwise,} \end{cases}$$
the perfect-square terms alone sum to $\sum_{k\geq 1} 1 = \infty$, so $\sum a_n$ diverges. But $\sum \frac{a_n}{1+na_n}$ converges: over the perfect squares $n = k^2$ the terms are $\frac{1}{1+k^2} < \frac{1}{k^2}$, summable in k; off the perfect squares $\frac{a_n}{1+na_n} < a_n = \frac{1}{n^2}$, summable as well. Both pieces converge (6.6.8, 6.6.5), so the whole series does.

Thus $\sum \frac{a_n}{1+na_n}$ may converge or diverge under the same hypotheses, and no general statement holds.

■

REFLECTIONS.

The fixed cast—positive terms a_n, partial sums $s_n = a_1 + \cdots + a_n$, and a *divergent* $\sum a_n$—hands you two facts to lean on before any part is read. Positivity makes (s_n) strictly increasing, and divergence of a positive-term series means exactly that those increasing partial sums run off to $+\infty$ (6.4.4); together they give $s_n \to +\infty$ and hence $1/s_n \to 0$ (Archimedean), the single quantitative input each later part quietly consumes. The four parts then ask the recurring Chapter 3 question—convergence or divergence—about series built by feeding a_n and s_n through various damping denominators, and the whole exercise is a tour of the standard tests: comparison (6.6.8), the Cauchy criterion (6.6.2), telescoping against the bounded-partial-sums criterion (6.6.6), and the p-series benchmark (6.6.5).

The unifying reading is that each denominator is a throttle on the term a_n, and the verdict turns on how hard it throttles relative to the divergence it is fighting. A throttle of bounded size ($1 + a_n$ once $a_n \to 0$) cannot kill the divergence; a throttle that grows like s_n slows it only logarithmically and still loses; a throttle that grows like s_n^2 finally wins. Believe this on the cleanest instance $a_n = 1$, where $s_n = n$: the three series become $\sum \frac{1}{2}$ (diverges), $\sum \frac{1}{n}$ (diverges, harmonically), and $\sum \frac{1}{n^2}$ (converges)—the harmonic-versus-$p = 2$ threshold in miniature, and the template the general argument upgrades.

The four deductions each rest on a named result:

(1) part (a) is a contradiction (0.5.5): a convergent $\sum \frac{a_n}{1+a_n}$ would force its terms to vanish (6.6.3), hence $a_n \to 0$, hence $1+a_n < 2$ eventually, so the comparison $\frac{a_n}{1+a_n} > \frac{a_n}{2}$ (6.6.8) would drag $\sum a_n$ into convergence—the inversion $a_n = \frac{b_n}{1-b_n}$ is the only nonobvious move, and it is what recovers $a_n \to 0$ from $\frac{a_n}{1+a_n} \to 0$;

(2) part (b) bounds each s_n in a block by the largest, s_{N+k}, so the block sum exceeds $\frac{s_{N+k}-s_N}{s_{N+k}} = 1 - \frac{s_N}{s_{N+k}}$; letting $k \to \infty$ with N fixed pushes this past $\frac{1}{2}$ because $s_{N+k} \to +\infty$, so a tail block stays large no matter where it starts—precisely the failure of the Cauchy criterion (6.6.2) that certifies divergence;

(3) part (c) telescopes: $\frac{1}{s_{n-1}} - \frac{1}{s_n} = \frac{a_n}{s_{n-1}s_n}$, and $s_{n-1} < s_n$ replaces $s_{n-1}s_n$ by the larger s_n^2 to give the inequality; the telescoping partial sums collapse to $\frac{1}{s_1} - \frac{1}{s_m} < \frac{1}{a_1}$, so the dominated nonnegative series has bounded partial sums and converges (6.6.6, 6.6.8);

(4) part (d) splits in two—$\sum \frac{a_n}{1+n^2a_n}$ is killed outright by dropping the 1 to get $< \frac{1}{n^2}$ and comparing to the convergent $p = 2$ series (6.6.5, 6.6.8), while $\sum \frac{a_n}{1+na_n}$ is answered by two examples (0.5.8), $a_n = 1$ for divergence and a sparse "spike" sequence for convergence.

The rigor lives in details easy to skate past. In (a) the case $a_n \not\to 0$ is not separate but *subsumed*: if the series converged its terms would vanish, so the proof never needs to suppose $a_n \to 0$ as a hypothesis—it derives it, then uses $a_n < 1$ only eventually, which is all comparison needs. In (b) the order of quantifiers is the whole point: the block can

be made large for *every* N, and that "for every N" is exactly the negation of the Cauchy criterion; proving it for one N would prove nothing. In (c) the strict inequality $s_{n-1} < s_n$ (positivity again) is what lets $s_{n-1}s_n < s_n^2$, and the telescoping bound $1/s_1$ is finite only because $1/s_m \to 0$, i.e. because $\sum a_n$ diverges—curiously, the *divergence* hypothesis is what makes this series converge *to* a clean bound. In (d), positivity is what licenses dropping the 1 to enlarge a denominator and shrink a term, and the convergent example must have $\sum a_n$ genuinely divergent—the perfect-square subsum $\sum 1$ supplies that, while everything off the squares is already $p = 2$-summable.

The transferable move is to read a denominator as a competition of growth rates: a damped positive series $\sum \frac{a_n}{d_n}$ converges or diverges according to whether d_n outgrows the divergence of $\sum a_n$, and the way to test it is to replace d_n by the dominant comparison—the bounded constant in (a), the block-maximal s_{N+k} in (b), the telescoping product $s_{n-1}s_n$ in (c), the clean n^2 in (d). The deepest of these is the telescoping trick of (c): writing a term as a difference $f(n-1) - f(n)$ turns an opaque series into a collapsing sum whose convergence you can read off the limit of f, the same idea that powers the integral-test intuition and the p-series threshold itself (6.6.5). And part (d) is the lasting caution—"$\sum a_n$ diverges" constrains the s_n^2-throttled series completely but says nothing about the na_n-throttled one, because a divergent series can hide its mass in a sparse subsequence that the linear weight na_n tames while the quadratic weight does not.

Problem 3.12 *Rudin, Chapter 3, Exercise 12*

Suppose $a_n > 0$ and $\sum a_n$ converges. Put

$$r_n = \sum_{m=n}^{\infty} a_m.$$

1. Prove that
$$\frac{a_m}{r_m} + \cdots + \frac{a_n}{r_n} > 1 - \frac{r_n}{r_m}$$
if $m < n$, and deduce that $\sum \frac{a_n}{r_n}$ diverges.

2. Prove that
$$\frac{a_n}{\sqrt{r_n}} < 2\left(\sqrt{r_n} - \sqrt{r_{n+1}}\right)$$
and deduce that $\sum \frac{a_n}{\sqrt{r_n}}$ converges.

SOLUTION.

Because $\sum a_n$ converges, each tail $r_n = \sum_{m=n}^{\infty} a_m$ is a finite real, and since every $a_m > 0$ each $r_n > 0$. The tails decrease strictly: $r_n - r_{n+1} = a_n > 0$, so

$$a_n = r_n - r_{n+1}, \qquad r_0 > r_1 > r_2 > \cdots > 0, \qquad r_n \to 0,$$

the last because r_n is the tail of a convergent series.

Part (a). Fix $m < n$. For each index k with $m \le k \le n$ the denominator satisfies $r_k \le r_m$ (the tails decrease), so $\dfrac{a_k}{r_k} \ge \dfrac{a_k}{r_m}$. Summing over $k = m, \dots, n$ and using $a_k = r_k - r_{k+1}$,

$$\sum_{k=m}^{n} \frac{a_k}{r_k} \;\ge\; \frac{1}{r_m} \sum_{k=m}^{n} (r_k - r_{k+1}) \;=\; \frac{r_m - r_{n+1}}{r_m} \;=\; 1 - \frac{r_{n+1}}{r_m},$$

the middle sum telescoping. Since $r_{n+1} < r_n$, we have $1 - \dfrac{r_{n+1}}{r_m} > 1 - \dfrac{r_n}{r_m}$, and therefore

$$\frac{a_m}{r_m} + \cdots + \frac{a_n}{r_n} > 1 - \frac{r_n}{r_m} \qquad (m < n).$$

This forces $\sum a_n/r_n$ to fail the Cauchy criterion (6.6.2). Fix m. As $n \to \infty$ the tail $r_n \to 0$, so $r_n/r_m \to 0$, and the block sum $\sum_{k=m}^{n} a_k/r_k > 1 - r_n/r_m$ exceeds $\frac{1}{2}$ for all large n. Thus for every m there is an $n > m$ with $\sum_{k=m}^{n} a_k/r_k > \frac{1}{2}$; the partial sums are not Cauchy, so $\sum a_n/r_n$ diverges.

Part (b). Factor $a_n = r_n - r_{n+1} = \big(\sqrt{r_n} - \sqrt{r_{n+1}}\big)\big(\sqrt{r_n} + \sqrt{r_{n+1}}\big)$, legitimate since $r_n, r_{n+1} > 0$. Dividing by $\sqrt{r_n}$,

$$\frac{a_n}{\sqrt{r_n}} = \big(\sqrt{r_n} - \sqrt{r_{n+1}}\big) \cdot \frac{\sqrt{r_n} + \sqrt{r_{n+1}}}{\sqrt{r_n}}.$$

From $r_{n+1} < r_n$ comes $\sqrt{r_{n+1}} < \sqrt{r_n}$, hence $\sqrt{r_n} + \sqrt{r_{n+1}} < 2\sqrt{r_n}$, so the fraction on the right is < 2. As $\sqrt{r_n} - \sqrt{r_{n+1}} > 0$, multiplying preserves the inequality and

$$\frac{a_n}{\sqrt{r_n}} < 2\big(\sqrt{r_n} - \sqrt{r_{n+1}}\big).$$

The right-hand side telescopes, which bounds the partial sums. For $N \le M$,

$$\sum_{n=N}^{M} 2\big(\sqrt{r_n} - \sqrt{r_{n+1}}\big) = 2\big(\sqrt{r_N} - \sqrt{r_{M+1}}\big) \le 2\sqrt{r_N},$$

since $r_{M+1} > 0$. Hence the partial sums of $\sum a_n/\sqrt{r_n}$, a series of positive terms, satisfy

$$\sum_{n=0}^{M} \frac{a_n}{\sqrt{r_n}} < \sum_{n=0}^{M} 2\big(\sqrt{r_n} - \sqrt{r_{n+1}}\big) = 2\big(\sqrt{r_0} - \sqrt{r_{M+1}}\big) < 2\sqrt{r_0},$$

bounded above by $2\sqrt{r_0}$. A series of nonnegative terms with bounded partial sums converges (6.6.6), so $\sum a_n/\sqrt{r_n}$ converges.

■

Reflections.

The data are a convergent series of positive terms and its tails $r_n = \sum_{m \geq n} a_m$, and the first move is to read what those tails are. The relation $a_n = r_n - r_{n+1}$ rewrites each term as a *difference* of consecutive tails, and a sum of consecutive differences is a telescoping sum (6.6.1)—the structural clue that governs both parts. Two facts about r_n are then forced and used throughout: the tails decrease strictly, because $r_n - r_{n+1} = a_n > 0$, and they vanish, $r_n \to 0$, precisely because r_n is the tail of a convergent series (6.6.2). The problem is a study in contrasts—the *same* weighting of a_n by a power of r_n diverges in (a) and converges in (b)—so each part ends not with a generic test but with a hand-built estimate that the tail structure makes available.

The two parts are deliberately paired on opposite shores: dividing a_n by r_n leaves too much mass and the series diverges, while dividing by the gentler $\sqrt{r_n}$ leaves little enough to converge. Believe the mechanism on the model $a_n = 2^{-n}$, where $r_n = 2^{-n+1}$: then $a_n/r_n = \frac{1}{2}$ is constant, so $\sum a_n/r_n$ is a sum of constants and diverges outright, while $a_n/\sqrt{r_n} = \frac{1}{2}\sqrt{r_n}$ shrinks geometrically and sums. Dividing by the full tail strips away exactly the decay that made $\sum a_n$ converge; taking a square root returns half of it.

Part (a) runs in three moves, each resting on the tail structure:

(1) over the block $m \leq k \leq n$ every denominator obeys $r_k \leq r_m$ because the tails decrease, so $a_k/r_k \geq a_k/r_m$—replacing each variable denominator by the single largest one r_m is what makes the block summable in closed form;

(2) the resulting sum $\frac{1}{r_m}\sum_{k=m}^{n}(r_k - r_{k+1})$ telescopes to $\frac{1}{r_m}(r_m - r_{n+1}) = 1 - r_{n+1}/r_m$, the one place $a_k = r_k - r_{k+1}$ is spent;

(3) $r_{n+1} < r_n$ weakens this to the stated $1 - r_n/r_m$, and now $r_n \to 0$ does the killing: holding m fixed and pushing $n \to \infty$ drives the block sum above $\frac{1}{2}$, so the partial sums are not Cauchy (6.6.2; a convergent series is Cauchy) and $\sum a_n/r_n$ diverges.

The verdict-then-justify shape matters: the Cauchy criterion is the right tool precisely because the divergence is not visible term by term—the terms a_n/r_n may individually tend to 0, yet *blocks* of them stay large, and only a criterion that watches blocks rather than single terms can see that.

Part (b) is the conjugate trick reincarnated. The term $a_n/\sqrt{r_n}$ resists direct comparison, but $a_n = r_n - r_{n+1}$ is a difference of squares of square roots, so factoring $a_n = (\sqrt{r_n} - \sqrt{r_{n+1}})(\sqrt{r_n} + \sqrt{r_{n+1}})$ exposes the very telescoping factor $\sqrt{r_n} - \sqrt{r_{n+1}}$ one wants to sum. The remaining factor $(\sqrt{r_n} + \sqrt{r_{n+1}})/\sqrt{r_n}$ is bounded by 2 for the same reason as in (a)—the tails decrease—and that single bound converts the awkward term into something whose sum collapses. The deduction is then a bounded-partial-sums argument (6.6.6): the majorant $\sum 2(\sqrt{r_n} - \sqrt{r_{n+1}})$ telescopes to $2(\sqrt{r_0} - \sqrt{r_{M+1}}) < 2\sqrt{r_0}$, a ceiling independent of M, so a positive series sitting beneath it must converge—this is the comparison test

(6.6.8) against a convergent telescoping series, with the lower bound $a_n/\sqrt{r_n} > 0$ supplied free by $a_n > 0$.

The argument is honest on each hypothesis. Positivity of the a_n is load-bearing twice over: it makes every $r_n > 0$ so that $\sqrt{r_n}$ and the divisions even parse, and it makes the tails strictly decreasing, which is the inequality $r_k \le r_m$ and $r_{n+1} < r_n$ that both parts run on; drop it and the tails need not be monotone and the estimates collapse. Convergence of $\sum a_n$ is what gives $r_n \to 0$, and that limit is the whole engine of (a)—without it the block sums need not approach 1 and divergence is lost. The reasoning is not circular: (a) never assumes $\sum a_n/r_n$ converges, it assumes the contrary and breaks the Cauchy condition, while (b) builds a convergent majorant by hand rather than borrowing the conclusion.

The transferable instinct is to let the tail r_n be the variable you compute with, not the terms a_n. Writing $a_n = r_n - r_{n+1}$ turns weighted sums of a_n into telescoping sums in r_n, and a monotone tail lets you replace a moving denominator by a fixed extreme one. The same difference-of-tails identity recurs whenever a series is reweighted by a function of its own tail, and the lesson of the pairing is quantitative: $\sum a_n/r_n^p$ inherits convergence from $\sum a_n$ when $p < 1$ and loses it when $p \ge 1$, with the conjugate factoring of (b) handling the borderline-beating exponent $p = \frac{1}{2}$.

Problem 3.13 *Rudin, Chapter 3, Exercise 13*

Prove that the Cauchy product of two absolutely convergent series converges absolutely.

Solution.

Let $\sum_{n\ge0} a_n$ and $\sum_{n\ge0} b_n$ be absolutely convergent series of real (or complex) numbers, and write

$$A = \sum_{n=0}^{\infty} |a_n| < \infty, \qquad B = \sum_{n=0}^{\infty} |b_n| < \infty$$

for the (finite) sums of their absolute values. Their Cauchy product is $\sum_{n\ge0} c_n$ with

$$c_n = \sum_{k=0}^{n} a_k\, b_{n-k}.$$

The claim is that $\sum c_n$ converges absolutely, i.e. that $\sum_n |c_n|$ converges.

Bound a single term by the triangle inequality:

$$|c_n| = \Big|\sum_{k=0}^{n} a_k b_{n-k}\Big| \le \sum_{k=0}^{n} |a_k|\,|b_{n-k}|.$$

Summing this over $n = 0, 1, \ldots, N$ gives, with every term nonnegative,

$$\sum_{n=0}^{N} |c_n| \le \sum_{n=0}^{N} \sum_{k=0}^{n} |a_k|\,|b_{n-k}| = \sum_{\substack{i,j\ge 0 \\ i+j\le N}} |a_i|\,|b_j|,$$

the last equality being the substitution $i = k$, $j = n - k$, which runs over exactly the pairs (i, j) of nonnegative integers with $i + j \le N$. Every such pair satisfies $i \le N$ and $j \le N$, so this triangle of indices sits inside the full square $\{(i, j) : 0 \le i, j \le N\}$; since all summands are nonnegative, enlarging the index set can only increase the sum:

$$\sum_{\substack{i,j\ge 0 \\ i+j\le N}} |a_i|\,|b_j| \le \sum_{i=0}^{N} \sum_{j=0}^{N} |a_i|\,|b_j| = \Big(\sum_{i=0}^{N} |a_i|\Big)\Big(\sum_{j=0}^{N} |b_j|\Big) \le A\,B.$$

Chaining the two displays,

$$\sum_{n=0}^{N} |c_n| \le A\,B \qquad \text{for every } N.$$

The partial sums of $\sum |c_n|$ are thus a nondecreasing sequence (the terms $|c_n| \ge 0$) bounded above by the constant $A\,B$. A nonnegative series whose partial sums are bounded converges (6.6.6)—its partial sums converge by monotone convergence (a bounded monotonic sequence converges; 6.2.2). Hence $\sum |c_n|$ converges, that is, the Cauchy product $\sum c_n$ converges absolutely.

■

REFLECTIONS.

The phrase that fixes everything is *converges absolutely*: the conclusion is not about the delicate value of $\sum c_n$ but only about $\sum |c_n|$, a series of *nonnegative* terms. That single observation steers the whole proof, because a nonnegative series has the friendliest convergence criterion in the notes—its partial sums climb monotonically, so it converges the instant they are bounded above (6.6.6), with no need to exhibit a limit or run an ε argument. The task therefore collapses to one inequality: produce a single constant that caps every partial sum of $\sum |c_n|$.

The word *Cauchy product* supplies the raw material, and reading its shape tells you where the cap comes from. The term $c_n = \sum_{k=0}^{n} a_k b_{n-k}$ collects all products $a_i b_j$ whose indices sum to exactly n; summing $|c_n|$ up to N therefore gathers all products $|a_i|\,|b_j|$ with $i + j \le N$. Picture those pairs (i, j) as lattice points: they fill a triangle below the diagonal line $i + j = N$. The two factors $A = \sum |a_i|$ and $B = \sum |b_j|$, multiplied out, gather instead the points of a *square*—every pair with $i \le N$ and $j \le N$. The triangle sits inside the

square, so the triangle's total cannot exceed the square's, and the square's total is just AB. Believe it on the smallest case: with $N = 1$ the triangle is $\{(0,0),(1,0),(0,1)\}$ and the square adds the corner $(1,1)$, so $|a_0||b_0| + |a_1||b_0| + |a_0||b_1| \le (|a_0| + |a_1|)(|b_0| + |b_1|)$ with the extra term $|a_1||b_1| \ge 0$ accounting for the slack.

The argument is then a short chain, each link resting on the one before:

(1) the triangle inequality turns $|c_n|$ into the dominating sum $\sum_k |a_k|\,|b_{n-k}|$, replacing signed (or complex) products by their nonnegative absolute values—the step that converts the problem into one about a nonnegative series;

(2) reindexing $(k, n-k) \mapsto (i, j)$ recognises $\sum_{n \le N} \sum_k |a_k||b_{n-k}|$ as the sum over the triangle $\{i + j \le N\}$, the bookkeeping that exposes the Cauchy product's diagonal structure;

(3) the triangle lies inside the square $\{i, j \le N\}$, and *because the terms are nonnegative* adding the missing corner pairs only enlarges the sum—this monotonicity of finite sums in their index set is the load-bearing move;

(4) the square sum factors as $\big(\sum_{i \le N} |a_i|\big)\big(\sum_{j \le N} |b_j|\big)$, each factor bounded by the full absolute sum, giving the uniform cap AB;

(5) bounded monotone partial sums converge (6.6.6, by monotone convergence a bounded monotonic sequence converges, 6.2.2), so $\sum |c_n|$ converges.

The rigor turns entirely on nonnegativity, and it is worth seeing exactly where each hypothesis is spent. Both series are assumed *absolutely* convergent precisely so that A and B are finite; mere convergence is not enough, since the rearrangement and the triangle-inside-square comparison are licensed only for nonnegative terms, and the Cauchy product of two conditionally convergent series can genuinely diverge—the classic witness is $a_n = b_n = (-1)^n/\sqrt{n+1}$, where each $|c_n| \ge 1$ and the product fails the term test (6.6.3). The comparison "triangle $\subseteq$ square" would be false for signed terms, where adding the corner could decrease the sum; it is the absolute values that make every step monotone. Nothing here is circular: A and B are finite by hypothesis, the bound AB is built by hand, and convergence is read off the bounded monotone partial sums rather than assumed.

The transferable instinct is to meet "absolutely convergent" by working entirely with nonnegative series, where bounded partial sums *are* convergence (6.6.6)—the same reflex that powers the comparison test (6.6.8) and settles Problem 3.7. The geometric heart, that a sum over the diagonals $i + j = n$ is dominated by the sum over the product grid $i \le N$, $j \le N$, is the discrete shadow of Fubini's theorem: when all terms are nonnegative a double sum may be totalled in any order, and the Cauchy product is exactly the diagonal slicing of the grid $\sum_{i,j} a_i b_j = \big(\sum_i a_i\big)\big(\sum_j b_j\big)$. That is also why the value of the product is AB's signed cousin $\big(\sum a_n\big)\big(\sum b_n\big)$ once convergence is secured—absolute convergence is

what buys the right to reorder.

Problem 3.14 *Rudin, Chapter 3, Exercise 14*

If $\{s_n\}$ is a complex sequence, define its arithmetic means σ_n by

$$\sigma_n = \frac{s_0 + s_1 + \cdots + s_n}{n+1} \qquad (n = 0, 1, 2, \ldots).$$

1. If $\lim s_n = s$, prove that $\lim \sigma_n = s$.
2. Construct a sequence $\{s_n\}$ which does not converge, although $\lim \sigma_n = 0$.
3. Can it happen that $s_n > 0$ for all n and that $\limsup s_n = \infty$, although $\lim \sigma_n = 0$?
4. Put $a_n = s_n - s_{n-1}$, for $n \geq 1$. Show that

$$s_n - \sigma_n = \frac{1}{n+1} \sum_{k=1}^{n} k a_k.$$

   Assume that $\lim(na_n) = 0$ and that $\{\sigma_n\}$ converges. Prove that $\{s_n\}$ converges. [This gives a converse of (a), but under the additional assumption that $na_n \to 0$.]

5. Derive the last conclusion from a weaker hypothesis: Assume $M < \infty$, $|na_n| \leq M$ for all n, and $\lim \sigma_n = \sigma$. Prove that $\lim s_n = \sigma$, by completing the following outline:

   If $m < n$, then

$$s_n - \sigma_n = \frac{m+1}{n-m}(\sigma_n - \sigma_m) + \frac{1}{n-m} \sum_{i=m+1}^{n} (s_n - s_i).$$

   For these i,

$$|s_n - s_i| \leq \frac{(n-i)M}{i+1} \leq \frac{(n-m-1)M}{m+2}.$$

   Fix $\varepsilon > 0$ and associate with each n the integer m that satisfies

$$m \leq \frac{n-\varepsilon}{1+\varepsilon} < m+1.$$

   Then $(m+1)/(n-m) \leq 1/\varepsilon$ and $|s_n - s_i| < M\varepsilon$. Hence

$$\limsup_{n\to\infty} |s_n - \sigma| \leq M\varepsilon.$$

   Since ε was arbitrary, $\lim s_n = \sigma$.

SOLUTION.

Write $\sigma_n = \frac{1}{n+1} \sum_{k=0}^{n} s_k$ throughout.

(a) Averaging preserves the limit. Suppose $s_n \to s$. Fix $\varepsilon > 0$. There is an index N with

$|s_k - s| < \varepsilon/2$ for every $k \geq N$ (5.1.1). For $n \geq N$, split the average about its limit:

$$\sigma_n - s = \frac{1}{n+1}\sum_{k=0}^{n}(s_k - s) = \frac{1}{n+1}\sum_{k=0}^{N-1}(s_k - s) + \frac{1}{n+1}\sum_{k=N}^{n}(s_k - s).$$

Let $C = \sum_{k=0}^{N-1} |s_k - s|$, a fixed number depending only on the frozen head. The tail terms each obey $|s_k - s| < \varepsilon/2$, and there are at most $n+1$ of them, so

$$|\sigma_n - s| \leq \frac{C}{n+1} + \frac{1}{n+1}\sum_{k=N}^{n}|s_k - s| < \frac{C}{n+1} + \frac{n-N+1}{n+1}\cdot\frac{\varepsilon}{2} \leq \frac{C}{n+1} + \frac{\varepsilon}{2}.$$

Since C is fixed, the Archimedean property (Archimedean property) gives an $N' \geq N$ with $C/(n+1) < \varepsilon/2$ for all $n \geq N'$. Then $|\sigma_n - s| < \varepsilon$ for $n \geq N'$, so $\sigma_n \to s$.

(b) Convergence of the means does not force convergence. Take $s_n = (-1)^n$. This sequence diverges, as it has the two distinct subsequential limits $+1$ and -1 (a limit is unique; 5.1.1). Its partial sums are

$$\sum_{k=0}^{n}(-1)^k = \begin{cases} 1, & n \text{ even}, \\ 0, & n \text{ odd}, \end{cases}$$

so $|s_0 + \cdots + s_n| \leq 1$ and $|\sigma_n| \leq \frac{1}{n+1} \to 0$. Hence $\sigma_n \to 0$ while $\{s_n\}$ does not converge.

(c) Yes—even with positive terms and an unbounded lim sup. The means vanish provided the spikes are sparse enough; the right tuning puts them at powers of 2. Define

$$s_n = \begin{cases} k, & n = 2^k \text{ for some integer } k \geq 1, \\ 2^{-n}, & \text{otherwise.} \end{cases}$$

Every term is strictly positive. The spike values $s_{2^k} = k$ form a subsequence tending to $+\infty$, so $\limsup_n s_n = \infty$ (6.4.5). For the means, fix n and split the partial sum into its non-spike part and its spikes. The non-spike terms are at most $\sum_{j=0}^{n} 2^{-j} < \sum_{j=0}^{\infty} 2^{-j} = 2$, a geometric series (6.6.7); the spikes present by index n are those with $2^k \leq n$, namely $k = 1, \ldots, L$ with $L = \lfloor \log_2 n \rfloor$, contributing $\sum_{k=1}^{L} k = \frac{L(L+1)}{2} \leq L^2$. Hence

$$0 < \sigma_n = \frac{1}{n+1}\sum_{k=0}^{n} s_k \leq \frac{2+L^2}{n+1}, \qquad L = \lfloor \log_2 n \rfloor \leq \log_2 n.$$

Now $(\log_2 n)^2$ grows far slower than $n+1$, so $\frac{2+L^2}{n+1} \to 0$ (the logarithm is dwarfed by any positive power, 6.5.1); by the squeeze (6.3.3) $\sigma_n \to 0$. Thus a sequence of strictly positive terms can have $\limsup s_n = \infty$ and yet $\lim \sigma_n = 0$: the spikes are real but so sparse that averaging drowns them.

(d) An identity, and a converse under $na_n \to 0$. With $a_k = s_k - s_{k-1}$ for $k \geq 1$, sum by

parts. Since $\sum_{k=1}^{n} k\,s_k - \sum_{k=1}^{n} k\,s_{k-1} = \sum_{k=1}^{n} k\,s_k - \sum_{j=0}^{n-1}(j+1)\,s_j = n\,s_n - \sum_{j=0}^{n-1} s_j$, we get

$$\sum_{k=1}^{n} k\,a_k = n\,s_n - \sum_{j=0}^{n-1} s_j = (n+1)s_n - \sum_{j=0}^{n} s_j = (n+1)(s_n - \sigma_n),$$

which rearranges to the claimed

$$s_n - \sigma_n = \frac{1}{n+1}\sum_{k=1}^{n} k\,a_k.$$

Now assume $na_n \to 0$ and $\sigma_n \to \sigma$. The numbers $b_k := k\,a_k$ satisfy $b_k \to 0$, so their own arithmetic means tend to 0 as well—this is precisely part (a) applied to $\{b_k\}$. But those means are

$$\frac{1}{n+1}\sum_{k=1}^{n} b_k = \frac{1}{n+1}\sum_{k=1}^{n} k\,a_k = s_n - \sigma_n$$

(the $k = 0$ term b_0 is absent, and dropping one fixed term changes a Cesàro average by $b_0/(n+1) \to 0$, so the conclusion of (a) is unaffected). Hence $s_n - \sigma_n \to 0$. Adding the convergent $\sigma_n \to \sigma$ (limits respect addition; 5.2.1) gives $s_n = (s_n - \sigma_n) + \sigma_n \to 0 + \sigma = \sigma$.

(e) The same conclusion from $|na_n| \le M$. Fix $m < n$. Averaging the identity of (d) over the block $i = m+1, \dots, n$ and isolating s_n yields the stated decomposition

$$s_n - \sigma_n = \frac{m+1}{n-m}\,(\sigma_n - \sigma_m) + \frac{1}{n-m}\sum_{i=m+1}^{n} (s_n - s_i). \tag{$*$}$$

To verify $(*)$, sum $s_i = \sigma_i + (s_i - \sigma_i)$ and use $\sum_{i=0}^{n} s_i = (n+1)\sigma_n$, $\sum_{i=0}^{m} s_i = (m+1)\sigma_m$: then $\sum_{i=m+1}^{n} s_i = (n+1)\sigma_n - (m+1)\sigma_m$, and

$$\frac{1}{n-m}\sum_{i=m+1}^{n} (s_n - s_i) = s_n - \frac{(n+1)\sigma_n - (m+1)\sigma_m}{n-m} = s_n - \sigma_n - \frac{m+1}{n-m}(\sigma_n - \sigma_m),$$

which is $(*)$ rearranged.

The block estimate. For $m < i \le n$, telescoping gives $s_n - s_i = \sum_{j=i+1}^{n} a_j$, and $|a_j| = |ja_j|/j \le M/j \le M/(i+1)$ for $j > i$, so

$$|s_n - s_i| \le \sum_{j=i+1}^{n} \frac{M}{j} \le (n-i)\cdot\frac{M}{i+1} = \frac{(n-i)M}{i+1} \le \frac{(n-m-1)M}{m+2},$$

the final step because $i \ge m+1$ makes the numerator $n - i \le n - m - 1$ no larger and the denominator $i + 1 \ge m + 2$ no smaller.

Choosing m. Fix $\varepsilon > 0$. For each n large enough that $n > \varepsilon$, pick the integer m with

$m \le \frac{n-\varepsilon}{1+\varepsilon} < m+1$; then $0 \le m < n$. From $m \le \frac{n-\varepsilon}{1+\varepsilon}$,

$$(m+1)(1+\varepsilon) \le m(1+\varepsilon) + (1+\varepsilon) \le (n-\varepsilon) + (1+\varepsilon) = n+1,$$

so $(m+1)\varepsilon \le n+1-(m+1) = n-m$, hence $\dfrac{m+1}{n-m} \le \dfrac{1}{\varepsilon}$. From $\frac{n-\varepsilon}{1+\varepsilon} < m+1$ we get $n-\varepsilon < (m+1)(1+\varepsilon)$, i.e. $n-m-1 < \varepsilon(m+2)$, hence $\dfrac{(n-m-1)M}{m+2} < M\varepsilon$. Combining with the block estimate, $|s_n - s_i| < M\varepsilon$ for every i in the block.

Passing to the limit. Apply these to $(*)$:

$$|s_n - \sigma_n| \le \frac{m+1}{n-m}\,|\sigma_n - \sigma_m| + \frac{1}{n-m}\sum_{i=m+1}^{n} |s_n - s_i| \le \frac{1}{\varepsilon}\,|\sigma_n - \sigma_m| + M\varepsilon.$$

As $n \to \infty$ the chosen $m = m(n) \to \infty$ as well (since $m+1 > \frac{n-\varepsilon}{1+\varepsilon} \to \infty$), and $\sigma_n \to \sigma$, $\sigma_m \to \sigma$, so $|\sigma_n - \sigma_m| \to 0$ (a convergent sequence is Cauchy; 5.4.2). Hence

$$\limsup_{n\to\infty} |s_n - \sigma_n| \le M\varepsilon, \qquad \text{so} \qquad \limsup_{n\to\infty} |s_n - \sigma| \le 0 + M\varepsilon = M\varepsilon,$$

where the equality used $s_n - \sigma = (s_n - \sigma_n) + (\sigma_n - \sigma)$ with $\sigma_n - \sigma \to 0$. Since $\varepsilon > 0$ was arbitrary, $\limsup_n |s_n - \sigma| = 0$, i.e. $\lim s_n = \sigma$.

■

Reflections.

The object the problem hands you is the running average σ_n, the *Cesàro mean* of $\{s_n\}$, and the verbs steering each part are different. Part (a) says "if $\lim s_n = s$, prove $\lim \sigma_n = s$"—a convergence claim to establish directly from the definition (5.1.1); part (b) says "construct," part (c) asks "can it happen," both demanding an explicit witness; and parts (d) and (e) ask for a partial *converse*—averages can converge without the sequence converging, by (b), so the converse can be true only with a brake on how fast s_n may move, which is exactly what $na_n \to 0$ and $|na_n| \le M$ supply.

The engine of part (a) is the head–tail split, the single most useful move for any limit attached to a growing average or sum. The reflex it answers is this: $s_k - s$ is small once $k \ge N$, but the first N terms can be anything, so quarantine them. Their total C is a *fixed* number, and a fixed number divided by $n+1$ is crushed by the Archimedean property (Archimedean property)—that is why the head is harmless. The tail, all of it within $\varepsilon/2$ of s, averages to within $\varepsilon/2$. Believe it on the slowest honest example, $s_n = 1/n$: the means are $\frac{1}{n+1}(1 + \frac{1}{2} + \cdots + \frac{1}{n}) \approx \frac{\ln n}{n} \to 0$, the limit preserved even though the averaging visibly drags.

Part (b) is the counterexample that makes the converse interesting, and it reads straight off (a)'s proof: averaging *smooths*, so a bounded oscillation whose partial sums stay bounded

must have means going to 0. The alternating $(-1)^n$ is the cleanest such signal—its partial sums never leave $\{0, 1\}$, so $|\sigma_n| \le 1/(n+1) \to 0$, while the sequence itself has two limits and so has none (a limit is unique). The lesson is that σ_n forgets oscillation.

Part (c) pushes that intuition to its limit: can the means vanish while the terms are positive and shoot to $+\infty$ along a subsequence? They can, provided the spikes are *sparse*, and the honest part of the answer is calibrating that sparsity. Spikes at the squares would still be too dense—heights $k = 1, \dots, L$ at $n = k^2$ (so $L \approx \sqrt{n}$) sum to $\frac{L(L+1)}{2} \approx n/2$, leaving $\sigma_n \to \frac{1}{2}$, a positive limit rather than 0—so spacing them geometrically, at 2^k with height k, makes their running total only $O((\log n)^2)$ against a denominator $n+1$. The squeeze (6.3.3), fed by the fact that a logarithm is dwarfed by any positive power (6.5.1), then sends σ_n to 0, while the geometric baseline 2^{-n} keeps the non-spike mass bounded by a convergent series (6.6.7). The transferable gauge is that a Cesàro mean sees a subsequence only in proportion to how densely its indices sit.

Part (d) is where the converse begins, and it turns on an exact identity proved by summation by parts:

(1) writing $a_k = s_k - s_{k-1}$ and reindexing $\sum k\, s_{k-1}$ collapses $\sum_{k=1}^{n} k\, a_k$ to $n\, s_n - \sum_{j=0}^{n-1} s_j = (n+1)(s_n - \sigma_n)$, which is the stated identity—the discrete analogue of integration by parts, with k playing the role of the slowly varying factor;

(2) reading the identity as $s_n - \sigma_n = \frac{1}{n+1} \sum_{k=1}^{n} b_k$ with $b_k = k\, a_k$ recognises the right side as the Cesàro mean of $\{b_k\}$, so the whole problem feeds on itself: $na_n \to 0$ means $b_k \to 0$, and part (a) (the limit 0 is preserved by averaging) forces $s_n - \sigma_n \to 0$;

(3) adding the assumed $\sigma_n \to \sigma$ back, by the algebra of limits (limits respect the operations; 5.2.1), recovers $s_n = (s_n - \sigma_n) + \sigma_n \to \sigma$.

The hypothesis $na_n \to 0$ is doing precise work: it is the rate condition that says successive terms $a_n = s_n - s_{n-1}$ shrink faster than $1/n$, just fast enough that their k-weighted average dies. Drop it and (b) already refutes the converse, so the condition is not decorative.

Part (e) trades the limit $na_n \to 0$ for the mere bound $|na_n| \le M$, and the gap is bridged by an identity over a *block* rather than the whole head. The decomposition $(*)$ writes $s_n - \sigma_n$ as a controlled multiple of $\sigma_n - \sigma_m$ plus an average of differences $s_n - s_i$ across a window $m < i \le n$. The bound $|na_n| \le M$ converts to $|a_j| \le M/j$, and telescoping over the window gives $|s_n - s_i| \le (n-i)M/(i+1)$: the terms cannot have drifted far if their increments are $O(1/j)$. The clever choice $m \approx \frac{n}{1+\varepsilon}$ tunes the window so that both error pieces are $O(\varepsilon)$ at once—the prefactor $\frac{m+1}{n-m} \le 1/\varepsilon$ keeps the first term in check once $\sigma_n - \sigma_m \to 0$, and the block bound becomes $< M\varepsilon$. Here is the one subtlety to guard: as $n \to \infty$ the chosen $m = m(n)$ also tends to ∞, so σ_m and σ_n chase the *same* limit and their difference vanishes—this is the Cauchy reading of a convergent sequence (convergent

sequences are Cauchy; 5.4.2), and it is where the assumption $\sigma_n \to \sigma$ is genuinely spent. The conclusion is then squeezed through $\limsup$ (6.4.5): $\limsup |s_n - \sigma| \le M\varepsilon$ for every $\varepsilon > 0$ forces the limsup to be 0, hence the limit to be σ.

What ties the five parts into one arc is the principle that averaging is a smoothing, information-losing operation: it always preserves a limit (a), it can manufacture one out of oscillation (b) or sparse blowup (c), and recovering the original sequence from its means is possible only with a quantitative leash on the increments—first $na_n \to 0$ (d), then the weaker $|na_n| \le M$ (e). These are the Cesàro and Tauberian theorems in miniature, and the reusable craftsmanship is the trio of moves: split a long average into a fixed head plus a uniformly small tail, read a weighted sum as a Cesàro mean of an auxiliary sequence, and when a single limit is awkward, control a sliding block whose width you tune to the error you are willing to tolerate.

Problem 3.15 *Rudin, Chapter 3, Exercise 15*

Definition 3.21 can be extended to the case in which the a_n lie in some fixed R^k. Absolute convergence is defined as convergence of $\sum |\mathbf{a}_n|$. Show that Theorems 3.22, 3.23, 3.25(a), 3.33, 3.34, 3.42, 3.45, 3.47, and 3.55 are true in this more general setting. (Only slight modifications are required in any of the proofs.)

Solution.

Throughout, $\mathbf{a}_n \in \mathbb{R}^k$, the partial sums are $\mathbf{s}_n = \sum_{j=0}^{n} \mathbf{a}_j$, and $|\cdot|$ is the Euclidean norm. Two facts about that norm are all the modifications need: it obeys the triangle inequality $|\mathbf{u}+\mathbf{v}| \le |\mathbf{u}| + |\mathbf{v}|$, and $\mathbb{R}^k$ is complete ($\mathbb{R}^k$ is complete; 5.5.3). Each theorem below is the real-line statement with the scalar absolute value read as the $\mathbb{R}^k$ norm; the proof in the notes is quoted and only the place where $|\cdot|$ enters is checked.

Theorem 3.22 (Cauchy criterion). $\sum \mathbf{a}_n$ converges iff for every $\varepsilon > 0$ there is N with $\left|\sum_{j=n}^{m} \mathbf{a}_j\right| \le \varepsilon$ whenever $m \ge n \ge N$. Convergence of the series *is* convergence of the sequence $(\mathbf{s}_n)$ in $\mathbb{R}^k$, and since $\mathbb{R}^k$ is complete that sequence converges iff it is Cauchy (5.4.3; Cauchy sequences); writing $\mathbf{s}_m - \mathbf{s}_{n-1} = \sum_{j=n}^{m} \mathbf{a}_j$ turns the Cauchy condition on $(\mathbf{s}_n)$ into the stated tail condition (6.6.2). Completeness of $\mathbb{R}^k$ is the only ingredient that must be invoked in place of completeness of $\mathbb{R}$.

Theorem 3.23 (terms vanish). If $\sum \mathbf{a}_n$ converges then $\mathbf{a}_n \to \mathbf{0}$. Take $m = n$ in 3.22: for $n \ge N$, $|\mathbf{a}_n| = \left|\sum_{j=n}^{n} \mathbf{a}_j\right| \le \varepsilon$, so $\mathbf{a}_n \to \mathbf{0}$ (6.6.3).

Theorem 3.25(a) (comparison). If $|\mathbf{a}_n| \le c_n$ for all $n \ge N_0$ and $\sum c_n$ converges, then $\sum \mathbf{a}_n$ converges. By the triangle inequality, for $m \ge n \ge N_0$,

$$\Big|\sum_{j=n}^{m} \mathbf{a}_j\Big| \le \sum_{j=n}^{m} |\mathbf{a}_j| \le \sum_{j=n}^{m} c_j.$$

Since $\sum c_n$ converges, its tails are $\le \varepsilon$ for $n \ge N$ by 3.22 applied to the real series, so the left side is $\le \varepsilon$ as well; the criterion 3.22 for the $\mathbb{R}^k$ series is met, and $\sum \mathbf{a}_n$ converges (6.6.8). The triangle inequality $|\sum \mathbf{a}_j| \le \sum |\mathbf{a}_j|$ is the one scalar step ($|\sum a_j| \le \sum |a_j|$) reread in $\mathbb{R}^k$.

Theorems 3.33 and 3.34 (root and ratio tests). If $\limsup_n |\mathbf{a}_n|^{1/n} < 1$, or if $|\mathbf{a}_{n+1}|/|\mathbf{a}_n| \le \beta < 1$ for all large n, then $\sum \mathbf{a}_n$ converges. Both tests are statements about the *scalar* series $\sum |\mathbf{a}_n|$ of nonnegative terms, to which the real root and ratio tests apply verbatim (6.6.9, 6.6.10): each produces a convergent geometric majorant $\sum c_n$ with $|\mathbf{a}_n| \le c_n$, and comparison 3.25(a) above then gives convergence of $\sum \mathbf{a}_n$. Divergence in the root test still follows from 3.23, since $\limsup |\mathbf{a}_n|^{1/n} > 1$ forces $|\mathbf{a}_n| \not\to 0$, hence $\mathbf{a}_n \not\to \mathbf{0}$.

Theorem 3.42 (Dirichlet's test). Let $\mathbf{a}_n \in \mathbb{R}^k$ have bounded partial sums $\mathbf{A}_n$, and let $b_0 \ge b_1 \ge \cdots$ be *real* with $b_n \to 0$; then $\sum b_n \mathbf{a}_n$ converges. Summation by parts is an algebraic identity valid coordinate by coordinate, so with $\mathbf{A}_{-1} = \mathbf{0}$,

$$\sum_{j=n}^{m} b_j \mathbf{a}_j = \sum_{j=n}^{m-1} (b_j - b_{j+1}) \mathbf{A}_j + b_m \mathbf{A}_m - b_n \mathbf{A}_{n-1}.$$

With $|\mathbf{A}_j| \le M$ for all j and $b_j - b_{j+1} \ge 0$, the triangle inequality gives

$$\Big|\sum_{j=n}^{m} b_j \mathbf{a}_j\Big| \le M\Big(\sum_{j=n}^{m-1} (b_j - b_{j+1}) + b_m + b_n\Big) = 2M b_n,$$

the sum telescoping. Since $b_n \to 0$ this is $\le \varepsilon$ for n large, so 3.22 yields convergence. Only the bound on $\mathbf{A}_j$ and the triangle inequality changed; the weights b_j stay real, which is what keeps the scalar multiples $b_j \mathbf{a}_j$ in $\mathbb{R}^k$.

Theorem 3.45 (absolute convergence implies convergence). If $\sum |\mathbf{a}_n|$ converges then $\sum \mathbf{a}_n$ converges. This is the comparison 3.25(a) with $c_n = |\mathbf{a}_n|$, since $|\mathbf{a}_n| \le c_n$ trivially and $\sum c_n = \sum |\mathbf{a}_n|$ converges by hypothesis.

Theorem 3.47 (linearity). If $\sum \mathbf{a}_n = \mathbf{A}$ and $\sum \mathbf{b}_n = \mathbf{B}$, then $\sum(\mathbf{a}_n + \mathbf{b}_n) = \mathbf{A} + \mathbf{B}$ and $\sum c\,\mathbf{a}_n = c\,\mathbf{A}$ for $c \in \mathbb{R}$. The partial sums satisfy $\sum_{j \le n}(\mathbf{a}_j + \mathbf{b}_j) = \mathbf{s}_n + \mathbf{t}_n$ and $\sum_{j \le n} c\,\mathbf{a}_j = c\,\mathbf{s}_n$, and limits in $\mathbb{R}^k$ respect sums and scalar multiples (the algebra of limits; 5.2.1), so these tend to $\mathbf{A} + \mathbf{B}$ and $c\,\mathbf{A}$.

Theorem 3.55 (rearrangement). If $\sum \mathbf{a}_n$ converges absolutely, every rearrangement $\sum \mathbf{a}_{\sigma(n)}$ converges to the same sum. The proof in the notes uses only the Cauchy tail of $\sum |\mathbf{a}_n|$ and the triangle inequality. Given $\varepsilon > 0$, absolute convergence and 3.22 applied to the real series $\sum |\mathbf{a}_n|$ give N with $\sum_{j>N} |\mathbf{a}_j| \le \varepsilon$. Choose p so large that $\{0, \dots, N\} \subseteq \{\sigma(0), \dots, \sigma(p)\}$;

then for the rearranged partial sums $\mathbf{s}'_n$ and any $q \geq p$, the indices $0, \ldots, N$ cancel and

$$|\mathbf{s}'_q - \mathbf{s}_q| \leq \sum_{j>N} |\mathbf{a}_j| \leq \varepsilon,$$

where $\mathbf{A} = \sum \mathbf{a}_n$ exists because absolute convergence forces convergence (3.45 above). Thus $\mathbf{s}'_q - \mathbf{s}_q \to \mathbf{0}$, and since $\mathbf{s}_q \to \mathbf{A}$ this gives $\mathbf{s}'_q \to \mathbf{A}$: the rearrangement converges to the same sum. The only $\mathbb{R}^k$ inputs are the norm reading of $|\mathbf{s}'_q - \mathbf{s}_q| \leq \sum_{j>N} |\mathbf{a}_j|$ and the existence of $\mathbf{A}$, both already secured above.

In every case the only edits are to read the absolute value as the $\mathbb{R}^k$ norm, to invoke completeness of $\mathbb{R}^k$ rather than of $\mathbb{R}$ in the Cauchy criterion, and to use the triangle inequality $\left|\sum \mathbf{a}_j\right| \leq \sum |\mathbf{a}_j|$ in place of its scalar twin. The two tests and absolute convergence reduce outright to the real series $\sum |\mathbf{a}_n|$.

■

Reflections.

The instruction *only slight modifications are required* is the whole exercise stated as a hint: the task is not to rebuild nine proofs but to locate, in each, the single line where the real number a_n was treated as a scalar and ask whether the same line survives when a_n becomes a vector $\mathbf{a}_n \in \mathbb{R}^k$. That reframing tells you to read each theorem looking for two things only, because those are the two places a scalar proof can secretly use that it lives on the line: where it manipulates $|a_n|$ as an absolute value, and where it concludes a Cauchy sequence converges. The first is governed by the triangle inequality, which the Euclidean norm shares; the second by completeness, which $\mathbb{R}^k$ shares ($\mathbb{R}^k$ is complete; 5.5.3). Pin those two down and the rest is transcription.

It is worth believing the principle on the smallest case before trusting it nine times. A series $\sum \mathbf{a}_n$ in $\mathbb{R}^k$ is nothing but its sequence of partial sums $\mathbf{s}_n = \sum_{j \leq n} \mathbf{a}_j$, and that sequence converges iff each of its k coordinate sequences does (coordinatewise convergence; 5.2.2)—so a vector series is, under the hood, k ordinary real series running in parallel. One could prove every part coordinatewise. The reason the norm-and-completeness route is cleaner is that it never has to split into coordinates: the single inequality $\left|\sum \mathbf{a}_j\right| \leq \sum |\mathbf{a}_j|$ does in one stroke what k separate triangle inequalities would do, which is exactly why Rudin can call the modifications slight.

The nine results sort into a small number of mechanisms, and walking them in dependency order shows that only the first two carry any genuine $\mathbb{R}^k$ content:

(1) Theorem 3.22 is the load-bearing one, because convergence of a series is by definition convergence of $(\mathbf{s}_n)$, and the only theorem that lets you certify that convergence without naming the limit is "Cauchy ⇒ convergent" (5.4.3, 6.6.2)—which holds in $\mathbb{R}^k$

precisely because $\mathbb{R}^k$ is complete (Cauchy sequences in a complete space); this is the one spot where completeness of $\mathbb{R}$ is swapped for completeness of $\mathbb{R}^k$;

(2) Theorem 3.25(a) is the second engine, and its only vector step is the triangle inequality $|\sum_{j=n}^{m} \mathbf{a}_j| \le \sum_{j=n}^{m} |\mathbf{a}_j|$, the norm reading of the scalar $|\sum a_j| \le \sum |a_j|$; with that in hand the convergent real majorant $\sum c_n$ forces the tail of the vector series small, and 3.22 closes it (6.6.8);

(3) everything else is downstream. Theorem 3.23 is 3.22 with $m = n$; the root and ratio tests (6.6.9, 6.6.10) are theorems about the *nonnegative real* series $\sum |\mathbf{a}_n|$ that build a geometric majorant and then call 3.25(a); Theorem 3.45 is 3.25(a) with $c_n = |\mathbf{a}_n|$; and Theorem 3.55 reuses the absolute-convergence tail of $\sum |\mathbf{a}_n|$ with one triangle inequality on the rearranged difference;

(4) two parts use a different but equally portable tool. Theorem 3.47 is pure algebra of limits on partial sums (the algebra of limits; 5.2.1), which already holds in $\mathbb{R}^k$; and Theorem 3.42 leans on summation by parts, an identity that holds coordinate by coordinate, after which the bounded partial sums $|\mathbf{A}_j| \le M$ and the triangle inequality telescope the estimate to $2Mb_n \to 0$.

The rigor check is to see which hypotheses are genuinely doing the work and which were never about the line at all. Completeness is not optional: it is exactly the property that fails in $\mathbb{Q}$, where a series of rationals can be Cauchy yet sum to an irrational, so 3.22 would be false—the slight modification "$\mathbb{R} \to \mathbb{R}^k$" is licensed only because $\mathbb{R}^k$ keeps completeness (complete). The triangle inequality is the other non-negotiable, and it is the *only* thing the norm needs to supply; nowhere does any proof multiply two vectors or order them, so the absence of a product or an order on $\mathbb{R}^k$ never bites. Dirichlet's test makes that visible: the weights b_n must stay *real* and monotone, since "$b_0 \ge b_1 \ge \cdots$" has no meaning for vectors and the scalar multiples $b_n\mathbf{a}_n$ are what keep the terms in $\mathbb{R}^k$. And the reduction of the root and ratio tests is honest rather than circular, because $\sum |\mathbf{a}_n|$ is a bona fide real series of nonnegative terms to which the unmodified real theorems apply; the vector series is reached only afterwards, through comparison.

The transferable lesson is what "slight modification" really certifies: a proof generalises exactly as far as the structure it actually uses. These nine arguments use only that the codomain is a complete normed space, so they hold verbatim in $\mathbb{R}^k$—and, for the same reason, in any Banach space, which is the setting they secretly belonged to all along. The two that quietly need more, Dirichlet's test with its monotone weights and any future result that multiplies or compares terms, are precisely the ones that keep a foot on the real line; spotting which is which, by auditing each proof for the properties of $|\cdot|$ it leans on, is the skill the exercise is training.

Problem 3.16 *Rudin, Chapter 3, Exercise 16*

Fix a positive number α. Choose $x_1 > \sqrt{\alpha}$, and define $x_2, x_3, x_4, \ldots$, by the recursion formula

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{\alpha}{x_n}\right).$$

1. Prove that $\{x_n\}$ decreases monotonically and that $\lim x_n = \sqrt{\alpha}$.

2. Put $\varepsilon_n = x_n - \sqrt{\alpha}$, and show that

$$\varepsilon_{n+1} = \frac{\varepsilon_n^2}{2x_n} < \frac{\varepsilon_n^2}{2\sqrt{\alpha}}$$

so that, setting $\beta = 2\sqrt{\alpha}$,

$$\varepsilon_{n+1} < \beta\left(\frac{\varepsilon_1}{\beta}\right)^{2^n} \qquad (n = 1, 2, 3, \ldots).$$

3. This is a good algorithm for computing square roots, since the recursion formula is simple and the convergence is extremely rapid. For example, if $\alpha = 3$ and $x_1 = 2$, show that $\varepsilon_1/\beta < \frac{1}{10}$ and that therefore

$$\varepsilon_5 < 4 \cdot 10^{-16}, \qquad \varepsilon_6 < 4 \cdot 10^{-32}.$$

Solution.

Throughout, $\sqrt{\alpha} > 0$ is the positive square root of $\alpha > 0$.

Part (a). First, every term is positive: $x_1 > \sqrt{\alpha} > 0$, and if $x_n > 0$ then $x_{n+1} = \frac{1}{2}(x_n + \alpha/x_n)$ is a sum of positive numbers divided by 2, hence positive. So $x_n > 0$ for all n by induction (0.6.2), and the recursion never divides by zero.

Next, $x_n > \sqrt{\alpha}$ for every n. This holds at $n = 1$ by hypothesis. The identity

$$x_{n+1} - \sqrt{\alpha} = \frac{1}{2}\Big(x_n + \frac{\alpha}{x_n}\Big) - \sqrt{\alpha} = \frac{x_n^2 - 2\sqrt{\alpha}\,x_n + \alpha}{2x_n} = \frac{(x_n - \sqrt{\alpha})^2}{2x_n}$$

shows the right-hand side is ≥ 0, with equality only when $x_n = \sqrt{\alpha}$. So if $x_n > \sqrt{\alpha}$ then $(x_n - \sqrt{\alpha})^2 > 0$ and $x_{n+1} > \sqrt{\alpha}$; by induction $x_n > \sqrt{\alpha}$ for all n.

The sequence decreases. From $x_n > \sqrt{\alpha}$ we get $x_n^2 > \alpha$, so

$$x_n - x_{n+1} = x_n - \frac{1}{2}\Big(x_n + \frac{\alpha}{x_n}\Big) = \frac{x_n^2 - \alpha}{2x_n} > 0,$$

i.e. $x_{n+1} < x_n$. Thus $\{x_n\}$ is monotonically decreasing (6.2.1) and bounded below by $\sqrt{\alpha}$, so it converges (a bounded monotone sequence converges; 6.2.2); call the limit L. Since

$x_n > \sqrt{\alpha}$ for all n, the order survives in the limit (6.3.1), giving $L \geq \sqrt{\alpha} > 0$. Passing to the limit in $x_{n+1} = \frac{1}{2}(x_n + \alpha/x_n)$ through the algebra of limits (limits respect the algebraic operations; 5.2.1)—legitimate because $L \neq 0$—yields $L = \frac{1}{2}(L + \alpha/L)$, whence $2L^2 = L^2 + \alpha$, so $L^2 = \alpha$. As $L > 0$, $L = \sqrt{\alpha}$. Hence $\lim x_n = \sqrt{\alpha}$.

Part (b). Put $\varepsilon_n = x_n - \sqrt{\alpha}$, so $\varepsilon_n > 0$ by part (a). The identity computed above is exactly

$$\varepsilon_{n+1} = x_{n+1} - \sqrt{\alpha} = \frac{(x_n - \sqrt{\alpha})^2}{2x_n} = \frac{\varepsilon_n^2}{2x_n}.$$

Since $x_n > \sqrt{\alpha}$, the denominator obeys $2x_n > 2\sqrt{\alpha}$, and as $\varepsilon_n^2 > 0$ this gives

$$\varepsilon_{n+1} = \frac{\varepsilon_n^2}{2x_n} < \frac{\varepsilon_n^2}{2\sqrt{\alpha}}.$$

Write $\beta = 2\sqrt{\alpha} > 0$. Dividing the last inequality by β rewrites it as

$$\frac{\varepsilon_{n+1}}{\beta} < \Big(\frac{\varepsilon_n}{\beta}\Big)^2,$$

since $\varepsilon_n^2/(2\sqrt{\alpha}) = \beta\,(\varepsilon_n/\beta)^2$. Set $t_n = \varepsilon_n/\beta > 0$, so $t_{n+1} < t_n^2$. We claim $t_{n+1} < t_1^{2^n}$ for $n \geq 1$, by induction (1.1.3). The base case $n = 1$ is $t_2 < t_1^2 = t_1^{2^1}$. If $t_{n+1} < t_1^{2^n}$ for some $n \geq 1$, then, both sides being positive, squaring preserves the inequality, and

$$t_{n+2} < t_{n+1}^2 < \big(t_1^{2^n}\big)^2 = t_1^{2^{n+1}}.$$

This is the claim at $n+1$. Multiplying $t_{n+1} < t_1^{2^n}$ by β restores

$$\varepsilon_{n+1} < \beta\Big(\frac{\varepsilon_1}{\beta}\Big)^{2^n} \qquad (n = 1, 2, 3, \dots).$$

Part (c). Take $\alpha = 3$ and $x_1 = 2$, so $\beta = 2\sqrt{3}$ and $\varepsilon_1 = x_1 - \sqrt{3} = 2 - \sqrt{3}$. The inequality $\varepsilon_1/\beta < \frac{1}{10}$ is, after clearing the positive denominator $\beta = 2\sqrt{3}$,

$$10(2 - \sqrt{3}) < 2\sqrt{3} \iff 20 < 12\sqrt{3} \iff \sqrt{3} > \tfrac{5}{3} \iff 3 > \tfrac{25}{9},$$

and $3 = \frac{27}{9} > \frac{25}{9}$ holds, so $\varepsilon_1/\beta < \frac{1}{10}$.

Now use part (b). Because $0 < \varepsilon_1/\beta < \frac{1}{10}$ and $2^n \geq 1$, raising to the power 2^n preserves the inequality, $(\varepsilon_1/\beta)^{2^n} < (\frac{1}{10})^{2^n} = 10^{-2^n}$, so

$$\varepsilon_{n+1} < \beta\Big(\frac{\varepsilon_1}{\beta}\Big)^{2^n} < \beta \cdot 10^{-2^n} = 2\sqrt{3} \cdot 10^{-2^n}.$$

Since $2\sqrt{3} = \sqrt{12} < \sqrt{16} = 4$, we have $\beta < 4$. Taking $n = 4$ gives $2^4 = 16$ and

$$\varepsilon_5 < \beta \cdot 10^{-16} < 4 \cdot 10^{-16};$$

taking $n = 5$ gives $2^5 = 32$ and

$$\varepsilon_6 < \beta \cdot 10^{-32} < 4 \cdot 10^{-32}.$$

■

Reflections.

The decisive word in part (a) is *decreases monotonically*: a sequence asked to be monotone and to have a limit is the exact shape the monotone convergence theorem was built for (a bounded monotone sequence converges; 6.2.2), so before touching the algebra one should aim to show the terms fall while staying above a floor. That theorem is the only tool in our notes that delivers a limit from a recursion you cannot solve in closed form, because it never asks you to name the limit in advance—it certifies existence from shape alone (6.2.1). What makes the floor obvious is the form of the recursion: $\frac{1}{2}(x+\alpha/x)$ is an average of x and α/x, two numbers whose product is α, so the average can never dip below their geometric mean $\sqrt{\alpha}$—the statement quietly hands you the lower bound by choosing $x_1 > \sqrt{\alpha}$.

It is worth believing this on the intended case before proving it. With $\alpha = 3$, $x_1 = 2$ the iteration reads $x_2 = \frac{1}{2}(2 + \frac{3}{2}) = 1.75$, then $x_3 = \frac{1}{2}(1.75 + 3/1.75) \approx 1.732143$, and $\sqrt{3} \approx 1.7320508$; the terms drop toward $\sqrt{3}$ from above and the agreement leaps from one decimal to several in a single step. That last observation—the explosive accuracy—is the whole point of parts (b) and (c).

The proof of (a) runs as a chain of inductions feeding the convergence theorem, each link resting on the one before:

(1) positivity of every x_n comes first by induction (0.6.2), since it is what licenses dividing by x_n and what keeps the later square-root manipulations honest;

(2) the single identity $x_{n+1} - \sqrt{\alpha} = (x_n - \sqrt{\alpha})^2/(2x_n)$, got by completing the square in the numerator, does double duty: a square over a positive number is ≥ 0, so $x_n > \sqrt{\alpha}$ propagates—this is the lower bound, proved by induction rather than quoted;

(3) monotonicity is then a one-line sign check, $x_n - x_{n+1} = (x_n^2 - \alpha)/(2x_n) > 0$, which is positive precisely because step (2) put x_n above $\sqrt{\alpha}$ (6.2.1);

(4) bounded below and decreasing, the sequence converges to some L (monotone convergence; 6.2.2), and the order $x_n > \sqrt{\alpha}$ passes to $L \geq \sqrt{\alpha} > 0$ in the limit (6.3.1);

(5) identifying L uses the standard fixed-point move: the limit of x_{n+1} equals the limit of x_n, so applying the algebra of limits (limits respect the operations; 5.2.1) to both sides of the recursion gives $L = \frac{1}{2}(L + \alpha/L)$, hence $L^2 = \alpha$ and $L = \sqrt{\alpha}$.

Step (5) is where rigour is easy to lose, and two guards are load-bearing. The algebra of limits may be applied to α/x_n only because $L \neq 0$, which is why step (4) bothered to record $L \geq \sqrt{\alpha} > 0$ rather than the weaker $L \geq \sqrt{\alpha}$; and the equation $L^2 = \alpha$ has two roots, so the sign $L > 0$ is needed to select $\sqrt{\alpha}$ over $-\sqrt{\alpha}$. The hypothesis $x_1 > \sqrt{\alpha}$ also earns its keep beyond seeding the bound: starting *at* $\sqrt{\alpha}$ would make the sequence constant, and starting below it (but positive) lands the very next term above $\sqrt{\alpha}$, so the strict inequality is what guarantees a genuine, strictly decreasing approach rather than a degenerate one.

Part (b) is the reason the method is celebrated, and its clue is the substitution $\varepsilon_n = x_n - \sqrt{\alpha}$: naming the error turns the recursion into a statement about how the error shrinks, and the identity from step (2) is already exactly $\varepsilon_{n+1} = \varepsilon_n^2/(2x_n)$. The error at each stage is the *square* of the previous error, scaled—this is quadratic convergence, and once the ratio ε_n/β drops below 1 the exponents 2^n make it plummet. The displayed bound is then a clean induction (1.1.3) on $t_n = \varepsilon_n/\beta$, where $t_{n+1} < t_n^2$ unwinds to $t_{n+1} < t_1^{2^n}$; squaring is order-preserving only because every $t_n > 0$, which is again the positivity of the errors from part (a).

Part (c) is purely a matter of feeding numbers through that bound, and the only subtlety is doing the two estimates without a calculator: $\varepsilon_1/\beta < \frac{1}{10}$ reduces by clearing denominators to $3 > \frac{25}{9}$, and $\beta = 2\sqrt{3} = \sqrt{12} < 4$ since $12 < 16$. With $\varepsilon_1/\beta < \frac{1}{10}$ the bound gives $\varepsilon_{n+1} < \beta \cdot 10^{-2^n}$, and reading off $n = 4$ and $n = 5$ produces the double-exponential decay 10^{-16} then 10^{-32}—each step roughly doubling the count of correct digits, the hallmark Rudin is advertising. The transferable move is the whole arc: when a recursion has a fixed point you cannot solve for, prove monotonicity and a bound to get convergence for free (monotone convergence), then pass to the limit in the recursion to identify it, and finally track the *error* rather than the term to read off the speed—an error that recurs as its own square is the signature of the Newton-type iteration this problem is.

Problem 3.17 *Rudin, Chapter 3, Exercise 17*

Fix $\alpha > 1$. Take $x_1 > \sqrt{\alpha}$, and define

$$x_{n+1} = \frac{\alpha + x_n}{1 + x_n} = x_n + \frac{\alpha - x_n^2}{1 + x_n}.$$

(a) Prove that $x_1 > x_3 > x_5 > \cdots$.

(b) Prove that $x_2 < x_4 < x_6 < \cdots$.

(c) Prove that $\lim x_n = \sqrt{\alpha}$.

(d) Compare the rapidity of convergence of this process with the one described in Exercise 16.

Solution.

Write $s = \sqrt{\alpha}$, so $s > 1$ and $s^2 = \alpha$. Two facts about the iteration come first.

Every term is positive. The start $x_1 > s > 0$ is positive, and if $x_n > 0$ then both $\alpha + x_n > 0$ and $1 + x_n > 0$, so $x_{n+1} = (\alpha + x_n)/(1 + x_n) > 0$. By induction $x_n > 0$ for all n.

The error obeys a single linear recursion. Put $\varepsilon_n = x_n - s$. Subtracting s from the recursion and clearing the denominator,

$$\begin{aligned}\varepsilon_{n+1} = x_{n+1} - s = \frac{\alpha + x_n}{1 + x_n} - s &= \frac{(\alpha - s) + (1 - s)x_n}{1 + x_n} \\ &= \frac{s(s-1) - (s-1)x_n}{1 + x_n} = -\,\frac{s-1}{1 + x_n}\,\varepsilon_n,\end{aligned}$$

where the last step used $\alpha - s = s(s-1)$ and factored out $-(s-1)$. Set

$$c_n = \frac{s-1}{1 + x_n}, \qquad \text{so} \qquad \varepsilon_{n+1} = -\,c_n\,\varepsilon_n.$$

Since $s > 1$ and $x_n > 0$, each $c_n > 0$. The sign and the size of the error are read straight off $\varepsilon_{n+1} = -c_n\varepsilon_n$:

$$\operatorname{sign}\varepsilon_{n+1} = -\operatorname{sign}\varepsilon_n, \qquad |\varepsilon_{n+1}| = c_n\,|\varepsilon_n|.$$

Because $\varepsilon_1 = x_1 - s > 0$, the signs alternate: $\varepsilon_n > 0$ for odd n and $\varepsilon_n < 0$ for even n. Equivalently, $x_n > s$ for odd n and $x_n < s$ for even n.

A clean two-step contraction governs the magnitudes. Multiplying out the denominators,

$$(1 + x_n)(1 + x_{n+1}) = (1 + x_n) + (\alpha + x_n) = 1 + \alpha + 2x_n,$$

using $1 + x_{n+1} = 1 + \dfrac{\alpha + x_n}{1 + x_n} = \dfrac{(1 + x_n) + (\alpha + x_n)}{1 + x_n}$. Hence

$$|\varepsilon_{n+2}| = c_{n+1}c_n\,|\varepsilon_n|, \qquad c_{n+1}c_n = \frac{(s-1)^2}{(1 + x_n)(1 + x_{n+1})} = \frac{(s-1)^2}{1 + \alpha + 2x_n}.$$

Since $x_n > 0$, the denominator exceeds $1 + \alpha = 1 + s^2$, and $(s-1)^2 = s^2 - 2s + 1 < s^2 + 1$ because $s > 0$; therefore

$$c_{n+1}c_n \;<\; \frac{(s-1)^2}{1 + s^2} \;=:\; r \;<\; 1,$$

a constant independent of n. Each two consecutive steps shrink the error by the fixed factor $r < 1$.

(a) and (b). For odd n, both ε_n and ε_{n+2} are positive while $|\varepsilon_{n+2}| = c_{n+1}c_n|\varepsilon_n| < |\varepsilon_n|$, so $0 < \varepsilon_{n+2} < \varepsilon_n$, i.e. $x_{n+2} < x_n$; this is $x_1 > x_3 > x_5 > \cdots$. For even n, both errors are

negative with $|\varepsilon_{n+2}| < |\varepsilon_n|$, so $\varepsilon_n < \varepsilon_{n+2} < 0$, i.e. $x_n < x_{n+2}$; this is $x_2 < x_4 < x_6 < \cdots$.

(c) Iterating $|\varepsilon_{n+2}| < r\,|\varepsilon_n|$ down to the first term of each parity gives, for $k \geq 1$,

$$|\varepsilon_{2k+1}| < r^k\,|\varepsilon_1|, \qquad |\varepsilon_{2k}| < r^{k-1}\,|\varepsilon_2|.$$

With $0 < r < 1$ one has $r^k \to 0$ (a standard limit, 6.5.1), so $|\varepsilon_n| \to 0$ along both parities, hence along the whole sequence: given any $\delta > 0$, all but finitely many odd terms and all but finitely many even terms satisfy $|\varepsilon_n| < \delta$, so all but finitely many terms do. Thus $\varepsilon_n = x_n - s \to 0$, that is $\lim x_n = s = \sqrt{\alpha}$.

The same conclusion follows from monotone convergence, which is worth recording. By (a) and (b) the odd-indexed subsequence decreases and is bounded below by s, while the even-indexed subsequence increases and is bounded above by s; each is bounded and monotone, so each converges (a bounded monotone sequence converges; 6.2.2). A limit L of either satisfies $L = (\alpha + L)/(1 + L)$, i.e. $L^2 = \alpha$, and $L \geq 0$ as a limit of positive terms (6.3.1), so $L = s$; both subsequences reach the same value s, and together they exhaust the indices.

(d) The errors here contract *linearly* (geometrically): from $\varepsilon_{n+1} = -c_n\varepsilon_n$ with $c_n = \dfrac{s-1}{1+x_n} \to \dfrac{s-1}{1+s}$ as $x_n \to s$, the ratio $|\varepsilon_{n+1}|/|\varepsilon_n|$ tends to the fixed constant $\dfrac{s-1}{1+s} \in (0,1)$, so each step multiplies the error by roughly this factor and the number of correct digits grows by a constant amount per step. The process of Problem 3.16 (Exercise 16) instead has $\varepsilon_{n+1} = \varepsilon_n^2/(2x_n) \sim \varepsilon_n^2/(2s)$, a *quadratic* contraction in which the error is squared at each step and the number of correct digits roughly doubles. Squaring beats any fixed multiplier once the error is small, so that process converges far more rapidly; the present iteration is the slower, geometric one.

■

REFLECTIONS.

The statement hands you a recursively defined real sequence and asks three things about it—two subsequences are monotone, and the whole sequence converges to $\sqrt{\alpha}$—so the words to seize on are $x_1 > x_3 > \cdots$, $x_2 < x_4 < \cdots$, and $\lim x_n = \sqrt{\alpha}$. The first two are monotone-subsequence claims, which name two bounded monotone subsequences and so summon monotone convergence (bounded monotone sequences converge; 6.2.2) as one honest route to the third; the value $\sqrt{\alpha}$ is exactly the positive solution of $x = (\alpha+x)/(1+x)$, the equation a limit would have to satisfy, so the target is the recursion's fixed point. That a single sequence splits into a decreasing odd part *and* an increasing even part is itself the loudest clue: the terms are not marching one way but *oscillating* across $\sqrt{\alpha}$, jumping to the far side each step while landing closer. The object that records oscillation cleanly is the error $\varepsilon_n = x_n - \sqrt{\alpha}$, and the whole solution is the discovery that this error satisfies one tidy linear rule.

Believe the oscillation before proving it. With $\alpha = 3$ and $x_1 = 2$ the iterates run $2,\ \frac{5}{3},\ \frac{7}{4},\ \frac{19}{11}, \ldots$, i.e. $2.000,\ 1.667,\ 1.750,\ 1.727, \ldots$, bouncing above and below $\sqrt{3} \approx 1.732$ and squeezing in from both sides—odd terms stepping down, even terms stepping up, both funnelling to $\sqrt{3}$. That picture of two pincers closing on the fixed point is the entire content of (a)–(c).

The engine is the identity $\varepsilon_{n+1} = -c_n \varepsilon_n$ with $c_n = (\sqrt{\alpha} - 1)/(1 + x_n) > 0$, and everything follows by reading off its sign and its size:

(1) subtracting $\sqrt{\alpha}$ from the recursion and using $\alpha - \sqrt{\alpha} = \sqrt{\alpha}(\sqrt{\alpha} - 1)$ collapses the difference into $\varepsilon_{n+1} = -\frac{\sqrt{\alpha}-1}{1+x_n}\varepsilon_n$; this algebraic factorisation is the crux—without it the three parts are unrelated, with it they are one statement;

(2) the minus sign forces the sign of ε_n to alternate, and since $\varepsilon_1 > 0$ (from $x_1 > \sqrt{\alpha}$) the odd errors are positive and the even ones negative, which is just "$x_n > \sqrt{\alpha}$ for odd n, $x_n < \sqrt{\alpha}$ for even n"—the bracketing the two monotone claims need;

(3) the factor c_n is positive, and multiplying out the denominators gives the clean two-step identity $c_{n+1}c_n = (\sqrt{\alpha}-1)^2/(1+\alpha+2x_n)$, bounded by the fixed $r = (\sqrt{\alpha}-1)^2/(1+\alpha) < 1$; so $|\varepsilon_{n+2}| < r|\varepsilon_n|$ and within each parity the magnitudes strictly shrink, which with the fixed sign is $x_1 > x_3 > \cdots$ and $x_2 < x_4 < \cdots$, parts (a) and (b);

(4) for (c) iterating $|\varepsilon_{n+2}| < r|\varepsilon_n|$ down each parity bounds $|\varepsilon_n|$ by a constant times $r^{\lfloor n/2 \rfloor}$, and $r^k \to 0$ since $0 < r < 1$ (a standard limit, 6.5.1), so the error vanishes along both parities and hence along the whole sequence—this is the quantitative route; equally, each subsequence is monotone (by (a),(b)) and bounded (odd terms above $\sqrt{\alpha}$, even below it), so monotone convergence (6.2.2) hands each a limit L with $L^2 = \alpha$ and $L \geq 0$ by order-preservation (6.3.1), forcing $L = \sqrt{\alpha}$;

(5) both parities reach the same $\sqrt{\alpha}$ and together exhaust the indices, so the full sequence converges to $\sqrt{\alpha}$.

The rigor turns on three points worth naming. Positivity of every x_n is not cosmetic: it keeps $1 + x_n > 0$, so $c_n > 0$ and the sign argument is valid, and it is what makes the limit nonnegative so that $L^2 = \alpha$ resolves to $+\sqrt{\alpha}$ rather than $-\sqrt{\alpha}$. The bracketing $x_n < \sqrt{\alpha}$ on even indices and $x_n > \sqrt{\alpha}$ on odd indices is what supplies the bounds the monotone-convergence theorem demands—monotone alone is not enough, an unbounded monotone sequence diverges. And the last step genuinely needs *both* subsequences to reach the same limit: knowing only the odd terms converge would leave the even terms unaccounted for, so the step that the odd and even indices exhaust $\mathbb{N}$ and share the limit is load-bearing, not bookkeeping. Nothing is circular—the value $\sqrt{\alpha}$ is forced out of the fixed-point equation, never assumed.

Part (d) is the payoff and the reason the error identity, not just a convergence verdict, was worth extracting. Here the error is multiplied by $c_n \to (\sqrt{\alpha}-1)/(1+\sqrt{\alpha})$, a fixed constant strictly between 0 and 1: convergence is *geometric*, adding a constant number of correct digits per step. The Newton-style iteration of Problem 3.16 squares the error each step, doubling the correct digits—*quadratic* convergence—which is incomparably faster once the error is small, since a square crushes a quantity below 1 far harder than any fixed multiplier. Reading the rate off the recursion for the error is the transferable move: for a fixed-point iteration, linearise near the fixed point and look at the multiplier on ε_n. A nonzero constant multiplier means linear (geometric) speed; a multiplier that itself vanishes with ε_n—an ε_n^2 law—means quadratic speed. The same $\varepsilon_{n+1} = -c_n\varepsilon_n$ that proved the convergence in (a)–(c) is precisely what measures its rapidity in (d).

Problem 3.18 *Rudin, Chapter 3, Exercise 18*

Replace the recursion formula of Exercise 16 by

$$x_{n+1} = \frac{p-1}{p}x_n + \frac{\alpha}{p}x_n^{-p+1}$$

where p is a fixed positive integer, and describe the behavior of the resulting sequences $\{x_n\}$.

SOLUTION.

Fix $\alpha > 0$ and a positive integer p, and let $g(x) = \dfrac{p-1}{p}\,x + \dfrac{\alpha}{p}\,x^{1-p}$ be the map driving the recursion, defined for $x > 0$. Write $L = \alpha^{1/p} > 0$ for the positive pth root of α, so that $L^p = \alpha$. The verdict: for $p \geq 2$ and any starting value $x_1 > 0$, the sequence satisfies $x_n \geq L$ for all $n \geq 2$, decreases monotonically from the second term on, and converges to $L = \alpha^{1/p}$; the degenerate case $p = 1$ is constant.

The case $p = 1$ is immediate: the formula reads $x_{n+1} = 0 \cdot x_n + \alpha\, x_n^0 = \alpha$, so $x_n = \alpha = \alpha^{1/1}$ for every $n \geq 2$, a constant sequence at its root. Assume $p \geq 2$ from here on.

First, a lower bound: $g(x) \geq L$ for every $x > 0$, with equality only at $x = L$. Substitute $x = tL$ with $t > 0$; using $\alpha = L^p$,

$$g(tL) = \frac{p-1}{p}\,tL + \frac{L^p}{p}\,(tL)^{1-p} = \frac{L}{p}\big[(p-1)t + t^{1-p}\big],$$

so $g(x) \geq L$ is equivalent to $(p-1)t + t^{1-p} \geq p$. Multiplying this by $p\,t^{p-1} > 0$ turns it into the polynomial inequality $(p-1)t^p - p\,t^{p-1} + 1 \geq 0$, and a direct expansion gives the factorization

$$(p-1)t^p - p\,t^{p-1} + 1 = (t-1)^2 \sum_{k=1}^{p-1} k\,t^{\,k-1}.$$

The second factor has only nonnegative coefficients, so it is positive for $t > 0$, while $(t-1)^2 \geq 0$; hence the product is ≥ 0, vanishing exactly when $t = 1$, i.e. $x = L$. Therefore

$g(x) \geq L$ for all $x > 0$, and the map sends $(0, \infty)$ into $[L, \infty)$.

Second, a decrease above the root: $g(x) \leq x$ whenever $x \geq L$. The difference is

$$x - g(x) = \frac{1}{p}\, x - \frac{\alpha}{p}\, x^{1-p} = \frac{1}{p}\, x^{1-p}(x^p - \alpha),$$

and for $x \geq L > 0$ one has $x^p \geq L^p = \alpha$, so $x - g(x) \geq 0$.

Now describe the sequence. Whatever $x_1 > 0$ is chosen, $x_2 = g(x_1) \geq L$ by the lower bound, and the same bound applied at each step keeps $x_n = g(x_{n-1}) \geq L$ for every $n \geq 2$. For $n \geq 2$ the term $x_n \geq L$ then feeds the decrease estimate, giving $x_{n+1} = g(x_n) \leq x_n$. So $\{x_n\}_{n \geq 2}$ is monotonically decreasing and bounded below by L, hence convergent by monotone convergence (a bounded monotone sequence converges; 6.2.2). Call the limit ℓ; since every $x_n \geq L$ for $n \geq 2$, order is preserved in the limit (6.3.1) and $\ell \geq L > 0$.

It remains to identify ℓ. Because $\ell > 0$, the algebra of limits (limits respect the algebraic operations; 5.2.1) applies to the recursion: $x_n^{p-1} \to \ell^{p-1}$ as a finite product of convergent sequences, and as $\ell^{p-1} \neq 0$ the reciprocals give $x_n^{1-p} \to \ell^{1-p}$. The shifted sequence $\{x_{n+1}\}$ has the same limit ℓ, so letting $n \to \infty$ in $x_{n+1} = \dfrac{p-1}{p} x_n + \dfrac{\alpha}{p} x_n^{1-p}$ yields

$$\ell = \frac{p-1}{p}\, \ell + \frac{\alpha}{p}\, \ell^{1-p}.$$

Multiplying through by $\dfrac{p}{\ell^{1-p}} = p\, \ell^{p-1}$ and simplifying gives $\ell^p = \alpha$, so $\ell = \alpha^{1/p} = L$. Thus the recursion is Newton's method for the root of $x^p - \alpha$: from any positive seed it converges monotonically (after the first step) to $\alpha^{1/p}$, recovering Exercise 16 at $p = 2$.

■

Reflections.

The statement is one of a family—Exercises 16, 17, 18—each handing you a recursion $x_{n+1} = g(x_n)$ and asking for its long-run behavior, and the phrase *describe the behavior of the resulting sequences* is the tell. A sequence defined by iterating a single map has no closed form to manipulate, so the only handle our notes offer is monotone convergence (6.2.2): show the sequence is monotone and bounded, harvest a limit, then pin the limit down. That two-move plan—first that the limit *exists*, separately what it *is*—is the spine of every iteration problem, and recognizing the recursion as the cue for the monotone convergence theorem is the first thing to read off.

Believe the answer before proving it. The formula is Newton's method applied to $f(x) = x^p - \alpha$: the update $x_{n+1} = x_n - f(x_n)/f'(x_n)$ works out to exactly $\frac{p-1}{p} x_n + \frac{\alpha}{p} x_n^{1-p}$, and a fixed point $\ell = g(\ell)$ forces $\ell^p = \alpha$. So the target must be $\alpha^{1/p}$, the pth-root generalization of the square-root algorithm in Problem 3.16, which is the case $p = 2$. Naming $L = \alpha^{1/p}$

up front converts every vague "it settles down" into a sharp inequality against a fixed number.

The argument then assembles from four moves, each resting on a specific fact:

(1) the lower bound $g(x) \geq L$ for all $x > 0$, which after the substitution $x = tL$ collapses to the single inequality $(p-1)t + t^{1-p} \geq p$, proved with no calculus by clearing denominators and factoring $(p-1)t^p - p\,t^{p-1} + 1 = (t-1)^2 \sum_{k=1}^{p-1} k\,t^{k-1}$—a perfect square times a polynomial with nonnegative coefficients, hence ≥ 0 on $(0, \infty)$;

(2) the decrease $g(x) \leq x$ for $x \geq L$, which is even cleaner, since $x - g(x) = \frac{1}{p}x^{1-p}(x^p - \alpha)$ is nonnegative precisely when $x^p \geq \alpha$, i.e. when $x \geq L$;

(3) combining the two, every x_n with $n \geq 2$ lies in $[L, \infty)$ and the sequence decreases there, so it is bounded-below and monotone and converges by monotone convergence (6.2.2), with limit $\ell \geq L > 0$ by order-preservation in the limit (6.3.1);

(4) passing to the limit in the recursion via the algebra of limits (the algebra of limits; 5.2.1)—legitimate because $\ell \neq 0$, so $x_n^{1-p} \to \ell^{1-p}$—gives $\ell^p = \alpha$, hence $\ell = \alpha^{1/p}$.

Reading where each hypothesis is spent shows why the description is shaped as it is. The lower bound is what guarantees the sequence enters $[L, \infty)$ *after one step regardless of the seed*, which is why the monotonicity is asserted only from $n \geq 2$: a seed $x_1 < L$ can overshoot upward on the first iterate, and the numerics bear this out, yet $x_2 \geq L$ is forced and the descent begins. This is a genuine difference from Exercise 16's framing, which assumes $x_1 > \sqrt{\alpha}$ to make the sequence decrease from the very start; here the lower bound does that work automatically. The decrease estimate in turn needs $x \geq L$ to make $x^p - \alpha \geq 0$—drop it and the sequence need not be monotone below the root. And the final limit step needs $\ell > 0$ to divide by ℓ^{1-p}; this is exactly what order-preservation supplies, since $\ell \geq L > 0$. Nothing is circular: existence of the limit is settled by monotone convergence before its value is computed, never the reverse.

The transferable move is the whole skeleton of an iteration problem. To analyze $x_{n+1} = g(x_n)$, locate the fixed point $\ell = g(\ell)$ first—it is the only candidate limit—then prove the iterates are trapped on one side of ℓ and monotone toward it, so monotone convergence hands you a limit that the algebra of limits forces to equal ℓ. The same template runs Exercise 16 (Problem 3.16) and, with a two-subsequence twist for the oscillation, Exercise 17 (Problem 3.17); the only craft that changes problem to problem is the inequality certifying boundedness, and here that craft was the telescoping factorization $(t-1)^2 \sum_k k\,t^{k-1}$, the algebraic heart of the weighted arithmetic–geometric mean inequality dressed for a single variable.

Problem 3.19 *Rudin, Chapter 3, Exercise 19*

Associate to each sequence $a = \{\alpha_n\}$, in which α_n is 0 or 2, the real number

$$x(a) = \sum_{n=1}^{\infty} \frac{\alpha_n}{3^n}.$$

Prove that the set of all $x(a)$ is precisely the Cantor set described in Sec. 2.44.

Solution.

The construction of Sec. 2.44 runs as follows. Set $E_0 = [0,1]$, and obtain E_m from E_{m-1} by deleting the open middle third of each of its intervals; then E_m is a union of 2^m closed intervals of length 3^{-m}, the sets are nested $E_0 \supseteq E_1 \supseteq \cdots$, and the Cantor set is

$$P = \bigcap_{m=0}^{\infty} E_m.$$

Write $C = \{x(a) : \alpha_n \in \{0,2\} \text{ for all } n\}$ for the set of values in question. The series defining $x(a)$ converges by comparison with the geometric series $\sum 2 \cdot 3^{-n}$ (each term lies in $[0, 2 \cdot 3^{-n}]$; 6.6.8, 6.6.7), so $x(a) \in [0,1]$ is well defined. We prove $C = P$ by two inclusions.

The key is to read off, from the intervals of E_m, which digits a point may carry. A direct induction shows that E_m is exactly the set of $x \in [0,1]$ admitting a representation

$$x = \sum_{n=1}^{m} \frac{\beta_n}{3^n} + r, \qquad \beta_n \in \{0,2\}, \quad 0 \le r \le 3^{-m}, \tag{$*$}$$

that is, a left endpoint $\sum_{n \le m} \beta_n 3^{-n}$ with all digits 0 or 2, plus a remainder no larger than one interval length. At $m = 0$ this reads $0 \le x \le 1$, which is E_0. Assume it for m. A stage-m interval $\left[t,\, t + 3^{-m}\right]$ with $t = \sum_{n \le m} \beta_n 3^{-n}$ has its open middle third $\left(t + 3^{-m-1},\, t + 2 \cdot 3^{-m-1}\right)$ deleted, leaving the two closed thirds

$$\left[t,\ t + 3^{-m-1}\right] \quad \text{and} \quad \left[t + 2 \cdot 3^{-m-1},\ t + 3^{-m}\right],$$

i.e. $\left[t + \frac{0}{3^{m+1}},\, t + \frac{0}{3^{m+1}} + 3^{-m-1}\right]$ and $\left[t + \frac{2}{3^{m+1}},\, t + \frac{2}{3^{m+1}} + 3^{-m-1}\right]$. So the stage-$(m+1)$ intervals are exactly those with left endpoint $t + \beta_{m+1} 3^{-m-1}$, $\beta_{m+1} \in \{0,2\}$, and length 3^{-m-1}, which is $(*)$ at $m+1$.

For the inclusion $C \subseteq P$, fix a with all $\alpha_n \in \{0,2\}$ and any $m \ge 0$, and split the series at m,

$$x(a) = \underbrace{\sum_{n=1}^{m} \frac{\alpha_n}{3^n}}_{=:t} + \underbrace{\sum_{n=m+1}^{\infty} \frac{\alpha_n}{3^n}}_{=:r},$$

the head t has digits in $\{0,2\}$, and the tail is bounded by the geometric estimate

$$0 \le r \le \sum_{n=m+1}^{\infty} \frac{2}{3^n} = \frac{2 \cdot 3^{-(m+1)}}{1 - 1/3} = 3^{-m}$$

(6.6.7). By $(*)$, $x(a) \in E_m$. As m was arbitrary, $x(a) \in \bigcap_m E_m = P$.

For the reverse inclusion $P \subseteq C$, fix $x \in P$, so $x \in E_m$ for every m. For each m choose, by $(*)$, digits $\beta_1^{(m)}, \ldots, \beta_m^{(m)} \in \{0,2\}$ with $x - \sum_{n \le m} \beta_n^{(m)} 3^{-n} \in [0, 3^{-m}]$. The stage-$m$ interval containing x is unique (distinct stage-m intervals are disjoint, being separated by the deleted middle thirds), so its left endpoint, hence each digit $\beta_n^{(m)}$, is determined by x; and a stage-$(m+1)$ interval lies inside a stage-m interval, so passing from m to $m+1$ keeps the earlier digits and appends one. The digits are therefore consistent: there is a single sequence $\alpha_n \in \{0,2\}$ with $\beta_n^{(m)} = \alpha_n$ for all $n \le m$. For each m,

$$\Big| x - \sum_{n=1}^{m} \frac{\alpha_n}{3^n} \Big| \le 3^{-m},$$

and $3^{-m} \to 0$ (Archimedean; 6.5.1), so the partial sums of $\sum \alpha_n 3^{-n}$ converge to x. By uniqueness of limits (a limit is unique) the sum equals x, that is $x = x(a) \in C$.

The two inclusions give $C = P$.

■

REFLECTIONS.

The statement hands you an infinite series and asks you to identify its range of values with a set built by a completely different recipe—repeatedly deleting middle thirds. The clue is the base: every denominator is a power of 3 and every numerator is a ternary digit (0 or 2, the two digits that survive when the middle third, digit 1, is thrown away). That is the tell that $x(a)$ is nothing but the base-3 expansion $0.\alpha_1\alpha_2\alpha_3\ldots$ with the digit 1 forbidden, and the Cantor set is exactly the points whose ternary expansion can avoid the digit 1. So the problem is the bridge between a *geometric* object (nested intervals) and an *arithmetic* one (digit strings), and the whole task is to make "$x \in E_m$" and "x's first m ternary digits avoid 1" literally the same statement.

Before any inclusion, the series itself must be shown to name a real number, and that is the first place the sequence machinery enters: each term lies in $[0, 2 \cdot 3^{-n}]$, so the series is dominated by the geometric series $\sum 2 \cdot 3^{-n}$ and converges by comparison (6.6.8, 6.6.7). The same geometric sum, $\sum_{n>m} 2 \cdot 3^{-n} = 3^{-m}$, is the quantitative heart of everything that follows: it says the entire tail of the expansion past digit m can shift the value by at most one stage-m interval length. Believe the correspondence on the smallest case first—$\alpha_1 = 0$ keeps you in $[0, \frac{1}{3}]$ and $\alpha_1 = 2$ in $[\frac{2}{3}, 1]$, which are exactly the two intervals of E_1, while $\alpha_1 = 1$ would land you in the forbidden middle $(\frac{1}{3}, \frac{2}{3})$.

The engine is the single inductive description $(*)$: E_m is the set of x that can be written as a length-m ternary head with digits in $\{0,2\}$ plus a remainder in $[0,3^{-m}]$. The induction is where the geometry becomes arithmetic, and it runs in clean steps:

(1) the base case is $E_0=[0,1]$, which is $(*)$ with an empty head and remainder $r=x$;

(2) the inductive step takes one stage-m interval $[t,t+3^{-m}]$ with t a $\{0,2\}$-head, deletes its open middle third, and observes that the two surviving closed thirds have left endpoints $t+0\cdot 3^{-m-1}$ and $t+2\cdot 3^{-m-1}$—appending precisely a new digit 0 or 2, which is $(*)$ one level deeper;

(3) with $(*)$ in hand, $C\subseteq P$ is the forward reading: split $x(a)$ into its first m digits plus a tail, bound the tail by the geometric 3^{-m} (6.6.7), and $(*)$ places $x(a)$ in every E_m, hence in the intersection P;

(4) $P\subseteq C$ is the reverse reading: a point of every E_m chooses a digit at each stage, those choices are forced and consistent because the stage-$(m+1)$ interval sits inside the stage-m interval, and the resulting partial sums approximate x to within 3^{-m}, so they converge to x by the standard limit $3^{-m}\to 0$ (6.5.1).

The rigor turns on two points that are easy to wave past. First, in $P\subseteq C$ the digits must be *consistent* across stages, not merely available at each: this is what nestedness of the intervals buys you—because each stage-$(m+1)$ interval lies in a unique stage-m interval, the choice at level $m+1$ cannot overwrite the earlier digits, so the per-stage digits assemble into one genuine sequence $\{\alpha_n\}$. Drop nestedness and you would have a different finite digit string at each stage with no single limit. Second, the passage "partial sums within 3^{-m} of x, and $3^{-m}\to 0$, therefore the sum is x" is exactly convergence of the series to x (5.1.1), made legitimate by the standard limit $3^{-m}\to 0$ (Archimedean, 6.5.1) and pinned down by uniqueness of the limit (a limit is unique)—the series cannot converge to its own value and also to something else. Nothing is circular: $(*)$ is proved by an induction that never mentions C or the series, and only afterward is it read in both directions.

One subtlety the digit-1 ban quietly handles is the ambiguity of expansions. A few reals carry two ternary names—$\frac{1}{3}=0.1000\cdots=0.0222\ldots$—and the Cantor set keeps such a point precisely because it admits the all-$\{0,2\}$ name $0.0222\ldots$; forbidding the digit 1 is what selects the surviving representation. This is the ternary twin of the decimal $0.4999\cdots=0.5000\ldots$ tail-dodging that made Cantor's diagonal argument honest in HW2 Problem 4, and it is why the endpoints of every deleted interval stay in P.

The transferable move is to recognise a digit expansion hiding inside a geometric construction: whenever a set is built by keeping a fixed subset of positions at each scale—ternary digits in $\{0,2\}$ here, the digits 4 and 7 in the digits-4-and-7 set of HW2 Problem 17—encode membership as a constraint on base-b digits, and a nested-interval geometry collapses into

an arithmetic statement about sequences. Once the correspondence is set up, the convergence of the defining series and the standard limit $b^{-m} \to 0$ do all the analytic work, and the topology of the Cantor set (it is uncountable, perfect, and nowhere dense) becomes a statement about $\{0, 2\}$-valued sequences, which is the cleanest way to see why P is as large as $\mathbb{R}$ yet contains no interval.

Problem 3.20 *Rudin, Chapter 3, Exercise 20*

Suppose $\{p_n\}$ is a Cauchy sequence in a metric space X, and some subsequence $\{p_{n_i}\}$ converges to a point $p \in X$. Prove that the full sequence $\{p_n\}$ converges to p.

SOLUTION.

Write d for the metric on X. A sequence is Cauchy when its terms eventually crowd together (5.4.1), and $\{p_n\}$ converges to p when every tolerance is met from some index on (5.1.1); the goal is to produce, for each $\varepsilon > 0$, an index past which $d(p_n, p) < \varepsilon$.

Fix $\varepsilon > 0$. Since $\{p_n\}$ is Cauchy, there is an N with

$$d(p_m, p_n) < \tfrac{\varepsilon}{2} \qquad \text{for all } m, n \geq N.$$

Since the subsequence $\{p_{n_i}\}$ converges to p, there is an I with $d(p_{n_i}, p) < \frac{\varepsilon}{2}$ for all $i \geq I$. The indices of a subsequence strictly increase, so $n_i \geq i$ for every i (5.3.1); choose a single index i large enough that both $i \geq I$ and $n_i \geq N$, for instance $i = \max\{I, N\}$. Then $q := p_{n_i}$ satisfies $d(q, p) < \frac{\varepsilon}{2}$, and $n_i \geq N$.

Now take any $n \geq N$. Both p_n and $q = p_{n_i}$ carry indices at least N, so the Cauchy estimate applies to the pair, and the triangle inequality routes p_n to p through q:

$$d(p_n, p) \leq d(p_n, p_{n_i}) + d(p_{n_i}, p) < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

Thus $d(p_n, p) < \varepsilon$ for every $n \geq N$, and since $\varepsilon > 0$ was arbitrary, $p_n \to p$ (5.1.1).

■

REFLECTIONS.

The statement hands you two pieces of information about the *same* sequence and asks you to merge them. The first, *Cauchy*, says the terms cluster together without naming a target (5.4.1); the second, that a *subsequence converges to p*, names the only candidate target in sight (5.3.1, 5.3.3). Convergence of $\{p_n\}$ needs both a destination and a guarantee the whole tail arrives there (5.1.1): the subsequence supplies the destination p, and the Cauchy condition supplies the discipline that drags the rest of the tail along. So the proof is a meeting of the two hypotheses, and the natural meeting place is the triangle inequality, the one tool that lets a fact about p_n near q and a fact about q near p combine into a fact about p_n near p. This is exactly the quotable packaging "if a subsequence of a Cauchy

sequence converges to p, so does the sequence"; here that fact is not invoked but proved.

Believe it first on the picture. A Cauchy sequence is a swarm that contracts to a speck: past some index every term sits within $\varepsilon/2$ of every other. The convergent subsequence pins where that speck is—it is sitting on p. A swarm contracting onto one of its own members that happens to rest at p cannot settle anywhere but p; the only thing missing is a single subsequential term deep enough in the tail to act as the anchor, and that is what choosing i large achieves.

The argument is short, and each move is load-bearing:

(1) from the Cauchy property pull an N with $d(p_m, p_n) < \varepsilon/2$ for all $m, n \geq N$ (5.4.1)—the half-ε is chosen now so the two contributions sum to ε later;

(2) from the subsequence's convergence pull an I with $d(p_{n_i}, p) < \varepsilon/2$ for $i \geq I$ (5.1.1, all but finitely many terms lie in each ball);

(3) secure one anchor term $q = p_{n_i}$ that is *both* close to p and deep in the tail, using $n_i \geq i$ (subsequence indices strictly increase, 5.3.1) to force $n_i \geq N$ by taking $i \geq \max\{I, N\}$—this is the step that fuses the two indices into one usable witness;

(4) for any $n \geq N$, the triangle inequality $d(p_n, p) \leq d(p_n, q) + d(q, p) < \varepsilon/2 + \varepsilon/2$ closes the gap, the first term controlled because both p_n and q have indices $\geq N$, the second because q is a far-enough subsequential term.

Each hypothesis earns its place, which is the way to see the result is sharp. Drop "Cauchy" and a convergent subsequence says nothing about the rest: $p_n = (-1)^n$ has the even subsequence converging to 1, yet $\{p_n\}$ diverges, because there is no contraction to spread that limit to the odd terms. Drop the converging subsequence and a Cauchy sequence in an incomplete space may have no limit at all to inherit—this is precisely why a Cauchy sequence *with* a convergent subsequence is the hypothesis that closes the gap, the mechanism behind completeness of $\mathbb{R}^k$ and of compact spaces (5.5.3). The reasoning is not circular: nowhere is $p_n \to p$ assumed; p is read off the subsequence, and the limit of the whole sequence is built from the half-ε estimates, not presupposed.

The transferable move is the triangle-inequality relay with a single well-chosen waypoint: to show p_n is near p when you only know p_n is near *its neighbours* and that *one* neighbour is near p, route through that neighbour and split the tolerance in half. The one subtlety worth banking is step (3)—a subsequence reaches arbitrarily far out ($n_i \geq i$), so it can always hand you an anchor lying as deep in the tail as the Cauchy index demands. This same fusion of "Cauchy" with "one convergent subsequence" is what powers the completeness theorems: a compact space forces a convergent subsequence out of any sequence (every sequence in a compact space has a convergent subsequence, 5.3.3), and this exercise then

upgrades a Cauchy sequence to a convergent one, so the space is complete (5.5.3).

Problem 3.21 *Rudin, Chapter 3, Exercise 21*

Prove the following analogue of Theorem 3.10(b): If $\{E_n\}$ is a sequence of closed nonempty and bounded sets in a complete metric space X, if $E_n \supset E_{n+1}$, and if

$$\lim_{n\to\infty} \operatorname{diam} E_n = 0,$$

then $\bigcap_1^\infty E_n$ consists of exactly one point.

SOLUTION.

Work in the metric space (X, d), and write $\operatorname{diam} E = \sup\{d(x, y) : x, y \in E\}$ for the diameter (5.5.1). The claim has two halves: the intersection $E = \bigcap_{n=1}^\infty E_n$ is nonempty, and it has at most one point.

First, E has at most one point. If $x, y \in E$, then $x, y \in E_n$ for every n, so $0 \le d(x, y) \le \operatorname{diam} E_n$ for every n. Letting $n \to \infty$ forces $d(x, y) \le \lim_n \operatorname{diam} E_n = 0$, hence $d(x, y) = 0$ and $x = y$—and this is the only place the diameter hypothesis is used.

It remains to show E is nonempty. Each E_n is nonempty, so choose a point $x_n \in E_n$ for every n. The sequence $\{x_n\}$ is Cauchy. Given $\varepsilon > 0$, the hypothesis $\operatorname{diam} E_n \to 0$ supplies an N with $\operatorname{diam} E_N < \varepsilon$. For any $m, n \ge N$ the nesting $E_n \subseteq E_N$ and $E_m \subseteq E_N$ (which follows from $E_k \supseteq E_{k+1}$) places both x_m and x_n in E_N, so

$$d(x_m, x_n) \le \operatorname{diam} E_N < \varepsilon.$$

Thus $\{x_n\}$ is Cauchy (5.4.1). Since X is complete (5.4.3), the sequence converges to some $x \in X$.

This limit lies in every E_m. Fix m. For each $n \ge m$ the nesting gives $x_n \in E_n \subseteq E_m$, so the tail $\{x_n\}_{n\ge m}$ lies entirely in E_m and still converges to x. A convergent sequence drawn from a set has its limit in that set's closure; as E_m is closed (3.2.4), the limit satisfies $x \in E_m$. Concretely, were $x \notin E_m$, then x would be neither a point nor a cluster point of E_m, so some ball $B_r(x)$ would miss E_m (3.2.5)—yet $x_n \to x$ puts $x_n \in B_r(x)$ for all large n (all but finitely many terms enter every ball about the limit), and these $x_n \in E_m$, a contradiction. Hence $x \in E_m$.

Since m was arbitrary, $x \in \bigcap_{n=1}^\infty E_n = E$, so E is nonempty. Combined with the first half, E consists of exactly one point.

■

REFLECTIONS.

The statement announces itself as an "analogue of Theorem 3.10(b)," and the first move is to recall what that theorem says: a nested sequence of nonempty *compact* sets has nonempty intersection (our nested nonempty compacts meet, 4.2.2). The present problem makes two changes at once, and reading each is the whole orientation. "Compact" has been weakened to "closed and bounded in a complete space"—and over a general metric space those are not the same (closed-and-bounded need not be compact), so the proof cannot lean on compactness and must instead earn nonemptiness from *completeness* (5.4.3). In exchange, the new diameter hypothesis $\operatorname{diam} E_n \to 0$ is handed to us, and a hypothesis that strong is exactly what upgrades the conclusion from "nonempty" to "exactly one point." So the word *complete* points at the machinery for producing the point, and the phrase *exactly one* splits the task cleanly in two.

It helps to believe the model case first. On the real line take $E_n = [0, 1/n]$: each is closed, nonempty, bounded, nested downward, with $\operatorname{diam} E_n = 1/n \to 0$, and the intersection is the single point $\{0\}$. That is the nested-interval theorem (4.2.3) in miniature, and the present exercise is its abstraction—the same picture of sets clamping down on one point, but with completeness standing in for the special structure of $\mathbb{R}$.

How does completeness produce a point when there is no ambient compactness to call on? By the one tool that conjures a limit out of internal data alone: the Cauchy criterion (5.4.1). The recipe is to manufacture a candidate sequence and prove it Cauchy, and the proof runs in these steps:

(1) picking one x_n from each E_n is legitimate precisely because every E_n is nonempty—the nonemptiness hypothesis is spent here, on the very existence of the sequence;

(2) the sequence is Cauchy because, once $\operatorname{diam} E_N < \varepsilon$, every later index $m, n \geq N$ has both terms trapped inside the single set E_N (this is where the nesting $E_n \subseteq E_N$ does its work), so $d(x_m, x_n) \leq \operatorname{diam} E_N < \varepsilon$;

(3) completeness (5.4.3; Cauchy sequences converge in a complete space) converts "Cauchy" into an actual limit $x \in X$—the step that has no analogue in an incomplete space;

(4) the limit lands in each E_m because, from index m onward, the whole tail of the sequence sits in the closed set E_m, and a closed set contains the limits of its convergent sequences (3.2.4, via the tail of a convergent sequence enters every ball about its limit);

(5) uniqueness is the diameter hypothesis again: two points of the intersection lie in every E_n, so their distance is bounded by $\operatorname{diam} E_n \to 0$, forcing distance zero.

Watching where each hypothesis is consumed shows the statement is sharp: drop any one and it fails. Without nonemptiness there is no sequence to build (and the intersection could

be empty outright); without nesting, terms from different E_n need not huddle together and the Cauchy estimate collapses ($E_n = \{0\}$ and $\{1\}$ alternately keep diameters small yet share no point); without closedness the limit can leak out the boundary—$E_n = (0, 1/n)$ in $\mathbb{R}$ has empty intersection though everything else holds; without completeness the Cauchy sequence may have nowhere to converge—inside $\mathbb{Q}$, the nested closed sets $E_n = \{q \in \mathbb{Q} : |q - \sqrt{2}| \leq 1/n\}$ shrink onto a hole and meet in nothing; and without $\operatorname{diam} E_n \to 0$ the intersection can be large, as $E_n = [0, 1] \subseteq \mathbb{R}$ shows, where it is the whole interval. The argument is not circular: the point is built from the Cauchy sequence, never assumed to exist in advance, and uniqueness is checked separately from existence.

The transferable lesson is the division of labour between the two new conditions. Completeness is the existence engine—it is the abstract replacement for compactness in any "nested sets must meet" theorem, accessed through the Cauchy criterion rather than through subsequences. The shrinking diameter is the uniqueness engine, the same device that pins a real number between nested intervals. Together they recover, in any complete space, the full force of the nested-interval theorem (4.2.3); and the diameter-of-nested-sets viewpoint here is exactly the one the notes use to prove that compact spaces and $\mathbb{R}^k$ are complete (5.5.2, 5.5.3), so the present exercise is that proof's mirror image—there completeness is the conclusion, here it is the tool.

Problem 3.22 *Rudin, Chapter 3, Exercise 22*

Suppose X is a nonempty complete metric space, and $\{G_n\}$ is a sequence of dense open subsets of X. Prove Baire's theorem, namely, that $\bigcap_1^\infty G_n$ is not empty. (In fact, it is dense in X.) *Hint:* Find a shrinking sequence of neighborhoods E_n such that $\overline{E}_n \subset G_n$, and apply Exercise 21.

SOLUTION.

Write d for the metric, $B(p, \rho) = \{x : d(p, x) < \rho\}$ for the open ball, and $\overline{B}(p, \rho) = \{x : d(p, x) \leq \rho\}$. Density of a set G means G meets every nonempty open set (4.4.1). It is enough to prove the stronger claim that $\bigcap_{n\geq 1} G_n$ is dense, for then, X being a nonempty open set, the intersection meets X and so is nonempty.

So fix a nonempty open $W \subseteq X$; the goal is a point of $\bigcap_{n\geq 1} G_n$ lying in W. One small shrinking lemma is used repeatedly: if $B(p, \rho) \subseteq V$ with V open, then $\overline{B}(p, \rho/2) \subseteq B(p, \rho) \subseteq V$, so any nonempty open set contains the closure of a ball about any of its points, of as small a radius as we please.

Build closed balls $\overline{E}_n = \overline{B}(x_n, r_n)$ by recursion. Since G_1 is dense and W is nonempty open, $W \cap G_1$ is nonempty; it is open, being the intersection of two open sets. Choose a point in it and, by the shrinking lemma, a radius with

$$\overline{E}_1 = \overline{B}(x_1, r_1) \subseteq W \cap G_1, \qquad 0 < r_1 < 1.$$

Suppose $\overline{E}_n = \overline{B}(x_n, r_n)$ has been chosen. The open ball $B(x_n, r_n)$ is nonempty and open (2.3.1), and G_{n+1} is dense and open, so $B(x_n, r_n) \cap G_{n+1}$ is nonempty and open. Applying the shrinking lemma inside it, choose

$$\overline{E}_{n+1} = \overline{B}(x_{n+1}, r_{n+1}) \subseteq B(x_n, r_n) \cap G_{n+1}, \qquad 0 < r_{n+1} < \tfrac{1}{n+1}.$$

Three properties of the family $\{\overline{E}_n\}$ follow at once. Each $\overline{E}_n$ is a closed ball, hence nonempty (it contains x_n), closed, and bounded, with $\operatorname{diam} \overline{E}_n \le 2r_n$ (5.5.1). They are nested, since $\overline{E}_{n+1} \subseteq B(x_n, r_n) \subseteq \overline{B}(x_n, r_n) = \overline{E}_n$, so $\overline{E}_n \supseteq \overline{E}_{n+1}$ for every n. And the diameters vanish: from $r_n < \frac{1}{n}$ (with $r_1 < 1 = \frac{1}{1}$),

$$0 \le \operatorname{diam} \overline{E}_n \le 2r_n < \tfrac{2}{n} \xrightarrow[n\to\infty]{} 0.$$

The family $\{\overline{E}_n\}$ is therefore a nested sequence of nonempty closed bounded sets in the complete space X with $\operatorname{diam} \overline{E}_n \to 0$. By Problem 3.21, the intersection $\bigcap_{n\ge1} \overline{E}_n$ consists of exactly one point; call it x.

This x lands where it should. For every n the inclusion $\overline{E}_{n+1} \subseteq B(x_n, r_n) \cap G_{n+1} \subseteq G_{n+1}$ gives $x \in \overline{E}_{n+1} \subseteq G_{n+1}$, and $x \in \overline{E}_1 \subseteq G_1$, so $x \in G_n$ for every $n \ge 1$; and $x \in \overline{E}_1 \subseteq W$. Hence

$$x \in W \cap \bigcap_{n=1}^{\infty} G_n.$$

Thus $\bigcap_{n\ge1} G_n$ meets the arbitrary nonempty open set W, so it is dense in X (4.4.1), and being dense in the nonempty space X it is nonempty.

■

Reflections.

The phrase to read first is *complete metric space*: completeness is the one hypothesis without which the conclusion is false, so the proof must spend it somewhere, and the explicit hint says where—through Problem 3.21, the analogue of nested closed intervals that turns a tower of shrinking closed sets into a single common point. That points the whole strategy at a construction: a point lying in *all* of the G_n cannot be exhibited by a formula, so it must be *trapped*, caught as the unique limit of a nested family. The other two words in the hypotheses say what the family must dodge at each stage—*open* is what lets a ball be seated inside a set at all (2.3.3), and *dense* is what guarantees the next G_n reaches into whatever ball has been built so far (4.4.1). And "in fact dense" tells you not to aim at a bare point of $\bigcap G_n$ but to make the construction start inside an arbitrary nonempty open W, since meeting every such W is exactly what density means (4.4.1) and nonemptiness then comes for free from $W = X$.

It is worth seeing the difficulty before the construction. To sit in G_1 is easy; to sit in $G_1 \cap G_2$

is still easy; the trouble is the *infinite* intersection, where no finite stage commits you to a point and the limit could slip out through a gap. Completeness is precisely the promise that a controlled limit cannot slip out, provided the closed sets shrink to nothing—which is why the radii are forced to zero rather than merely kept positive.

The construction is a short recursive loop, each step resting on the one before:

(1) density of G_1 makes $W \cap G_1$ nonempty and intersection-of-opens keeps it open, so the shrinking lemma seats a closed ball $\overline{E}_1 \subseteq W \cap G_1$ of radius < 1—the base of the tower, planted inside W so the final point will land in W;

(2) at stage n the *open* ball $B(x_n, r_n)$ is a legitimate nonempty open target (2.3.1), so density of G_{n+1} lets it reach in and openness of $B(x_n, r_n) \cap G_{n+1}$ lets the shrinking lemma seat the next closed ball $\overline{E}_{n+1}$ inside it, with radius $< \frac{1}{n+1}$; one shrinks into the *open* ball $B(x_n, r_n)$, not the closed ball $\overline{E}_n$, exactly because a ball can only be seated inside an open set, and $B(x_n, r_n) \cap G_{n+1}$ is open while $\overline{E}_n \cap G_{n+1}$ need not be;

(3) nesting $\overline{E}_{n+1} \subseteq \overline{E}_n$ then comes for free from $B(x_n, r_n) \subseteq \overline{B}(x_n, r_n) = \overline{E}_n$, and the radius cap forces $\operatorname{diam} \overline{E}_n \leq 2r_n < \frac{2}{n} \to 0$ (5.5.1);

(4) the family is now nested, nonempty, closed, bounded, with diameters shrinking to zero, so Problem 3.21 hands back a single point $x \in \bigcap_n \overline{E}_n$, and $\overline{E}_n \subseteq G_n$ together with $\overline{E}_1 \subseteq W$ places it in $W \cap \bigcap_n G_n$.

The rigor turns on three things being genuinely present. The diameters must *vanish*, not merely the sets be nested: that is the whole hypothesis of Problem 3.21 and the reason it returns *one* point rather than possibly none—a nested sequence of nonempty closed sets can have empty intersection in a complete space (the closed sets $\{n, n+1, \dots\} \subseteq \mathbb{R}$ are a witness) unless their size is squeezed to zero. Completeness is load-bearing in the same breath, since Problem 3.21 is exactly where it is consumed; drop it and Baire fails outright, as $\mathbb{Q}$ shows—it is the countable union of its closed, empty-interior singletons, and the matching dense open sets $\mathbb{Q} \setminus \{q\}$ have empty intersection. And openness is what licenses every seating of a ball, while density is what keeps each G_{n+1} from missing the current ball; remove either and the recursion stalls at its first step.

The decisive contrast is with the $\mathbb{R}^k$ version, Problem 30, where the same tower is pinned down instead by nested nonempty *compact* sets (nested compacts meet), and the radii were pure decoration. That shortcut is unavailable here: in a general metric space a closed ball need not be compact, so there is no nested-compact property to fall back on, and the only substitute is completeness packaged as Problem 3.21—which is precisely why the diameters that were decorative there become essential now (completeness of X; 5.4.3). The transferable move is the one the hint is really teaching: to produce a point lying in infinitely many sets at once, never seek it directly—nest a tower of closed sets so that membership in

> the n-th set is built into the n-th ball, shrink the diameters to zero, and let completeness deliver the unique common point. This is the skeleton of every Baire-category argument.

Problem 3.23 *Rudin, Chapter 3, Exercise 23*

Suppose $\{p_n\}$ and $\{q_n\}$ are Cauchy sequences in a metric space X. Show that the sequence $\{d(p_n, q_n)\}$ converges. *Hint:* For any m, n,

$$d(p_n, q_n) \leq d(p_n, p_m) + d(p_m, q_m) + d(q_m, q_n);$$

it follows that

$$|d(p_n, q_n) - d(p_m, q_m)|$$

is small if m and n are large.

Solution.

Write $a_n = d(p_n, q_n)$; this is a sequence of real numbers, and the claim is that it converges in $\mathbb{R}$. We show $\{a_n\}$ is Cauchy and invoke the completeness of $\mathbb{R}$.

The key estimate compares a_n with a_m by routing each through the other pair of indices. The triangle inequality (2.2.1), applied twice, gives the hint's bound

$$d(p_n, q_n) \leq d(p_n, p_m) + d(p_m, q_m) + d(q_m, q_n),$$

which rearranges to

$$a_n - a_m = d(p_n, q_n) - d(p_m, q_m) \leq d(p_n, p_m) + d(q_m, q_n).$$

The roles of m and n are symmetric, so swapping them yields $a_m - a_n \leq d(p_m, p_n) + d(q_n, q_m)$. Since a metric is symmetric, the two right-hand sides agree, and combining the inequalities,

$$|a_n - a_m| = |d(p_n, q_n) - d(p_m, q_m)| \leq d(p_n, p_m) + d(q_n, q_m).$$

Now fix $\varepsilon > 0$. Because $\{p_n\}$ is Cauchy (5.4.1), there is an N_1 with $d(p_n, p_m) < \varepsilon/2$ whenever $m, n \geq N_1$; because $\{q_n\}$ is Cauchy, there is an N_2 with $d(q_n, q_m) < \varepsilon/2$ whenever $m, n \geq N_2$. Put $N = \max\{N_1, N_2\}$. For all $m, n \geq N$ the displayed bound gives

$$|a_n - a_m| \leq d(p_n, p_m) + d(q_n, q_m) < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

Thus $\{a_n\}$ is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete (every Cauchy sequence in $\mathbb{R}$ converges; 5.5.3), the sequence $\{a_n\} = \{d(p_n, q_n)\}$ converges to a real number.

■

REFLECTIONS.

Two clues in the statement decide the whole approach. The hypotheses hand you *Cauchy* sequences (5.4.1) but never a limit, and the goal is to show a derived sequence *converges* without naming what it converges to—and a request to prove convergence with no candidate limit in sight is the standing signal to test the Cauchy condition instead (6.1.3). The second clue is where the target sequence lives: $d(p_n, q_n)$ is a real number, so $\{a_n\} = \{d(p_n, q_n)\}$ is a sequence in $\mathbb{R}$, and $\mathbb{R}$ is the space where "Cauchy" is allowed to mean "convergent." Those two readings fit together into a single plan—show $\{a_n\}$ is Cauchy, then let completeness ($\mathbb{R}$ is complete; 5.5.3) supply the limit for free.

It is worth seeing why the limit cannot simply be written down. The space X need not be complete, so $\{p_n\}$ and $\{q_n\}$ may have no limits in X at all—imagine $X = \mathbb{Q}$ with $p_n \to \sqrt{2}$ and $q_n \to 0$ "from outside." There is then no point $p = \lim p_n$ to feed into $d(\cdot, \cdot)$, so any argument that first locates the limits of the two sequences is doomed. What survives is that the *distances* between the moving points settle down even when the points themselves have nowhere to land, and that settling-down is exactly Cauchyness of $\{a_n\}$.

The hint is really the proof, and unpacking it is the exercise:

(1) the triangle inequality (2.2.1), applied along the path $p_n \rightsquigarrow p_m \rightsquigarrow q_m \rightsquigarrow q_n$, gives the hint's three-term bound and so controls $a_n - a_m$ by $d(p_n, p_m) + d(q_m, q_n)$;

(2) swapping m and n controls $a_m - a_n$ by the same quantity (the metric is symmetric, so the bound is unchanged), and the two one-sided estimates combine into the two-sided $|a_n - a_m| \le d(p_n, p_m) + d(q_n, q_m)$;

(3) feeding in the two Cauchy conditions with tolerance $\varepsilon/2$ apiece, and taking the larger of the two thresholds, forces $|a_n - a_m| < \varepsilon$ for all large m, n—so $\{a_n\}$ is Cauchy in $\mathbb{R}$;

(4) completeness of $\mathbb{R}$ (5.5.3) converts Cauchy into convergent, finishing the proof.

The rigor turns on three quiet points, each load-bearing. The two-sided absolute value in step (2) is what a Cauchy estimate requires, and getting it needs *both* directions of the triangle inequality together with symmetry of d; with only the one-sided bound from the hint one would have controlled $a_n - a_m$ but not $|a_n - a_m|$. The split of ε into halves and the $N = \max\{N_1, N_2\}$ in step (3) are the standard device for honouring "for all m, n large" from two separate Cauchy hypotheses at once; drop either sequence's Cauchyness and the bound has an uncontrolled term, after which $\{d(p_n, q_n)\}$ can oscillate—take $p_n = 0$ and $q_n = 1 + (-1)^n$ in $\mathbb{R}$, where $\{p_n\}$ is Cauchy but $\{q_n\}$ is not and $d(p_n, q_n)$ runs $0, 2, 0, 2, \ldots$ without settling. Completeness in step (4) is not optional decoration: it is invoked for $\mathbb{R}$, the codomain of the metric, precisely because X itself may fail to be complete, which is why the conclusion is stated for the real sequence and not for any limit inside X.

The transferable move is to recognise when "prove it converges" should be answered by "prove it is Cauchy." Whenever the candidate limit is unavailable—unnamed, or living in a space that may not contain it—the Cauchy criterion (5.4.1) lets you certify convergence from the terms alone, provided you finish in a complete space. Here the trick is to notice that although the points may wander out of X, their pairwise distance is a real number, and $\mathbb{R}$ is always complete: the quotable relationship between convergent and Cauchy sequences run in reverse, pushed through the one inequality that every metric guarantees.

Problem 3.24 *Rudin, Chapter 3, Exercise 24*

Let X be a metric space.

1. Call two Cauchy sequences $\{p_n\}$, $\{q_n\}$ in X *equivalent* if

$$\lim_{n\to\infty} d(p_n, q_n) = 0.$$

   Prove that this is an equivalence relation.

2. Let X^* be the set of all equivalence classes so obtained. If $P \in X^*$, $Q \in X^*$, $\{p_n\} \in P$, $\{q_n\} \in Q$, define

$$\Delta(P, Q) = \lim_{n\to\infty} d(p_n, q_n);$$

   by Exercise 23, this limit exists. Show that the number $\Delta(P, Q)$ is unchanged if $\{p_n\}$ and $\{q_n\}$ are replaced by equivalent sequences, and hence that Δ is a distance function in X^*.

3. Prove that the resulting metric space X^* is complete.

4. For each $p \in X$, there is a Cauchy sequence all of whose terms are p; let P_p be the element of X^* which contains this sequence. Prove that

$$\Delta(P_p, P_q) = d(p, q)$$

   for all $p, q \in X$. In other words, the mapping φ defined by $\varphi(p) = P_p$ is an isometry (i.e., a distance-preserving mapping) of X into X^*.

5. Prove that $\varphi(X)$ is dense in X^*, and that $\varphi(X) = X^*$ if X is complete. By (d), we may identify X and $\varphi(X)$ and thus regard X as embedded in the complete metric space X^*. We call X^* the *completion* of X.

Solution.

Write $\mathcal{C}$ for the set of Cauchy sequences in (X, d), and for $\{p_n\}, \{q_n\} \in \mathcal{C}$ write $\{p_n\} \sim \{q_n\}$ when $\lim_n d(p_n, q_n) = 0$. Each such limit exists by Problem 3.23, whose estimate $|d(p_n, q_n) - d(p_m, q_m)| \le d(p_n, p_m) + d(q_n, q_m)$ makes $\{d(p_n, q_n)\}$ a Cauchy, hence convergent, sequence of reals.

(a) An equivalence relation. Reflexivity is $\lim_n d(p_n, p_n) = \lim_n 0 = 0$. Symmetry is the symmetry of the metric: $d(p_n, q_n) = d(q_n, p_n)$, so the two limits agree. For transitivity, suppose $\{p_n\} \sim \{q_n\}$ and $\{q_n\} \sim \{r_n\}$. The triangle inequality gives $0 \le d(p_n, r_n) \le d(p_n, q_n) + d(q_n, r_n)$, and the right-hand side tends to $0 + 0 = 0$, so the squeeze (6.3.3) forces $d(p_n, r_n) \to 0$, i.e. $\{p_n\} \sim \{r_n\}$. Thus $\sim$ is reflexive, symmetric, and transitive (0.4.16).

(b) Δ is well defined and a metric. Let $\{p_n\} \sim \{p'_n\}$ and $\{q_n\} \sim \{q'_n\}$. Applying the triangle inequality twice,

$$d(p_n, q_n) \le d(p_n, p'_n) + d(p'_n, q'_n) + d(q'_n, q_n),$$

and the symmetric estimate with the roles reversed, gives

$$\big|\, d(p_n, q_n) - d(p'_n, q'_n) \,\big| \le d(p_n, p'_n) + d(q_n, q'_n).$$

The right-hand side tends to 0 by the two equivalences, so $\lim_n d(p_n, q_n) = \lim_n d(p'_n, q'_n)$: the value depends only on the classes P, Q, not on the chosen representatives. Hence $\Delta(P, Q) = \lim_n d(p_n, q_n)$ is a well-defined number, and it inherits the metric axioms (2.2.1) from d in the limit. Each $\Delta(P, Q) = \lim_n d(p_n, q_n) \ge 0$; it is symmetric because each $d(p_n, q_n)$ is; the triangle inequality $d(p_n, r_n) \le d(p_n, q_n) + d(q_n, r_n)$ passes to the limit by the order properties of limits (6.3.1), giving $\Delta(P, R) \le \Delta(P, Q) + \Delta(Q, R)$. Finally $\Delta(P, Q) = 0$ means $\lim_n d(p_n, q_n) = 0$, which is exactly $\{p_n\} \sim \{q_n\}$, i.e. $P = Q$. So Δ is a distance function on X^*.

(c) X^* is complete. Let $\{P^{(k)}\}_{k \ge 0}$ be a Cauchy sequence in (X^*, Δ); the goal is a class $P \in X^*$ with $\Delta(P^{(k)}, P) \to 0$. For each k pick a representative Cauchy sequence $\{p_n^{(k)}\}_n \in P^{(k)}$. Since $\{p_n^{(k)}\}_n$ is Cauchy in X, choose an index N_k with

$$d(p_m^{(k)}, p_n^{(k)}) < \tfrac{1}{k+1} \qquad \text{for all } m, n \ge N_k,$$

and set $r_k := p_{N_k}^{(k)}$, the "N_k-th term" of the k-th representative. We show the diagonal sequence $\{r_k\}_k$ is Cauchy in X and that its class is the limit of $\{P^{(k)}\}$.

First, $\Delta\big(\varphi(r_k), P^{(k)}\big) \le \frac{1}{k+1}$, where $\varphi(r_k)$ is the class of the constant sequence at r_k. The distance is $\lim_n d\big(r_k, p_n^{(k)}\big)$, and for $n \ge N_k$ the choice of N_k gives $d\big(r_k, p_n^{(k)}\big) = d\big(p_{N_k}^{(k)}, p_n^{(k)}\big) < \frac{1}{k+1}$; passing to the limit in n (6.3.1) yields the bound. Now

$$\begin{aligned} d(r_j, r_k) &= \Delta\big(\varphi(r_j), \varphi(r_k)\big) \\ &\le \Delta\big(\varphi(r_j), P^{(j)}\big) + \Delta\big(P^{(j)}, P^{(k)}\big) + \Delta\big(P^{(k)}, \varphi(r_k)\big) \\ &\le \tfrac{1}{j+1} + \Delta\big(P^{(j)}, P^{(k)}\big) + \tfrac{1}{k+1}. \end{aligned}$$

using that φ preserves distance (part (d) below). Given $\varepsilon > 0$, take K with $\frac{1}{K+1} < \frac{\varepsilon}{3}$ and, by the Cauchyness of $\{P^{(k)}\}$, an index M with $\Delta(P^{(j)}, P^{(k)}) < \frac{\varepsilon}{3}$ for $j, k \geq M$; then $d(r_j, r_k) < \varepsilon$ for all $j, k \geq \max(K, M)$. So $\{r_k\}$ is Cauchy in X (5.4.1), and we let $P \in X^*$ be its equivalence class.

Finally $\Delta(P^{(k)}, P) \to 0$. By the triangle inequality in X^*,

$$\Delta\big(P^{(k)}, P\big) \leq \Delta\big(P^{(k)}, \varphi(r_k)\big) + \Delta\big(\varphi(r_k), P\big) \leq \tfrac{1}{k+1} + \Delta\big(\varphi(r_k), P\big).$$

The second term is $\lim_j d(r_k, r_j)$, and the Cauchy estimate for $\{r_k\}$ makes this small: given $\varepsilon > 0$, for k large enough $d(r_k, r_j) < \varepsilon$ for all large j, so $\Delta(\varphi(r_k), P) \leq \varepsilon$. Hence $\Delta(P^{(k)}, P) \to 0$, every Cauchy sequence in X^* converges, and X^* is complete (5.4.3; a complete space is one in which Cauchy sequences converge).

(d) φ is an isometry. For $p \in X$ the constant sequence $(p, p, p, \dots)$ is Cauchy, and $P_p = \varphi(p)$ is its class. For $p, q \in X$ the representatives are the constant sequences, so

$$\Delta(P_p, P_q) = \lim_n d(p, q) = d(p, q).$$

Thus φ preserves distance; in particular $\varphi(p) = \varphi(q)$ forces $d(p, q) = 0$, i.e. $p = q$, so φ is injective and is an isometry of X into X^*.

(e) $\varphi(X)$ is dense, and equals X^* when X is complete. Let $P \in X^*$ with representative $\{p_n\}$, and fix $\varepsilon > 0$. Being Cauchy, $\{p_n\}$ has an index N with $d(p_m, p_n) < \varepsilon$ for all $m, n \geq N$. Then $\varphi(p_N)$ approximates P:

$$\Delta\big(P, \varphi(p_N)\big) = \lim_n d(p_n, p_N) \leq \varepsilon,$$

since $d(p_n, p_N) < \varepsilon$ for all $n \geq N$ and the bound survives the limit (6.3.1). Every ball about P thus meets $\varphi(X)$, so $\varphi(X)$ is dense in X^*.

Suppose now X is complete and let $P \in X^*$ have representative $\{p_n\}$. As a Cauchy sequence in a complete space, $\{p_n\}$ converges to some $p \in X$ (Cauchy sequences converge in a complete space). Then $\{p_n\} \sim (p, p, \dots)$, because $d(p_n, p) \to 0$ by convergence (5.1.1); so P is the class of the constant sequence at p, that is $P = \varphi(p) \in \varphi(X)$. Hence $\varphi(X) = X^*$ when X is complete.

■

Reflections.

The exercise is, line for line, the construction our notes record as the *Cauchy completion* (5.4.4)—and reading it as a construction rather than a list of five drills is the whole orientation. The driving idea is the one from 6.1.3: a Cauchy sequence is a packet of data that "ought to" name a limit even when the space has no point to receive it, so if

X is missing points, the cleanest way to supply them is to let the Cauchy sequences *be* the points. Two Cauchy sequences that close in on each other should name the same new point, which is precisely why the problem opens by quotienting Cauchy sequences under $\lim_n d(p_n, q_n) = 0$. The words *equivalence relation*, *equivalence classes*, *distance function*, *complete*, and *isometry* are a checklist: build the set of points, put a metric on it, prove no Cauchy sequence escapes it, and show the original space sits inside undistorted.

It helps to believe the target before building it. The model case is $X = \mathbb{Q}$, where the missing points are the irrationals: the Cauchy sequence $1, 1.4, 1.41, 1.414, \ldots$ has no rational limit, yet it deserves to name $\sqrt{2}$, and any other rational sequence creeping toward the same gap is declared equivalent to it. Quotienting by "their distance tends to 0" glues exactly the sequences pointing at one hole into a single new point, and the completion $\mathbb{Q}^*$ is a copy of $\mathbb{R}$. Every abstract step below is this picture in disguise.

The five parts divide the labour, and each rests on a specific tool:

(1) the relation is reflexive and symmetric from the metric axioms, and transitive by the squeeze $0 \le d(p_n, r_n) \le d(p_n, q_n) + d(q_n, r_n) \to 0$ (6.3.3)—the triangle inequality is what lets "close to a common sequence" chain into "close to each other," so $\sim$ is genuinely an equivalence relation (0.4.16);

(2) before Δ can be a metric it must be *well defined on classes*, the recurring worry whenever a rule is stated through representatives (0.4.20); the four-term triangle estimate $|d(p_n, q_n) - d(p'_n, q'_n)| \le d(p_n, p'_n) + d(q_n, q'_n)$ shows swapping to equivalent representatives moves the limit by 0, and the metric axioms (2.2.1) then pass from d to Δ in the limit by the order properties of limits (6.3.1), with $\Delta(P, Q) = 0$ decoding to $P = Q$ exactly because $\sim$ was defined by vanishing distance;

(3) completeness is the heart, and the move is a diagonal one: from the k-th representative pull a single term $r_k = p^{(k)}_{N_k}$ chosen past the point where that sequence has settled to within $\frac{1}{k+1}$, so that $\varphi(r_k)$ sits within $\frac{1}{k+1}$ of $P^{(k)}$; the diagonal $\{r_k\}$ is then Cauchy in X (5.4.1), its class P is a point of X^*, and $P^{(k)} \to P$ because $P^{(k)}$ is squeezed between itself and P through the nearby $\varphi(r_k)$;

(4) the isometry is immediate—constant representatives give $\Delta(P_p, P_q) = \lim_n d(p, q) = d(p, q)$ directly—and distance preservation forces injectivity, which is the one place φ being an embedding (not merely a map) is used;

(5) density is the same approximation as (3) read backwards: a representative $\{p_n\}$ of P is Cauchy, so a late term p_N has $\Delta(P, \varphi(p_N)) = \lim_n d(p_n, p_N) \le \varepsilon$, putting a point of $\varphi(X)$ in every ball about P; and if X is already complete the representative *converges* in X (Cauchy sequences converge in a complete space), so its class is already a constant class and nothing new is added.

The argument earns the name "completion" only if no step quietly assumes what it is proving, and the delicate point is (3). The limit class P is not posited; it is built from the diagonal sequence $\{r_k\}$, which lives entirely inside the *original* X, so its Cauchyness is a statement about d, not about the new metric Δ—there is no circular appeal to a completeness of X^* we are trying to establish. The chosen indices N_k are exactly what converts each abstract class $P^{(k)}$ into a concrete nearby point $r_k \in X$ with controlled error $\frac{1}{k+1} \to 0$ (Archimedean property), and the triangle inequality, used through the isometry $d(r_j, r_k) = \Delta(\varphi(r_j), \varphi(r_k))$, transfers the Cauchyness of $\{P^{(k)}\}$ to that of $\{r_k\}$. Each hypothesis is load-bearing: drop "$\{p_n\}$ Cauchy" from the definition and the representative-distance of Problem 3.23 need not converge, so Δ would not even be a number; the constant sequences are Cauchy automatically, which is why φ is defined on all of X.

The transferable move is the completion recipe itself, and it is worth carrying whole: to repair a space that fails completeness, take its Cauchy sequences as raw points, glue together the ones whose distance vanishes, transport the metric by a limit, and recover the old space as the constant sequences. This is the abstract engine behind building $\mathbb{R}$ from $\mathbb{Q}$ (1.6.3), now run in any metric space; the density in (e) is what makes X "almost all" of its completion, and the isometry in (d) is what lets us stop distinguishing X from its image and simply regard X as living inside the complete space X^*.

Problem 3.25 *Rudin, Chapter 3, Exercise 25*

Let X be the metric space whose points are the rational numbers, with the metric $d(x, y) = |x-y|$. What is the completion of this space? (Compare Exercise 24.)

Solution.

The completion of $(\mathbb{Q}, |\cdot|)$ is $(\mathbb{R}, |\cdot|)$.

A completion of (X, d) is a complete metric space into which X embeds isometrically as a dense subspace, and it is unique up to isometry (5.4.4). So three things must be checked for the candidate $\mathbb{R}$: that it is complete, that the inclusion $\mathbb{Q} \hookrightarrow \mathbb{R}$ is an isometry, and that $\mathbb{Q}$ is dense in $\mathbb{R}$.

The metric on $\mathbb{R}$ is the restriction of $|\cdot|$, so for $x, y \in \mathbb{Q}$ the distance $|x - y|$ is the same whether computed in X or in $\mathbb{R}$; the inclusion map is therefore an isometry, and $\mathbb{Q}$ sits inside $\mathbb{R}$ as a subspace.

$\mathbb{R}$ is complete: every Cauchy sequence of real numbers converges, since $\mathbb{R} = \mathbb{R}^1$ is complete ($\mathbb{R}^k$ is complete; 5.5.3).

$\mathbb{Q}$ is dense in $\mathbb{R}$: every real number is a limit of rationals. Given $x \in \mathbb{R}$, density of $\mathbb{Q}$ ($\mathbb{Q}$ is dense in $\mathbb{R}$; 1.7.2) supplies, for each $n \geq 1$, a rational q_n with $|q_n - x| < \frac{1}{n}$, so every ball $B_{1/n}(x)$ meets $\mathbb{Q}$; hence $q_n \to x$ and $x \in \overline{\mathbb{Q}}$, and since x was arbitrary $\overline{\mathbb{Q}} = \mathbb{R}$.

Thus $\mathbb{R}$ is a complete metric space containing $\mathbb{Q}$ isometrically as a dense subspace, which

is exactly a completion of X. By the essential uniqueness in 5.4.4, any completion of $\mathbb{Q}$ is isometric to $\mathbb{R}$, so the completion is $\mathbb{R}$ with the metric $|x-y|$.

The abstract completion of Exercise 24 produces the same space in disguise. There $\widehat{X}$ is the set of Cauchy sequences of rationals modulo $\{p_n\} \sim \{q_n\} \iff |p_n - q_n| \to 0$, with $\widehat{d}([p_n],[q_n]) = \lim_n |p_n - q_n|$, and each $q \in \mathbb{Q}$ embeds as the class of the constant sequence. The map $[p_n] \mapsto \lim_n p_n \in \mathbb{R}$ is a well-defined bijection onto $\mathbb{R}$: the limit exists because $\{p_n\}$ is Cauchy in the complete space $\mathbb{R}$ (Cauchy sequences in $\mathbb{R}$ converge); it is independent of the representative because $|p_n - q_n| \to 0$ forces equal limits; it preserves distance, since $\widehat{d}([p_n],[q_n]) = \lim_n |p_n - q_n| = |\lim_n p_n - \lim_n q_n|$; and it is onto because every real is the limit of some rational sequence, by the density just shown. So Exercise 24's $\widehat{X}$ *is* $\mathbb{R}$.

■

REFLECTIONS.

The word that fixes everything is *completion*, and our notes give it a single precise meaning: a complete space into which X embeds isometrically as a *dense* subspace, unique up to isometry (5.4.4). Reading that definition is already most of the work, because it tells you that to *identify* a completion you do not have to build one—you only have to recognise a familiar space that meets the three conditions, and uniqueness does the rest. The parenthetical "compare Exercise 24" is the second clue: that exercise constructs the completion abstractly out of Cauchy sequences, so Exercise 25 is asking you to put a name to the object Exercise 24 manufactures.

The reason $\mathbb{R}$ is the right name is the relationship our notes record between $\mathbb{Q}$ and $\mathbb{R}$: $\mathbb{R}$ is the complete ordered field in which $\mathbb{Q}$ is dense (1.6.2), and $\mathbb{Q}$ is famously *not* complete—a sequence of rationals approaching $\sqrt{2}$ is Cauchy with no rational limit (6.1.2). That single failure is the whole motive for completing $\mathbb{Q}$ in the first place: completion is the operation that supplies the missing limits, and $\mathbb{R}$ is precisely $\mathbb{Q}$ with those limits filled in (1.6.1).

Believe it on the canonical hole first. The decimal truncations 1, 1.4, 1.41, 1.414, ... are rationals, they bunch together (Cauchy), yet inside $\mathbb{Q}$ they converge to nothing. Pass to $\mathbb{R}$ and the same sequence acquires the limit $\sqrt{2}$. The completion is exactly the act of adjoining one new point for each such rational Cauchy sequence that had nowhere to land.

The verification then walks the three clauses of the definition in turn, each a short citation:

(1) the inclusion $\mathbb{Q} \hookrightarrow \mathbb{R}$ is an isometry, because the metric on $\mathbb{R}$ restricts to the very same $|x-y|$ on rationals—there is nothing to compute, only to notice that the distance does not change when the ambient space grows;

(2) $\mathbb{R}$ is complete ($\mathbb{R}^k$ is complete; 5.5.3), which is the one substantive theorem in play and the reason a candidate exists at all—an incomplete target could never be a completion;

(3) $\mathbb{Q}$ is dense in $\mathbb{R}$ ($\mathbb{Q}$ is dense; 1.7.2): every real x has rationals $q_n \to x$ with $|q_n - x| < \frac{1}{n}$, so $x \in \overline{\mathbb{Q}}$ and the closure is all of $\mathbb{R}$.

With the three conditions met, the essential-uniqueness half of 5.4.4 converts "$\mathbb{R}$ *is a* completion" into "$\mathbb{R}$ *is the* completion"—the step that makes the answer a clean identification rather than merely an example.

The rigor check is to confirm all three clauses are genuinely load-bearing, since dropping any one breaks the identification. Without density, $\mathbb{R}^2$ (with $\mathbb{Q}$ on the first axis) would be a complete space containing $\mathbb{Q}$ isometrically, but it is far too large to be *the* completion; density is what forbids spurious extra points. Without completeness, $\mathbb{Q}$ itself trivially contains $\mathbb{Q}$ densely, but it is the very space we set out to repair. And the isometry clause is what pins the metric, not just the set: the completion of $\mathbb{Q}$ depends on which metric it carries—this is the force of the cross-reference to Exercise 24 and to other metrics on $\mathbb{Q}$ (under a p-adic metric the same set $\mathbb{Q}$ completes to a different space entirely). Matching the abstract construction of Exercise 24 to $\mathbb{R}$ by the map $[p_n] \mapsto \lim_n p_n$ closes the loop without circularity: that limit exists precisely because $\mathbb{R}$ is already known complete (Cauchy sequences in $\mathbb{R}$ converge), so the concrete answer is consistent with the abstract one rather than assumed from it.

The transferable move is to treat "what is the completion / closure / compactification of X?" as a *recognition* problem governed by a uniqueness theorem: find any concrete space satisfying the defining clauses—here complete, isometric, dense—and uniqueness promotes that example to the answer. It is also the cleanest way to see what $\mathbb{R}$ *is*: not a list of axioms but the metric completion of $\mathbb{Q}$, the smallest complete space holding the rationals densely, which is the construction 5.4.4 says is hiding inside the passage from $\mathbb{Q}$ to $\mathbb{R}$ in Lecture 1.

Rudin Chapter 4 – Continuity

Problem 4.1 *Rudin, Chapter 4, Exercise 1*

Suppose f is a real function defined on R^1 which satisfies

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

for every $x \in R^1$. Does this imply that f is continuous?

SOLUTION.

No. Define f on R^1 by $f(0)=1$ and $f(x)=0$ for $x \neq 0$. Fix x and take $h \neq 0$ small: if $x \neq 0$ then for $|h|<|x|$ both $x+h$ and $x-h$ avoid 0, while if $x=0$ then $x \pm h = \pm h \neq 0$. In either case $f(x+h)=f(x-h)=0$, so $f(x+h)-f(x-h)=0$ and the limit is 0 for every x. Yet f is not continuous at 0: $f(x)\to 0$ as $x\to 0$, whereas $f(0)=1$ (R 4.5; 7.2.1). So the condition does not imply continuity.

■

REFLECTIONS.

The expression $f(x+h)-f(x-h)$ only compares f at points placed symmetrically about x, and it never involves $f(x)$ itself. So it measures whether f is becoming balanced around x, not whether it is approaching the right value there. A lone spike — a function equal to 0 everywhere except for one raised point — is perfectly balanced around every x, since once h is small both sample points $x \pm h$ sit at height 0, so the symmetric limit vanishes everywhere even though the spike breaks continuity at that one point. The condition is genuinely weaker than continuity: it is a statement about symmetry, while continuity is the statement that $f(x)$ matches the values nearby.

Problem 4.2 *Rudin, Chapter 4, Exercise 2*

If f is a continuous mapping of a metric space X into a metric space Y, prove that

$$f(\overline{E}) \subset \overline{f(E)}$$

for every set $E \subset X$. ($\overline{E}$ denotes the closure of E.) Show, by an example, that $f(\overline{E})$ can be a proper subset of $\overline{f(E)}$.

SOLUTION.

Let $y \in f(\overline{E})$, say $y=f(x)$ with $x \in \overline{E}$. There is a sequence $\{x_n\}$ in E with $x_n \to x$ (take $x_n = x$ if $x \in E$, or a sequence in E approaching x if x is a limit point of E). By continuity $f(x_n)\to f(x)=y$ (R 4.2; R 4.6), and each $f(x_n) \in f(E)$, so y is a limit of points of $f(E)$, i.e. $y \in \overline{f(E)}$. Hence $f(\overline{E}) \subset \overline{f(E)}$.

For a proper inclusion, take $X=Y=R^1$, $f(x)=1/(1+x^2)$, and $E=\{1,2,3,\dots\}$. Then

E is closed, so $\overline{E} = E$ and $f(\overline{E}) = f(E) = \{1/(1+n^2) : n \geq 1\}$. Since $1/(1+n^2) \to 0$, the point 0 lies in $\overline{f(E)}$ but not in $f(E)$; thus $f(\overline{E})$ is a proper subset of $\overline{f(E)}$.

■

Reflections.

Closure adds the points reachable as limits from a set, and continuity is exactly the property that carries limits across the map: $x_n \to x$ forces $f(x_n) \to f(x)$. So a value $f(x)$ with x in the closure of E is automatically a limit of values coming from E — that is the whole inclusion. It can be *proper* because $\overline{f(E)}$ may acquire limit points that no point of $\overline{E}$ produces: in the example the outputs $1/(1+n^2)$ drift down toward 0, so 0 sits in the closure of the image, yet nothing in E (whose closure is E again, as E is closed) maps to 0. The map is continuous, but "0 at infinity" is a limit of outputs without being an output.

Problem 4.3 *Rudin, Chapter 4, Exercise 3*

Let f be a continuous real function on a metric space X. Let $Z(f)$ (the *zero set* of f) be the set of all $p \in X$ at which $f(p) = 0$. Prove that $Z(f)$ is closed.

Solution.

Let p be a limit point of $Z(f)$ and take $p_n \in Z(f)$ with $p_n \to p$. By continuity, $f(p_n) \to f(p)$ (R 4.6); but $f(p_n) = 0$ for all n, so $f(p) = 0$, i.e. $p \in Z(f)$. Thus $Z(f)$ contains all of its limit points and is closed (R 2.18). Equivalently, $Z(f) = f^{-1}(\{0\})$ is the preimage of the closed set $\{0\} \subset R^1$ under a continuous map, hence closed (R 4.8).

■

Reflections.

A zero set is a level set, $Z(f) = f^{-1}(\{0\})$, and the clean way to see it is closed is the general principle that continuous preimages of closed sets are closed. Concretely: if the points where f vanishes pile up at some p, continuity forces the value at p to be the limit of those zeros, namely 0, so p is itself a zero and the set cannot shed a boundary point. This is the same pattern by which closedness and openness of target sets pull back through continuous maps, which is the content of the open-set characterization of continuity (7.4.3).

Problem 4.4 *Rudin, Chapter 4, Exercise 4*

Let f and g be continuous mappings of a metric space X into a metric space Y, and let E be a dense subset of X. Prove that $f(E)$ is dense in $f(X)$. If $g(p) = f(p)$ for all $p \in E$, prove that $g(p) = f(p)$ for all $p \in X$. (In other words, a continuous mapping is determined by its values on a dense subset of its domain.)

Solution.

First, $f(E)$ is dense in $f(X)$. Let $y \in f(X)$, say $y = f(x)$. Since E is dense, there is a sequence $\{e_n\}$ in E with $e_n \to x$, and by continuity $f(e_n) \to f(x) = y$ (R 4.6; 7.2.1). Thus

y is a limit of points of $f(E)$, so every point of $f(X)$ lies in $\overline{f(E)}$, i.e. $f(E)$ is dense in $f(X)$.

Now suppose $g(p) = f(p)$ for all $p \in E$. Given $x \in X$, choose $e_n \in E$ with $e_n \to x$. Then $f(e_n) = g(e_n)$ for every n, while $f(e_n) \to f(x)$ and $g(e_n) \to g(x)$ by continuity. Limits in a metric space are unique (R 3.2(b)), so $f(x) = g(x)$. Hence $f = g$ on all of X.

■

REFLECTIONS.

A continuous map is completely determined by its values on a dense set, for the simple reason that every point of the space is a limit of dense-set points and continuity transports that limit. To compare two continuous maps that agree on a dense E, run both along a sequence from E approaching an arbitrary x: the two maps give identical values all along the sequence, each value-sequence converges to the map's value at x, and uniqueness of limits forces those two values to agree. The density-of-image statement is the same machinery read forward — the outputs over E already crowd against every output of the map. This is why a continuous function on R^1 is fixed once you know it on the rationals.

Problem 4.5 *Rudin, Chapter 4, Exercise 5*

If f is a real continuous function defined on a closed set $E \subset R^1$, prove that there exist continuous real functions g on R^1 such that $g(x) = f(x)$ for all $x \in E$. (Such functions g are called *continuous extensions* of f from E to R^1.) Show that the result becomes false if the word "closed" is omitted. Extend the result to vector-valued functions. *Hint:* Let the graph of g be a straight line on each of the segments which constitute the complement of E (compare Exercise 29, Chap. 2). The result remains true if R^1 is replaced by any metric space, but the proof is not so simple.

SOLUTION.

Since E is closed, $R^1 \setminus E$ is open, hence a countable union of disjoint open intervals (R 2.29). Define $g = f$ on E; on each bounded complementary interval (a, b) — whose endpoints a, b lie in E — let g be the line joining $(a, f(a))$ to $(b, f(b))$,

$$g(x) = f(a) + \frac{f(b) - f(a)}{b - a}\,(x - a) \qquad (a \le x \le b);$$

on an unbounded complementary interval $(-\infty, b)$ or (a, ∞) set $g \equiv f(b)$ or $g \equiv f(a)$, respectively.

Then g is real-valued on R^1 and equals f on E. It is continuous on $R^1 \setminus E$ (linear or constant near each such point) and at every interior point of E (where $g = f$ near the point). The one case to check is a point $p \in E$ that is an endpoint of complementary intervals. Given $\varepsilon > 0$, continuity of f at p gives $\delta > 0$ with $|f(q) - f(p)| < \varepsilon$ for $q \in E$ with $|q - p| < \delta$. If $|x - p| < \delta$, then x lies in E or in a complementary interval whose

endpoints are within δ of p; in the latter case $g(x)$ lies between the two endpoint values $f(\cdot)$, each within ε of $f(p)$. Either way $|g(x) - f(p)| < \varepsilon$, so g is continuous at p. Thus g is a continuous extension of f.

If "closed" is omitted the result fails: on the non-closed set $E = (0, 1)$ the function $f(x) = 1/x$ is continuous but unbounded near 0, so no function continuous on R^1 — being locally bounded — can agree with it on E.

For vector-valued $f = (f_1, \ldots, f_k) : E \to R^k$, extend each component f_i as above to g_i; then $g = (g_1, \ldots, g_k)$ is continuous and extends f.

■

Reflections.

A closed subset of the line is the whole line with a family of open "gaps" cut out, and those gaps are disjoint intervals. The extension simply connects the dots: across each finite gap draw the straight segment between the two known endpoint heights, and beyond an infinite gap hold the value flat. Nothing can go wrong inside a gap, where g is a line, or well inside E, where g is just f; the only sensitive place is a point of E at the mouth of a gap, and there the interpolated values are trapped between nearby values of f, so f's own continuity pulls them to the right limit. Closedness is precisely what forces the gap endpoints to belong to E, so f provides heights to interpolate between — drop it and the function can escape to infinity at a missing endpoint, as $1/x$ does on $(0, 1)$, leaving nothing to extend. The argument is one-dimensional, resting on the gap structure of open sets in R^1, which is why the general metric-space version takes more work.

Problem 4.6 *Rudin, Chapter 4, Exercise 6*

If f is defined on E, the *graph* of f is the set of points $(x, f(x))$, for $x \in E$. In particular, if E is a set of real numbers, and f is real-valued, the graph of f is a subset of the plane.
Suppose E is compact, and prove that f is continuous on E if and only if its graph is compact.

Solution.

Suppose first that f is continuous. The map $\Phi : E \to E \times f(E)$, $\Phi(x) = (x, f(x))$, has continuous coordinate functions and so is continuous, and the graph is $\Phi(E)$. As the continuous image of the compact set E, the graph is compact (R 4.14; 7.5.1).

Conversely, suppose the graph $G = \{(x, f(x)) : x \in E\}$ is compact, and let $x_n \to x$ in E; we show $f(x_n) \to f(x)$. If not, then $d_Y(f(x_{n_k}), f(x)) \geq \varepsilon$ for some $\varepsilon > 0$ and some subsequence. The points $(x_{n_k}, f(x_{n_k}))$ lie in the compact set G, so a further subsequence converges to a point $(a, b) \in G$, where $b = f(a)$ (R 3.6(a)). Its first coordinate is $\lim x_{n_k} = x$, so $a = x$ and $b = f(x)$; but then $f(x_{n_{k_j}}) \to b = f(x)$, contradicting $d_Y(f(x_{n_k}), f(x)) \geq \varepsilon$. Hence $f(x_n) \to f(x)$, and f is continuous (R 4.6).

■

REFLECTIONS.

The graph is a relabeled copy of the domain sitting up in the product $X \times Y$. One direction is the packaging fact that a map into a product is continuous exactly when its coordinates are, so $x \mapsto (x, f(x))$ is continuous and sends the compact domain to a compact graph. The other direction is where compactness earns its keep: if f tried to jump, you could follow a sequence on the graph whose heights stay ε away from the target, yet compactness of the graph forces a subsequence to settle onto an actual graph point; since the base coordinates head to x, that limit can only be $(x, f(x))$, which drags the heights back to $f(x)$ and contradicts the jump. The compactness of E is essential: it is what supplies the convergent subsequence that traps the limit on the graph.

Problem 4.7 *Rudin, Chapter 4, Exercise 7*

If $E \subset X$ and if f is a function defined on X, the *restriction* of f to E is the function g whose domain of definition is E, such that $g(p) = f(p)$ for $p \in E$. Define f and g on R^2 by: $f(0,0) = g(0,0) = 0$, $f(x,y) = xy^2/(x^2+y^4)$, $g(x,y) = xy^2/(x^2+y^6)$ if $(x,y) \neq (0,0)$. Prove that f is bounded on R^2, that g is unbounded in every neighborhood of $(0,0)$, and that f is not continuous at $(0,0)$; nevertheless, the restrictions of both f and g to every straight line in R^2 are continuous!

SOLUTION.

f is bounded: if $x = 0$ or $y = 0$ (in particular at the origin) then $f(x,y) = 0$. Otherwise $|x|\,y^2 > 0$, and $(|x| - y^2)^2 \geq 0$ gives $x^2 + y^4 \geq 2|x|y^2$, so

$$|f(x,y)| = \frac{|x|\,y^2}{x^2+y^4} \leq \frac{|x|\,y^2}{2|x|\,y^2} = \frac{1}{2}.$$

Hence $|f| \leq \frac{1}{2}$ on R^2.

g is unbounded in every neighborhood of the origin: along the curve $x = y^3$ $(y \neq 0)$,

$$g(y^3, y) = \frac{y^3 \cdot y^2}{y^6 + y^6} = \frac{y^5}{2y^6} = \frac{1}{2y} \xrightarrow[y \to 0]{} \infty,$$

and such points occur in every neighborhood of $(0,0)$.

f is not continuous at the origin: along $x = y^2$ $(y \neq 0)$, $f(y^2, y) = y^4/(y^4+y^4) = \frac{1}{2}$, so $f \to \frac{1}{2}$ along this approach while $f(0,0) = 0$; the limit does not equal the value (R 4.5; 7.3.6).

Finally, the restriction to any straight line is continuous. Away from the origin both f and g are continuous, being quotients with nonvanishing denominators, so only a line through the origin needs checking. Parametrize it as $(x,y) = (at, bt)$, $t \in R$, with $(a,b) \neq (0,0)$.

For $t \neq 0$,

$$f(at, bt) = \frac{ab^2 t}{a^2 + b^4 t^2}, \qquad g(at, bt) = \frac{ab^2 t}{a^2 + b^6 t^4},$$

and if $a = 0$ the line is the y-axis, on which $f = g = 0$. In every case the value $\to 0 = f(0,0) = g(0,0)$ as $t \to 0$, so each restriction is continuous.

■

Reflections.

The surprise is that being continuous along every straight line through a point is much weaker than being continuous there. Both functions are built so that any *linear* approach to the origin kills the numerator faster than the denominator — on a line $x = at$, $y = bt$ the ratio behaves like a constant times t, which runs to 0 — so every line sees a continuous function taking the value 0 at the origin. But continuity at a point demands that *all* approaches agree, and the denominators $x^2 + y^4$ and $x^2 + y^6$ are tuned to special curved approaches: along the parabola $x = y^2$ the two terms of f's denominator balance and f holds at $\frac{1}{2}$, while along $x = y^3$ the function g even blows up. So testing continuity in several variables means controlling every path into the point, not just the straight ones; checking lines can confirm a limit only once you already know the limit exists.

Problem 4.8 *Rudin, Chapter 4, Exercise 8*

Let f be a real uniformly continuous function on the bounded set E in R^1. Prove that f is bounded on E.

Show that the conclusion is false if boundedness of E is omitted from the hypothesis.

Solution.

By uniform continuity, for $\varepsilon = 1$ there is $\delta > 0$ with $|f(x) - f(y)| < 1$ whenever $x, y \in E$ and $|x - y| < \delta$ (R 4.18). Since E is bounded, $E \subset [a, b]$ for some $a < b$. Partition $[a, b]$ into finitely many subintervals $I_1, \ldots, I_N$ each of length $< \delta$, and for every k with $E \cap I_k \neq \varnothing$ fix a point $e_k \in E \cap I_k$. Any $x \in E$ lies in some such I_k, where $|x - e_k| < \delta$, so $|f(x)| \leq |f(e_k)| + 1 \leq \max_k |f(e_k)| + 1$. The right-hand side is a finite bound independent of x, so f is bounded on E.

If boundedness of E is dropped, the conclusion fails: $f(x) = x$ on $E = R^1$ is uniformly continuous (take $\delta = \varepsilon$) yet unbounded.

■

Reflections.

Uniform continuity hands you one δ that works at every point at once, and that uniformity is exactly what converts a bounded domain into a bound on f. Tile the bounded set with finitely many pieces, each narrower than δ; across any single piece f can change by less than a fixed amount, so starting from the finitely many sample values it can climb only a bounded distance and has no room to run to infinity. Finiteness of the tiling is where

boundedness of the domain is used; on an unbounded domain there are infinitely many pieces and nothing stops a uniformly continuous function from growing without end, as $f(x) = x$ shows. Ordinary continuity would not be enough either, since its δ could shrink from piece to piece and the tiling argument would fall apart.

Problem 4.9 *Rudin, Chapter 4, Exercise 9*

Show that the requirement in the definition of uniform continuity can be rephrased as follows, in terms of diameters of sets: To every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\operatorname{diam} f(E) < \varepsilon$ for all $E \subset X$ with $\operatorname{diam} E < \delta$.

SOLUTION.

By definition, f is uniformly continuous if for every $\varepsilon > 0$ there is $\delta > 0$ with $d_Y(f(p), f(q)) < \varepsilon$ whenever $d_X(p, q) < \delta$ (R 4.18), and $\operatorname{diam} A = \sup_{p,q \in A} d(p, q)$ (R 3.9).

Suppose f is uniformly continuous, and let $\varepsilon > 0$. Choose $\delta > 0$ so that $d_X(p, q) < \delta$ implies $d_Y(f(p), f(q)) < \varepsilon/2$. If $A \subset X$ has $\operatorname{diam} A < \delta$, then any $p, q \in A$ satisfy $d_X(p, q) < \delta$, hence $d_Y(f(p), f(q)) < \varepsilon/2$; taking the supremum over $p, q \in A$ gives $\operatorname{diam} f(A) \leq \varepsilon/2 < \varepsilon$.

Conversely, suppose the diameter condition holds, and let $\varepsilon > 0$; take the corresponding δ. For any p, q with $d_X(p, q) < \delta$, the two-point set $A = \{p, q\}$ has $\operatorname{diam} A = d_X(p, q) < \delta$, so $d_Y(f(p), f(q)) = \operatorname{diam} f(A) < \varepsilon$. Hence f is uniformly continuous, and the two formulations coincide.

■

REFLECTIONS.

This is the same definition in different dress: uniform continuity says small input distances force small output distances, with a threshold δ that does not depend on where you are. Stating it through diameters makes that location-independence vivid — "every set small enough, of diameter $< \delta$, has a small image, of diameter $< \varepsilon$" speaks about all sets at once, with no basepoint singled out. To recover the ordinary form, feed in the smallest interesting set, the pair $\{p, q\}$, whose diameter is exactly $d(p, q)$. The set-level phrasing is useful precisely because it talks about whole sets rather than points, which is what streamlines arguments like the extension theorem, where one follows the images of shrinking neighborhoods.

Problem 4.10 *Rudin, Chapter 4, Exercise 10*

Complete the details of the following alternative proof of Theorem 4.19: If f is not uniformly continuous, then for some $\varepsilon > 0$ there are sequences $\{p_n\}$, $\{q_n\}$ in X such that $d_X(p_n, q_n) \to 0$ but $d_Y(f(p_n), f(q_n)) > \varepsilon$. Use Theorem 2.37 to obtain a contradiction.

SOLUTION.

Suppose f is not uniformly continuous. Negating the definition, there is $\varepsilon > 0$ such that for every $\delta > 0$ some pair $p, q \in X$ has $d_X(p, q) < \delta$ yet $d_Y(f(p), f(q)) > \varepsilon$. Taking $\delta = 1/n$ produces sequences $\{p_n\}, \{q_n\}$ in X with $d_X(p_n, q_n) \to 0$ but $d_Y(f(p_n), f(q_n)) > \varepsilon$ for all n.

Since X is compact, $\{p_n\}$ has a subsequence $p_{n_k} \to p$ for some $p \in X$ (R 3.6(a), which rests on R 2.37). As $d_X(p_n, q_n) \to 0$, also $q_{n_k} \to p$, because $d_X(q_{n_k}, p) \le d_X(q_{n_k}, p_{n_k}) + d_X(p_{n_k}, p) \to 0$. By continuity of f at p, both $f(p_{n_k}) \to f(p)$ and $f(q_{n_k}) \to f(p)$ (R 4.6), so

$$d_Y(f(p_{n_k}), f(q_{n_k})) \le d_Y(f(p_{n_k}), f(p)) + d_Y(f(p), f(q_{n_k})) \to 0,$$

contradicting $d_Y(f(p_{n_k}), f(q_{n_k})) > \varepsilon$. Hence f is uniformly continuous (R 4.19).

■

Reflections.

This is the sequential route to the fact that continuity becomes uniform on a compact space. The failure of uniform continuity is really a statement about two moving points: however close you require the inputs to be, the outputs still refuse to come within ε. Record that refusal as two sequences p_n, q_n that close in on each other while their images stay ε apart. Compactness then supplies a single point p that a subsequence of the p_n approaches, and the q_n are dragged to the same p since they shadow the p_n; continuity at that one point pulls both image-streams to $f(p)$, so the images cannot stay apart, and the contradiction closes. Compactness does the essential work here by guaranteeing the common limit point that ties the two sequences together — on a noncompact space the construction survives and uniform continuity can genuinely fail.

Problem 4.11 *Rudin, Chapter 4, Exercise 11*

Suppose f is a uniformly continuous mapping of a metric space X into a metric space Y and prove that $\{f(x_n)\}$ is a Cauchy sequence in Y for every Cauchy sequence $\{x_n\}$ in X. Use this result to give an alternative proof of the theorem stated in Exercise 13.

Solution.

Let $\{x_n\}$ be a Cauchy sequence in X. Given $\varepsilon > 0$, uniform continuity provides $\delta > 0$ with $d_Y(f(x), f(y)) < \varepsilon$ whenever $d_X(x, y) < \delta$ (R 4.18). Since $\{x_n\}$ is Cauchy, there is N with $d_X(x_n, x_m) < \delta$ for all $n, m \ge N$, and then $d_Y(f(x_n), f(x_m)) < \varepsilon$. Thus $\{f(x_n)\}$ is Cauchy in Y.

This yields the extension theorem of Exercise 13 whenever Y is complete: if E is dense in X and $f : E \to Y$ is uniformly continuous, then for $p \in X$ pick $e_n \in E$ with $e_n \to p$. The sequence $\{e_n\}$ is Cauchy, so $\{f(e_n)\}$ is Cauchy and hence converges in the complete space Y; set $g(p) = \lim_n f(e_n)$. This limit does not depend on the approximating sequence (interleaving two sequences tending to p shows their images share a limit), g agrees with

f on E, and g is continuous — so g is the desired extension, with the estimates carried out in detail in Problem 13.

■

REFLECTIONS.

Uniform continuity is exactly the strength needed to carry Cauchy sequences to Cauchy sequences, and ordinary continuity is not: $f(x) = 1/x$ is continuous on $(0, 1)$ and sends the Cauchy sequence $1/n$ to n, which is not Cauchy. What makes uniform continuity succeed is its single location-free δ — once the inputs are eventually within δ of one another, the outputs are within ε, with no dependence on where the sequence is heading. This is the property that lets a uniformly continuous function be extended from a dense set to the whole space: a point outside the set is the limit of a Cauchy sequence from inside, its images form a Cauchy sequence, and in a complete range that sequence has a limit to serve as the extended value.

Problem 4.12 *Rudin, Chapter 4, Exercise 12*

A uniformly continuous function of a uniformly continuous function is uniformly continuous. State this more precisely and prove it.

SOLUTION.

Precisely: if $f : X \to Y$ and $g : Y \to Z$ are uniformly continuous mappings of metric spaces, then $g \circ f : X \to Z$ is uniformly continuous.

Let $\varepsilon > 0$. By uniform continuity of g there is $\eta > 0$ with $d_Z(g(y), g(y')) < \varepsilon$ whenever $d_Y(y, y') < \eta$. By uniform continuity of f there is $\delta > 0$ with $d_Y(f(x), f(x')) < \eta$ whenever $d_X(x, x') < \delta$ (R 4.18). Then $d_X(x, x') < \delta$ gives $d_Y(f(x), f(x')) < \eta$, which gives $d_Z(g(f(x)), g(f(x'))) < \varepsilon$. Hence $g \circ f$ is uniformly continuous.

■

REFLECTIONS.

The proof is just chaining the two moduli of continuity in the right order: begin with the target tolerance ε, let g turn it into a tolerance η on its inputs, then let f turn η into a tolerance δ on its inputs. What makes the conclusion *uniform* is that at each link the tolerance depends only on the previous tolerance and never on the location, so the final δ depends on ε alone. The identical chaining proves that a composition of merely continuous functions is continuous (R 4.7); there the tolerances may depend on the point, and the argument is simply run one point at a time.

Problem 4.13 *Rudin, Chapter 4, Exercise 13*

Let E be a dense subset of a metric space X, and let f be a uniformly continuous *real* function defined on E. Prove that f has a continuous extension from E to X (see Exercise 5 for terminology). (Uniqueness follows from Exercise 4.) *Hint:* For each $p \in X$ and each positive integer n,

let $V_n(p)$ be the set of all $q \in E$ with $d(p,q) < 1/n$. Use Exercise 9 to show that the intersection of the closures of the sets $f(V_1(p))$, $f(V_2(p))$, ..., consists of a single point, say $g(p)$, of R^1. Prove that the function g so defined on X is the desired extension of f.
Could the range space R^1 be replaced by R^k? By any compact metric space? By any complete metric space? By any metric space?

Solution.

Let $f : E \to R^1$ be uniformly continuous on the dense set $E \subset X$. For $p \in X$, density gives $e_n \in E$ with $e_n \to p$; being convergent, $\{e_n\}$ is Cauchy, so by Problem 11 $\{f(e_n)\}$ is Cauchy in R^1 and therefore converges (R 3.11). Define $g(p) = \lim_n f(e_n)$.

This is independent of the sequence: if $e_n \to p$ and $e'_n \to p$, the interleaving $e_1, e'_1, e_2, e'_2, \ldots$ also converges to p, hence is Cauchy, so its image is Cauchy and convergent, forcing $\lim f(e_n) = \lim f(e'_n)$. For $p \in E$ the constant sequence gives $g(p) = f(p)$, so g extends f.

To see g is uniformly continuous, let $\varepsilon > 0$ and choose $\delta > 0$ from the uniform continuity of f so that $a, b \in E$ with $d(a,b) < \delta$ give $|f(a) - f(b)| \le \varepsilon/2$. Suppose $p, p' \in X$ with $d(p,p') < \delta/3$, and take $e_n \to p$, $e'_n \to p'$. For large n, $d(e_n, p) < \delta/3$ and $d(e'_n, p') < \delta/3$, so $d(e_n, e'_n) < \delta$ and $|f(e_n) - f(e'_n)| \le \varepsilon/2$; letting $n \to \infty$ gives $|g(p) - g(p')| \le \varepsilon/2 < \varepsilon$. Thus g is uniformly continuous, and uniqueness follows from Problem 4, since two continuous extensions agree on the dense set E.

The only property of the range used is that Cauchy sequences converge there. So R^1 may be replaced by R^k, by any compact metric space, or by any complete metric space — each is complete. It cannot be replaced by an arbitrary metric space: with $X = R^1$, $E = \mathbb{Q}$, and $f : \mathbb{Q} \to \mathbb{Q}$ the identity (uniformly continuous), there is no continuous extension $g : R^1 \to \mathbb{Q}$, for a continuous map from the connected space R^1 into $\mathbb{Q}$ has connected image (R 4.22), hence is constant, while g must restrict to the identity on $\mathbb{Q}$.

■

Reflections.

The extension is forced on you: a point p outside E is approached by points e_n of the dense set, and if g is to be continuous it has no choice but to take the value $\lim f(e_n)$. The only question is whether that limit exists, and uniform continuity guarantees it — the e_n form a Cauchy sequence, uniform continuity (through Problem 11) makes $f(e_n)$ a Cauchy sequence, and in a complete range that sequence converges. This is why the range must be complete: completeness is the promise that the manufactured Cauchy sequence of values has a limit to become the new value. R^k, compact spaces, and complete spaces all keep that promise; the rationals do not, and the identity on $\mathbb{Q}$ has nowhere to send the irrational limits. Uniform rather than ordinary continuity is essential, since ordinary continuity need not preserve Cauchyness and the extension can genuinely fail to exist.

Problem 4.14 *Rudin, Chapter 4, Exercise 14*

Let $I = [0, 1]$ be the closed unit interval. Suppose f is a continuous mapping of I into I. Prove that $f(x) = x$ for at least one $x \in I$.

SOLUTION.

Define $h : I \to R^1$ by $h(x) = f(x) - x$; it is continuous as the difference of continuous functions (R 4.9). Since f maps I into $I = [0, 1]$, we have $h(0) = f(0) \geq 0$ and $h(1) = f(1) - 1 \leq 0$. If $h(0) = 0$ then 0 is a fixed point, and if $h(1) = 0$ then 1 is; otherwise $h(0) > 0 > h(1)$, and by the intermediate value theorem (R 4.23; 7.5.5) there is $x \in (0, 1)$ with $h(x) = 0$. In every case $f(x) = x$ for some $x \in I$.

■

REFLECTIONS.

This is the one-dimensional Brouwer fixed-point theorem, and the move that does the real work is to stop hunting for a fixed point and instead hunt for a zero of $h(x) = f(x) - x$. A fixed point of f is exactly a root of h, and h is easy to read at the two ends: because f cannot leave $[0, 1]$, at the left end $f(0) \geq 0$ puts h at or above zero, and at the right end $f(1) \leq 1$ puts h at or below zero. A continuous function that is nonnegative at one end and nonpositive at the other must vanish somewhere between, which is the intermediate value theorem. Geometrically, the graph of f sits in the unit square and has to cross the diagonal $y = x$, since it begins on or above the diagonal and finishes on or below it.

Problem 4.15 *Rudin, Chapter 4, Exercise 15*

Call a mapping of X into Y *open* if $f(V)$ is an open set in Y whenever V is an open set in X. Prove that every continuous open mapping of R^1 into R^1 is monotonic.

SOLUTION.

We first show f is injective. Suppose $f(x_1) = f(x_2)$ with $x_1 < x_2$. By the extreme value theorem (R 4.16), f attains a maximum and a minimum on $[x_1, x_2]$. If either is attained at an interior point $p \in (x_1, x_2)$ with $f(p) \neq f(x_1)$, then $f(p)$ is the largest (or smallest) value of f on the open interval (x_1, x_2), so $f((x_1, x_2))$ contains $f(p)$ but no value beyond it, and $f(p)$ is not an interior point of that image — contradicting that f carries the open set (x_1, x_2) to an open set. Otherwise both extrema equal $f(x_1)$, so f is constant on $[x_1, x_2]$ and $f((x_1, x_2))$ is a single point, again not open. Hence f is injective.

A continuous injective function on R^1 is strictly monotonic. Indeed, were it not, there would be $a < b < c$ with $f(b)$ not between $f(a)$ and $f(c)$; as the three values are distinct, $f(b)$ either exceeds both or lies below both. Say $f(a) < f(c) < f(b)$ (the remaining cases are symmetric). Choosing y with $f(c) < y < f(b)$, the intermediate value theorem (R 4.23; 7.5.5) yields $s \in (a, b)$ and $t \in (b, c)$ with $f(s) = f(t) = y$, and $s \neq t$ contradicts injectivity. Therefore f is monotonic.

■

Reflections.

The picture is that an open map cannot fold the line back on itself. A fold would place a local maximum or minimum at some interior point, and there the function turns around, so just past the peak the values stop climbing — the image acquires a top edge it never crosses, and a set with a top edge is not open. So openness forbids interior extrema, which is the same as saying f never repeats a value, since a repeat would force an interior extremum in between; that is, f is injective. The final step is the familiar fact that a continuous injection on an interval must be strictly monotonic, because any out-of-order triple of values would let the intermediate value theorem manufacture two inputs sharing one output. The extreme value theorem and the intermediate value theorem are the two workhorses here, one forbidding folds and the other detecting repeats.

Problem 4.16 *Rudin, Chapter 4, Exercise 16*

Let $[x]$ denote the largest integer contained in x, that is, $[x]$ is the integer such that $x-1 < [x] \leq x$; and let $(x) = x - [x]$ denote the fractional part of x. What discontinuities do the functions $[x]$ and (x) have?

Solution.

Both functions are continuous away from the integers — on each interval $[n, n+1)$ the floor $[x] = n$ is constant and $(x) = x - n$ is linear — and each has a simple discontinuity (one of the first kind) at every integer n (R 4.26).

For the floor: as $x \to n^-$ one has $[x] = n-1$, so $[\,n^-] = n-1$, while $[\,n^+] = [n] = n$. Both one-sided limits exist and differ, a jump of 1 at each integer.

For the fractional part $(x) = x - [x]$: $(n) = 0$, and as $x \to n^-$ the relation $[x] = n-1$ gives $(x) = x - (n-1) \to 1$, so $(n^-) = 1$, whereas $(n^+) = 0 = (n)$. Again both one-sided limits exist, now with a jump from 1 down to 0 at each integer. Neither function has any discontinuity other than these.

■

Reflections.

Both functions are driven by the integer count that $[x]$ performs, so they can misbehave only where that count ticks over — at the integers — and between consecutive integers each is as tame as possible, constant or a straight line of slope 1. At an integer the floor jumps up by 1, and the fractional part, being what is left after subtracting the floor, drops from 1 back down to 0: the familiar sawtooth reset. Each of these is a discontinuity of the first kind, meaning both one-sided limits exist and are finite and the function fails to be continuous only because the two sides, or the assigned value, disagree. There are no oscillatory or infinite discontinuities anywhere.

Problem 4.17 *Rudin, Chapter 4, Exercise 17*

Let f be a real function defined on (a, b). Prove that the set of points at which f has a simple discontinuity is at most countable. *Hint:* Let E be the set on which $f(x-) < f(x+)$. With each point x of E, associate a triple (p, q, r) of rational numbers such that

1. $f(x-) < p < f(x+)$,
2. $a < q < t < x$ implies $f(t) < p$,
3. $x < t < r < b$ implies $f(t) > p$.

The set of all such triples is countable. Show that each triple is associated with at most one point of E. Deal similarly with the other possible types of simple discontinuities.

SOLUTION.

At a simple discontinuity the one-sided limits $f(x-)$ and $f(x+)$ both exist (R 4.26); the possibilities are $f(x-) < f(x+)$, $f(x-) > f(x+)$, or $f(x-) = f(x+) \neq f(x)$. It is enough to show each type forms an at most countable set.

Take first the set E where $f(x-) < f(x+)$. For each $x \in E$ choose rationals p, q, r with

$$f(x-) < p < f(x+), \qquad a < q < x < r < b,$$

such that $f(t) < p$ for $q < t < x$ and $f(t) > p$ for $x < t < r$. These exist: p is any rational between the two one-sided limits, and since $f(t) \to f(x-) < p$ as $t \to x^-$ and $f(t) \to f(x+) > p$ as $t \to x^+$, the inequalities $f(t) < p$ and $f(t) > p$ hold on suitable one-sided neighborhoods, in which we place rationals q and r. The assignment $x \mapsto (p, q, r)$ is injective on E: if $x < x'$ shared the same triple, choose t with $x < t < x'$; then $t < r$ forces $f(t) > p$ (from x), while $q < t$ forces $f(t) < p$ (from x'), a contradiction. So E injects into $\mathbb{Q}^3$ and is at most countable.

The type $f(x-) > f(x+)$ is identical with the inequalities reversed. For $f(x-) = f(x+) = L \neq f(x)$, say $L < f(x)$ (the case $L > f(x)$ is symmetric), choose rationals p, q, r with $L < p < f(x)$, $a < q < x < r < b$, and $f(t) < p$ for all $t \in (q, r) \setminus \{x\}$ (possible since $f(t) \to L < p$ from both sides); if $x < x'$ shared this triple, then $x' \in (q, r) \setminus \{x\}$ would force $f(x') < p$, contradicting $f(x') > p$. Hence each type is at most countable, and the set of all simple discontinuities, a union of finitely many such sets, is at most countable.

■

REFLECTIONS.

The strategy is to give every simple discontinuity its own rational "address" and then count the addresses. At a jump with $f(x-) < f(x+)$ there is a real gap between the two one-sided limits; drop a rational p into that gap and bracket x by rationals q, r tight enough that f stays below p just to the left of x and above p just to the right. No second

discontinuity can reuse that triple, since two of them would demand both $f > p$ and $f < p$ at a shared point between them. As there are only countably many rational triples, there can be only countably many such points. The existence of the one-sided limits is exactly what opens the gap for the rational p — which is why the count works for simple discontinuities and says nothing about wilder ones, and why a monotonic function, whose discontinuities are all jumps, can have at most countably many.

Problem 4.18 *Rudin, Chapter 4, Exercise 18*

Every rational x can be written in the form $x = m/n$, where $n > 0$, and m and n are integers without any common divisors. When $x = 0$, we take $n = 1$. Consider the function f defined on R^1 by

$$f(x) = \begin{cases} 0 & (x \text{ irrational}), \\ \dfrac{1}{n} & \left(x = \dfrac{m}{n}\right). \end{cases}$$

Prove that f is continuous at every irrational point, and that f has a simple discontinuity at every rational point.

Solution.

We show $\lim_{x \to x_0} f(x) = 0$ for every $x_0 \in R^1$. Fix x_0 and $\varepsilon > 0$, and choose an integer N with $1/N < \varepsilon$. The interval $(x_0 - 1, x_0 + 1)$ contains only finitely many rationals m/n in lowest terms with $n \leq N$, since each such denominator permits only finitely many numerators in a bounded range. Let $\delta \in (0, 1)$ be less than the distance from x_0 to each of those rationals other than x_0 itself. Then for $0 < |x - x_0| < \delta$: if x is irrational, $f(x) = 0 < \varepsilon$; if $x = m/n$ in lowest terms, then x is none of the excluded rationals, so $n > N$ and $f(x) = 1/n < 1/N < \varepsilon$. Hence $\lim_{x \to x_0} f(x) = 0$.

If x_0 is irrational, then $f(x_0) = 0 = \lim_{x \to x_0} f(x)$, so f is continuous at x_0 (R 4.6). If $x_0 = m/n$ is rational, then $f(x_0) = 1/n > 0$ while both one-sided limits equal $\lim_{x \to x_0} f(x) = 0 \neq f(x_0)$, so f has a simple discontinuity (of the first kind) at x_0 (R 4.26).

■

Reflections.

This is Thomae's function, and its whole behavior rests on one counting fact: the rationals with a *small* denominator — exactly the ones where f takes a *large* value — are sparse, only finitely many in any bounded stretch. So to make f small near a point you only have to dodge those finitely many spikes, which a small enough δ does, while everything else (the irrationals, and the rationals with large denominators) already has f tiny. That forces the limit to be 0 at every point, regardless of whether the point is rational. The rational-versus-irrational split then decides continuity by the value alone: at an irrational the value is 0 and matches the limit, so f is continuous; at a rational the value $1/n$ pokes up above the limit 0, a clean removable-type discontinuity. It is the standard example

of a function continuous at every irrational and discontinuous at every rational — and a theorem of Baire says no function can do the reverse.

Problem 4.19 *Rudin, Chapter 4, Exercise 19*

Suppose f is a real function with domain R^1 which has the intermediate value property: If $f(a) < c < f(b)$, then $f(x) = c$ for some x between a and b.
Suppose also, for every rational r, that the set of all x with $f(x) = r$ is closed.
Prove that f is continuous.
Hint: If $x_n \to x_0$ but $f(x_n) > r > f(x_0)$ for some r and all n, then $f(t_n) = r$ for some t_n between x_0 and x_n; thus $t_n \to x_0$. Find a contradiction. (N. J. Fine, *Amer. Math. Monthly*, vol. 73, 1966, p. 782.)

SOLUTION.

Suppose, for contradiction, that f is discontinuous at some x_0. Then there are $\varepsilon > 0$ and a sequence $x_n \to x_0$ with $|f(x_n) - f(x_0)| \geq \varepsilon$ for all n. Passing to a subsequence we may assume $f(x_n) > f(x_0) + \varepsilon$ for all n, or else $f(x_n) < f(x_0) - \varepsilon$ for all n; take the first case (the second is symmetric).

Choose a rational r with $f(x_0) < r < f(x_0) + \varepsilon$, so that $f(x_0) < r < f(x_n)$ for every n. By the intermediate value property there is a point t_n between x_0 and x_n with $f(t_n) = r$; since t_n lies between x_0 and x_n and $x_n \to x_0$, we get $t_n \to x_0$. The points t_n all belong to $\{x : f(x) = r\}$, which is closed by hypothesis, so its limit x_0 belongs to it too (R 2.18); that is, $f(x_0) = r$. But $r > f(x_0)$, a contradiction. Hence f is continuous at every point.

■

REFLECTIONS.

Neither hypothesis alone forces continuity — the intermediate value property permits badly behaved functions, and closed level sets are easy to arrange — yet together they pin f down. The idea is to assume a jump at x_0 and watch what the intermediate value property does with it. If f leaps above $f(x_0)$ along a sequence approaching x_0, then for any rational level r just above $f(x_0)$ the function must, between x_0 and each x_n, pass through r, producing solutions of $f = r$ that march right up to x_0. Closedness of the level set $\{f = r\}$ then traps x_0 inside it, forcing $f(x_0) = r$, impossible since r sits strictly above $f(x_0)$. The rationals appear only so that countably many closed-set hypotheses are enough; the real mechanism is that the intermediate value property turns a jump into nearby solutions of $f = r$ that the closed level set refuses to release.

Problem 4.20 *Rudin, Chapter 4, Exercise 20*

If E is a nonempty subset of a metric space X, define the distance from $x \in X$ to E by

$$\rho_E(x) = \inf_{z \in E} d(x, z).$$

1. Prove that $\rho_E(x) = 0$ if and only if $x \in \overline{E}$.

2. Prove that ρ_E is a uniformly continuous function on X, by showing that

$$|\rho_E(x) - \rho_E(y)| \leq d(x, y)$$

for all $x \in X$, $y \in X$. *Hint:* $\rho_E(x) \leq d(x, z) \leq d(x, y) + d(y, z)$, so that

$$\rho_E(x) \leq d(x, y) + \rho_E(y).$$

Solution.

1. If $\rho_E(x) = 0$, then for each n there is $z_n \in E$ with $d(x, z_n) < 1/n$, so $z_n \to x$ and $x \in \overline{E}$. Conversely, if $x \in \overline{E}$ then either $x \in E$, whence $\rho_E(x) \leq d(x, x) = 0$, or x is a limit point of E, so every ball about x meets E and the infimum $\rho_E(x) = 0$. Thus $\rho_E(x) = 0$ if and only if $x \in \overline{E}$ (R 2.18).

2. For any $z \in E$ and any $x, y \in X$, the triangle inequality gives $\rho_E(x) \leq d(x, z) \leq d(x, y) + d(y, z)$. Taking the infimum over $z \in E$ on the right,

$$\rho_E(x) \leq d(x, y) + \rho_E(y), \qquad \text{that is,} \qquad \rho_E(x) - \rho_E(y) \leq d(x, y).$$

By symmetry $\rho_E(y) - \rho_E(x) \leq d(x, y)$, so $|\rho_E(x) - \rho_E(y)| \leq d(x, y)$. Given $\varepsilon > 0$, taking $\delta = \varepsilon$ then shows ρ_E is uniformly continuous (R 4.18; 7.7.1).

■

Reflections.

The function ρ_E measures how far a point sits from the set E, and the two parts confirm the two things you would hope for. Part (a): being at distance zero from E means there are points of E arbitrarily close, which is exactly what it means to lie in the closure — so ρ_E vanishes precisely on $\overline{E}$, not merely on E. Part (b): moving from x to y can change the distance-to-E by no more than the distance you moved, because any route from x to a point of E may detour through y. That single inequality, $|\rho_E(x) - \rho_E(y)| \leq d(x, y)$, is Lipschitz continuity with constant 1, which is far stronger than plain continuity and is automatically uniform. This little function is the standard device for manufacturing continuous functions tailored to a set — it is what lets one separate disjoint closed sets and prove that a metric space is normal in the problems that follow.

Problem 4.21 *Rudin, Chapter 4, Exercise 21*

Suppose K and F are disjoint sets in a metric space X, K is compact, F is closed. Prove that there exists $\delta > 0$ such that $d(p, q) > \delta$ if $p \in K$, $q \in F$. *Hint:* ρ_F is a continuous positive function on K.

Show that the conclusion may fail for two disjoint closed sets if neither is compact.

SOLUTION.

Consider $\rho_F(p) = \inf_{q\in F} d(p,q)$ on K. By Problem 20, ρ_F is continuous and $\rho_F(p) = 0$ exactly when $p \in \overline{F} = F$. Since $K \cap F = \varnothing$, we have $\rho_F(p) > 0$ for every $p \in K$. As ρ_F is continuous on the compact set K, it attains a minimum (R 4.16): there is $p_0 \in K$ with $\rho_F(p_0) = \min_{p\in K} \rho_F(p) =: m$, and $m = \rho_F(p_0) > 0$. Put $\delta = m/2 > 0$. For any $p \in K$ and $q \in F$, $d(p,q) \geq \rho_F(p) \geq m > \delta$.

The conclusion can fail for two disjoint closed sets when neither is compact. In R^2 let $K = \{(x,0) : x \in R\}$ (the x-axis) and $F = \{(x, 1/x) : x > 0\}$. Both are closed and disjoint, and neither is compact (both are unbounded), yet $d((x,0),(x,1/x)) = 1/x \to 0$, so no $\delta > 0$ separates them.

■

REFLECTIONS.

The distance from a point to a closed set is positive unless the point lies in the set, and ρ_F records that distance continuously. On the compact set K this positive function attains its minimum — it cannot merely approach zero without reaching it — so the minimum is a genuine positive number, and that number is a uniform gap between K and F. Compactness is doing the essential work: it upgrades "positive at every point" to "bounded below by a positive constant." Without it the gap can shrink to zero in the limit even though the sets never meet, as the x-axis and the graph of $1/x$ show — they stay disjoint while drawing arbitrarily close out toward infinity.

Problem 4.22 *Rudin, Chapter 4, Exercise 22*

Let A and B be disjoint nonempty closed sets in a metric space X, and define

$$f(p) = \frac{\rho_A(p)}{\rho_A(p) + \rho_B(p)} \qquad (p \in X).$$

Show that f is a continuous function on X whose range lies in $[0,1]$, that $f(p) = 0$ precisely on A and $f(p) = 1$ precisely on B. This establishes a converse of Exercise 3: Every closed set $A \subset X$ is $Z(f)$ for some continuous real f on X. Setting

$$V = f^{-1}\left(\left[0, \tfrac{1}{2}\right)\right), \qquad W = f^{-1}\left(\left(\tfrac{1}{2}, 1\right]\right),$$

show that V and W are open and disjoint, and that $A \subset V$, $B \subset W$. (Thus pairs of disjoint closed sets in a metric space can be covered by pairs of disjoint open sets. This property of metric spaces is called *normality*.)

SOLUTION.

First, $\rho_A(p)+\rho_B(p)>0$ for every p: if it vanished then $\rho_A(p)=\rho_B(p)=0$, so $p\in\overline{A}=A$ and $p\in\overline{B}=B$ (Problem 20), contradicting $A\cap B=\varnothing$. Hence f is defined everywhere, and as a quotient of the continuous functions ρ_A and $\rho_A+\rho_B$ with nonvanishing denominator it is continuous (R 4.9). Since $0\le\rho_A\le\rho_A+\rho_B$, the range of f lies in $[0,1]$. Moreover $f(p)=0 \iff \rho_A(p)=0 \iff p\in A$, and $f(p)=1 \iff \rho_B(p)=0 \iff p\in B$. In particular $A=\{p: f(p)=0\}=Z(f)$, so every closed set is a zero set of a continuous function — the converse of Problem 3 (indeed ρ_A alone already has zero set A).

Finally, $V=f^{-1}([0,\frac12))=f^{-1}((-\infty,\frac12))$ and $W=f^{-1}((\frac12,1])=f^{-1}((\frac12,\infty))$ are preimages of open sets under the continuous f, hence open (R 4.8). They are disjoint, and $A\subset V$ (since $f=0<\frac12$ on A) while $B\subset W$ (since $f=1>\frac12$ on B). Thus disjoint closed sets are separated by disjoint open sets, i.e. X is normal.

■

Reflections.

This builds, by hand, a continuous function that is 0 on A, 1 on B, and slides between them in between — a metric-space version of Urysohn's lemma. The construction is the natural ratio: weigh the distance to A against the total distance to both sets, so a point of A scores 0, a point of B scores 1, and the denominator never vanishes precisely because no point can touch both disjoint closed sets at once. Once such a function exists, normality is immediate — split the values at $\frac12$ and pull back the two open halves — and so is the converse to "zero sets are closed," since A is exactly where f vanishes. The distance function ρ_A from the previous problems is the single ingredient that makes all of this run.

Problem 4.23 *Rudin, Chapter 4, Exercise 23*

A real-valued function f defined in (a,b) is said to be *convex* if

$$f(\lambda x+(1-\lambda)y)\le\lambda f(x)+(1-\lambda)f(y)$$

whenever $a<x<b$, $a<y<b$, $0<\lambda<1$. Prove that every convex function is continuous. Prove that every increasing convex function of a convex function is convex. (For example, if f is convex, so is e^f.)

If f is convex in (a,b) and if $a<s<t<u<b$, show that

$$\frac{f(t)-f(s)}{t-s}\le\frac{f(u)-f(s)}{u-s}\le\frac{f(u)-f(t)}{u-t}.$$

Solution.

We prove the slope inequality first, since continuity follows from it. Fix $a<s<t<u<b$

and write $t = \lambda s + (1-\lambda)u$ with $\lambda = \frac{u-t}{u-s} \in (0,1)$, so $1-\lambda = \frac{t-s}{u-s}$. Convexity gives

$$f(t) \le \lambda f(s) + (1-\lambda) f(u) = \frac{(u-t)f(s) + (t-s)f(u)}{u-s}.$$

Subtracting $f(s)$ and dividing by $t-s>0$ gives $\frac{f(t)-f(s)}{t-s} \le \frac{f(u)-f(s)}{u-s}$; subtracting the displayed bound from $f(u)$ and dividing by $u-t>0$ gives $\frac{f(u)-f(s)}{u-s} \le \frac{f(u)-f(t)}{u-t}$. Together these are the asserted chain.

Continuity. Fix $x_0 \in (a,b)$ and choose s,u with $a<s<x_0<u<b$. Applying the chain to $s<x_0<x$ and to $x_0<x<u$, for any $x \in (x_0,u)$,

$$\frac{f(x_0)-f(s)}{x_0-s} \le \frac{f(x)-f(x_0)}{x-x_0} \le \frac{f(u)-f(x_0)}{u-x_0}.$$

The difference quotient is thus bounded by constants independent of x, so $|f(x)-f(x_0)| \le C\,(x-x_0) \to 0$; the same bound holds for $x \in (s,x_0)$. Hence f is continuous at x_0 (R 4.5; 7.2.1).

Increasing convex of convex. Let φ be increasing and convex on an interval containing the range of f. For x,y and $\lambda \in (0,1)$, convexity of f gives $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$; applying φ (increasing), then its convexity,

$$\varphi\big(f(\lambda x + (1-\lambda)y)\big) \le \varphi\big(\lambda f(x) + (1-\lambda) f(y)\big) \le \lambda\, \varphi(f(x)) + (1-\lambda)\, \varphi(f(y)),$$

so $\varphi \circ f$ is convex. Taking $\varphi = \exp$ shows e^f is convex whenever f is.

■

REFLECTIONS.

Convexity says the graph lies below each of its chords, and everything here is a way of reading that off the slopes of those chords. The three-term inequality says exactly that chord slopes increase as the chord slides right: the slope from s to t is at most that from s to u, which is at most that from t to u. That monotone-slope picture is the key relation. Continuity falls out because near any interior point the chord slopes are trapped between two fixed values, making f Lipschitz there — and this needs an *open* interval, since a convex function on a closed interval can jump at an endpoint. The composition fact is almost a formality once seen the right way: f convex puts its value below the chord height, an increasing φ preserves that order, and φ's own convexity bounds φ of the chord height by the chord of $\varphi \circ f$.

Problem 4.24 *Rudin, Chapter 4, Exercise 24*

Assume that f is a continuous real function defined in (a, b) such that

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}$$

for all $x, y \in (a, b)$. Prove that f is convex.

Solution.

We prove $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ first for dyadic λ, then for all $\lambda \in [0,1]$ by continuity.

Dyadic λ. Induct on n, proving the inequality for every $\lambda = j/2^n$ with $0 \leq j \leq 2^n$. For $n = 1$ the only nontrivial case is $\lambda = \frac{1}{2}$, which is the hypothesis. Assume the claim for n, and let $\lambda = j/2^{n+1}$ with j odd (if j is even, λ is already a multiple of $1/2^n$). Set $\mu = \frac{j-1}{2^{n+1}}$, $\nu = \frac{j+1}{2^{n+1}}$; as j is odd, μ and ν are multiples of $1/2^n$, and $\lambda = \frac{1}{2}(\mu+\nu)$. With $p = \mu x + (1-\mu)y$ and $q = \nu x + (1-\nu)y$ we have $\lambda x + (1-\lambda)y = \frac{1}{2}(p+q)$, so midpoint convexity gives

$$f(\lambda x + (1-\lambda)y) = f\left(\frac{p+q}{2}\right) \leq \frac{f(p)+f(q)}{2}.$$

The inductive hypothesis bounds $f(p) \leq \mu f(x)+(1-\mu)f(y)$ and $f(q) \leq \nu f(x)+(1-\nu)f(y)$; averaging these and using $\frac{1}{2}(\mu+\nu) = \lambda$,

$$\frac{f(p)+f(q)}{2} \leq \frac{\mu+\nu}{2} f(x) + \Big(1 - \frac{\mu+\nu}{2}\Big) f(y) = \lambda f(x) + (1-\lambda) f(y).$$

All λ. The dyadic rationals are dense in $[0,1]$. Given $\lambda \in [0,1]$, take dyadic $\lambda_k \to \lambda$; then $\lambda_k x + (1-\lambda_k)y \to \lambda x + (1-\lambda)y$, and continuity of f (R 4.6) lets us pass to the limit in $f(\lambda_k x + (1-\lambda_k)y) \leq \lambda_k f(x) + (1-\lambda_k)f(y)$, giving $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. Hence f is convex.

■

Reflections.

Midpoint convexity is the convexity inequality at the single ratio $\frac{1}{2}$, and the question is how much that one case already forces. Taking midpoints repeatedly reaches all the dyadic ratios — halving generates $\frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \ldots$ — so the inequality spreads from $\frac{1}{2}$ to a dense set of λ by an induction that splits an odd dyadic into the average of its two even neighbors. Density finishes the job only because f is continuous: a continuous function obeying an inequality on a dense set obeys it everywhere. Continuity is essential rather than cosmetic — without it, midpoint-convex functions can be wildly discontinuous (built from a non-measurable additive function) and fail to be convex, so the hypothesis is exactly what excludes those pathologies.

Problem 4.25 *Rudin, Chapter 4, Exercise 25*

If $A \subset R^k$ and $B \subset R^k$, define $A + B$ to be the set of all sums $\mathbf{x} + \mathbf{y}$ with $\mathbf{x} \in A$, $\mathbf{y} \in B$.

1. If K is compact and C is closed in R^k, prove that $K + C$ is closed. *Hint:* Take $\mathbf{z} \notin K + C$, put $F = \mathbf{z} - C$, the set of all $\mathbf{z} - \mathbf{y}$ with $\mathbf{y} \in C$. Then K and F are disjoint. Choose δ as in Exercise 21. Show that the open ball with center $\mathbf{z}$ and radius δ does not intersect $K + C$.

2. Let α be an irrational real number. Let C_1 be the set of all integers, let C_2 be the set of all $n\alpha$ with $n \in C_1$. Show that C_1 and C_2 are closed subsets of R^1 whose sum $C_1 + C_2$ is *not* closed, by showing that $C_1 + C_2$ is a countable dense subset of R^1.

SOLUTION.

1. Let $\mathbf{z} \notin K+C$; we find a ball about $\mathbf{z}$ missing $K+C$. Put $F = \mathbf{z}-C = \{\mathbf{z}-\mathbf{y} : \mathbf{y} \in C\}$, closed since C is closed and $\mathbf{y} \mapsto \mathbf{z} - \mathbf{y}$ is a homeomorphism. Then $K \cap F = \varnothing$: if $\mathbf{w} \in K \cap F$, then $\mathbf{w} = \mathbf{z} - \mathbf{y}$ for some $\mathbf{y} \in C$, giving $\mathbf{z} = \mathbf{w} + \mathbf{y} \in K + C$, a contradiction. By Problem 21 there is $\delta > 0$ with $|\mathbf{p} - \mathbf{q}| > \delta$ for all $\mathbf{p} \in K$, $\mathbf{q} \in F$. If some $\mathbf{w}$ with $|\mathbf{w} - \mathbf{z}| < \delta$ lay in $K + C$, write $\mathbf{w} = \mathbf{k} + \mathbf{c}$ with $\mathbf{k} \in K$, $\mathbf{c} \in C$; then $\mathbf{z} - \mathbf{c} \in F$ and $|\mathbf{k} - (\mathbf{z} - \mathbf{c})| = |\mathbf{w} - \mathbf{z}| < \delta$, contradicting the choice of δ. So the ball of radius δ about $\mathbf{z}$ misses $K + C$, and $K + C$ is closed.

2. $C_1 = \mathbb{Z}$ and $C_2 = \{n\alpha : n \in \mathbb{Z}\}$ are closed: each consists of isolated points (consecutive elements differ by 1 and by $|\alpha|$, respectively), so neither has a limit point. Their sum $C_1 + C_2 = \{m + n\alpha : m, n \in \mathbb{Z}\}$ is countable. It is dense: given $\varepsilon > 0$, choose an integer $N > 1/\varepsilon$ and split $[0, 1)$ into N subintervals of length $1/N$; among the $N + 1$ fractional parts of $0, \alpha, 2\alpha, \ldots, N\alpha$ two lie in one subinterval, so for some $0 \le i < j \le N$ there is an integer m with $0 < |(j - i)\alpha - m| < 1/N < \varepsilon$ (nonzero since α is irrational). Then $\beta = (j - i)\alpha - m \in C_1 + C_2$ satisfies $0 < |\beta| < \varepsilon$, and its integer multiples, all in $C_1 + C_2$, are spaced $|\beta|$ apart, so $C_1 + C_2$ meets every interval of length ε. Being dense, $C_1 + C_2$ has closure all of R^1; but $C_1 + C_2 \neq R^1$ since it is countable, so $C_1 + C_2$ does not contain all its limit points and is not closed.

■

REFLECTIONS.

Part (a) is the Minkowski sum of a compact set with a closed set, and the reason it stays closed is the uniform gap from the previous problem. To show a point $\mathbf{z}$ outside the sum has room around it, reflect the closed set through $\mathbf{z}$ to turn "$\mathbf{z} \notin K + C$" into "K and the reflected set are disjoint"; one is compact and one is closed, so a positive distance δ holds them apart, and that same δ is the radius of a ball about $\mathbf{z}$ the sum cannot reach. Part (b) shows the compactness is not optional — two closed sets can sum to something badly non-closed. The integers together with the integer multiples of an irrational α give $\mathbb{Z} + \mathbb{Z}\alpha$, the countable dense set behind irrational rotations: dense because an irrational

slope forces the fractional parts of $n\alpha$ to crowd into every subinterval, and a countable dense set can never be closed.

Problem 4.26 *Rudin, Chapter 4, Exercise 26*

Suppose X, Y, Z are metric spaces, and Y is compact. Let f map X into Y, let g be a continuous one-to-one mapping of Y into Z, and put $h(x) = g(f(x))$ for $x \in X$.
Prove that f is uniformly continuous if h is uniformly continuous. *Hint:* g^{-1} has compact domain $g(Y)$, and $f(x) = g^{-1}(h(x))$.
Prove also that f is continuous if h is continuous.
Show (by modifying Example 4.21, or by finding a different example) that the compactness of Y cannot be omitted from the hypotheses, even when X and Z are compact.

Solution.

Since Y is compact and $g : Y \to Z$ is continuous and one-to-one, g is a homeomorphism of Y onto $g(Y)$; in particular $g^{-1} : g(Y) \to Y$ is continuous (R 4.17). Also $g(Y)$ is compact (R 4.14), so g^{-1}, being continuous on a compact space, is uniformly continuous (R 4.19). Because $g \circ f = h$, we have $f = g^{-1} \circ h$.

If h is uniformly continuous, then $f = g^{-1} \circ h$ is a composition of uniformly continuous maps (Problem 12), hence uniformly continuous. If h is merely continuous, then $f = g^{-1} \circ h$ is a composition of continuous maps, hence continuous (R 4.7).

Compactness of Y cannot be dropped, even with X and Z compact. Let $X = Z$ be the unit circle (compact) and $Y = [0, 2\pi)$ (not compact), with $g : [0, 2\pi) \to Z$, $g(t) = (\cos t, \sin t)$, the continuous one-to-one map of Example 4.21 (R 4.21). Take $h = \mathrm{id}$ on the circle, which is uniformly continuous. The relation $g \circ f = h$ forces $f = g^{-1}$, the inverse parametrization, discontinuous at $(1, 0)$ (its value jumps from near 2π to 0). So f fails to be continuous although h is uniformly continuous, showing the compactness of Y is essential.

■

Reflections.

The point is that an injective continuous map out of a compact space is automatically a homeomorphism onto its image, so its inverse is not just continuous but uniformly continuous. That lets you write $f = g^{-1} \circ h$ and inherit whatever regularity h has — uniform continuity from uniform continuity, continuity from continuity — because composing with the well-behaved g^{-1} cannot spoil it. Everything hinges on g^{-1} being continuous, and that is precisely where compactness of Y is spent: drop it and g can be a continuous bijection whose inverse jumps, like the wrapping of the half-open interval $[0, 2\pi)$ onto the circle. There h can be as smooth as the identity while $f = g^{-1}$ tears apart at the point where the interval's two ends are glued together.

PART III

Resources and References

The Course List of Facts

These are the results the instructor lists as quotable on homework and exams. The list is not exhaustive. Each fact has a stable key and an automatically generated identifier, and each carries a pointer to where it lives in our notes. "Quotable" means you may use it without reproving it, but you should still know *why* it holds.

Number Systems

- **F01.** $\mathbb{Q}$ is an ordered field: all the field and order rules, and valid manipulations of (in)equalities, may be used without proof. (1.3.4)
- **F02.** There exist nonempty subsets of $\mathbb{Q}$ that are bounded above but have no supremum in $\mathbb{Q}$. (1.4.2)
- **F03.** $\mathbb{R}$ is an ordered field containing $\mathbb{Q}$ as a subfield and satisfying the least-upper-bound property: every nonempty bounded-above subset has a supremum in $\mathbb{R}$. (1.6.1)
- **F04.** $\mathbb{R}$ is Archimedean: if $x > 0$ then $nx > y$ for some $n \in \mathbb{N}$; equivalently $\inf\{1/n\} = 0$. (1.7.1)
- **F05.** $\mathbb{Q}$ is dense in $\mathbb{R}$: every real interval contains a rational. (1.7.2)

Countable and Uncountable Sets

- **F06.** $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countable. (1.7.2; 2.1.4, 2.1.7)
- **F07.** A subset of a countable set is finite or countable. (2.1.9)
- **F08.** A countable union of countable sets is countable. (2.1.15)
- **F09.** A finite product of countable sets is countable. (2.1.17)
- **F10.** The set of all sequences of 0's and 1's is uncountable. (2.1.10)
- **F11.** $\mathbb{R}$ is uncountable. (2.1.13)

Open and Closed Subsets

- **F12.** The open balls of (X, d) are open. (2.3.5)
- **F13.** If p is a cluster point of E, every ball around p contains infinitely many points of E. (3.2.1)
- **F14.** A subset is open in (X, d) iff its complement is closed. (3.2.8)
- **F15.** Any union of open sets is open; a finite intersection of open sets is open. (3.1.2)
- **F16.** Any intersection of closed sets is closed; a finite union of closed sets is closed. (3.2.10)
- **F17.** If E' is the set of cluster points of E, then $E \cup E'$ is closed. (4.4.1)

Compactness

- **F18.** Compactness is absolute: K is compact relative to X iff relative to any $Y \supseteq X$; one may simply say (K, d) is compact. (3.4.4)
- **F19.** Compact subsets of metric spaces are closed and bounded. (4.1.1)
- **F20.** Closed subsets of compact spaces are compact. (4.1.3)
- **F21.** The intersection of nonempty, nested compact sets is nonempty. (4.2.2)
- **F22.** An infinite subset of a compact space has a cluster point. (4.3.3)
- **F23.** A subset of $\mathbb{R}^k$ is compact iff it is closed and bounded. (4.3.5)

Sequences

- **F24.** If $x_n \to p$, every ball around p contains all but finitely many x_n. (5.1.3)
- **F25.** A sequence in a metric space has at most one limit. (5.1.3)
- **F26.** A convergent sequence is bounded. (5.1.3)
- **F27.** If p is a cluster point of E, some sequence in E converges to p. (5.1.3)
- **F28.** Limits in $\mathbb{R}^k$ respect the algebraic operations. (5.2.1)
- **F29.** A sequence in $\mathbb{R}^k$ converges iff each coordinate converges. (5.2.2)
- **F30.** Every sequence in a compact metric space has a convergent subsequence; every bounded sequence in $\mathbb{R}^k$ does too. (5.3.3)
- **F31.** Every convergent sequence is Cauchy; every Cauchy sequence is bounded; if a subsequence of a Cauchy sequence converges to p, so does the sequence. (5.4.2, 5.5.3)
- **F32.** Every Cauchy sequence in $\mathbb{R}^k$ or in a compact metric space converges: these are complete. (5.5.3)
- **F33.** A bounded, monotonic sequence in $\mathbb{R}$ converges. (6.2.2)

Series

- **F34.** *Cauchy criterion:* $\sum a_n$ converges iff for every $\varepsilon > 0$ there is N with $\left|\sum_{k=m}^{n} a_k\right| < \varepsilon$ for all $n > m \geq N$. (6.6.2)
- **F35.** If $\sum a_n$ converges then $a_n \to 0$; the converse fails (the harmonic series $\sum 1/n$ diverges). (6.6.3, 6.6.4)
- **F36.** A series of nonnegative terms converges iff its partial sums are bounded. (6.6.6)
- **F37.** *Comparison test:* if $|a_n| \leq c_n$ eventually and $\sum c_n$ converges, then $\sum a_n$ converges; if $a_n \geq d_n \geq 0$ and $\sum d_n$ diverges, then $\sum a_n$ diverges. (6.6.8)

- **F38.** *Geometric series:* $\sum_{n\geq 0} x^n = \frac{1}{1-x}$ for $|x| < 1$, and the series diverges for $|x| \geq 1$.(6.6.7)
- **F39.** *Root test:* with $\alpha = \limsup_n |a_n|^{1/n}$, the series converges if $\alpha < 1$, diverges if $\alpha > 1$, and the test is inconclusive if $\alpha = 1$. (6.6.9)
- **F40.** *Ratio test:* if $\limsup_n |a_{n+1}/a_n| < 1$ the series converges; if $|a_{n+1}/a_n| \geq 1$ for all large n it diverges. (6.6.10)

Continuity

- **F41.** $\lim_{x\to p} f(x) = q$ iff $f(p_n) \to q$ for every sequence $p_n \to p$ with $p_n \neq p$; in particular a function limit, when it exists, is unique. (7.1.6)
- **F42.** For real-valued f, g: $\lim(f+g) = \lim f + \lim g$, and likewise for products and quotients (the latter when $\lim g \neq 0$), provided the right-hand limits exist. (7.1.6; cf. 5.2.1)
- **F43.** f is continuous at p iff $\lim_{x\to p} f(x) = f(p)$. (7.2.1)
- **F44.** $f : X \to Y$ is continuous iff $f^{-1}(V)$ is open in X for every open $V \subseteq Y$. (7.4.3)
- **F45.** The composition of continuous functions is continuous. (7.2.4, 7.4.4)
- **F46.** A map into $\mathbb{R}^k$ is continuous iff each of its coordinate functions is continuous. (via the sequential criterion (7.4.1) and 5.2.2)
- **F47.** The continuous image of a compact metric space is compact. (7.5.1)
- **F48.** A real-valued continuous function on a compact metric space attains its maximum and minimum. (7.5.2)
- **F49.** If K is compact and $f : K \to Y$ is a continuous bijection, then f^{-1} is continuous.(7.6.2)
- **F50.** A continuous function on a compact set is uniformly continuous. (7.7.2)
- **F51.** The continuous image of a connected set is connected. (7.5.4)
- **F52.** *Intermediate Value Theorem:* a continuous $f : [a, b] \to \mathbb{R}$ takes every value between $f(a)$ and $f(b)$. (7.5.5)

Differentiation

The statements in this section concern functions $\mathbb{R} \to \mathbb{R}$.

- **F53.** If f is differentiable at x_0 then f is continuous at x_0. (8.2.1)
- **F54.** Sums, products, and quotients of differentiable functions are differentiable. (8.2.3)
- **F55.** *Chain rule:* a composition of differentiable functions is differentiable, with $(f \circ g)'(x) = f'(g(x))\, g'(x)$. (8.2.5)
- **F56.** If f has a local extremum at an interior point x_0 and is differentiable there, then $f'(x_0) = 0$ (this can fail at endpoints). (8.4.3)

- **F57.** *Mean Value Theorem:* if f is continuous on $[a,b]$ and differentiable on (a,b), then $f'(x) = \dfrac{f(b)-f(a)}{b-a}$ for some $x \in (a,b)$. (8.4.7)
- **F58.** If $f' \geq 0$ on an interval then f is increasing there, with the corresponding statements for $\leq 0, > 0, < 0, = 0$. (8.4.9)
- **F59.** If f' exists on $[a,b]$ then f' has the intermediate value property: a derivative has no jump discontinuities. (8.5.1)
- **F60.** *Taylor's theorem:* if f is n times differentiable on (a,b) and $x_0, x_0+h \in (a,b)$, then for some ξ strictly between x_0 and x_0+h, $f(x_0+h) = \sum_{k=0}^{n-1} \frac{f^{(k)}(x_0)}{k!}h^k + \frac{f^{(n)}(\xi)}{n!}h^n$. (8.6.3)

Riemann Integration

The statements in this section concern functions $\mathbb{R} \to \mathbb{R}$.

- **F61.** Refining a partition does not decrease the lower Riemann sum and does not increase the upper Riemann sum. (9.2.2)
- **F62.** The lower integral is at most the upper integral. (9.2.5)
- **F63.** *Integrability criterion:* f is integrable iff for every $\varepsilon > 0$ there is a partition P with $U(P,f) - L(P,f) < \varepsilon$. (9.2.7)
- **F64.** If f is continuous on $[a,b]$ then f is integrable on $[a,b]$. (9.3.1)
- **F65.** If f is monotone on $[a,b]$ then f is integrable on $[a,b]$. (9.3.2)
- **F66.** If f is bounded on $[a,b]$ with only finitely many discontinuities, then f is integrable. (9.3.3)
- **F67.** If f is integrable and φ is continuous, then $\varphi \circ f$ is integrable. (9.4.1)
- **F68.** Properties of the integral: linearity, monotonicity, additivity over adjacent intervals, and the bound $\left|\int_a^b f\right| \leq M(b-a)$ when $|f| \leq M$. (9.4.2)
- **F69.** A product of integrable functions is integrable. (9.4.2)
- **F70.** If f is integrable then $|f|$ is integrable and $\left|\int_a^b f\right| \leq \int_a^b |f|$. (9.4.2)
- **F71.** *Fundamental Theorem of Calculus I:* for $f \in \mathcal{R}[a,b]$, $F(x) = \int_a^x f(t)\,dt$ is continuous on $[a,b]$; and if f is continuous at x_0, then $F'(x_0) = f(x_0)$. (9.4.3)
- **F72.** *Fundamental Theorem of Calculus II:* if $f \in \mathcal{R}[a,b]$ and $F' = f$ on $[a,b]$, then $\int_a^b f = F(b) - F(a)$. (9.4.4)
- **F73.** *Integration by parts:* if $F' = f \in \mathcal{R}[a,b]$ and $G' = g \in \mathcal{R}[a,b]$, then $\int_a^b Fg = F(b)G(b) - F(a)G(a) - \int_a^b fG$. (to be proved in Lecture 10)
- **F74.** *Change of variables:* if $\varphi : [A,B] \to [a,b]$ is differentiable and strictly increasing and $f \in \mathcal{R}[a,b]$, then $f \circ \varphi \in \mathcal{R}[A,B]$ and $\int_a^b f(x)\,dx = \int_A^B f(\varphi(y))\,\varphi'(y)\,dy$. (to be proved in Lecture 10)

Index of Objects

General Index