

Real Analysis

Noah Jussila

January 22, 2023

Contents

0 Introduction	6
0.1 Prerequisites	6
0.2 Organization and Sources	6
0.3 Presentation	7
I The Basics	8
1 The Real Numbers	8
1.1 Natural Numbers, Integers, and Rational Numbers	8
1.2 “Holes” in \mathbb{Q}	10
1.3 sup and inf	12
1.4 The Real Numbers	15
1.5 Properties of \mathbb{R}	19
1.6 Cardinality	20
1.7 Exercises	25
2 Point-Set Topology in Metric Spaces	27
2.1 Metric Spaces	27
2.2 Open Sets, Closed Sets, and Boundaries	29
2.3 Properties of Open and Closed Sets	36
2.4 Closures, Interiors, Dense Sets, and Perfect Sets	39
2.5 Compact Sets	40
2.6 Properties of Compact Sets	43
2.7 Compact Sets in \mathbb{R}^n	48
2.8 Exercises	53
3 Sequences and Series	55
3.1 Convergence	55
3.2 Properties Related to Convergence	58
3.3 Subsequences	64
3.4 Cauchy Sequences	66
3.5 Monotonic Sequences	71

3.6 Sequences and Infinity	74
3.7 \limsup and \liminf	74
3.8 Series	77
3.9 Tests for Convergent Series	78
3.10 Exercises	80
4 Continuity	81
4.1 Limits of Functions	81
4.2 Continuous Functions	84
4.3 Uniform Continuity	90
4.4 Continuity and Compactness	92
4.5 Intermediate Value Theorem	97
4.6 Discontinuities	99
4.7 Monotonicity	101
4.8 Exercises	105
5 Differentiation	106
5.1 The Definition of a Derivative	106
5.2 Familiar Properties of the Derivative	108
5.3 Local Extrema	113
5.4 Mean Value Theorems	115
5.5 L'Hôpital's Rule	117
5.6 Higher Order Derivatives	118
5.7 Approximation	119
5.8 Exercises	125
6 Riemann Integration	126
6.1 Partitions	126
6.2 Upper and Lower Riemann Integrals	127
6.3 An Alternative Interpretation: Simple Functions (Very Optional)	130
6.4 Verifying Riemann Integrability	136
6.5 Properties of Riemann Integration	140
6.6 Riemann Integration and Continuity	146
6.7 Riemann Integration, Monotonicity, and Discontinuities	150
6.8 The Riemann-Stieltjes Integral (Optional)	151
6.9 The Fundamental Theorem Of Calculus and Consequences	156
6.10 Change of Variables	160
6.11 Null Sets and Lebesgue's Criterion	163
6.12 Shortcomings of Riemann Integration	172
7 Sequences and Series of Functions	174
7.1 Metric Spaces of Functions	174
7.2 Pointwise Convergence	176
7.3 Uniform Convergence	180
7.4 Properties of Uniform Convergence	184

7.5	Approximation with Polynomials	189
7.6	Series	192
7.7	Power Series	196
7.8	Taylor Series	204
7.9	Arzelà–Ascoli Theorem	207
II	Higher Dimensions	209
8	Real Functions of Several Variables	209
8.1	Euclidean Space	210
8.2	Linear Transformations	215
8.3	Nonlinear Transformations and Functions	218
8.4	Continuity	222
8.5	The Space of Bounded Linear Transformations	224
9	Differentiation with Several Variables	232
9.1	The Derivative as a Linear Map	232
9.2	Partial Derivatives and the Jacobian	237
9.3	The Chain Rule	245
9.4	Mean Value Theorems	248
9.5	Clairaut's Theorem	251
9.6	The Inverse Function Theorem	256
9.7	The Implicit Function Theorem	266
10	Riemann Integration with Several Variables	274
10.1	Integration over a Rectangle	274
10.2	Riemann and Lebesgue's Criteria	280
10.3	Iterated Integrals and Fubini's Theorem	284
10.4	Jordan Content	293
10.5	Integration over General Regions	299
10.6	Change of Variables	301
10.7	Change of Variables, Proof	305
11	Manifolds in Euclidean Space	307
11.1	Motivation – Smooth Curves	307
11.2	Smooth Manifolds	307
11.3	Tangent Spaces	307
12	Multilinear Algebra and Differential Forms	307
12.1	Familiar Examples	307
12.2	Manifolds	307
12.3	Tensors	307
12.4	Wedge Product	307
12.5	Tangent Vectors and Differential Forms	307

12.6 Stokes' Theorem	307
III Measure Theory	308
13 Point-Set Topology Revisited	308
13.1 A Beautiful Day in the Neighborhood	308
13.2 Countability and Separation Axioms	317
13.3 Continuity	318
13.4 Nets	318
13.5 Filters	318
13.6 Various Notions of Compactness	318
13.7 The Product Topology	318
13.8 Pointwise and Uniform Convergence	318
14 Measures	318
14.1 Motivating Example and Problem Statement	318
14.2 Constructing Measures	321
14.3 σ -Algebras	328
14.4 Measures	328
14.5 Measures on \mathbb{R} , Lebesgue Measure	330
14.6 Translation Invariance	330
15 Integration Revisited	330
15.1 Measurable Functions	330
15.2 Integration of Simple Functions	330
15.3 Integration of Nonnegative Functions	330
15.4 Integration of Real Functions	330
15.5 Convergence Revisited	330
15.6 Product Measures	330
15.7 Lebesgue Integration in n -Dimensions	330
16 Differentiation with Measures	330
16.1 Absolute Continuity in \mathbb{R}	330
16.2 Signed Measures	330
16.3 Radon-Nikodym Derivative	330
IV Functional Analysis	330
17 Foundations	330
17.1 Normed Vector Spaces	330
17.2 Important Examples	330
17.3 Dual Spaces and Hahn-Banach	330
17.4 The Baire Category Theorem	332
17.5 Topological Vector Spaces	332

17.6 Hilbert Spaces	332
18 L^p Spaces	332
18.1 Basic Theory	332
18.2 The Dual of L^p	332
18.3 Inequalities	332
19 Riesz Representation Theorem	332
19.1 Continuous Functions with Compact Support	332
19.2 Radon Measures	332
19.3 Main Theorem	332
19.4 Related Results and Corollaries	332
20 Foundations of Fourier Analysis	332
20.1 Convolutions	332
21 Operator Theory	332
22 Distribution Theory	332
V Probability Theory	332
23 Introduction to Probability	332
23.1 Probability Spaces	332
23.2 Random Variables	332
23.3 Independence	332
23.4 Distributions Functions and Densities	332
23.5 Convergence	332
23.6 Expectation and Moments	332
23.7 Characteristic Functions	332
24 Laws of Large Numbers	332
25 Weak Convergence	332
26 Central Limit Theorems	332
27 Conditional Expectation and Martingales	332
28 Markov Chains	332
VI Stochastic Analysis	332
29 Stochastic Processes	332
30 Integration of Stochastic Processes	332

VII Theory of Optimization	332
31 Convexity	332
32 Optimizing Functionals	332
33 Constrained Optimization	332
34 Applications	332

0 Introduction

This collection of notes is an ongoing project that aims to combine four semesters of real analysis notes into one document. My goal in doing so is to not only stave off the boredom resulting from quarantining due to Covid-19, but also pick and choose from different presentations of standard material in real analysis.

0.1 Prerequisites

These notes will assume familiarity with differential and integral calculus, linear algebra, multivariable calculus, basic set theory, and properties of functions (image, preimage, injectivity, etc.) At times, some material, particularly concepts in linear algebra, will be reviewed. Two very good video series that serve as reviews for calculus and linear algebra can be found at this amazing [YouTube channel](#).

I also will assume familiarity with basic styles of proof (contradiction, contrapositive, etc.). I will try *very hard* to be as detailed as possible with every proof though.

0.2 Organization and Sources

If these notes are ever finished (which is unlikely), they will span about four semesters worth of real analysis: two at an honors undergraduate level, and two at a introductory PhD level.

My selection of sources is motivated not only by the books I really enjoy, but also those that are considered standard. For material presented in an undergraduate course, I absolutely love [Tao \(2016a\)](#) and [Tao \(2016b\)](#). Tao starts at the most basic level of arithmetic and numbers are works all the way up to Lebesgue Integration in his two part series. He motivates nearly every topic, provides examples, and gives the reader a big picture of the subject. That being said, his approach left me confused at times when I first read his books. In the first volume, Tao eschews point-set topology entirely and opts to work exclusively in \mathbb{R} . This is great because most people who learn real analysis only ever care about \mathbb{R} , but it also makes some concepts overly technical. For instance, I think that sequences and continuity are easier to present and motivate in general metric spaces. My hunch is that I'm in the minority here, as Terry Tao went the other direction, and he has one more Fields Medal than I do. Tao also doesn't include any figures. Another text that is awesome is the infamous "Baby" [Rudin \(1976\)](#). This book is so widely used for a reason.¹ It's concise, clear, and presents the material with no frills. It makes a great textbook if you have a professor who motivates the material, provides examples, and draws the occasional illustration. If you do not have such a professor, than [Rudin \(1976\)](#) becomes pure torture. This book is very much the outline of the first 7 sections of these notes. Sections 8-11 pull mostly from...

¹My uneducated hypothesis about this is that [Rudin \(1976\)](#) really only perfect makes sense if you already know all the

Standard Topics	Sections 1 - 7	Tao (2016a,b) , Rudin (1986)
Multivariable Calculus	Sections 8 - 12	
Measure Theory and Integration	Sections 13 - 15	
Functional Analysis	Sections 16-20	

0.3 Presentation

Many standard math textbooks assume the reader can create their own novel examples, fill in the blanks purposefully left in proofs, understand the motivation for the material, and pick up on the subtle “tricks” used in proofs. A good professor will help students do this during a lecture. I want my presentation to do this. In doing this, I’m very much inspired by a professor I had for a second semester honors analysis course I took at Boston College. His presentation of some of the material that will be treated not only will be replicated here, but also motivates how I approach other topics. This means lots of illustrations, footnotes, remarks, and asking hypothetical questions to motivate material. It may seem very pedantic if you’re familiar with analysis, but in that case just read Rudin’s books.

material well, so professors who do not remember what it is like to learn analysis just assume everyone understands this book.

Part I

The Basics

1 The Real Numbers

We begin by returning to the most basic concepts in math. What exactly is a number? We begin with the most basic possible set of numbers, and use those to define more complex sets of numbers, with our goal being to define the real numbers. Lastly, we will look at the “size” of these sets, and explore the concept of infinity.

1.1 Natural Numbers, Integers, and Rational Numbers

First, a cursory overview of several sets of numbers is in order. It is given for the sake of exposition, and to illustrate how we define sets of numbers using previously defined sets. For the sake of time, the formal definition of the standard operations (addition, multiplication, etc.) on these sets will be forgone. Rest assured that the operations we are all familiar with are well defined on these sets, and this can be shown rigorously. An excellent reference for this within the context of real analysis can be found in [Tao \(2016a\)](#), who takes nothing as given.

The most basic numbers are those we use to count. We will call these natural numbers.

Definition 1.1. Define the set $\mathbb{N} := \{0, 1, 2, \dots\}$ to be the *natural numbers*.

The operations of addition and multiplication are well defined on \mathbb{N} , in that when adding or multiplying natural numbers, the result is a natural number. A more succinct way of putting this is saying that \mathbb{N} is *closed* under addition and multiplication. While the natural numbers are great for things such as counting (a fact we will return to), it fails to be useful for much more. In particular, two basic operations we are familiar with are not well defined on \mathbb{N} .

Example 1.1. Suppose we want to find the difference in 2 and 5, both elements in \mathbb{N} . The difference in question would be $2 - 5$, but this is not an element of \mathbb{N} !

In order to address this shortcoming, we need to broaden our view. In effect, we need to “add more” numbers to \mathbb{N} . We want to enlarge the set of natural numbers by the amount necessary for subtraction to be well defined. This can be done by taking the set of all differences of natural numbers, giving the set of integers \mathbb{Z} . While it’s tempting to define this set as

$$\{a - b \mid (a, b) \in \mathbb{N}^2\},$$

this doesn’t account for multiple pairs $(a, b) \in \mathbb{N}^2$ giving the same integer. For example, $4 - 6$ and $8 - 10$ gives the same integer. To address this, we’ll define the equivalence relation

$$(a, b) \sim (c, d) \iff a + c = b + d$$

on the set \mathbb{N}^2 . If we want to write the set of differences between natural numbers such that each difference is its own unique element, regardless of the specific pair of natural numbers which give the difference, we use the notation

$$\{a - b \mid (a, b) \in \mathbb{N}^2\} / \sim .$$

The technical details of this type of construction come from group theory, and can be discarded at the moment,² but an informal explanation is useful. All equivalence relations form a partition, and in this case, each block of the partition corresponds to a single integer value. We want to consolidate the similar elements in each block, and “factor” out the rest of the structure, which is why the notation is reminiscent of division. We can do this with *any* equivalence relation, so let’s look at an example that only relies on basic arithmetic.

Example 1.2 (Modular Arithmetic). Suppose we have some integer (the formal definition of which follows this example) $a \in \mathbb{Z}$ which can be written as $a = kn + b$ for some $k, n, b \in \mathbb{Z}$. This is a fancy way of saying that we have a remainder of b when dividing a by n , which can be written as

$$a \equiv b \pmod{n}.$$

For simplicity, let $n = 3$. You may have seen the set of possible remainders given as

$$\mathbb{Z}/3\mathbb{Z} = \{0, 1, 2\}.$$

This is the same exact “factoring” which shrinks equivalence classes. Each element of $\mathbb{Z}/3\mathbb{Z}$ correspond to a block in a partition of \mathbb{Z} given by modular arithmetic for $n = 3$.

$$\begin{aligned} a \equiv 0 \pmod{3} &\iff a \in \{-3, 0, 3, 6, \dots\} \iff 0 \in \mathbb{Z}/3\mathbb{Z} \\ a \equiv 1 \pmod{3} &\iff a \in \{-2, 1, 4, 7, \dots\} \iff 1 \in \mathbb{Z}/3\mathbb{Z} \\ a \equiv 2 \pmod{3} &\iff a \in \{-1, 2, 5, 8, \dots\} \iff 2 \in \mathbb{Z}/3\mathbb{Z} \end{aligned}$$

Now let’s present the formal definition of the integers.

Definition 1.2. Define the set of *integers* as

$$\mathbb{Z} := \{a - b \mid (a, b) \in \mathbb{N}^2\} / \sim$$

for the equivalence relation

$$(a, b) \sim (c, d) \iff a + c = b + d$$

on \mathbb{N}^2 .

We will take the ordering of \mathbb{Z} , the negation of elements of \mathbb{Z} , and all arithmetic properties of \mathbb{Z} to be given. There are two things worth noting. The first is that $\mathbb{N} \subseteq \mathbb{Z}$. The identity element $0 \in \mathbb{N}$ gives $a - 0 = a$ for each $a \in \mathbb{N}$, so any natural number can be written as the difference of two natural numbers. Secondly, we defined \mathbb{Z} only by using \mathbb{N} . This is crucial, as we will define the rationals only by using \mathbb{Z} , and in turn define the real numbers only by using the rationals.

While subtraction is well defined for \mathbb{Z} , the same does not hold for division.

Example 1.3. Take the integers -3 and 6 , and suppose we are interested in the ratio of the prior to the latter. Obviously,

$$\frac{-3}{6} = -\frac{1}{2},$$

but this is not an element of \mathbb{Z} .

²For details, look up “quotient group”, and “quotient space”

We need to “extend” the integers to accommodate for division in a similar fashion to when we defined the integers using the natural numbers. This will give the set of rational numbers \mathbb{Q} , which, loosely speaking, are the set of all fractions formed from integers. Of course, we cannot divide by zero, so we want to only consider pairs of integers from the set $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$. We also need to account for the fact that multiple pairs of integers may give the same rational number, so we need to define an equivalence relation and “factor” it out like we did when defining \mathbb{Z} .

Definition 1.3. Define the set of *rational numbers* as

$$\mathbb{Q} := \{a/b \mid (a, b) \in \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})\} / \sim$$

for the equivalence relation

$$(a, b) \sim (c, d) \iff ac = bd$$

on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$.

Constructing one set by looking at all the fractions formed from another is not unique to the construction of \mathbb{Q} . This general idea gives the definition of a *field of fractions* in abstract algebra.

1.2 “Holes” in \mathbb{Q}

We now begin where the canonical [Rudin \(1976\)](#) opens. Our goal has been, and continues to be, to define the most comprehensive set of numbers possible. It may help to visualize what we have done so far with a number line. We can illustrate any “gaps” or “holes” by using red. This can be seen in Figure 1. Clearly

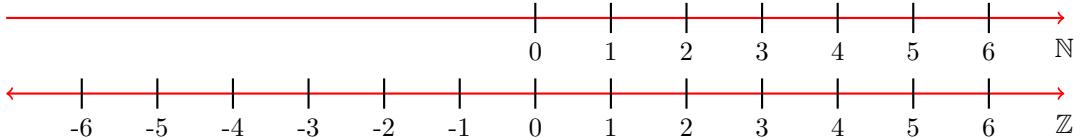


Figure 1: The natural numbers and integers ordered number lines.

the natural numbers and integers are not “comprehensive” in that they have many gaps. This is what led us to define the rational numbers \mathbb{Q} . It isn’t immediate just how well the rationals do at covering the holes in the integers. We can get a sense of this by introducing a property of the rationals.

Proposition 1.1. (Interspersing of integers by rationals) For any $x, y \in \mathbb{Q}$ where $x < y$, there exists a third rational number $z \in \mathbb{Q}$ such that $x < z < y$.

Proof. Let there be two rationals $x, y \in \mathbb{Q}$ such that $x < y$. We can define the third rational number of interest as $z = (x + y)/2$. We can show that $x < z < y$ by using arithmetic.

$$\begin{aligned} x &< y \\ \frac{x}{2} &< \frac{y}{2} \\ \frac{x}{2} + \frac{y}{2} &< \frac{y}{2} + \frac{y}{2} \\ z &< y \end{aligned}$$

And we can arrive at $x < z$ by adding $x/2$ to each side of the given inequality.

$$\begin{aligned} x &< y \\ \frac{x}{2} &< \frac{y}{2} \\ \frac{x}{2} + \frac{x}{2} &< \frac{y}{2} + \frac{x}{2} \\ x &< z \end{aligned}$$

□

Example 1.4. Take the rational numbers 0 and 1. Using the construction given in the previous proof we have

$$\frac{0+1}{2} = \frac{1}{2}$$

is between 0 and 1. We can now repeat this process using the pairs $(0, 1/2)$ and $(1/2, 1)$.

$$\begin{aligned} \frac{0+1/2}{2} &= \frac{1}{4} \\ \frac{1/2+1}{2} &= \frac{3}{4} \end{aligned}$$

We could repeat this process an infinite number of times, in effect “filling in” gaps in \mathbb{Z} by successively taking the average of two rational numbers. Figure 2 shows this process on the unit interval in the rationals.

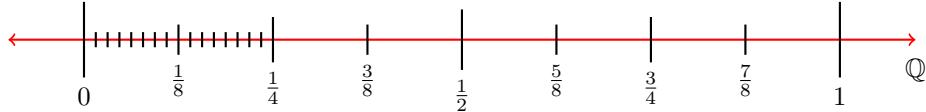


Figure 2:

The key question is whether or not this fills *all* the gaps in the integers.

While , and more formally , may lead us to believe that the rational numbers have no gaps, this is unfortunately not the case. There are two classic examples that arise from two of the most basic geometric constructions.

Example 1.5. Suppose we have a circle with diameter d and circumference c . In this case, the ratio give by c/d is not an element of the rational numbers. This familiar ratio is written as π . For the moment, we can take this as fact. We have not yet developed the tools required to proof that $\pi \notin \mathbb{Q}$, but we will return to this.

Example 1.6. Suppose there is an isosceles right triangle with legs of length 1, as shown in Figure 3. We want to find the length of the hypotenuse x .

This is a simple application of the Pythagorean Theorem.

$$\begin{aligned} 1^2 + 1^2 &= x^2 \\ 2 &= x^2 \end{aligned}$$

But this equation has no rational solution, something we can formally prove.

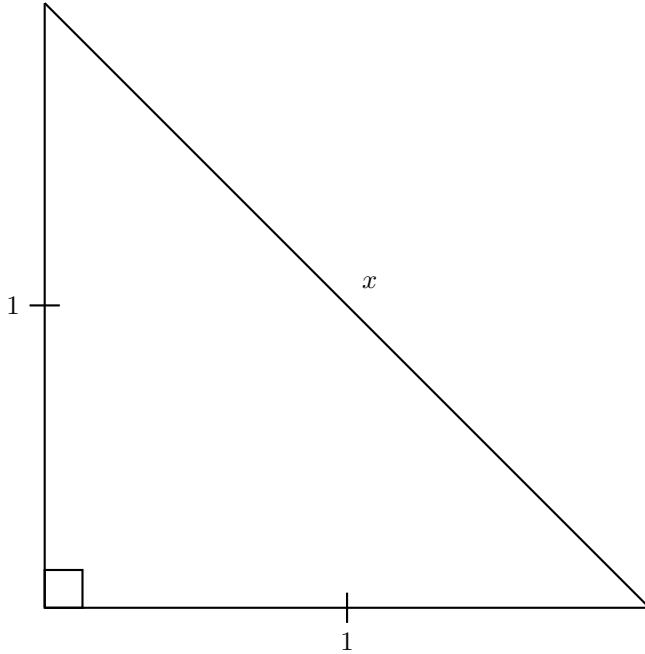


Figure 3:

Proposition 1.2. There exists no rational number x which satisfies $x^2 = 2$.

Proof. For the sake of contradiction, suppose that there exists a rational x which satisfies $x^2 = 2$. If this were the case, we could write $x = m/n$ for some $m, n \in \mathbb{Z}$, where m and n are not both even.³ \square

Any x which does satisfy $x^2 = 2$ would be *irrational*, in that it is not an element of \mathbb{Q} .

Definition 1.4. A number is *irrational* if it is not an element of \mathbb{Q} .

There are *many* irrational numbers, each of which is a gap in the rationals (see Figure 4).

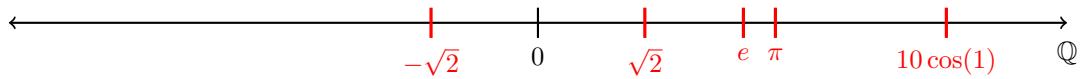


Figure 4:

Our goal now becomes defining a set of numbers that includes not only the rationals, but also all of the irrationals. We began with the natural numbers, and then defined a set \mathbb{Z} which included the additive inverses of the natural numbers. Then we filled more of the gaps in the integers by taking the ratios of integers. We are now faced with the task of defining a set which eliminates the gaps caused by irrational numbers, and doing so entirely with the set \mathbb{Q} .

1.3 sup and inf

Before informally constructing the real numbers, it is worth thinking about why \mathbb{Q} has these “holes”, and how it relates to a specific property of sets. It goes without saying that, all the sets of numbers we’ve discussed up until now have some ordering to them. We can make this formal by defining an ordered set.

³Otherwise we could write x in simpler terms as m and n would have a common factor of 2.

Definition 1.5. An *ordered set* is some set S with a binary relation, denoted by $<$,⁴ which satisfies the following properties:

1. If $x, y \in S$, then exactly one of the statements

$$x < y, \quad x = y, \quad y < x$$

is true.

2. If $x, y, z \in S$, and both $x < y$ and $y < z$, then $x < z$.

The statement “ $x < y$ ” is read as “ x is less than y .” We could also write $y > x$ instead of $x < y$. If we were to negative $y < x$ (“ y is not less than x ”), we would arrive at “ y is either greater than x or equal to x .” This is denoted as $y \geq x$.

Example 1.7. The set \mathbb{Q} is a well ordered set if we define $<$ in the following way for $x, y \in \mathbb{Q}$:

$$x < y := y - x \text{ is a positive rational number.}$$

Note that we can only relate objects that belong to \mathbb{Q} . This means that we have no way of comparing rational numbers and the solution to the equation $x^2 = 2$.

We can use the order relation on an ordered set to define bounds on sets.

Definition 1.6. Suppose S is an ordered set, and $E \subseteq S$. If there exists a $\beta \in S$ such that $x \leq \beta$ for all $x \in E$, E is *bounded above*, and β is an *upper bound* of E .

Definition 1.7. Suppose S is an ordered set, and $E \subseteq S$. If there exists a $\beta \in S$ such that $x \geq \beta$ for all $x \in E$, E is *bounded below*, and β is a *lower bound* of E .

A subtlety in both definitions that is extremely important, is that upper and lower bounds must be elements of the ordered set S . The next example highlights this.

Example 1.8. Take \mathbb{Z} to be an ordered set with the natural order. Pick the subset $E = \{-2, -1, 2\} \subseteq \mathbb{Z}$. This set has many upper and lower bounds. For upper bounds we have $2, 3, 4, \dots$. For lower bounds we have $-2, -3, -4, \dots$. It may be tempting to say that a fraction such as $5/2$ is an upper bound of E , but it is not. This follows from the fact that $5/2 \notin \mathbb{Z}$, so we have no means of relating it to elements in \mathbb{Z} . In this particular case, there are upper and lower bounds of the set are included in the set. This need not be the case, as the next example shows.

Example 1.9. Let’s look at the ordered set \mathbb{Q} , and subset $E = \{x \in \mathbb{Q} \mid 0 < x < 1\} \subseteq \mathbb{Q}$.⁵ In this case, each element of $\{x \in \mathbb{Q} \mid x \leq 0\}$ is a lower bound of E , and each element of $\{x \in \mathbb{Q} \mid x \geq 1\}$ is an upper bound of E . Even though $0, 1 \notin E$, they are still least and upper bounds of E respectively.

Remark 1.1. It is often obvious what exact order we are talking about when referring to an ordered set, like in the case of \mathbb{Q} and \mathbb{Z} . In these cases, we’ll just assume we’re using the natural order.

We now will introduce two definitions that correspond to a special type of upper and lower bound.

⁴In this case, “ $<$ ” can mean *any* order. It just so happens that we use the same symbol as the familiar “less than” order, because it is the canonical example of such a relation.

⁵The use of the familiar interval notation of $(0, 1)$ will be properly defined and restricted to the real numbers in the following section.

Definition 1.8. Suppose S is an ordered set, $E \subseteq S$, and E is bounded above. We say that α is a *least-upper-bound* of E if:

1. α is an upper bound of E .
2. If $\gamma < \alpha$, then γ is not an upper bound of E .

Alternatively, we can refer to α as the *supremum* of E , and write $\alpha = \sup E$

Definition 1.9. Suppose S is an ordered set, $E \subseteq S$, and E is bounded below. We say that α is a *greatest-lower-bound* of E if:

1. α is a lower bound of E .
2. If $\gamma > \alpha$, then γ is not a lower bound of E .

Alternatively, we can refer to α as the *infimum* of E , and write $\alpha = \inf E$

Remark 1.2. Both definitions use the definite article *the* before supremum and infimum. This is because they are unique. This is also implied by the use of the superlative *least* and *greatest*. Nevertheless, this is a result of the definition, and can be properly proven.

Example 1.10. If we return to Example 1.8, where $E = \{-2, -1, 2\} \subseteq \mathbb{Z}$, we have $\sup E = 2$ and $\inf E = -2$.

Example 1.11. In example 1.9, $\inf E = 0$ and $\sup E = 1$.

Example 1.12. Sticking with the set \mathbb{Q} , consider the subset $E = \{x \in \mathbb{Q} \mid x^2 \leq 2\} \subseteq \mathbb{Q}$. This set has no supremum, because the number satisfying $x^2 = 2$ is not an element of \mathbb{Q} (as shown in Proposition 1.2). We will formally prove this fact shortly.

It is no coincidence that a subset of \mathbb{Q} fails to have a supremum, because of one of the “holes” in \mathbb{Q} . The following definition will help us formalize this relationship.

Definition 1.10. Let S be an ordered set. If for all $E \subseteq S$, where E is nonempty and bounded from above, $\sup E$ exists, then S has the *least-upper-bound* property.

The least-upper-bound property ensures that any nontrivial subset of an ordered set has a supremum in that ordered set. We could define an equivalent property known as the greatest-lower-bound property. The next two propositions serve as nice examples of the least-upper-bound property, or lack thereof, in action.

Proposition 1.3. The set \mathbb{Z} has the least-upper-bound property.

The idea behind the following proof takes advantage of the fact that \mathbb{Z} is discrete. For some set $E \subseteq \mathbb{Z}$, we can always just look at an upper bound of it, and keep subtracting 1 until the resulting number is in E . Then we will have found our upper bound.

Proof. We will show that an arbitrary nontrivial set $E \subseteq \mathbb{Z}$ has a supremum. Let $x \in E$, and β be an upper bound of E . We know that $\beta \geq x$ for all $x \in E$. We can show that $\sup E$ exists via induction on $\beta - x$ for our arbitrary $x \in E$. Our base case is when $\beta - x = 0$. If this holds, then $\beta \in E$, so $\beta \in \mathbb{Z}$ and $\sup E = \beta$. Now suppose that this statement holds when $\beta - x = k$ for $k \in \mathbb{N}$ (this is our induction hypothesis). It is either the case that $\beta \in E$ or $\beta \notin E$. If $\beta \in E$, then $\sup E = \beta$. If $\beta \notin E$, then let $\beta' = \beta - 1$. Then β' is an upper bound of E , and

$$\beta' - x = \beta - 1 - x = \beta - x - 1 = k + 1 = k.$$

By the induction hypothesis, $\sup E$ exists. \square

Proposition 1.4. The set \mathbb{Q} does not have the least-upper-bound property.

To prove this, we will first establish that 2 is an upper bound of the set defined in Example 1.12, and then show the set has no supremum via contradiction.

Proof. It suffices to find a single subset of \mathbb{Q} which fails to have a supremum. Let that set be $E = \{x \in \mathbb{Q} \mid x^2 \leq 2\}$.

1. Suppose for contradiction that 2 is not an upper-bound of E . Then there exists an $x \in E$ such that $x > 2$. This would imply that $x^2 > 4$, which contradicts the assumption that $x \in E$.
2. Suppose for contradiction that E has a supremum, and that $\sup E = \alpha$ for $\alpha \in \mathbb{Q}$. Define a new rational number $y \in \mathbb{Q}$ as

$$y = \alpha - \frac{\alpha^2 - 2}{x + 2} = \frac{2(\alpha + 1)}{\alpha + 2}. \quad (1)$$

Squaring this and subtracting 2 gives

$$y^2 - 2 = \frac{4(\alpha + 1)^2}{(\alpha + 2)^2} - \frac{2(\alpha + 2)^2}{(\alpha + 2)^2} = \frac{2(\alpha^2 - 2)}{(\alpha + 2)^2}. \quad (2)$$

We can use y to reach a contradiction in each possible case, those being: $\alpha^2 < 2$, $\alpha^2 = 2$, $\alpha^2 > 2$.

- (a) Suppose that $\alpha^2 < 2$. This means that $\alpha^2 - 2 < 0$, so Equation (1) implies that $y > \alpha$. At the same time, Equation (2) implies that $y^2 - 2 < 0$, which means $y^2 < 2$. This gives that $y \in E$, despite the fact that $\alpha < y$. This contradicts the fact that α is an upper-bound of E .
- (b) Suppose $\alpha^2 = 2$. We already know this cannot be the case by Proposition 1.2.
- (c) Finally, assume that $\alpha^2 > 2$, giving $\alpha^2 - 2 = 0$. Equation (1) implies $y < \alpha$ while Equation (2) implies $y^2 - 2 > 0$, meaning $y^2 > 2$. This establishes y as an upper bound for E , but $y < \alpha$, which contradicts $\sup E = \alpha$.

□

The “holes” in \mathbb{Q} are a result of \mathbb{Q} not having the least-upper-bound property. In order to perform calculus, we need a set that has this property, otherwise things like continuity and differentiation would not work. This property is not sufficient in and of itself though. If that were the case then we would have stopped extending our set of numbers at \mathbb{Z} . We want a set of numbers as “comprehensive” as \mathbb{Q} , but with the least-upper-bound property. It turns out, that this (and a whole lot more) is what we will get from the real numbers.

1.4 The Real Numbers

We will now construct the real numbers using only \mathbb{Q} . First, we will define the algebraic structure that the real numbers will take on.

Definition 1.11. A *field* is a set F with two operations, addition and multiplication, which satisfy the following axioms for all $x, y, z \in F$:

1. Axioms for addition:
 - (a) $x + y \in F$ (closed under addition)

- (b) $x + y = y + x$ (commutative)
- (c) $(x + y) + z = x + (y + z)$ (associative)
- (d) There exists an element $0 \in F$ such that $0 + x = x$ (identity element)
- (e) There exists an element $-x \in F$ such that $x + (-x) = 0$ (inverse element)

2. Axioms for multiplication:

- (a) $xy \in F$ (closed under multiplication)
- (b) $xy = yx$ (commutative)
- (c) $(xy)z = x(yz)$ (associative)
- (d) There exists an element $1 \in F$ such that $1x = x$ (identity element)
- (e) If $x \neq 0$, there exists an element $1/x \in F$ such that $x(1/x) = 1$ (inverse element)

3. The distributive property:

$$x(y + z) = xy + xz$$

The study of fields is its own entire subject in math, and lives within the discipline of abstract algebra. For more details about fields, see [Dummit and Foote \(2004\)](#). These axioms can be used to reach several familiar conclusions about arithmetic in fields, and can be found as formal propositions in [Rudin \(1976\)](#). A more specific type of field is that which is also an ordered set.

Definition 1.12. An *ordered field* is a field F such that

- 1. $x + y < x + z$ if $x, y, z \in F$ and $y < z$,
- 2. $xy > 0$ if $x, y \in F$, $x > 0$, and $y > 0$.

Example 1.13. The set \mathbb{Q} is an ordered field.

Our goal is now to construct an ordered field which not only contains \mathbb{Q} , but also has the least-upper-bound property. In order to do this we'll use the fact that \mathbb{Q} has “holes” in it. We'll form a pair of sets (A, B) that partition \mathbb{Q} such that each of these partitions corresponds to a real number.

Definition 1.13. A *Dedekind cut* $x = (A, B)$ is a pair of subsets $A, B \subseteq \mathbb{Q}$ satisfying the following:

- 1. $A \cup B = \mathbb{Q}$, $A \cap B = \emptyset$, $A \neq \emptyset$, and $B \neq \emptyset$.
- 2. If $a \in A$ and $b \in B$, then $a < b$.
- 3. A contains no largest element.

Example 1.14. Let $A = \{y \in \mathbb{Q} \mid y < 0\}$ and $B = \{y \in \mathbb{Q} \mid y \geq 0\}$. Our cut is $x = (A, B)$, and can be seen in Figure 5. This cut uniquely represents $0 \in \mathbb{Q}$, as no other cut can be defined in this way “at” 0.

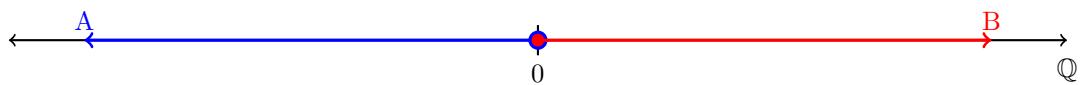


Figure 5: Dedekind cut corresponding to $0 \in \mathbb{Q}$.

Example 1.15. Perhaps a better example is the cut defined by $A = \{q \in \mathbb{Q} \mid q \leq 0 \text{ or } q^2 < 2\}$ and $B = \{q \in \mathbb{Q} \mid q > 0 \text{ or } q^2 > 2\}$. This cut corresponds to the solution of the equation $x^2 = 2$.

Definition 1.14. A *real number* is a Dedekind cut in \mathbb{Q} . The set of real numbers is denoted by \mathbb{R} .

Definition 1.15. A real number $x = (A, B)$ is a *rational number* if B contains a smallest element (namely x).

Definition 1.16. A real number $x = (A, B)$ is a *irrational number* if B contains no smallest element.

Example 1.16. The cut defined by $A = \{y \in \mathbb{Q} \mid y < 0\}$ and $B = \{y \in \mathbb{Q} \mid y \geq 0\}$ is rational, as B has a smallest element in the form of 0.

Example 1.17. The cut defined by $A = \{q \in \mathbb{Q} \mid q \leq 0 \text{ or } q^2 < 2\}$ and $B = \{q \in \mathbb{Q} \mid q > 0 \text{ or } q^2 > 2\}$ has no smallest element. Therefore it is an irrational number. We will denote this particular number as $\sqrt{2}$.

Now that we have properly defined \mathbb{R} , we can *finally* refer to the quantity $\sqrt{2}$! It is no longer a mysterious solution to an equation, and is well defined in \mathbb{R} . This is a relatively small payoff, but the real rewards are the two following theorems. These are the main results of this section, and are of the utmost importance. The first will allow us to perform operations on \mathbb{R} , and the second will play a key role in proving familiar theorems from calculus. The proof of the first result is rather long, and not very informative, so it is omitted. It is important to understand *how* it would be proved though. Before the big reveal, we will define an order on \mathbb{R}

Definition 1.17. Given real numbers $x = (A, B)$ and $y = (C, D)$, we define the following order:

$$x \leq y := A \subseteq C.$$

The inequality is strict if $A \subsetneq C$.

Example 1.18. Let $x = 2 = (A, B) = (\{y \in \mathbb{Q} \mid y < 2\}, \{y \in \mathbb{Q} \mid y \geq 2\})$, and $y = 3 = (C, D) = (\{z \in \mathbb{Q} \mid z < 3\}, \{z \in \mathbb{Q} \mid z \geq 3\})$. It should come as no surprise that $2 < 3$, but this is because $A \subseteq C$.

Theorem 1.1. The set \mathbb{R} is an ordered field containing \mathbb{Q} .

Proof. See the appendix of chapter 1 in [Rudin \(1976\)](#). The idea is that addition and multiplication of cuts must be defined, and all the axioms of fields and ordered fields must be verified using the cut definition of a real number. \square

Theorem 1.2 (Completeness of the real numbers). The set \mathbb{R} has the least-upper-bound property.

Proof. We will show an arbitrary nonempty subset of \mathbb{R} has a supremum. Let $E \subseteq \mathbb{R}$, where $E \neq \emptyset$, have the upper bound $\beta \in \mathbb{R}$. We may write β as a Dedekind cut, $\beta = (A, B)$. Additionally, we may express each $\alpha \in E$ as a cut $\alpha = (L_\alpha, U_\alpha)$. Now we will construct a real number by taking the union of all L_α .

$$\gamma = \left(\bigcup_{\alpha \in E} L_\alpha, \mathbb{Q} \setminus \bigcup_{\alpha \in E} L_\alpha \right) = (L, \mathbb{Q} \setminus L) = (L, U)$$

I claim that $\sup E = \gamma$.

First we must verify that $\gamma \in \mathbb{R}$ by showing that (L, U) is a valid Dedekind cut, and satisfies the requirements of [Definition 1.13](#):

1. The set E is nonempty, so there exists at least one $\alpha = (L_\alpha, U_\alpha) \in E$. Because $L_\alpha \neq \emptyset$ and $U_\alpha \neq \emptyset$ by definition 1.13, we have $L \neq \emptyset$ and $U \neq \emptyset$. By the definition of U as $\mathbb{Q} \setminus L$, we have that $L \cup U = \mathbb{Q}$ and $L \cap U = \emptyset$. Therefore $\gamma \in \mathbb{R}$. By construction, $\alpha \leq \gamma$ for all $\alpha \in E$, making γ an upper bound.
2. To show that it is the least-upper-bound, we will now show any number lesser than it cannot be an upper bound. Now suppose $\delta < \gamma$. This means $C \subseteq L$, where δ is expressed as a cut $\delta = (C, D)$. This means there exists some $s \in L$ such that $s \notin C$. But $s \in L$, so it is in L_α for some $\alpha \in E$. Hence, $C \subseteq L_\alpha$, giving $\delta < \alpha$. This shows that δ is not an upper bound, meaning $\sup E = \gamma$.

□

Example 1.19. Consider the set of real numbers $E = \{-1, -1/2, -1/3, -1/4, \dots\}$. What is the supremum of this set? Intuitively, it should be 0, but we can verify this by constructing it like we did in the previous proof. Each number in E corresponds to a cut (L_n, U_n) , for $L_n = \{x \in \mathbb{Q} \mid x < -1/n\}$ and $U_n = \{x \in \mathbb{Q} \mid x \geq -1/n\}$, where $n \in \mathbb{N}$. Our supremum is

$$\gamma = \left(\bigcup_{n \in \mathbb{N}} \{x \in \mathbb{Q} \mid x < -1/n\}, \mathbb{Q} \setminus \bigcup_{n \in \mathbb{N}} \{x \in \mathbb{Q} \mid x < -1/n\} \right) = (\{x \in \mathbb{Q} \mid x < 0\}, \{x \in \mathbb{Q} \mid x \geq 0\}).$$

Therefore, $\gamma = 0$.

You will often here the real numbers referred to as “complete” because they have the least-upper-bound property. This is because the least-upper-bound property ensures there are no “gaps” in the real line like there are in \mathbb{Q} . Theorem 1.2 may be the most important result in real analysis. Remember it well, as it will be used often. Most disciplines in math build on themselves over time, and the fact that \mathbb{R} has the least-upper-bound property will be our foundation. One could argue it is \mathbb{R} ’s defining property.

Finally, we will adopt the familiar notation of intervals in \mathbb{R} , and add make an important addition to \mathbb{R} .

Definition 1.18. We will use the following notation to refer to *intervals* of \mathbb{R} :

$$\begin{aligned}(a, b) &= \{x \in \mathbb{R} \mid a < x < b\} \\ [a, b) &= \{x \in \mathbb{R} \mid a \leq x < b\} \\ (a, b] &= \{x \in \mathbb{R} \mid a < x \leq b\} \\ [a, b] &= \{x \in \mathbb{R} \mid a \leq x \leq b\}\end{aligned}$$

for $a < b$.

Definition 1.19. The *extended real number system* consists of the real field \mathbb{R} and two symbols: ∞ , and $-\infty$. The original order of \mathbb{R} is preserved, and we define

$$-\infty < x < \infty$$

for all $x \in \mathbb{R}$. We will denote the extended real numbers as $\overline{\mathbb{R}}$.

The extended real numbers do not form a proper field, but we can adopt some conventions for arithmetic using ∞ and $-\infty$ for $x \in \mathbb{R}$:

1. $x + \infty = \infty$, $x - \infty = -\infty$, $x/\infty = x/-\infty = 0$.
2. For $x > 0$, $x(\infty) = \infty$, and $x(-\infty) = -\infty$.

3. For $x < 0$, $x(\infty) = -\infty$, and $x(-\infty) = \infty$.

The addition of an upper and lower bound on \mathbb{R} make the set $\overline{\mathbb{R}}$ easier to work with in certain situations.⁶ The most immediate result of working in $\overline{\mathbb{R}}$ is that *every* subset of \mathbb{R} has a supremum, not just bounded ones (the latter case being the only one stipulated by the least-upper-bound property).

1.5 Properties of \mathbb{R}

The importance of Theorem 1.2 can not be understated. It is perhaps *the* defining property of \mathbb{R} , and it gives rise to numerous results in analysis. For now, we can use it to prove two additional properties of \mathbb{R} .

Theorem 1.3 (Archimedean property of \mathbb{R}). For $x, y \in \mathbb{R}$ where $x > 0$, there exists an $n \in \mathbb{N}$ such that $nx > y$.

Proof. Let A be the set of all nx for $x \in \mathbb{R}$ and $n \in \mathbb{N}$, where $x > 0$. For contradiction, suppose that there exists no such $n \in \mathbb{N}$ such that $nx > y$ for $y \in \mathbb{R}$. This makes y an upper bound of A . By the completeness of \mathbb{R} , $\sup A = \alpha$ exists. Since $x > 0$, $\alpha - x < \alpha$, and $\alpha - x$ is not an upper bound of A . This means there exists an $m \in \mathbb{N}$ such that $\alpha - x < mx$. But this would imply $\alpha < mx + x = m(x + 1)$, where $(m + 1)x \in A$. This contradicts the fact that α is an upper bound of A . \square

Example 1.20. Suppose $x = 10$ and $y = 213$. By the Archimedean property of \mathbb{R} , we know there exists a multiple of 10 that is greater than 213.

$$10(22) = 220 > 213$$

Theorem 1.4 (\mathbb{Q} is dense in \mathbb{R}). For $x, y \in \mathbb{R}$ where $x < y$, there exists a $p \in \mathbb{Q}$ such that $x < p < y$.

Proof. We have $x < y$, giving $y - x > 0$. By the Archimedean property (Theorem 1.3), there exists an $n \in \mathbb{N}$ such that

$$n(y - x) > 1. \tag{3}$$

We can use Theorem 1.3 to find $m_1, m_2 \in \mathbb{N}$ for which:

$$\begin{aligned} m_1 &> nx, \\ m_2 &> -nx. \end{aligned}$$

We can combine these two inequalities to conclude $-m_2 < nx < m_1$. This implies there exists an $m \in \mathbb{N}$ (with $-m_2 \leq m \leq m_1$) such that

$$m - 1 \leq nx < m. \tag{4}$$

If we combine (3) and (4) we get

$$nx < m \leq 1 + nx < ny.$$

Dividing by n (which is positive) gives $x < \frac{m}{n} < y$. \square

The density of \mathbb{Q} in \mathbb{R} is both surprising and useful for constructing examples. In practice, it means that every irrational number has a rational number arbitrarily close to it. We can approximate any number in \mathbb{R} arbitrarily well with a rational number.

⁶What we are really doing is working with $\overline{\mathbb{R}}$ because it is a complete lattice. A complete lattice is a partially ordered set in which every subset has an infimum or supremum. The real line is not a complete lattice, as any set of the form $(a, \infty) \subset \mathbb{R}$ has no supremum.

Example 1.21. We will now mimic the proof of Theorem 2.4 with actual numbers. We will find a rational $p \in \mathbb{Q}$ such that $e < p < \pi$.⁷ We have $\pi - e > 0$. We know

$$3(\pi - e) > 1.$$

Next we pick whole numbers $m_1 = 9$ and $m_2 = 8$, and get the inequality $8 < 3e < 9$. Now take m to be 9 and reach our final inequality of

$$3e < 9 < 1 + 3e < 3\pi.$$

Dividing by $n = 3$ gives our desired rational number is $9/3 = 3$.

Example 1.22. Let $\sqrt{2} \in \mathbb{R}$, and let $\varepsilon > 0$.⁸ By the density of \mathbb{Q} in \mathbb{R} , there exists $p \in \mathbb{Q}$ such that $\sqrt{2} - \varepsilon < p < \sqrt{2}$. This will hold for all $\varepsilon > 0$ so as we let ε become smaller and smaller, we will have an increasingly accurate rational approximation of $\sqrt{2}$.

1.6 Cardinality

So far, I've been intentional in avoiding any discussion of the size of the sets we have been working with. When constructing \mathbb{Z} from \mathbb{N} , it was never stated that \mathbb{Z} was somehow "bigger" than \mathbb{N} . All we know is that \mathbb{Z} has elements that \mathbb{N} does not. The same can be said for \mathbb{Z} and \mathbb{Q} , or \mathbb{Q} and \mathbb{R} . It is now time that we turn our attention to this matter, and more generally the size, or "cardinality" of sets.

Determining the size of a set amounts to counting the number of elements in that set. But how do we make the notion of counting formal? We will do this with functions. Before formally defining anything, consider how you may count something. If you are tasked with counting the number of elements in the set $X = \{a, b, c\}$, your answer will surely be 3. How did you get that number? You said assigned the number 1 to a , 2 to b , and 3 to c . We should note three different things about this process:

1. Each number we use is from \mathbb{N} .
2. Each element of X is assigned a number. We wouldn't have counted properly if we skipped some element.
3. No number in \mathbb{N} is assigned to multiple elements in X . We do not want to count multiple elements as a single element.

This process of assigning elements in \mathbb{N} to those in X is shockingly similar to the notion of a function, as we are mapping elements from one set to those in another set. Furthermore, the properties we must obey while counting have their own analogous forms with functions: surjectivity, and injectivity. For this reason, we will use functions to formalize the size of a set.

First, we will address the abstract case of when two sets have the same number of elements, and then we will transition to the size of sets.

Definition 1.20. The *cardinality* of a set X , denoted $|X|$, is the number of elements that belong to the set.

Definition 1.21. Two sets X and Y have *the same cardinal number* if there exists a bijection $f : X \rightarrow Y$ from X to Y .

⁷You shouldn't even need to perform the construction to arrive at an answer, as e and π are not particularly "close" to each other. We could for instance take $p = 3$.

⁸This is the first time we're using the infamous ε . It just stands in for any arbitrarily small positive number.

Proposition 1.5. Define the relation $X \sim Y$ only if X and Y have the same cardinal number ($|X| = |Y|$). The relation \sim is an equivalence relation.

Proof. We have that $X \sim X$ by letting $f : X \rightarrow X$ be $f(x) = x$, so \sim is reflexive. If $X \sim Y$, there exists a bijection $f : X \rightarrow Y$. Since f is a bijection, it has an inverse $f^{-1} : Y \rightarrow X$. This inverse is itself a bijection, so $Y \sim X$, making \sim symmetric. Lastly, assume $X \sim Y$ and $Y \sim Z$. We have bijections $f : X \rightarrow Y$ and $g : Y \rightarrow Z$. The composition of two bijections is a bijection, so $h : X \rightarrow Z$ is a bijection. This makes \sim transitive. \square

Example 1.23. Let $X = \{1, 2, 3\}$ and $Y = \{\sqrt{2}, e, \pi\}$. We can define $f : X \rightarrow Y$ as

$$f(x) = \begin{cases} \sqrt{2} & \text{if } x = 1 \\ e & \text{if } x = 2 \\ \pi & \text{if } x = 3 \end{cases}.$$

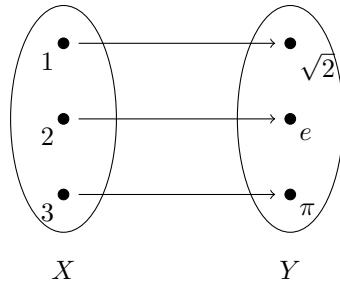


Figure 6: Bijection $f : X \rightarrow Y$.

This function is clearly a bijection, and we have that $|X| = |Y|$.

Example 1.24. Let $\mathbb{Z}^- = \{x \in \mathbb{Z} \mid x < 0\}$ and $\mathbb{Z}^+ = \{x \in \mathbb{Z} \mid x > 0\}$. There exists a very natural bijection between these sets, namely that which maps each element in \mathbb{Z}^+ to its negative counterpart in \mathbb{Z}^- . Formally, $f : \mathbb{Z}^+ \rightarrow \mathbb{Z}^-$ is defined as $f(x) = -x$. This function is clearly a bijection, and its existence shows that $|\mathbb{Z}^+| = |\mathbb{Z}^-|$. There are the same number of positive integers as negative integers.

Remark 1.3. Because $f : X \rightarrow Y$ is a bijection, it doesn't matter which set is the domain and which set is the codomain. A function is invertible if and only if it is a bijection, and a functions inverse is a bijection, so we would just have $f^{-1} : Y \rightarrow X$ if we picked the sets in the other order. In example 2.23, we could have instead assigned each negative integer to its positive counterpart and had $f : \mathbb{Z}^- \rightarrow \mathbb{Z}^+$. In this case we would still have $f(x) = -x$, as this particular function is its own inverse!

As discussed earlier, counting is intrinsically linked to the set of natural numbers \mathbb{N} . We will now make this formal by defining three types of sets: finite sets, countably infinite sets, and uncountably infinite sets.

Definition 1.22. A set X is *finite* if there exists a subset of the whole numbers $N \subseteq \mathbb{N}$ for which X and N have the same cardinal number.

Definition 1.23. A set X is *countably infinite* if X has the same cardinal number as \mathbb{N} . Alternatively, X is countably infinite if there exists a bijection $f : X \rightarrow \mathbb{N}$. This is sometimes denoted as $|X| = \aleph_0$.⁹

⁹This symbol is an “aleph”, and is the first letter of the Hebrew alphabet.

Definition 1.24. A set X is *countable* if it is countably infinite or finite.

Remark 1.4. From here on out, I'm going to use countable and countably infinite interchangeably, as nearly all the sets we are interested in are infinite.

Definition 1.25. A set X is *uncountably infinite* (or uncountable) if it is neither finite nor countably infinite.

Before jumping into examples, let's unpack some of this. Finiteness and countable infiniteness depend on whether a set has the same cardinal number as a subset of \mathbb{N} or \mathbb{N} itself. This means we can find a bijection between the set and a subset of \mathbb{N} or \mathbb{N} itself. [Definition 1.22](#) and [Definition 1.23](#) are often presented in terms of this hypothetical bijection. Secondly, two of these definitions involve infinity. A set can be either countably infinite or uncountably infinite. In a sense, some infinite sets have so many elements that they cannot even be counted, and are “bigger” than other uncountable sets! These two concepts of infinity will show up constantly in real analysis. Hopefully examples will make this clear. Some of the following examples are so important that they will be presented as formal results.

Example 1.25. We will modify Example 2.22. Let $X = \{\sqrt{2}, e, \pi\}$ and $N = \{1, 2, 3\} \subseteq \mathbb{N}$. Take our bijection to be the inverse of the function defined previously in Example 2.22. This shows that X is finite, and $|X| = 3$.

Proposition 1.6. The set \mathbb{N} is countably infinite.

Proof. Define $f : \mathbb{N} \rightarrow \mathbb{N}$ as $f(n) = n$. This function is clearly a bijection. \square

Proposition 1.7. The set \mathbb{Z} is countably infinite.

Before proving this result, it's worth acknowledging that it seem paradoxical. How can it be that \mathbb{Z} is the same size as \mathbb{N} , despite \mathbb{Z} being defined as \mathbb{N} plus more elements? It would make more sense for \mathbb{Z} to have twice the cardinality as \mathbb{N} , but this is not the case. The result may sit better if you consider just how we would count \mathbb{Z} . If you started at $1 \in \mathbb{Z}$, followed by $2 \in \mathbb{Z}$, etc. then you would miss all the negative numbers! Instead, we need to be clever in the order in which we count \mathbb{Z} . We will instead count in the following order: $0, 1, -1, 2, -2, \dots$. We could count like this forever and never run out of \mathbb{N} , and never miss any elements of \mathbb{Z} . This is why we have $|\mathbb{N}| = |\mathbb{Z}|$.

Proof. Recall that we can select either \mathbb{Z} or \mathbb{N} to be the domain of our bijection. We will define $f : \mathbb{N} \rightarrow \mathbb{Z}$ as

$$f(n) = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ -\frac{n-1}{2} & \text{if } n \text{ is odd} \end{cases}.$$

This function counts \mathbb{Z} in the aforementioned manner of alternating between positive and negative integers. We now will verify that f is a bijection, by showing it is injective and surjective.¹⁰ Let $y \in \mathbb{Z}$, and pick $x \in \mathbb{N}$ such that $x = 2y$ if y is even, and $x = -2y + 1$ if y is odd. This choice of $x \in \mathbb{N}$ gives $f(x) = y$, so f is surjective. Now suppose that $f(x_1) = f(x_2)$. If $x_1/2 = x_2/2$, then $x_1 = x_2$. If $-(x_1 - 1)/2 = -(x_2 - 1)/2$, then $x_1 = x_2$. Therefore, f is injective. \square

An even more surprising result is that not only $|\mathbb{N}| = |\mathbb{Q}|$, but also $|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}|$! Even if we add every possible fraction to the integers, the size of our set remains the same.

¹⁰It would be quicker to show f has an inverse, but that approach is not as instructive.

Theorem 1.5. The set \mathbb{Q} is countably infinite.

Proof. We can enumerate the rational numbers in the following way:

$$\frac{0}{1}, \frac{1}{1}, \frac{-1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{2}{1}, \frac{-2}{1}, \frac{-1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{-1}{4}, \frac{2}{3}, \dots$$

This particular ordering can be seen in Figure 7. The red arrows in Figure 7 show the order in which we count, and it becomes evident that we will eventually count every possible fraction. Note that we only count fractions which are expressed in simplest terms, with others in gray.

□

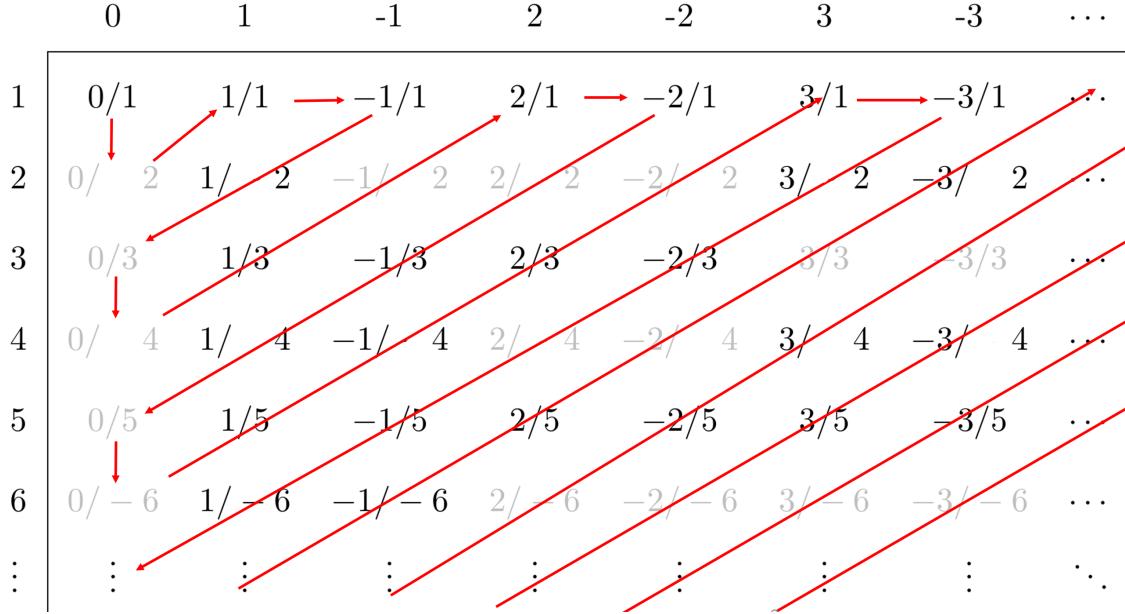


Figure 7:

It may not be a surprise that it is not possible to count \mathbb{R} . This puts \mathbb{R} in our second category of infinite sets, uncountably infinite.

Theorem 1.6. The set \mathbb{R} is uncountably infinite.

The proof of this theorem is a classic, and is due to Cantor.

Proof. Suppose for contradiction that \mathbb{R} were countably infinite. There exists a bijection $f : \mathbb{N} \rightarrow \mathbb{R}$, and we make a table of values the function takes on (see Table 1). Table 1 is just an example of what such a bijection may look like, and the exact values are moot. Because f is a bijection, this table should go on forever, and count every element of \mathbb{R} . To reach a contradiction, we will simply show there exists a real number that was not counted.¹¹ “Construct” this uncounted real number in the following way: let the n^{th} digit (decimal places included) take on the value of the n^{th} digit of $f(n)$ minus one (if it is 0, set it to 9). For Table 1, the first digit would be the first digit of $f(1)$ minus 1, which is $4 - 1 = 3$. The second digit would be the second digit of $f(2)$ minus 1, which is $4 - 1 = 3$. We repeat this process for all $n \in \mathbb{N}$, and in our case we get

3.308690...

¹¹There in fact exist many that would go uncounted, but it suffices to find just one.

$n \in \mathbb{N}$	$f(n) \in \mathbb{R}$
1	4.3214875...
2	1.4918401...
3	3.0194510...
4	9.0194510...
5	0.3917293...
6	5.9184017...
7	1.9284010...
\vdots	\vdots

By construction, the n^{th} digit of this number is different from at least one of the n^{th} digits of $f(n)$. This holds for every $n \in \mathbb{N}$, so this number is different from every value of $f(n)$, and was therefore not counted. \square

Corollary 1.1. Every infinite subset of \mathbb{R} is uncountable.

Example 1.26. Every interval $[a, b] \subseteq \mathbb{R}$ is uncountable.

Now that we've seen which familiar sets are and are not countable, there are several key results involving the cardinality of sets that deserve attention. These will establish what happens to the cardinality of sets when different set operations are performed.

Proposition 1.8. Let $\{E_n\}$, $n \in \mathbb{N}$, be a sequence of countably infinite sets, and let $E = \cup_{n \in \mathbb{N}} E_n$ be a countable union. The set E is countably infinite.

Proof. We will prove via induction. Our base case is $n = 2$. The sets E_1 and E_2 are countably infinite, so there exists bijections $f : \mathbb{N} \rightarrow E_1$ and $g : \mathbb{N} \rightarrow E_2$. Without loss of generality, assume $E_1 \cap E_2 = \emptyset$.¹² Define $h : \mathbb{N} \rightarrow E_1 \cup E_2$ as

$$h(k) = \begin{cases} f(k/2) & \text{if } k \text{ is even} \\ g((k+1)/2) & \text{if } k \text{ is odd} \end{cases}.$$

This function counts the elements in the set by alternating between those in E_1 and E_2 (like in the proof of Proposition 1.7). The function h is a bijection, so $E_1 \cup E_2$ is countably infinite. Now suppose this holds for E_1, \dots, E_{n-1} . We can write E as a union of two countably infinite sets by taking the union over E_1, \dots, E_{n-1} , which is countably infinite by the induction hypothesis.

$$\begin{aligned} E &= \bigcup_{n \in \mathbb{N}} E_n \\ &= E_1 \cup E_2 \cup \dots \cup E_{n-1} \cup E_n \\ &= (E_1 \cup E_2 \cup \dots \cup E_{n-1}) \cup E_n \end{aligned}$$

Therefore E is countably infinite. \square

Corollary 1.2. If X is uncountable, and $E \subseteq X$ is countably infinite, then $X \setminus E$ is uncountably infinite.

¹²Otherwise, we could replace E_1 with $E_1 \setminus E_2$.

Example 1.27. Let $E_n = \{m/n \mid m \in \mathbb{Z}\}$ for $n \in \mathbb{N}$. Each E_n is countable as there is a bijection $f_n : E_n \rightarrow \mathbb{Z}$ defined as $f_n(x) = nx$, and \mathbb{Z} is countably infinite.¹³ Note that

$$\bigcup_{n \in \mathbb{N}} E_n = \mathbb{Q},$$

which is indeed countably infinite.

Example 1.28. The set \mathbb{R} is uncountable. We have that $\mathbb{Q} \subseteq \mathbb{R}$ is countable. By Corollary 2.1, $\mathbb{R} \setminus \mathbb{Q}$ (the set of irrational numbers) is uncountable. This means that in a certain sense, there are more gaps in \mathbb{Q} than there aren't! We are not even capable of counting all the gaps, whereas we can count \mathbb{Q} .

Proposition 1.9. Let X be a countable set. Any subset $Y \subseteq X$ is countable

Proof. There exists a bijection $f : \mathbb{N} \rightarrow X$. If we restrict the codomain of f to be Y , f is still a bijection. \square

Example 1.29. Every subset of \mathbb{Q} is countable, because \mathbb{Q} is countable. We already know two such examples: \mathbb{N} and \mathbb{Z} .

Proposition 1.10. Let $\{E_n\}$, $n = 1, \dots, m$, be a finite sequence of countable sets, and let $E = \times_{n=1}^m E_n$ be a countable Cartesian product. The set E is countably infinite.

Proof. It suffices to show the result for two sets E_1 and E_2 , and then apply induction using the same argument used in the proof of Proposition 1.8. We have bijections $f : E_1 \rightarrow \mathbb{N}$ and $g : E_2 \rightarrow \mathbb{N}$. Define $h : E_1 \times E_2$ as

$$h((a, b)) = 2^{f(a)}3^{g(a)},$$

where $(a, b) \in E_1 \times E_2$. Each element in $h(E_1 \times E_2)$ is a whole number with a prime factorization comprised of only 2 and/or 3. Because each element of \mathbb{N} is uniquely determined by its prime factorization, h is injective. Unfortunately, h is no surjective, as there exist many elements of \mathbb{N} with prime factorizations that include more than 2 and/or 3. If we restrict the codomain of h to just its image, we have a bijection $h' : E_1 \times E_2 \rightarrow h(E_1 \times E_2)$. We do have that $h(E_1 \times E_2) \subseteq \mathbb{N}$, so by Proposition 1.9, $|h(E_1 \times E_2)| = \aleph_0$. By transitivity, $|E_1 \times E_2| = \aleph_0$. The aforementioned induction can be applied to conclude $|E| = \aleph_0$. \square

Example 1.30. The set of all pairs of rational numbers \mathbb{Q}^2 is countable.

1.7 Exercises

Exercise 1.1. Show that $\sqrt{3}$ is irrational.

Exercise 1.2. Let $(S, <)$ be an ordered set, and T be a nonempty subset of S . Verify that T has at most one supremum.

Exercise 1.3. Let $(S, <)$ be an ordered set, and A and B be nonempty subsets of T . Show that if $A \subset B$, then $\sup A \leq \sup B$ and $\inf B \leq \inf A$.

Exercise 1.4. 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 $\alpha \leq \beta$.

Exercise 1.5. Let E be a nonempty subset of \mathbb{R} which is bounded below. Let $-E = \{-x \mid x \in E\}$. Show that $\inf E = -\sup(-E)$.

¹³This allows us to use the transitivity of sets having the same cardinality.

Exercise 1.6. A set has the least-upper-bound property *if and only if* it has the greatest-lower-bound property.

2 Point-Set Topology in Metric Spaces

One of the main goals of calculus is to study rates of change and limiting behavior. Both of these concepts require some notion of distance, and to that end we will study *metric spaces*, sets equipped with a distance function. Any such space has induced “topological” properties. This is a fancy way of saying that we can use distance to categorize different types of sets. Of particular interest, will be the different types of sets in \mathbb{R} and \mathbb{R}^n , as the properties these sets have will have major implications down the road.

2.1 Metric Spaces

Our first definition will outline how we endow a set with some notion of distance.

Definition 2.1. A *metric space* is an ordered pair (M, d) where M is a set and $d : M \times M \rightarrow [0, \infty]$ is a function which satisfies:

1. $d(x, y) = 0 \iff x = y$.
2. $d(x, y) = d(y, x)$ for all $x, y \in M$ where $x \neq y$.
3. $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in M$. (Triangle Inequality)

The function d is often called the metric, and most of its properties are compatible with our everyday understanding of distance. Firstly, distance cannot be negative. There is no distance between a point and itself. The distance from x to y is the same from y to x . The final property may not be as immediate, but it is extremely important.

Suppose you are traveling from point x to z . If you decide to take a detour to point y before heading to z , then the triangle inequality ensures that you travel a weakly greater distance. An illustration shown in Figure 8 of this gives rise to the inequalities name. Geometrically, this is equivalent to saying that the length

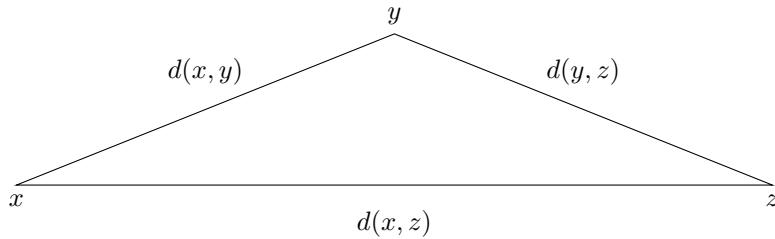


Figure 8: The triangle inequality.

of any side of a triangle cannot be greater than the sum of the other two lengths. Whenever presented with a weak inequality, it is often helpful to ask “when does this hold with equality”? In this case the answer is when y is on the line segment formed by x and z . In this case going to y isn’t a detour at all, but just a trivial stop on the way from x to z !

Example 2.1 (Euclidean Metric). The real line \mathbb{R} is a metric space when equipped with the metric $d(x, y) = |x - y|$.

Example 2.2 (Euclidean Metric). Euclidean space \mathbb{R}^n is a metric space when equipped with the metric

$$d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

The Euclidean metric is intimately linked to the concept of a *norm*. Recall from linear algebra that Euclidean space is a vector space where vectors are elements of \mathbb{R}^n , and scalars are elements of \mathbb{R} . This space is equipped with function $\|\cdot\|_2 : \mathbb{R}^n \rightarrow \mathbb{R}$ that measures the length of vectors.

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$$

We can write the Euclidean metric as $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$. Right now, don't worry too much about norms.¹⁴

Example 2.3 (Taxi-Cab Metric). If our set is \mathbb{R}^2 we can let $d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2|$. This is often referred to as the taxi-cab metric, as it is how you would measure distance if driving a car on a grid. We can extend to \mathbb{R}^n and let $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|$.

Example 2.4 (p -Adic Metric). The previous examples are easy to verify, but this may not always be the case. Suppose our set is \mathbb{Z} and $d : \mathbb{Z} \times \mathbb{Z} \rightarrow [0, \infty]$ is defined as

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ p^{-\max\{m \in \mathbb{N} \mid p^m \mid (x-y)\}} & \text{otherwise} \end{cases}$$

for some prime number p . Before we verify this is a metric, it's worth getting a feel for how the metric actually works. If $x \neq y$, then the distance between two points is p raised to some negative power. That negative power is defined to be the maximum whole number m such that $(x - y)$ is divisible by p^m . This gives us the vague idea that distance between points x and y is somehow related to how many times p shows up in the prime factorization of $(x - y)$ (where m is the number of times). Let's take $p = 3$, and pick several points in \mathbb{Z} to measure the distance between.

x	y	$x - y$	prime factorization of $x - y$	m	p^{-m}
100	19	81	3^4	4	$1/81$
368	8	360	$2^3 \cdot 5 \cdot 9$	0	1
35	5	30	$2 \cdot 3 \cdot 5$	1	$1/3$

It turns out that the more factors of p that go into the prime factorization of $(x - y)$, the closer x and y are. Furthermore, the maximum distance between any two points is 1, as $p^0 = 1$ for all p . We will now verify that this is indeed a metric.

1. The function $d(x, y)$ is defined such that $d(x, y) = 0$ if and only if $x = y$.
2. We have $(x - y) = -(y - x)$. Therefore, the prime factorization of each number differ only in sign, and give the same value m . This implies that $d(x, y) = d(y, x)$.
3. Note that to show $d(x, z) \leq d(x, y) + d(y, z)$ for all points in \mathbb{Z} , it suffices to show that $d(x, z) \leq \max\{d(x, y), d(y, z)\}$. This inequality happens to be a stronger condition than implies the triangle inequality. Suppose $p^m \mid (x - y)$ and $p^n \mid (y - z)$. For some $s, r \in \mathbb{Z}$, we having

$$\begin{aligned} x - y &= p^m r \\ y - z &= p^n s. \end{aligned}$$

¹⁴I admittedly am not certain of when it is best to introduce the concept of a norm. I don't like how Rudin (1976) talks about it in passing when reviewing Euclidean space. Introducing it latter on when covering functional analysis is also problematic, because we're going to use the sup-norm before that to measure the distance between two functions.

We can combine these equations to conclude

$$x - z = (x - y) + (y - z) = p^m r + p^n s.$$

If $m > n$, then $x - z = p^n(p^{m-n}r + s)$ and $d(x, z) = d(y, z)$. Similarly, if $n > m$, $d(x, z) = d(x, y)$.

Finally if $n = m$, then

$$x - z = p^n(r + s) = p^m(r + s),$$

and $d(x, z) = d(x, y) = d(y, z)$. These three cases gives the desired inequality.

Definition 2.2. Let X be a metric space. A set $E \subseteq X$ is *bounded* if there is a positive number $M \in \mathbb{R}$ and a point $x \in X$ such that $d(x, y) < M$ for all $y \in E$. If a set is no bounded, we say it is *unbounded*.

Boundedness insures that a set doesn't "go off to infinity".

Example 2.5. The sets \mathbb{N} , \mathbb{Z} , \mathbb{Q} , and \mathbb{R} are all unbounded.

Example 2.6. Both the intervals $[a, b]$ and (a, b) are bounded in \mathbb{R} . For any $x, y \in [a, b]$, $d(x, y) < d(a, b) + 1$. The same holds for (a, b) .

The metric space we are most interested in is of course \mathbb{R}^n equipped with the familiar Euclidean metric. We can use this metric to define the notion of an open or closed ball in \mathbb{R}^n .

Definition 2.3. If $\mathbf{x} \in \mathbb{R}^n$ and $r > 0$, the *open ball* with center \mathbf{x} and radius r is defined as

$$B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n \mid |\mathbf{y} - \mathbf{x}| < r\}.$$

Definition 2.4. If $\mathbf{x} \in \mathbb{R}^n$ and $r > 0$, the *closed ball* with center \mathbf{x} and radius r is defined as

$$\bar{B}_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n \mid |\mathbf{y} - \mathbf{x}| \leq r\}.$$

Open and closed balls in \mathbb{R}^n are a generalization of the open and closed intervals you were first introduced to in high school, and Figure 9 provides an illustration in \mathbb{R}^2 .

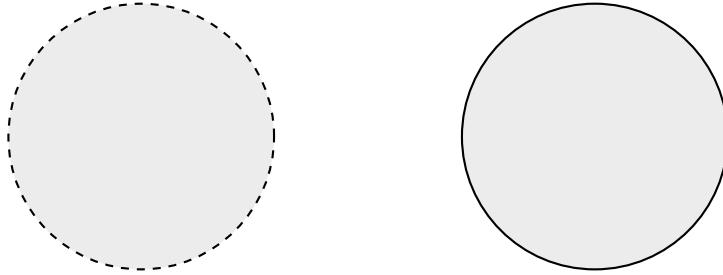


Figure 9: Open and closed balls in \mathbb{R}^2 .

2.2 Open Sets, Closed Sets, and Boundaries

We want to generalize the notion of open and closed balls in \mathbb{R}^n to any metric space. In order to do these, we'll need to outline a couple preliminary definitions that classify the elements of a metric space.

Definition 2.5. An *open ball* centered at x with radius r in a metric space X is defined as

$$B_r(x) = \{y \in X \mid d(x, y) < r\}$$

for a radius $r > 0$.¹⁵

An open ball is its own set, and we will use them constantly. They are sort of like “sets of utility”, because we will use them as tools to analyze the properties of other sets. If there is a set E in a metric space X , we can use open balls in X to learn about the points in E . Are some open balls subsets of E ? Do some open balls intersect E ? Will the answers to these questions change if we make r really big or really small?

Example 2.7. Let our metric space be \mathbb{Z}^2 equipped with the taxi-cab metric. Figure 10 shows the open ball centered at the origin of radius 3.

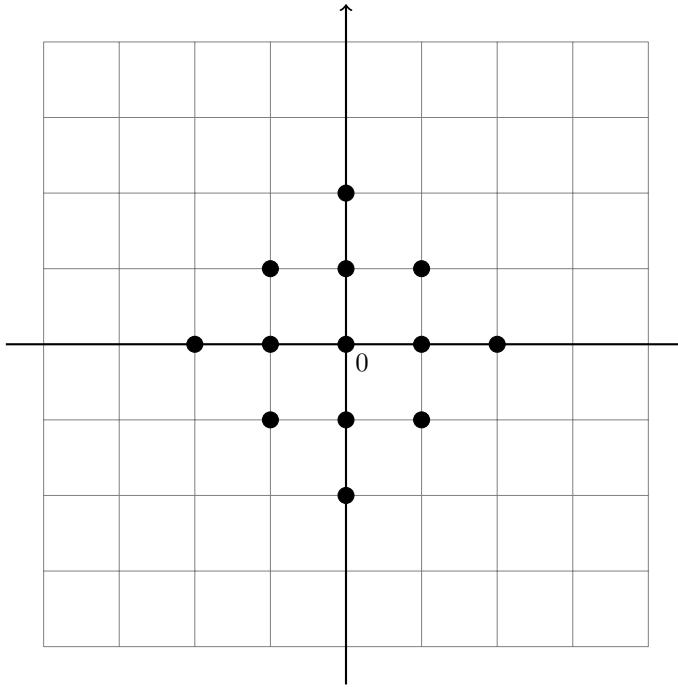


Figure 10: The set $B_3(\mathbf{0}) = \{\mathbf{y} \in \mathbb{Z}^2 \mid |\mathbf{y}_1| + |\mathbf{y}_2| < 3\}$.

Example 2.8. Let $d_2 : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow [0, \infty]$ be the Euclidean metric, and $d_1 : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow [0, \infty]$ be the taxi-cab metric. An open ball in (\mathbb{R}^2, d_1) may have a different “shape” than it would in (\mathbb{R}^2, d_2) . We will denote $B_3(\mathbf{0}) \in \mathbb{R}^2$ as E and F , in (X, d_2) and (X, d_1) respectively. These open balls are shown in Figure 11.

Definition 2.6. Let X be a metric space. A point $x \in X$ is a *limit point* of the set $E \subseteq X$ if *every* open ball of x contains a point $y \in E$, where $y \neq x$. We will denote the *set of all limit points* of E as

$$E' = \{x \in X \mid x \text{ is a limit point of } E\} = \{x \in X \mid \exists (B_r(x) \cap E) \setminus \{x\} \neq \emptyset \forall r > 0\}$$

¹⁵Rudin (1976) calls this a neighborhood, but this isn't quite right once we consider topology outside the context of metric spaces.

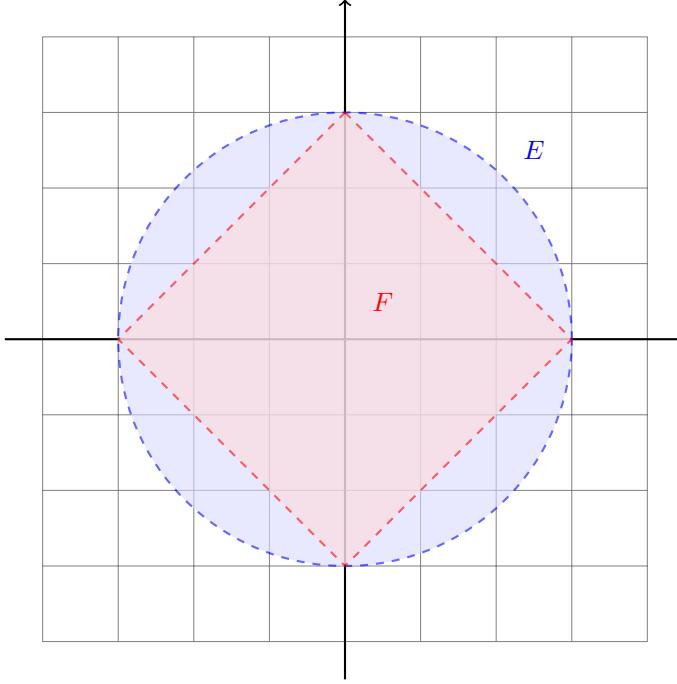


Figure 11: $B_3(\mathbf{0}) \in \mathbb{R}^2$ in (\mathbb{R}^2, d_2) and (\mathbb{R}^2, d_1) .

Notation 2.1. For the remainder of Section 2, we will use X to denote a metric space, and E as some subset of X .

A limit point of a set is in some sense always “close” to points of the set. If x is a limit point of $E \subseteq X$, then $B_r(x)$ will always include points other than x , no matter what we take r to be! We could make r smaller and smaller, but the set $B_r(x)$ will never just be x . In this sense, a limit point can always be “approximated” by elements in E .

Remark 2.1. Definition 3.5 never specifically said that a limit point of some set belongs to the set. As the next example shows, being a limit point has nothing to do with whether or not a point is included in the set in question.

Example 2.9. \mathbb{R}^2 is a good starting place. Suppose we have a set $E \subseteq \mathbb{R}$ that for the most part forms a rectangle. The “border” of the rectangle is not included in E . Also note that E includes an “isolated” point z (see Figure 12). Let’s consider three points in \mathbb{R}^2 : x , y , and z .

The point x belongs to E . Furthermore, no matter what we take r to be, $B_r(x)$ will never become a singleton of just $\{x\}$. For the sake of argument, suppose $x = (2, 2)$. If $r = 0.5$, then $(2, 2.49) \in B_{0.5}(x)$. If $r = 0.01$, then we still have $(2, 2.001) \in B_{0.01}(x)$. In fact, for every r , we have $(2, 2 + r/2) \in B_r(x)$. This means that x is a limit point of E .

Now consider y . This point does not belong to E , but it is still a limit point! We could repeat the same argument we made for x without running into trouble, because every open ball around y will include points “just below” y , all of which are in E ! What matters with limit points is not what set the point belongs to, but what set the points nearby it belong to.

Lastly, the point z is not a limit point. If we took r to be sufficiently large, then $B_r(z)$ would include points in E that form the rectangle. Unfortunately, we could easily take r to be so small that $B_r(z) = \{z\}$.

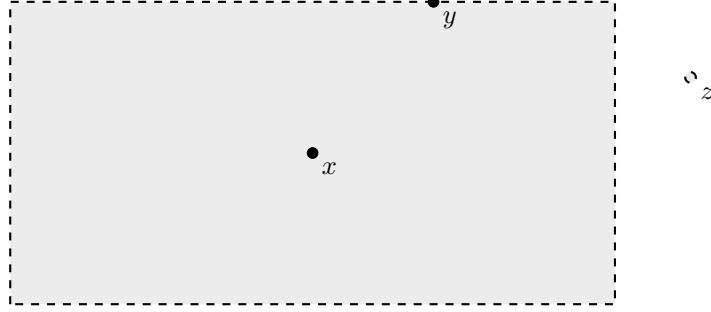


Figure 12: The set $E \subseteq \mathbb{R}^2$.

It only takes one such r to rule out the chance of z being a limit point. We can provide a definition that corresponds to points like z .

Definition 2.7. Let X be a metric space. For a set $E \subseteq X$, $x \in E$ is an *isolated point* if it is not a limit point. That is, there exists an $r > 0$ such that $B_r(x) = \{x\}$.

By the definition of an isolated point, it is the opposite of a limit point, rendering the two definitions mutually exclusive. This definition also means any point $x \in X$ is *either* a limit point *or* an isolated point. An isolated point of any set is also included in the set, which is not the case for limit points.

Definition 2.8. Let X be a metric space. A point $x \in X$ is an *interior point* of $E \subseteq X$ if there exists a single $r > 0$ such that $B_r(x) \subseteq E$.

Example 2.10. Again, let's look at an example in \mathbb{R}^2 . Let $E \subseteq \mathbb{R}^2$ be a closed ball that is "punctured" at $z \in \mathbb{R}^2$ such that $z \notin E$. This can be seen in figure 13. The point x is in an interior point, as we could find some small r for which $B_r(x) \subseteq E$. The point y is not an interior point of E , because every single $B_r(y)$ will contain some point outside of E , meaning $B_r(y) \not\subseteq E$. Finally, the point z is not an interior point, as each open ball $B_r(z)$ contains z , and $z \notin E$. Even though we can make r small enough to guarantee the only point in $B_r(z)$ which is not in E is z ($B_r(z) \setminus E = \{z\}$), this point is all it takes to guarantee $B_r(z) \not\subseteq E$ for all r .

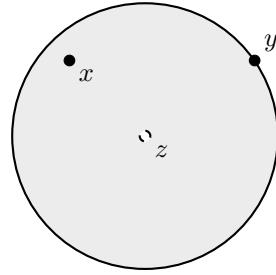


Figure 13: The set $E \subseteq \mathbb{R}^2$.

Remark 2.2. While a limit point of $E \subseteq X$ need not be a point in E , an interior point of E must be an element of E . If x is an interior point, then $x \in B_r(x) \subseteq E$ for some r , so $x \in E$.

Example 2.11. It may be tempting to conclude that an interior point must be a limit point, after all, if we can find an $B_r(x) \subseteq E$, then it is likely each open ball would contain infinite points of X . This logic makes

the dangerous assumption that X is infinite, and $d(x, y)$ “behaves like” the Euclidean metric. Consider a metric space X with the discrete metric

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}.$$

For any $x \in X$, $B_{1/2}(x) = \{x\} \subseteq X$. We have that x is an interior point, but not a limit point.

We now briefly introduce the idea of an exterior point. The only difference between the definition of an interior point and exterior point, will be that the open ball around a point will be in the complement of E for an exterior point. This small change in language will make a big difference in meaning.

Definition 2.9. Let X be a metric space. A point $x \in X$ is an *exterior point* of $E \subseteq X$ if there exists a single $r > 0$ such that $B_r(x) \subseteq E^c$.

Remark 2.3. Any point $x \in X$ is *either* an interior point *or* an exterior point.

Example 2.12. Let $[0, 1] \in \mathbb{R}$. The point $2 \in \mathbb{R}$ is an exterior point of $[0, 1]$.

Remark 2.4 (VERY IMPORTANT THEME). Nearly every definition in this section specifies a metric space X . This means the metric space we work in could affect how we classify a point (and later sets). If we have two metric spaces X and Y where $X \subseteq Y$, a point $x \in E \subseteq X$ may be a limit point/interior point/exterior point in X but not in Y .

We will see this come up again, and again. How a set/point behaves or is classified is contingent on what space we are in. A small change, whether it be the inclusion of some additional points, or changing the metric, can make a big difference. This means it is important to specify what space we’re in if it is ever unclear. On the bright side, this all makes for great examples!

Example 2.13. The set \mathbb{R} has no limit points when equipped with the discrete metric

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}.$$

We already generalized the idea of some open or closed interval in \mathbb{R} to the concept of an open or closed ball in \mathbb{R}^n . Now we will go one step further, by bringing these concepts to any metric space.

Definition 2.10. Let X be a metric space. A set $E \subseteq X$ is *open* if every point of E is an interior point.

Definition 2.11. Let X be a metric space. A set $E \subseteq X$ is *closed* if it contains all its limit points. That is, $E' \subseteq E$.

Example 2.14. Let $(a, b) \subseteq \mathbb{R}$. This set is open, as for all $x \in (a, b)$, we can find an r such that $B_r(x) \subseteq (a, b)$. If $d(a, x) \geq d(x, b)$, let $r = d(x, b)/2$. If $d(x, b) > d(a, x)$ let $r = d(x, a)/2$. Figure 14 shows this open ball around x where $d(a, x) \geq d(b, x)$. By construction, our open ball will always be a proper subset of (a, b) , so each point of x is an interior point. Therefore (a, b) is open. On the other hand, (a, b) is not closed, because a and b are limit points, but neither are in the set (a, b) .

Example 2.15. Let $[a, b] \subseteq \mathbb{R}$. This set is not open, as a and b are not interior points. For instance, $a - r/2 \in B_r(a)$ for all $r > 0$. The number $a - r/2 \notin [a, b]$, so $B_r(a) \not\subseteq [a, b]$ for all r . While $[a, b]$ is not open, it is closed. Every point in $[a, b]$ is a limit point because \mathbb{R} is complete. The interval $[a, b]$ trivially contains itself, so it contains all its limit points and is closed.

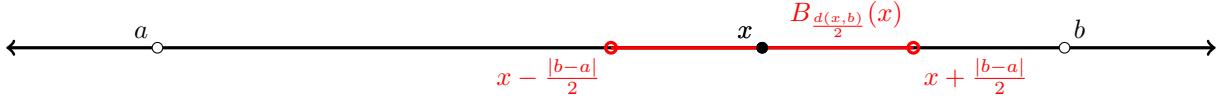


Figure 14: The open interval $(a, b) \subseteq \mathbb{R}$.

Example 2.16. Let X be a metric space, and $E \subseteq X$. Any set with no limit points, $E' = \emptyset$, is closed, because $\emptyset \subseteq E$. This means any finite set is closed, as no finite set has limit points. Let $X = \{x_1, \dots, x_n\}$, and $E \subseteq X$. If we let $y \in E$, and set $r = \min_{x \in X} d(x, y)$, then $B_r(y) = \{y\}$. This means y fails to be a limit point for all $y \in E$.

Remark 2.5. The definitions of open and closed sets never imply that a set is either closed or open. It is possible for a set to be both closed and open, or be neither closed nor open.

Example 2.17. The set \emptyset in any metric space X is closed and open. This set has no limit points ($\emptyset' = \emptyset$), and $\emptyset \subseteq \emptyset$, so it is closed. The set also has no points, so every point is an interior point, making \emptyset open.

Example 2.18. The set of rationals \mathbb{Q} is neither open nor closed in \mathbb{R} . For all $x \in \mathbb{Q}$, $B_r(x)$ will contain irrational numbers for all r , meaning $B_r(x) \not\subseteq \mathbb{Q}$. Therefore no elements of \mathbb{Q} are interior points. The set \mathbb{Q} also does not contain all of its limit points, as any irrational number is a limit point as a result of Theorem 2.4. For example, any open ball around $\sqrt{2}$ will contain elements of \mathbb{Q} (which are not $\sqrt{2}$), making it a limit point.

Remark 2.6 (Relative to Which Space). The point made in Remark 2.4 is especially relevant for open and closed sets. A set could be open in one metric space, but closed in another. For example, if we have the space of real numbers $[0, 1]$ with the Euclidean metric, $[0, 0.5)$ is open in $[0, 1]$. We also have that $[0, 1]$ is open in $[0, 1]$! In most cases it's clear what metric space we are working in, but sometimes it is not. In cases where it is vague, it's always best to say a set is *open in X* or *closed in X* . For this reason it is a good practice to either specify $E \subseteq X$, or include “in X .” Many topics in analysis concern the metric space \mathbb{R}^n or \mathbb{R} , so if you say a set is closed or open in conversation, it is usually assumed the metric space is one of these spaces. For example, if you were to ask someone “are the integers closed or open?”, they would most likely assume you mean “are the integers closed or open in \mathbb{R} ?“

Example 2.19. Suppose we want to determine if \mathbb{Z} is open or closed in \mathbb{Z} . Every point is an interior point as $B_{1/2}(x) = x \subseteq \mathbb{Z}$ for all $x \in \mathbb{Z}$, so \mathbb{Z} is open in \mathbb{Z} . The set \mathbb{Z} has no limit points in \mathbb{Z} , as $B_{1/2}(x)$ does not include any points $y \in \mathbb{Z}$ where $y \neq x$. This gives that $\mathbb{Z}' = \emptyset$,¹⁶ so $\mathbb{Z} \subseteq \mathbb{Z}'$, and \mathbb{Z} is closed in \mathbb{Z} .

Now let our metric space be \mathbb{R} . Is \mathbb{Z} open in \mathbb{R} ? Let $x \in \mathbb{Z}$. For any $B_r(x)$ such that $r < 1$, $x - r/2 \in B_r(x)$, where $x - r/2 \notin \mathbb{Z}$. If $r \geq 1$, then $x - 1/2 \in B_r(x)$, where $x - 1/2 \notin \mathbb{Z}$.¹⁷ Therefore, there exists no r such that $B_r(x) \subseteq \mathbb{Z}$, so \mathbb{Z} is not open in \mathbb{R} . We have that \mathbb{Z} is closed in \mathbb{R} , as each point of \mathbb{Z} is still isolated.

Before proving some useful properties of open and closed sets, there is one more definition that can prove helpful at times. It formalizes the notion of points in a set that are just on the border of a set, like the endpoints of $[a, b] \subseteq \mathbb{R}$.

¹⁶This also means that every point of \mathbb{Z} is isolated.

¹⁷The case where $r \geq 1$ handles the situation where $r/2 \in \mathbb{Z}$. If this were the case, then $x - r/2 \in \mathbb{Z}$. This is not a problem, as $B_r(x)$ would still contain an uncountably infinite number of real numbers, but it makes explicitly finding one of those reals a little tricky. It's easier to just add or subtract $1/2$ from x and call it a day.

Definition 2.12. Let X be a metric space, and $E \subseteq X$. The *boundary* of E , denoted ∂E , is the set of points in X such that every open ball around p contains at least one point of E and at least one point not of E .

$$\partial E = \{x \in X \mid B_r(x) \cap E \neq \emptyset \text{ and } B_r(x) \cap E^c \neq \emptyset \forall r > 0\}$$

Any element of ∂E is a *boundary point*.

There are several equivalent definitions of ∂E , many of which are more popular than this specific one. These other definitions use terms that we will cover in Section 3.4, so we will circle back then and discuss the boundary of a again. The first example one's mind should jump to are open and closed balls in \mathbb{R}^n .

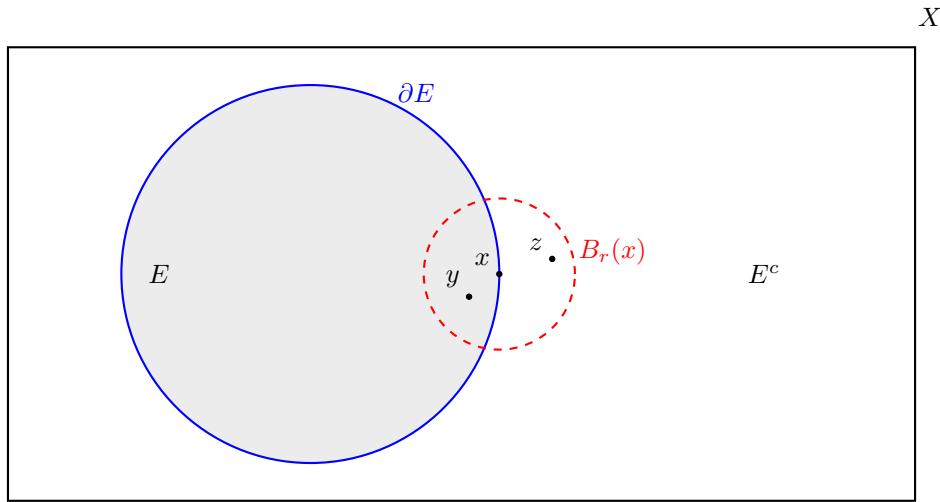


Figure 15: The boundary of a set E is shown in blue. Let $x \in X$. No matter how small we make r , $B_r(x)$ will always contain some $y \in E$ and some $z \in E^c$, so $x \in \partial E$.

Example 2.20. Let $B_r(\mathbf{x})$ be the open ball of radius r centered at $\mathbf{x} \in \mathbb{R}^n$ (we could also denote this as $B_r(\mathbf{x})$). As the name suggests, the boundary is just all the points that are exactly a distance of r away from \mathbf{x} , meaning $\partial B_r(\mathbf{x}) = \{\mathbf{y} \in X \mid |\mathbf{x} - \mathbf{y}| = r\}$. In this particular case, no points in the boundary are in $B_r(\mathbf{x})$, so we have $B_r(\mathbf{x}) \cap \partial B_r(\mathbf{x}) = \emptyset$. If we take $\bar{B}_r(\mathbf{x})$ to be the closed ball, then we have the same boundary.

$$\partial \bar{B}_r(\mathbf{x}) = \partial B_r(\mathbf{x}) = \{\mathbf{y} \in X \mid |\mathbf{x} - \mathbf{y}| = r\}$$

We also have $\partial \bar{B}_r(\mathbf{x}) \cap \bar{B}_r(\mathbf{x}) = \bar{B}_r(\mathbf{x})$, and $B_r(\mathbf{x}) \cup \partial B_r(\mathbf{x}) = \bar{B}_r(\mathbf{x})$.

Remark 2.7. It is very tempting to think all boundary points are limit points. At first glance, the definition of a boundary point seems to imply a point $x \in \partial E$ is not only a limit point of E , but also a limit point of E^c . This is not true! Suppose $x \in E$ is a limit point. The definition of a limit point not only requires that $B_r(x) \cap E \neq \emptyset$ for all r , but also requires that there are points *other than* x in $B_r(x)$. A boundary point needn't satisfy this second requirement, so even if $B_r(x) \cap E = \{x\}$ for all r , x can still be a boundary point!

Example 2.21. Let $E \subseteq \mathbb{R}^2$ be the union of a disk punctured at z and an isolated point y (Figure 16).

The points y and z are both boundary points. Despite this, y is not a limit point of E , and z is not a limit point of E^c . This follows from the reasoning in the previous remark.

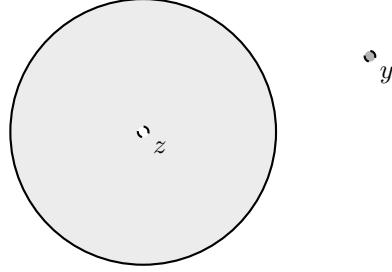


Figure 16: The set $E \subseteq \mathbb{R}^2$.

Remark 2.8. We have now introduced five different definitions that classify points: limit points, isolated points, interior points, boundary points, and exterior points. This is *a lot* to take in all at once. By far the most important concepts introduced here were open and closed sets. Being able to determine if a set is open and/or closed is one of the most important skills to have for this section, and those that follow.

2.3 Properties of Open and Closed Sets

Open set and closed sets will play a role in many of the results and theorems to come, so it is important to be able to identify which sets are open and which are closed. We will now introduce several tools that make this easier.

Proposition 2.1. Every open ball is an open set.

Proof. Suppose we have a metric space X , and some point $x \in X$. We will show that any point $y \in B_r(x)$ is an interior point. There exists some h such that

$$d(x, y) = r - h.$$

I claim that $B_{r'}(y) \subseteq E$ for $r' < h$. For all points z such that $d(y, z) = r' < h$, the triangle inequality gives

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

so $z \in B_r(x) = E$ for all z by the definition of $B_r(x)$. This implies that $B_{r'}(y) \subseteq E$, making y an interior point. (Figure 17) \square

Proposition 2.2. If $x \in X$ is a limit point of E ($x \in E'$), then every open ball around x contains infinitely many points of E .

Proof. Let $x \in X$. Suppose for contradiction, there exists some $B_r(x)$ which contains only a finite number of points of E . Let this finite set of points be $\{y_1, \dots, y_n\} \subseteq B_r(x) \cap E$. Pick the radius of $B_r(x)$ to be the distance between x and the point to which it is closest in the finite set $\{y_1, \dots, y_n\}$:

$$r = \min_{1 \leq m \leq n} d(x, y_m).$$

By construction, $B_r(x)$ contains no point $y \in E$ such that $y \neq x$, so x is not a limit point of E . This is a contradiction. (Figure 18) \square

Corollary 2.1. A finite set has no limit points. (see Example 3.13)

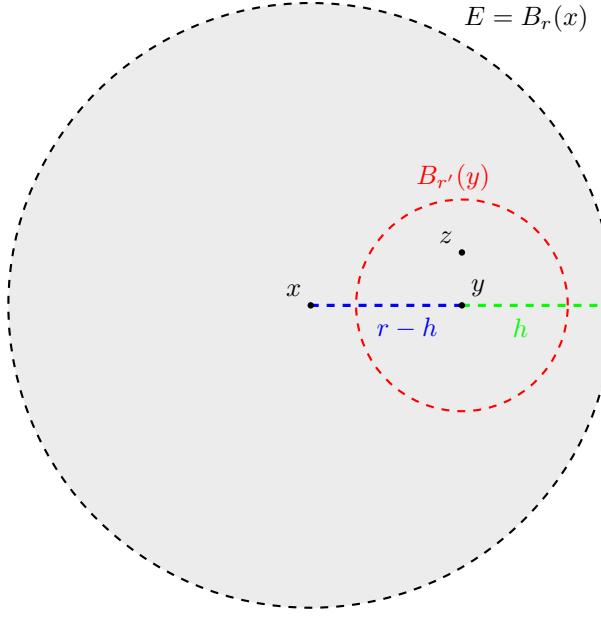


Figure 17: The set $E \subseteq X$. We constructed a open ball $B_{r'}(y)$ for an arbitrary $y \in B_r(x)$ such that $B_{r'}(y) \subseteq E$.

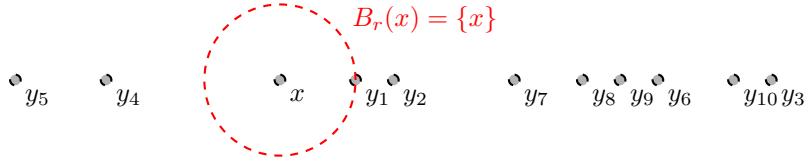


Figure 18: If our finite set of points of E is $\{y_1, \dots, y_{10}\}$, then we can reach a contradiction by constructing an open ball around x with $r = \min_{1 \leq m \leq n} d(x, y_m) = d(x, y_1)$. This will hold no matter where the points $\{y_1, \dots, y_{10}\}$ happen to be in E . In this case $x \in E$, but remember that this isn't a requirement.

The next theorem and its corollary allows us to determine if a set is open or closed based on its complement. At first, this may not seem helpful, but if it is not clear if E is open or closed, one can just use E^c ! We will provide examples where using E^c is easier.

Theorem 2.1. A set E is open if and only if its complement is closed.

Proof.

(\Rightarrow) Suppose E is open. Let $x \in X$ be a limit point of E^c . Every open ball $B_r(x)$ contains a point of E^c , so $B_r(x) \not\subseteq E$, meaning x is not an interior point of E . But we have assumed every point of E is an interior point, so $x \in E^c$. Therefore E^c includes all its limit points and is closed.

(\Leftarrow) Suppose E^c is closed. Let $x \in E$. We have $x \notin E^c$, so x is not a limit point of E^c (otherwise it would be in E^c , as E^c is closed). If x is not a limit point of E^c , then there exists an $B_r(x) \cap E^c = \emptyset$, giving $B_r(x) \subseteq E$. Thus $x \in E$ is an interior point, and E is open.

□

Corollary 2.2. A set E is closed if and only if its complement is open.

One practical consequence of these results, is that if you find it more difficult to check if a set is open or closed (or vice versa), you can always just work with the complement.

Example 2.22. Let X be any metric space. Recall from Example 3.14 that \emptyset is closed and open. This means that $\emptyset^c = X$ is closed and open as well. This allows us to conclude that \mathbb{R} in \mathbb{R} is open and closed.

Example 2.23. The set $[a, b] \subseteq \mathbb{R}$ is closed. This implies that $[a, b]^c = (-\infty, a) \cup (b, \infty)$ is open.

We often define some set of interest as a union or intersection of a collection of sets. For instance, the proof of Theorem 1.2, the supremum of a set in \mathbb{R} was defined as the union of the Dedekind cuts that comprise the set. In Example 1.27, \mathbb{Q} was written as the countably infinite union of intervals. In situations like this, it is possible to know if a set is open or closed if the sets over which we take the union/intersection are open or closed.

Theorem 2.2. Let $\{G_\alpha\}$ and $\{F_\alpha\}$ be an arbitrary collection of open sets and closed sets respectively. Let G_1, \dots, G_n and F_1, \dots, F_n be a finite collection of open sets and closed sets respectively. In this case we have:

1. $\bigcup_\alpha G_\alpha$ is open.
2. $\bigcap_\alpha F_\alpha$ is closed.
3. $\bigcap_{i=1}^n G_i$ is open.
4. $\bigcup_{i=1}^n F_i$ is closed.

Proof.

1. Suppose G_α is open for all α . Let $x \in \bigcup_\alpha G_\alpha$. For some α , $x \in G_\alpha$. There exists an open ball $B_r(x)$ such that $B_r(x) \subseteq G_\alpha$, because G_α is open. Therefore $\bigcup_\alpha G_\alpha$ is open, as $B_r(x) \subseteq G_\alpha \subseteq \bigcup_\alpha G_\alpha$.
2. Suppose F_α is closed for all α . It suffices to show that $(\bigcap_\alpha F_\alpha)^c$ is open using Theorem 2.1. By the aforementioned theorem, F_α^c is closed for all α . Therefore the union of F_α^c is open by part (1). This completes our proof, as De Morgan's Law gives

$$\left(\bigcap_\alpha F_\alpha \right)^c = \bigcap_\alpha F_\alpha^c.$$

3. Suppose the sets G_1, \dots, G_n are open. Let $x \in \bigcap_{i=1}^n G_i$. For all $x \in \bigcap_{i=1}^n G_i$, there exists open balls $B_{r_i}(x)$ with radii r_i , such that $B_{r_i}(x) \subseteq G_i$ for all i . Let $r = \min\{r_1, \dots, r_n\}$. This radius gives us $B_r(x) \subseteq G_i$ for all i , meaning $B_r(x) \subseteq \bigcap_{i=1}^n G_i$. Therefore x is an interior point, and $\bigcap_{i=1}^n G_i$ is open. (Figure 19)
4. Suppose the sets F_1, \dots, F_n are closed. It suffices to show that $(\bigcap_{i=1}^n F_i)^c$ is open using Theorem 2.1. By the aforementioned theorem, F_i^c is open for all i . By part (2) the the intersection of all F_i^c is open. This complete out proof, as De Morgan's Law gives

$$\left(\bigcap_{i=1}^n F_i \right)^c = \bigcap_{i=1}^n F_i^c.$$

□

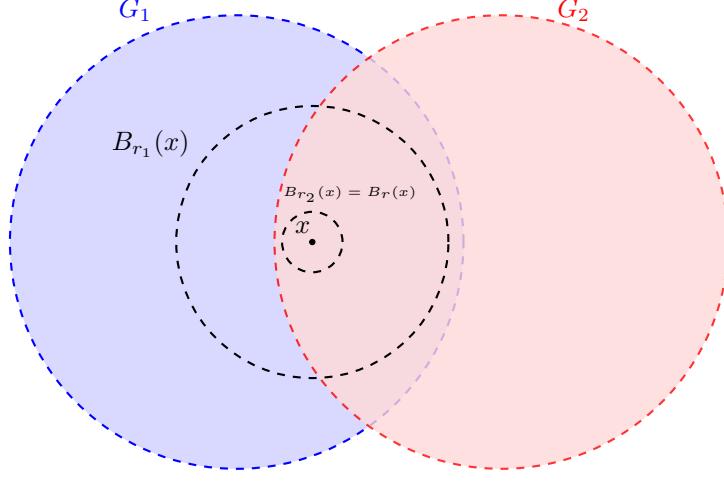


Figure 19: In this simplified setting, we have two open sets: G_1 , and G_2 . By letting $r = \min\{r_1, r_2\} = r_2$, we find an open ball $B_r(x)$ such that $B_r(x) \subseteq G_1 \cap G_2$.

Part (2) and (4) of Theorem 2.2 require that the collection of sets is finite, otherwise the minimum over the finite number of radii of open balls many not be well defined. The following two examples show that Theorem 2.2 does not hold if we take these collections to be infinite.

Example 2.24. Let $G_n = (1/n, 1 + 1/n)$ for all $n \in \mathbb{N}$. The set G_n is open in \mathbb{R} for all n . Taking the intersection gives

$$\bigcap_n G_n = [0, 1].$$

The interval $[0, 1]$ is closed, despite each G_n being open.

Example 2.25. Let $F_n = [1/n, \infty)$ for all $n \in \mathbb{N}$. The set F_n is closed in \mathbb{R} for all n . Taking the union gives

$$\bigcup_n F_n = (0, \infty).$$

This interval is open in \mathbb{R} , despite each F_n being closed.

2.4 Closures, Interiors, Dense Sets, and Perfect Sets

There are a handful of other definitions related to open set and closed sets that deserve a bit of attention.

Definition 2.13. Let X be a metric space. The *interior* of a set $E \subseteq X$, denoted E° , is the set of all interior points of E .

$$E^\circ = \{x \in X \mid x \text{ is an interior point of } E\}$$

The set E° is clearly open, as by definition it is comprised only of interior points. Because interior points of E must be in E , we have $E^\circ \subseteq E$. Informally, we can think of E° as the smallest open set contained within E . This interpretation leads to the conclusion that if E is open, then $E = E^\circ$.

Example 2.26. Let $[a, b] \subseteq \mathbb{R}$. The interior of this set is (a, b) .

Definition 2.14. Let X be a metric space. The *closure* of a set $E \subseteq X$ is $\bar{E} = E \cup E'$.

The closure \bar{E} is the opposite of the interior in a certain sense. The closure can be thought of as the smallest closed set that E is contained in. If E is closed, then $E' \subseteq E$, and $E = \bar{E}$.

Example 2.27. Let $(a, b) \subseteq \mathbb{R}$. The closure of this set is $[a, b]$.

Definition 2.15. Let X be a metric space. The set $E \subseteq X$ is *dense in X* if every point of X is a limit point of E , or a point of E . ($\bar{E} = X$)

Informally, if E is dense in X , then we can approximate any point of X with a point in E arbitrarily well. This follows from the fact that any point in X is either in E , or a limit point of E' , or both. We have already seen one example of this with Theorem 1.4.

Example 2.28. The set \mathbb{Q} is dense in \mathbb{R} , that is $\bar{\mathbb{Q}} = \mathbb{R}$. One implication of this fact is that the irrational numbers $\mathbb{R} \setminus \mathbb{Q}$ are limit points of \mathbb{Q} . We also have that the irrationals $\mathbb{R} \setminus \mathbb{Q}$ are dense in \mathbb{Q} !

Many results involving approximation can be stated in terms of dense sets. One of these is the Weierstrass Approximation Theorem. This will be formally treated and proved in Section 7, but for now we will give the result as an example of a dense set.

Example 2.29 (Weierstrass Approximation Theorem). Let $C([a, b]) = \{f \mid f : [a, b] \rightarrow \mathbb{R} \text{ and } f \text{ continuous}\}$ be the set of real valued continuous functions with domain $[a, b]$. Now let $\mathcal{P}([a, b])$ be the set of all real valued polynomials with domain $[a, b]$.¹⁸ The set $\mathcal{P}([a, b])$ is dense in $C([a, b])$. We will always be able to approximate a continuous function on a bounded interval arbitrarily well with polynomials. This result, and spaces of functions, will be discussed again in Section 7.

Definition 2.16. Let X be a metric space. The set $E \subseteq X$ is *perfect* if every point of E is a limit point of E ($E = E'$).

If E is perfect, every point in E can be approximated arbitrarily well by other points in E .

Example 2.30. The real line \mathbb{R} is a perfect set.

2.5 Compact Sets

We have encountered infinity several times now. Sets can have an infinite number of elements, in which case they are either countable or uncountable. A set can “take up an infinite amount of space” if it is unbounded. Each limit point of a set contains an infinite number of points in the set. These factors can result in sets that are difficult to work with. For example, suppose a set is unbounded. It can be hard to determine how functions or sequences behave on sets like this, because the distance between points can become arbitrarily large. What kind of headaches do limit points cause? Any open ball around a limit point will contain an infinite number of points in that set (Proposition 2.2). If the limit point is not in the set, this can also pose problems. As we will see later on, it could be possible to get arbitrarily close to that limit point while never leaving the set. In a sense, we would be getting closer to a point in a set, where the destination is not even included in the set. To prevent this from happening, all limit points should be included in a set, i.e. the set should be closed. Later on when working with sequences and continuous functions, two concepts intrinsically linked by the idea of getting arbitrarily close to a point, sets that are “nice”, and will not illicit

¹⁸This set is traditionally denoted as $\mathbb{R}[x]$ in abstract algebra.

the two mentioned complications, will lead to nice results. Our goal now is to characterize these sets, and attempt to motivate their characterization.¹⁹

One somewhat trivial way to guarantee a set is both closed and bounded is by restricting our attention to finite sets. If E is finite, it has no limit points and is trivially closed. A finite set must also be bounded. While finite may have the nice properties we are looking for, they are not that interesting. Real analysis almost always involves the real numbers, an uncountably infinite set. So what criteria would guarantee an infinite set, whether it be countable or uncountable, will behave like a finite set?

The way we will go about defining these sets is by looking how we can “cover” them with a collection of open sets. The idea is that a set may be infinite, but perhaps we can cover it with a finite collection of open sets.

Definition 2.17. Let X be a metric space. An *open cover* of a set $E \subseteq X$ is a collection of open sets $\{G_\alpha\} \subseteq X$ such that $E \subseteq \cup_\alpha G_\alpha$.

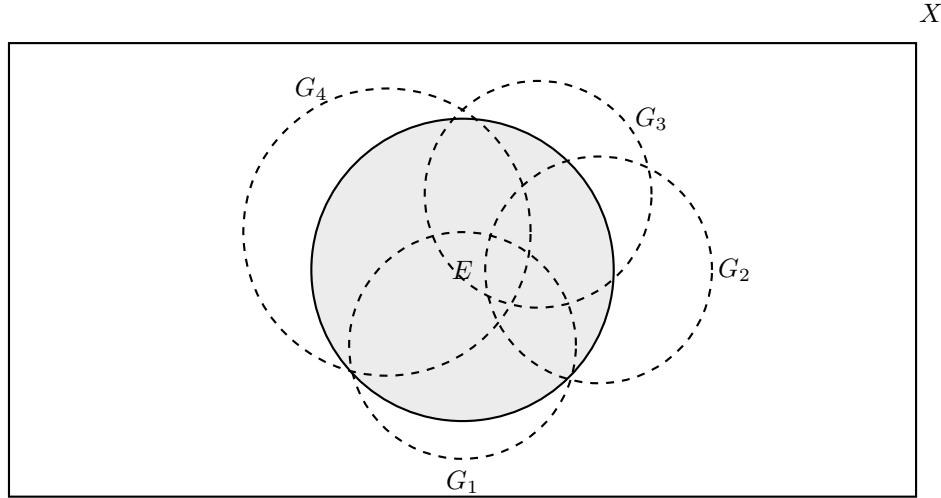


Figure 20: The collection of set $\{G_1, \dots, G_4\}$ forms an open cover of E .

An open cover is a *collection* of sets. Each element of an open cover is a single open set. This means that the cardinality of an open set has *nothing* to do with the cardinality of the sets it is comprised of. The collection $\{\mathbb{R}, (0, 1), (0, 2)\}$ is 3. We do not care whatsoever about the fact that each of the sets in the collection are uncountably infinite. The emphasis here is due to the fact that the cardinality of these covers (and a second type we will define soon) will be the bases of our criteria of what makes a set “nice”.

Example 2.31. Let $E = B_r(x)$ be a subset of a metric space X . One open cover of the set E is the single set $B_{r+1}(x)$. Another would be $B_{r+2}(x)$.

Example 2.32. Let \mathbb{R} be the entire real line. We can cover this with the union of all sets of the form $G_n = (-n, n)$ for $n \in \mathbb{N}$.

$$\mathbb{R} \subseteq \bigcup_{n \in \mathbb{N}} (-n, n)$$

¹⁹Motivating the main definition of this subsection is infamously difficult, as it is not clear how it will be used in the future. It would be like explaining what a hammer is to someone who has no idea what a nail is.

It is not enough to require that a set has a finite open cover. Any set E can trivially be covered by itself, forming an open cover consisting of one element. We could require that every open cover is finite, but this could never be satisfied. For example, take the closed interval $(a, b) \subseteq \mathbb{R}$. We could cover this with a finite open cover $\{(a, b)\}$. We could also cover it with the infinite open cover $\{(a, b), (-1, 1), (-2, 2), (-3, 3), \dots\}$. We can just take the cover $\{(a, b)\}$ and throw in an infinite number of random intervals of \mathbb{R} and still end up with an open cover. This is complete “overkill” when it comes to covering (a, b) ! To address the fact that we will always have infinite open covers, we will introduce a new type of open cover that is both finite, and limits any redundant additions to the cover.

Definition 2.18. Let X be a metric space, and $E \subseteq X$. A *finite subcover* of an open cover $\{G_\alpha\}$ of E is a collection of open sets $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$ such that

$$E \subseteq \bigcup_{i=1}^n G_{\alpha_i} \subseteq \bigcup_\alpha G_\alpha.$$

Example 2.33. Let X be a metric space and $E \subseteq X$. The collection $\{E\}$ is a trivial open cover. We also have a trivial finite open subcover in $\{E\}$.

Example 2.34. Let $\{(a, b), (-1, 1), (-2, 2), (-3, 3), \dots\}$ be an open cover of $(a, b) \subseteq \mathbb{R}$. One finite subcover of this open cover is $\{(a, b)\}$. Another finite subcover is $\{(a, b), (-1, 1)\}$. In fact, any set $\{(a, b), (-1, 1), \dots, (-n, n)\}$ is a finite subcover.

This simple example shows that some open covers actually have an infinite number of finite subcovers. The introduction of finite subcovers may be a bit confusing, so it is worth recapping what we have done before presenting the main definition of this subsection:

- We have some metric space X and some set $E \subseteq X$. We can cover this set with a collection of sets $\{G_\alpha\}$ called an open cover, the cardinality of which is determined by the number of sets in the collection $\{G_\alpha\}$. We like the idea of finite open covers.
- There exists an infinite number of open covers of a set. There also always exists a finite open cover of a set, so we need to do better than just having a finite open cover.
- A finite subcover $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$ of the open cover $\{G_\alpha\}$ is a subset of $\{G_\alpha\}$, which also covers E . The finite subcover of one open cover $\{G_\alpha\}$ is *not necessarily* a finite subcover of another open cover $\{G'_\alpha\}$, so whenever we talk about finite subcovers, it is with respect to some fixed open cover. Some open covers of sets have an infinite number of finite subcovers (Example 2.29).

We are now ready give a proper definition and name to the “nice” sets we want to characterize. In doing so, we will answer an important question about finite subcovers that may have arisen by now – some open covers of sets have an infinite number of finite subcovers, but do *all* open covers of a set have *at least one* finite subcover?

Definition 2.19. Let X be a metric space, and K be a subset of X . The set K is *compact* if *every* open cover of K contains *at least one* finite subcover.

The answer to our question turns out to be no. If it were yes, then there would be no need to define compactness, because every set would be compact. Compact sets turn out to be the nice sets we were looking for. As we’ll see, they can be infinite, but they do not cause the complications with infinity that we discussed at the open of this subsection. In some sense, compact sets are the next best thing to finite sets.

Example 2.35. The set $(a, b) \subseteq \mathbb{R}$ is not compact. In order verify this, we just need to find a single open cover that has no finite subcover. Let $\{G_n\}$ be an open cover where $G_n = (a + 1/n, b)$. We have that

$$(a, b) \subseteq \bigcup_{n \in \mathbb{N}} G_n = \bigcup_{n \in \mathbb{N}} (a + 1/n, b) = (a, b).$$

No finite subset of $\{G_n\}$ will be a finite subcover of (a, b) . If we had a finite subset of $\{G_n\}$ then there would exist some $N \in \mathbb{N}$ such that $(a, a + 1/N)$ is “uncovered”. Therefore any open interval in \mathbb{R} is not compact.

Example 2.36. The real line \mathbb{R} is not compact. The open cover $\{G_n\}$ where $G_n = (-n, n)$ has no finite subcover. If we had a finite subset of $\{G_n\}$, then there would exist some $N \in \mathbb{N}$ such that $(-\infty, -N) \cup (N, \infty)$ is “uncovered”. Therefore the real line \mathbb{R} is not compact.

These two examples should not be entirely surprising. When compactness was motivated, complications involving two types of sets were cited: unbounded sets, and sets that were not closed. (a, b) and \mathbb{R} both fall into exactly one of these categories.

Example 2.37. Suppose $X = \{x_1, \dots, x_n\}$ is a finite metric space. The entire space X is compact. Let $\{G_\alpha\}$ be an open cover of X . For each $x_i \in X$, there exists an G_{α_i} such that $x_i \in G_{\alpha_i}$. Therefore, the open cover has a finite subcover in $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$.

Using the definition of compactness to verify that a set is not compact takes some creativity, but only requires one to find a single counterexample. This is opposed to verifying a set is compact. This requires we somehow verify that every single open cover has a finite subcover. This is prohibitively difficult to do in most cases. This is why we want to find conditions that are easy to verify and that imply compactness. Of chief concern, is doing this for subsets of \mathbb{R}^n .

2.6 Properties of Compact Sets

Before restricting our attention to \mathbb{R}^n , we need to establish some properties of compact sets that will allow come in handy when working with them. Unfortunately, we still do not have any nontrivial examples of compact sets, so some of these results will not have examples presented alongside them. Once we are able to identify compact sets in \mathbb{R}^n by means other then the definition of compactness, these results can be verified.

Lemma 2.1. Suppose $Y \subseteq X$. A subset $E \subseteq Y$ is open in Y if and only if $E = Y \cap G$ for some open $G \subseteq X$.

Proof.

(\Rightarrow) Suppose $E \subseteq Y$ is open in Y . For each $x \in E$, there exists a r_x such that $B_{r_x}(x) \subseteq E$. By the definition of an open ball, for all $y \in Y$ satisfying $d(x, y) < r_x$, we have $y \in E$. Denote $B_{r_x}(x) = V_x$ for all $x \in E$, and define

$$G = \bigcup_{x \in E} V_x.$$

The set G is a union of open sets, so it is open (Figure 21). I now claim that $E = G \cap Y$, which is our desired result. For $x \in E$, we have $x \in V_x$ for all $x \in X$, so $x \in E$ and $x \in Y$. This gives $E \subseteq G \cap Y$. Now let $x \in G \cap Y$. For the corresponding V_x , $V_x \cap Y \subseteq E$. This implies $G \cap Y \subseteq E$.

(\Leftarrow) Suppose $E = Y \cap G$ for some open G in X . The set G is open, so for all $x \in E$ there exists an open ball $B_r(x) \subseteq E$. This gives $B_r(x) \subseteq G$, as $E = Y \cap G \subseteq G$. Intersecting $B_r(x)$ with Y yields $B_r(x) \cap Y \subseteq E$, so E is open in Y .

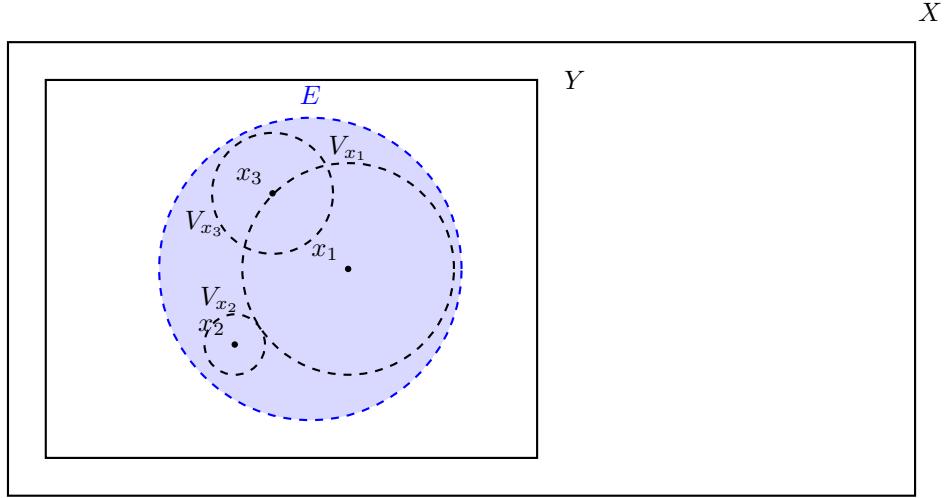


Figure 21: We take V_x to be the open ball around x contained in E . We will always be able to find such an open ball because E is open in Y . In this case, we have shown only three such open balls. The (possibly infinite) union of all such balls is G , which itself is open.

□

Proposition 2.3. Suppose $K \subseteq Y \subseteq X$, where Y and X are metric spaces. The subset K is compact in X if and only if K is compact in Y .

Proof.

(\Rightarrow) Suppose K is compact in X . Let $\{V_\alpha\}$ be an arbitrary collection of open sets in Y which cover K . By Lemma 2.1, there exist sets G_α , open in X , such that $V_\alpha = Y \cap G_\alpha$ for all α . By the compactness of K , there exists a finite subcover $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$. We have

$$K \subseteq \bigcup_{i=1}^n G_{\alpha_i},$$

but $K \subseteq Y$, so

$$K \subseteq \bigcup_{i=1}^n V_{\alpha_i}.$$

This makes $\{V_{\alpha_1}, \dots, V_{\alpha_n}\}$ a finite subcover, so K is compact in X .

(\Leftarrow) Suppose K is compact in Y . Let $\{G_\alpha\}$ be an open cover of K in X . If we let $V_\alpha = G_\alpha \cap Y$, then $\{V_\alpha\}$ is an open cover of K in Y . By the compactness of K we have a finite subcover $\{V_{\alpha_1}, \dots, V_{\alpha_n}\}$ in Y . Since $V_\alpha \subseteq G_\alpha$ for all α , then $\{G_\alpha\}$ has a finite subcover in $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$, so K is compact in X .

□

This result does not seem particularly interesting, but it is novel if you consider open and closed sets. We gave several examples of sets that may be open or closed in some intermediate space, but not a larger space. For instance, \mathbb{Z} is open in \mathbb{Z} , but it is not open in \mathbb{R} . This result tells us that results like this are not possible with compactness!

Example 2.38. Suppose $E \subseteq \mathbb{Z}$ is compact in \mathbb{Z} . This implies that E is compact in \mathbb{Q} and \mathbb{R} . If we had another compact set $F \subseteq \mathbb{R}$ which is compact in \mathbb{R} , then it is compact in \mathbb{Z} and \mathbb{Q} as well.

The next two theorems establish that all compact sets are both closed and bounded. This should feel somewhat natural, as compactness can be interpreted as a generalization of closed and bounded sets.

Theorem 2.3. Let X be a metric space, and $K \subseteq X$ be compact. The set K is closed.

Proof. Let K be a compact subset of a metric space X . It suffices to show that K^c is open in X . Suppose $x \in K^c$, and $y \in K$. For $r < \frac{1}{2}d(x, y)$, let $V_y = B_r(x)$ and $W_y = B_r(y)$ (Figure 24). For our fixed $x \in K^c$, we will repeat this process for multiple points in K . By compactness, we know there exists a finite set of $\{y_1, \dots, y_n\}$ such that $\{W_{y_1}, \dots, W_{y_n}\}$ is a finite subcover of K . In constructing this subcover, we also constructed the corresponding sets $\{V_{y_1}, \dots, V_{y_n}\}$ (Figure 23). If we let $V = \cap_{i=1}^n V_{y_i}$, then we have $x \in V \subseteq K^c$, making x an interior point of K^c . Therefore K^c is open. \square

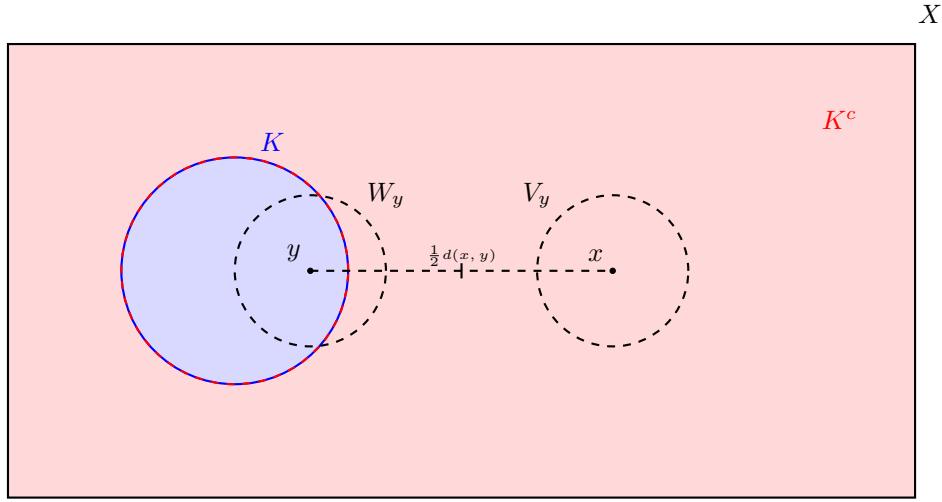


Figure 22: For some points $x \in K^c$ and $y \in K$, we construct open balls $W_y = B_r(y)$ and $V_y = B_r(x)$ such that $r < \frac{1}{2}d(x, y)$. This choice of radius ensures $V_y \cap W_y = \emptyset$.

Example 2.39. The set \mathbb{Q} in \mathbb{R} is not closed (Example 2.18), so it is not compact in \mathbb{R} .

Theorem 2.4. Let X be a metric space, and $K \subseteq X$ be compact. The set K is bounded.

Proof. Let $x \in K$. The collection of open balls $\{B_r(x)\}$ for $r \in \mathbb{N}$ forms an open cover of K . By compactness, this open cover has a finite subcover $\{B_{r_1}(x), \dots, B_{r_n}(x)\}$. If we take $r^* = \max\{r_1, \dots, r_n\}$, then $K \subseteq B_{r^*}(x)$, and $d(x, y) < r^*$ for all $y \in K$. (Figure 24) \square

Example 2.40. The set $(0, \infty)$ in \mathbb{R} is not bounded so it is not compact in \mathbb{R} .

Remark 2.9. While compactness implies closed and bounded, the converse is not necessarily true. Soon, we will see that the converse will hold in \mathbb{R}^n , but in general, this is not the case. The next example illustrates this.

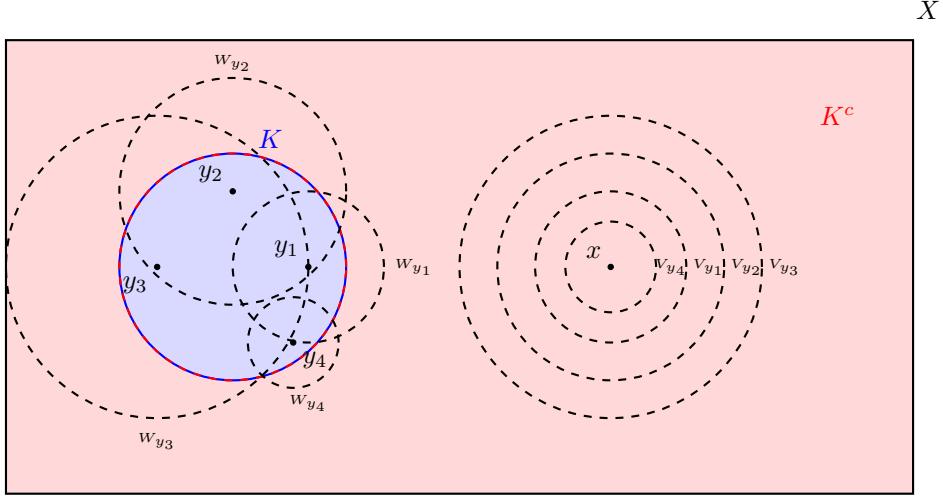


Figure 23: For the fixed value $x \in K^c$, repeat the process illustrated in Figure 21 until we have a finite subcover of K , $\{W_{y_1}, \dots, W_{y_4}\}$. If we let V be the intersection of all V_{y_i} , then $x \in V \subseteq K^c$, rendering x an interior point of K^c .

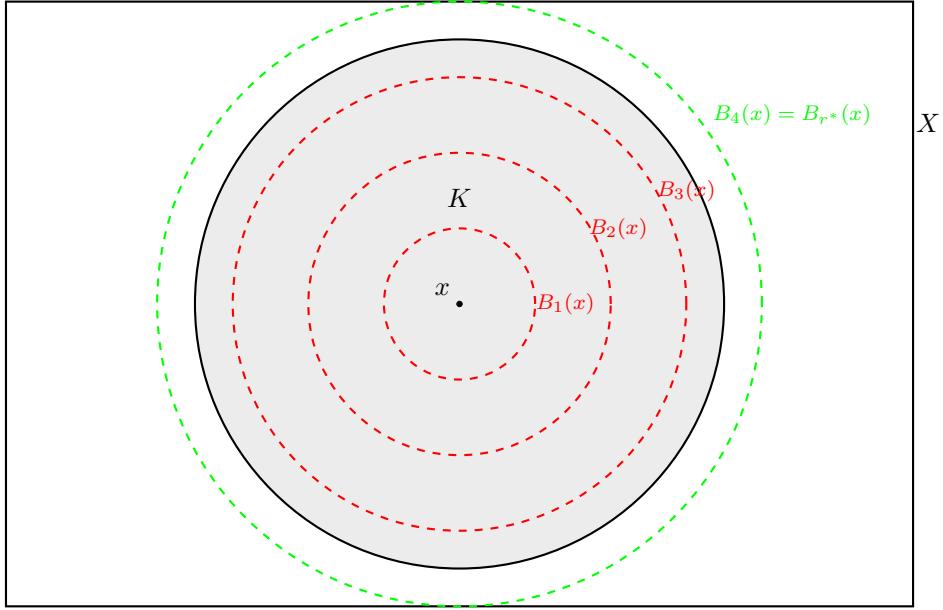


Figure 24: We cover the set K with an infinite open cover comprised of open balls of radii in \mathbb{N} . By compactness there is an open subcover, such as $\{B_1(x), \dots, B_4(x)\}$. The set is bounded by the maximum radii $r^* = 4$ in this finite collection.

Example 2.41. Let $X = \{1/n \mid n \in \mathbb{N}\}$ be a metric space equipped with

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}.$$

The set X is closed, as it is the whole space. It is also bounded, as $d(x, y) \leq 1$ for all $x, y \in X$. Let $G_n = \{1/n\}$. Each G_n is open, as $B_{1/2}(1/n) \subseteq G_n$. This makes G_n an open cover of X , as $X \subseteq \bigcup_{n \in \mathbb{N}} G_n = X$.

The set X fails to be compact, because this open cover has no finite subcover. Any finite cover $\{G_1, \dots, G_N\}$ would not “cover” $\{1/(N+1), 1/(N+2), \dots\} \subseteq X$.

We often restrict our attention to subsets of compact sets, so it would be nice to know if subsets of compact sets are compact. Unfortunately, this is not true in general, but it becomes true if we require the subset satisfy one condition.

Proposition 2.4. Closed subsets of compact sets are compact.

Proof. Suppose K in X is compact, and $F \subseteq K$ is closed in X . Let $\{V_\alpha\}$ be an arbitrary open cover of F . If we add F^c to this collection of open sets, we have an open cover $\Omega = \{V_{\alpha_1}, V_{\alpha_2}, \dots, F^c\}$ of K , because $K = F \cup F^c$.

$$K \subseteq \left(\bigcup_{\alpha} V_{\alpha} \right) \cup F^c.$$

The set K is compact, so there exists a finite open cover of Ω , $\Phi = \{V_{\alpha_1}, \dots, V_{\alpha_n}, F^c\}$. We can now remove F^c from Φ , resulting in a finite subcover for $\{V_\alpha\}$. \square

Corollary 2.3. If F is closed and K is compact, $F \cap K$ is compact.

An interesting property of compact sets is that if we have a decreasing sequence of nested compact intervals, than their intersection is nonempty. This result follows as a corollary of a more general result.

Proposition 2.5. If $\{K_\alpha\}$ is a collection of compact subsets of a metric space X such that the intersection of every finite subcollection of $\{K_\alpha\}$ is nonempty, then $\bigcap_{\alpha} K_\alpha \neq \emptyset$.

Proof. For the sake of contradiction, assume that $\bigcap_{\alpha} K_\alpha = \emptyset$. This means there is some fixed $K_1 \in \{K_\alpha\}$ such that no point of K_1 belongs to every K_α . Let $G_\alpha = K_\alpha^c$. The collection $\{G_\alpha\}$ forms an open cover of K_1 . Since K_1 is compact, there exists a finite subcover $\{G_{\alpha_1}, \dots, G_{\alpha_n}\}$.

$$K_1 \subseteq \bigcup_{i=1}^n G_{\alpha_i}$$

This inclusion, along with the definition of G_α implies that $K_1 \cap (\bigcap_{i=1}^n K_{\alpha_i}) = \emptyset$. This contradicts our assumption that every finite subcollection of $\{K_\alpha\}$ is nonempty. \square

Corollary 2.4 (Cantor’s Intersection Theorem). If $\{K_\alpha\}$ is a sequence of nonempty compact sets such $K_n \supseteq K_{n+1}$ for $n \in \mathbb{N}$, then $\bigcap_{i=1}^{\infty} K_n$ is not empty.

Example 2.42. Recall that open intervals in \mathbb{R} are not compact (Example 2.35), and do not satisfy the requirement of Cantor’s Intersection Theorem. Let $G_n = (0, 1/n)$. We have $G_{n+1} \subseteq G_n$ for all $n \in \mathbb{N}$, so $\{G_n\}$ is a decreasing nested sequence of intervals.

$$\dots \subseteq \left(0, \frac{1}{4} \right) \subseteq \left(0, \frac{1}{3} \right) \subseteq \left(0, \frac{1}{4} \right) \subseteq \left(0, \frac{1}{2} \right) \subseteq (0, 1)$$

The intersection of these sets is empty.

Proposition 2.6 (Bolzano-Weierstrass Property). If E is an infinite subset of a compact set K , then E has a limit point in K .

Proof. For the sake of contradiction, assume that no point of K is a limit point of E . For each $x \in K$, there exists some r such that $V_x = B_r(x) = \{x\}$ if $x \in E$, or $V_x = B_r(x) = \emptyset$ if $x \notin E$. The collection $\{V_x\}$ is an open cover of E . The set E is infinite, so we cannot find a finite subcover for $\{V_x\}$, as each set is at most a singleton. Because $E \subseteq K$, the same is true with respect to K , which contradicts the assumption that K is compact. \square

Informally, the Bolzano-Weierstrass Property tells us that we can approximate some point of K with points in an infinite subset E . Right now, this may not seem significant, but it will become important when we work in \mathbb{R}^n .

Remark 2.10. The Bolzano-Weierstrass Property is often referred to as *limit point compactness*. In the context of metric spaces, limit point compactness and compactness are equivalent²⁰, so in most real analysis texts the prior is never even given a name. As we'll see *much* later on, the two are not equivalent when we explore point-set topology in general (Section 16).

2.7 Compact Sets in \mathbb{R}^n

Now we can start working towards sufficient conditions for compactness in \mathbb{R}^n . This will culminate in the famed Heine-Borel Theorem. This theorem establishes sufficient conditions for compactness in \mathbb{R}^n . Specifically, the converses of Theorem 2.3 and Theorem 2.4 will hold in \mathbb{R}^n .

In order to make the proof of this result a bit more clear, we will show a series of results beforehand which build to the Heine-Borel Theorem.

Lemma 2.2. If $\{I_n\}$ is a sequence of closed intervals in \mathbb{R} , such that $I_n \supset I_{n+1}$ for $n \in \mathbb{N}$, then $\cap_{i=1}^n I_n \neq \emptyset$.



Figure 25: The first three intervals in the type of sequence $\{I_n\}$ described in Proposition 2.7.

Proof. Suppose $I_n = [a_n, b_n]$. Define E to be the set of all the a_n . The set E is nonempty and bounded above by b_1 , because

$$\cdots a_3 \leq a_2 \leq a_1 \leq b_1 \leq b_2 \leq b_3 \leq \cdots .$$

We have $E \subseteq \mathbb{R}$ and is bounded, so it has a supremum in \mathbb{R} . Let $x = \sup E$. For all $m, n \in \mathbb{N}$,

$$a_n \leq a_{m+n} \leq b_{m+n} \leq b_m,$$

so $x \leq b_m$ for all m . By the definition of $\sup E$, $a_m \leq x$. If $x \leq b_m$ and $a_m \leq x$, then $x \in [a_m, b_m] = I_m$ for all $m \in \mathbb{N}$. Therefore $x \in \cap_{n \in \mathbb{N}} I_n$. \square

Example 2.43. If we modify Example 2.42 so the open intervals are closed, then we have a sequence $\{I_n\}$ where $I_n = [0, 1/n]$. This gives $\cap_{n \in \mathbb{N}} I_n = \{0\} \neq \emptyset$. If we worked through the proof of Lemma 2.2 with this particular example, we would find that 0 is the supremum of the set of lower bounds of I_n .

²⁰Proving this requires some concepts that are beyond this treatment of metric spaces.

This result should look familiar. Cantor's Intersection Theorem states a similar result for compact sets. This in and of itself does not show that closed intervals are compact, but it should catch our attention, as closed intervals and compact sets share a noteworthy property. It should also be noted that this property of closed intervals follows from the least-upper-bound property of \mathbb{R} . This is one of the magical results we get because \mathbb{R} is complete. We can generalize Proposition 2.7 to *k-cells* in \mathbb{R}^k . An *k-cell* is the set of all points $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$ which satisfy $a_i \leq x_i \leq b_i$ for $i = 1, \dots, k$, where $a_i, b_i \in \mathbb{R}$, and $a_i \leq b_i$.

Lemma 2.3. Let $k \in \mathbb{N}$. If $\{I_n\}$ is a sequence of k -cells such that $I_n \supset I_{n+1}$ for all $n \in \mathbb{N}$, then $\cap_{n \in \mathbb{N}} I_n \neq \emptyset$.

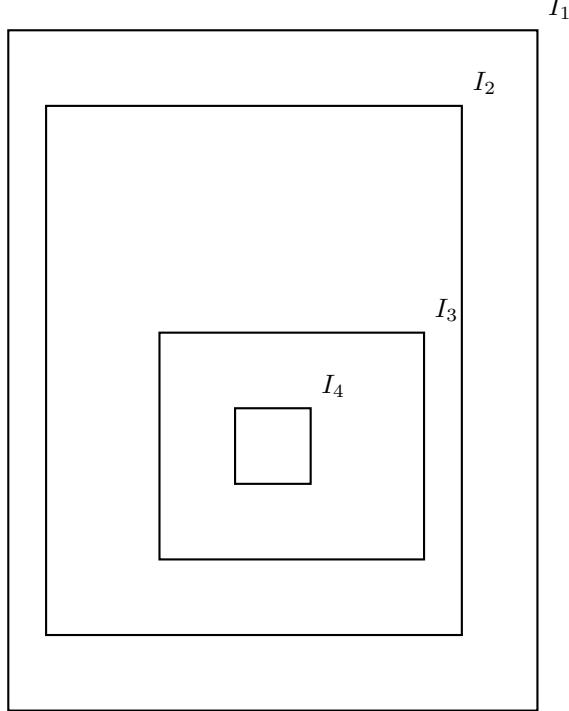


Figure 26: The first four 2-cells in the type of sequence $\{I_n\}$ described in Proposition 2.8.

Proof. Let I_n be the set of points $\mathbf{x} = (x_1, \dots, x_k)$ such that $a_{n,j} \leq x_j \leq b_{n,j}$ for $j = 1, \dots, k$ and $n \in \mathbb{N}$, and write $I_{n,j} = [a_{n,j}, b_{n,j}]$. By Lemma 2.2, for each $I_{n,j}$, $\cap_{n \in \mathbb{N}} I_{n,j} \neq \emptyset$, so there is some $x_j^* \in \cap_{n \in \mathbb{N}} I_{n,j}$ which satisfies

$$a_{n,j} \leq x_j^* \leq b_{n,j}$$

for $j = 1, \dots, k$ and $n \in \mathbb{N}$. If we set $\mathbf{x}^* = (x_1^*, \dots, x_k^*)$, then $\mathbf{x}^* \in I_n$ for all $n \in \mathbb{N}$. Therefore $\cap_{n \in \mathbb{N}} I_n \neq \emptyset$. \square

We will now prove that each k -cell is compact. The definition of a k -cell is equivalent to a closed and bounded set in \mathbb{R} , so this result will give us our sufficient conditions for compactness in \mathbb{R}^n . This result will give rise to the Heine-Borel Theorem which is an equivalence result, which will follow immediately from the compactness of k -cells, Theorem 2.3, and Theorem 2.4. That being said, the proof that each k -cell is compact is not immediate, and is on the more difficult side for a proof in an introductory analysis course.

Lemma 2.4. Every k -cell is compact.

Proof. Let $I = \{\mathbf{x} \in \mathbb{R}^k \mid a_j \leq x_j \leq b_j \ j = 1, \dots, k\}$ be a k -cell. Let δ be the maximum distance between any two points in I .

$$\max_{\mathbf{x}, \mathbf{y} \in I} d(\mathbf{x}, \mathbf{y}) = \delta = \left(\sum_{j=1}^k (b_j - a_j)^2 \right)^{1/2}$$

For all $\mathbf{x}, \mathbf{y} \in I$, $|\mathbf{x} - \mathbf{y}| \leq \delta$ (Figure 27 shows this for $k = 2$).

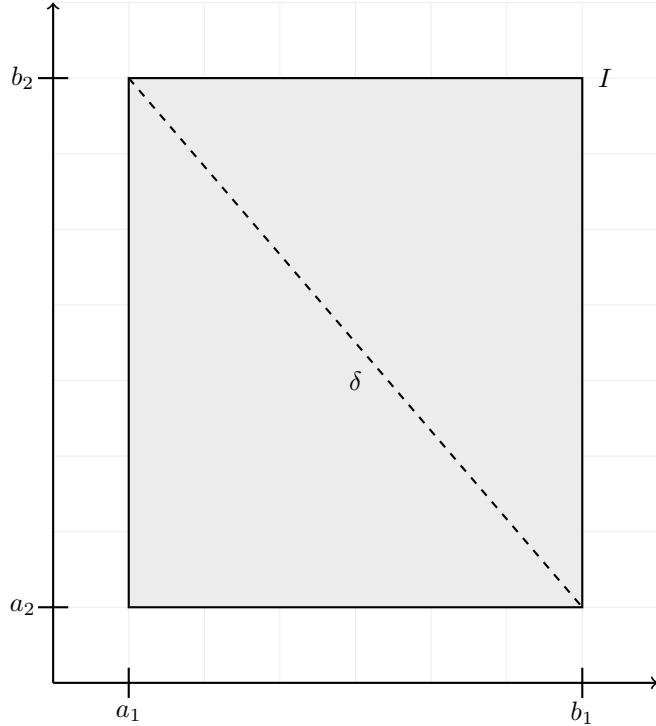


Figure 27: If I is a 2-cell.

For the sake of contradiction, assume that there exists some arbitrary open cover $\{G_\alpha\}$ of I which contains no finite subcover of I . Let $c_j = (a_j + b_j)/2$. The intervals $[a_j, c_j]$ and $[c_j, b_j]$ give rise to 2^k k -cells Q_i whose union is I . If each of these cells Q_i could be covered by a finite subcollection of $\{G_\alpha\}$, then I could be covered by the union of all these finite subcollections, which is finite. Since we've assumed K is not compact, then it must be that there exists at least one Q_i , call it I_1 , that cannot be covered by any finite subcollection of $\{G_\alpha\}$. Now we divide I_1 into 2^k k -cells and repeat this process indefinitely, giving us a sequence $\{I_n\}$ (Figure 29). This sequence of k -cells was constructed to have three properties:

1. $I \supset I_1 \supset I_2 \supset I_3 \supset \dots$
2. I_n cannot be covered by any finite subcollection of $\{G_\alpha\}$, otherwise K would not be compact.²¹
3. If $\mathbf{x}, \mathbf{y} \in I_n$, then $|\mathbf{x} - \mathbf{y}| \leq \delta/2^n$.²²

By the first property of this sequence, we can invoke Lemma 2.3 to determine $\cap_{n \in \mathbb{N}} I_n \neq \emptyset$. This means there exists some $\mathbf{x}^* \in \mathbb{R}^k$ such that $\mathbf{x}^* \in I_n$ for all $n \in \mathbb{N}$. There must be some α such that $\mathbf{x}^* \in G_\alpha$, otherwise

²¹This follows from the same reasoning applied to the 2^k k -cells Q_i we initially chose I_1 from.

²²When we divided I , the length of the diagonal for I_1 became half of δ . When we divide I_1 , the length of the diagonal of I_2 became half of that of I_1 . This means we can always write the diagonal of I_n in terms of powers of $1/2$ and δ .

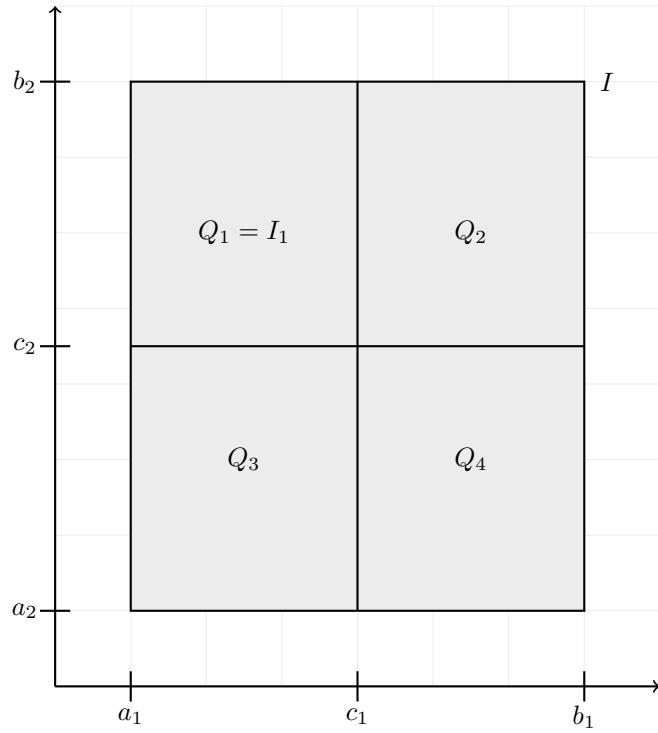


Figure 28: We partition I into 2^k cells Q_i . If I is not compact, then there is some Q_i which is not compact. Suppose in this case Q_1 is not compact, and call it I_1 .

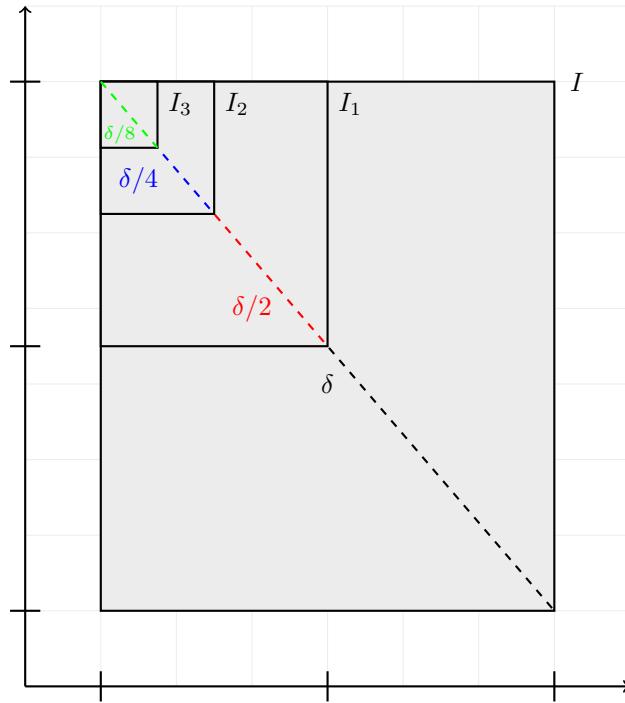


Figure 29: The first three 2-cells in the sequence $\{I_n\}$.

$\{G_\alpha\}$ would not be an open cover of I . Since G_α is open, there exists some $r > 0$ such that $B_r(x) \subseteq G_\alpha$. Alternatively, we could say that $|\mathbf{y} - \mathbf{x}^*| < r$ implies $y \in G_\alpha$ by the definition of $B_r(x)$. If we take n to be so large that $\delta/2^n < r$,²³ then $I_n \subseteq G_\alpha$, but this would mean I_n has a finite subcover. This contradicts property 2 of our sequence $\{I_n\}$, thereby contradicting the assumption that K is not compact. \square

Example 2.44. In order to make this proof a bit more concrete we'll walk through it with the 2-cell define by $I = [0, 1] \times [0, 1]$, and a specific open cover.²⁴ Let our open cover $G_\alpha = B_{0.01}(\alpha)$ for $\alpha \in I$.

$$I \subseteq \bigcup_{\alpha \in I} G_\alpha = \bigcup_{\alpha \in I} B_{0.01}(\alpha)$$

Assume that this open cover has no finite subcover. We have

$$\delta = ((1-0)^2 + (1-0)^2)^{1/2} = \sqrt{2}.$$

We divide I into four 2-cells: $Q_1 = [0, 1/2] \times [0, 1/2]$, $Q_2 = [0, 1/2] \times [1/2, 1]$, $Q_3 = [0.5, 1] \times [0, 1/2]$, and $Q_4 = [1/2, 1] \times [1/2, 1]$. If $\{G_\alpha\}$ has no finite subcover for I , then the same can be said for one of these Q_i . Suppose this is the case for Q_1 , and let $I_1 = Q_1 = [0, 1/2] \times [0, 1/2]$. Repeat this process seven times until we arrive at $I_7 = [0, 1/256] \times [0, 1/256]$. The maximum distance between any two points in I_7 is

$$((1/256 - 0)^2 + (1/256 - 0)^2)^{1/2} \approx 0.0055 < 0.01.$$

Therefore, we can cover I_7 with a single $B_{0.01}(\alpha) \in \{G_\alpha\}$ for any $\alpha \in I_7$. But this means we can cover I_6 with four elements in $\{G_\alpha\}$,²⁵ and cover I_5 with 4^2 elements in $\{G_\alpha\}$, etc. We can cover I with 4^7 elements in $\{G_\alpha\}$. This contradicts the assumption that $\{G_\alpha\}$ has no finite subcover.

One of the subtler, but nevertheless important, parts of this process is that we could repeatedly divide the cells until we found a cell that fit in a open ball of radius 0.01. We will always be able to do this because of the Archimedean property as discussed in Footnote 22. It is important, that you do not associate this particular “trick” with the fact that I is a subset of \mathbb{R}^2 . We are discussing the Archimedean property in the context of distances and radii, so we're actually using the fact that \mathbb{R}^2 is equipped with a distance function $d : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow [0, \infty]$. We're using the fact that the codomain of d has the Archimedean property!

We have now shown that every k -cell is compact. The proof is made more manageable with illustrations, but is still rather technical. Our contradiction came from the fact that as n becomes large, I_n becomes small. Eventually I_n will be so small, that it must have a finite open subcover. This contradicts the assumption that I is not compact.

Theorem 2.5 (Heine-Borel Theorem). A set E in \mathbb{R}^n is compact if and only if it is closed and bounded.

Proof.

(\Rightarrow) All compact sets are closed and bounded by Theorems 2.3 and 2.4.

²³We can always find such an n . If not, $2^n \leq \delta/r$ for all $n \in \mathbb{N}$. This can't be though because \mathbb{R} has the Archimedean property (Theorem 1.3).

²⁴In order to give such an example, we need to specify an open cover to work with. In doing so, we're sort of shooting ourselves in the foot. The whole point of compactness is that *every* open cover has a finite subcover. What we're really proving in this example, is this specific open cover has no finite subcover.

²⁵I'm playing a little fast and loose here with which exact elements, because I'm not specifying where the open balls are centered.

(\Leftarrow) If E is closed and bounded then $E \subseteq I$ for some n -cell. Any closed subset of a compact set is compact by Proposition 2.4, so E is compact.

□

Example 2.45. Any closed interval $[a, b] \subseteq \mathbb{R}$ is compact.

Theorem 2.6 (Bolzano–Weierstrass Theorem). Every bounded infinite subset of \mathbb{R}^n has a limit point in \mathbb{R}^n .

Proof. Let $E \subseteq \mathbb{R}^n$ be bounded and infinite. There is some n -cell $I \subseteq \mathbb{R}^k$, such that $E \subseteq I$. By Lemma 2.4 I is compact. We know apply Proposition 2.6 to conclude that E has a limit point in I , which is also in \mathbb{R}^k . □

Example 2.46. The set $(a, b) \subseteq \mathbb{R}$ is infinite and bounded. It has an infinite number of limit points in \mathbb{R}^n ,²⁶ including a and b .

2.8 Exercises

Exercise 2.1. Show that the set of all algebraic numbers is countably infinite.

Exercise 2.2. Show that the set of all binary numbers with infinite digits is uncountably infinite.

Exercise 2.3. Verify that the taxi-cab metric on \mathbb{R}^n is a valid metric space.

Exercise 2.4. Let X be an infinite set with the metric,

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}.$$

Prove that (X, d) is a metric space. Which subsets of X are open? Which are closed?

Exercise 2.5. Prove that E° is open.

Exercise 2.6. Prove that E is open if and only if $E = E^\circ$.

Exercise 2.7. If $G \subseteq E$ and G is open, prove that $G \subseteq E^\circ$

Exercise 2.8. Prove that $(E^\circ)^c = \overline{E^c}$.

Exercise 2.9. Find an example of a set E in a metric space such that $E^\circ \neq (\bar{E})^\circ$.

Exercise 2.10. Find an example of a set E in a metric space such that $\bar{E} \neq \overline{E^\circ}$.

Exercise 2.11. Prove that ∂E is closed.

Exercise 2.12. Prove that $\partial(E^\circ) \subseteq \partial E$, and $\partial(\bar{E}) \subseteq \partial E$

Exercise 2.13. Prove that $\partial E = \partial(E^c)$.

Exercise 2.14. Suppose E is closed. Show that $(\partial E)^\circ = \emptyset$.

Exercise 2.15. Prove that $\partial(\partial E) \subseteq \partial E$. When will the sets be equal?

²⁶In fact, the set of limit points is $[a, b] \subseteq \mathbb{R}$

Exercise 2.16. Prove that E is closed if and only if $E \cap \partial E = \emptyset$.

Exercise 2.17. Prove that $\partial E = \emptyset$ if and only if E is closed and open.

Exercise 2.18. Prove that $\bar{E} = E \cup \partial E$.

Exercise 2.19. Prove that $(\partial \bar{E})^\circ = \emptyset$.

3 Sequences and Series

Now that we are intimately familiar with the behavior of metric spaces, we can discuss a topic that may be familiar from calculus – sequences and series. Metric spaces will allow us to rigorously define convergence, and the properties related to the convergence of sequences and series. While we will derive some results and examples in general metric spaces, we will also start introducing results specific to \mathbb{R} .

3.1 Convergence

Definition 3.1. Let X be a metric space. A *sequence* $\{x_n\}$ is a function from $f : \mathbb{N} \rightarrow X$. We will sometimes refer to an entire sequence as $x_n = f(n)$.

Using an arbitrary metric space X in this definition means that we, once again, always need to pay attention to what metric space we are in. We saw this with open sets, closed sets, and compact sets, and we will see it again. It will become especially relevant when determining if sequences converge.

Example 3.1. Let $x_n = 1/n$ be a sequence in \mathbb{R} . The first several terms in this sequence are

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$$

This set is also a sequence in \mathbb{Q} . This sequence is not defined in \mathbb{Z} or \mathbb{N} , as neither of these sets has fractions.

Example 3.2. Let $x_n = 2$ be a sequence in \mathbb{R} . This constant sequence always takes on the value 2. This sequence is also a sequence in \mathbb{N}, \mathbb{Z} , and \mathbb{Q} .

Example 3.3. Let $x_n = (-1)^n$ be a sequence in \mathbb{R} . This sequence alternates between -1 and 1 for all values in \mathbb{N} .

Now we are ready to formalize what it means for a sequence to converge. When the idea of convergence is first introduced, you often hear phrases like “arbitrarily close”. If a sequence converges to some point $x \in X$, we can *always* get closer to x . For any value in $\{x_n\}$, we can find other points “later” in the sequence $\{x_n\}$ that is even closer. If convergence is a recipe, then these are the ingredients:

1. No matter how “close” we get, we can always get closer with another point in $\{x_n\}$. Fortunately, we’re in a metric space (X, d) , so we can use d to determine how close we are.
2. Well actually, it cannot be *any* other points “later” in $\{x_n\}$. For instance, suppose we have the following sequence:

$$1, \frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \dots$$

The even terms of this series are getting closer to 0, while the odd terms are growing. The latter fact means this sequence doesn’t converge. This happens because not *all* the points “later” in $\{x_n\}$ are closer.

These two ingredients will correspond to the ε and N in our definition.

Definition 3.2. A sequence $\{x_n\}$ in a metric space X *converges (in X)* if there exists an $x \in X$ such that *for all* $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ *for all* $n \geq N$. We will call x the *limit* of $\{x_n\}$, and write either $x_n \rightarrow x$, or

$$\lim_{n \rightarrow \infty} x_n = x.$$

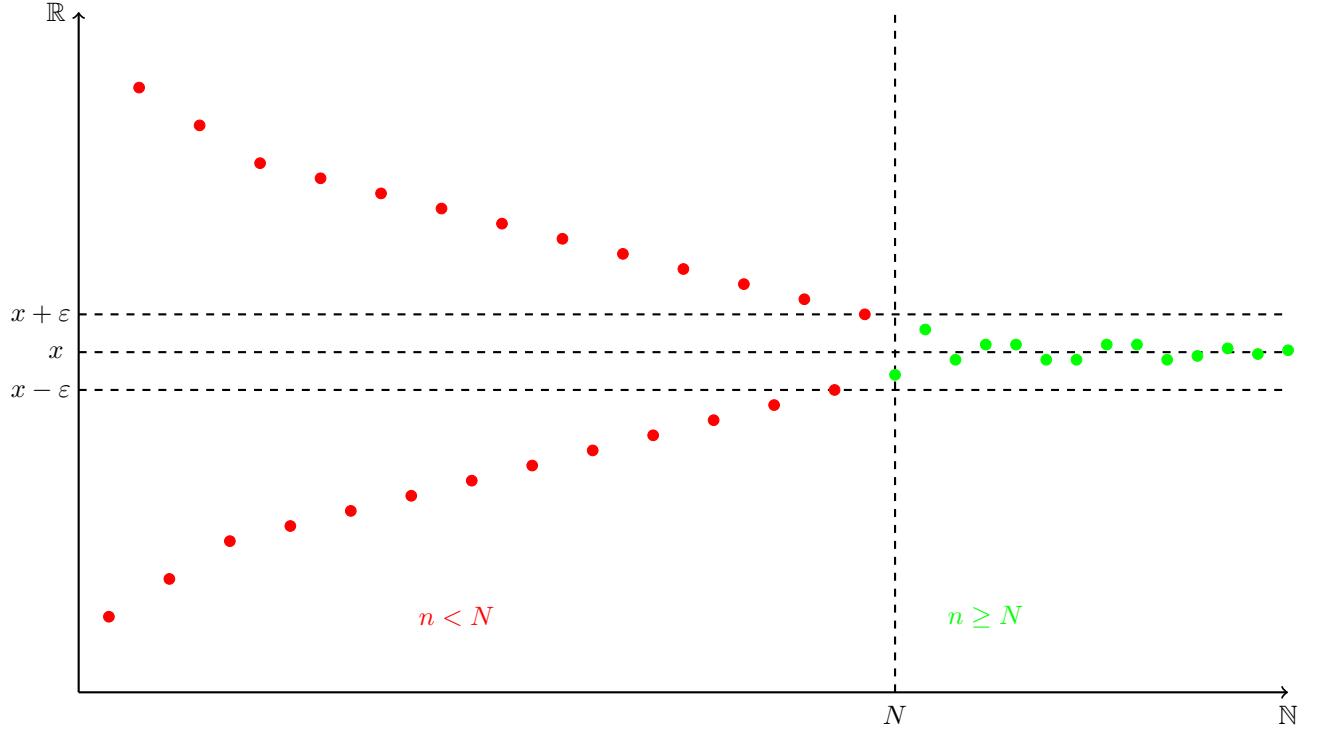


Figure 30: A convergent sequence $\{x_n\}$ in \mathbb{R} . No matter how small we take ε to be, we can always find some N such that all $d(x_n, x) = |x_n - x| < \varepsilon$ for all $n \geq N$. We could also write $x_n \in B_\varepsilon(x) = (x - \varepsilon, x + \varepsilon)$.

We can think of the convergence of a sequence in the context of a hypothetical game. Suppose you and a friend have some convergent sequence $\{x_n\}$ in X . Your friend says some small number ε , and challenges you to find an N such that $d(x_n, x) < \varepsilon$ for all $n \geq N$. By the definition of convergence, you can do this. This frustrates your friend, so he demands you do it for an even smaller value of ε . Unfortunately for them, you will always be able to find such an N . No matter how small ε is, you will be able to do this.

Remark 3.1. We can formulate an equivalent definition of convergence using open balls. If for all $\varepsilon > 0$, $d(x_n, x) < \varepsilon$ whenever $n \geq N$, then we could also say $x_n \in B_\varepsilon(x)$ for all $n \geq N$.

Example 3.4. The sequence $x_n = 1/n$ converges to 0 in \mathbb{R} . Suppose your friend lets $\varepsilon = 0.01$, and asks you to find an $N \in \mathbb{N}$ such that

$$d(1/n, 0) = |1/n - 0| = 1/n < \varepsilon = 0.01$$

for all $n \geq N$. If you let $N = 101$, then you have done this.

$$1/101 < 0.01$$

$$1/102 < 0.01$$

$$1/103 < 0.01$$

⋮

Your friend then gives you $\varepsilon = 0.001$. In this case, let $N = 1001$.

$$1/1001 < 0.01$$

$$1/1002 < 0.01$$

$$1/1003 < 0.01$$

⋮

You're already bored of this game, so you get an idea. Maybe you can find some function of ε that will give you your value of N . In order to do this, you just manipulate the inequality you must satisfy.

$$\begin{aligned} d(1/n, 0) &< \varepsilon \\ |1/n - 0| &< \varepsilon \\ 1/n &< \varepsilon \\ n &> \varepsilon^{-1} \end{aligned}$$

We just let $N = \varepsilon^{-1} + 1$. This way, for all $n \geq N$, we have $n > \varepsilon^{-1}$, which implies $d(1/n, 0) < \varepsilon$.

Remark 3.2. The definition of convergence has two inequalities. The inequality $d(x_n, x) < \varepsilon$ is strict, while $n \geq N$ is not. In the grand scheme of things, it doesn't matter if these are strict or not. If we instead had $d(x_n, x) \leq \varepsilon$, then we could just have taken $N = \varepsilon^{-1}$ in the previous example. Alternatively, if we had $d(x_n, x) < \varepsilon$ and $n > N$, then $N = \varepsilon^{-1}$ would work as well. I'm going to try very hard to stick with the inequalities in the definition, but I may make a mistake. Just know that it doesn't change the results of proofs at all. It does mean we need to be a little careful when using open balls though, because they are open sets.

Example 3.5. We can verify that the sequence $x_n = (n+1)/(n-1)$ converges to 1 in \mathbb{R} . Let $\varepsilon > 0$. We want to find the value of N in terms of ε that satisfies $d(x_n, x) < \varepsilon$ for all $n \geq N$. First notice that

$$d(x_n, x) = \left| \frac{n-1}{n+1} - 1 \right| = \left| \frac{-2}{n+1} \right| = \frac{2}{n+1}.$$

We can satisfy $\frac{2}{n+1} < \varepsilon$ with $N = 2/\varepsilon - 1$.²⁷ We have that $d(x_n, x) < \varepsilon$ for all $n \geq 2/\varepsilon - 1$.

Example 3.6. Let \mathbb{Z} be a metric space equipped with the p -adic metric (see Example 2.4). The sequence $x_n = p^n$ converges to 0 for any prime p . We have

$$d(x_n, 0) = p^{-\max\{m \in \mathbb{N} \mid p^m \mid (p^n - 0)\}} = p^{-n}.$$

If we let $N = \ln(2/\varepsilon)/\ln(p)$, then for all $\varepsilon > 0$ we have

$$d(x_n, 0) \leq d(x_N, 0) = p^{-\frac{\ln(2/\varepsilon)}{\ln(p)}} = p^{\frac{\ln(2/\varepsilon)}{\ln(p)}} = p^{\log_p(2/\varepsilon)} = \frac{2}{\varepsilon} < \varepsilon$$

for all $n \geq N$. Let's let $p = 2$. The sequence $x_n = 2^n$ will converge in \mathbb{Z} with the 2-adic metric.

$$\{2^n\} = \{2, 4, 8, 16, 32, 64, 128, \dots\}$$

²⁷You may be thinking that it isn't always true that this N will be in \mathbb{N} . That's fine. We could just round the answer to get a whole number that satisfies the inequality. Generally, we're not too worried about this.

n	x_n	$d(x_n, 0)$
1	2	1/2
2	4	1/4
3	8	1/8
4	16	1/16
5	32	1/32
6	64	1/64

Example 3.7. The series $x_n = (1 + 1/n)^n$ converges to e in \mathbb{R} . This series does not converge in \mathbb{Q} , because $e \notin \mathbb{Q}$.

Example 3.8. Let $X = (0, 1]$ be equipped with the Euclidean metric. The sequence $x_n = 1/n$ does not converge in X , because $0 \notin X$.

Example 3.9. Let $x_n = c$ for some constant $c \in \mathbb{R}$. This type of sequence is often called a *constant sequence*. All constant sequences converge to c in \mathbb{R} , as

$$|x_n - c| = |c - c| = 0 < \varepsilon$$

for all ε . Constant sequences are the only sequences that converge in \mathbb{Z} equipped with the Euclidean metric.

Remark 3.3 (Where's My Limit?!). These last two examples really emphasize the fact that the limit must be in the same metric space as our sequence. If our sequence is defined by some $f : \mathbb{N} \rightarrow X$, we need $x \in X$! Note that this is different from requiring that the value x is actually realized by our sequence. That is, it needn't be the case that f is in the image of X . We just need it to be in the codomain of X . In Example 3.4, $0 \notin f(\mathbb{N})$ (the image/range of \mathbb{N}), but it is in \mathbb{R} , and that's all that matters.

3.2 Properties Related to Convergence

Now we'll cover some basic properties of sequences and convergence. We will also prove several results that are specific to sequences in \mathbb{R} , many of which should be familiar from calculus.

Proposition 3.1. Let $\{x_n\}$ be a sequence in a metric space X . The sequence $\{x_n\}$ converges to $x \in X$ if and only if every open ball around p contains x_n for all but finitely many n .

Proof.

(\Rightarrow) Suppose $x_n \rightarrow x$, and let $B_r(x)$ be some open ball around x . For some $\varepsilon = r > 0$, $x_n \in B_r(x)$ for all $n \geq N$ (see Remark 3.1). Therefore $B_r(x)$ contains x_n for all but finitely n , those being $\{x_1, \dots, x_{N-1}\}$.

(\Leftarrow) Suppose every open ball around x contains all but finitely many x_n . Fix $\varepsilon > 0$, and observe $B_\varepsilon(x)$. By our assumption, there exists an N such that $x_n \in B_\varepsilon(x)$ for $n \geq N$. Therefore we have $d(x_n, x) < \varepsilon$ if $n \geq N$, so $x_n \rightarrow x$.

□

Remark 3.4 (Limit vs. Limit Point). This result implies that if $x_n \rightarrow x$, then x is a limit point of the range of x_n . The converse is not necessarily true. Take the sequence $x_n = (-1)^n(1 + 1/n)$ as an example.

$$-2, \frac{3}{2}, -\frac{4}{3}, \frac{5}{4}, -\frac{6}{5}, \dots$$

The range of this sequence has two limit points, $\{-1, 1\}$. Neither of these are limits of the sequence, as this sequence fails to converge because it alternates.

Example 3.10. Let $x_n = 1/n$ be a sequence in \mathbb{R} . Example 3.4 showed that $x_n \rightarrow 0$. For the open ball $B_{0.01}(0)$ we have:

$$B_{0.01}(0) = \left\{ \frac{1}{101}, \frac{1}{102}, \frac{1}{103}, \dots \right\},$$

where our finite set of points not in $B_{0.01}(0)$ is

$$\{x_n\} \setminus B_{0.01}(0) = \left\{ 1, \frac{1}{2}, \dots, \frac{1}{100} \right\}.$$

Proposition 3.2 (Uniqueness of Limits). Let $\{x_n\}$ be a sequence in a metric space X . If $x, x' \in X$, and if $\{x_n\}$ converges to x and x' , then $x = x'$.

Remark 3.5 (Playing with ε). Before we prove this, we should highlight a “trick” we will use. It is the most common technique used in proofs involving ε . If we know for all $\varepsilon > 0$, $d(x_n, x) < \varepsilon$ for all $n \geq N$, then we can use *any* $\varepsilon \in (0, \infty]$. This means we have $d(x_n, x) < f(\varepsilon)$ for all $n \geq N$, where $f : (0, \infty] \rightarrow (0, \infty]$. Return to the hypothetical game you are playing with your friend. At first, your friend wants to let $\varepsilon = 0.01$, but then he thinks “no that’s too big, let me cut it in half”, and he uses $\varepsilon/2 = 0.005$. The number $\varepsilon/2 > 0$ so it still is valid. He could even say “let me square it, divide it by 4, and then add π ”. In this case $f : (0, \infty] \rightarrow (0, \infty]$ is defined as $f(\varepsilon) = \varepsilon^2/4 + \pi$, and is still valid because $f(\varepsilon) > 0$.

Why would we want to do this? We may want to show something converges, and do so using inequalities we already know that involve ε . We want our end result to show that $d(x_n, x) < \varepsilon$, so we need to choose our initial inequalities involving ε such that they yield a single ε . This probably isn’t too clear right now, but this next proof, and that for Theorem 3.2 will hopefully make this more clear.

Proof. Assume that $x_n \rightarrow x$, and $x_n \rightarrow x'$. Let $\varepsilon > 0$. There exists $N', N'' \in \mathbb{N}$ such that

$$\begin{aligned} n \geq N &\implies d(x_n, x) < \varepsilon/2, \\ n \geq N' &\implies d(x_n, x') < \varepsilon/2. \end{aligned}$$

If we let $N = \max\{N', N''\}$, then using the triangle inequality gives

$$d(x, x') \leq d(x, x_n) + d(x_n, x') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

for all $n \geq N$. If this holds for all $\varepsilon > 0$, then it must be that $d(x, x') = 0$, so $x = x'$. \square

Definition 3.3. A set $\{x_n\}$ in a metric space X is *bounded* if its range is bounded. That is, if $f : \mathbb{N} \rightarrow \mathbb{R}$ is the function corresponding to x_n , the set $f(\mathbb{N})$ is bounded in X .

Example 3.11. The sequence $x_n = 1/n$ in \mathbb{R} is bounded. We have a range of $f(\mathbb{N}) \subseteq (0, 1]$, so the range is clearly bounded.

Example 3.12. The sequence 2^n in the 2-adic metric (see Example 3.6) is bounded. The range of this sequence is $\{2, 4, 8, 16, \dots\}$, but we have $d(x, y) \leq 1/2$ for all $x, y \in \{2, 4, 8, 16, \dots\}$.

Proposition 3.3. Let $\{x_n\}$ be a sequence in a metric space X . If $\{x_n\}$ converges, then $\{x_n\}$ is bounded

²⁸Because we picked $\varepsilon/2$, they added to the desired ε . If we hadn’t done this, then we would have $d(x, x') < 2\varepsilon$. Sometimes, people are fine with this and just argue “well $2\varepsilon \rightarrow 0$ as $\varepsilon \rightarrow 0$ ”. This isn’t wrong, but it’s not exactly kosher. It’s best to just satisfy the definition without having to use a limiting process with ε .

Proof. Suppose $x_n \rightarrow x$. For all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n > N$. We have $1 > 0$, so we can let $\varepsilon = 1$. There is some N such that $d(x_n, x) < 1$ for all $n > N$. Let

$$r = \max\{1, d(p_1, p), \dots, d(p_N, p)\}.$$

We have $d(p_n, p) \leq r$ for all $n \in \mathbb{N}$, so the sequence is bounded. \square

Example 3.13. Let's work through this proof with an actual example. Let $x_n = 1/n$ in \mathbb{R} .²⁹ We know $x_n \rightarrow 0$, so there exists an N such that $d(x_n, 0) < 1$ for all $n \geq N$. In this case $N = 2$. We let $r = \max\{1, d(p_1, p)\} = \max\{1, 1\} = 1$. We have $d(x_n, 0) \leq 1$ for all $n \in \mathbb{N}$.

This proof is short, but interesting. We know every open ball around our limit 0 contains all but finitely many n (Proposition 3.1). The maximum distance becomes that between the limit and the finite points excluded from some arbitrary $B_\varepsilon(0)$.³⁰

Example 3.14. The converse of Proposition 3.3 is not true. Take $x_n = (-1)^n$ in \mathbb{R} . The sequence is bounded, as the range of the sequence is $\{-1, 1\}$. Despite being bounded, the sequence does not converge as it alternates between 1 and -1 .

Recall that in Section 2, we sometimes discussed approximating a limit point of some set E with elements in E . This was not the most formal of discussions, but our next theorem will make this fact explicit. The theorem is a statement about existence, and does not provide an actual construction of the sequence in claims exists. It's *very* important to be able to distinguish when a result does one of these, but not the other. It can often have practical implications for problem solving.

Theorem 3.1. Let $\{x_n\}$ be a sequence in a metric space X . If $E \subseteq X$ and if x is a limit point of E , then there is a sequence $\{x_n\}$ in E such that $x_n \rightarrow x$ in X .

Proof. The point $x \in X$ is a limit point of E , so all $B_r(x)$ contain some points of E that are not x . If we let $r = 1/n$, then we have that there exists some point $x_n \in B_{1/n}(x)$. Equivalently, $d(x_n, x) < 1/n$. For all $\varepsilon > 0$, choose N such that $N\varepsilon > 1$.³¹ This way, we have

$$d(x_n, x) < \frac{1}{1/\varepsilon} = \varepsilon$$

for all $n > N$. Therefor $x_n \rightarrow x$. \square

Example 3.15. Let $E = (0, 1] \subseteq \mathbb{R}$ and, $x_n = 1/n$ in E . The point $0 \in \mathbb{R}$ is a limit point of E . The sequence $\{x_n\}$ converges to 0 in X . Note that $\{x_n\}$ does not converge in E (Example 3.7).

Corollary 3.1. Let E be a subset of a metric space X . If E is dense in X , then for all $x \in X$, there exists some sequence $\{x_n\}$ in E such that $x_n \rightarrow x$.

Corollary 3.1 is an amazingly useful result once we become more comfortable with limits (and continuity). We may want to prove that some set X has a certain property, which could require we verify some condition for each $x \in X$. If X has some dense subset Y , then we could just prove that the property exists for a limit of a sequence in E , because each point in X is a limit of such a sequence! This sounds like it would be more a more complicated method of proof, but that is because we are just starting to build the toolkit required

²⁹Yes, I will keep using this very trite example.

³⁰In this case we just took $\varepsilon = 1$, but any other number would work just fine.

³¹This gives $N > 1/\varepsilon$.

to work with limits. If points in E are easier to work with than those in X ,³² then it may just be easier to take limits of them.

Example 3.16. The set \mathbb{Q} is dense in \mathbb{R} . This means that every number in \mathbb{R} is the limit of a sequence in \mathbb{Q} , including irrational numbers. We saw this already in Example 3.6. In this specific case, we actually can write down the sequence. Even if we do not know the explicit form of the sequence, we still know it at least exists.³³ You probably don't know any sequence in \mathbb{Q} that converges to π , but you do know that such a sequence exists because of Corollary 3.1.³⁴

The next two results pertain only to sequences in \mathbb{R}^n . The first is a set of familiar results from calculus, and the second gives us the means to determine if sequences of real vectors converge.

Theorem 3.2. Suppose $\{x_n\}$ and $\{y_n\}$ are sequences in \mathbb{R} , $\lim_{n \rightarrow \infty} x_n = x$, and $\lim_{n \rightarrow \infty} y_n = y$. Then

1. $\lim_{n \rightarrow \infty} (x_n + y_n) = x + y$;
2. $\lim_{n \rightarrow \infty} cx_n = cx$, for all $c \in \mathbb{R}$;
3. $\lim_{n \rightarrow \infty} (x_n y_n) = xy$, for all $c \in \mathbb{R}$;
4. $\lim_{n \rightarrow \infty} 1/x_n = 1/x$, provided $s_n \neq 0$ for all $n \in \mathbb{N}$, and $s \neq 0$.

Proof.

1. There exists $B_1, B_2 \in \mathbb{N}$ such that

$$\begin{aligned} n \geq B_1 &\implies |x_n - x| < \frac{\varepsilon}{2}, \\ n \geq B_2 &\implies |y_n - y| < \frac{\varepsilon}{2}. \end{aligned}$$

If we set $N = \max\{B_1, B_2\}$, then for all $n \geq N$ we have

$$|(s_n + t_n) - (s + t)| = |(s_n - s) + (t_n - t)| \leq |(s_n - s)| + |(t_n - t)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Therefore $x_n + y_n \rightarrow x + y$.

2. Given $\varepsilon > 0$, there exists a $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon/c$ for all $n \geq N$. This means for all $n \geq N$, we have

$$|cx_n - cx| = c|x_n - x| < c \cdot \frac{\varepsilon}{c} = \varepsilon,$$

so $cx_n \rightarrow cx$.

3. There exists $B_1, B_2 \in \mathbb{N}$ such that

$$\begin{aligned} n \geq B_1 &\implies |x_n - x| < \frac{\sqrt{2}}{\varepsilon}, \\ n \geq B_2 &\implies |y_n - y| < \frac{\sqrt{2}}{\varepsilon}. \end{aligned}$$

³²Nealy every time it's actually the points in $X \setminus E$ that are the ones that are harder to work with.

³³This could be considered a drawback of this proof. It is not a proof via construction, so we don't have some blueprint that tells us how to find the sequence.

³⁴I cannot think of a sequence that does this off the top of my head, but the series $\sum_{n=0}^{\infty} \frac{4(-1)^k}{2k+1}$ will. We'll prove this in Section 7.

If we let $N = \max\{B_1, B_2\}$, then

$$\begin{aligned}
|x_n y_n - xy| &= |x_n y_n - xy + 0 + 0 + 0| \\
&= |x_n y_n - xy + (xy - yx) + (x_n y - yx_n) + (y_n x - xy_n)| \\
&= |(x_n y_n - xy_n - x_n y + xy) + (yx_n - yx) + (xy_n - xy)| \\
&= |(x_n - x)(y_n - y) + y(x_n - x) + x(y_n - y)| \\
&\leq |(x_n - x)(y_n - y)| + |y(x_n - x) + x(y_n - y)| \\
&\leq |(x_n - x)(y_n - y)| \\
&< \sqrt{\varepsilon} \sqrt{\varepsilon} \\
&= \varepsilon,
\end{aligned}$$

for all $n \geq N$. This gives $x_n y_n \rightarrow xy$.³⁵

4. There exists an $B_1, B_2 \in \mathbb{N}$ such that

$$\begin{aligned}
n \geq B_1 &\implies |x_n - x| < \frac{|x|}{2}, \\
n \geq B_2 &\implies |x - x_n| < \frac{x^2}{2} \varepsilon.
\end{aligned}$$

Note that we can manipulate one of this inequalities using the triangle inequality:

$$\begin{aligned}
|x_n - x| &< \frac{|x|}{2} \\
|x| - |x_n| &< \frac{|x|}{2} \\
||x| - |x_n|| &< \frac{|x|}{2} \\
-\frac{|x|}{2} &< |x| - |x_n| < \frac{|x|}{2} \\
\frac{|x|}{2} &< |x_n| < \frac{3|x|}{2} \\
\frac{1}{|x_n|} &< \frac{2}{|x|}
\end{aligned}$$

If we let $N = \max\{B_1, B_2\}$, then

$$\begin{aligned}
\left| \frac{1}{x_n} - \frac{1}{x} \right| &= \left| \frac{1}{x} \frac{1}{x_n} (x - x_n) \right| \\
&= \frac{1}{|x|} \frac{1}{|x_n|} |x - x_n| \\
&< \frac{1}{|x|} \frac{1}{|x|} \frac{|x|^2 \varepsilon}{2} \\
&= \varepsilon,
\end{aligned}$$

for all $n \geq N$. This gives $1/x_n \rightarrow 1/x$.

³⁵A very common “trick” used in almost all nontrivial proofs involving ε is to add and subtract the same term within an absolute value, in effect adding zero. You can then rearrange the terms and use the triangle inequality to “separate” terms in the absolute value. I call it a “trick”, because when I first saw someone do it I thought “well how the hell was I ever supposed to know to do that without ever seeing it done!”

□

Remark 3.6. Part 1 and 2 of Theorem 3.2 allow us immediately to perform subtraction using limits. Parts 3 and 4 allow us to do the same with division.

Remark 3.7 (Linearity of Limits). The first two parts of Theorem 3.2 allow us to conclude that limits are linear. This shouldn't be news, as it comes up in any standard calculus course, but it has powerful implication. We will use limiting processes to define differentiation and integration, so the linearity of limits will give rise to the linearity of these two operations as well.

Proposition 3.4. Suppose $\{\mathbf{x}_n\}$ is a sequence in \mathbb{R}^k where $\mathbf{x}_n = (x_{1,n}, \dots, x_{k,n})$. The sequence $\{\mathbf{x}_n\}$ converges to $\mathbf{x} = (x_1, \dots, x_n)$ if and only if $x_{j,n} \rightarrow x_j$ for $j = 1, \dots, k$.

Proof.

(\Rightarrow) Suppose $\mathbf{x}_n \rightarrow \mathbf{x}$. For all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$|\mathbf{x}_n - \mathbf{x}| < \varepsilon$$

for all $n \geq N$. This gives that $x_{j,n} \rightarrow x_j$, as

$$\begin{aligned} |x_{j,n} - x_j| &< ((x_{1,n} - x_1)^2 + \dots + (x_{k,n} - x_k)^2)^{1/2} \\ &= |\mathbf{x}_n - \mathbf{x}| \\ &< \varepsilon \end{aligned}$$

for $j = 1, \dots, k$.

(\Leftarrow) Suppose $x_{j,n} \rightarrow x_j$ for $j = 1, \dots, k$. For all $\varepsilon > 0$, there is some $N \in \mathbb{N}$ such that

$$|x_{j,n} - x_j| < \frac{\varepsilon}{\sqrt{k}}$$

for all $n \geq N$. For all $n \geq N$ we have

$$\begin{aligned} |\mathbf{x}_n - \mathbf{x}| &= ((x_{1,n} - x_1)^2 + \dots + (x_{k,n} - x_k)^2)^{1/2} \\ &< ((\varepsilon/\sqrt{k})^2 + \dots + (\varepsilon/\sqrt{k})^2)^{1/2} \\ &= \left(\frac{k\varepsilon^2}{k}\right)^{1/2} \\ &= \varepsilon. \end{aligned}$$

Therefore $\mathbf{x}_n \rightarrow \mathbf{x}$.

□

Proposition 3.4 establishes that a sequence of vectors converge if and only if each component converges. This is not terribly surprising.

3.3 Subsequences

Given some sequence $\{x_n\}$ in X , it's possible to restrict our attention to only certain terms of $\{x_n\}$. For example, if we have $x_n = (-1)^n$ in \mathbb{R} , we could look at every other term and in effect have a sequence of all 1's or all -1's. Clearly, both these sequences that are "contained" in $x_n = (-1)^n$ would converge. This case is particularly interesting because the sequence $x_n = (-1)^n$, despite us being able to "throw away" certain terms and have a convergent sequence as a result. As we'll see, this is not some fluke occurrence, and it's related to the fact that this sequence lives in a compact space! This will be the first of many nice/cool results that will follow from compactness.

Definition 3.4. Given a sequence $\{x_n\}$ in X , consider a sequence $\{n_k\}$ such that $n_1 < n_2 < \dots$. The sequence $\{x_{n_k}\}$ is called a *subsequence* of $\{x_n\}$. If $\{x_{n_k}\}$ converges, its limit is called a *subsequential limit* of $\{x_n\}$.

Remark 3.8. If $\{x_n\}$ corresponds to the function $f : \mathbb{N} \rightarrow X$, then we can think of a subsequence $\{x_{n_k}\}$ corresponding to some function $g : \mathbb{N} \rightarrow f(\mathbb{N})$, where $f(\mathbb{N}) \subseteq X$. In this sense, a subsequence of $\{x_n\}$ is a sequence in the range/image of $\{x_n\}$.

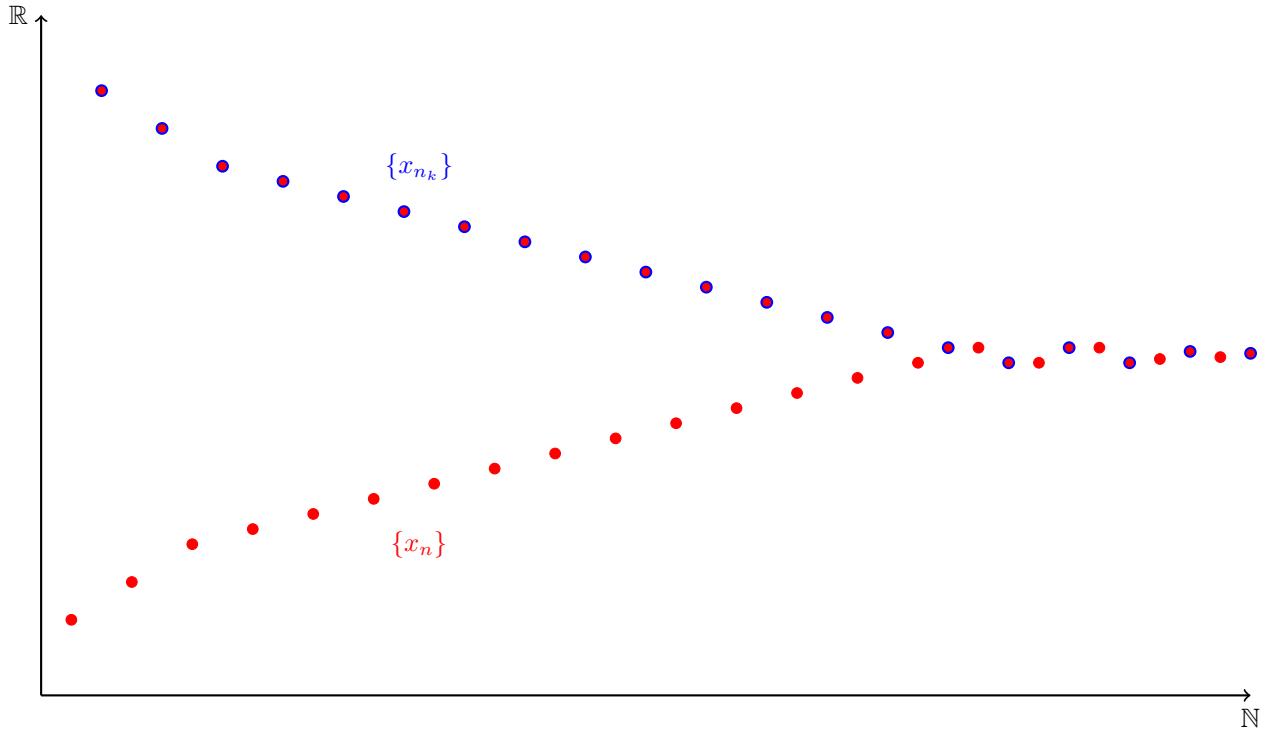


Figure 31: A sequence $\{x_n\}$ in \mathbb{R} , with a subsequence $\{x_{n_k}\}$.

A logical question to ask is how do we relate the limit of some convergent sequence with its subsequential limits? Is there at least one subsequential limit of a sequence that is the same as the limit of a sequence? Is the even stronger result that every subsequential limit of a sequence the same as the limit of the sequence possible?

Theorem 3.3. A sequence $\{x_n\}$ converges to x if and only if every subsequence of $\{x_n\}$ converges to x .

Proof.

(\Rightarrow) Suppose $x_n \rightarrow x$. For all ε , there exists an N such that $|x_n - x| < \varepsilon$ for all $n \geq N$. If x_{n_k} is some arbitrary subsequence of $\{x_n\}$, then $|x_{n_k} - x| < \varepsilon$ for all $n_k \geq N$. Therefore $x_{n_k} \rightarrow x$.

(\Leftarrow) Suppose every subsequence of $\{x_n\}$ converges to x . This means that the trivial subsequence of $\{x_{n_k}\} = \{x_n\}$ converges.³⁶

□

Corollary 3.2. A sequence x_n in a metric space X diverges if and only if it has a divergent subsequence, or more than one subsequential limit.

Theorem 3.3 and its corollary relate the limit of a convergent sequence to its subsequential limits, and says they are all the same. The proof of this is very brief, but the result is sweeping. Every single possible subsequence, a collection which is uncountably infinite, has the same limit.

Example 3.17. Let $\{x_n\} = \{1, 2, 3, 1, 2, 3, 1, 2, 3, \dots\}$ be a bounded sequence in \mathbb{R} . This sequence has subsequences that converge to 1, 2, and 3, so $\{x_n\}$ diverges. Alternatively, we could say that $\{x_n\}$ diverges because it has a divergent subsequence in $\{x_{n_k}\} = \{1, 2, 1, 2, 1, 2, \dots\}$.

A more interesting case to consider than that of a convergent sequence, is a divergent sequence. How do subsequences of divergent sequences behave. Do divergent sequences have subsequential limits? This question is answered with one of the most important theorems in analysis.³⁷

Theorem 3.4. If a metric space X is compact, then every sequence $\{x_n\}$ in X has a convergent subsequence.

Proof.

Suppose X is compact. Let E be the range of $\{x_n\}$. If E is finite,³⁸ then we can construct a constant subsequence $\{x_{n_k}\}$ such that

$$x_{n_1} = x_{n_2} = \dots = x,$$

where $x \in E$. This constant sequence converges.

If E is infinite, then E has some limit point $x \in X$ by the Bolzano-Weierstrass Property (Proposition 2.6). By Theorem 3.1, there exists a sequence $\{y_n\}$ in E such that $y_n \rightarrow x$. But E is just the range of $\{x_n\}$, so any sequence $\{y_n\}$ in E can be expressed as a subsequence $\{x_{n_k}\}$ (see Remark 3.8). Therefore x is a subsequential limit of x_n .

□

Corollary 3.3 (Bolzano-Weierstrass Theorem). Every bounded sequence in \mathbb{R}^k has a convergent subsequence.

Proof. A bounded subset of \mathbb{R}^k can always be contained in a closed and bounded subset of \mathbb{R}^k , the latter of which is compact by the Heine Borel Theorem. This means that the bounded sequence is a sequence in a compact space and we can apply Theorem 3.4. □

³⁶This trivial subsequence comes from taking $\{n_k\} = \mathbb{N}$.

³⁷How can you tell if a result is important? A good rule of thumb is that something is important if it has some specific name.

³⁸Remember that a sequence is infinite even if it ranges is finite. If its range is finite, that just means it takes on at least one of these values an infinite number of times.

Remark 3.9 (Equivalence Forms of Compactness in Metric Space). Recall Proposition 2.6 and Remark 2.10. These pertained to the fact that in a compact metric space, any subset has a limit point in the compact set. It was then discussed that there are more general settings in point-set topology where this may not be true for compact sets, and for that reason the property given by Proposition 2.6 is often called limit point compactness. A similar situation holds for Theorem 3.4. If every convergent sequence in some space has at least one subsequential limit, the space is sometimes called *sequentially compact*. In metric spaces, sequential compactness is equivalent to compactness, so knowing the difference isn't important.³⁹ Many times, the definition of sequential compactness is given as the definition of compactness, as it is easier to digest than the actual definition. In sum: a *metric space* is compact *if and only if* it is sequentially compact *if and only if* it is limit point compact. All of this will be made formal in Section 13.

Example 3.18. The sequence $x_n = 1/n$ in $(0, 1]$ diverges (see Example 3.8). It also has no convergent subsequence, so by Theorem 3.4, $(0, 1]$ is not compact. We already knew this though, as $(0, 1]$ is not closed.

Example 3.19. The sequence $x_n = (-1)^n$ in \mathbb{R} is bounded (see Example 3.14). Therefore it has at least one subsequential limit. Those limits are 1 and -1 .

Remark 3.10. Much like the proof of Theorem 3.1, we do not find an explicit formula for a convergent subsequence when proving The Bolzano-Weierstrass Theorem. As of now, we don't have a blueprint for finding some convergent subsequence for a sequence in \mathbb{R}^k , we only know it is out there. Later on, we'll introduce the concept of \limsup and \liminf , and show a specific type of subsequence that will *always* converge.

3.4 Cauchy Sequences

Up until now, showing a sequence converges has been purely a theoretical exercise. Proving x_n converges to x using the definition of convergence requires we know x , but how would we know the limit of a sequence if we didn't know it converged?⁴⁰

Example 3.20. Let $x_n = (1 + 1/n)^n$ be a sequence in \mathbb{R} . In Example 3.7, I claimed this converged to the constant e , but this is not clear at all from looking at the sequence.

In this subsection we'll develop a way of verifying of a sequence converges without knowing the limit. It's hard to emphasize just how useful this is in actual applications of analysis, as it's rarely clear what some relevant sequence would converge to. The method we develop will not work in every metric space, but will work in the most common spaces. We start by introducing a second type of convergence.

Definition 3.5. A sequence $\{x_n\}$ in a metric space X is a *Cauchy sequence (in X)* if for all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ if $n \geq N$ and $m \geq N$.

The terms in a Cauchy sequence get arbitrarily closer to each other over time. This is opposed to a convergent sequence, where the terms get arbitrarily close to some limit. This difference is the reason we never refer to a Cauchy sequence having a limit.⁴¹

³⁹This is why no introductory analysis course distinguishes them.

⁴⁰This is similar to something you may have seen in an intro stats course. You can use a *z*-test for the population mean, but that requires you know the population standard deviation. In what world would you know the population standard deviation but not the population mean?

⁴¹Unless of course it also happens to be convergent. No matter how small we let ε be, we can always find an $N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ for all $n, m \geq N$.

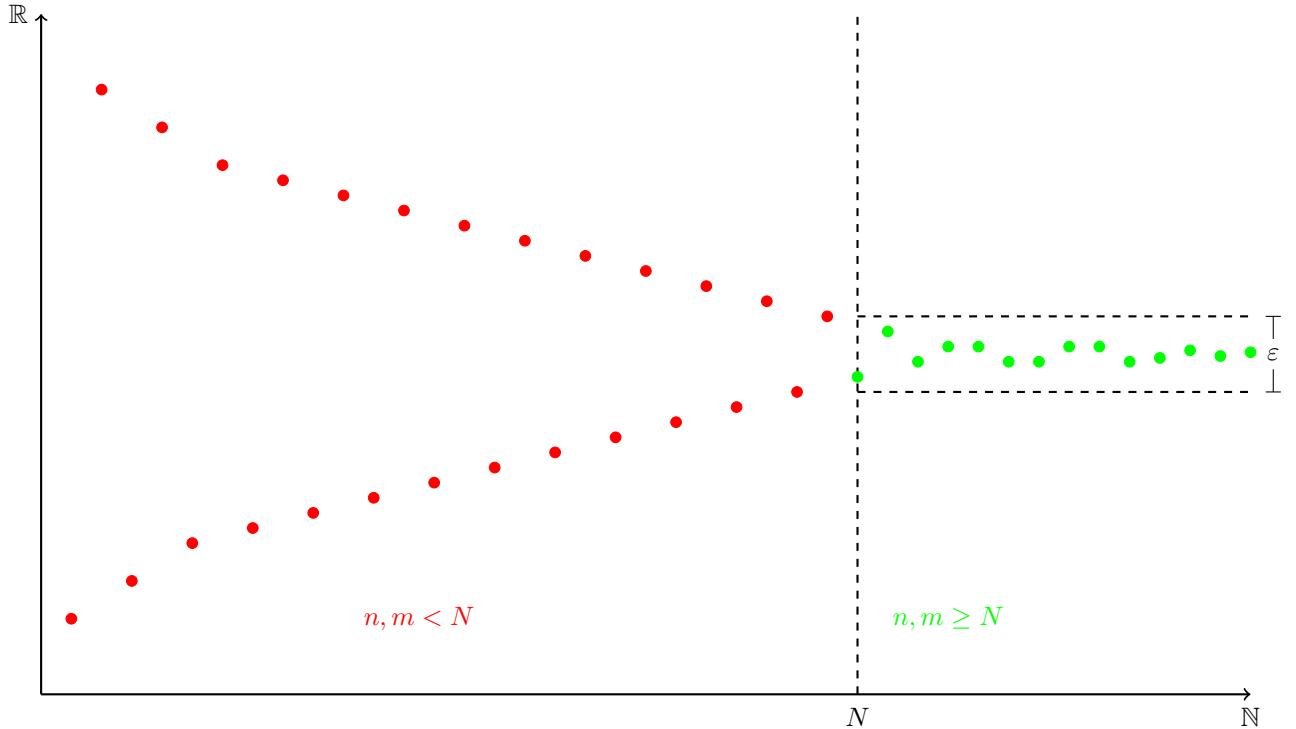


Figure 32: A Cauchy sequence in \mathbb{R} .

Example 3.21. Let $x_n = 1/2^n$ be a sequence in \mathbb{R} . We can show that this sequence is a Cauchy sequence. If we let $N = -\ln(\varepsilon/4)/\ln(2)$, then we have

$$|x_n - x_m| = \left| \frac{1}{2^n} - \frac{1}{2^m} \right| \leq \frac{1}{2^N} + \frac{1}{2^N} = \frac{1}{2^{-\ln(\varepsilon/4)/\ln(2)}} + \frac{1}{2^{-\ln(\varepsilon/4)/\ln(2)}} = 2 \cdot 2^{\log_2(\varepsilon/4)} = \frac{\varepsilon}{2} < \varepsilon$$

for all $n, m \geq N$. Therefore $x_n = 1/2^n$ is a Cauchy sequence.

While a Cauchy sequence does not have a limit, we can formulate a convergent sequence using a Cauchy sequence. If the terms in a Cauchy sequence are getting arbitrarily close, then one may suspect that if we look at the “tail” of such a sequence, the maximum distance between them is shrinking. This idea gives rise to the next definition and theorem, which will prove an alternate definition for a Cauchy sequence that uses the definition of a convergent sequence.

Definition 3.6. Let E be a nonempty subset of a metric space X . The *diameter* of E is the supremum of the set of $d(x, y)$ for all $x, y \in E$.

$$\text{diam } E = \sup\{d(x, y) \mid x, y \in E\}$$

Remark 3.11. Whenever you see a supremum, it’s worth asking yourself “does this necessarily exist? If so, how do I know this?” In this case, the answer is yes. In [Definition 3.6](#), the set E is a subset of $[0, \infty]$, because $d(x, y) \in [0, \infty]$ for all $d(x, y)$ by the definition of d . This in turn gives $E \subseteq \mathbb{R}$, and any subset of \mathbb{R} has a supremum.

Example 3.22. Let $E = [a, b] \subseteq \mathbb{R}$. We have $\text{diam } E = b - a$. If we instead have $E = (b - a)$, then we still have $\text{diam } E = b - a$, even though this diameter is never realized, because a supremum of a set needn’t be in the set.

Lemma 3.1. If x_n is a sequence in X , and $E_n = \{x_N, x_{N+1}, x_{N+2}, \dots\}$, then $\{x_n\}$ is a cauchy sequence if and only if $\lim_{N \rightarrow \infty} \text{diam } E_n = 0$.

Proof.

(\Rightarrow) Suppose $\{x_n\}$ is a Cauchy sequence. For all $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < \frac{\varepsilon}{2}$$

for all $n, m \geq N$. We use the supremum to define $\text{diam } E_n$, so we know there exists $x', y' \in E_n$ such that

$$0 \leq \text{diam } E_N = \sup_{x, y \in E_N} d(x, y) < d(x', y') + \frac{\varepsilon}{2}.$$

By the definition of E_n , x' and y' take the form x_n and p_m for some $n, m \geq N' \geq N$. Therefore

$$\text{diam } E_n < d(x_n, x_m) + \frac{\varepsilon}{2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This means that $\lim_{N \rightarrow \infty} \text{diam } E_n = 0$.

(\Leftarrow) Suppose $\lim_{N \rightarrow \infty} \text{diam } E_n = 0$. For all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$d(\text{diam } E_n, 0) = \text{diam } E_n < \varepsilon$$

for all $n \geq N$. If we have $n, m \geq N$, then $x_n, x_m \in E_N$. Therefore

$$d(x_n, x_m) \leq \sup_{x, y \in E_N} d(x, y) = \text{diam } E_N < \varepsilon$$

for all $n, m \geq N$. This gives that $\{x_n\}$ is a Cauchy Sequence

□

Example 3.23. Let $x_n = 1/n$ in \mathbb{R} . For $E_n = \{1/N, 1/(N+1), 1/(N+2), \dots\}$, we have

$$\text{diam } E_n = \sup\{d(x, y) \mid x, y \in E_n\} = \sup\{d(1/n, y) \mid y \in E_n\} = 1/N.$$

This means $\lim_{N \rightarrow \infty} \text{diam } E_n = 0$, so $\{x_n\}$ is a Cauchy sequence by Theorem 3.5.

Considering the diameter of a Cauchy sequence may seem redundant and overly complicated, but it allows us to prove many results relating to Cauchy sequences. Their introduction is more a means to an end than a definition that is important in its own right. That being said, we do need to introduce a couple of properties of the diameter in order to get mileage out of the concept when proving results.

Lemma 3.2 (Properties of Diameter).

1. If \bar{E} is the closure of a set E in a metric space X , then $\text{diam } E = \text{diam } \bar{E}$.
2. If K_n is a sequence of compact sets in X such that $K_{n+1} \subseteq K_n$ for all $n \in \mathbb{N}$ and if

$$\lim_{N \rightarrow \infty} \text{diam } K_n = 0,$$

then $\cap_{n=1}^{\infty} K_n$ consists of exactly one point.

Proof.

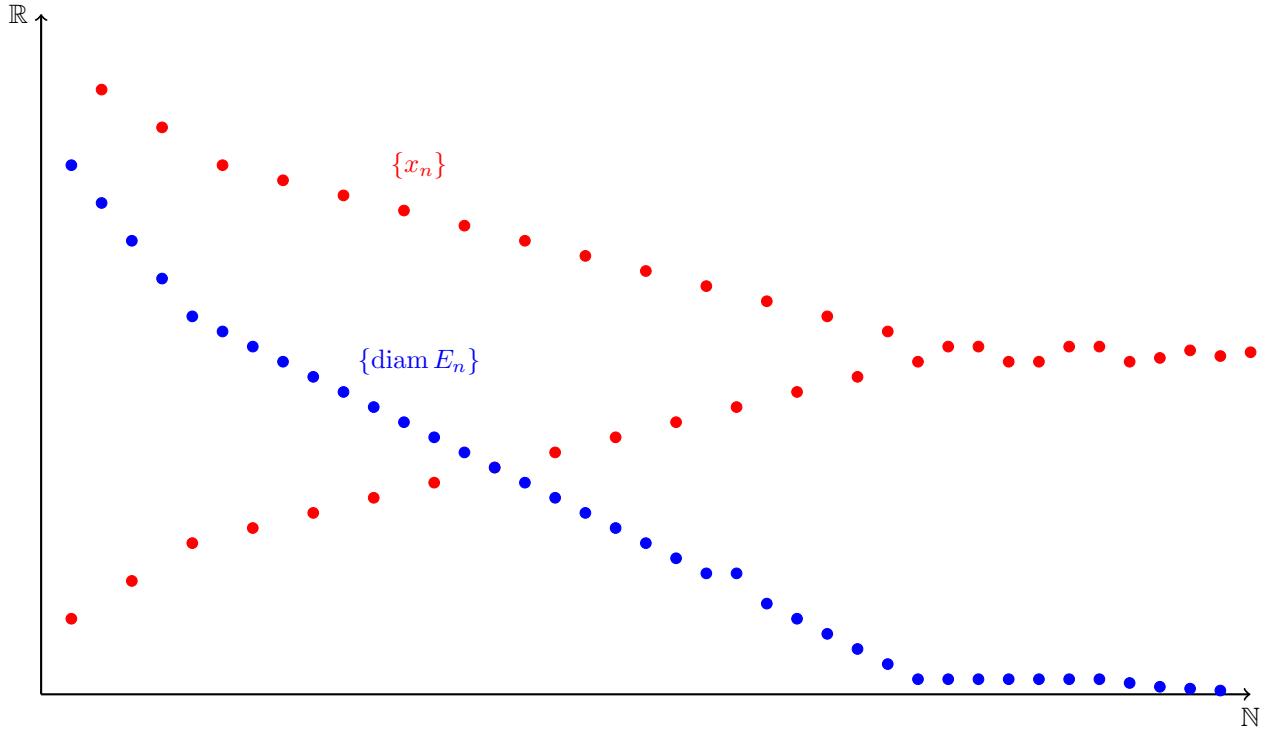


Figure 33: A Cauchy sequence in $\{x_n\}$ in \mathbb{R} , and the sequence $\{\text{diam } E_n\}$ as defined in Theorem 3.5

1. We have that $E \subseteq \bar{E}$, so $\text{diam } E \leq \text{diam } \bar{E}$.⁴² Now we will show that $\text{diam } E \geq \text{diam } \bar{E}$, which gives the equality. Fix a specific $\varepsilon > 0$, and let $x, y \in \bar{E}$. The point x and y are limit points of E , so there exists $x', y' \in E$ such that $d(x, x') < \varepsilon$ and $d(y, y') < \varepsilon$. The triangle inequality gives

$$d(x, y) \leq d(x, x') + d(x', y') + d(y', y) < 2\varepsilon + \text{diam } E.$$

Our selection of $x, y \in \bar{E}$ and ε were arbitrary, so $\text{diam } \bar{E} \leq \text{diam } E$.

2. By Lemma 2.3, $\cap_{i=1}^{\infty} K_n \neq \emptyset$. For the sake of contradiction, suppose $\cap_{i=1}^{\infty} K_n \neq \emptyset$ contains more than one point. If this is the case, then $\text{diam } \cap_{i=1}^{\infty} K_n > 0$. For all $n \in \mathbb{N}$, $\cap_{i=1}^{\infty} K_n \not\subseteq K_n$, so $\text{diam } K_n \geq \text{diam } \cap_{i=1}^{\infty} K_n > 0$. This contradicts the assumption that $\lim_{N \rightarrow \infty} \text{diam } K_n \rightarrow 0$.

□

With the introduction of Cauchy sequences comes the million dollar question that underlies this whole subsection. Do Cauchy sequences converge? The answer is in fact, no. Most of the time, Cauchy sequences do converge, in general this is not true. What is true however, is that all convergent sequences are Cauchy sequences. The task of finding a method to prove convergence without knowing a limit becomes a matter of finding a general class of functions for which all Cauchy sequences converge. Once we know this, we can simply verify such a sequence is Cauchy and know it converges.

Example 3.24. Recall the sequence $x_n = 1/n$ in the space $(0, 1]$. This sequence does not converge because $0 \notin (0, 1]$. Nevertheless, it is a Cauchy sequence. Cauchy sequences do not have an explicit limit, so excluding 0 from the space $\{x_n\}$ is in does not affect whether or not $\{x_n\}$ is Cauchy.

⁴²As much as I hate to borrow condescending adjectives from Rudin (1976), this is “clear”.

Example 3.24 shows that a Cauchy sequence need not be convergent, but we can show that any convergent sequence is a Cauchy sequence.

Theorem 3.5. Let $\{x_n\}$ be a sequence in a metric space X . If x_n converges to some limit $x \in X$, then $\{x_n\}$ is a Cauchy sequence.

Proof. For all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon/2$ for all $n \geq N$. The triangle inequality gives

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $m, n \geq N$. Thus $\{x_n\}$ is Cauchy. \square

Example 3.25. Example 3.4, 3.5, 3.6, 3.7, and 3.9 also double as examples of Cauchy sequences by Theorem 3.6.

We now turn to one of the most important results involving sequences. If we restrict our attention to compact spaces, then all Cauchy sequences converge. This allows us to prove a sequence in a compact space converges without having any knowledge of its limit!

Theorem 3.6. If X is a compact metric space and if $\{x_n\}$ is a Cauchy sequence in X , then $\{x_n\}$ converges to some point of X .

Proof. Let $\{x_n\}$ be a Cauchy sequence in a compact space X . For $n \in \mathbb{N}$, define $E_n = \{x_N, x_{N+1}, x_{N+2}, \dots\}$. By Lemma 3.1 and Lemma 3.2,

$$\lim_{N \rightarrow \infty} \text{diam } E_n = \lim_{N \rightarrow \infty} \text{diam } \bar{E}_n = 0.$$

The set \bar{E}_n is a closed subset (Lemma 3.2) of a compact space, so it is compact (Proposition 2.4). We also have $E_{n+1} \subseteq E_n$ for all $n \in \mathbb{N}$ so $\bar{E}_{n+1} \subseteq \bar{E}_n$.

If $\bar{E}_{n+1} \subseteq \bar{E}_n$ for all $n \in \mathbb{N}$, where \bar{E}_n is compact and $\lim_{N \rightarrow \infty} \text{diam } \bar{E}_n = 0$, there is a unique $x \in X$ such that $x \in \cap_{n=1}^{\infty} \bar{E}_n$ (Lemma 3.2).

Let $\varepsilon > 0$. By $\lim_{N \rightarrow \infty} \text{diam } \bar{E}_n = 0$, There exists some $N' \in \mathbb{N}$ such that $\text{diam } \bar{E}_N < \varepsilon$ for all $N \geq N'$. Our unique point x is in \bar{E}_N , so $d(x, y) \leq \varepsilon$ for all $y \in \bar{E}_N$, and hence for every $y \in E_n$. This is equivalent to saying $x_n \in B_\varepsilon(x)$ for all $n \geq N'$. This gives $x_n \rightarrow x$. \square

Theorem 3.7 (Cauchy Criterion). In \mathbb{R}^k , every Cauchy sequence converges.

To show this result, we will show that a Cauchy sequence in \mathbb{R}^k is bounded.

Proof. FINISH \square

The Cauchy Criterion is what we have been building to. To prove a sequence converges in \mathbb{R}^k , we just need to show it is a Cauchy sequence. This is often much easier and not as time consuming. It's important to remember that Theorem 3.6 is a more general result. It's so useful, we even have a name for metric spaces where Cauchy sequences always converge.

Definition 3.7. A metric space is *complete* if every Cauchy sequence converges.

Example 3.26. The real line \mathbb{R} is complete by the Cauchy Criterion. This example also shows that the converse of Theorem 3.6 does not hold. The real line is not compact, but is nevertheless complete.

Example 3.27. The rationals \mathbb{Q} are not complete. For example, the sequence $x_n = (1 + 1/n)^n$ does not converge in \mathbb{Q} , but it is Cauchy.

Example 3.28. The set of all real valued continuous functions defined on the domain $[a, b] \subseteq \mathbb{R}$, denoted $C([a, b])$, is complete. Right now, this example seems very abstract, as we haven't even discussed sequences of functions. We will return to this later on.

Remark 3.12. Early on, we referred to \mathbb{R} as "complete" in the sense that it had no gaps. While this is related to [Definition 3.7](#), they're not the same. Formally defining the first and distinguishing it from [Definition 3.7](#) would require a long detour. If it is ever unclear which one is being referred to, then I'll try to specify.

3.5 Monotonic Sequences

While the Cauchy Criterion simplifies showing convergence in \mathbb{R}^n , is it possible to arrive at even stronger results by restricting our attention to a smaller class of real sequences? The answer is yes.

Many of the sequences in \mathbb{R} we are first introduced to share a common element in the ordering of the elements. Many basic sequences either "grow" or "shrink" at each step. We will refer to this behavior as monotonicity.

Definition 3.8. A sequence $\{x_n\}$ in \mathbb{R} is *monotonically increasing* if $x_n \leq x_{n+1}$ for all $n \in \mathbb{N}$.

Example 3.29. The sequence $x_n = n$ in \mathbb{R} is not only monotonically increasing, but also divergent.

Example 3.30. The sequence $x_n = 1 - 1/n$ in \mathbb{R} is monotonically increasing and converges to 1.

Definition 3.9. A sequence $\{x_n\}$ in \mathbb{R} is *monotonically decreasing* if $x_n \geq x_{n+1}$ for all $n \in \mathbb{N}$.

Example 3.31. The sequence $x_n = 1/n$ in \mathbb{R} is monotonically decreasing and converges to 0.

Example 3.32. The sequence $x_n = -n$ in \mathbb{R} is monotonically decreasing and diverges.

If you take some time to think about the examples presented, it may become clear what the convergent sequences have in common. They are bounded! This is not coincidence. If a monotonic sequence is bounded, it "moves in one direction" towards its bound. It can neither "overcome" this bound, nor "change direction", so the only option is that it converges to it! As it turns out, the converse also happens to hold – Any convergent monotonic sequence is bounded.

Theorem 3.8 (Monotone Convergence Theorem). Suppose $\{x_n\}$ in \mathbb{R} is monotonic. Then $\{x_n\}$ converges if and only if it is bounded.

Proof.

(\Rightarrow) Suppose $\{x_n\}$ is monotonically increasing and bounded.⁴³ We have $x_n \leq x_{n+1}$, and $x_n \leq x$ for all $n \in \mathbb{N}$, where $x \in \mathbb{R}$ is the supremum of the range of $\{x_n\}$. For all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$x - \varepsilon < x_N \leq x,$$

otherwise $s - \varepsilon$ would be the supremum of the range of $\{x_n\}$. Since $\{x_n\}$ is monotonically increasing,

$$x - \varepsilon < x_n \leq x$$

⁴³If $\{x_n\}$ is monotonically decreasing, then the proof is analogous.

for all $n \geq N$. But this gives that

$$|x_n - x| < \varepsilon$$

for all $n \geq N$. Therefore $x_n \rightarrow x$.

(\Leftarrow) Suppose $\{x_n\}$ is bounded and monotonic. All bounded sequences converge (Proposition 3.3).

□

Not only have we shown our result, but in doing so, we showed that a bounded and monotonic sequence will converge to its supremum/infimum.

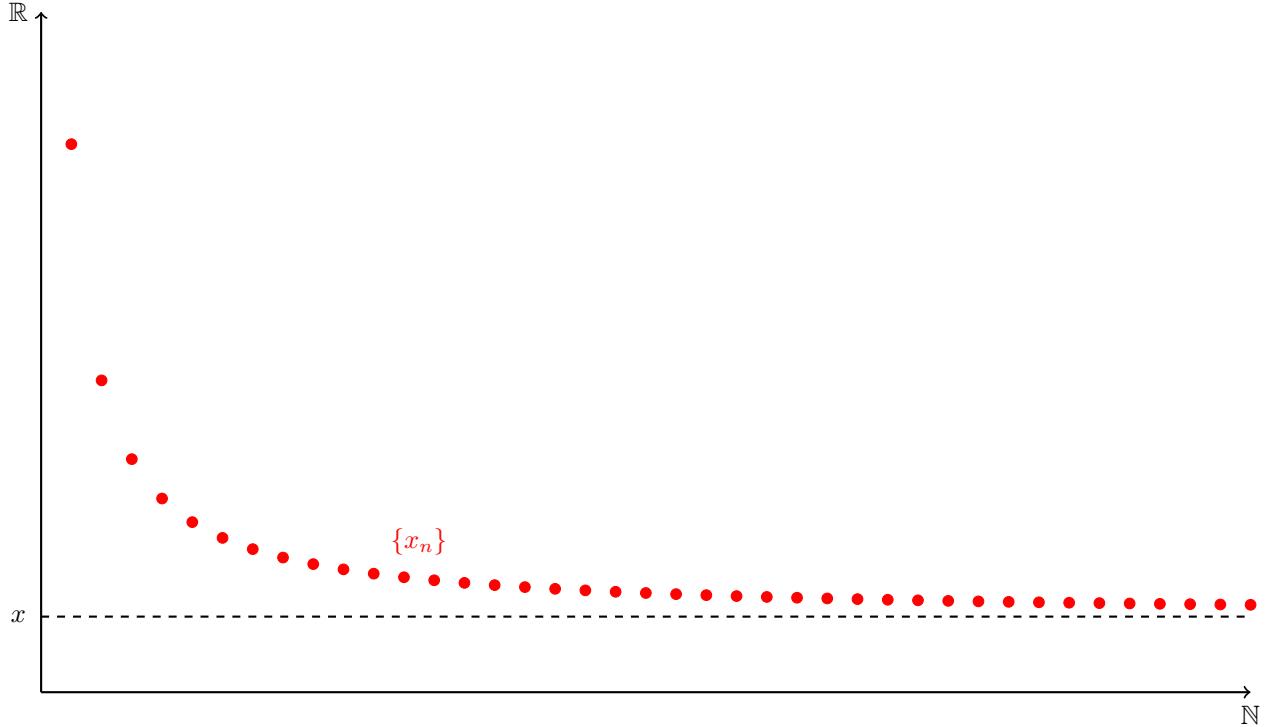


Figure 34: A monotonic and bounded sequence $\{x_n\}$ in \mathbb{R} . The sequence converges to the infimum of the sequences range, x .

Example 3.33. We can verify that $x_n = (1 + 1/n)^n$ converges by showing it is monotonically increasing and bounded. We can show it is bounded by using the inequality $\ln(1 + x) \leq x$.

$$\left(1 + \frac{1}{n}\right)^n = \exp\left(n \ln\left(1 + \frac{1}{n}\right)\right) \leq \exp\left(n \cdot \frac{1}{n}\right) = e.$$

To show the sequence is monotonic, we can use the inequality relating the arithmetic mean to the geometric mean (AM-GM inequality)⁴⁴ which gives,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq (x_1 x_2 \cdots x_n)^{1/n}.$$

⁴⁴This is a very useful inequality to know when working with proves involving boundedness and monotonicity.

If we let $x_1 = 1$, $x_2 = x_3 = \dots = x_{n+1} = 1 + 1/n$, then

$$\begin{aligned}
(x_1 x_2 \dots x_{n+1})^{\frac{1}{n+1}} &\leq \frac{1}{n+1}(x_1 + x_2 + \dots + x_{n+1}) \\
\left(1\left(1+\frac{1}{n}\right)\dots\left(1+\frac{1}{n}\right)\right]^{\frac{1}{n+1}} &\leq \frac{1}{n+1}\left(1+\left(1+\frac{1}{n}\right)+\dots+\left(1+\frac{1}{n}\right)\right) \\
\left(1+\frac{1}{n}\right)^{\frac{n}{n+1}} &\leq \frac{1+n\left(1+\frac{1}{n}\right)}{n+1} \\
&= 1 + \frac{1}{n+1} \\
\left(1+\frac{1}{n}\right)^n &\leq \left(1+\frac{1}{n+1}\right)^{n+1}.
\end{aligned}$$

Hence $x_n \leq x_{n+1}$. We've only shown that $\{x_n\}$ converges. In order to show that $x_n \rightarrow e$, we would need to show that e is a least upper bound.

Remark 3.13 (Inequalities). Proofs involving inequalities can be an acquired taste. They often rely on really clever algebraic manipulations and using a set of common inequalities such as the AM-GM inequality. For this reason, I think they're a complete pain in the ass. Instead of relying on mathematical intuition and insight, you just need to make some guess about how to manipulate an equation.

Remark 3.14. It's worth recapping what we've done with sequences up until this point, and the motivation behind our results. One of the main goals when working with sequences is proving/disproving convergence. Unfortunately, showing a sequence converges only using the definition of convergence can be hard, as it requires us to know the limit of a function. If we know our sequence is in a specific type of metric space, or has a certain property, then you can use Theorem 3.6, Theorem 3.7, and/or Theorem 3.8 to verify convergence!

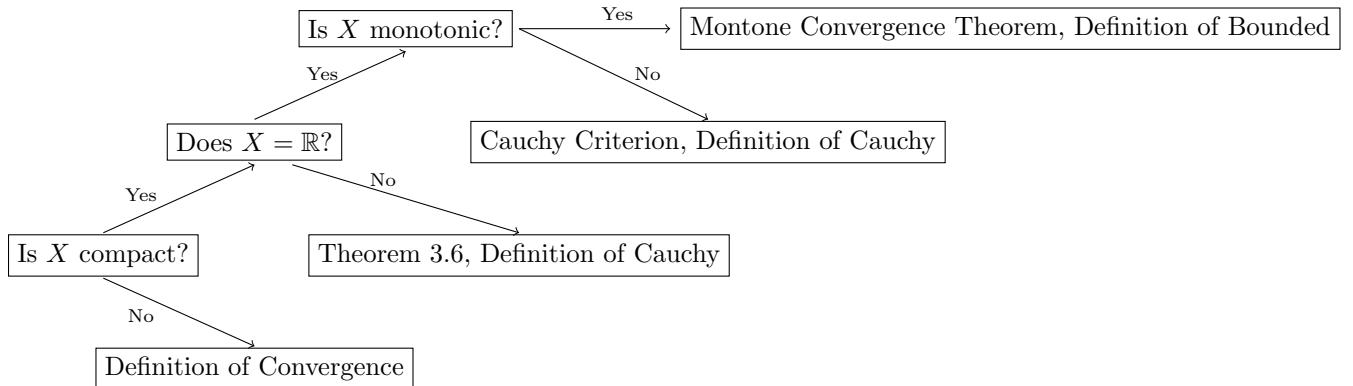


Figure 35: If $\{x_n\}$ is a convergent sequence in X , how do we prove that $\{x_n\}$ converges?

3.6 Sequences and Infinity

We have yet to discuss what happens when divergent sequences “go off” to infinity.

Definition 3.10. Let $\{x_n\}$ be a sequence in \mathbb{R} . If for all $M \in \mathbb{R}$ there exists an $N \in \mathbb{N}$ such that $x_n \geq M$, we write

$$\lim_{n \rightarrow \infty} x_n = \infty.$$

Similarly, if for every $M \in \mathbb{R}$ there exists an $N \in \mathbb{N}$ such that $n \geq N$ implies $x_n \leq M$, we write

$$\lim_{n \rightarrow \infty} x_n = -\infty.$$

These definition make a slight abuse of notation, as they refer to the limit of a sequence despite the sequence in question diverging.

Example 3.34. Let $x_n = n$ be a sequence in \mathbb{R} . For all $M \in \mathbb{R}$, let $N = M + 1$. We have

$$s_n \geq s_N = N = M + 1 > M$$

for all $n \geq N$. Therefore $x_n \rightarrow \infty$.

Remark 3.15 (Extended Real Numbers). If we were instead in the extended real numbers (Definition 1.19) $\overline{\mathbb{R}}$, then the type of sequences given in Definition 3.10 would converge, because $\{-\infty, \infty\} \subset \overline{\mathbb{R}}$. In fact, because every set in $\overline{\mathbb{R}}$ is bounded, many of the results pertaining to sequences would be more general in $\overline{\mathbb{R}}$. Any proposition or theorem that requires a sequence to be bounded in \mathbb{R} , would not make this requirement in $\overline{\mathbb{R}}$, as it would be implicitly met.

3.7 \limsup and \liminf

It can be insightful to study how the bounds of sequences change as n grows. For sequences in \mathbb{R} , doing this illuminates interesting relationships between converge, bounds, limits of bounds, and subsequential limits.

A sequence is bounded if its range is a bounded set. If this sequence is in \mathbb{R} , then we know something very powerful about the range of the sequence – its supremum and infimum exist in \mathbb{R} (Theorem 1.2). The same can be said for the supremum and infimum of the set of all elements in the range of x_n that have yet to be realized, $\{x_m \mid m \geq n\}$. The set $\{x_m \mid m \geq n\}$ is bounded above or below by x_n , so the supremum or the infimum (or both) of $\{x_m \mid m \geq n\}$ exist. We will define the supremum and infimum of a set using this set of values $\{x_m \mid m \geq n\}$.

Definition 3.11. Let x_n be a sequence in X . The *infimum of a sequence* is

$$\inf x_n = \inf_{m \geq n} x_m = \inf\{x_m \mid m \geq n\}.$$

Definition 3.12. Let x_n be a sequence in X . The *supremum of a sequence* is

$$\sup x_n = \sup_{m \geq n} x_m = \sup\{x_m \mid m \geq n\}.$$

The supremum and infimum of a sequence form their own sequences as the set $\{x_m \mid m \geq n\}$ depends on the value of n . Figure 36 illustrate the supremum and infimum of a sequence in \mathbb{R} .

Example 3.35. Let $x_n = 1/n$ be a sequence in \mathbb{R} . We have $\{x_m \mid m \geq n\} = \{1/m \mid n \geq m\}$. This gives $\sup x_n = 1/n$ and $\inf x_n = 0$, each being its own sequence in \mathbb{R} .

Example 3.36. If x_n be a monotonically increasing sequence, then $\inf x_n = x_n$ as $x_n \leq x_m$ for all $m \geq n$. Similarly, if x_n is monotonically decreasing, then $\sup x_n = x_n$.

Example 3.37. If $x_n = \sin x$, then $\inf x_n = -1$ and $\sup x_n = 1$ for all $n \in \mathbb{N}$.

Now we will define the limits of $\sup x_n$ and $\inf x_n$.

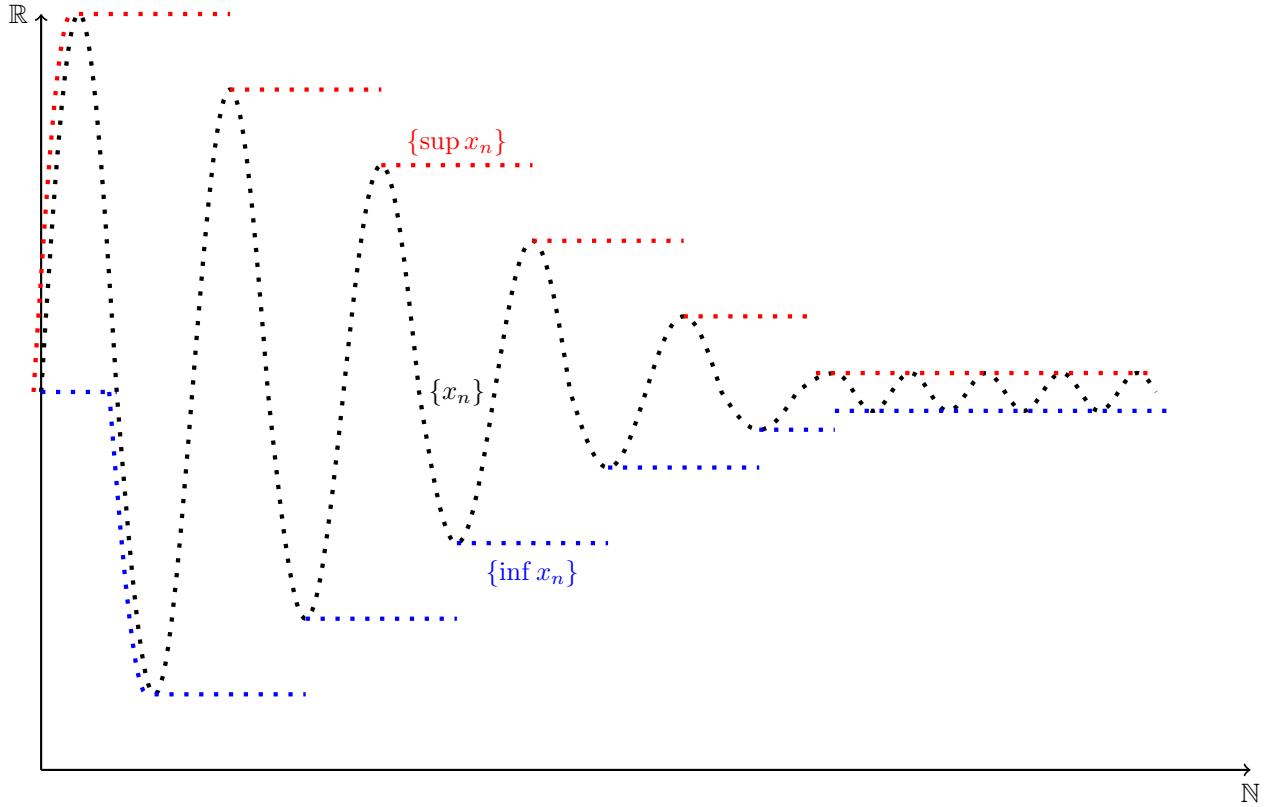


Figure 36: A sequence $\{x_n\}$ in \mathbb{R} , with the sequence $\{\sup x_n\}$ and $\{\inf x_n\}$ shown in red and blue, respectively. For any $n \in \mathbb{N}$, we take the supremum and infimum of all the points of $\{x_m \mid m \geq n\}$. Graphically this means we take the supremum and infimum of all the points to the right of a certain value of \mathbb{N} .

Definition 3.13. Let $\{x_n\}$ be a sequence in X . The *limit inferior* of $\{x_n\}$ is the limit of $\{\inf x_n\}$, and is written as $\liminf_{n \rightarrow \infty} x_n$ or $\underline{\lim}_{n \rightarrow \infty} x_n$.

Definition 3.14. Let $\{x_n\}$ be a sequence in X . The *limit superior* of $\{x_n\}$ is the limit of $\{\sup x_n\}$, and is written as $\limsup_{n \rightarrow \infty} x_n$ or $\overline{\lim}_{n \rightarrow \infty} x_n$.

Example 3.38. The sequence $x_n = 1/n$ has a limit inferior of 0, and a limit superior of 0.

Example 3.39. The sequence $x_n = \sin x$ has a limit inferior of -1, and a limit superior of 1.

You may have noticed two things: the sequences $\{\sup x_n\}$ and $\{\inf x_n\}$ seem to always converge, and sometimes the limit inferior and limit superior are equal. The observations are consequences of our next results.

Proposition 3.5. If $\{x_n\}$ is a bounded sequence in \mathbb{R} , then $\liminf_{n \rightarrow \infty} x_n$ and $\limsup_{n \rightarrow \infty} x_n$ exist. That is, the sequences $\{\inf x_n\}$ and $\{\sup x_n\}$ converge.

Proof. We will prove the result for the limit superior.⁴⁵ Let $\{x_n\}$ be a bounded sequence in \mathbb{R} . By the least-upper-bound property, $\sup x_n$ exists for all $n \in \mathbb{N}$, as the range of x_n is bounded. The sequence $\sup x_n$

⁴⁵The proof for the limit inferior is very similar.

is bounded because x_n is bounded. By the Monotone Convergence Theorem, it suffices to show that $\sup x_n$ is monotonically decreasing. We have $\{x_n \mid m \geq n+1\} \subset \{x_n \mid m \geq n\}$, so any upper bound of $\{x_n \mid m \geq n\}$ is an upper bound of $\{x_n \mid m \geq n+1\}$. This includes the least-upper-bound, which in this case is $\sup x_n$. By the definition of the least-upper-bound, $\sup x_{n+1}$ is less than all other upper-bounds of $\{x_n \mid m \geq n+1\}$, therefore $\sup x_n \geq \sup x_{n+1}$. This gives that $\{\sup x_n\}$ is monotonically decreasing. \square

Remark 3.16 (Monotonicity and Limits of Bounds). The proof of the last result took advantage of the fact that $\sup x_n$ was monotonically decreasing and bounded, allowing the use of the Monotone Convergence Theorem (Theorem 3.8). When we proved this result, we showed that a monotonically decreasing sequence will converge to its infimum. This means that $\{\sup x_n\}$ converges to its infimum. For this reason the limit superior is sometimes written as $\inf \sup x_n$. Similarly, the limit inferior can be written as $\sup \inf x_n$.

Proposition 3.6. Let $\{x_n\}$ be a bounded subsequence in \mathbb{R} . Then

1. There exist subsequences which converges to $\limsup x_n$ and $\liminf x_n$.
2. For all $c \in \mathbb{R}$, if there exists a subsequence which converges to c , then $\liminf x_n \leq c \leq \limsup x_n$.

If we let E be the set of subsequential limits corresponding to a bounded $\{x_n\}$ in \mathbb{R} , then Proposition 3.6 concludes that $\{\liminf x_n, \limsup x_n\} \subset E$, $\liminf x_n = \inf E$, and $\limsup x_n = \sup E$.

Example 3.40. Let $x_n = 1/n$ be a sequence in \mathbb{R} . It is convergent and bounded. We know from Example 3.37 that $\liminf x_n = \limsup x_n = 0$. By Proposition 3.6, 0 is the only subsequential limit of $\{x_n\}$. We already knew this though, because that is precisely what Theorem 3.3 asserts.

Example 3.41. For $x_n = \sin x$ in \mathbb{R} , the set of subsequential limits is a subset of $[-1, 1]$.

Proposition 3.7. Let $\{x_n\}$ be a bounded sequence in \mathbb{R} . The sequence converges if and only if

$$\liminf_{n \rightarrow \infty} x_n = \limsup_{n \rightarrow \infty} x_n.$$

Proof.

(\Rightarrow) Suppose $\{x_n\}$ converges to $x \in \mathbb{R}$. By Theorem 3.3, every subsequence of $\{x_n\}$ converges to x . By Proposition 3.6, $\liminf x_n$ and $\limsup x_n$ are subsequential limits of $\{x_n\}$. This means they both are x , as x is the only subsequential limit of $\{x_n\}$. Therefore $\liminf x_n = \limsup x_n$.

(\Leftarrow) Assume that $\limsup x_n = \liminf x_n = x$. For all $\varepsilon > 0$,

$$x - \varepsilon < \liminf x_n \leq \limsup x_n < x + \varepsilon.$$

This allows us to choose an $B_1, B_2 \in \mathbb{N}$ such that

$$s - \varepsilon < x_n < s + \varepsilon$$

for all $n \geq \max\{B_1, B_2\}$. This gives that $|x_n - x| < \varepsilon$ for all such n , so $x_n \rightarrow x$. \square

3.8 Series

Now we consider series. As we'll see, everything we did with sequences carries over nicely to series. In fact, working with series will be even simpler.

A series is simply a special type of sequence which results from summing each element of a different sequence.

Definition 3.15. Given a sequence $\{x_n\}$ in a metric space X , we have a *series* in the form of

$$\sum_{n=1}^{\infty} x_n.$$

Example 3.42 (Geometric Series). If we let $x_n = 1/2^n$, then

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$$

In order to define the convergence of a series, we will look at the behavior of the first n terms. That is, we are interested in

$$\begin{aligned} s_1 &= \sum_{k=1}^1 x_k = x_1 \\ s_2 &= \sum_{k=1}^2 x_k = x_1 + x_2 \\ s_3 &= \sum_{k=1}^3 x_k = x_1 + x_2 + x_3 \\ s_4 &= \sum_{k=1}^4 x_k = x_1 + x_2 + x_3 + x_4 \\ &\vdots \end{aligned}$$

If the sequence $\{s_n\}$ converges, then we will say the series $\sum_{n=1}^{\infty} x_n$ converges.

Definition 3.16. Let $\sum_{n=1}^{\infty} x_n$ be a series in X . Define the *partial sums* of the series to be the sequence

$$s_n = \sum_{k=1}^n x_k.$$

A series *converges* to $x \in X$ if the sequence $\{s_n\}$ converges to $x \in X$. In this case, we write

$$\sum_{n=1}^{\infty} x_n = x.$$

Remark 3.17 (Everything from Sequences Carries Over). Every single result we established for sequences can be formulated in terms of series, because every single series can be expressed as a sequence of partial sums.

Showing a series converges using this definition is unnecessarily difficult. In \mathbb{R} we can use a reformulation of the Cauchy Criterion

Theorem 3.9 (Cauchy Criterion for Series). Let $\sum_{n=1}^{\infty} x_n$ be a sequence in \mathbb{R} . The series converges if and only if for all $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that

$$\left| \sum_{k=n}^m x_k \right| \leq \varepsilon$$

for all $m \geq n \geq N$.

Proof. This follows from the fact that

$$|s_n - s_m| = \left| \sum_{k=1}^n x_n - \sum_{k=1}^m x_m \right| = \left| \sum_{k=n}^m x_k \right|.$$

□

Corollary 3.4. If $\sum_{n=1}^{\infty} x_n$ converges, then $x_n \rightarrow 0$.

Proof. Let $m = n$ and use Theorem 3.9

□

Remark 3.18 (No Limit Needed!). Theorem 3.9 is a direct consequence of the Cauchy Criterion, so we don't need to know the limit of a convergent sequence to prove it converges.

Example 3.43. The converse of Corollary 3.4 is not true! The harmonic series $\sum_{n=1}^{\infty} 1/n$ diverges despite the fact that $1/n \rightarrow 0$.

Example 3.44. The series $\sum_n 1/x!$ converges to e in \mathbb{R} . This sequence does not converge in \mathbb{Q} , as $e \notin \mathbb{Q}$.

3.9 Tests for Convergent Series

Whenever we work with a series $\sum_{n=1}^{\infty} x_n$, we want to find some way to relate a sequence involving x_n to the convergence of the series. There are three tests we can use to verify that a sequence in \mathbb{R} converges by doing just this: the comparison test, the ratio test, and the root test. Another benefit of these tests, is none require us to know the limit of a series to prove it converges!

Theorem 3.10 (The Comparison Test). Let $\sum_{n=1}^{\infty} x_n$ be a series in \mathbb{R} . If $|x_n| \leq y_n$ for $n \geq B_0$, where B_0 is some fixed integer, and if $\sum_{n=1}^{\infty} c_n$ converges, then $\sum_{n=1}^{\infty} x_n$ converges.

Proof. For all $\varepsilon > 0$, there exists an $N \geq B_0$ such that $m \geq n \geq N$ implies

$$\sum_{k=n}^m y_k \leq \varepsilon,$$

according to The Cauchy Criterion. Therefore we have

$$\left| \sum_{k=n}^m x_k \right| \leq \sum_{k=n}^m |x_k| \leq \sum_{k=n}^m y_k \leq \varepsilon,$$

so again by the Cauchy criterion, $\sum_{k=n}^m x_k$ converges.

□

Example 3.45. The geometric series $\sum_{n=1}^{\infty} 1/3^n$ converges to $3/2$. We have

$$\frac{1}{3^n + n} < \frac{1}{3^n}$$

for all $n \in \mathbb{N}$, so the series $\sum_{n=1}^{\infty} 1/(3^n + n)$ converges by the Comparison Test.

Example 3.46. The harmonic series $\sum_{n=1}^{\infty} 1/n$ diverges. We have

$$\frac{n}{n^2 - \cos^2 n} > \frac{n^2}{n^2} = \frac{1}{n}.$$

Therefore by the converse of Theorem 3.10, the series $\sum_{n=1}^{\infty} \frac{n}{n^2 - \cos^2 n}$ diverges.

Theorem 3.11 (The Ratio Test). If $\sum_{n=1}^{\infty} x_n$ is a series in \mathbb{R} , then

1. the series converges if $\lim_{n \rightarrow \infty} |x_{n+1}/x_n| < 1$;
2. the series diverse if $|x_{n+1}/x_n| \geq 1$ for all $n \geq B_0$ for a fixed $B_0 \in \mathbb{N}$.

Proof. Suppose $\lim |x_{n+1}/x_n| = \beta < 1$. For ε This means we can find an integer $N \in \mathbb{N}$ such that $|x_{n+1}/x_n| < \beta$ for all $n \geq N$. This can be rewritten as $|x_{n+1}| < \beta|x_n|$. If we do this for $n = N, N + 1, N + 2, \dots, N + k$, then

$$\begin{aligned} |x_{N+1}| &< \beta|x_N|, \\ |x_{N+2}| &< \beta|x_{N+1}| < \beta^2|x_N|, \\ &\vdots \\ |x_{N+k}| &< \beta^k|x_N|. \end{aligned}$$

This gives

$$\sum_{k=N+1}^{\infty} |x_k| = \sum_{k=1}^{\infty} |x_{N+k}| < \sum_{k=1}^{\infty} \beta^k |x_N| = |x_N| \sum_{k=1}^{\infty} \beta^k = |x_N| \frac{\beta}{1-\beta}.$$

Therefore by the Comparison Test, $\sum_{n=1}^{\infty} x_n$ converges, because

$$\sum_{n=1}^{\infty} |x_n| \leq \sum_{k=N+1}^{\infty} |x_k|.$$

Now suppose that $\limsup |x_{n+1}/x_n| = \beta > 1$. This means that $|x_{n+1}| > |x_n|$ for a sufficiently large n , meaning $\lim_{n \rightarrow \infty} x_n \neq 0$. By the converse of Corollary 3.4, the series diverges. \square

Example 3.47. Does $\sum_{n=0}^{\infty} n!/5^n$ converge or diverge?

$$\lim_{n \rightarrow \infty} \left| \frac{x_{n+1}}{x_n} \right| = \lim_{n \rightarrow \infty} \left| \frac{(n+1)!}{5^{n+1}} \frac{5^n}{n!} \right| = \lim_{n \rightarrow \infty} \frac{(n+1)!}{5n!} = \lim_{n \rightarrow \infty} \frac{(n+1)n!}{5n!} = \lim_{n \rightarrow \infty} \frac{n+1}{5} = \infty$$

By the Ratio Test, the series diverges.

Example 3.48. If we have $\lim_{n \rightarrow \infty} |x_{n+1}/x_n| = 1$, then we aren't able to conclude anything about the convergence of a series. For example $\sum_{n=1}^{\infty} 1/n$ diverges, whereas $\sum_{n=1}^{\infty} 1/n^2$ converges. In both cases $\lim_{n \rightarrow \infty} |x_{n+1}/x_n| = 1$.

Theorem 3.12 (The Root Test). If $\sum_{n=1}^{\infty} x_n$ is a series in \mathbb{R} , then

1. the series converges if $\lim_{n \rightarrow \infty} \sqrt[n]{|x_n|} < 1$;
2. the series diverges if $\lim_{n \rightarrow \infty} \sqrt[n]{|x_n|} > 1$;
3. we cannot determine anything if $\lim_{n \rightarrow \infty} \sqrt[n]{|x_n|} = 1$.

Proof. Let $\alpha = \lim_{n \rightarrow \infty} \sqrt[n]{|x_n|}$. If $\alpha < 1$, we can find a $\beta \in (\alpha, 1)$, and an $N \in \mathbb{N}$ such that $\sqrt[n]{|x_n|} < \beta$ for $n \geq N$. This inequality can be expressed as $|x_n| < \beta^n$. Because $\sum_{n=0}^{\infty} \beta^n$ converges, the Comparison Test gives that $\sum_{n=N}^{\infty} |x_n|$ converges. Since

$$\sum_{n=1}^{\infty} |x_n| = \sum_{n=N}^{N-1} |x_n| + \sum_{n=N}^{\infty} |x_n|,$$

the series $\sum_{n=1}^{\infty} |x_n|$ must converge, as the first sum is finite.

If $\alpha > 1$, then $|x_n| > 1^n = 1$, so $\lim_{n \rightarrow \infty} x_n \neq 0$. By Corollary 3.4, the series diverges.

The series $\sum_{n=1}^{\infty} 1/n$ and $\sum_{n=1}^{\infty} 1/n^2$ both have $\alpha = 1$, despite the prior diverging and the latter converging. Therefore if $\alpha = 1$, we cannot make any statement about convergence. \square

Example 3.49. Does the series $\sum_{n=0}^{\infty} \left(\frac{5n-3n^3}{7n^3+3}\right)^n$ converge or diverge?

$$\lim_{n \rightarrow \infty} \sqrt[n]{|x_n|} = \lim_{n \rightarrow \infty} \left| \left(\frac{5n-3n^3}{7n^3+2} \right)^n \right|^{1/n} = \lim_{n \rightarrow \infty} \left| \frac{5n-3n^3}{7n^3+2} \right| = \frac{3}{7} < 1$$

By the Root Test, the series converges.

Remark 3.19 (Slightly More General Tests). We can make the Ratio Test and Root Test a bit more powerful by using \limsup instead of just the limit. The proofs would be modified slightly. This will rarely make a difference.

Remark 3.20 (Ratio vs. Root). The Ratio Test is much easier to use than the Root Test. Calculating ratios tends to be way easier than the n th root of a value. That being said, the Root Test is more powerful in the sense that the Ratio Test will always agree with the Root Test. Furthermore, there are cases when the Ratio Test is inconclusive, but the Root Test is not. A good example of this is found in Example 3.35 in Rudin (1976).

3.10 Exercises

rudin, 3.4 part b

constant sequences in \mathbb{Z}

subsequential limits closed

series of nonnegative terms converges iff partial sums bounded

4 Continuity

We spent a fair amount of time studying what it means for a sequence or series to get “arbitrarily close” to some point. We now want to do something similar with more general functions, and rigorously define continuity. Continuity makes nearly all of analysis possible, and without it we would be hopeless. Knowing the properties related to continuity will prove useful when proving many results, as many results will pertain exclusively to continuous functions.

4.1 Limits of Functions

We begin by defining what a limit of a function is.

Definition 4.1. Let X and Y be metric spaces, and $E \subseteq X$. Suppose there is some function $f : E \rightarrow Y$, and a limit point x_0 of E and some $L \in Y$ satisfying: for all $\varepsilon > 0$ there exists a corresponding $\delta > 0$ such that $d_Y(f(x), L) < \varepsilon$ for all points $x \in E$ which satisfy $d_X(x, x_0) < \delta$. We say that L is *the limit of f as x approaches x_0* , and write

$$\lim_{x \rightarrow x_0} f(x) = L.$$

This definition seems much more complicated than that given for convergence, but the concepts conveyed are fairly similar. It may be helpful to return to the somewhat lame example of playing a game with your friend involving ε . Instead of a sequence you now play with some function $f : E \rightarrow Y$, and a fixed limit point $x_0 \in E$.⁴⁶ Your friend gives you a value of ε , and he challenges you to find a δ such that $d_Y(f(x), L) < \varepsilon$ whenever $d_X(x_0, x) < \delta$. Note that we never require that $f(x) = L$ (see Figure 37). Our next several examples will make this more concrete.

Example 4.1. Suppose we define $f : \mathbb{R} \rightarrow \mathbb{R}$ as some linear function $f(x) = mx + b$. For $L = mx_0 + b$ and $x_0 \in \mathbb{R}$, your friend gives you $\varepsilon = 0.1$. You need to find a δ such that

$$d_{\mathbb{R}}(f(x), L) = |mx + b - (mx_0 + b)| = |m(x - x_0)| = |m||x_0 - x| < 0.1.$$

Therefore you let $\delta = 0.1/|m|$. Instead of letting him keep naming ε , you tell him that you’ll just set $\delta = \varepsilon/|m|$. This way whenever $d_{\mathbb{R}}(f(x), L) < \varepsilon$ we have

$$d_{\mathbb{R}}(f(x), L) = |mx + b - (mx_0 + b)| = |m(x - x_0)| = |m||x_0 - x| = |m| \cdot d_{\mathbb{R}}(x, x_0) < |m| \cdot \frac{\varepsilon}{|m|} = \varepsilon$$

whenever $d_{\mathbb{R}}(x, x_0) < \delta$. Therefore $\lim_{x \rightarrow x_0} f(x) = L$.

Example 4.2. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

We can show that the limit at $x_0 = 0$ fails to exist by finding one such ε where Definition 4.1 does not hold. Let $\varepsilon = 1/2$, and suppose for contradiction $\lim_{x \rightarrow x_0} f(x) = L$. We have that $|f(x) - L| < \varepsilon = 1/2$ for all x

⁴⁶It’s important that x_0 is a limit point. This means every open ball around p has infinitely many points, so we can always get arbitrarily close to it.

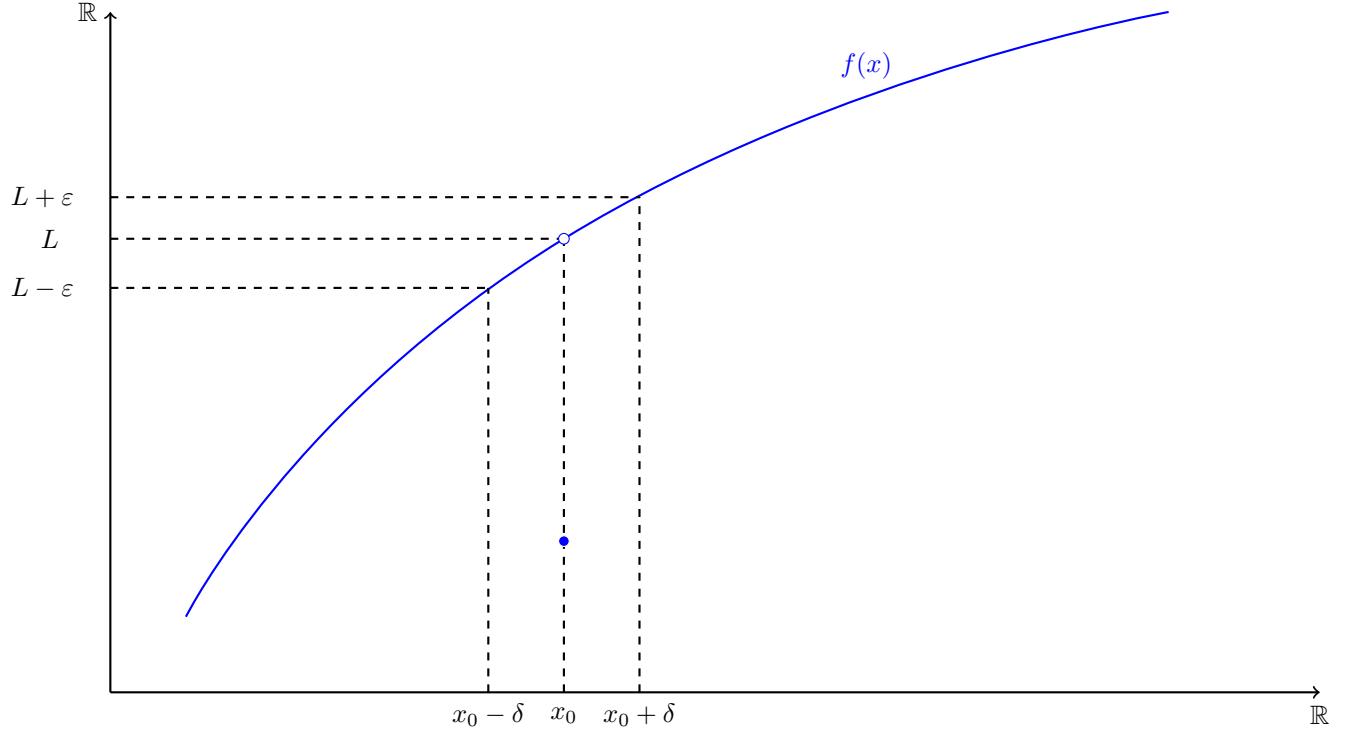


Figure 37: We have $\lim_{x \rightarrow x_0} f(x) = L$ for a real function. No matter how small ε gets, we can find a δ such that $|f(x) - L| < \varepsilon$ for all x satisfying $|x - x_0| < \delta$.

such that $|x - 0| < \delta$. But we also have that

$$\begin{aligned}
2 &= |1 - (-1)| \\
&= |f(\delta/2) - f(-\delta/2)| \\
&= |f(\delta/2) - L + L - f(-\delta/2)| \\
&\leq |f(\delta/2) - L| + |L - f(-\delta/2)| \\
&\leq \frac{1}{2} + \frac{1}{2} \\
&= 1,
\end{aligned}$$

which is a clear contradiction.

Example 4.3. Suppose we define $f : \mathbb{R} \rightarrow \mathbb{R}$ as $f(x) = \sin x$. We can verify that $\lim_{x \rightarrow 0} f(x) = 1$ despite the fact that $f(x)$ is undefined at 0. If we let $\delta = \sqrt{\varepsilon}$ and restrict our attention to the interval $(-\pi/2, \pi/2)$,⁴⁷ then for all $\varepsilon > 0$ we have

$$d_{\mathbb{R}}(f(x), 1) = \left| \frac{\sin x}{x} - 1 \right| < 1 - \cos x < 2 \sin^2 \frac{x}{2} < \frac{x^2}{2} = \frac{|x - 0|^2}{2} = \frac{d_{\mathbb{R}}(x, 0)^2}{2} < \frac{(\sqrt{\varepsilon})^2}{2} < \frac{\varepsilon}{2} < \varepsilon$$

whenever $d_{\mathbb{R}}(x, 0) < \delta$.

As you may suspect, there is a very strong link between the convergence of a sequence and the limit of a function. If we have $\lim_{x \rightarrow x_0} f(x) = L$, then x_0 is a limit point of $E \subset X$ as defined in [Definition 4.1](#).

⁴⁷This insures that $0 < \cos x < \sin x/x < 1$.

This limiting process lets x become arbitrarily close to x_0 , much like a sequence would to its limit. As it turns out, there is a way to reformulate limits in terms of sequences if we use a sequence $\{x_n\}$ in E which converges to x_0 .⁴⁸

Theorem 4.1. Let X and Y be metric spaces, and $E \subseteq X$. Suppose there is some function $f : E \rightarrow Y$, and a limit point x_0 of E . Then $\lim_{x \rightarrow x_0} f(x) = L$ if and only if $f(x_n) \rightarrow L$ (alternatively written as $\lim_{n \rightarrow \infty} f(x_n) = L$) for every non-constant sequence $\{x_n\}$ in E such that $x_n \rightarrow x_0$.⁴⁹

Proof.

(\Rightarrow) Suppose $\lim_{x \rightarrow x_0} f(x) = L$. By Theorem 3.1, we can choose some arbitrary sequence $\{x_n\}$ in E such that $x_n \rightarrow x_0$. For all $\varepsilon > 0$ there exists a $\delta > 0$ such that $d_Y(f(x), L) < \varepsilon$ for all $x \in E$ satisfying $d_X(x, x_0) < \delta$. By the convergence of x_n , there exists an $N \in \mathbb{N}$ such that $d_X(x_n, x_0) < \delta$ for all $n > N$.⁵⁰ Therefore, for all $n > N$, we have $d_Y(f(x_n), L) < \varepsilon$. This gives

$$\lim_{n \rightarrow \infty} f(x_n) = L.$$

(\Leftarrow) We will prove the contrapositive. Suppose that $\lim_{x \rightarrow x_0} f(x) \neq L$. There exists some $\varepsilon > 0$ such that for all $\delta > 0$, there exists some point $x \in E$ (which depends on δ) for which $d_Y(f(x), L) \geq \varepsilon$ but $d_X(x, x_0) < \delta$.⁵¹ This holds for all $\delta > 0$, so define $\delta_n = 1/n$. There will exist some point $x_n \in E$ for which $d_Y(f(x_n), L) \geq \varepsilon$ but

$$d_X(x_n, x_0) < \delta_n = 1/n.$$

If we repeat this for all n , we have a sequence of $\{x_n\} \subseteq E$ which satisfies $x_n \rightarrow x_0$, even though $d_Y(f(x_n), L) \geq \varepsilon$. In other words, despite $x_n \rightarrow x_0$, $\lim_{n \rightarrow \infty} f(x_n) \neq L$.

□

Example 4.4. For the real function $f(x) = x$, it's clear that $\lim_{x \rightarrow 0} f(x) = 0$. By Theorem 4.1 we know that

$$\lim_{n \rightarrow \infty} f(1/n) = 0,$$

as $1/n \rightarrow 0$. Furthermore we know this is the case for any $\{x_n\}$ which converges to 0.

So what's the big deal with writing limits in terms of sequences? This just seems like another complex relationship to remember. Theorem 4.1 is important because it allows us to prove many results involving limits (and later continuity) using properties we already know from sequences. For example, when using Theorem 4.1 we know right away that limits of functions are unique because the limits of sequences are unique.

Corollary 4.1. If f has a limit at x_0 , the limit is unique.

Proof. If $\lim_{x \rightarrow x_0} f(x) = L$, then by Theorem 4.1 the sequence $\{f(x_n)\}$ converges to L for all $\{x_n\}$ which converge to x_0 . By Proposition 3.2, L is unique. □

The proof really amounts to nothing more than saying “combine Theorem 4.1 and Proposition 3.2”. We can combine Theorem 4.1 with Theorem 3.2 to arrive at familiar properties of functions.

⁴⁸Such a sequence is guaranteed to exist because x_0 is a limit point of E (see Theorem 3.1).

⁴⁹The use of the same limit notation here may obscure the fact that $\{f(x_n)\}$ is a sequence such that $f(x_n) \rightarrow L$.

⁵⁰We take the δ from the definition of a limit of a function and we use it as the ε in the definition of a limit of a sequence.

⁵¹This is just the negation of Definition 4.1.

Theorem 4.2. Let X and Y be metric spaces, and $E \subseteq X$. Suppose there are some functions $f : E \rightarrow Y$ and $g : E \rightarrow Y$, and a limit point p of E . If $\lim_{x \rightarrow x_0} f(x) = A$ and $\lim_{x \rightarrow x_0} g(x) = B$, then

1. $\lim_{x \rightarrow x_0} (f + g)(x) = A + B$;
2. $\lim_{x \rightarrow x_0} (fg)(x) = AB$;
3. $\lim_{x \rightarrow x_0} (f/g)(x) = A/B$ if $B \neq 0$.

Proof. This follows from Theorem 4.1 and Theorem 3.2. \square

4.2 Continuous Functions

The definition of a limit of a function as given in [Definition 4.1](#) has two quirks. The first, which has been acknowledged twice, is that a function needn't actually take on the value of its limit at a point. We never require $f(x_0) = L$. Secondly, we do not require that $x_0 \in E$, we instead require that x_0 is a limit point of E . This means that not only can f be undefined at x_0 , but we shouldn't even be surprised that it is undefined at x_0 when $x_0 \notin E$, as $f : E \rightarrow Y$. We want to rule both of these out. Essentially, we want to define “nice” functions at a point x_0 to be those where $\lim_{x \rightarrow x_0} f(x) = L$. This is how we get our definition of continuity.

Definition 4.2. Let X and Y be metric spaces, $E \subset X$, $x_0 \in E$, and $f : E \rightarrow Y$. Then f is [continuous at \$p\$](#) if for all $\varepsilon > 0$, there exists a corresponding $\delta > 0$ such that

$$d_Y(f(x), f(x_0)) < \varepsilon$$

for all points $x \in E$ which satisfy $d_X(x, x_0) < \delta$. If f is continuous at every point of E , then f is [continuous \(on \$E\$ \)](#).

This definition is illustrated in Figure 38. Just like in calculus, a function is continuous at a point x if and only if

$$\lim_{t \rightarrow x} f(t) = f(x)$$

Remark 4.1. Go back and review Remark 2.4. This issue came up with sequences. We saw sequences that converged in one metric space, but not another. Similarly, a function can be continuous in one metric space but not another. It also can be continuous on one subset of a metric space, but not another. If a subset $E \subset X$ is never specified, then we assume that $E = X$. This is why we usually consider all of \mathbb{R} when determining if a real function is continuous.

Example 4.5. Suppose we define $f : \mathbb{R} \rightarrow \mathbb{R}$ as some linear function $f(x) = mx + b$. Let $x_0 \in \mathbb{R}$. We can mimic our work in Example 4.1 to show this function is continuous on all of \mathbb{R} . For $x_0 \in \mathbb{R}$ satisfying $|x - x_0| < \varepsilon/|m|$, we have

$$d_{\mathbb{R}}(f(x), f(x_0)) = |mx + b - (mx_0 + b)| = |m(x - x_0)| = |m||x - x_0| = |m| \cdot d_{\mathbb{R}}(x, p) < |m| \cdot \frac{\varepsilon}{|m|} = \varepsilon.$$

Therefore f is continuous at x_0 , for all $x_0 \in \mathbb{R}$

Example 4.6. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined as

$$\begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0 \end{cases}.$$

This function is continuous on the set $\mathbb{R} \setminus \{0\}$. It fails to be continuous on all of \mathbb{R} , as there is a “jump” from -1 to 1 at $x = 0$.

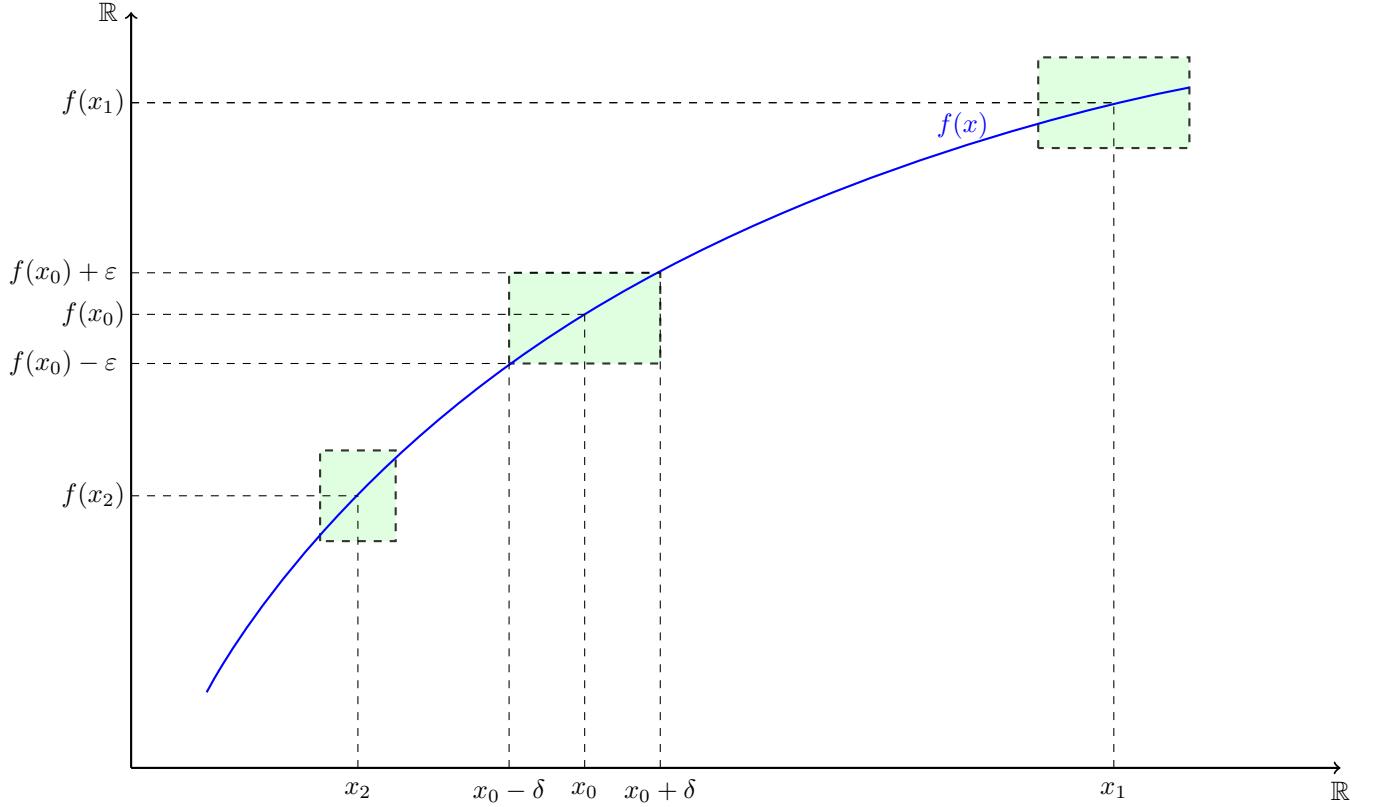


Figure 38: The function f is continuous at x_0 , x_1 , and x_2 . Given some $\varepsilon > 0$, there exist a delta such that $|f(x) - f(y)| < \varepsilon$ for all $|x - y| < \delta$. Each green box has height of 2ε , and the width for the 2δ at the corresponding point. Note that the value of δ is the same for x_0 and x_1 , but different for x_2 . For a fixed ε , the delta for one point may differ depending on the value of x . If some value of δ is too large, then the box would be too wide, and the function $f(x)$ would “escape” from the top and/or bottom of the box. This graphical feature would imply a violation of continuity, so we need to shrink δ until the function “escapes” from the sides (which is what we did for x_2).

Example 4.7. The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as $f(x) = x^2$ is continuous on \mathbb{R} . Let $x_0 \in \mathbb{R}$. For all $\varepsilon > 0$, define $\delta = \min\{1, \varepsilon/(2|x_0| + 1)\}$. We have

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - p| \cdot |x + p| < \delta(2|x_0| + 1) = \varepsilon,$$

for $x \in \mathbb{R}$ which satisfy $|x - x_0| < \delta$. Therefore f is continuous at x_0 , for all $x_0 \in \mathbb{R}$.

Example 4.8 (Thomae's function). Define the real function $f : [0, 1] \rightarrow [0, 1]$ as

$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ for } p, q \in \mathbb{Z} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases},$$

where we assume $p/q \in \mathbb{Q}$ is in simplest terms. The function f is continuous for the irrational numbers in $[0, 1]$ but discontinuous for the rational numbers in $[0, 1]$.

Suppose $r \in [0, 1]$ is irrational, giving $f(r) = 0$. Fix $\varepsilon > 0$. By the Archimedean Property, there is some

$m \in \mathbb{N}$ such that $1/m\varepsilon$, and some $k_m \in$ such that

$$r \in \left(\frac{k_m}{m}, \frac{k+1}{m} \right).$$

Define the following values:

$$d_m = \min \left\{ \left| r - \frac{k}{m} \right|, \left| r - \frac{k+1}{m} \right| \right\},$$

$$\delta = \min\{d_1, \dots, d_m\}.$$

If $x \in [0, 1]$ is a rational number with $|x - r| < \delta$, then $x = p/q$ and $q > m$, so

$$|f(x) - f(r)| = \left| \frac{1}{q} - 0 \right| = \frac{1}{q} < \frac{1}{m} < \varepsilon,$$

whenever $|x - r| < \delta$. If $x \in [0, 1]$ is irrational, then

$$|f(x) - f(r)| = |0 - 0| = 0 < \varepsilon,$$

regardless of δ . Thus, for any $x \in [0, 1]$ satisfying $|x - r| < \delta$, $|f(x) - f(r)| < \varepsilon$. The function is continuous at any irrational r .

Now suppose $r = p/q \in [0, 1]$ is rational. We have $f(r) = 1/q$. Let $\varepsilon = 1/2q$ and define $r_k = r + \frac{1}{k\sqrt{2}}$. The number r_k is irrational, so $f(r_k) = 0$. Can we find a δ such that

$$|f(x) - f(r)| = \left| f(x) - \frac{1}{q} \right| < \varepsilon = \frac{1}{2q}?$$

No we cannot. We can always find and x_k such that

$$|f(x_k) - f(r)| = \frac{1}{q} > \frac{1}{2q}.$$

Remark 4.2 (Pathological Examples). Thomae's function is constructed for the express purpose of being an unintuitive example. Such examples are known as pathological in math, and are particularly common in analysis. While such examples would never arise in the real world, they often give insight into just what is possible. We'll see several more: a function that is continuous everywhere but differentiable nowhere, a set that we can not possibly know the size of, a set that is uncountably infinite but has no size, etc.

Remark 4.3 (Can δ depend on x_0 ?). Yes! In Example 4.5, our selection of δ was independent of x , and works for every point in \mathbb{R} . This was not the case for Example 4.7, but that is fine. Our selection of δ will change for each value of $x_0 \in \mathbb{R}$. All that matters is *for each* $\varepsilon > 0$, we can find a correspond δ . We'll come back to this idea shortly.

Remark 4.4 (It Only Takes One ε , Discontinuity). Showing a function is *not* continuous at $x_0 = 0$ only requires we find a single ε such that $|f(x) - f(x_0)| \geq \varepsilon$ for all $|x - x_0| < \delta$ for $y \in E$. This follows from the negation of **Definition 4.1**. If we let $x \rightarrow 0$, then we end up with $\delta = 0$, a clear contradiction.

Our next makes explicit the fact that if f is continuous at a point, then the limit of the function exists at that point.

Theorem 4.3. Let X and Y be metric spaces, $E \subseteq X$, $x_0 \in E$ be a limit point of E , and $f : E \rightarrow Y$. Then f is continuous if and only if $\lim_{x \rightarrow x_0} f(x) = f(x_0)$.

Proof. Simply compare the definition of a limit (Definition 4.1), with the definition of continuity (Definition 4.2). \square

Corollary 4.2 (Continuous Functions Preserve Limits). Let X and Y be metric spaces, and $E \subset X$. Suppose there is some function $f : E \rightarrow Y$, and a $x_0 \in E$ which is a limit point of E . Then f is continuous at px_0 if and only if $f(x_n) \rightarrow f(x_0)$ for every non-constant sequence $\{x_n\}$ in E such that $x_n \rightarrow x_0$. That is

$$\lim_{n \rightarrow \infty} f(x_n) = f\left(\lim_{n \rightarrow \infty} x_n\right) = f(x_0).$$

Proof. This follows from Theorem 4.1 and Theorem 4.3. \square

Corollary 4.3 (Properties of Continuous Functions). Let X and Y be metric spaces, and $E \subseteq X$. Suppose there are some functions $f : E \rightarrow Y$ and $g : E \rightarrow Y$, and a point $x_0 \in E$. If f and g are continuous at p , then

1. $f + g$ is continuous at x_0 ;
2. fg is continuous at x_0 ;
3. f/g is continuous at x_0 if $x_0 \neq 0$.

Proof. This follows from Theorem 4.2 and Theorem 4.3. \square

Remark 4.5 (Moving Around Limits). Corollary 4.2 is the first time we've encountered a type of question analysis will provide answers for again and again – when can I move a limit into a function/operation? The majority of Section 7 will be dedicated to these type of questions. Corollary 4.2 tells us that a function is continuous if and only if the limit of a function is the function of the limits.

Example 4.9. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\begin{cases} x & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

This function is continuous at 0, because for all real $x_n \rightarrow 0$, $f(x_n) \rightarrow f(0) = 0$. This happens to be the only point in \mathbb{R} at which f is continuous. Let $c \in \mathbb{Q} \setminus \{0\}$. There is a sequence of irrational numbers $\{x_n\}$ such that $x_n \rightarrow c$. Despite this, $f(x_n) = 0 \rightarrow 0 \neq c$, so f is not continuous at c . Similarly, if $c \in \mathbb{R} \setminus \mathbb{Q}$, then there exists a sequence of rationals x_n such that $x_n \rightarrow c$. Despite this, $f(x_n) = x_n \rightarrow c \neq f(c) = 0$, so f is not continuous at c .

While Corollary 4.3 proves valuable when showing a function is continuous, when combined with our next result, we can show that almost any function we can write down is continuous.

Theorem 4.4 (Composition Preserves Continuity). Let X , Y , and Z be metric spaces, and $E \subseteq X$. If $f : E \rightarrow Y$ and $g : f(E) \rightarrow Z$ are continuous at $x_0 \in E$ and $f(x_0) \in f(E)$ respectively, then the function $h : g(f(E)) \rightarrow Z$ defined as

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

is continuous at x_0 .

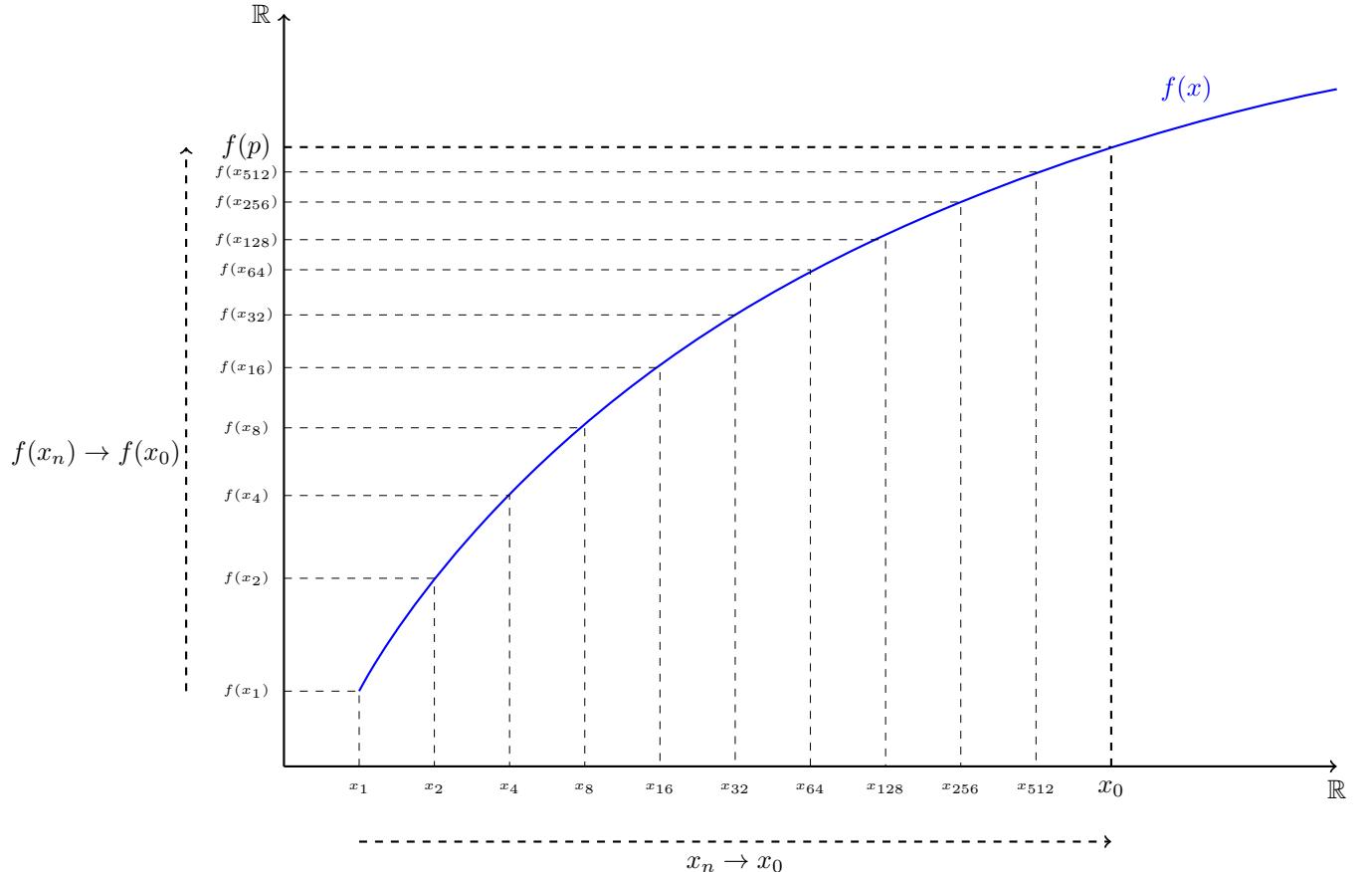


Figure 39: For $p_n \rightarrow p$, we have $f(p_n) \rightarrow f(p)$, therefore f is continuous at p .

Proof. Fix $\varepsilon > 0$. Since g is continuous at $f(x_0)$, there exists a δ' such that

$$d_Z(g(y), g(f(x_0))) < \varepsilon$$

if $d_Y(y, f(x_0)) < \delta'$ for $x_0 \in f(E)$. Since f is continuous at x_0 , there exists $\delta > 0$ such that

$$d_Y(f(x), f(x_0)) < \delta'$$

if $d_X(x, x_0) < \delta$ and $x_0 \in E$.⁵² These inequalities give

$$d_Z(g(f(x)), g(f(x_0))) = d_Z(h(x), h(x_0)) < \varepsilon$$

if $d_X(x, x_0) < \delta$ and $x \in E$. Therefore $h = g \circ f$ is continuous at x_0 . \square

If $h = g \circ f$ is continuous, then any composition of h with another continuous function will also be continuous. If we keep composing the result with continuous functions, we will always have a continuous function. For this reason, nearly every function encountered in actual applications will be continuous.

Example 4.10. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined as

$$f(x) = \left(\frac{e^{\sin x}}{\cos x} \right)^4.$$

⁵²We took the δ' from the definition of g 's continuity, and let $\varepsilon = \delta'$ for the definition of f 's continuity. This means that any $f(x)$ which satisfy $d_Y(f(x), f(x_0)) < \delta'$ will satisfy $d_Y(y, f(x_0)) < \delta'$.

If we let $g(x) = \cos x$, $h(x) = \sin x$, $j(x) = e^x$, $k(x) = x^4$ (all continuous functions from \mathbb{R} to \mathbb{R}), then

$$f(x) = k \circ ((j \circ h)/g).$$

By Theorem 4.4 and Corollary 4.3, $f(x)$ is continuous.

If a continuous function is a “nice” mapping from some metric space $E \subseteq X$ to another metric space Y , it’s worth asking what happens to sets when they mapped using f ? If E is open, will the image $f(E)$ be open? Does this hold for closed sets? What about our favorite sets, compact sets? We will now begin to put continuity in conversation with Section 2, and the point-set topology of metric spaces. We’ll end this subsection by introducing a *very important* theorem related to the topology of a metric space. We will then introduce a special type of continuity, before seeing how continuity interacts with compactness.

Theorem 4.5. Let X and Y be metric spaces. The function $f : X \rightarrow Y$ is continuous on X if and only if $f^{-1}(V)$ is open in X for every open set V in Y .

Proof.

(\Rightarrow) Suppose f is continuous on X and V is an open set in Y . We need to show that $f^{-1}(V)$ is open by showing every point of the set is an interior point. Let $x_0 \in f^{-1}(V) \subset X$, and $f(x_0) \in V$. The set V is open, so there exists an $r = \varepsilon > 0$ such that $B_\varepsilon(f(x_0)) \subset V$. Alternatively we may write, $y \in V$ for all $d_Y(f(x_0), y) < \varepsilon$. The function f is continuous at x_0 because $x_0 \in X$, so there exists a $\delta > 0$ such that $d_Y(f(x), f(x_0)) < \varepsilon$ is $d_X(x, x_0) < \delta$. In terms of sets, this means that $B_\delta(x_0) \subset f^{-1}(V)$, so $x_0 \in f^{-1}(V)$ is an interior point. Therefore $f^{-1}(V)$ is open.

(\Leftarrow) Suppose $f^{-1}(V)$ is open in X for all open sets V in Y . For $x_0 \subset X$ and $\varepsilon > 0$, let $V = B_\varepsilon(f(x_0))$. Alternatively, V is the set of all y such that $d_Y(y, f(x_0)) < \varepsilon$. The set V is an open ball, so it is open, hence $f^{-1}(V)$ is open. If $f^{-1}(V)$ is open, then there exists a δ such that $B_\delta(x_0) \subset f^{-1}(V)$. In other words, as soon as $d_X(x_0, x) < \delta$, $x \in f^{-1}(V)$. Because $x \in f^{-1}(V)$, $f(x) \in f(V)$, so we have $d_Y(f(x), f(x_0)) < \varepsilon$ as soon as $d_X(x_0, x) < \delta$. This means f is continuous on X .

□

Corollary 4.4. Let X and Y be metric spaces. The function $f : X \rightarrow Y$ is continuous on X if and only if $f^{-1}(V)$ is closed in X for every closed set V in Y .

Proof. The complement of an open set is closed, and $f^{-1}(V^c) = (f^{-1}(V))^c$. □

Theorem 4.5 is illustrated in Figure 40 for a real function.

Remark 4.6 (Topological Definition of Continuity). We’ve already referenced several times the fact that we are restricting our attention to metric spaces, but that point-set topology can be studied in a more general setting. As it turns out, when working in a topological space that is not a metric space, we define a continuous function as one with the property established in Theorem 4.5. If you were to ever take a proper topology course, continuity is never discussed in terms of $\varepsilon - \delta$, but is instead related to open sets.

Example 4.11. Theorem 4.5 only says that the preimage of an open set will be open when a function is continuous. It never says that the image of an open set is open. If we define $f : \mathbb{R} \rightarrow \mathbb{R}$ as $f(x)^2$, and let $E = (-1, 1)$ be an open set, then

$$f(E) = [0, 1).$$

Therefore $f(E)$ is not open, despite f being continuous on \mathbb{R} .

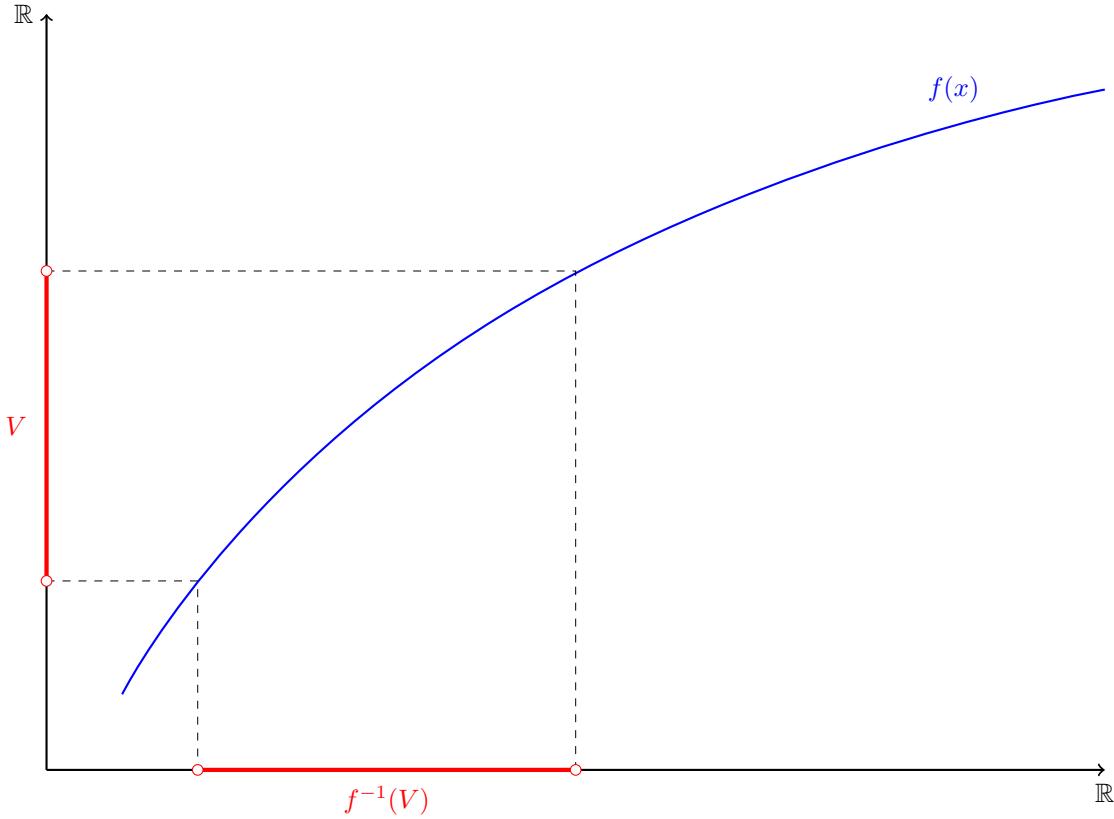


Figure 40: A continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$, and the open interval $V \subset \mathbb{R}$. Because f is continuous, $f^{-1}(V)$ is open in \mathbb{R} .

Example 4.12. Theorem 4.5 greatly simplifies the proof of Theorem 4.4. For an open $E \subset Z$, $g^{-1}(E) \subset Y$ is open by the continuity of g . The set $f^{-1}(g^{-1}(E)) \subset X$ is open because f is continuous. Therefore, for $h = g(f(x))$, $h^{-1}(E) \subset X$ is open, making h continuous.

4.3 Uniform Continuity

Before exploring compactness and continuity, we need to address an earlier remark, namely Remark 4.3. We noticed that the value of δ which corresponds to ε sometimes depends on where we are in the domain of a function. While does not violate continuity, it is a bit odd. If f is continuous on a set E , then we can always find a corresponding δ for each ε , but that δ may not work for all values of E . What we are really doing is find a corresponding δ for each ε and for each $x_0 \in E$. However, Example 4.5 showed us that sometimes our corresponding δ works for all values on the domain. This is a special case that merits its own definition.

Definition 4.3. Let X and Y be metric spaces. The function $f : X \rightarrow Y$ is *uniformly continuous (on X)* if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, x_0) < \delta$ implies $d_Y(f(x), f(x_0)) < \varepsilon$ for all $x, x_0 \in X$.

There are two major differences between uniform continuity and continuity. We already discussed the first – we find a single value δ which works for all elements of f 's domain. This difference arises from the fact that in Definition 4.2, we only looked at $x_0 \in E$ which satisfy $d_X(x, x_0) < \delta$. We don't do this in definition 4.3, as we look at all $x_0 \in E$. The second is that uniform continuity pertains to sets, not points. While a

function can be continuous at a single point, saying f is uniformly continuous at a point it vacuously true and meaningless, as the concept deals with multiple points in a domain. It also should be clear that any function which is uniformly continuous is continuous.

It cannot be emphasized enough how much stronger uniform continuity is than continuity. If $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, then for every $\varepsilon > 0$ (of which there are an uncountably infinite amount), we may need to find a corresponding δ for each point in \mathbb{R} ! It is feasible that we have an uncountably infinite amount of δ for each ε . If instead f is uniformly continuous, for each fixed $\varepsilon > 0$ (of which there are still an uncountably infinite amount), then we know there is one “silver bullet” δ that works for all of \mathbb{R} . This difference gives rise to an interesting geometric interpretation presented in Figure 41 and Figure 42.

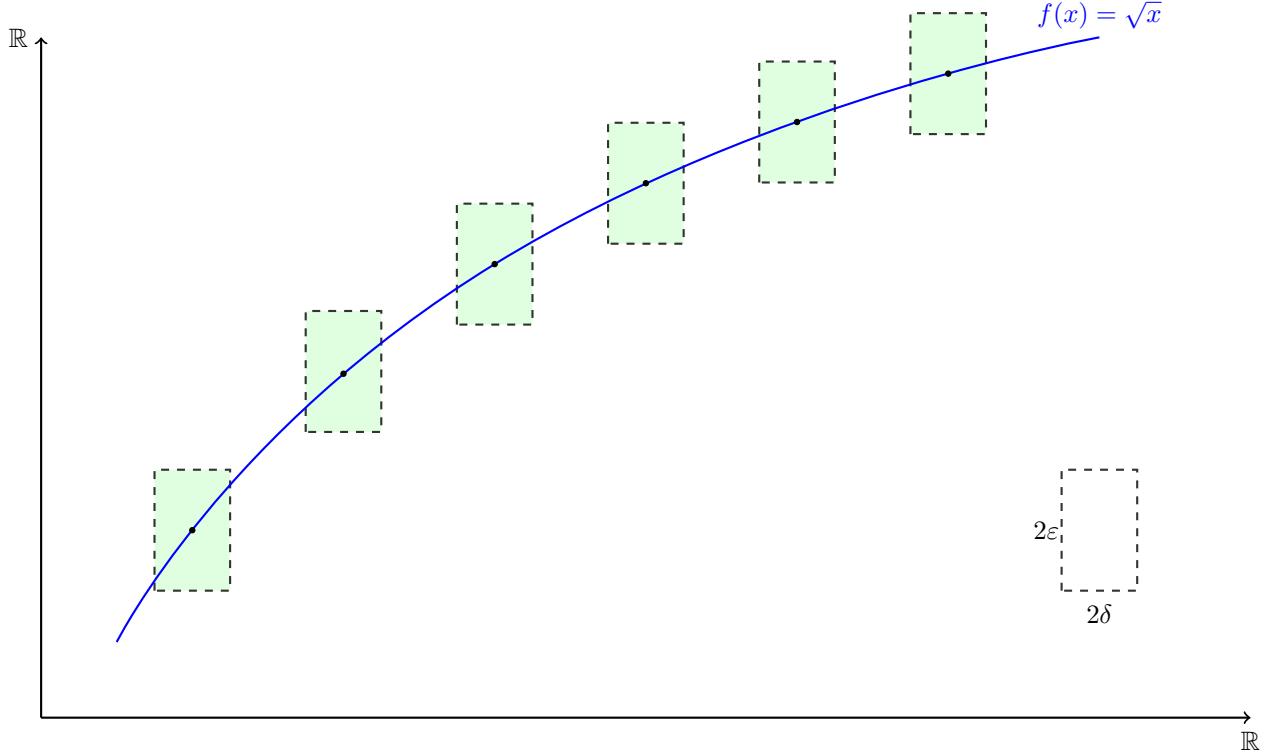


Figure 41: Let $f : (0, 1] \rightarrow \mathbb{R}$ be defined as $f(x) = \sqrt{x}$. The function is uniformly continuous. For each fixed ε , we can center a box of height ε at a point $(x, f(x))$. We can find some width of the box δ such that the function will never “escape” from the top or bottom of the box, no matter the point $(x, f(x))$.

Example 4.13. Let $f : [0, \infty) \rightarrow \mathbb{R}$ be defined as $f(x) = \sqrt{x}$. The function is uniformly continuous on $[0, \infty)$. If we let $\delta = \varepsilon^2$, then for all $|x - x_0| < \delta$, we have

$$|f(x) - f(x_0)|^2 = |\sqrt{x} - \sqrt{x_0}|^2 \leq |\sqrt{x} - \sqrt{x_0}| \cdot |\sqrt{x} + \sqrt{x_0}| = |x - x_0| < \varepsilon^2.$$

Taking the square root of this inequality gives $|f(x) - f(x_0)| < \varepsilon$ for all $|x - x_0| < \delta$. Our choice of δ does not depend on $x_0 \in \mathbb{R}$, so f is uniformly continuous.

Example 4.14. The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as $f(x) = 1/x$ is not uniformly continuous on $(0, 1)$. Suppose for contradiction that f is uniformly continuous on $(0, 1)$. Let $\varepsilon = 1/2$. There exists a δ such that

$$|f(x) - f(x_0)| = \left| \frac{1}{x} - \frac{1}{x_0} \right| < \varepsilon = \frac{1}{2}$$

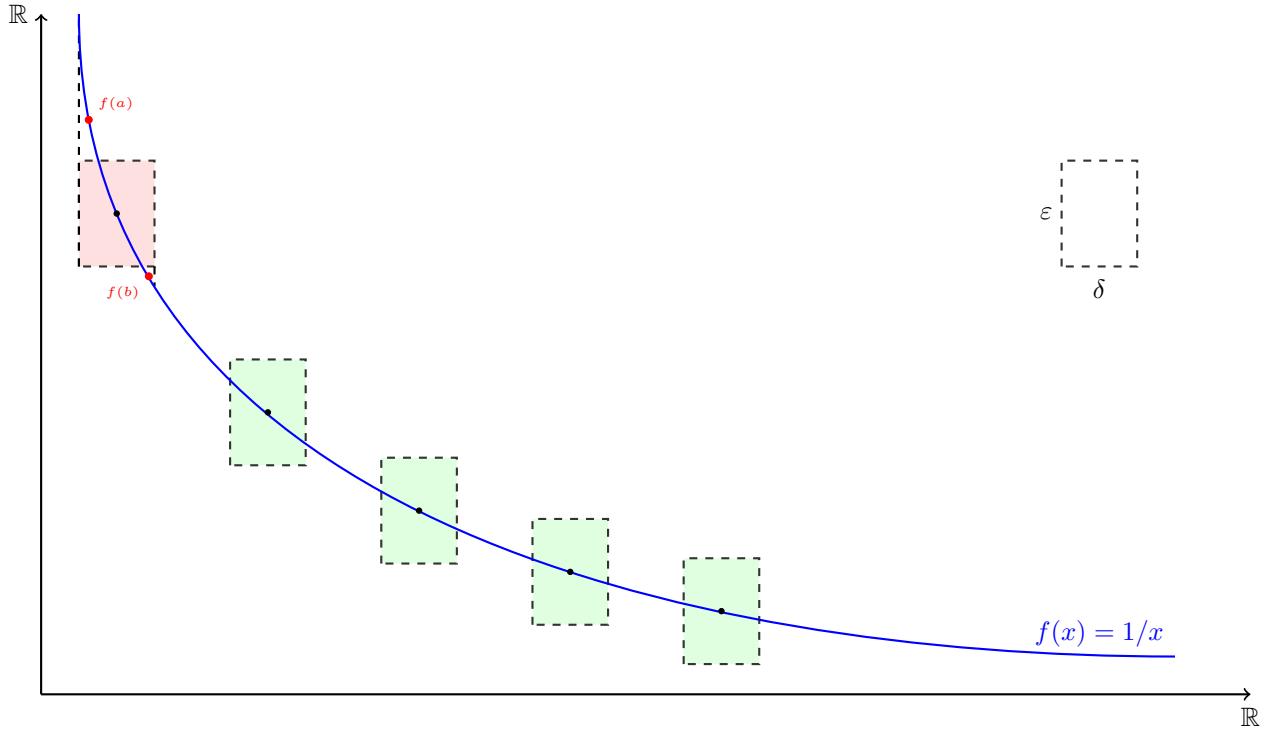


Figure 42: The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as $f(x) = 1/x$ is not uniformly continuous. For our set value of ε and δ the function “escapes” from the top and bottom of the red rectangle. This is because there exists $a, b \in \mathbb{R}$ such that $|f(a) - f(b)| \geq \varepsilon$ despite the fact that $|a - b| < \delta$. We could make the corresponding δ smaller, but the function eventually “escape” as we move to the left.

whenever $|x - x_0| < \delta$. Let $x < \delta$ and $x_0 = x/2$. In this case we have $|x - x_0| < \delta$ as

$$|x - x_0| = \left| x - \frac{x}{2} \right| = \left| \frac{x}{2} \right| < |x| < \delta,$$

but we also have

$$\left| \frac{1}{x} - \frac{1}{x_0} \right| = \left| \frac{1}{x} - \frac{2}{x} \right| = \left| -\frac{1}{x} \right| = \frac{1}{x} > \varepsilon$$

since $x \in (0, 1)$. This is a contradiction.

4.4 Continuity and Compactness

If continuous functions are “nice”, and compact sets are “nice”, then it should not come as a surprise that continuous functions behave very well with compact sets. We will first define what it means for a function to be bounded, and then we will jump into a series of results about compactness.

Definition 4.4. Let $E \subseteq \mathbb{R}$. The real function $f : E \rightarrow \mathbb{R}$ is *bounded* if there exists an $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in E$.

Figure 43 shows a bounded real function.

Example 4.15. The function $\sin(x)$ is bounded on all of \mathbb{R} , as $|\sin x| \leq 1$ for all $x \in \mathbb{R}$.

Example 4.16. The function $f(x) = x^2$ on \mathbb{R} is continuous but not bounded.



Figure 43: A bounded function $f : [a, b] \rightarrow \mathbb{R}$

Theorem 4.6 (Continuity Preserves Compactness). Let $f : X \rightarrow Y$ be a continuous function where X and Y are metric spaces. If X is compact, then $f(X)$ is compact.

Proof. We will show that an arbitrary open cover $\{V_\alpha\}$ of $f(X)$ has a finite subcover. Since f is continuous, $f^{-1}(V_\alpha)$ is open for all α (Theorem 4.5). The set X is compact, so there exists finitely many indices $\alpha_1, \dots, \alpha_n$ such that

$$X \subset f^{-1}(V_{\alpha_1}) \cup \dots \cup f^{-1}(V_{\alpha_n}).$$

But we have $f(f^{-1}(E)) \subset E$ for any $E \subset Y$, so we can take the image of the finite subcover of X and conclude

$$f(X) \subset V_{\alpha_1} \cup \dots \cup V_{\alpha_n}.$$

Therefore $f(X)$ is compact. □

This result is similar to Theorem 4.5, but they “go in different directions”. The preimage of an open set is open for a continuous f . Now we are saying that the image of a compact set is compact for a continuous f . It’s tempting to say this holds for open sets, but it is not true. In general openness is not preserved by continuous functions (Example 4.11).

Corollary 4.5. Let $E \subset \mathbb{R}$. If the real function $f : E \rightarrow \mathbb{R}$ is continuous and E is closed and bounded, then $f(E)$ is closed and bounded.

Example 4.17. Let $f : (0, \infty) \rightarrow \mathbb{R}$ be the continuous function $f(x) = 1/x$. The set $(1, \infty)$ is neither closed nor bounded, so $f^{-1}((1, \infty)) = (0, 1)$ is neither closed nor bounded.

Our next theorem proves useful in many applications, as it allows us to determine when a function has a maximum and minimum on an intervals.

Theorem 4.7 (Extreme Value Theorem). Suppose f is a continuous real function on a compact metric space X , and

$$M = \sup_{x \in X} f(x),$$

$$m = \inf_{x \in X} f(x).$$

Then there exists points $x, y \in X$ such that $f(x) = M$ and $f(y) = m$.

Proof. Continuous functions preserve compactness (Theorem 4.6), so $f(X) \subset \mathbb{R}$ is closed and bounded. A closed and bounded set contain their infimum and supremum, so $f(X)$ contains M and m . \square

Remark 4.7 (sup or max?). It may not be clear when we use maximum and when we use supremum. A maximum of some set or function is always attained. This is not always the case for a supremum. If we know that $\sup E \in E$, then we are free to write $\max E = \sup E$, but this won't always hold. For instance, $(0, 1)$ has a supremum of 1, but no well defined maximum. If you're ever unsure if the supremum is attained in the set or by the function, use sup. A maximum is always a supremum, so it technically is not incorrect. The same holds for inf and min.

Theorem 4.7 tells us that if a function is defined on a compact space, then it *must* achieve a maximum and minimum on that domain. Just knowing that such points exist is a great deal of information.

Example 4.18. The function $\sin(x)$ does not achieve a maximum or minimum on the interval $(0, \pi/2)$, despite it being bounded. We can always find some larger value of $f(x)$ as we get arbitrarily close to $\pi/2$, or some smaller value of $f(x)$ as we get arbitrarily close to 0. This is the exact type of behavior we ruled out when defining compactness! By defining $\sin(x)$ on $[0, \pi/2]$, we now achieve a maximum and minimum on the interval.

Example 4.19. The function $f(x) = x^2$ achieves a minimum on \mathbb{R} , but it fails to achieve both a maximum and a minimum on all of \mathbb{R} . This is because \mathbb{R} is neither closed nor bounded, rendering it not compact.

Recall the fact that $f(x) = 1/x$ is not uniformly continuous on $(0, 1)$. In Figure 42 we argued that no matter the value of δ given for a fixed ε , we could move the rectangle of height ε and length δ to the left until the function “escaped” from the top and bottom. Us being able to keep moving the rectangle to the left is a result of $(0, 1)$ not being closed. We can always get a little bit closer to 0 without leaving the domain. As we do this, the function will take on larger values indefinitely because $f((0, 1)) = (1, \infty)$ is unbounded. These two observations seem to hint at the fact that if eliminate these behaviors, a function will always be uniformly continuous. As it turns out, compactness does just this, and the proof of this is one of the more elegant proofs in analysis.⁵³

Theorem 4.8. Let f be a continuous function which maps a compact metric space X to a metric space Y . Then f is uniformly continuous on X .

Proof. Fix $\varepsilon > 0$. The function f is continuous on X , so for each point $p \in X$ there is a $\delta(p)$ such that

$$d_Y(f(p), f(x)) < \varepsilon/2$$

⁵³Full disclosure: this is my favorite proof from basic real analysis.

whenever $x \in X$ and $d_X(p, x) < \frac{1}{2}\delta(p)$.⁵⁴ Let

$$J(p) = \left\{ x \in X \mid d_X(x, p) < \frac{1}{2}\delta(p) \right\}.$$

We have $p \in J(p)$ for all $p \in X$, so

$$X \subset \bigcup_{p \in X} J(p),$$

where $J(p)$ is open for all $p \in X$.⁵⁵ That is to say, $\{J(p)\}_{p \in X}$ is an open cover of X . The space X is compact, so there exists a finite set of points p_1, \dots, p_n such that

$$X \subset J(p_1) \cup \dots \cup J(p_n).$$

If we let

$$\delta = \frac{1}{2} \min\{\delta(p_1), \dots, \delta(p_n)\},$$

then $\delta > 0$.⁵⁶

Let $x, p \in X$ such that $d_X(x, p) < \delta$. The point p is in X , so it must be “covered” by one of the open sets in the finite subcover $\{J(p_1), \dots, J(p_n)\}$. That is, there is an $m \in \mathbb{N}$, where $1 \leq m \leq n$, such that $p \in J(p_m)$. Hence, $d_X(x, p_m) < \frac{1}{2}\delta(p_m)$. The triangle inequality gives

$$d_X(x, p_m) \leq d_X(x, p) + d_X(p, p_m) < \delta + \frac{1}{2}\delta(p_m) \leq \delta(p_m).$$

But then, we can use the triangle inequality and the continuity of f to conclude

$$d_Y(f(x), f(p)) \leq d_Y(f(x), f(p_m)) + d_Y(f(p_m), f(p)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

so f is uniformly continuous. □

This proof may look technical, but the underlying idea is much simpler. Think about the function $f(x) = 1/x$ on $(0, 1]$, and Figure 42. For a fixed ε , we can always move the rectangles to the left until the function “escapes” from the top or bottom. The closer we get to 0 on the x -axis, the smaller we need to make δ , and we can always move closer and closer to 0, so there isn’t one δ that will work at every point. Suppose instead we looked at $f(x) = 1/x$ on the interval $[0.001, 1]$. We may be able to get closer and closer to 0.001, making δ smaller as we go to satisfy continuity, but because 0.001 is defined by f , we could just take the δ that works at 0.001 and use it for all of $[0.001, 1]$. If the δ needs to become small as we move our rectangle to the left, then the “leftmost” point must have the smallest δ , a choice of δ that will work for all of $[0.001, 1]$ and our fixed ε . The next example will walk through this process, and the proof of Theorem 4.8, with an actual function.

Example 4.20. Suppose $f : [-10, 10] \rightarrow \mathbb{R}$ is defined as $f(x) = x^2$. Example 4.7 showed that f is continuous. In this case we had $\delta = \min\{1, \varepsilon/(2|p| + 1)\}$. Fix $\varepsilon = 1$, so $\varepsilon/2 = 1/2$.⁵⁷ We can let $\delta(p) = 1/(4|p| + 1)$, and

⁵⁴We are writing δ as a function of $p \in X$ to emphasize that all we know now if f is continuous.

⁵⁵This follows from the fact that $J(p)$ is an open ball of p , and all open balls are open sets. We could also write $J(p) = B_{1/2\delta(p)}(p)$.

⁵⁶This finite set achieves its infimum so we use min. A minimum of a finite set of positive numbers is positive. This is not necessarily the case with an infimum of an infinite set of positive numbers even if that set is compact. For us to be able to conclude $\delta > 0$, it is crucial that the set $\{\delta(p_1), \dots, \delta(p_n)\}$ is finite, that way we can take the minimum over a finite set corresponding to the finite subcover.

⁵⁷We use $\varepsilon = 1$ as the final value of ε used to show uniform continuity. The value $\varepsilon/2$ is used in the definition of continuity of x^2 .

have

$$|f(x) - f(p)| < \varepsilon = 1/2$$

for all $p \in [-10, 10]$ satisfying $|x - p| < \frac{1}{2}\delta(p) < \delta(p)$. First, we should explicitly show that $\delta(p)$ will change with p . Suppose $p = 1$, giving $\delta(1) = 1/5$. For all $x \in [-10, 10]$ such that $|x - 1| < 1/5$, we do indeed have $|x^2 - 1| < 1/2$. Now let's try to use $\delta(1)$ for $p = 9$. The set of all $x \in [-10, 10]$ which satisfy $|x - 9| < 1/5$ is $(44/5, 46/5)$. For $91/10 \in (44/5, 46/5)$, we have

$$|f(91/10) - f(9)| = \left(\frac{91}{10}\right)^2 - 9^2 = 1.81 > \frac{1}{2} = \frac{\varepsilon}{2}.$$

This means that $\delta(1)$ won't work for $p = 9$! We now will mimic the proof of Theorem 4.8 and find a δ that will work for all $p \in [-10, 10]$.

Define

$$J(p) = \left\{ x \in [-10, 10] \mid |x - p| < \frac{1}{2}\delta(p) = \frac{1}{8|p| + 2} \right\} = B_{1/(2\delta(p))} = B_{1/(8|p|+2)}(p).$$

We have $p \in J(p)$ for all $p \in [-10, 10]$, so

$$[-10, 10] \subset \bigcup_{p \in [-10, 10]} J(p).$$

For the finite set of points $P = \{-10, -9.9, \dots, 9.9, 10\} = \{-10 + (0.1)n \mid n = 0, \dots, 200\}$,⁵⁸ we have

$$[-10, 10] \subset \bigcup_{p \in P} J(p),$$

making $\{J(p)\}_{p \in P}$ a finite subcover of $[-10, 10]$. Define

$$\delta = \min_{p \in P} \{\delta(p)\} = \delta(10) = \delta(-10) = \frac{1}{81}.$$

We will show that this value of δ will work for a random point in $[-10, 10]$, say $p = 1.95$. We could verify this right away, but we'll instead follow the steps of the proof, even though they become painfully redundant when working with actual numbers. Now let $x \in [-10, 10]$ such that $|x - 1.95| < 1/81$.⁵⁹ We have $x \in [-10, 10]$, so it must be in one of the elements of the finite subcover $\{J(p)\}_{p \in P}$. We in fact have $1.95 \in J(1.9)$, hence

$$|1.95 - 1.9| < \frac{1}{2}\delta(1.95) = 0.0581,$$

so the continuity of f at $p = 1.95$ gives

$$|f(1.95) - f(1.9)| = 0.1925 < \frac{\varepsilon}{2} = \frac{1}{2}. \tag{5}$$

We also have

$$|x - 1.9| \leq |1.95 - x| + |1.95 - 1.9| < \delta + \frac{1}{2}\delta(1.95) = \frac{1}{81} + 0.0581 = 0.0704 < \delta(1.95) = .104,$$

⁵⁸Where the heck do I get this set? Well the minimum value of $1/(8|p| + 2)$ on the interval $[-10, 10]$ is about 0.012. If we round this down to 0.01, then all the $J(p)$ centered at points in $[-10, 10]$ that are distance 0.01 apart will be guaranteed to cover $[-10, 10]$, as we took the distance between points to be less than the smallest radius of $J(p)$.

⁵⁹For the rest of this paragraph, whenever we refer to x , we mean these particular x satisfying $|x - 1.95| < 1/81$

so by the continuity of f at 1.9,

$$|f(x) - 1.9| < \frac{\varepsilon}{2} = \frac{1}{2}. \quad (6)$$

Combining Equation (5) and (6) gives

$$|f(x) - f(1.95)| \leq |f(x) - f(1.9)| + |f(1.9) - f(1.95)| < \frac{1}{2} + \frac{1}{2} = 1 = \varepsilon$$

for all $x \in [-10, 10]$ such that $|x - 1.95| < \delta = 1/81$. Our choice of δ does in fact work for $\varepsilon = 1$ and $p = 1.95$!

The choice of $p = 1.95$ does not matter. We could have picked *any* value in $[-10, 10]$, and $\delta = 1/81$ still would have worked for $\varepsilon = 1$. This should not come as a surprise, because we just picked *one* of the smallest value of $\delta(p)$ for $p \in [-10, 10]$. Why is it that we need compactness if we could have just let $\delta = \min_{p \in [-10, 10]} \delta(p)$? This would work in this specific case because the range $\delta([-10, 10])$ is compact, due to $\delta(p)$ being continuous for $\varepsilon = 1$ (Theorem 4.6), and we the infimum of a compact set in \mathbb{R} is an element of the set. In general this won't work, because we never explicitly said that $\delta(p)$ is a function, let alone a continuous function.

Example 4.21. Any real continuous function $f : [a, b] \rightarrow \mathbb{R}$ is uniformly continuous on $[a, b]$, because $[a, b] \subset \mathbb{R}$ is compact.

Example 4.22. A compact domain is a sufficient condition for uniform continuity, but it is not a necessary condition. Any linear real function $y = mx + b$ is uniformly continuous on all of \mathbb{R} , and \mathbb{R} is not compact.

4.5 Intermediate Value Theorem

We now will turn our attention to one of the major theorems presented in a calculus course. The Intermediate Value Theorem is perhaps the quintessential result of continuity. In calculus, you may have been taught that a continuous function is any function you could draw without picking up your pencil. If this is the case, and the range of your function starts at $f(a)$ and ends at $f(b)$, then your function will of course take on every value between $f(a)$ and $f(b)$.

Theorem 4.9. Let f be a real function continuous on the interval $[a, b]$. If $f(a) < f(b)$ and if c is a number satisfying $f(a) < c < f(b)$, then there is a $x \in (a, b)$ such that $f(x) = c$.

Proof. Suppose $f : [a, b] \rightarrow \mathbb{R}$ is continuous, and $f(a) < f(b)$. Let $c \in \mathbb{R}$ such that $f(a) < c < f(b)$. Define the set

$$S = \{x \in [a, b] \mid f(x) \leq c\}.$$

The set $S \subset [a, b]$ is nonempty⁶⁰ and bounded above by b , so $s = \sup S$ exists by the least-upper-bound property. We claim that $f(s) = c$. We will show that $f(s) \not< c$ and $f(s) \not> c$.

Suppose $f(s) > c$, and let $\varepsilon = f(s) - c > 0$. By the continuity of f , there exists a $\delta > 0$ such that

$$|f(x) - f(s)| < f(s) - c = \varepsilon$$

for all $x \in [a, b]$ which satisfy $|x - s| < \delta$. For all such x we can conclude $f(x) > c$,⁶¹ so $x \notin S$. But if this holds for any $x \in [a, b]$ such that $|x - s| < \delta$, then $s - \delta$ is an upper bound of S . This contradicts $s = \sup S$.

⁶⁰ $a \in S$

⁶¹ $|f(x) - f(s)| < f(s) - c$ implies $f(x) - f(s) < f(s) - c$ or $f(s) - f(x) < f(s) - c$. Either way, $f(x) > c$.

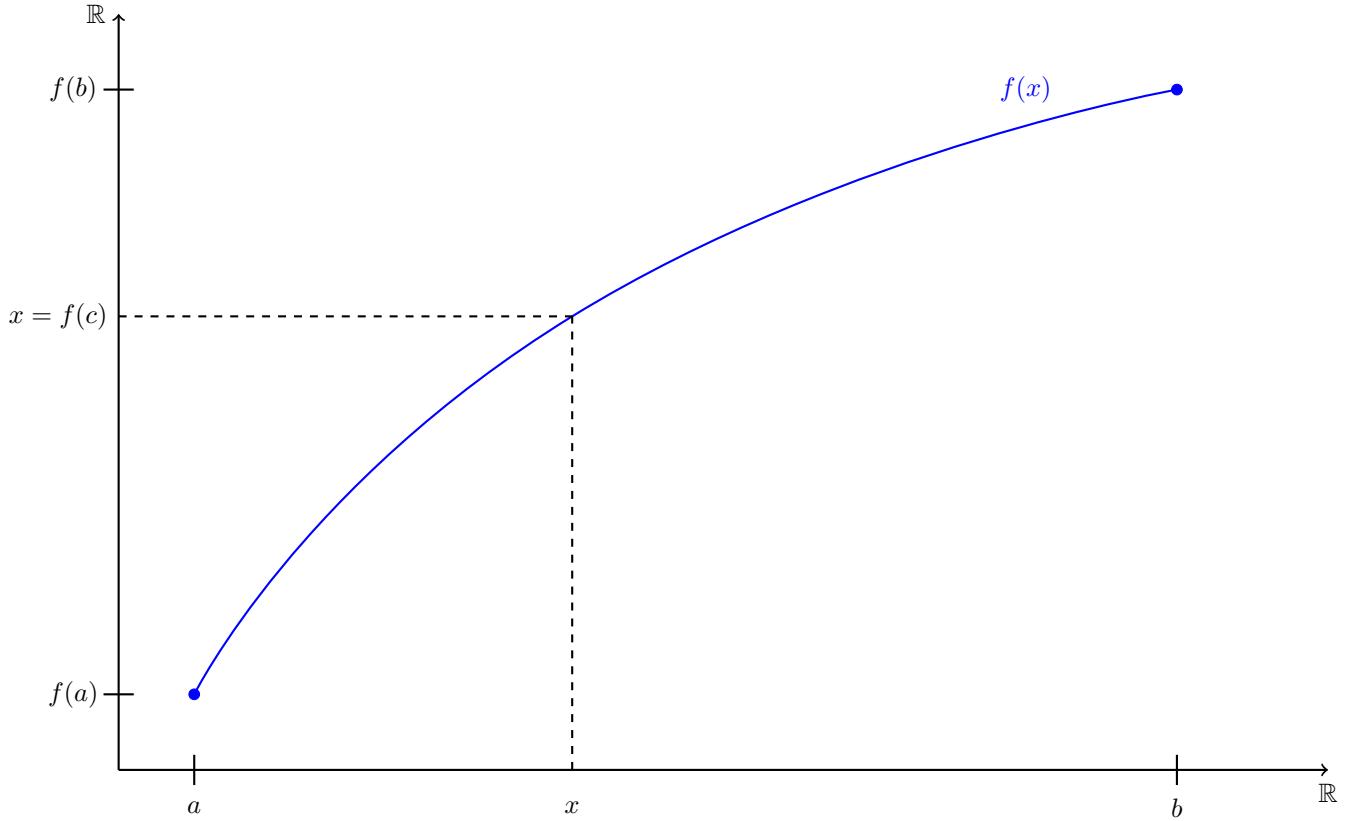


Figure 44: We have a real function $f : [a, b] \rightarrow \mathbb{R}$. The Intermediate Value Theorem says for all $f(a) < c < f(b)$, there is a $x \in (a, b)$ such that $f(x) = c$.

Suppose $f(s) > c$, and let $\varepsilon = c - f(s) > 0$. By the continuity of f , there exists a $\delta > 0$ such that

$$|f(x) - f(s)| < c - f(s) = \varepsilon$$

for all $x \in [a, b]$ which satisfy $|x - s| < \delta$. For all such x we can conclude $f(x) < y$, so $x \in S$. This implies that $s + \delta/2 \in S$ is an upper bound, which contradicts $s = \sup S$ being an upper bound. \square

The Intermediate Value Theorem may seem like a parlor trick, but it is useful in many proofs. There are many proofs that rely on us being able to find a certain value in an interval. If we have information about some function, and know the Intermediate Value Theorem holds, then we may be able to use it to find the value we're interested in. Many proofs are about “converting” information. We may have information about f , but need information about points in the domain of f . In a sense, Theorem 4.9 allows us to convert information about f *into* the information about the domain.

Remark 4.8 (Bounded Intervals, Figures). Figure 44 is one of the first times where it's been important that a function is shown on a bounded interval $[a, b]$. For all future figures, if a function is on some bounded interval, or we are restricting our attention to a bounded interval, I will try to denote that with solid dots at the “beginning” and “end” of the function, or mark a and b on the x -axis. We saw this with Figure 43 as well. If no such markings are present, then I'm trying to convey that f is defined on all of \mathbb{R} .

Example 4.23 (Continuous Functions on \mathbb{Q}). Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous such that $f([a, b]) \subset \mathbb{Q}$.⁶² The function f must be constant, that is for all $x \in [a, b]$, $f(x) = c$ for some $c \in \mathbb{R}$. Without loss of generality, assume $f(a) < f(b)$. By the Intermediate Value Theorem, the function takes on every value in the interval $(f(a), f(b))$. This interval contains irrational numbers, as the irrational numbers are dense in \mathbb{R} . This is a contradiction, so f is constant.

Example 4.24 (Roots of a Polynomial). Suppose

$$p(x) = a_0 + a_1x + \dots + a_kx^k$$

is an odd polynomial (k is an odd number) with real coefficients $a_i \in \mathbb{R}$. Taking the limit of $p(x)$ as x goes to infinity and negative infinity gives

$$\begin{aligned}\lim_{x \rightarrow -\infty} f(x) &= -\infty, \\ \lim_{x \rightarrow \infty} f(x) &= \infty.\end{aligned}$$

This is the first time we've seen limits that go off to infinity, but if we combine what we know about sequences that diverge to infinity and Theorem 4.1, we can conclude that the limit does not exist, despite us using notation that would indicate it does. These attempts to take limits also lets us know that at some point in \mathbb{R} , $p(x)$ switches signs. Therefore by the Intermediate Value Theorem, there exists at least one root $x_0 \in \mathbb{R}$ such that $p(x_0) = 0$.⁶³

Example 4.25. The converse of the intermediate value theorem is not true. That is, just because for any two points $x_1 < x_2$ and a number c in between $f(x_1)$ and $f(x_2)$ we are able to find a point $x \in (x_1, x_2)$ such that $f(x) = c$, that *does not mean* f is continuous. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ to be

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}.$$

This function satisfies the aforementioned property, but is not continuous on \mathbb{R} , as there is a discontinuity at 0.

4.6 Discontinuities

We've used the word discontinuous several times, but now we'll take the time to define it, and classify two types of discontinuities. A function that is discontinuous at a point is simply not continuous at the point. The formal definition we will present negates the statement given in Definition 4.2

Definition 4.5. Let X and Y be metric spaces, $E \subset X$, $p \in E$, and $f : E \rightarrow Y$. Then f is *discontinuous at p* if for all $\delta > 0$, there exists a single $\varepsilon > 0$ such that

$$d_Y(f(x), f(p)) \geq \varepsilon$$

for a single $x \in E$ which satisfies $d_X(x, p) < \delta$. If f is discontinuous at least one point in E , then f is *discontinuous (on E)*.

⁶²The function is still real, as $\mathbb{Q} \subset \mathbb{R}$.

⁶³We do not know for sure how many roots of $p(x)$ are in \mathbb{R} , but we know it's at least one. If we wanted to look for the other roots, we would find exactly k of them in \mathbb{C} ! This follows from the Fundamental Theorem of Algebra which says that any complex polynomial of degree k has k roots in \mathbb{C} . If these sorts of facts interest you, see [Dummit and Foote \(2004\)](#).

We can classify three different types of discontinuities if we introduce the notion of a right-hand limit and a left-hand limit for a real function.

Definition 4.6. Let f be a real function defined on (a, b) , and consider any point p such that $a \leq p < b$. If for all sequence $\{p_n\}$ in (p, b) such that $p_n \rightarrow p$, we have $f(p_n) \rightarrow L$, then we write $f(p+) = L$, or

$$\lim_{x \rightarrow p^+} f(x) = \lim_{x \searrow p} f(x) = \lim_{x \downarrow p} f(x) = p.$$

We say that L is the right-hand limit of $f(x)$ at x .

Definition 4.7. Let f be a real function defined on (a, b) , and consider any point p such that $a < p \leq b$. If for all sequence $\{p_n\}$ in (a, p) such that $p_n \rightarrow p$, we have $f(p_n) \rightarrow L$, then we write $f(p-) = L$, or

$$\lim_{x \rightarrow p^-} f(x) = \lim_{x \nearrow p} f(x) = \lim_{x \uparrow p} f(x) = p.$$

We say that L is the left-hand limit of $f(x)$ at x .

An alternate definition of these limits would use ε and δ . For example, $f(p-) = L$ if for all $\varepsilon > 0$, we have $|f(p) - L| < \varepsilon$ for all $x \in (a, b)$ satisfying $p - \delta < x < p$. The only difference between this and the definition of a limit is we restrict our attention to the x that are within a distance of δ to the left of p . The definition for the right hand limit would be the same, except we look at the interval $p < x < x + \delta$. It can be proven that if $f(x+) = f(x-)$, then $\lim_{x \rightarrow x_0} f(x) = p$. The first type of discontinuity is the type we saw in Figure 37. This type is special, as it does not prohibit the existence of a limit at a point.

Definition 4.8. Let f be a real function defined on (a, b) . If f is discontinuous at a point x , and if $f(x+) = f(x-)$, then we say f has a removable discontinuity at x .

With removable discontinuities, it's important to remember that f still needs to be defined at the point of discontinuity. It doesn't mean anything to say that f has a discontinuity at a point where it is not defined. Our second type of discontinuity is the type shown in Figure 45.

Definition 4.9. Let f be a real function defined on (a, b) . If f is discontinuous at a point x , and if $f(x+) \neq f(x-)$, then we say f has a jump discontinuity at x .

Lastly we treat the case shown in Figure 46.

Definition 4.10. Let f be a real function defined on (a, b) . If f is discontinuous at a point x , and if either $f(x-)$, or $f(x+)$ (or both) do not exist, then we say f has an essential discontinuity at x .

Example 4.26 (Dirichlet Function, Nowhere Continuous). Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

This function takes the value 1 at every rational, and 0 at every irrational. This function happens to be discontinuous on all of \mathbb{R} ! Let ε be any number in $(0, 1]$ such that $1/2$. If $x \in \mathbb{Q}$, for any value of $\delta > 0$, we can always find some irrational $y \notin \mathbb{Q}$ such that $|x - y| < \delta$, as the irrational numbers are dense in the rationals. This means we would have

$$|f(x) - f(y)| = |0 - 1| = 1 > \frac{1}{2} = \varepsilon,$$

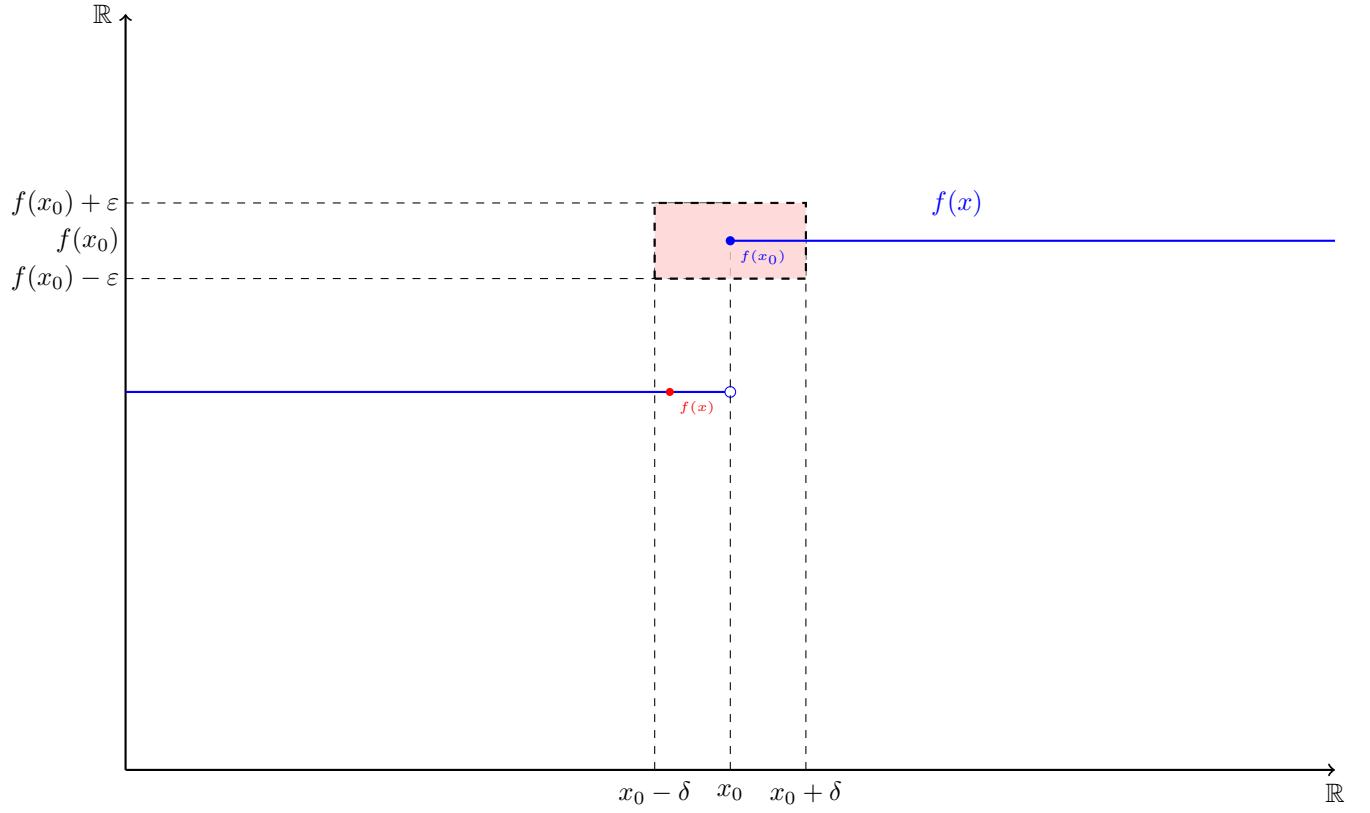


Figure 45: The real function $f : \mathbb{R} \rightarrow \mathbb{R}$ shown here is discontinuous at x_0 . No matter small we take δ to be, we can find at least one $\varepsilon > 0$ and at least one x satisfying $|x - x_0| < \delta$ such that $|f(x) - f(x_0)| \geq \varepsilon$. This is also an example of a jump discontinuity.

for all δ ! A similar argument holds if $x \notin \mathbb{Q}$, as the rational numbers are dense in the reals. Another way to see this is by letting x_n be a sequence of rationals which converge to a real number x .⁶⁴ We have $f(x_n) = 1$ for all x_n , so

$$f\left(\lim_{n \rightarrow \infty} x_n\right) = f(x) = 0 \neq 1 = \lim_{n \rightarrow \infty} 1 = \lim_{n \rightarrow \infty} f(x_n),$$

so f is not continuous at x by Corollary 4.3. We could show a similar result by using the density of the irrationals in the rationals to construct a sequence of irrationals which converge to an arbitrary rational. We also have that each discontinuity is an essential discontinuity.

4.7 Monotonicity

Finally, we'll discuss a special group of functions that are either always weakly increasing or weakly decreasing. These functions will play a very important role in the study of integration, and will have an interesting property related to differentiation.

Definition 4.11. Let f be a real function defined on (a, b) . The function f is *monotonically increasing* on (a, b) if $a < x < y < b$ implies $f(x) \leq f(y)$.

⁶⁴We know such a sequence exists by Corollary 3.1.

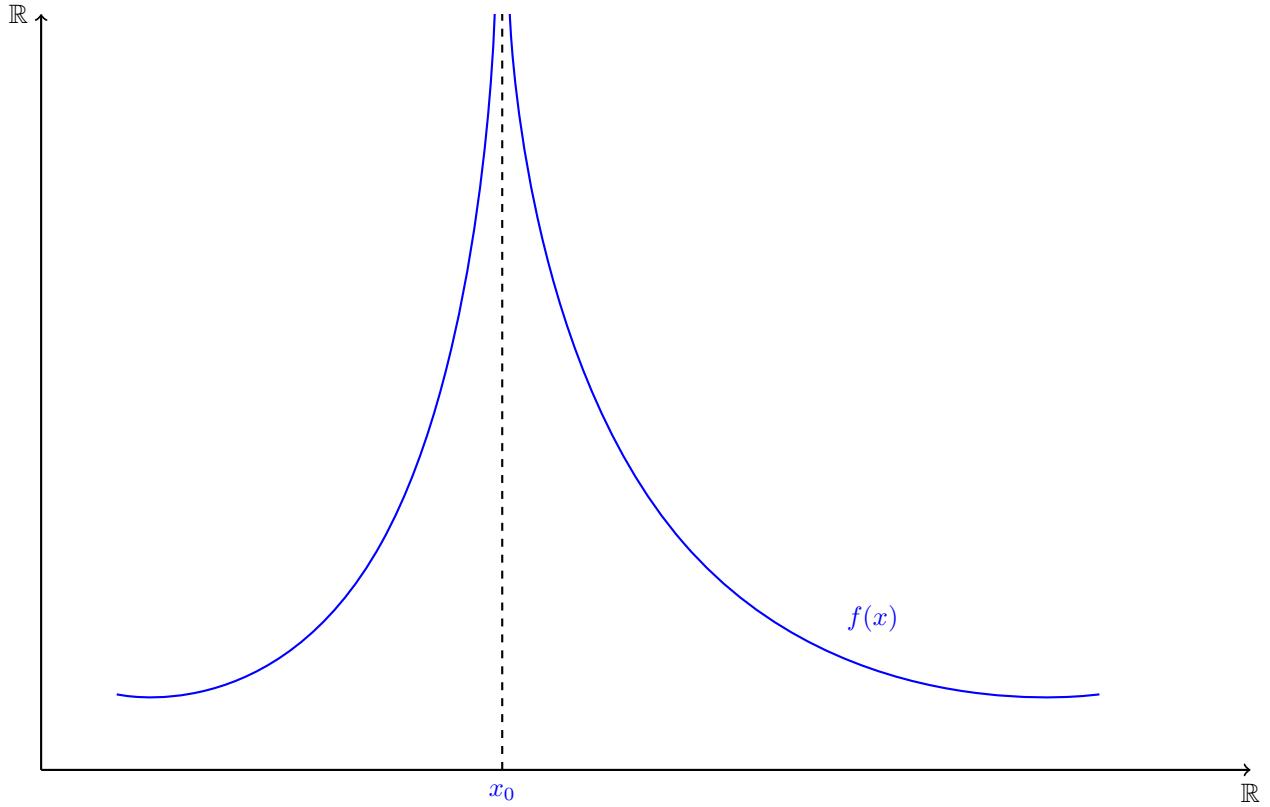


Figure 46: This real function has an essential discontinuity at the point x_0 . In this particular instance, neither $f(x_0-)$, nor $f(x_0+)$ exist.

Definition 4.12. Let f be a real function defined on (a, b) . The function f is *monotonically decreasing* on (a, b) if $a < x < y < b$ implies $f(x) \geq f(y)$.

Example 4.27. The following real functions are monotonically increasing on the entirety of their domains: $f(x) = e^x$, $f(x) = \ln x$, $f(x) = x$, $f(x) = \sqrt{x}$. The following functions are monotonically decreasing on the entirety of their domains: $f(x) = -x$, $f(x) = 1/x$, $f(x) = \arccos(x)$.⁶⁵

If a function is either monotonically increasing or monotonically decreasing, we may just refer to it as *monotonic*. Neither Definition 4.11 nor Definition 4.12 make any mention of continuity. As Figure 47 shows, monotonicity does not imply continuity. In fact, neither of the presented definitions seem to have anything whatsoever to do with continuity. Monotonicity alone is not a strong enough condition for us to make meaningful statements about continuity, *but* it does provide us with information about discontinuities. If a function is monotonic, we can deduce three facts about any discontinuities it may have.

Proposition 4.1. Let f be a monotonically increasing real function on (a, b) . The limits $f(x+)$ and $f(x-)$ exist at every point of (a, b) . Furthermore,

$$\sup_{a < t < x} f(t) = f(x-) \leq f(x) \leq f(x+) = \inf_{x < t < b} f(t).$$

We also will have $f(x+) \leq f(y-)$ for all $a < x < y < b$.

⁶⁵You should avoid writing $\cos^{-1}(x)$, as it is not clear if that either means $\arcsin(x)$ or $1/\cos(x)$.

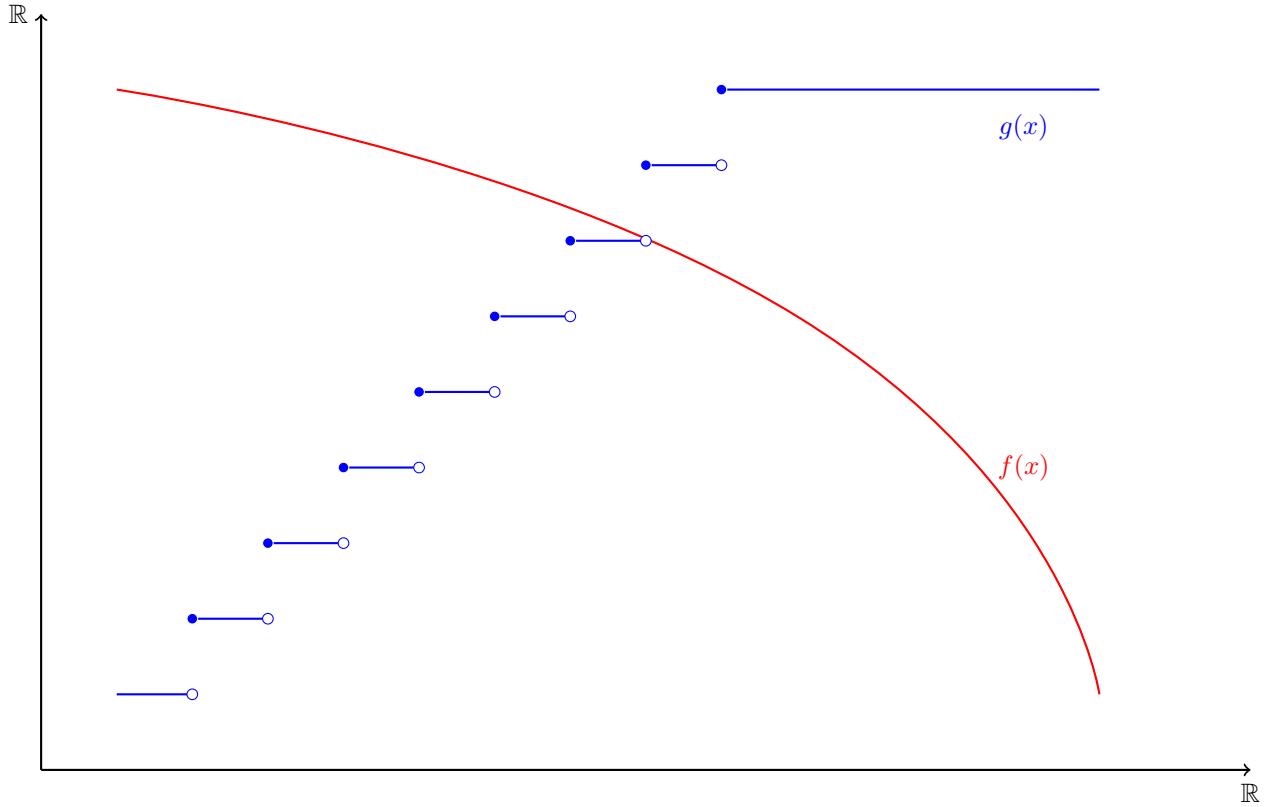


Figure 47: We have two real functions, $f : \mathbb{R} \rightarrow \mathbb{R}$, and $g : \mathbb{R} \rightarrow \mathbb{R}$. The function f is monotonically decreasing, whereas the function $f(x)$ is monotonically increasing. Note that g is not continuous on \mathbb{R} .

An analogous result holds for monotonically decreasing functions, and the proof is similar.

Proof. The range $f((a, x)) \subset \mathbb{R}$ has an upper bound in $f(x)$. By the completeness of \mathbb{R} , there exists some $A = \sup f((a, x))$. By the definition of the supremum, $A \leq f(x)$. We will show that $A = f(x-)$.

Fix $\varepsilon > 0$. By the definition of A as a least-upper-bound, there exists some $\delta > 0$ such that $a < x - \delta < x$ and

$$A - \varepsilon < f(x - \delta) \leq A. \text{⁶⁶}$$

The function f is monotonic, so for all t satisfying $x - \delta < t < x$, we have

$$f(x - \delta) \leq f(t) \leq A.$$

These two inequalities imply that for all $t \in (a, b)$ satisfying $x - \delta < t < x$, we have

$$|f(t) - A| < \varepsilon,$$

so

$$A = \sup f((a, b)) = \sup_{a < t < x} f(t) = f(x-)$$

⁶⁶We found a value $x - \delta$ such that $f(x - \delta)$ is in $f((a, x))$, but just barely below the set's supremum or equal to it. It's so close to that supremum, that $f(x - \delta) \in (A - \varepsilon, A]$ for our arbitrarily small ε .

as desired. To show that $\sup_{x < t < b} f(t) = f(x+)$, we let $B = \inf f((x, b))$, and repeat this process. This gives the inequality presented in the proposition.

If we now let $a < x < y < b$, then we can apply the presented inequality to the point $x \in (a, y)$ and write

$$f(x+) = \inf_{x < t < y} f(t).$$

Applying it to $y \in (x, b)$ gives

$$f(y-) = \sup_{x < t < y} f(t).$$

This infimum and supremum are taken over the same set, so the infimum is less than the supremum, which gives the second desired inequality,

$$f(x+) = \inf_{x < t < y} f(t) \leq \sup_{x < t < y} f(t) = f(y-).$$

□

The first inequality of Proposition 4.1 is shown in Figure 48. Perhaps more useful than Proposition 4.1, is one of its immediate consequences. The sup and inf in question will always exist, implying the existence of both $f(x-)$ and $f(x+)$. In light of definition 4.10, we arrive at a corollary.

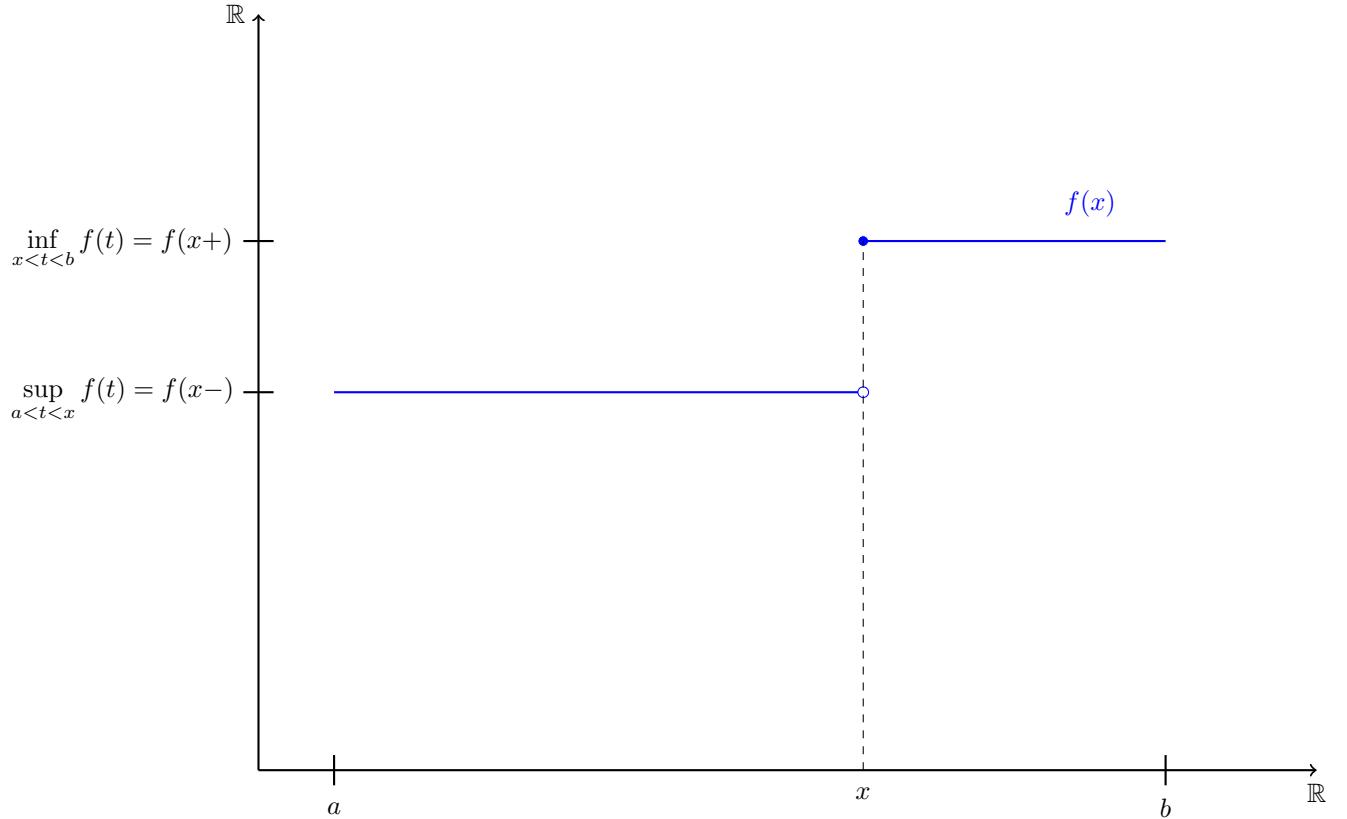


Figure 48: The first inequality of Proposition 4.1 illustrated.

Corollary 4.6. A monotonic function in \mathbb{R} has no essential discontinuous.

This result makes sense if we think about Figure 46. For a function to have an essential discontinuity like this, we need it to increase to infinity from the left or right. This means the function must decrease once we move past the discontinuity. This would contradict monotonicity because the function would increase for certain values of its domain, and then decrease for others.

Perhaps the most interesting result pertaining to discontinuities of monotonic functions, is that they will always form a countable set. We saw functions with an uncountable number of discontinuities (Example 4.26), but we'll see that monotonicity does not allow for this. This fact will become especially useful when we develop the theory of Riemann integration.

Proposition 4.2. Let f be a monotonic function on (a, b) . The set of points (a, b) at which f is discontinuous is at most countable.

Proof. We will prove this result for a monotonically increasing function. Let E be the set of points at which f is discontinuous. To show E is countable, it suffices to show that we can write a bijection from E to some countable set.⁶⁷

The rationals are dense in \mathbb{R} , so we can always find a rational in number in between $f(x-)$ and $f(x+)$. For all $x \in r(x)$ we can associate a rational number $r(x)$ such that

$$f(x-) < r(x) < f(x+).$$

Since $x_1 < x_2$ implies $f(x_1+) \leq f(x_2-)$ by Proposition 4.1, $r(x_1) \neq r(x_2)$ if $x_1 = x_2$. We therefore have a bijection from E to a subset of \mathbb{Q} which is countable.⁶⁸ \square

4.8 Exercises

min and max continuous

examples of open close compact bounded and complete not being preserved in wrong direction

proof of 4.8 with real functions, examples where inf won't work for a general compact X .

Uniform cont and cauchy

left and right limits equal

⁶⁷This works because sets having equal cardinality is transitive. See Proposition 1.5.

⁶⁸We used the density of \mathbb{Q} in \mathbb{R} to make the mapping surjective, and then used Proposition 4.1 to insure injectivity.

5 Differentiation

We now will treat the first of two major topics in calculus – differentiation. For this whole section, we will be dealing with real functions of a single variable. Most of the results should be very familiar, and as such the number of examples will be limited. For the most part, differentiation as you saw it in calculus is rigorously defined. The only things that should be new are the proofs of results, and the emphasis put on certain topics. When people first take calculus, most of it is just learning how to take derivatives. That won't be so important here. The emphasis will instead be placed on the theorems, and how to use them to prove certain results.

5.1 The Definition of a Derivative

We begin by defining the rate of change of a function between two points.

Definition 5.1. Let f be a real function defined on $X \subset \mathbb{R}$. For any $x \in X$ we define the *difference quotient* $\phi(t)$ as

$$\phi(t) = \frac{f(t) - f(x)}{t - x}$$

where $t \in X$ and $t \neq 0$.

For a fixed value x_0 , the difference quotient captures the rate of change of a function between points x_0 and t . If X is an open interval, then ϕ will not be defined at the endpoints of X , as there are excluded from the set. We arrive at the definition of a derivative by letting t tend towards x_0 , thereby let $|x_0 - t|$ going to 0.

Definition 5.2. Let f be a real function defined on X , and x an interior point of X . If $\lim_{t \rightarrow x} \phi(t)$ exists, then we write

$$f'(x) = \lim_{t \rightarrow x} \phi(t),$$

and call $f'(x)$ the *derivative of f at the point x* . The derivative f' is a function defined at every point where $\lim_{t \rightarrow x} \phi(t)$ exists. If f' exists at a point x we say f is *differentiable at x* . If f' exists at every point of $E \subset X$, we say f is *differentiable on E* .

We will sometimes refer to $f'(x)$ as the “instantaneous rate of change” of $f(x)$ at x .

Remark 5.1 (Is the Derivative a Paradox?). As [this video](#) points out, calling the derivative the “instantaneous rate of change” can be interpreted as a paradox. The phrase “rate of change” implies that something actually changes over a time period. The derivative correspond to a single point. Does something change at one given point in time? The phrase “instantaneous rate of change” should be meaningless. In using it in reference to the derivative, we are giving it a definition that it would otherwise not have in the physical world.

Remark 5.2 (An Equivalent Definition). The definition of $f'(x)$ as given in Definition 5.2, may not be the way you first saw the derivative defined in calculus. Instead, you may have seen

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}.$$

This is *perfectly valid*, and the *exact same definition*. We arrive at this equation by letting $t - x = h$, and taking $h \rightarrow 0$. The variable h just represents the “distance” (or displacement) over which we are calculating

the rate of change. The choice of presenting it as an alternative is based off the fact that [Rudin \(1976\)](#) and [Tao \(2016a\)](#) give the definition of $f'(x)$ in terms of the difference quotient. While all the results will hold regardless of what definition we use, sometimes one definition more be more suitable for a proof or example. In all these cases, we could simply take Definition 5.1, let $t - x = h$, and take $h \rightarrow 0$. This step won't be explicitly shown after this, so make sure this sits well. As we'll see later, using the definition where $h \rightarrow 0$ will help us build intuition when using the derivative to approximate functions. Much later on when we consider differentiation with multiple variables, we will also opt to use a definition with $h \rightarrow 0$.

Notation 5.1. There are many different ways to notate the derivative, the two most common being Lagrange's notation of f' , and Leibniz's notation of $\frac{df}{dx}$. We will exclusively use Lagrange's notation for real functions of a single variable. This decision is motivated by the fact that f' does not make reference to the variable used to denote the input of f . We are free to write $f(x)$, $f(t)$, $f(\theta)$, etc. If we use $\frac{df}{dx}$, then we must always write $f(x)$. This is a matter of purely notational flexibility. Another reason to shy away from $\frac{df}{dx}$, is that it encourages one to interpret df and dx as infinitesimally small numbers. Historically, this is how Leibniz interpreted the derivative, but it was never formal in this context. We will make it formal when treating differential forms.

Example 5.1 (The Power Rule). Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined as $f(x) = x^n$. We will show that $f'(x) = nx^{n-1}$ for all $x \in \mathbb{R}$. First, we should point out that

$$(t - x)^n = (t - x) (x^{n-1} + tx^{n-2} + \cdots + t^{n-2}x + t^{n-1}) = (t - x) \sum_{k=0}^{n-1} t^k x^{(n-1)-k}.$$

We will need to use this to rewrite the denominator of the difference quotient.

$$\begin{aligned} f'(x) &= \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} \\ &= \lim_{t \rightarrow x} \frac{t^n - x^n}{t - x} \\ &= \lim_{t \rightarrow x} \frac{(t - x)(x^{n-1} + tx^{n-2} + \cdots + t^{n-2}x + t^{n-1})}{t - x} \\ &= \lim_{t \rightarrow x} (x^{n-1} + tx^{n-2} + \cdots + t^{n-2}x + t^{n-1}) \\ &= x^{n-1} + x \cdot x^{n-2} + \cdots + x^{n-2} \cdot x \\ &= nx^{n-1} \end{aligned}$$

Example 5.2. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined as e^x . What is $f'(x)$?

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{e^{x+h} - e^x}{h} = \lim_{h \rightarrow 0} \frac{e^x e^h - e^x}{h} = e^x \cdot \lim_{h \rightarrow 0} \frac{e^h - 1}{h}.$$

At this point we will let $u = e^h - 1$ and perform a substitution, noting that $h = \ln(u+1)$.⁶⁹ We also have $u \rightarrow 0$ as $h \rightarrow 0$.⁷⁰

$$f'(x) = e^x \cdot \lim_{u \rightarrow 0} \frac{e^h - 1}{h} = e^x \cdot \lim_{u \rightarrow 0} \frac{u}{\ln(u+1)} = e^x \cdot \lim_{u \rightarrow 0} \frac{1}{\frac{1}{u} \cdot \ln(u+1)} = e^x \cdot \lim_{u \rightarrow 0} \frac{1}{\left(\ln(u+1)^{\frac{1}{u}}\right)} = e^x \cdot \frac{1}{\ln\left(\lim_{u \rightarrow 0} (u+1)^{\frac{1}{u}}\right)}$$

⁶⁹I find substitutions like this very unsettling. It feels like another “trick” used in proofs that you would only know if you’ve proved the result before. I also could be very alone in this regard, and just don’t think “maybe I should try substitution” as much as I should.

⁷⁰The limit of $u = e^h - 1$ as $h \rightarrow 0$ is $e^0 - 1 = 0$ because e^x is continuous. This gives $u \rightarrow 0$ as $h \rightarrow 0$.

Why are we able to pass the limit into the natural log?⁷¹ Finally note that

$$\lim_{u \rightarrow 0} (u+1)^{1/u} = \lim_{n \rightarrow \infty} (1+1/n)^{1/n} = e,$$

which we first took as a fact in Example 3.7.

$$f'(x) = e^x \cdot \frac{1}{\ln\left(\lim_{u \rightarrow 0} (u+1)^{\frac{1}{u}}\right)} = e^x \cdot \frac{1}{\ln(e)} = e^x \cdot \frac{1}{1} = e^x$$

We have that the derivative of e^x is e^x .

Theorem 5.1 (Differentiability implies Continuity). Let f be defined on X . If f is differentiable at $x \in X$, then f is continuous at x .

Proof. We can use Theorem 4.3, and show $\lim_{t \rightarrow x} f(t) = f(x)$ if f is differentiable at x . We have

$$\begin{aligned} f(t) - f(x) &= \frac{f(t) - f(x)}{t - x}(t - x) \\ \lim_{t \rightarrow x} [f(t) - f(x)] &= \lim_{t \rightarrow x} \phi(t)(t - x) \\ \lim_{t \rightarrow x} f(t) - f(x) &= f'(x) \cdot 0 \\ &= 0. \end{aligned}$$

If $\lim_{t \rightarrow x} f(t) - f(x) = 0$, then $\lim_{t \rightarrow x} f(t) = f(x)$. □

Where would this proof go wrong if we did not know f is differentiable at x ?⁷² It is quite easy to come up with examples where the converse of Theorem 5.1 fails.

Example 5.3. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be defined as $f(x) = |x|$. This function is continuous at 0 (and all of \mathbb{R}), but is not differentiable at 0. We have

$$\phi(t) = \frac{f(t) - f(0)}{t - 0} = \frac{|t|}{t}.$$

If we take the left-hand limit and right-hand limit of the difference quotient, we will find they do not agree.

$$\begin{aligned} \lim_{t \rightarrow 0^+} \phi(t) &= \lim_{t \rightarrow 0^+} \frac{|t|}{t} = \lim_{t \rightarrow 0^+} \frac{t}{t} = \lim_{t \rightarrow 0^+} 1 = 1 \\ \lim_{t \rightarrow 0^-} \phi(t) &= \lim_{t \rightarrow 0^-} \frac{|t|}{t} = \lim_{t \rightarrow 0^-} \frac{-t}{t} = \lim_{t \rightarrow 0^-} -1 = -1 \end{aligned}$$

These two limits agreeing is both a necessary and sufficient condition for the limit existing. We therefore have that $\lim_{t \rightarrow 0} \phi(t)$ does not exist, so f' is undefined at 0. Note that f is still differentiable on $\mathbb{R} \setminus \{0\}$.

5.2 Familiar Properties of the Derivative

We will now prove the rules of differentiation that are presented in a calculus course. After proving the most basic properties, we will introduce the Chain Rule, and a basic version of the Inverse Function Theorem, two results that are not immediately obvious.

Theorem 5.2. Suppose f and g are defined on X , and are differentiable at $x \in X$.

⁷¹Because the natural log is continuous, and you can bring limits into continuous functions, as continuity preserves limits.

⁷²If $\lim_{t \rightarrow x} \phi(t)$ did not exist, then there would have been no way to calculate $(\lim_{t \rightarrow x} \phi(t)) \cdot 0$.

1. If $f(x) = c$ for some constant $c \in \mathbb{R}$, then $f'(x) = 0$.

2. $(cf)'(x) = cf'(x)$.

3. (Sum Rule) $(f + g)'(x) = f'(x) + g'(x)$.

4. (Product Rule) $(fg)'(x) = f'(x)g(x) + g'(x)f(x)$.

5. (Quotient Rule) $\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2}$.

Proof. The functions f and g are differentiable at x . By Theorem 5.1, these functions are continuous at x , so we will have $\lim_{t \rightarrow x} g(t) = g(x)$ and $\lim_{t \rightarrow x} f(t) = f(x)$ by Theorem 4.3. We will use these equalities in the proof of the Product Rule and the Quotient Rule.

1.

$$\lim_{t \rightarrow x} \phi(t) = \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} = \lim_{t \rightarrow x} \frac{c - c}{t - x} = \lim_{t \rightarrow x} \frac{0}{t - x} = \lim_{t \rightarrow x} 0 = 0$$

2.

$$\lim_{t \rightarrow x} \phi(t) = \lim_{t \rightarrow x} \frac{cf(t) - cf(x)}{t - x} = \lim_{t \rightarrow x} c \cdot \frac{f(t) - f(x)}{t - x} = c \cdot \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} = cf'(x)$$

3.

$$\begin{aligned} \lim_{t \rightarrow x} \phi(t) &= \lim_{t \rightarrow x} \frac{(f + g)(t) - (f + g)(x)}{t - x} \\ &= \lim_{t \rightarrow x} \frac{[f(t) + g(t)] - [f(x) + g(x)]}{t - x} \\ &= \lim_{t \rightarrow x} \frac{f(t) - f(x) + g(t) - g(x)}{t - x} \\ &= \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} + \lim_{t \rightarrow x} \frac{g(t) - g(x)}{t - x} \\ &= f'(x) + g'(x) \end{aligned}$$

4.

$$\begin{aligned} \lim_{t \rightarrow x} \phi(t) &= \lim_{t \rightarrow x} \frac{(fg)(t) - (fg)(x)}{t - x} \\ &= \lim_{t \rightarrow x} \frac{f(t)g(t) - f(x)g(x) + 0}{t - x} \\ &= \lim_{t \rightarrow x} \frac{f(t)g(t) - f(x)g(x) + [f(t)g(x) - f(t)g(x)]}{t - x} \\ &= \lim_{t \rightarrow x} \frac{g(x)[f(t) - f(x)] + f(t)[g(t) - g(x)]}{t - x} \\ &= g(x) \cdot \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} + \lim_{t \rightarrow x} f(t) \cdot \frac{g(t) - g(x)}{t - x} \\ &= g(x)f'(x) + \lim_{t \rightarrow x} f(t) \cdot \lim_{t \rightarrow x} \frac{g(t) - g(x)}{t - x} \\ &= f'(x)g(x) + g'(x)f(x) \end{aligned}$$

5.

$$\begin{aligned}
\lim_{t \rightarrow x} \phi(t) &= \lim_{t \rightarrow x} \frac{(f/g)(t) - (f/g)(x)}{t - x} \\
&= \lim_{t \rightarrow x} \frac{\frac{f(t)}{g(t)} - \frac{f(x)}{g(x)} + 0}{t - x} \\
&= \lim_{t \rightarrow x} \frac{\frac{f(t)}{g(t)} - \frac{f(x)}{g(x)} + \left(\frac{f(x)}{g(t)} - \frac{f(x)}{g(t)} \right)}{t - x} \\
&= \lim_{t \rightarrow x} \frac{\frac{f(t)}{g(t)} - \frac{f(x)}{g(t)} + \frac{f(x)}{g(t)} - \frac{f(x)}{g(x)}}{t - x} \\
&= \lim_{t \rightarrow x} \frac{\frac{1}{g(x)} \left(\frac{f(t)g(x)}{g(t)} - \frac{f(x)g(x)}{g(t)} \right) - \left(\frac{f(x)g(t)}{g(x)g(t)} - \frac{f(x)g(x)}{g(t)g(x)} \right)}{t - x} \\
&= \lim_{t \rightarrow x} \frac{\frac{g(x)}{g(x)g(t)} (f(t) - f(x)) - \frac{f(x)}{g(x)g(t)} (g(t) - g(x))}{t - x} \\
&= \lim_{t \rightarrow x} \frac{1}{g(x)g(t)} \left(g(x) \cdot \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} - f(x) \cdot \lim_{t \rightarrow x} \frac{g(t) - g(x)}{t - x} \right) \\
&= \frac{1}{g(x)g(x)} [g(x)f'(x) - f(x)g'(x)] \\
&= \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2}
\end{aligned}$$

□

Example 5.4 (Linearity of Derivatives). If we combine the Sum Rule and Part 2 of Theorem 5.2, we have that the operation of differentiation is linear. If f and g are defined on $[a, b]$ and are differentiable at $x \in [a, b]$, then for constants $c, d \in \mathbb{R}$ we have

$$(cf + dg)'(x) = cf'(x) + dg'(x).$$

Example 5.5 (Difference Rule). Suppose f and g are defined on $[a, b]$, and are differentiable at $x \in [a, b]$. We can derive⁷³ the Difference Rule by using the linearity of differentiation.

$$(f - g)'(x) = (f + (-1)g)'(x) = f'(x) + (-1)g'(x) = f'(x) - g'(x)$$

Example 5.6 (Polynomials). If $\mathcal{P}(\mathbb{R})$ is the set of all polynomials with real coefficients, then every element of $\mathcal{P}(\mathbb{R})$ is differentiable on all of \mathbb{R} . We can show this using the Power Rule, Sum Rule, and Part 2 of Theorem 5.2. We can write $p(x) \in \mathcal{P}(\mathbb{R})$ as

$$p(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n = \sum_{k=0}^n a_kx^k$$

for $a_k \in \mathbb{R}$ for all k . We have

$$p'(x) = a_1 + 2a_2x + \cdots + (n-1)a_{n-1}x^{n-2} + na_nx^{n-1} = \sum_{k=1}^n ka_kx^{k-1}.$$

This is define for all $x \in \mathbb{R}$. The ease at which we can differentiate polynomials is one of the reasons we like working with them.

⁷³Pun intended.

One of the benefits of composition preserving the continuity of two functions (Theorem 4.4) is that it means most function we work with are continuous. If you were tasked with writing down some random function, it could most likely be the result of a composition of simpler functions. The ubiquity of functions composed via other functions makes it useful to know how to differentiate such functions. This is where the Chain Rule comes in. Before formally stating and proving the Chain Rule, it's worth building a *non-technical*⁷⁴ intuition as to why it works.

Suppose you have a real function $g(y)$, and you want to find the rate of change with. We know that this rate is given as $g'(y)$. Now let there be a second function $f(x)$. You're now challenged with finding the rate of change of g with respect to x , after taking $y = f(x)$ to be an input of g . That is, we need to find $h'(x)$ for $h(x) = g(f(x))$. We have two pieces of information: $g'(y)$ (how g changes with respect to its input), and $f'(x)$ (how f changes with respect to its input). To measure the change of g at $f(x)$ with respect to x , we need to somehow convert a measure of change in $y = f(x)$ to a change in x . This can be achieved by using the change in y with respect to x as an exchange rate. If we think of this as a problem related to converting between two measures of change, then it seems the logical thing to do would be multiply g 's change with respect to $y = f(x)$ by the conversion rate of $y = f(x)$'s change with respect to x .

$$g'(y) \cdot f'(x) = \frac{\text{change in } g}{\text{change with respect to } y = f(x)} \times \frac{\text{change in } y = f(x)}{\text{change with respect to } x} = \frac{\text{change in } g}{\text{change with respect to } x}$$

Writing everything in terms of x gives us $g'(f(x))f'(x)$.

Remark 5.3 (I Broke the Rules). Okay so this explanation of the Chain Rule is kind of wrong. In Notation 5.1, I said that I would be avoiding the notation of $\frac{dy}{dx}$, as it encourages people to think of dy and dx as their own numbers which form a fraction. This is not what derivatives are, but I just treated them as fractions when explaining the Chain Rule. In fact, had I used Leibniz's notation, the Chain Rule becomes very aesthetically pleasing and easy to remember:

$$\frac{dg}{dx} = \frac{dg}{dy} \frac{dy}{dx}.$$

Contrary to what it looks like, we aren't multiplying fractions when we write this. Treating dx and dy like their own numbers can lead to problems and at this point lacks the rigor that analysis aims to achieve. Remark 10.1.17 in [Tao \(2016a\)](#) makes similar points about this issue.

Theorem 5.3 (The Chain Rule). Let X and Y be subsets of \mathbb{R} . If $f : X \rightarrow \mathbb{R}$ is continuous on X , $f'(x)$ exists at some point $x \in X$, and $g : Y \rightarrow \mathbb{R}$ is differentiable at $f(x)$. If we define $h(t) = g(f(t))$, then h is differentiable at x , and

$$h'(x) = g'(f(x))f'(x).$$

Proof. Let $y = f(x)$. The functions f and g are differentiable at x and y respectively.

$$\begin{aligned} f'(x) &= \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} \\ g'(y) &= \lim_{s \rightarrow y} \frac{g(s) - g(y)}{s - y} \end{aligned}$$

⁷⁴I'll explain why my explanation is technically wrong afterwards.

Define $u(t)$ and $v(s)$ as

$$u(t) = \begin{cases} \frac{f(t)-f(x)}{t-x} - f'(x) & \text{if } t \neq x \\ 0 & \text{if } t = x \end{cases},$$

$$v(s) = \begin{cases} \frac{g(s)-g(y)}{s-y} - g'(s) & \text{if } s \neq y \\ 0 & \text{if } s = y \end{cases},$$

for $t \in X$ and $s \in Y$. We have $\lim_{t \rightarrow x} u(t) = 0$ and $\lim_{s \rightarrow y} v(s) = 0$. These functions allow to rewrite the derivatives of g and f as

$$f(t) - f(x) = (t - x)[f'(x) + u(t)], \quad (7)$$

$$g(s) - g(y) = (s - y)[g'(y) + v(s)], \quad (8)$$

as we take the limits $t \rightarrow x$ and $s \rightarrow y$.⁷⁵ If we let $f(t) = s$ in (8), we can use (7) and (8) to write

$$\begin{aligned} h(t) - h(x) &= g(f(t)) - g(f(x)) \\ &= [f(t) - f(x)][g'(y) + v(s)] \\ &= (t - x)[f'(x) + u(t)][g'(y) + v(f(t))] \\ \frac{h(t) - h(x)}{(t - x)} &= [f'(x) + u(t)][g'(y) + v(f(t))] \end{aligned}$$

for $t \neq x$. If we let $t \rightarrow x$, then

$$\begin{aligned} h'(x) &= \lim_{t \rightarrow x} \frac{h(t) - h(x)}{(t - x)} \\ &= \lim_{t \rightarrow x} [f'(x) + u(t)][g'(y) + v(f(t))] \\ &= \left(f'(x) + \lim_{t \rightarrow x} u(t) \right) \left(g'(y) + \lim_{t \rightarrow x} v(f(t)) \right) \\ &= (f'(x))(g'(y) + v(y)) \\ &= f'(x)g'(y) \\ &= g'(f(x))f'(x). \end{aligned}$$

Note that to conclude $\lim_{t \rightarrow x} v(f(t)) = v(y)$, it must be the case that $\lim_{t \rightarrow x} f(t) = f(x) = y$. This follows by the assumed continuity of f on X . \square

Our last rule concerning the calculation of derivatives pertains to the inverse of a function. If we let X and Y be subsets of \mathbb{R} , we may want to know the derivative of $f^{-1} : Y \rightarrow X$ at a point $y = f(x)$, but only know the derivative of $f : X \rightarrow Y$ at x . If we know that f^{-1} is differentiable, then this is a straightforward application of the chain rule and composition of f^{-1} and f .

$$\begin{aligned} (f^{-1} \circ f)'(x) &= (f^{-1})'(f(x))f'(x) \\ 1 &= (f^{-1})'(y)f'(x) \\ (f^{-1})'(y) &= \frac{1}{f'(x)} \end{aligned}$$

⁷⁵Equation (78) is simply $f(t) - f(x) = (t - x) \left(f'(x) + \frac{f(t) - f(x)}{t - x} - f'(x) \right)$ when $t \neq x$. When we let $t \rightarrow x$ we have $u(t) \rightarrow 0$, and we arrive at $f(t) - f(x) = (t - x)f'(x)$ as $t \rightarrow x$. In other words, $f'(x) = \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x}$, which is true. Equation (7) is therefore perfectly valid. The same argument works for Equation (7).

There is one shortcoming in this approach. In order to use the Chain Rule, we needed to assume that f^{-1} was differentiable. If we do not know beforehand that f^{-1} is invertible, then perhaps we cannot conclude $(f^{-1})'(y) = \frac{1}{f'(x)}$. As it turns out, we still can by The Inverse Function Theorem, which says that knowing f^{-1} is continuous suffices to conclude $(f^{-1})'(y) = \frac{1}{f'(x)}$. This result is rarely given this name in the context of a single variable, and in fact it may not show up in an analysis course to begin with. It's not very special for functions in \mathbb{R} . A far more general version of the Inverse Function Theorem will hold with functions in several variables. The introduction here is just to give some early exposure to the idea of the theorem (which I'm calling a proposition here as an allusion to it not being the full-fledged Inverse Function Theorem).

Proposition 5.1 (The Inverse Function Theorem). Let X and Y be subsets of \mathbb{R} , and $f : X \rightarrow Y$ be an invertible function with inverse $f^{-1} : Y \rightarrow X$. Suppose $x \in X$ and $y \in Y$ such that $f(x) = y$. If f is differentiable at x_0 , f^{-1} is continuous at y_0 , and $f'(x_0) \neq 0$, then f^{-1} is differentiable and

$$(f^{-1})'(y) = \frac{1}{f'(x)}.$$

Proof. We need to show that

$$\lim_{s \rightarrow y} \frac{f^{-1}(s) - f^{-1}(y)}{s - y} = \frac{1}{f'(x)}.$$

By the assumed continuity of f^{-1} , we can show

$$\lim_{n \rightarrow \infty} \frac{f^{-1}(s_n) - f^{-1}(y)}{s_n - y} = \frac{1}{f'(x)},$$

for a sequence $\{s_n\}$ in Y which converges to s . Let $t_n = f^{-1}(s_n)$ be a sequence in X . By continuity, we have

$$\lim_{n \rightarrow \infty} t_n = \lim_{n \rightarrow \infty} f^{-1}(s_n) = f^{-1}\left(\lim_{n \rightarrow \infty} s_n\right) = f^{-1}(s) = t.$$

Because f is differentiable (and therefore continuous) at x , we have

$$\lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} = \lim_{n \rightarrow \infty} \frac{f(t_n) - f(x)}{t_n - x} = f'(x).$$

This difference quotient is not zero since $t_n \neq x$, and $f(t_n) \neq f(x)$ (f is a bijection after all). We can use the limit laws of Theorem 3.2 to invert the difference quotient and arrive at

$$\lim_{n \rightarrow \infty} \frac{t_n - x}{f(t_n) - f(x)} = \frac{1}{f'(x)}.$$

But since $t_n = f^{-1}(s_n)$ and $x = f^{-1}(y)$, we have $\lim_{n \rightarrow \infty} \frac{f^{-1}(s_n) - f^{-1}(y)}{s_n - y} = \lim_{s \rightarrow y} \frac{f^{-1}(s) - f^{-1}(y)}{s - y} = \frac{1}{f'(x)}$. \square

5.3 Local Extrema

The derivative gives us a useful tool for finding *local* minima and maxima of a function.

Definition 5.3. Let X be a subset of \mathbb{R} . The function $f : X \rightarrow Y$ has a *local maximum* at a point $p \in X$ if there exists $\delta > 0$ such that $f(x) \leq f(p)$ for all $x \in X$ with $d(p, x) < \delta$.

Definition 5.4. Let X be a subset of \mathbb{R} . The function $f : X \rightarrow Y$ has a *local minimum* at a point $p \in X$ if there exists $\delta > 0$ such that $f(x) \geq f(p)$ for all $x \in X$ with $d(p, x) < \delta$.

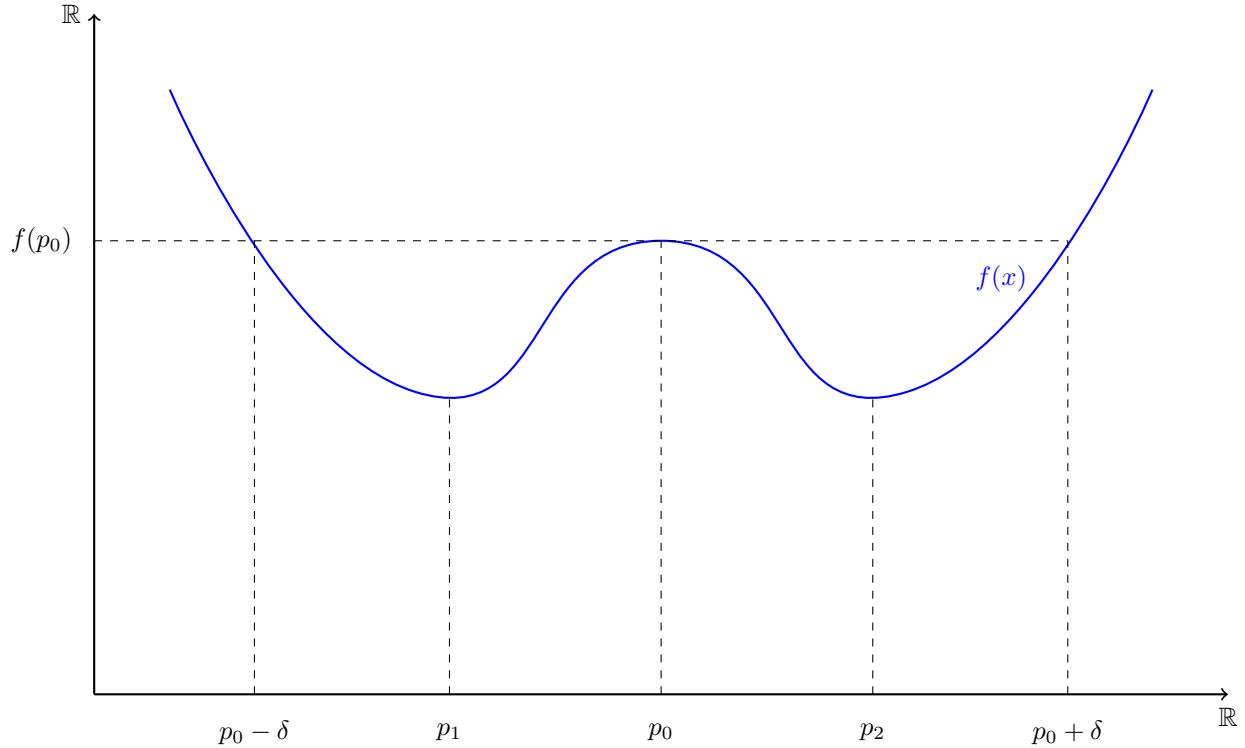


Figure 49: The function has a local maximum at p_0 , and local minima at p_1 and p_2 . The points $p_0 - \delta$ and $p_0 + \delta$ correspond to the endpoints of the interval $(p_0 - \delta, p_0 + \delta)$. We have $f(x) \leq f(p_0)$ for all $x \in (p_0 - \delta, p_0 + \delta)$, as prescribed in Definition 5.3.

We use the word local to emphasize the fact that local extrema need not be *the* maximum or *the* minimum on X . We see this in Figure 49, f has a local maximum at p_0 , despite not having a *global* maximum at p_0 . All we care about is that there is some small open ball around p_0 over f attains a maximum at p_0 . This open ball is the set of all $x \in X$ such that $d(p_0, x) < \delta$ for $\delta > 0$, which can also be written as $B_\delta(p_0) \subset X$. We can relate the local extrema of a function to its derivative. Note that the proposition will present is a necessary condition for local extrema, but not a sufficient condition. Our main use for this will be proving other theorems. In many applied settings which rely on optimization, this result is assigned more significance.

Proposition 5.2 (Fermat's Theorem for Extrema). Let $X = [a, b]$ be a subset set of \mathbb{R} . If $f : X \rightarrow \mathbb{R}$ has a local extrema at a point $p \in (a, b)$, and if $f'(p)$ exists, then $f'(p) = 0$.

Proof. We will prove the case for local maxima. Let $p \in (a, b)$ be a local maximum. There exists some δ such that $f(p) \geq f(x)$ for all x satisfying

$$x < p - \delta < x < p + \delta < b.$$

If $p - \delta < t < p$, then

$$\frac{f(t) - f(p)}{t - p} \geq 0$$

as $f(p) \geq f(t)$, and $p > t$. This means

$$\lim_{t \rightarrow p} \frac{f(t) - f(p)}{t - p} = f'(p) \geq 0.$$

If instead $p < t < p + \delta$, then

$$\frac{f(t) - f(p)}{t - p} \leq 0$$

as $f(p) \geq f(t)$, and $p < t$. This means

$$\lim_{t \rightarrow p} \frac{f(t) - f(p)}{t - p} = f'(p) \leq 0.$$

If $f'(p) \leq 0$ and $f'(p) \geq 0$, then $f'(p) = 0$. □

As the next three examples show, there are several drawbacks to Fermat's Theorem.

Example 5.7 (Saddle Point). Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be the function $f(x) = x^3$. We have $f'(0) = 0$, but the point 0 is neither a local minimum nor a local maximum. This follows from the fact that Lemma 5.1 is a necessary condition, but not a sufficient condition.

Example 5.8 (Endpoints). Let $f : [0, 1] \rightarrow \mathbb{R}$ be the function $f(x) = x$. This function has a local maximum at 1 and local minimum at 0. Lemma 5.1 does not hold in this case, because it assumes that $p \in (a, b)$.

Example 5.9 (Not Differentiable). Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be the function $f(x) = |x|$. This function has a local minimum at 0, but $f'(0)$ is undefined. Lemma 5.1 does not hold in this case, as it assumes that $f'(p)$ exists.

5.4 Mean Value Theorems

As far as analysis is concerned, the Mean Value Theorem is the most important result involving differentiation. When discussing the Intermediate Value Theorem, it was noted that it gave us a way to “convert” information about f to information about the domain of f . A similar metaphor holds for the Mean Value Theorem, as it allows us to gather some information about f' if we know information about f (or vice-versa). If we know that f is differentiable on some interval $[a, b]$, then the Mean Value Theorem tells us the value of $f'(c)$ for $c \in [a, b]$, namely it is

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

We will present three versions of the Mean Value Theorem, each more general than the last.

Theorem 5.4 (Rolle's Theorem). Let $a < b$, and let $f : [a, b] \rightarrow R$ be a continuous function which is differentiable on (a, b) . Suppose also that $f(a) = f(b)$. Then there exists an $x \in (a, b)$ such that $f'(x) = 0$.

Proof. We have that f is continuous on $[a, b]$. By the Extreme Value Theorem (Theorem 4.7), f attains a maximum M at some point $x \in [a, b]$, and a minimum m at some point $y \in [a, b]$.

Case 1. If $x, y \in \{a, b\}$,⁷⁶ then f is constant on $[a, b]$ because $f(a) = f(b)$, so $f'(z) = 0$ for all $z \in (a, b)$.

Case 2. Suppose x is not an endpoint of $[a, b]$. We have a local maximum of f at x , so by Fermat's Theorem (Proposition 5.2), $f'(x) = 0$.

Case 3. Suppose y is not an endpoint of $[a, b]$. We have a local minimum of f at y , so by Fermat's Theorem (Proposition 5.2), $f'(y) = 0$. □

⁷⁶That is, they are the endpoints of the interval $[a, b]$.

Corollary 5.1 (Mean Value Theorem). Let $a < b$, and let $f : [a, b] \rightarrow \mathbb{R}$ be a continuous function which is differentiable on (a, b) . Then there exists an $c \in (a, b)$ such that

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

Proof. Define a new function as

$$h(x) = f(x) - \frac{f(b) - f(a)}{b - a} \cdot x.$$

The function h is continuous on $[a, b]$, differentiable on (a, b) ,⁷⁷ and was defined so $h(a) = h(b)$. By Rolle's Theorem, there exists a $c \in (a, b)$ such that $h'(c) = 0$.

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

□

A geometric interpretation of the Mean Value Theorem is found in Figure 50. The Mean Value Theorem

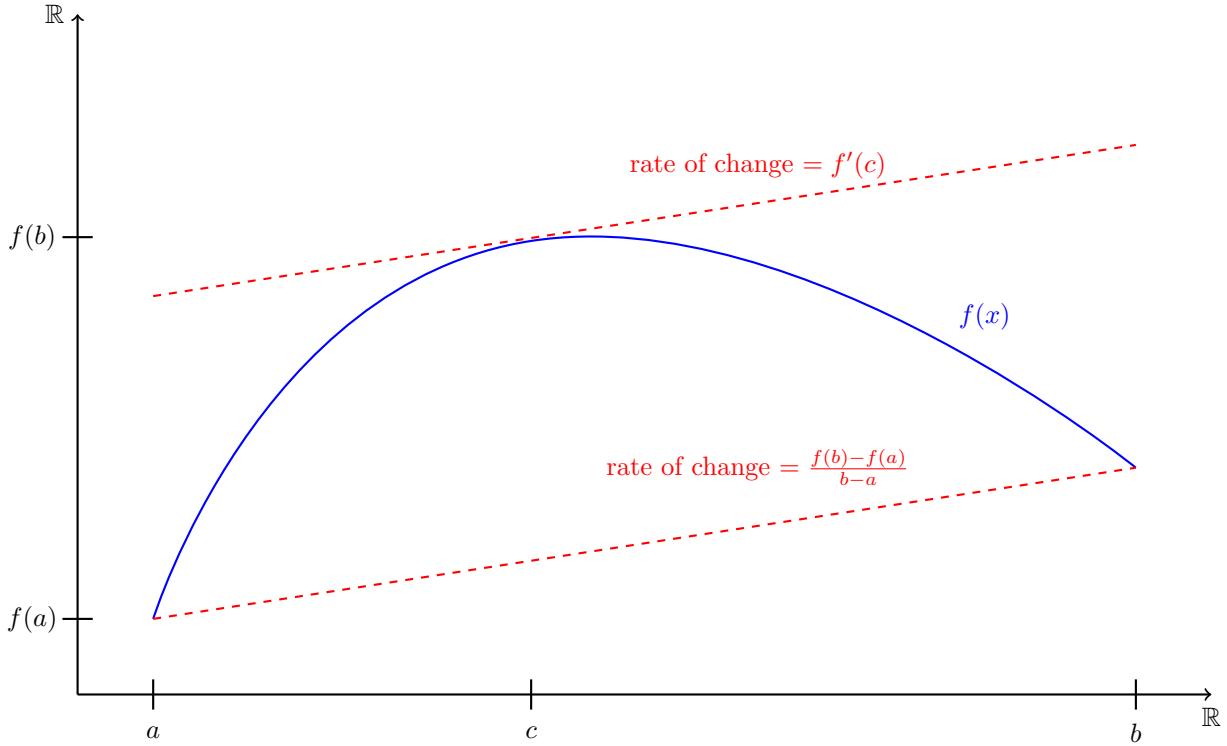


Figure 50: The Mean Value Theorem.

can be extended to the case of parametric curves.

Proposition 5.3 (Cauchy's Mean Value Theorem). Let $a < b$, and let $f : [a, b] \rightarrow \mathbb{R}$ and $g : [a, b] \rightarrow \mathbb{R}$ be continuous functions which are differentiable on (a, b) . Then there exists an $c \in (a, b)$ such that

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

⁷⁷The continuity and differentiability of h follow immediately from f .

Proof. Like we did while proving the Mean Value Theorem, define a function

$$h(x) = f(x) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot g(x).$$

If we apply Rolle's Theorem, the result follows. \square

Figure 51 illustrates Cauchy's Mean Value Theorem.

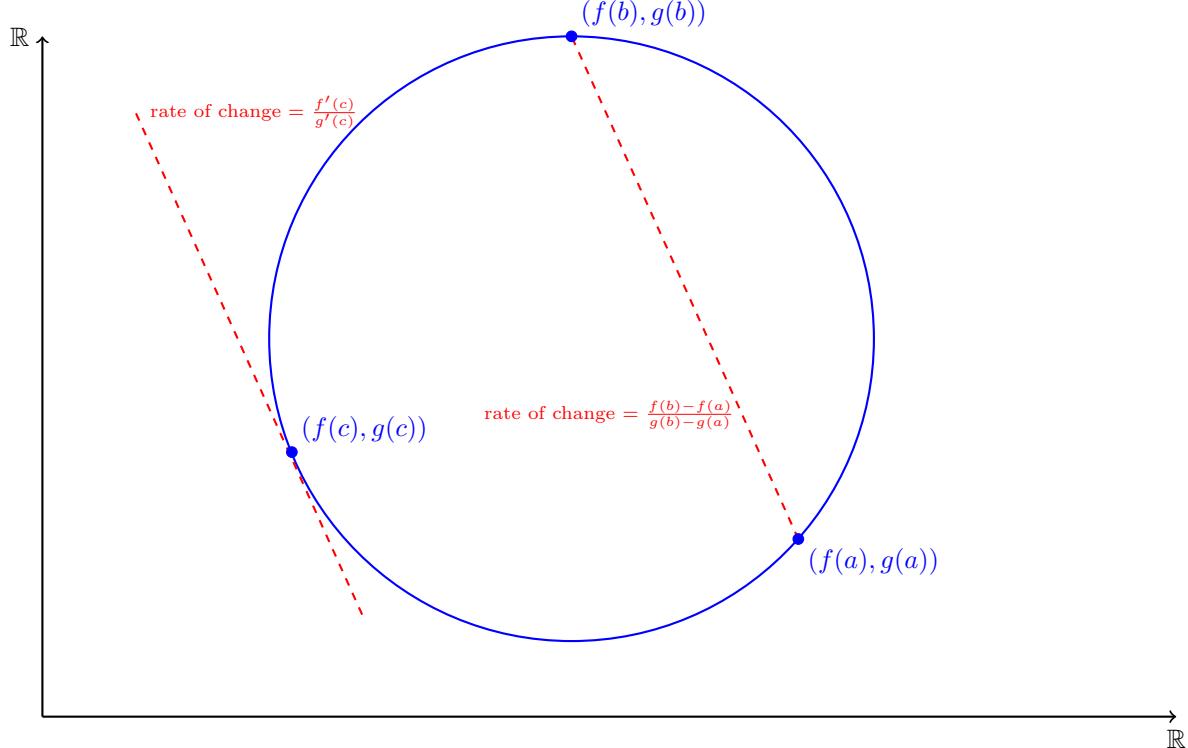


Figure 51: Cauchy's Mean Value Theorem. Note that the parameterization begins at $(f(a), g(a))$ and then is orientated clockwise.

Example 5.10. Let $f(t) = \cos(t)$ and $g(t) = \sin t$. If we let $a = 0$ and $b = \pi/2$. Then we can find a value of $c \in (0, \pi/2)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{-\sin c}{\cos c} = \frac{0 - 1}{1 - 0} = -1.$$

You can verify that $c = \pi/4$ works.

5.5 L'Hôpital's Rule

Now we will prove a classic result from calculus that allows us to calculate a limit using derivatives.

Proposition 5.4 (L'Hôpital's Rule). Suppose f and g are real and differentiable in (a, b) , and $g'(x) \neq 0$ for all $x \in (a, b)$, where $-\infty \leq a < b \leq \infty$. Suppose we have

$$\lim_{x \rightarrow a} \frac{f'(x)}{g'(x)} = A.$$

If $\lim_{x \rightarrow a} f(x) = 0$ and $\lim_{x \rightarrow a} g(x) = 0$, or if $\lim_{x \rightarrow a} g(x) = \infty$, then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = A.$$

Proof. FINISH □

5.6 Higher Order Derivatives

The derivative f' is a function in its own right. This means we could not only differentiate f' itself, but differentiate the resulting function. This idea gives rise to derivatives of higher order.

Definition 5.5. If f has a derivative f' on an interval, and if f' is itself differentiable, we write $(f')' = f''$, and call f'' the *second derivative of f*. Continuing in this manner gives

$$f, f', f'', f^{(3)}, \dots, f^{(n)},$$

where $f^{(n)}$ is the *nth derivative of f*.

Remark 5.4 (Smoothness). For $f^{(n)}(x)$ to exist at a point x , we need to be able to take the limit

$$\lim_{t \rightarrow x} \frac{f^{(n-1)}(t) - f^{(n-1)}(x)}{t - x}.$$

This means that not only is a requirement that $f^{(n-1)}(t)$ is differentiable at x , but also we need $f^{(n-1)}(t)$ to be defined in some open ball around x . What does it mean for $f^{(n-1)}(t)$ to exist in an open ball around x ? It would mean that $f^{(n-2)}$ is differentiable in that open ball. This reasoning follows for all $f^{(n-2)}, \dots, f$. For this reason, the highest order derivative you are able to take of a function f is often associated with how “smooth” it is.

A continuous function is “nice”, but it doesn’t need to be smooth.⁷⁸ A function that is differentiable is smooth. A function that is twice differentiable is even smoother, as it means that f' is differentiable, and therefore continuous! To be formal we say that f is a *smooth function* if it has continuous derivatives up to some desired order. Having a continuous derivative of order $n - 1$, is implied by being n times differentiable. In many applied setting this becomes important. For example, you may see theorems assume that a certain function has a continuous second derivative.⁷⁹ The smoothest functions are those which are infinitely differentiable.

Example 5.11. Every real polynomial $p(x)$ is infinitely differentiable. Let

$$p(x) = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n = \sum_{k=0}^n a_kx^k$$

⁷⁸Think of $|x|$ at the points $x = 0$

⁷⁹For example, in statistics and econometrics there is a method of estimation known as maximum likelihood estimation. This method is only “efficient” if we assume that a probability distribution is three times differentiable, i.e it’s second derivative is continuous.

for $a_k \in \mathbb{R}$ be a polynomial of degree n . We have

$$\begin{aligned}
p'(x) &= a_1 + 2a_2x + \cdots + (n-1)a_{n-1}x^{n-2} + na_nx^{n-1} \\
p''(x) &= 2a_2 + \cdots + (n-1)(n-2)a_{n-1}x^{n-3} + n(n-1)a_nx^{n-2} \\
&\vdots \\
p^{(n-1)}(x) &= n!a_nx \\
p^{(n)}(x) &= n!a_n \\
p^{(n+1)}(x) &= 0 \\
&\vdots
\end{aligned}$$

The polynomial is still infinitely differentiable even if we keep getting zero, as the zero function is differentiable.

Example 5.12. The functions $f(x) = \sin x$ and $g(x) = \cos x$ are infinitely differentiable, and each of their derivatives are given by

$$f^{(n)}(x) = \begin{cases} \cos x & \text{if } n = 1, 5, 9, \dots \\ -\sin x & \text{if } n = 2, 6, 10, \dots \\ -\cos x & \text{if } n = 3, 7, 11, \dots \\ \sin x & \text{if } n = 4, 8, 12, \dots \end{cases} \quad g^{(n)}(x) = \begin{cases} -\sin x & \text{if } n = 1, 5, 9, \dots \\ -\cos x & \text{if } n = 2, 6, 10, \dots \\ -\sin x & \text{if } n = 3, 7, 11, \dots \\ \cos x & \text{if } n = 4, 8, 12, \dots \end{cases}$$

Remark 5.5 (Polynomials and Trig Functions). Polynomials, cos, and sin are not only infinitely differentiable, but they're also really easy to differentiate. This makes working with them preferable. If we have a complicated function, it would be nice if we could find some way to express it as a polynomial or in terms of sin and cos. This is more of a longterm goal we will return to later.

5.7 Approximation

We now turn to a subject involving differentiation that will return several times – approximation. First we will discuss the derivative in the context of linear approximation of a function, then we will explore how we can get increasingly accurate approximations of a function by using higher order derivatives.

In Remark 5.2, an alternate definition of the derivative was presented.

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

A reformulation of this allows us to write the derivative as the difference quotient (with h) if we introduce some “remainder term”.

$$f'(x_0) + r(h)/h = \frac{f(x_0 + h) - f(x_0)}{h} \tag{9}$$

As long the ratio $r(h)/h \rightarrow 0$ as $h \rightarrow 0$, then this is a valid equation, as

$$\begin{aligned}
f'(x_0) &= \frac{f(x_0 + h) - f(x_0)}{h} + \frac{r(h)}{h} \\
f'(x_0) &= \lim_{h \rightarrow 0} \left(\frac{f(x_0 + h) - f(x_0)}{h} + \frac{r(h)}{h} \right) \\
f'(x_0) &= \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} + 0.
\end{aligned}$$

A valid question is, why do we the ratio $r(h)/h$ goes to zero and not just that $r(h) \rightarrow 0$? We need the remainder $r(h)$ to “shrink” faster than h , otherwise $r(h)/h \not\rightarrow 0$ as $h \rightarrow 0$. In fact, $r(h)/h$ could very well “blow up” if we don’t stipulate $r(h)/h \rightarrow 0$ as $h \rightarrow 0$ ⁸⁰. If this were to happen, then we would not end up with the definition of the derivative. If we multiply by h and rearrange some terms, (9) becomes

$$f(x_0 + h) = f(x_0) + f'(x_0)h + hr(h). \quad (10)$$

Equation (10) may seem random, but it’s *very* interesting. We know that $hr(h)$ is going to be small, so we could be informal and write

$$f(x_0 + h) \approx f(x_0) + f'(x_0)h.$$

For a given x_0 , this tells us that $f(x_0) + f'(x_0)h$ is an approximation for $f(x_0 + h)$ (f near some point x_0). Furthermore, this approximation is linear, as $f(x_0)$ and $f'(x_0)$ are constants, $f(x_0) + f'(x_0)h$ is a linear function of h . Not only is this a linear approximation, it is the *best* linear approximation we are able to achieve, a fact we can prove.

Notation 5.2 (“Little o ”). We say that $f(h) = o(g(h))$ (read as “ $f(h)$ is little o of $g(h)$ as $h \rightarrow 0$ ”) if

$$\lim_{h \rightarrow 0} \frac{f(h)}{g(h)} = 0.$$

This means that $f(h)$ tends to zero faster than $g(h)$ as $h \rightarrow 0$. For example,

$$\lim_{h \rightarrow 0} \frac{h^3}{h} = 0,$$

so we could write $o(h^3) = o(h)$.

When we noted that $r(h)/h \rightarrow 0$, what we really were saying was $r(h) = o(h)$.⁸¹ In the case of linear approximation, having a remainder which is $o(h)$ is as good as it gets, because h enters the equation linearly. For example, we cannot end up with some $h^n r(h)$ that vanishes in (10), giving $r(h) = o(h^n)$, because (10) would no longer be a linear function of h . We now can state and prove the result we have been building to.

Theorem 5.5 (Best Linear Approximation). If f is differentiable at x_0 , then

$$f(x_0 + h) = f(x_0) + f'(x_0)h + o(h).$$

Conversely, if there exists some linear approximation

$$f(x_0 + h) = B + Ah + o(h),$$

then not only is f differentiable at x_0 , but also $B = f(x_0)$ and $A = f'(x_0)$.

Proof. While we did not use the phrase “if and only if” in Theorem 5.5, this theorem is a biconditional statement. A function will have a best linear approximation if and only if it is differentiable.

(\Rightarrow) Suppose f is differentiable at x_0 . While building up to this theorem, we have already shown that $f(x_0 + h) = f(x_0) + f'(x_0)h + o(h)$, as $r(h) = o(h)$ in (10).

⁸⁰This would happen if $h \rightarrow 0$ “faster” than $r(h) \rightarrow 0$.

⁸¹One of the weird things about little o notation is we use “=” as a stand in for “is”. There is no actual equality when we write $o(h)$ instead of $r(h)$. In general this means when you see $o(h)$ you should think “something here is $o(h)$, the particular form of which is not that important because it is going to zero.”

(\Leftarrow) Suppose f has a (best) linear approximation of $f(x_0 + h) = B + Ah + o(h)$, for scalers $A, B \in \mathbb{R}$. Setting $h = 0$ gives $f(x_0) = B$. We therefore have

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - B}{h} = \lim_{h \rightarrow 0} \frac{Ah + o(h)}{h} = \lim_{h \rightarrow 0} \left(A + \frac{o(h)}{h} \right) = A.$$

□

The first part of Theorem 5.5 tells us that (10) is a good approximation, while the second part tells us that if such a good approximation exists, f is differentiable, and the approximation is given by (10). There is no other way to say it – the second part of this theorem is *****ing* amazing. The simple concept of approximating any function with a line, gives rise to the concept of the derivative.⁸² Is this a surprise though?

If you want to approximate a nonlinear function f with a line at a point x_0 , you have two degrees of freedom: a point on the line, and the slope of the line.⁸³ The choice of point is easy. You want to approximate f at x , so you should pick $(x_0, f(x_0))$. Now you have an infinite choice of slopes, which one do you pick? Pick the slope to be the rate of change at that point! This is just $f'(x_0)$ though. You literally⁸⁴ cannot make a better choice of point and slope.

Remark 5.6 (Another Alternate Definition). Theorem 5.5 inspires yet another equivalent definition of differentiation. We can say that f is differentiable at a point x if there exists some scalar $A \in \mathbb{R}$ such that

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x) - A \cdot h}{h} = 0.$$

In this case we write $f'(x) = A$. You sometimes hear the phrase “locally linear” associated with the existence of the derivative, and it comes from this definition. All we did was subtract $f(x)$ from both sides of the definition given in Remark 5.2. We now are expressing $f'(x)$ as some linear map where $h \mapsto f'(x)h$. This coincides with (10).

Example 5.13. We can show that $f(x) = |x|$ is not differentiable at $x_0 = 0$ using Theorem 5.5 and Remark 5.4. Suppose we had some linear approximation of $|x|$ at the point 0.

$$\begin{aligned} f(x_0 + h) &= B + Ah + r(h) \\ |0 + h| &= B + Ah + r(h) \\ |h| &= B + Ah + \underbrace{(|h| - B - Ah)}_{r(h)} \end{aligned}$$

But do we have $r(h) = o(h)$?

$$\lim_{h \rightarrow 0} \frac{r(h)}{h} = \lim_{h \rightarrow 0} \frac{(|h| - B - Ah)}{h} = \lim_{h \rightarrow 0} \frac{(|h| - B - A)}{h} \neq 0$$

This limit is undefined, as $|h|/h$ has no limit as $h \rightarrow 0$. This is the *exact* same limit we attempted to take in Example 5.3, but could not, render $|x|$ not differentiable at 0.

⁸²In fact, this is how the derivative is introduced to many people in calculus courses.

⁸³Any point and slope uniquely define a line, hence point-slope form from middle school: $y - y_1 = m(x - x_1)$.

⁸⁴And I mean the definition of the word “literally”, not “figuratively”.

Example 5.14. The equation of the line given in Theorem 5.5 looks a little strange. In calculus, the equation for a tangent line looks a bit like this, but something is off. For a fixed x_0 , it's a function of h , where $x_0 + h$ is a point close to x_0 . Instead, we can write this nearby point as $x = x_0 + h$. Equation (10) now becomes

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + r(h) \\ f(x) &\approx f(x_0) + f'(x_0)(x - x_0). \end{aligned}$$

This final equation is what most people see in calculus. Figure 52 shows the approximation for $x_0 + h$ and x .

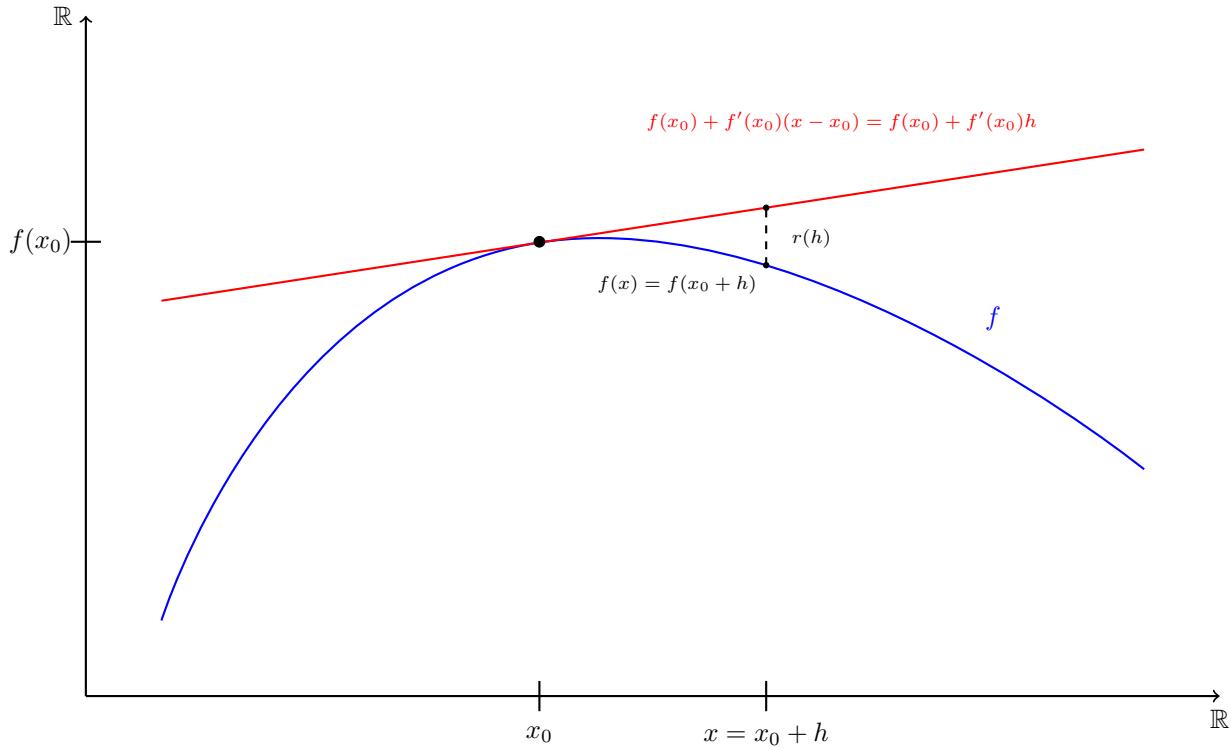


Figure 52: The equation of a tangent line to f at point x_0 is given as $f(x_0) + f'(x_0)(x - x_0)$ or $f(x) + f'(x)h$. The difference between the value of the function and the value of the approximation at a nearby point $x = x_0 + h$ is given by $r(h)$. If we add $r(h)$ to the approximation, then we have an equation for f as given in Equation (10).

Our next example not only calculates the best linear approximation for a function at a point, but also serves as motivation for our next theorem.

Example 5.15. Suppose we want to approximate e^x at the point $x_0 = 0$ using Theorem 5.5 (which we can

do because e^x is differentiable). Note in this case $x = h + x_0 = h + 0 = h$

$$\begin{aligned} f(x_0 + h) &= f(x_0) + f'(x_0)h + r(h) \\ e^{0+h} &= e^0 + e^0h + r(h) \\ e^h &= 1 + h + (e^h - 1 - h) \\ e^h &\approx 1 + h \\ e^x &\approx 1 + x. \end{aligned}$$

We won't be able to get a better *linear* approximation than this, but can we get a better approximation?

The linear approximation for f at x_0 is a result of picking the best point and slope to define the approximation. It makes sense that we should pick the point $(x_0, f(x_0))$ and the slope $f'(x_0)$. This is all the information about f we can incorporate into our approximation as far as derivatives go, because the equation of a line *must* have a second derivative of zero.⁸⁵ What if we allowed our approximation to have a nonzero second derivative? For e^x , such an approximation would be quadratic. It would look like

$$e^x \approx 1 + x + Cx^2,$$

for a scalar C . We can now incorporate more information about e^x into our approximation, namely we can pick a value of C such that the second derivative of our approximation matches the second derivative of e^x at our point of interest $x_0 = 0$. For $f(x) = e^x$, we have $f''(x) = e^x$, giving $f''(x_0) = f''(0) = 1$. Perhaps we should try $C = 1$.

$$\begin{aligned} e^x &\approx 1 + x + Cx^2 \\ e^x &\approx 1 + x + x^2 \end{aligned}$$

But if we differentiate our approximation, then we have $(1 + x + x^2)''(0) = 2 \neq 1$. Letting $C = f''(x_0)$ failed to account for the power rule, so we need to scale it by $1/2$ to account for this. Our updated quadratic approximation is

$$e^x \approx 1 + x + \frac{1}{2}x^2.$$

This approximation has the same value, derivative, and second derivative as e^x at $x_0 = 0$. But in what sense is this better than our linear approximation? We can address this with $r(h)$.

$$e^x \approx 1 + x + \frac{1}{2}x^2 + \underbrace{(e^x - 1 - x - \frac{1}{2}x^2)}_{r(h)}$$

Because $h = x$ in this case, we have $r(h) = e^h - 1 - h - \frac{1}{2}h^2$. Not only do we have $r(h) = o(h)$, but we have $r(h) = o(h^2)$!

$$\lim_{h \rightarrow 0} \frac{r(h)}{h^2} = \lim_{h \rightarrow 0} \frac{e^h - 1 - h - \frac{1}{2}h^2}{h^2} = 0$$

The remainder is moving to 0 faster than h^2 . In a sense, it's shrinking *twice* as fast as h . This makes our quadratic approximation superior to our linear approximation, as our quadratic remainder converges to 0 even faster than that of the linear approximation. But why stop at quadratic, why not cubic, or quartic, or

⁸⁵All subsequent derivatives will also be 0.

quintic? If we repeated this process we would arrive at the following equations:

$$\begin{aligned} e^x &= 1 + x + o(x) \\ e^x &= 1 + x + \frac{1}{2}x^2 + o(x^2) \\ e^x &= 1 + \frac{1}{2}x^2 + \frac{1}{6}x^3 + o(x^3) \\ e^x &= 1 + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + o(x^4) \\ e^x &= 1 + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \frac{1}{120}x^5 + o(x^5) \end{aligned}$$

Each time we increase the degree of our polynomial approximation, we can incorporate the information given by a higher order derivative of f .

The idea of approximating a differentiable function with a polynomial of some fixed degree is formalized in Taylor's Theorem. For this theorem, we will opt to use $(x - x_0)$ instead of h .

Theorem 5.6 (Taylor's Theorem I). Suppose f is a real function on $[a, b]$, n is a positive integer, and f is n times differentiable at a point x_0 . Then we have

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + o((x - x_0)^n).$$

We refer to the polynomial (excluding the remainder term) as the *n -th order Taylor Polynomial of f at x_0* , and write

$$P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

The proof just amounts to applying L'Hôpital's Rule $n - 1$ times

Proof. We just need to show that the remainder term is in fact $o((x - x_0)^n)$. If

$$r((x - x_0)) = f(x) - P_n(x)$$

is our remainder term, then we need to verify that

$$\lim_{(x-x_0) \rightarrow 0} \frac{f(x) - P_n(x)}{(x - x_0)^n} = 0.$$

By construction, $f^{(k)}(x_0) = P_n^{(k)}(x_0)$ for $k = 0, 1, \dots, n$. This means that our desired limit will give the indeterminate form of $0/0$, so we must use L'Hôpital's Rule.⁸⁶ But if we use it once, we still get an

⁸⁶You should verify that we meet the conditions required of Proposition 5.4.

indeterminate form of 0/0. We in fact must apply it $n - 1$ times before we do not end up with 0/0:

$$\begin{aligned}
\lim_{(x-x_0) \rightarrow 0} \frac{f(x) - P_n(x)}{(x-x_0)^n} &= \lim_{(x-x_0) \rightarrow 0} \frac{f'(x) - P'_n(x)}{n(x-x_0)^{n-1}} \\
&= \dots \\
&= \lim_{(x-x_0) \rightarrow 0} \frac{f^{(n-1)}(x) - P_n^{(n-1)}(x)}{n!(x-x_0)} \\
&= \frac{1}{n!} \lim_{x \rightarrow x_0} f^{(n)}(x) - P_n^{(n)}(x) \\
&= \frac{1}{n!} f^{(n)}(x_0) - P_n^{(n)}(x_0) \\
&= 0
\end{aligned}$$

Therefore we have $o((x-x_0)^n)$. \square

Remark 5.7 (Taylor's Theorem and the MVT). The presentation of Taylor's Theorem Rudin (1976) is slightly different. I suspect his (equally valid) treatment was motivated by the connection between the Mean Value Theorem and Taylor's Theorem. Rudin (1976) explicitly writes “For $n = 1$, the [Taylor's Theorem] is just the mean value theorem.” If we use Taylor's Theorem to find $P_n(x_0)$ for an $n+1$ differentiable function, we have

$$f(x) = \underbrace{f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n}_{P_n(x)} + \underbrace{\frac{f^{(n+1)}(x_0)}{(n+1)!}(x-x_0)^{n+1}}_{R_n(x)},$$

where $R_n(x)$ is a remainder term. Rudin's version says that under stronger conditions, we can always find some value between $\xi \in (x, x_0)$ which satisfies

$$f(x) = P_n(x) + R_n(\xi).$$

If we do this for $P_0(x)$ and $R_0(x)$, we get

$$f(x) = f(x_0) + f'(\xi)(x-x_0) \implies f'(\xi) = \frac{f(x) - f(x_0)}{x - x_0},$$

which is the Mean Value Theorem.

A natural question that arises with Taylor's Theorem is what happens when we let the $n \rightarrow \infty$. If f is infinitely differentiable at x_0 , then what is stopping us from writing the Taylor Polynomial as an infinite series? If we do this, will the series have any remainder? If it has no remainder does that mean it simply equals the function at every point it is defined? We will answer all these questions in Section 7!

5.8 Exercises

uniform continuity and second derivative

derivative of x^n using h definition

derivative of a^x

taylor approx is best

6 Riemann Integration

The theory of differentiation may at times have seemed painless. It's eerily similar to calculus. Isn't analysis supposed to be hard and confusing? Fear not, because here comes integration to bring us back to reality! The definition of Riemann integration cannot be given immediately like that of the derivative, as we need to do some prep work. Even once it is defined, proofs involving the Riemann integral are a bit more sophisticated than the average proof up until this point. This is a constant theme in math – integration is *way* harder than differentiation. Our goal is to develop a form of integration that works reasonably well for real functions. Fortunately, this simplifies things, as we're comfortable with real functions. There is good news. The Riemann integral mostly achieves our goal!⁸⁷ Nevertheless, there will be some drawbacks of Riemann integration, and we will only be able to go so far with it. For this reason, we will return to integration again in Sections 13-15. Consider this a first pass at what turns out to be a much more sophisticated problem.

Why are we even interested in integration though? Many problems are concerned with accumulation over time, which is captured by the area underneath a curve. For example, the probability of a certain event can be interpreted as the area under a probability distribution function. Similar problems exist in nearly every field of science, so it's important that we have some tool for measuring area under the curve of some function.

First, we should ask, when is it even possible to calculate the area under the curve of a function? Is it possible to measure the area over some infinite interval, such as all of \mathbb{R} ? It is not immediately clear how to do this, as we have not yet developed a way of measuring the length of \mathbb{R} ,⁸⁸ so we should stick to some bounded interval $[a, b] \subset \mathbb{R}$. We also want f to be bounded on \mathbb{R} , otherwise the area under the curve would not be well defined. For these reasons, **we will restrict our attention to bounded real functions on an interval $[a, b]$** for this whole section.

6.1 Partitions

First we need to develop some notation about partitioning an interval of the real line into smaller intervals.

Definition 6.1. Let $[a, b] \subset \mathbb{R}$. A *partition* P of $[a, b]$ is a finite set of points $P = \{x_0, x_1, \dots, x_n\}$ such that

$$a = x_0 \leq x_1 \leq \dots \leq x_{n-1} \leq x_n = b.$$

We will write $\Delta x_i = x_i - x_{i-1}$ for $i = 1, \dots, n$. We will denote the set of all partitions of $[a, b]$ as $\mathbf{P}([a, b])$.⁸⁹

Figure 53 shows a partition of the interval $[a, b]$. Contrary to what this figure shows, the points which comprise a partition need not be evenly spaced out along $[a, b]$. We can refine a partition by adding more points to $P = \{x_0, x_1, \dots, x_n\}$.

Definition 6.2. A partition P^* is *refinement* of P if $P \subset P^*$. Given two partitions, P_1 and P_2 , we say P^* is their *common refinement* if $P^* = P_1 \cup P_2$.

⁸⁷There is a reason that it is the only form of integration most people ever need to know and use.

⁸⁸We'll do this in Section 13.

⁸⁹ $\mathbf{P}([a, b])$ is *not* standard notation.

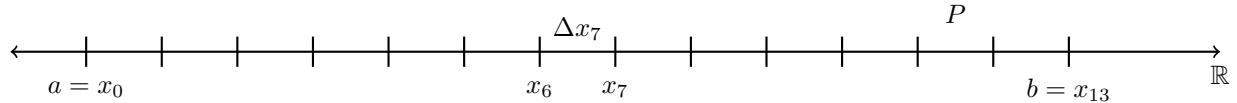


Figure 53: A Partition $P = \{x_0, x_1, \dots, x_{13}\}$ of the interval $[a, b]$.

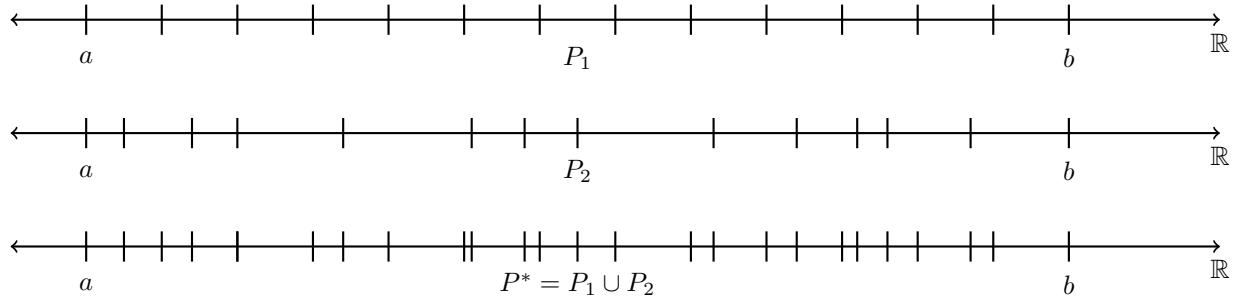


Figure 54: The partition P^* is a common refinement of P_1 and P_2 .

6.2 Upper and Lower Riemann Integrals

We will now use partitions to approximate the area under a curve. This will be nearly identical to the method developed in calculus, but with one small difference. Calculus courses normally use left and right Riemann sums, which use the points of a partition to determine the height of the rectangles used to approximate the area under a function. We will instead determine the height of the rectangles using the infimum and supremum of a function.

Definition 6.3. Suppose f is a bounded real function defined on $[a, b]$. Corresponding to each partition P of $[a, b]$, we have

$$\begin{aligned} M_i &= \sup_{x \in [x_{i-1}, x_i]} f(x) \\ m_i &= \inf_{x \in [x_{i-1}, x_i]} f(x) \\ U(P, f) &= \sum_{i=1}^n M_i \Delta x_i, \\ L(P, f) &= \sum_{i=1}^n m_i \Delta x_i, \end{aligned}$$

We call $U(P, f)$ an *upper Riemann sum* and $L(P, f)$ a *lower Riemann sum*.

Figure 55 shows what $m_i \Delta x_i$ and $M_i \Delta x_i$ may look like for a given function and partition. Figure 56 and Figure 57 show the upper Riemann sum and lower Riemann sum for the partition of $[a, b]$ that was shown in Figure 53, respectively. So far, this is fairly similar to the way integration was developed in calculus, but now we're going to do something a little different. Instead of taking a limit of Riemann sums, we will simply say a function is Riemann integrable if the supremum of all possible upper Riemann sums coincides with the infimum of all the possible lower Riemann sums.

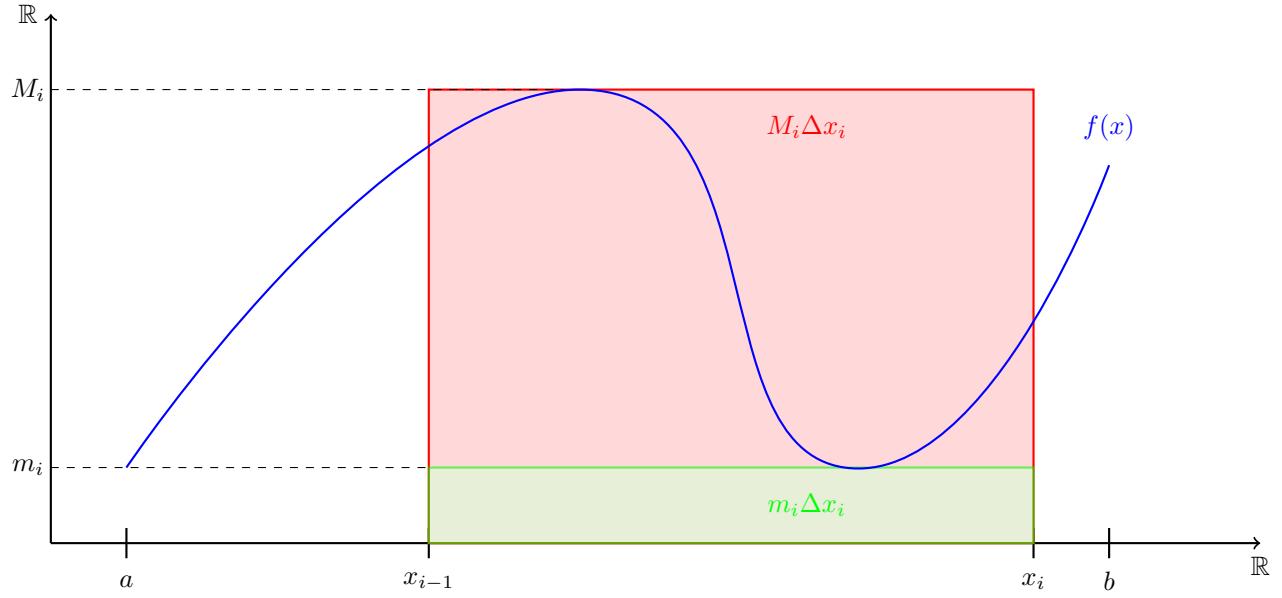


Figure 55: The particular values $m_i \Delta x_i$ and $M_i \Delta x_i$ for some function f and partition of $[a, b]$.

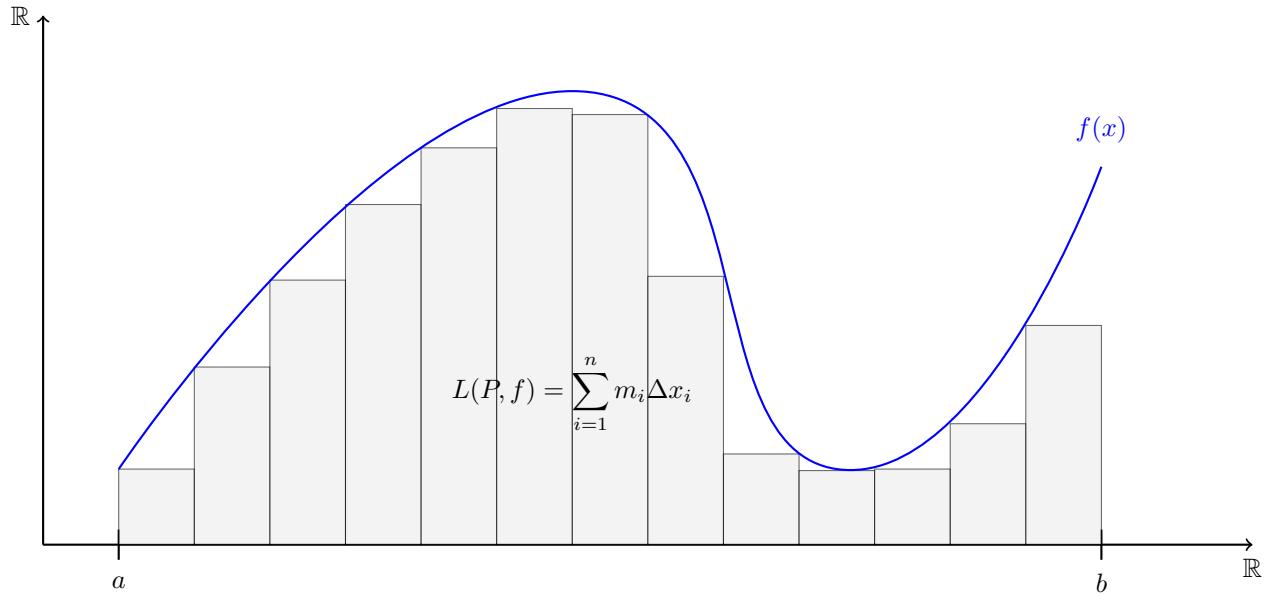


Figure 56: A lower Riemann sum.

Definition 6.4. Suppose f is a bounded real function on the interval $[a, b]$. Define the values

$$\begin{aligned}\underline{\int}_a^b f(x) dx &= \sup_{P \in \mathbf{P}([a, b])} L(P, f), \\ \bar{\int}_a^b f(x) dx &= \inf_{P \in \mathbf{P}([a, b])} U(P, f).\end{aligned}$$

We refer to these as the *lower Riemann integral of f over $[a, b]$* and *upper Riemann integral of f over $[a, b]$* respectively.

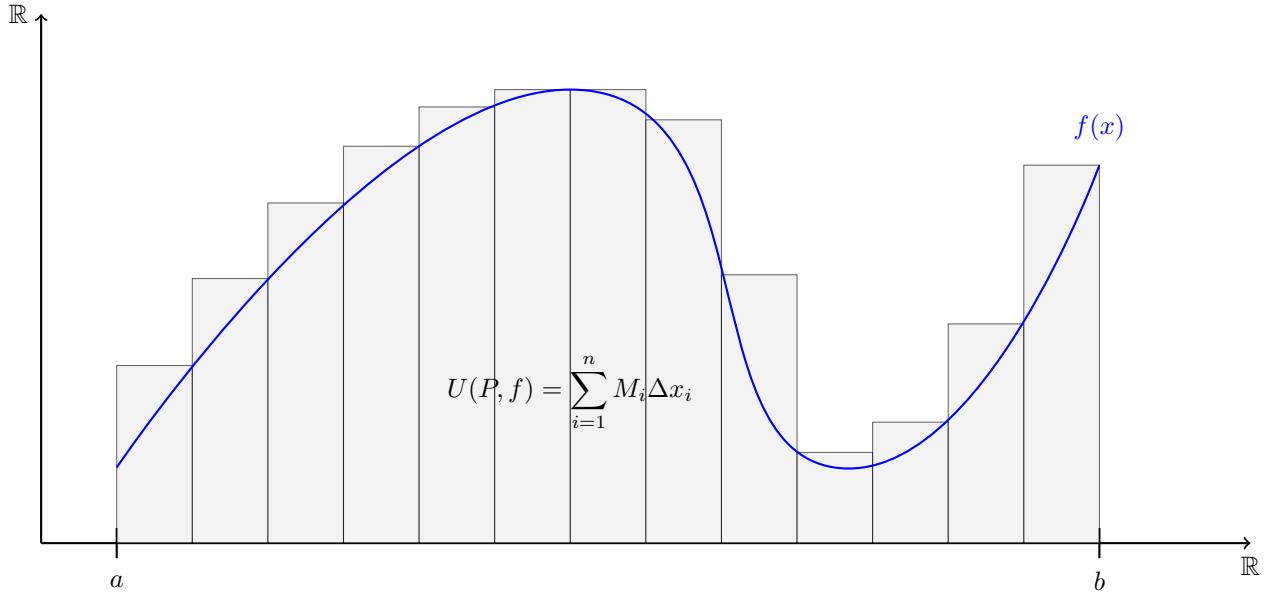


Figure 57: An upper Riemann sum.

The upper and lower Riemann integral will *always* exist for any bounded function on $[a, b]$. The set $\mathbf{P}([a, b]) \neq \emptyset$, as any interval $[a, b]$ has a trivial partition of $\{x_0 = a, x_1 = b\}$. The set is also bounded, as for all $P \in \mathbf{P}([a, b])$,

$$\inf_{x \in [a, b]} f(x) \cdot (b - a) \leq L(P, f) \leq U(P, f) \leq \sup_{x \in [a, b]} f(x) \cdot (b - a).$$

We know the infimum and supremum of f exist on $[a, b]$, because f is bounded. This gives us two nonempty bounded subsets of \mathbb{R} in the form of $\{U(P, f)\}_{P \in \mathbf{P}([a, b])}$ and $\{L(P, f)\}_{P \in \mathbf{P}([a, b])}$. The supremum and infimum of these sets are guaranteed to exist by the completeness of \mathbb{R} .

We define the Riemann integral using these upper and lower integrals.

Definition 6.5. Suppose f is a bounded real function on the interval $[a, b]$. If

$$\int_a^b f(x) dx = \bar{\int}_a^b f(x) dx,$$

then we say f is *Riemann integrable (on $[a, b]$)* and we write the common value of the upper and lower Riemann integral as

$$\int_a^b f(x) dx.$$

We refer to this common value as the *Riemann integral of f on $[a, b]$* .

Figure 58 shows a Riemann integrable function, and the value of the integral on $[a, b]$.

Remark 6.1 (Bounded Functions on a Closed Interval). It cannot be stressed enough that we are only able to define the Riemann integral for real functions that are bounded, and can only do so on a closed interval.

While Definition 6.5 may make sense on a theoretical level, it's not at all practical to use it to verify a function is actually integrable. How are we supposed to possibly find the supremum of all possible lower Riemann sums, and find the infimum of all possible upper Riemann sums?! A simple example will be

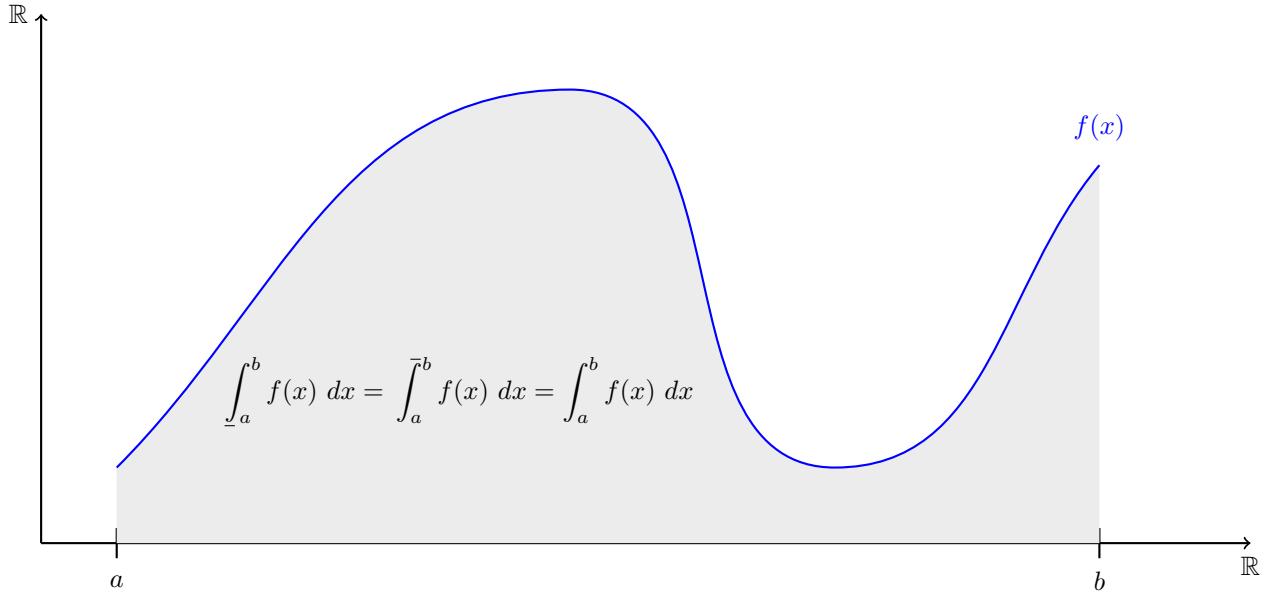


Figure 58: The function f is Riemann integrable on $[a, b]$.

presented, but it will be the only time we actually use Definition 6.5 to verify integrability. We will therefore need to find some alternate criterion for Riemann integrability.

Example 6.1. Let $f(x) = c$ on $[0, 1]$ for some constant $c \in \mathbb{R}$. Suppose $P = \{0, x_1, \dots, x_{n-1}, 1\}$ is a partition of $[0, 1]$. We have

$$\begin{aligned} M_i &= \sup_{x \in [x_{i-1}, x_i]} f(x) = c, \\ m_i &= \inf_{x \in [x_{i-1}, x_i]} f(x) = c, \end{aligned}$$

For $i = 1, \dots, n$. The lower and upper Riemann sums will agree, as $M_i = m_i$ for $i = 1, \dots, n$.

$$U(P, f) = L(P, f) = \sum_{i=1}^n c \cdot (x_k - x_{k-1}) = c(x_n - x_{n-1}) + \dots + c(x_1 - x_0) = c(x_n - x_0) = c(1 - 0) = c.$$

We let P be arbitrary, so these sums will always be 1.

$$\begin{aligned} \int_0^1 f(x) dx &= \sup_{P \in \mathbf{P}([0,1])} L(P, f) = \sup\{c\} = c \\ \int_0^1 f(x) dx &= \inf_{P \in \mathbf{P}([0,1])} U(P, f) = \inf\{c\} = c. \end{aligned}$$

These values agree, so f is Riemann integrable, and

$$\int_0^1 c dx = c.$$

6.3 An Alternative Interpretation: Simple Functions (Very Optional)

Before we explore the properties of Riemann integration and develop a way to verify we can integrate a function, we can give an alternate formulation of the upper and lower Riemann integral. Not only is this

how Tao (2016a) opts to introduce the Riemann integral, but it is how we will go about defining a superior form of integration in Section 14. The general idea is that we define the Riemann integral for very “simple” functions, and then relate the integration of these functions to bounded real functions on $[a, b]$. **Disclaimer:** this introduction to simple functions is slightly informal for reasons that will be acknowledged afterwards.

We begin by defining a special type of step function.

Definition 6.6. Let X be a subset of \mathbb{R} , and $E \subset X$. We define the *characteristic function on E* as $\chi_E : X \rightarrow \{0, 1\}$, where

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

This function is sometimes called the indicator function, as it indicates whether or not an element of X is in the subset E . We will be interested in characteristic functions in \mathbb{R} , and an example of such a function is seen Figure 59. We can use characteristic functions to write *any* step function. Step functions written

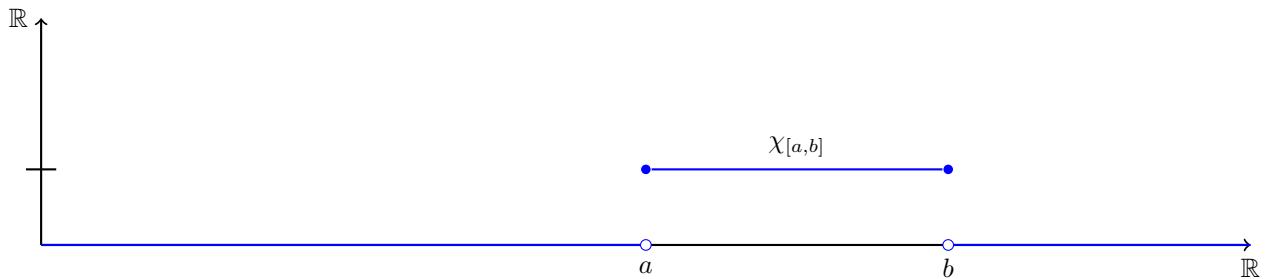


Figure 59: The characteristic function $\chi_{[a,b]} : \mathbb{R} \rightarrow \{0, 1\}$.

this way will be the “simple” functions which we are interested in.

Definition 6.7. Let X be a subset of \mathbb{R} . A *simple function* $\varphi : X \rightarrow \mathbb{R}$ is a function which can be written as a *finite* linear combination of characteristic functions on $E \subset X$. That is there exists a set of finite scalars $\{c_1, \dots, c_n\} \subset \mathbb{R}$ and finite sets $E_1, \dots, E_n \subset X$ such that

$$\varphi(x) = \sum_{k=1}^n c_k \chi_{E_k}(x).$$

A simple functions domain will always be X , even if the sets on which the characteristic functions are defined do not cover X , i.e $X \not\subset \bigcup_{k=1}^n E_k$. We simply have $\varphi(x) = 0$ on the set $X \setminus \bigcup_{k=1}^n E_k$. The codomain of a simple function will *always* be \mathbb{R} , as the output of a simple function is determined by the set of real scalars $\{c_1, \dots, c_n\}$. Where the function attains each value c_k in the range is determined by the set E_k . A concrete example where $X \subset \mathbb{R}$ may make things clearer.

Example 6.2. Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ be the simple function

$$\varphi(x) = 2\chi_{[1,3.5)} + 5\chi_{[3.5,4)} - 2\chi_{[4,7)} + 4\chi_{[7,8)} - \chi_{[8,10)} + \chi_{[10,14)} + 2\chi_{[14,15]}$$

on \mathbb{R} , where each χ_E is defined on a subset $E \subset [1, 15]$. This function is shown in Figure 60. In this case, $E_1 \cup \dots \cup E_8 \neq \mathbb{R}$, but $\varphi(x)$ is still defined on all of \mathbb{R} . The function takes on the value 0 outside of $E_1 \cup \dots \cup E_7$, but is still defined there. We could also define $\varphi : [1, 15] \rightarrow \mathbb{R}$ in a similar fashion. The only difference would be that φ would not be defined as zero on $(-\infty, 1) \cup (15, \infty)$.

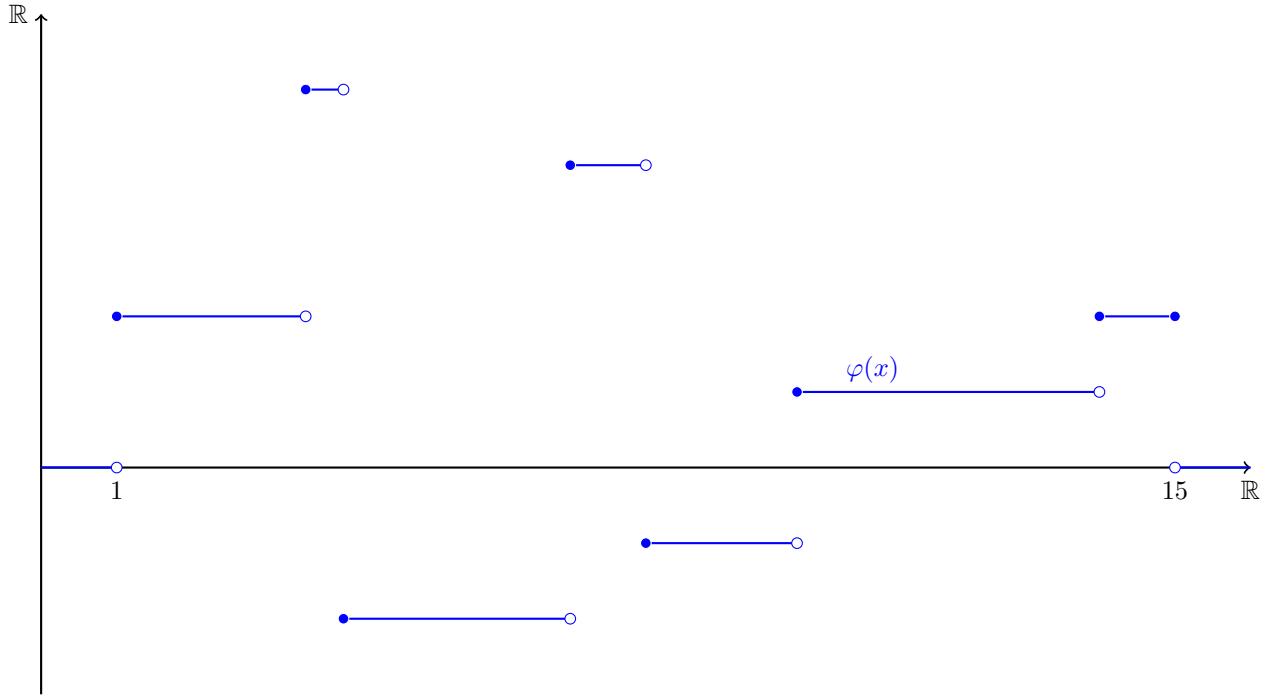


Figure 60: A simple function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$.

In this particular case the set $\{E_k\}$ are piecewise disjoint. This need not be the case. Just know that if the intervals are not piecewise disjoint, we can always rewrite $\varphi(x)$ so they are. We will prove this later on when we treat simple functions with a bit more care.

Simple functions get their name, as in the context of integration they are simple. Because a simple function is written as a finite linear combination, it is bounded on the entirety of its domain.⁹⁰ If φ is only defined on some interval $[a, b]$, then every set $E_k \subset [a, b]$ is bounded for all k . These facts allow us to appeal directly to geometry to calculate the Riemann integral of a simple function.

Definition 6.8. Suppose $\varphi : [a, b] \rightarrow \mathbb{R}$ is a simple function which can be expressed as $\sum_{k=1}^n c_k \chi_{E_k}(x)$ for $E_k = (a_k, b_k) \subset [a, b]$ and $c_k \in \mathbb{R}$. Then we define the *Riemann integral of a simple function (on $[a, b]$)* as

$$\int_a^b \varphi(x) dx = \sum_{k=1}^n c_k \cdot (b_k - a_k).$$

For each “step” of the simple function, we multiply the width of the interval, $(b_k - a_k)$, by the height the function achieves on that interval, c_k . Adding the area of these rectangles up gives the integral. In Definition 6.8, E_k is open, but it really does not matter. It could be closed, or half-open. We will freely interchange them in this section, as it won’t make a difference. The length of the interval would still be $b_k - a_k$.

Example 6.3. Define $\varphi : [1, 15] \rightarrow \mathbb{R}$ as

$$\varphi(x) = 2\chi_{[1, 3.5)} + 5\chi_{[3.5, 4)} - 2\chi_{[4, 7)} + 4\chi_{[7, 8)} - \chi_{[8, 10)} + \chi_{[10, 14)} + 2\chi_{[14, 15]}.$$

⁹⁰The function is bounded by $\max\{|c_1|, \dots, |c_n|\}$.

Definition 6.8 gives

$$\begin{aligned}
\int_1^{15} \varphi(x) dx &= \sum_{k=1}^7 c_k \cdot (b_k - a_k) \\
&= 2(3.5 - 1) + 5(4 - 3.5) - 2(7 - 4) + 4(8 - 7) - (10 - 8) + (14 - 10) + 2(15 - 14) \\
&= 9.5.
\end{aligned}$$

This integral can be seen in Figure 61.

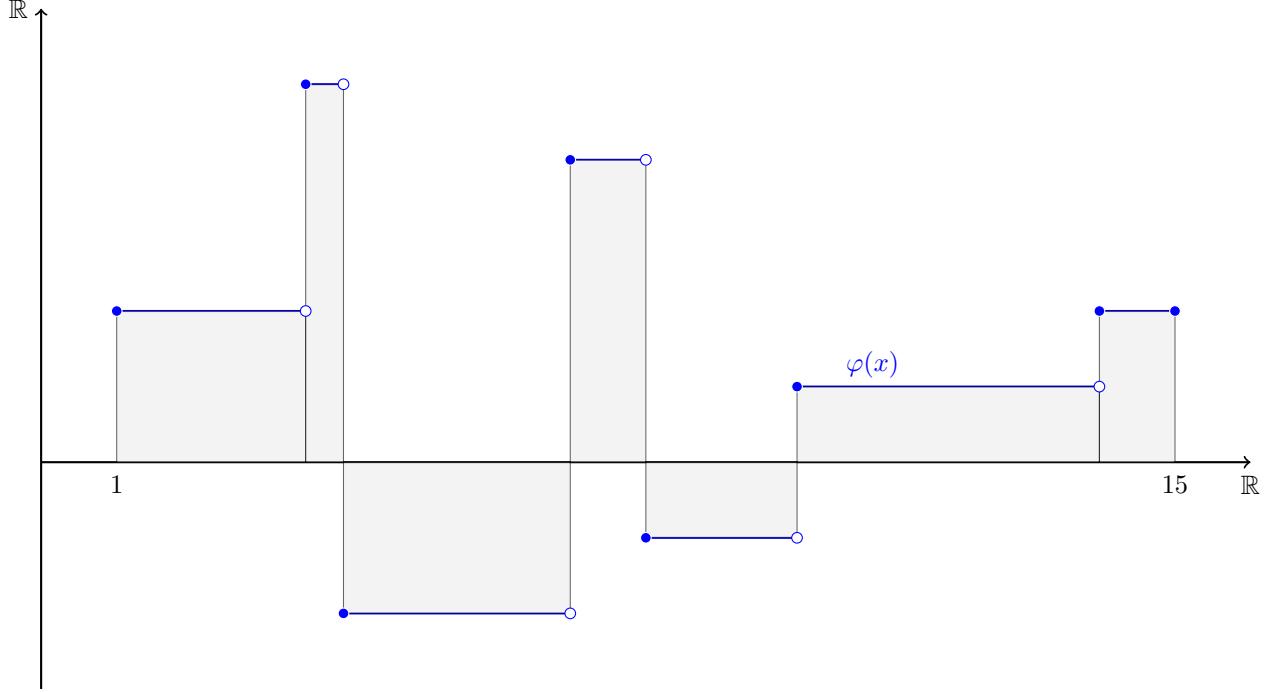


Figure 61: The integral of a simple function $\varphi : [1, 15] \rightarrow \mathbb{R}$

Example 6.4. Let $\varphi : [a, b] \rightarrow \mathbb{R}$ be the trivial simple function of $\varphi(x) = \chi_{[a,b]}$. We have

$$\int_a^b \varphi(x) dx = b - a.$$

Now we that we can integrate simple functions, how do we extend this to any bounded real function on $[a, b]$?

Definition 6.9. Let X be a set, Y be an ordered set, $f : X \rightarrow Y$, and $g : X \rightarrow Y$. We say f is greater than or equal to g (on X) if $f(x) \geq g(x)$ for all $x \in X$, and write $f \geq g$. Similarly, we say f is less than or equal to g (on X) if $f(x) \leq g(x)$ for all $x \in X$, and write $f \leq g$.

Like many of the definitions that have been introduced, it will be important to specify a domain when we say a function is greater than or less than another. It should always be specified if it is unclear.

Let there be some real valued function f that is bounded on the interval $[a, b]$. Suppose for two simple functions defined on $[a, b]$, call them φ and ψ , we have $\varphi \leq f \leq \psi$ on $[a, b]$, as seen in Figure 62. Can we somehow use the well defined integrals of φ and ψ to approximate that of f ? If you compare Figure 62 to

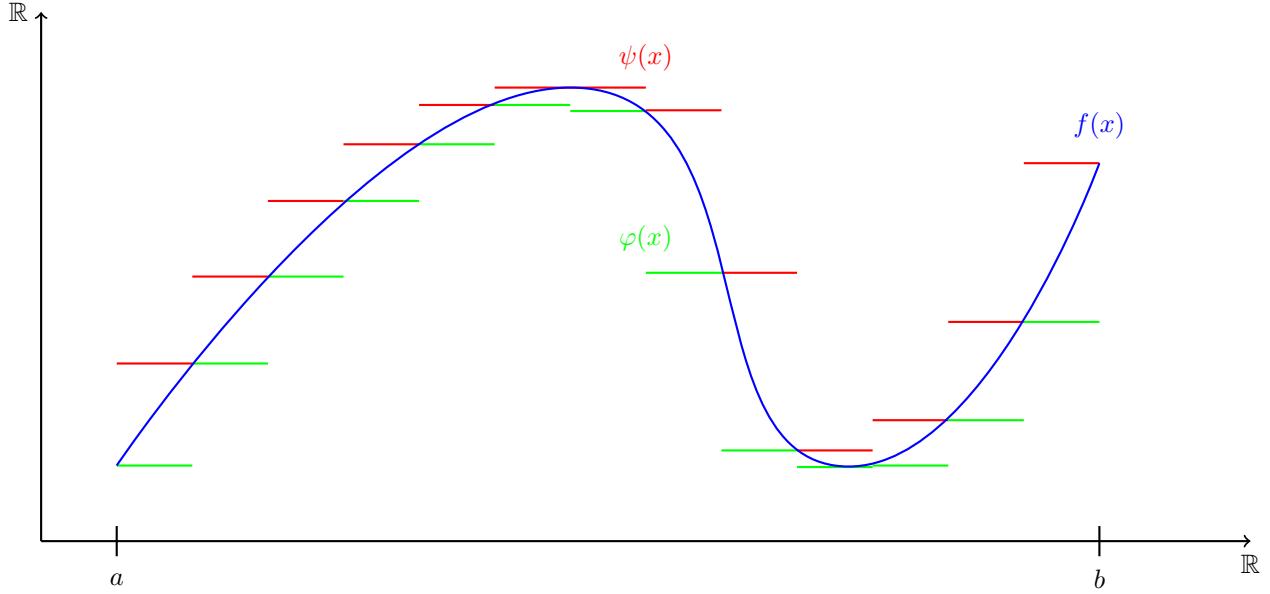


Figure 62: The function f is real and bounded on the interval $[a, b]$. We have $\varphi \leq f \leq \psi$ on $[a, b]$ for two simple functions defined on that interval.

Figure 56 and Figure 57, then things become clearer. The integrals of φ and ψ on $[a, b]$ coincide with the lower and upper Riemann sums shown in Figure 56 and Figure 57! In fact, we can think of a Riemann sum as the integral of a simple function. For a partition $P = \{x_0, \dots, x_n\}$ of $[a, b]$ we may write

$$\begin{aligned}\varphi(x) &= \sum_{i=1}^{n-1} m_i \chi_{[x_{i-1}, x_i)} + m_n \chi_{[x_{n-1}, x_n]}, \\ \psi(x) &= \sum_{i=1}^{n-1} M_i \chi_{[x_{i-1}, x_i)} + M_n \chi_{[x_{n-1}, x_n]},\end{aligned}$$

which gives

$$\begin{aligned}\int_a^b \varphi(x) dx &= \sum_{i=1}^n m_i \Delta x_i, \\ \int_a^b \psi(x) dx &= \sum_{i=1}^n M_i \Delta x_i.\end{aligned}$$

Thus we have established the following set inclusions:

$$\{L(P, f)\}_{P \in \mathbf{P}([a, b])} \subset \left\{ \int_a^b \varphi(x) dx \mid \varphi \leq f \text{ on } [a, b], \varphi \text{ simple} \right\}, \quad (11)$$

$$\{U(P, f)\}_{P \in \mathbf{P}([a, b])} \subset \left\{ \int_a^b \psi(x) dx \mid \psi \geq f \text{ on } [a, b], \psi \text{ simple} \right\}. \quad (12)$$

I'm now going to make what may seem like a bold claim, but is certainly corroborated by Figure 62 and our

geometric understanding of the area under a curve:

$$\begin{aligned}\underline{\int}_a^b f(x) dx &= \sup\{L(P, f)\}_{P \in \mathbf{P}([a, b])} = \sup \left\{ \int_a^b \varphi(x) dx \mid \varphi \leq f \text{ on } [a, b], \varphi \text{ simple} \right\}, \\ \bar{\int}_a^b f(x) dx &= \inf\{U(P, f)\}_{P \in \mathbf{P}([a, b])} = \inf \left\{ \int_a^b \varphi(x) dx \mid \varphi \geq f \text{ on } [a, b], \varphi \text{ simple} \right\}.\end{aligned}$$

Let's provide an informal proof sketch as to why this works for the lower Riemann integral.⁹¹

We know that $\sup A \leq \sup B$, for $A \subset B$, so (11) gives us the first half of the result. We just need to show that

$$\sup\{L(P, f)\}_{P \in \mathbf{P}([a, b])} \geq \sup \left\{ \int_a^b \varphi(x) dx \mid \varphi \leq f \text{ on } [a, b], \varphi \text{ simple} \right\}.$$

We can show that any integral of such a simple function is less than or equal to a lower Riemann sum, implying that the supremum of the set must take the form of a lower Riemann sum. Suppose $\varphi \leq f$ on $[a, b]$ and is written as

$$\varphi(x) = \sum_{i=k}^n c_k \chi_{E_k} = \sum_{k=1}^n c_k \chi_{[a_k, b_k)}$$

for $c_k \in \mathbb{R}$ and $a_k, b_k \in [a, b]$. Assume that $a_{k+1} = b_k$, so $\cup_{k=1}^n [a_k, b_k) = [a, b]$. We have $\varphi \leq f$, so $c_k \leq \inf_{x \in [a_k, b_k)} f(x)$. Therefore

$$\int_a^b \varphi(x) dx = \sum_{k=1}^n c_k (b_k - a_k) \leq \sum_{k=1}^n \inf_{x \in [a_k, b_k)} f(x) (b_k - a_k).$$

But this final sum is just the integral of a simple function corresponding to a lower Riemann sum. Therefore for any integral of a simple function $\varphi \leq f$, there exists lower Riemann sum that is greater than or equal to it.

This all justifies the following alternate definition of the upper and lower Riemann integrals:

$$\begin{aligned}\underline{\int}_a^b f(x) dx &= \sup \left\{ \int_a^b \varphi(x) dx \mid \varphi \leq f \text{ on } [a, b], \varphi \text{ simple} \right\}, \\ \bar{\int}_a^b f(x) dx &= \inf \left\{ \int_a^b \varphi(x) dx \mid \varphi \geq f \text{ on } [a, b], \varphi \text{ simple} \right\}.\end{aligned}$$

If this makes no sense, that is fine. For the rest of this section, we will use the definitions presented in Subsection 6.1. In fact, we will never formally use any proofs involving Riemann integrals. We will however use simple functions in certain examples. This was only presented in an effort to make Section 14 easier. When we return to the idea of simple functions then, it will actually be even clearer than this case, as Section 13 will develop a formal concept of measuring an interval.

Remark 6.2 (About That Disclaimer). Before the introduction of simple functions, I warned that it would not be totally formal. I made a very strong assumption that the length of an interval $[a, b]$ is $b - a$, and I made many assumptions about the properties of this length, and the length of the analogous open interval (a, b) . Once again, we haven't defined any notion of length on \mathbb{R} , so this was careless. Tao (2016a) does a very good job of introducing the basic idea of length on \mathbb{R} when discussing partitions and simple functions.

⁹¹The case for the upper Riemann integral is similar.

6.4 Verifying Riemann Integrability

Okay, that detour is now over. We return to the more pressing problem of us not having any easy way of verifying a bounded real function on $[a, b]$ is Riemann integrable. After introducing a series of lemmas, we will arrive at a familiar criterion that we will use to verify Riemann integrability.

Lemma 6.1. Let P be a partition of the interval $[a, b]$. If P^* is a refinement of P , then

$$\begin{aligned} L(P, f) &\leq L(P^*, f) \\ U(P^*, f) &\leq U(P, f) \end{aligned}$$

Proof. We will show the result for the first inequality, as the proof for the second is analogous. Suppose that P^* contains exactly one more point than P^* , call it x^* . There are two consecutive points of P , x_{i-1} and x_i , such that $x_{i-1} < x^* < x_i$. Define

$$\begin{aligned} w_1 &= \inf_{x \in [x_{i-1}, x^*]} f(x), \\ w_2 &= \inf_{x \in [x^*, x_i]} f(x). \end{aligned}$$

For $m_i = \inf_{x \in [x_{i-1}, x_i]} f(x)$, we have $w_1 \geq m_i$ and $w_2 \geq m_i$.⁹² We therefore have

$$\begin{aligned} L(P^*, f) - L(P, f) &= w_1(x^* - x_{i-1}) + w_2(x_i - x^*) - m_i[x_i - x_{i-1}] \\ &= (w_1 - m_i)(x^* - x_{i-1}) + (w_2 - m_i)(x_i - x^*) \end{aligned}$$

This is the desired result if P^* only has one additional point.⁹³ If instead P^* contains k more points than P , we simply repeat this process k times to arrive at our result. \square

Example 6.5. Let f be the bounded real function on $[0, 1]$ defined as $f(x) = x^2$. Suppose $P = 0, 1$, and $P^* = 0, 1/2, 1$. We have

$$\begin{aligned} L(P, f) &= \inf_{x \in [0, 1]} x^2 = 0 \\ L(P^*, f) &= \inf_{x \in [0, 1/2]} x^2 + \inf_{x \in [1/2, 1]} x^2 = 0 + \frac{1}{4} = \frac{1}{4} \end{aligned}$$

This agrees with Lemma 6.1 (as it should).

Our next lemma provides an inequality we would hope holds for the upper and lower Riemann integral.

Lemma 6.2. Let f be a bounded real function on $[a, b]$. We have

$$\int_a^b f(x) dx \leq \int_a^b f(x) dx.$$

⁹²The bound m_i is the original infimum on the interval $[x_{i-1}, x_i]$. When we add x^* to our partition, this interval gets split into two intervals: $[x_{i-1}, x^*]$, and $[x^*, x_i]$. We're saying that the infima of these two new subintervals are weakly greater than that of the original interval. Drawing a picture of this can be helpful!

⁹³How did we end up with such a simple expression for $L(P^*, f) - L(P, f)$? Well the only difference between P^* and P is that P^* has the intervals $[x_{i-1}, x^*]$, and $[x^*, x_i]$ instead of $[x_{i-1}, x_i]$. All the other intervals are the same and cancel out. We're left with the two additional intervals, and need to subtract the interval which they replaced.

Proof. Let P_1 and P_2 be partitions of $[a, b]$, and P^* be a common refinement of P_1 and P_2 . By Lemma 6.1,

$$L(P_1, f) \leq L(P^*, f) \leq U(P^*, f) \leq U(P_2, f).$$

If we fix P_1 and P_2 , we can take the supremum and infimum over all of $\mathbf{P}([a, b])$.

$$\begin{aligned} \sup_{P_1 \in \mathbf{P}([a, b])} L(P_1, f) &\leq \inf_{P_2 \in \mathbf{P}([a, b])} U(P_2, f) \\ \int_a^b f(x) \, dx &\leq \bar{\int}_a^b f(x) \, dx \end{aligned}$$

□

When does Lemma 6.2 hold with equality? By Definition 6.5, this occurs precisely when f is Riemann integrable, because this is how we defined Riemann integrability. This would imply that we have

$$\underline{\int}_a^b f(x) \, dx < \bar{\int}_a^b f(x) \, dx$$

if and only if f is *not* Riemann integrable. At the end of this section, we will see some examples of this.

We are able to present the first of several results that will allow us to determine if a function is Riemann integrable. This theorem not only gives us sufficient conditions for integration, but it also gives a necessary condition for integrability. This gives it much more bite.

Theorem 6.1 (Riemann's Criterion). Suppose f is a bounded real function on $[a, b]$. The function f is Riemann integrable if and only if for all $\varepsilon > 0$ there exists a partition such that

$$U(P, f) - L(P, f) < \varepsilon.$$

The proof of this crucial theorem amounts to just shuffling around inequalities. It only looks so long due to liberal typesetting.

Proof.

(\Rightarrow) Suppose f is Riemann integrable, and let $\varepsilon > 0$. By the completeness of the real numbers and properties of the supremum, there exists a partition P_1 such that

$$\begin{aligned} U(P_1, f) - \sup_{P \in \mathbf{P}([a, b])} U(P, f) &< \frac{\varepsilon}{2}, \\ U(P_1, f) - \bar{\int}_a^b f(x) \, dx &< \frac{\varepsilon}{2}, \\ U(P_1, f) - \underline{\int}_a^b f(x) \, dx &< \frac{\varepsilon}{2}. \end{aligned}$$

Similarly, there exists a partition P_2 such that

$$\bar{\int}_a^b f(x) \, dx - L(P_2, f) < \frac{\varepsilon}{2}.$$

If we choose P^* to be the common refinement of P_1 and P_2 ,⁹⁴ then by Lemma 6.1,

$$\begin{aligned} U(P^*, f) - \int_a^b f(x) dx &< U(P_1, f) - \int_a^b f(x) dx < \frac{\varepsilon}{2}, \\ \int_a^b f(x) dx - L(P^*, f) &< \int_a^b f(x) dx - L(P_2, f) < \frac{\varepsilon}{2}. \end{aligned}$$

Combining these two inequalities for P^* yields

$$\begin{aligned} U(P^*, f) - L(P^*, f) &= U(P^*, f) - L(P^*, f) + 0 \\ &= U(P^*, f) - L(P^*, f) + \left(- \int_a^b f(x) dx + \int_a^b f(x) dx \right) \\ &= \left(U(P^*, f) - \int_a^b f(x) dx \right) + \left(\int_a^b f(x) dx - L(P^*, f) \right) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon. \end{aligned}$$

(\Leftarrow) Suppose for all $\varepsilon > 0$ there exists a partition such that $U(P, f) - L(P, f) < \varepsilon$. By the definition of supremum and infimum we have

$$\begin{aligned} L(P, f) &\leq \sup_{P \in \mathbf{P}([a, b])} U(P, f) = \int_a^b f(x) dx, \\ \int_a^b f(x) dx &\leq \inf_{P \in \mathbf{P}([a, b])} U(P, f) \leq U(P, f), \end{aligned}$$

for all $P \in \mathbf{P}([a, b])$. Combining this inequalities with Lemma 6.2 gives

$$L(P, f) \leq \int_a^b f(x) dx \leq \bar{\int}_a^b f(x) dx \leq U(P, f).$$

This implies that

$$0 \leq \int_a^b f(x) dx - \bar{\int}_a^b f(x) dx \leq \varepsilon,$$

but if this holds for all $\varepsilon > 0$,

$$\int_a^b f(x) dx = \bar{\int}_a^b f(x) dx.$$

Therefore f is Riemann integrable. □

Riemann's Criterion should not catch us off guard. If f is integrable then the lower and upper Riemann integrals are equal. By the definition of those values, we need that the upper and lower Riemann sums become arbitrarily close to each other. We've seen this type of behavior with sequences, limits, and continuity. This time instead of some n or δ term which correspond to ε , we have a partition P that corresponds to ε . As we take ε to be smaller and smaller, we will need to find finer and finer partitions P , but we will always be able to do this in order to satisfy $U(P, f) - L(P, f) < \varepsilon$.

⁹⁴Remember all the proofs with convergent sequences where we had two values B_1 and B_2 , and took $N = \max\{B_1, B_2\}$ (see the proof of Proposition 3.2 for the first example of this)? This is essentially what we're doing by taking the common refinement of P_1 and P_2 . We know that the all the ε inequalities will hold simultaneously by Lemma 6.1, allowing us to combine them.

Remark 6.3 (Riemann Criterion and Cauchy Criterion). The Riemann Criterion may have nothing to do with sequences, but it shares one important similarity with the Cauchy Criterion. Recall that the Cauchy Criterion is so useful because it does not require us to know the limit of a sequence. The same is true with the Riemann Criterion. We don't need to know the integral of f on $[a, b]$ to prove it exists!

Example 6.6. Let $f : [0, 1] \rightarrow \mathbb{R}$ be

$$f(x) = \begin{cases} 1/n & \text{if } x = 1/n \\ 0 & \text{otherwise} \end{cases}.$$

We can use the Riemann Criterion to show this function is integrable on $[0, 1]$. Let $\varepsilon > 0$. The sequence $\{1/n\}$ converges to 0, so for all $\varepsilon/2 > 0$ there exists an N such that $1/n \in [0, \varepsilon/2]$ for all $n \geq N$. Only a finite number of values of the form $1/n$ are in the interval $[\varepsilon/2, 1]$ (Proposition 3.1). We can cover these finite values by intervals $[x_1, x_2], \dots, [x_{m-1}, x_m]$ such that $x_i \in [\varepsilon/2, 1]$ for all $i = 1, \dots, m$. The finite length of these m intervals is less than $\varepsilon/2$. Let $P = \{0, \varepsilon, x_1, \dots, x_m\}$. We have

$$U(P, f) - L(P, f) = (\varepsilon/2 - 0) + \underbrace{(x_2 - x_1) + \dots + (x_m - x_{m-1})}_{<\varepsilon/2} < \varepsilon.$$

The function is Riemann integrable.⁹⁵

The Riemann Criterion gives way to several related results that put integrability in conversation with limit processes.

Proposition 6.1. Suppose f is a bounded real function on the interval $[a, b]$.

1. If $U(P, f) - L(P, f) < \varepsilon$ for some P , then it holds for every refinement of P .
2. If $U(P, f) - L(P, f) < \varepsilon$ for some $P = \{x_0, \dots, x_n\}$, and if $s_i, t_i \in [x_{i-1}, x_i]$, then

$$\sum_{i=1}^n |f(s_i) - f(t_i)| \Delta x_i < \varepsilon.$$

3. If f is Riemann integrable, $U(P, f) - L(P, f) < \varepsilon$ for some $P = \{x_0, \dots, x_n\}$, and $t_i \in [x_{i-1}, x_i]$, then

$$\left| \sum_{i=1}^n f(t_i) \Delta x_i - \int_a^b f(x) dx \right| < \varepsilon.$$

Proof.

1. Let P^* be a refinement of P . By Lemma 6.1,

$$U(P^*, f) - L(P^*, f) < U(P, f) - L(P, f) < \varepsilon.$$

2. We have $f(s_i), f(t_i) \in [m_i, M_i]$, as f is bounded on any interval by the supremum and infimum on that interval. This means that

$$|f(s_i) - f(t_i)| \leq M_i - m_i.$$

Therefore

$$\sum_{i=1}^n |f(s_i) - f(t_i)| \Delta x_i \leq \sum_{i=1}^n (M_i - m_i) \Delta x_i = \sum_{i=1}^n M_i \Delta x_i - \sum_{i=1}^n m_i \Delta x_i = < U(P, f) - L(P, f) < \varepsilon.$$

⁹⁵If you keep going with this example, you can verify that $\int_0^1 f(x) dx = 0$.

3. We will always have the two following inequalities:

$$L(P, f) \leq \sum_{i=1}^n f(t_i) \Delta x_i \leq U(P, f),$$

$$L(P, f) \leq \int_a^b f(x) dx \leq U(P, f).$$

These combine to give

$$\left| \sum_{i=1}^n f(t_i) \Delta x_i - \int_a^b f(x) dx \right| \leq U(P, f) - L(P, f) < \varepsilon.$$

□

This proposition seems like a random string of inequalities,⁹⁶ but we will use them to prove many important results.

6.5 Properties of Riemann Integration

Now we can present some familiar properties of the integral. Just like when we did this with limits and derivatives, most of these may be familiar, so examples won't be furnished for every single possible property.

Theorem 6.2. Suppose f and g are both bounded real functions on $[a, b]$ which are integrable. Then

1. $f + g$ is integrable, and

$$\int_a^b f(x) + g(x) dx = \int_a^b f(x) dx + \int_a^b g(x) dx;$$

2. cf is integrable for $c \in \mathbb{R}$, and

$$\int_a^b cf(x) dx = c \int_a^b f(x) dx;$$

3. (Monotonicity) if $f \leq g$ on $[a, b]$ we have

$$\int_a^b f(x) dx \leq \int_a^b g(x) dx.$$

4. (Additivity) for $c \in [a, b]$, f is integrable on $[a, c]$ and $[c, b]$, and

$$\int_a^c f(x) dx + \int_c^b f(x) dx = \int_a^b f(x) dx$$

5. if $|f(x)| \leq M$ on $[a, b]$ we have

$$\left| \int_a^b f(x) dx \right| \leq M(b - a)$$

The proofs of these are all really similar and monotonous, so fair warning.⁹⁷

Proof.

⁹⁶To be honest, most of real analysis seems like a random string of inequalities sometimes.

⁹⁷There is a reason that Rudin (1976) omits most of them.

1. Let P be a partition of $[a, b]$. We have

$$L(P, f) + L(P, g) \leq L(P, f + g) \leq U(P, f + g) \leq U(P, f) + U(P, g). \quad (13)$$

For $\varepsilon > 0$ there are partitions P_1 and P_2 such that

$$\begin{aligned} U(P_1, f) - L(P_1, f) &< \frac{\varepsilon}{2}, \\ U(P_2, g) - L(P_2, g) &< \frac{\varepsilon}{2}. \end{aligned}$$

If we take P^* to be the common refinement of P_1 and P_2 , then

$$\begin{aligned} U(P^*, f) - L(P^*, f) &< \frac{\varepsilon}{2}, \\ U(P^*, g) - L(P^*, g) &< \frac{\varepsilon}{2}. \end{aligned}$$

These inequalities combined with (13) gives $U(P^*, f + g) - L(P^*, f + g) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$, so $f + g$ is Riemann integrable. For this same P^* ,

$$\begin{aligned} U(P^*, f) &< \int_a^b f(x) dx + \frac{\varepsilon}{2}, \\ U(P^*, g) &< \int_a^b g(x) dx + \frac{\varepsilon}{2}. \end{aligned}$$

Again by (13),

$$\int_a^b f(x) + g(x) dx \leq U(P^*, f + g) \leq U(P^*, f) + U(P^*, g) < \int_a^b f(x) dx + \int_a^b g(x) dx + \varepsilon.$$

This holds for all $\varepsilon > 0$, so

$$\int_a^b f(x) + g(x) dx \leq \int_a^b f(x) dx + \int_a^b g(x) dx.$$

Now for the reverse direction all at once:

$$\begin{aligned} L(P^*, f) &> \int_a^b f(x) dx - \frac{\varepsilon}{2}, \\ L(P^*, g) &> \int_a^b g(x) dx - \frac{\varepsilon}{2}, \\ \int_a^b f(x) + g(x) dx &\geq L(P^*, f + g) \geq L(P^*, f) + L(P^*, g) > \int_a^b f(x) dx + \int_a^b g(x) dx - \varepsilon \\ \int_a^b f(x) + g(x) dx &\geq \int_a^b f(x) dx + \int_a^b g(x) dx. \end{aligned}$$

Therefore we have

$$\int_a^b f(x) + g(x) dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

2. First suppose $c \geq 0$. For any set $A \subset [a, b]$,

$$\begin{aligned} \sup_{x \in A} cf(x) &= c \sup_{x \in A} f(x), \\ \inf_{x \in A} cf(x) &= c \inf_{x \in A} f(x). \end{aligned}$$

This gives

$$\begin{aligned}\int_a^b f(x) dx &= \inf_{P \in \mathbf{P}([a,b])} U(P, cf) = c \inf_{P \in \mathbf{P}([a,b])} U(P, f) = c \int_a^b f(x) dx, \\ \underline{\int}_a^b f(x) dx &= \sup_{P \in \mathbf{P}([a,b])} L(P, cf) = c \sup_{P \in \mathbf{P}([a,b])} L(P, f) = c \int_a^b f(x) dx.\end{aligned}$$

Since f is Riemann integrable, its upper and lower Riemann integrals are equal. This establishes the integrability of f , as

$$c \int_a^b f(x) dx = c \bar{\int}_a^b f(x) dx = c \int_a^b f(x) dx.$$

Now let $c = -1$. In this case, we can't "factor" out a constant from a supremum and infimum. Instead, for any $A \subset [a, b]$, we have

$$\begin{aligned}\sup_{x \in A} -f(x) &= -\inf_{x \in A} f(x), \\ \inf_{x \in A} -f(x) &= -\sup_{x \in A} f(x).\end{aligned}$$

We will have $U(P, -f) = -L(P, f)$ and $L(P, -f) = -U(P, f)$ for any partition. This gives

$$\begin{aligned}\int_a^b -f(x) dx &= \inf_{P \in \mathbf{P}([a,b])} U(P, -f) = \inf_{P \in \mathbf{P}([a,b])} -L(P, f) = -\sup_{P \in \mathbf{P}([a,b])} L(P, f) = -\int_a^b f(x) dx, \\ \underline{\int}_a^b -f(x) dx &= \sup_{P \in \mathbf{P}([a,b])} L(P, -f) = \sup_{P \in \mathbf{P}([a,b])} -U(P, f) = -\inf_{P \in \mathbf{P}([a,b])} U(P, f) = -\int_a^b f(x) dx.\end{aligned}$$

Since f is Riemann integrable, its negative upper and lower Riemann integrals are equal, so

$$\int_a^b -f(x) dx = c \int_a^b -f(x) dx = c \int_a^b -f(x) dx.$$

In general, if $c < 0$, we can write it as $-1 \cdot |c|$ and apply the first two cases.

3. By the two previous parts, we can show that the integral of $g - f$ is not negative, as

$$\int_a^b g(x) dx + \int_a^b -f(x) dx = \int_a^b g(x) - f(x) dx.$$

We know $g - f \geq 0$, so $\inf_{x \in A}(f(x) - g(x)) \geq 0$ for any $A \subset [a, b]$. This means that for any P , $L(P, f) \geq 0$, so

$$\int_a^b g(x) - f(x) dx \geq L(P, f) \geq 0.$$

4. For all $\varepsilon > 0$, there exists a partition P such that $U(P, f) - L(P, f) < \varepsilon$. Define $P^* = P \cup \{c\}$ to be the refinement of P which results from adding c .⁹⁸ Let $Q = P^* \cap [a, c]$ and $R = P^* \cap [c, b]$ be partitions of $[a, b]$ and $[c, b]$ respectively. For these partitions

$$\begin{aligned}U(P^*, f) &= U(Q, f) + U(R, f), \\ L(P^*, f) &= L(Q, f) + L(R, f).\end{aligned}$$

⁹⁸It could be the case that $c \in P$. This would just mean that $P^* = P$.

We can conclude

$$\begin{aligned} U(Q, f) - L(Q, f) &= U(P^*, f) - L(P^*, f) - [U(R, f) - L(R, f)] \leq U(P, f) - L(P, f) < \varepsilon, \\ U(R, f) - L(R, f) &= U(P^*, f) - L(P^*, f) - [U(Q, f) - L(Q, f)] \leq U(P, f) - L(P, f) < \varepsilon. \end{aligned}$$

This shows that f is integrable on $[a, c]$ and $[c, b]$ by Riemann's Criterion.

We have

$$\begin{aligned} \int_a^b f(x) dx &\leq U(P, f) = U(Q, f) + U(R, f) < L(Q, f) + L(R, f) + \varepsilon < \int_a^c f(x) dx + \int_c^b f(x) dx + \varepsilon, \\ \int_a^b f(x) dx &\geq U(P, f) = U(Q, f) + U(R, f) > U(Q, f) + U(R, f) - \varepsilon > \int_a^c f(x) dx + \int_c^b f(x) dx - \varepsilon, \end{aligned}$$

which combine to give

$$\int_a^c f(x) dx + \int_c^b f(x) dx - \varepsilon < \int_a^b f(x) dx < \int_a^c f(x) dx + \int_c^b f(x) dx + \varepsilon.$$

If this for all $\varepsilon > 0$, then

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx$$

5. Treat M as a constant function on $[a, b]$. Example 6.1 showed that

$$\int_a^b M dx = M(b - a).$$

If $|f(x)| \leq M$, then $-M \leq f(x) \leq M$. By monotonicity,

$$-M(b - a) = \int_a^b -M dx \leq \int_a^b f(x) dx \leq \int_a^b M dx = M(b - a),$$

which can be written as

$$\left| \int_a^b f(x) dx \right| \leq M(b - a).$$

□

Example 6.7 (Linearity). The first 2 parts of Theorem 6.2 give that Riemann integration is linear. For any $c, d \in \mathbb{R}$, we will have

$$\int_a^b cf(x) + dg(x) dx = c \int_a^b f(x) dx + d \int_a^b g(x) dx.$$

This will turn out to be *very very* important in later sections. We will not be working with the Riemann integral than, but we'll see that integration in general is intrinsically linked to linearity.

While the Mean Value Theorem deals with derivatives, a similar result holds for integrals. It asserts that a continuous function must take on its average value on an interval.

Proposition 6.2 (Mean Value Theorem for Integrals I). Let f be a bounded real function on $[a, b]$. If f is continuous, then there exists a $c \in [a, b]$ such that

$$f(c) = \frac{1}{b - a} \int_a^b f(x) dx.$$

Proof. Since f is continuous, we can use the Extreme Value Theorem. The function f must attain a maximum M and minimum m on $[a, b]$. By monotonicity,

$$\begin{aligned}\int_a^b f(m) dx &\leq \int_a^b f(x) dx \leq \int_a^b f(M) dx, \\ f(m)(b-a) &\leq \int_a^b f(x) dx \leq f(M)(b-a), \\ f(m) &\leq \frac{1}{b-a} \int_a^b f(x) dx \leq f(M).\end{aligned}$$

By the Intermediate Value Theorem, f takes on every value in $[m, M]$, so there must be some c such that

$$f(c) = \frac{1}{b-a} \int_a^b f(x) dx.$$

□

The integral in Proposition 6.2 corresponds to the average value attained by the function on $[a, b]$. This proposition has a nice geometric interpretation if the equality is written as

$$f(c)(b-a) = \int_a^b f(d) dx.$$

As shown in Figure 63, we can find some rectangle of height $f(c)$ on the interval $[a, b]$ that equals the area under f . A second form of the Mean Value Theorem for Integrals is even more useful, as it allows us to find

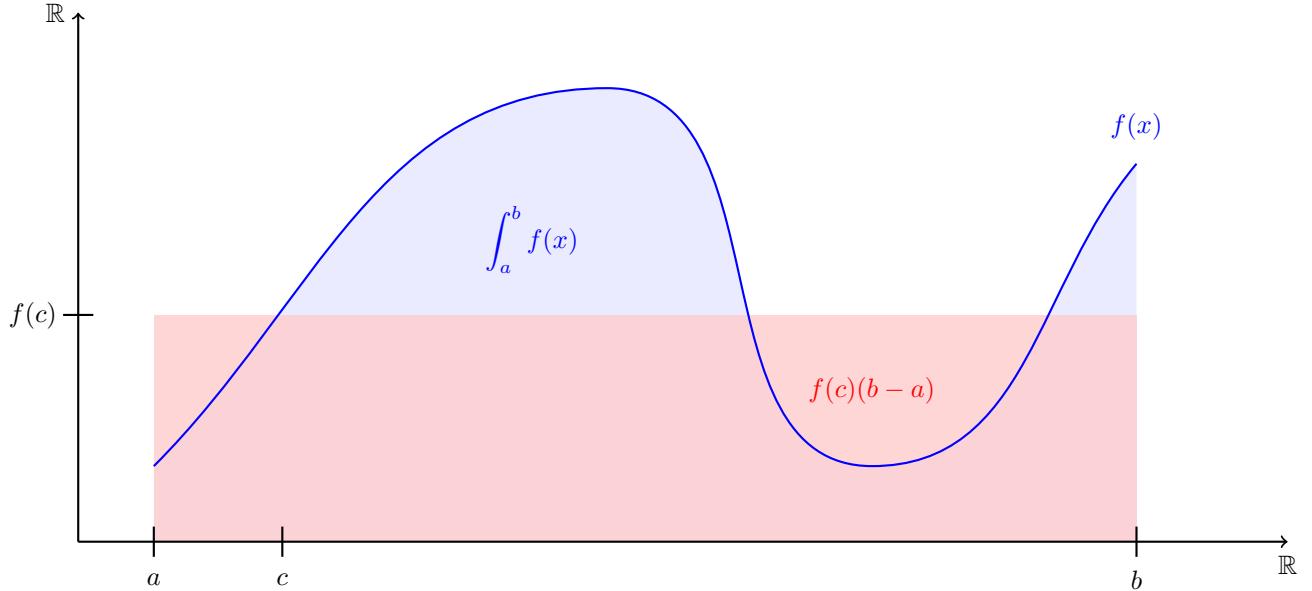


Figure 63: The Mean Value Theorem for Integrals says that we can find a value $c \in [a, b]$ such that the red rectangle is equal in area to the integral shown in blue.

a value in $[a, b]$ that allows us to “factor” a function out of the integral of a product.

Proposition 6.3 (Mean Value Theorem for Integrals II). Let f be a bounded real function on $[a, b]$. If f is continuous, and g is integrable function $[a, b]$ and does not change signs, then there exists a $c \in [a, b]$ such that

$$\int_a^b f(x)g(x) dx = f(c) \int_a^b g(x) dx.$$

Proof. If $g(x) = 0$ for all $x \in [a, b]$, then the result holds trivially. Assume instead that $g(x) > 0$. Since f is continuous, we can use the Extreme Value Theorem. The function f must attain a maximum M and minimum m on $[a, b]$.

$$\begin{aligned} f(m) &\leq f(x) \leq f(M), \\ f(m)g(x) &\leq f(x)g(x) \leq f(M)g(x). \end{aligned}$$

Multiply by $g(x)$ will not change the inequality, as $g(x) > 0$ for all $x \in [a, b]$.⁹⁹ By monotonicity and the linearity of integration,

$$\begin{aligned} f(m)g(x) &\leq f(x)g(x) \leq f(M)g(x), \\ \int_a^b f(m)g(x) \, dx &\leq \int_a^b f(x)g(x) \, dx \leq \int_a^b f(M)g(x) \, dx, \\ f(m) \int_a^b g(x) \, dx &\leq \int_a^b f(x)g(x) \, dx \leq f(M) \int_a^b g(x) \, dx. \end{aligned}$$

If we divide by the integral of g , then

$$f(m) \leq \frac{\int_a^b f(x)g(x) \, dx}{\int_a^b g(x) \, dx} \leq f(M).$$

By the continuity of f and the Intermediate Value Theorem, there exists a $c \in \mathbb{R}$ such that

$$f(c) = \frac{\int_a^b f(x)g(x) \, dx}{\int_a^b g(x) \, dx},$$

which can be expressed as

$$\int_a^b f(x)g(x) \, dx = f(c) \int_a^b g(x) \, dx.$$

As mentioned in Footnote 99, the case where $g < 0$ is virtually the same. The only difference is that all the inequalities are reversed, but we're still able to apply the Intermediate Value Theorem. \square

Example 6.8. For Proposition 6.3, let $g(x) = 1$ on $[a, b]$. In this case we recover Proposition 6.2.

$$\begin{aligned} \int_a^b f(x)g(x) \, dx &= f(c) \int_a^b g(x) \, dx, \\ \int_a^b f(x) \cdot 1 \, dx &= f(c) \int_a^b 1 \, dx, \\ \int_a^b f(x)g(x) \, dx &= f(c)(b - a). \end{aligned}$$

This observation shows that Proposition 6.3 is a generalized version of its predecessor.

Example 6.9 (What if g Changes Sign?). Proposition 6.3 stipulates that the function g cannot change sign on $[a, b]$. This is essential. If this is not the case, then the inequality

$$f(m)g(x) \leq f(x)g(x) \leq f(M)g(x)$$

⁹⁹The other case where $g(x) < 0$ would reverse the order of the inequality, but that is fine. What's important is that we have an upper and lower bound on $f(x)g(x)$, not what those bounds happen to be.

will not hold for all $x \in [a, b]$. Suppose $g(x) = x$ on $[-1, 1]$, and $f(x) = x$. In this case

$$f(x) \int_{-1}^1 g(x) dx = 0$$

for all x , but $f(x)g(x) = x^2$, so

$$\int_{-1}^1 f(x)g(x) dx \neq 0.$$

Therefore Proposition 6.3 does not hold.

6.6 Riemann Integration and Continuity

With the Riemann Criterion in hand, we're able to prove that a function is integrable. We will now use it to show that two very general classes of functions are integrable, the first being continuous functions.

The fact that all continuous functions (on $[a, b]$) are integrable should not come as a surprise. Not only is nearly every function integrated in a calculus course continuous, but continuity and integrability are both the results of limiting behavior with an arbitrary $\varepsilon > 0$ (the latter being the case as a result of Riemann's Criterion).

Theorem 6.3 (Continuity implies Integrability). If f is a real continuous function on $[a, b]$, then f is Riemann integrable on $[a, b]$.

Proof. First note that a continuous function on $[a, b]$ is in fact bounded on $[a, b]$ by the Extreme Value Theorem.¹⁰⁰ Furthermore, $[a, b]$ is compact (Heine-Borel), so f is uniformly continuous on $[a, b]$ (Theorem 4.8). For all $\varepsilon > 0$, there exists a δ such that

$$|f(x) - f(y)| < \frac{\varepsilon}{n(b-a)},$$

for all $x, y \in [a, b]$ which satisfy $|x - y| < \delta$. Choose a partition P of $[a, b]$ such that $|x_i - x_{i-1}| < \delta$ for $i = 1, \dots, n$.¹⁰¹

Our function is continuous, so it achieves its maximum and minimum on each $[x_{i-1}, x_i] \subset P$, but in this case those are M_i and m_i . We therefore have

$$|f(s_i) - f(t_i)| < |M_i - m_i| < \frac{\varepsilon}{b-a}$$

for $|s_i - t_i| < \delta$, where $f(s_i) = M_i$ and $f(t_i) = m_i$.¹⁰² This allows us to verify that the Riemann Criterion holds for the partitions f :

$$U(P, f) - L(P, f) = \sum_{i=1}^n M_i \Delta x_i - \sum_{i=1}^n m_i \Delta x_i = \sum_{i=1}^n (M_i - m_i) \Delta x_i < n \cdot \frac{\varepsilon}{n(b-a)} \cdot (b-a) < \varepsilon.$$

□

Example 6.10 (Integrable But Not Continuous). The converse of Theorem 6.3 is not true. Example 6.6 showed a function that failed to be continuous on $[0, 1]$ but is still integrable on $[0, 1]$.

¹⁰⁰This is important to point out because we only can define the Riemann integral for bounded real functions.

¹⁰¹For example, you could take form P with n intervals of length $(b-a)/n$ for $n > (b-a)/\delta$

¹⁰²That is, s_i and t_i are where the function achieves its maximum M_i and minimum m_i respectively.

Proposition 6.4 (Continuous Composition Preserves Integrability). Suppose f is a bounded real function on $[a, b]$, $m \leq f \leq M$, ϕ is continuous on $[m, M]$, and $h(x) = \phi \circ f(x) = \phi(f(x))$ on $[a, b]$. Then h is integrable on $[a, b]$.

For the proof of this result, it's important to remember that the domain of ϕ is the range of f . It can be easy to get bogged down with the notation in this one. This also will be the most intense proof involving ε yet.

Proof. Let $\varepsilon > 0$. The function ϕ is continuous on the compact set $[a, b]$, so it is uniformly continuous. There exists a $\delta > 0$ such that $\delta < \varepsilon$ and

$$|\phi(s) - \phi(t)| < \frac{\varepsilon}{(b-a) + 2 \sup_{v \in [m, M]} |\phi(v)|}$$

whenever $|s - t| < \delta$ for $s, t \in [m, M]$.¹⁰³

Since f is integrable, there is a partition $P = \{x_0, \dots, x_n\}$ on $[a, b]$ for which

$$U(P, f) - L(P, f) < \delta^2 \quad (14)$$

by Riemann's Criterion. This partition gives a partition of $[m, M]$ in $P^* = \{m = f(x_0), f(x_1), \dots, f(x_n) = M\}$.¹⁰⁴ Let

$$M_i = \sup_{x \in [x_{i-1}, x_i]} f(x), \quad m_i = \inf_{x \in [x_{i-1}, x_i]} f(x), \quad M_i^* = \sup_{t \in [f(x_{i-1}), f(x_i)]} \phi(t), \quad m_i^* = \inf_{t \in [f(x_{i-1}), f(x_i)]} \phi(t).$$

Figure 64 shows what a partition M_i^* and m_i^* will look like. We can divide the indices for the partition P

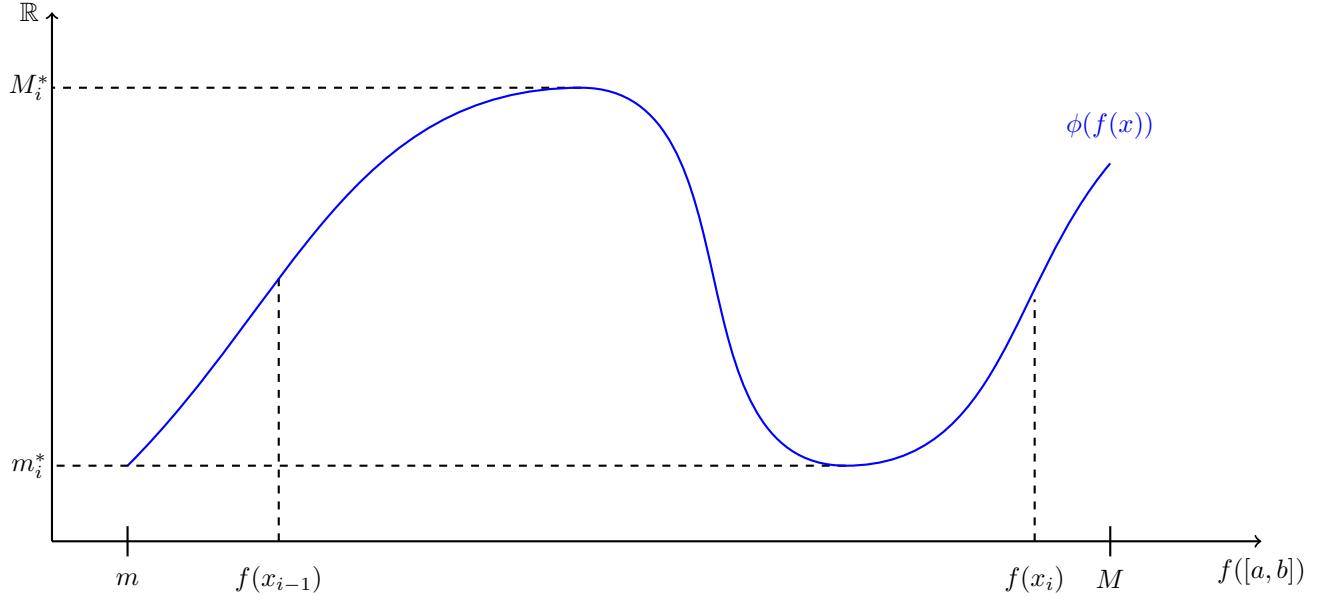


Figure 64: An interval $[f(x_{i-1}), f(x_i)] \subset P^*$, and the values M_i^* and m_i^* .

¹⁰³Woah, what is up with us saying $\delta < \varepsilon$! Well if the inequalities hold for δ , then it should hold for any number less than $\delta' = \min\{\delta, \varepsilon\}$. If we have $\delta' < \varepsilon$, then $\delta' = \delta$. Therefore we lost no generality in saying $\delta' < \varepsilon$. We simply introduce it as δ instead of explicitly going through this though process with δ' .

¹⁰⁴This is the partition that we will use with $\phi : [m, M] \rightarrow \mathbb{R}$.

into two sets

$$\begin{aligned} A &= \{i \mid M_i - m_i < \delta\}, \\ B &= \{i \mid M_i - m_i \geq \delta\}. \end{aligned}$$

If $i \in A$, then by the uniform continuity of ϕ ,

$$|\phi(M_i) - \phi(m_i)| = |M_i^* - m_i^*| < \varepsilon.$$

If $i \in B$ then

$$|M_i^* - m_i^*| \leq 2K,$$

for $K = \sup |\phi(t)|$ and $m \leq t \leq M$.¹⁰⁵ By (14),

$$\begin{aligned} U(P, f) - L(P, f) &= \sum_{i=1}^n M_i \Delta x_i - \sum_{i=1}^n m_i \Delta x_i = \sum_{i=1}^n (M_i - m_i) \Delta x_i < \delta^2 \\ \implies \sum_{i \in A} (M_i - m_i) \Delta x_i + \sum_{i \in B} (M_i - m_i) \Delta x_i &< \delta^2 \\ \implies \sum_{i \in B} \underbrace{(M_i - m_i)}_{\geq \delta} \Delta x_i &< \delta^2 \\ \implies \delta \sum_{i \in B} \Delta x_i &\leq \sum_{i \in B} (M_i - m_i) \Delta x_i < \delta^2 \\ \implies \sum_{i \in B} \Delta x_i &< \delta. \end{aligned}$$

This final inequality allows us to show the Riemann condition holds for $\phi(f(x))$, as required:

$$\begin{aligned} U(P, \phi(f)) - L(P, \phi(f)) &= \sum_{i \in A} \underbrace{(M_i^* - m_i^*)}_{< \varepsilon} \Delta x_i + \sum_{i \in B} \underbrace{(M_i^* - m_i^*)}_{2K} \underbrace{\Delta x_i}_{< \delta} \\ &< \frac{\varepsilon}{(b-a) + \sup |\phi(t)|} (b-a) + 2K \underbrace{\delta}_{< \frac{\varepsilon}{(b-a) + 2 \sup |\phi(t)|}} \\ &< \frac{\varepsilon}{(b-a) + 2K} [(b-a) + 2K] \\ &= \varepsilon. \end{aligned}$$

□

Along with keeping track of if we're working in $[a, b]$ or $[m, M]$, I think another part of this proof that is not clear is how to handle the case where $M_i - m_i \geq \delta$. Splitting the partition of $[a, b]$ into two parts, one corresponding to the case where $M_i - m_i < \delta$ (indexed by A), the other where $M_i - m_i \geq \delta$ (indexed by B), is a pretty clear answer, but it goes against a lot on the instincts developed when doing analysis. Many proofs require you to show some inequality holds in every possible case, so if you were trying to develop this proof on your own, when you realize $M_i - m_i \geq \delta$ for some i , it may seem like you're doing something wrong.

Proposition 6.4 allows us to prove two additional properties of integration.

¹⁰⁵In the context of Figure 64, this inequality makes a bit more sense. The distance between the sup and inf of $\phi(t)$ on the interval $[f(x_{i-1}), f(x_i)]$ must be less than twice the absolute value of the sup and of ϕ on the whole domain being partitioned $[m, M]$. If $m_i^* = -M_i^*$, then $|M_i^* - m_i^*| = 2M_i^*$. If M_i^* happens to be the supremum on all of $[m, M]$, that is $K = M_i^*$, then we have the equality $|M_i^* - m_i^*| = 2K$. Understanding this inequality may be the toughest part of this proof, and it's worth drawing some pictures if it is not clear.

Proposition 6.5. Suppose f and g are both bounded real functions on $[a, b]$ which are integrable. Then fg is integrable on $[a, b]$.

Proof. By Theorem 6.2, $f \pm g$ is integrable. Proposition 6.4 tells us that $\phi(f \pm g)$ will be integrable as long as ϕ is continuous. If we let $\phi(t) = t^2$, then

$$\begin{aligned}\phi(f - g) &= (f - g)^2, \\ \phi(f + g) &= (f + g)^2,\end{aligned}$$

are integrable. Again by Theorem 6.2, their difference scaled should be integrable.

$$fg = \frac{(f + g)^2 + (f - g)^2}{4}$$

This gives that fg is integrable. □

The converse of Proposition 6.5 is not true. Being able to factor an integrable function does not guarantee the factors are integrable as the next example shows.

Example 6.11. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ -1 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

This function is not integrable on $[0, 1]$ (see Subsection 6.11), but f^2 is, as $f^2(x) = 1$ for all $x \in \mathbb{R}$.

Proposition 6.6. Suppose f is bounded real functions on $[a, b]$ which is integrable. Then $|f|$ is integrable on $[a, b]$, and

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f| dx.$$

Proof. Proposition 6.4 tells us that $\phi(f)$ will be integrable as long as ϕ is continuous, so let $\phi(t) = |t|$. This shows that $\phi(f) = |f|$ is integrable. Choose $c \in \{-1, 1\}$ such that

$$c \int_a^b f(x) dx \geq 0.$$

We have $cf \leq |f|$, so by the monotonicity and linearity of integration,

$$\left| \int_a^b f(x) dx \right| = c \int_a^b f dx = \int_a^b cf dx \leq \int_a^b |f| dx.$$

□

Example 6.12 (Which Direction is the Inequality?). If you need to cram for an exam, you're not too concerned with having a deep understanding of propositions and theorems, so remembering the direction of the inequality of Proposition 6.6 may come down to memorization.¹⁰⁶ In this case just remember the simple examples of $f(x) = x$ on the interval $[-1, 1]$. In this case we have

$$\left| \int_{-1}^1 x dx \right| = 0 \leq 1 = \int_{-1}^1 |x| dx = 0.$$

¹⁰⁶This *may* be slightly autobiographical.

6.7 Riemann Integration, Monotonicity, and Discontinuities

The second class of functions that are integrable are monotonic functions. The integrability on continuous functions was not a surprise, but this is. Proposition 4.2 limited the number of discontinuities of a monotonic function to a countably infinite number, but this is still an infinite number. The fact that such a function could feasibly be integrated is not something we should have expected.

Proposition 6.7. If f is monotonic on $[a, b]$, then f is integrable.

Proof. Let $\varepsilon > 0$. For any $n \in \mathbb{N}$, define a partition of $[a, b]$ such that

$$\Delta x_i = \frac{b - a}{n}.$$

We will show the result for a monotonically increasing function. Because the function is monotonically increasing, $M_i = f(x_i)$, and $m_i = f(x_i)$.¹⁰⁷ We have

$$\begin{aligned} U(P, f) - L(P, f) &= \sum_{i=1}^n (M_i - m_i) \Delta x_i \\ &= \sum_{i=1}^n (f(x_i) - f(x_{i-1})) \frac{b - a}{n} \\ &= \frac{b - a}{n} \sum_{i=1}^n (f(x_i) - f(x_{i-1})) \end{aligned}$$

By the Archimedean Property of \mathbb{R} , we can find an n large enough such that

$$U(P, f) - L(P, f) = \frac{b - a}{n} \sum_{i=1}^n (f(x_i) - f(x_{i-1})) < \varepsilon$$

for all $\varepsilon > 0$. Because this n corresponds to a choice of a partition, we have found a partition that gives $U(P, f) - L(P, f) < \varepsilon$, so Riemann's Criterion is satisfied. \square

Example 6.13. FINISH

Proposition 6.8. Suppose f is bounded on $[a, b]$ and has only finitely many points of discontinuity on $[a, b]$. Then f is Riemann integrable on $[a, b]$.

Proof. Let $\varepsilon > 0$. Set $M = \sup |f(x)|$, and $E = \{x \in [a, b] \mid f \text{ discontinuous}\}$. The set E is finite by assumption, so we can cover E by finitely many disjoint intervals $[u_j, v_j] \subset [a, b]$.¹⁰⁸ We can pick these intervals such that the sum of their lengths is ε :

$$\sum_{j=1}^m v_j - u_j < \frac{\varepsilon}{(b - a) + 2M}.$$

Now we will remove the open segments (u_j, v_j) from $[a, b]$, giving a set K .

$$K = [a, b] \setminus \bigcup_{j=1}^m (u_j, v_j)$$

¹⁰⁷If f is always increasing, then the infimum on $[x_{i-1}, x_i]$ will be achieved at smallest value in the interval, namely x_{i-1} . Similarly, the supremum on the interval will be achieved at largest value in the interval, namely x_i .

¹⁰⁸So if f is discontinuous at x_0 , we form a little interval of length ε around it such that no other point in the small interval is a point of discontinuity.

Because we removed open segments, K is still closed, so it is compact.¹⁰⁹ On the compact set K , f is continuous,¹¹⁰ giving uniform continuity. There exists a $\delta > 0$ such that

$$|f(s) - f(t)| < \frac{\varepsilon}{(b-a) + 2M}$$

for any $s, t \in K$ satisfying $|s - t| < \delta$.

Now construct a partition $P = \{x_0, \dots, x_n\}$ of $[a, b]$ such that: $u_j \in P$ for all j , $v_j \in P$ for all j , $(u_j, v_j) \cap P = \emptyset$. We will have $\Delta x_i < \delta$, as $x_i, x_{i-1} \in K$...FINISH

□

Remark 6.4 (Integrability is Weaker than Continuity). Integrability is often seen as equivalent to continuity in calculus courses, but this section goes to show this is not at all correct. Any monotonic function is integrable, even though it may have countably infinite discontinuities. Any bounded function with finite points of discontinuity is bounded. These facts make integrability a far weaker condition than continuity.

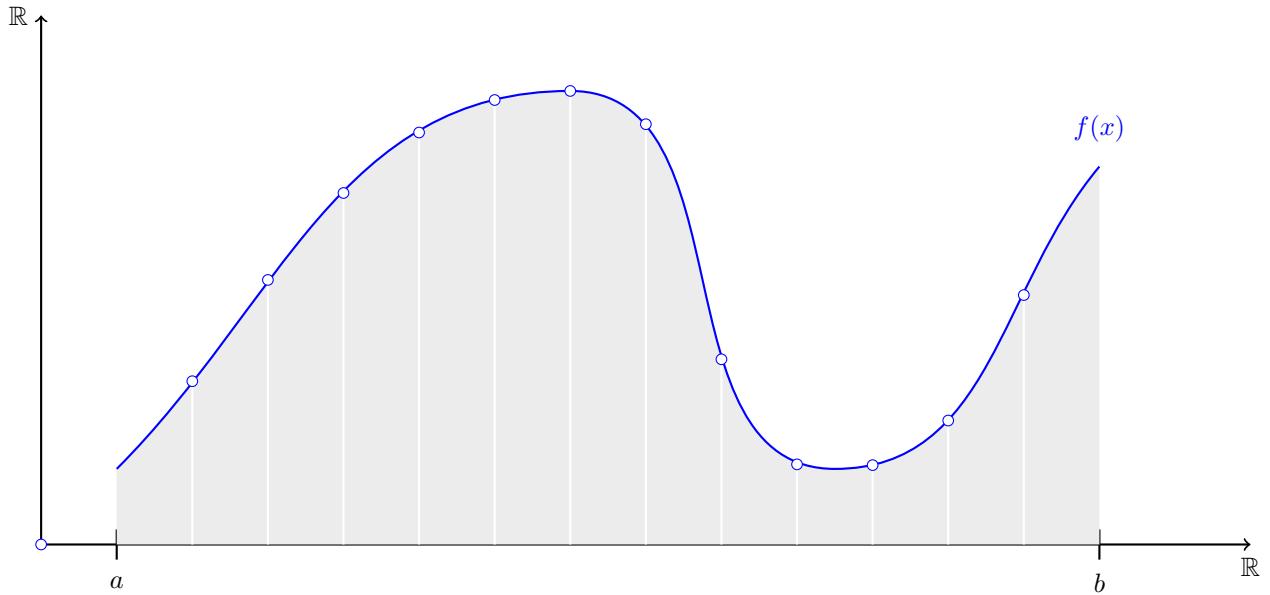


Figure 65: Despite having 12 points of discontinuity on $[a, b]$, f is still integrable.

Example 6.14 (Simple Functions). Any simple function (Definition 6.6) is Riemann integrable, despite them often being discontinuous. Example 6.3 and Figure 61 show an example of this type of discontinuous, yet integrable, function.

6.8 The Riemann-Stieltjes Integral (Optional)

Disclaimer: Again, we need to be a bit informal here. The purpose of this section is to build intuition for Sections 13-15, so I'm not as worried about using proper verbiage. Again, I'll be assuming that the “actual length”, or simply “length”, of an interval $[x_{i-1}, x_i]$ is $|x_i - x_{i-1}|$.

¹⁰⁹Removing segments never put the boundedness of $[a, b]$ in jeopardy, so clearly K is bounded.

¹¹⁰Remember, we just removed all the point pf discontinuity by removing (u_j, v_j) .

If you think back to the definition of the upper and lower Riemann sums, we always treated the width of the rectangle at $[x_{i-1}, x_i] \subset P$ as $|x_{i-1} - x_i|$. Is there anything stopping us from defining the width of the corresponding rectangle as $|2x_{i-1} - 2x_i|$? Or what about $|x_{i-1}^2 - x_i^2|$? It turns out, we can still perform Riemann integration if we decide to assign different widths to the rectangles of a Riemann sum.

Let $\alpha : [a, b] \rightarrow \mathbb{R}$ be any monotonic function. We can redefine the upper and lower Riemann sums as

$$U(P, f, \alpha) = \sum_{i=1}^n M_i \Delta\alpha(x_i),$$

$$L(P, f, \alpha) = \sum_{i=1}^n m_i \Delta\alpha(x_i),$$

where $\Delta\alpha(x_i) = |\alpha(x_i) - \alpha(x_{i-1})|$. The function α is determining how we assign length to different intervals in P . We can think of it as a *weighted* version of the original upper and lower sums, where α determine the weight assigned to different parts of $[a, b]$, and each interval is assigned an “artificial length” of $|\alpha(x_i) - \alpha(x_{i-1})|$. By “weight”, I’m talking about how much “artificial” length we assign to an interval relative to its “actual” length. It is meant in the same sense as a “weighted” average, where you take some observations to be more significant, and assign them more weight when calculating the mean. This clearly merits several examples.

Example 6.15 ($\alpha(x) = x$). Let $\alpha(x) = x$. In this case we have

$$\alpha(\Delta x_i) = |\alpha(x_i) - \alpha(x_{i-1})| = |x_i - x_{i-1}|,$$

so each interval is assigned its “true” length (Figure 66). This is a special case for two reasons. Firstly, each

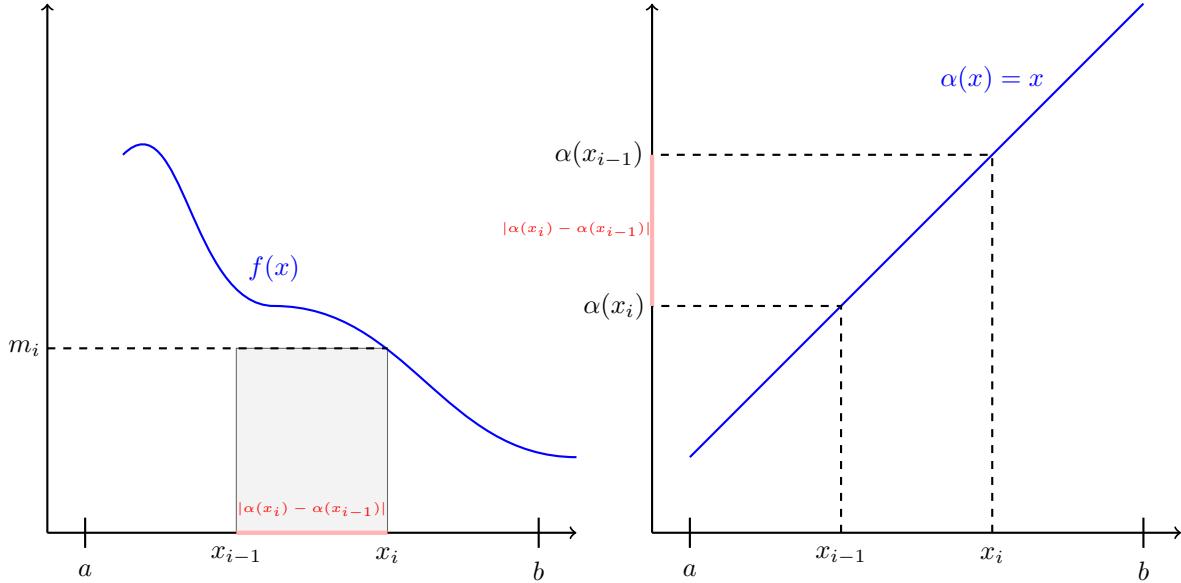


Figure 66: To determine how we weight the interval $[x_{i-1}, x_i]$, we calculate $|\alpha(x_i) - \alpha(x_{i-1})|$, and then assign it to the interval.

interval is weighted equally.¹¹¹ Secondly, each interval is weighted such that $|\alpha(x_i) - \alpha(x_{i-1})| = |x_i - x_{i-1}|$, so we get back the original Riemann sums. The weighted area assigned to the gray rectangle is

$$m_i \Delta\alpha(x_i) = m_i \Delta x_i.$$

¹¹¹Meaning that $|\alpha(x_i) - \alpha(x_{i-1})| \propto |x_i - x_{i-1}|$ for all i .

Example 6.16 ($\alpha(x) = 2x$). Let $\alpha(x) = 2x$. In this case we have

$$\alpha(\Delta x_i) = |\alpha(x_i) - \alpha(x_{i-1})| = |2x_i - 2x_{i-1}|,$$

so each interval is assigned twice its “actual” length (Figure 67). Now we have

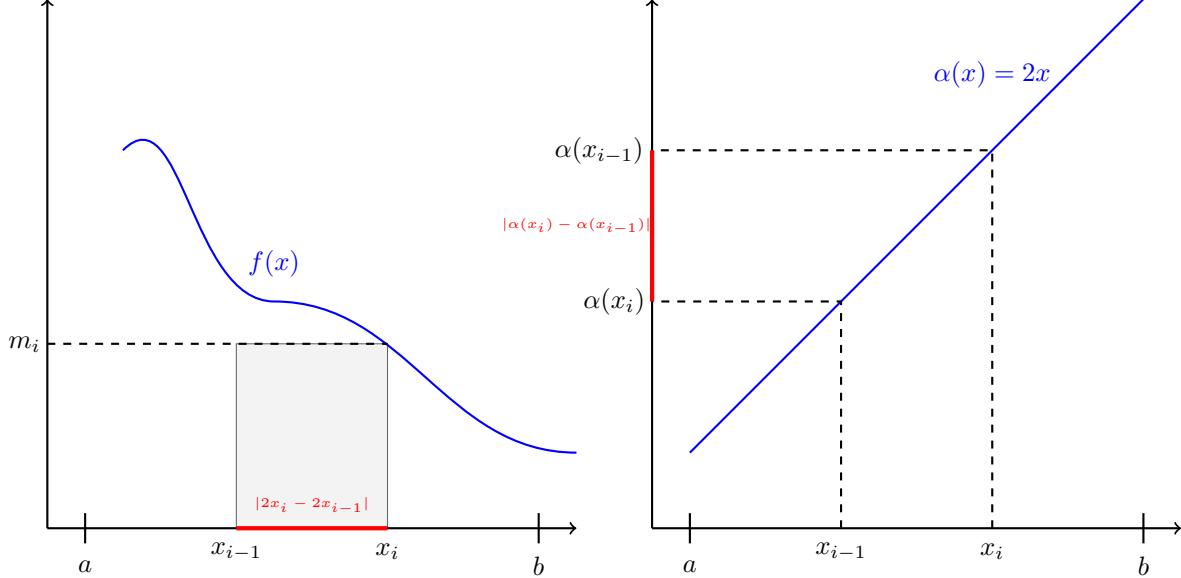


Figure 67: We now assign more weight to $[x_{i-1}, x_i]$ than we did with $\alpha(x) = x$.

$$m_i \Delta \alpha(x_i) = 2m_i \Delta x_i.$$

This is still a somewhat special case though, because we are assigning “artificial” lengths that are proportional to the “actual” length of an interval. For example, this choice of α gives $|\alpha(1) - \alpha(0)| = 2$ and $|\alpha(2) - \alpha(0)| = 4$. In both cases, the “artificial” length is double the “actual” length. In fact, if $\alpha(x)$ is linear, then each interval is weighted equally. For $\alpha(x) = \beta x + \gamma$,

$$\Delta \alpha(x_i) = |(\beta \Delta x_{i-1} + \gamma)| - |(\beta \Delta x_i + \gamma)| = \beta |\Delta x_{i-1}| = \beta \Delta x_i \propto \Delta x_i,$$

so all we’re doing is scaling the “actual” length.

Example 6.17. $\alpha(x) = x^2$ Now let’s see what happens for a nonlinear choice of $\alpha(x)$. Suppose we have two intervals in a partition, $[0, 0.1]$ and $[1, 10]$. With this choice of α we have

$$|\alpha(0.1) - \alpha(0)| = 0.01,$$

$$|\alpha(10) - \alpha(1)| = 81.$$

This value of α assigns a lot of weight to large intervals, and a small amount of weight to small intervals. This is the first example we’ve seen where α truly assigns different weights to different intervals, and not just scaling up every length by a constant.

Using α , we can redefine the Riemann integral as developed in Subsection 6.2.

$$\int_a^b f(x) d\alpha = \sup_{P \in \mathbf{P}([a,b])} L(P, f, \alpha), \quad \bar{\int}_a^b f(x) d\alpha = \inf_{P \in \mathbf{P}([a,b])} U(P, f, \alpha).$$

Definition 6.10. Suppose f is a bounded real function on the interval $[a, b]$ and α is a monotonic function on $[a, b]$. If

$$\int_a^b f(x) \, d\alpha = \bar{\int}_a^b f(x) \, d\alpha,$$

then we say f is *Riemann–Stieltjes integrable (on $[a, b]$)* and we write the common value of the upper and lower Riemann integral as

$$\int_a^b f(x) \, d\alpha.$$

We refer to this common value as the *Riemann–Stieltjes integral of f on $[a, b]$* .

The Riemann–Stieltjes (RS) is simply a weighted integral. Nearly all the results developed up until now can be established for the RS integral. In fact many standard texts such as [Rudin \(1976\)](#) opt to work exclusively with the RS integral for this reason.

Remark 6.5 (Why Monotonic?). If α is not monotonic, then it could assign negative “artificial length” to an interval $[x_{i-1}, x_i]$ (which requires $x_i > x_{i-1}$).

Example 6.18 (Probability Distributions). A random variable X has a cumulative distribution function (CDF) $F : [a, b] \rightarrow [0, 1]$ such that

$$\Pr(c \leq X \leq d) = F(d) - F(c).$$

The CDF is assigning a probability to $[c, d]$. Depending on our choice of F , the weight assigned to each interval in probability will differ. If F is a uniform distribution, every interval is assigned equal probability, i.e equal weight. But this sounds *just like* RS integration! If we let $\alpha(x) = F(x)$, then we have

$$\int_c^d dF = \int_c^d 1 \, dF = 1(F(d) - F(c)) = \Pr(c \leq X \leq d).$$

Now suppose that the random variable X only takes on values $x \in [a, b]$. If we want to find the expected value of X over, then we calculate a the RS integral of x over all of $[a, b]$.

$$E[x] = \int_a^b x \, dF$$

This is akin to a weighted average, because we are averaging every single possible realization of X , but we first assign weight to the interval $[a, b]$ in accordance with the probability of each event occurring.

Example 6.19 (Inertia). Suppose there is a rod of unit length where the mass contained in $[0, x]$ is given by $m(x)$. The inertia I is given as

$$I = \int_0^1 x^2 \, dm.$$

If you have taken a probability or physics course, these last two examples may look somewhat familiar. Using integration to calculate the expected value of a random variable, or the moment of inertia are standard applications, but it is normally introduced as a Riemann integral of a different quantity, not a RS integral. What is the connection between the two? Is it possible to calculate a RS integral as a Riemann integral? The next theorem allows us to do just this.

Theorem 6.4 (RS Integral as Riemann Integral). Assume α increase monotonically and α' is integrable on $[a, b]$. Let f be a bounded real function on $[a, b]$. The function f is RS integrable on $[a, b]$ if and only if $f\alpha'$ is Riemann integrable on $[a, b]$. In this case we have

$$\int_a^b f(x) d\alpha = \int_a^b f(x)\alpha'(x) dx.$$

Before we prove this, we should rationalize why this equation holds. We want to integrate $f(x)$ as if the interval $[a, b]$ is not weighted. To do this we need to convert between $\alpha(x)$ and x somehow. For the RS sums, $\alpha(x)$ is manipulating the width of a rectangle, thereby manipulating its area. But couldn't we instead manipulate the height of the rectangle to achieve the same change in area? If we were to do this, we would need to multiply the height of the rectangle by the same factor that we change the width by. But what quantity captures the change in width? That would be α' . What determines the height of the rectangle? That is f (vicariously through its supremum or infimum). So if we let the height of the rectangle be $f(x)\alpha'(x)$, we achieve the same area that resulted from reweighting the width using α .

Let's accompany this argument with a quick example. Using a constant function would be best, as the integral in this case just amounts to a rectangle. Let $f(x) = 10$ on $[5, 10]$, and $\alpha(x) = 2x$. If we calculate the area using a RS integral, we have

$$\int_5^{10} f(x) d\alpha = \int_5^{10} 10 d\alpha = 10[2(10) - 2(5)] = 100.$$

Alternatively, we could achieve the same area under $f(x)$ by multiply $f(x)$ by 2, which just happens to be α' (which we are about to prove is no coincidence).

$$\int_5^{10} f(x)\alpha'(x) d\alpha = \int_5^{10} 10 \cdot 2 dx = 20(10 - 5) = 100.$$

We get the same exact value. The actual proof, which is on the harder side, is not nearly as important as understanding the intuition behind the result.

Proof. The function α is integrable, so it satisfies the Riemann Criterion. That is for $\varepsilon > 0$ there exists a partition $P = \{x_0, \dots, x_n\}$ of $[a, b]$ such that

$$U(P, \alpha') - L(P, \alpha') < \frac{\varepsilon}{\sup |f(x)|}.$$

By the differentiability of α' , we can apply the Mean Value Theorem on each interval $[x_{i-1}, x_i]$. There exists a $t_i \in [x_{i-1}, x_i]$ such that

$$\begin{aligned} \alpha_i(x_{i-1}) - \alpha_i(x_i) &= \alpha'(t_i)(x_{i-1} - x_i), \\ \Delta\alpha_i &= \alpha'(t_i)\Delta x_i, \end{aligned} \tag{15}$$

for $i = i, \dots, n$. By Part 2 of Proposition 6.1 and the integrability of α' , for $s_i \in [x_{i-1}, x_i]$,¹¹²

$$\sum_{i=1}^n |\alpha'(s_i) - \alpha'(t_i)|\Delta x_i < \frac{\varepsilon}{|\sup f(x)|}. \tag{16}$$

¹¹²Part 2 of Proposition 6.1 requires two points $s_i, t_i \in [x_{i-1}, x_i]$. We will use the t_i that we found using the Mean Value Theorem.

By (15),

$$\sum_{i=1}^n f(s_i) \Delta \alpha_i = \sum_{i=1}^n f(s_i) \alpha'(t_i) \Delta x_i. \quad (17)$$

If we set $M = \sup |f(x)|$, then (17) and (16) give

$$\left| \underbrace{\sum_{i=1}^n f(s_i) \Delta \alpha_i}_{U(P, f, \alpha)} - \underbrace{\sum_{i=1}^n f(s_i) \alpha'(t_i) \Delta x_i}_{U(P, f \alpha')} \right| \leq M \varepsilon.$$

This will hold for all $s_i \in [x_{i-1}, x_i]$, so

$$|U(P, f, \alpha) - U(P, f \alpha')| < M \frac{\varepsilon}{|f(x)|} = \varepsilon.$$

This all will hold for any refinement of P , so we can conclude

$$\left| \int_a^b f(x) d\alpha - \int_a^b f(x) \alpha'(x) dx \right| \leq \varepsilon.$$

This holds for all $\varepsilon > 0$, so

$$\int_a^b f(x) d\alpha = \int_a^b f(x) \alpha'(x) dx.$$

This *exact* reasoning will give

$$\int_a^b f(x) d\alpha = \int_a^b f(x) \alpha'(x) dx,$$

so we can conclude

$$\int_a^b f(x) d\alpha = \int_a^b f(x) \alpha'(x) dx.$$

□

Example 6.20 (Probability Distributions). A random variable X with CDF F , has a probability density function (PDF) defined as $F' = f$. If we apply Theorem 6.4 to Example 6.16 we get

$$\begin{aligned} \Pr(c \leq X \leq d) &= \int_c^d dF = \int_c^d F'(x) dx = \int_c^d f(x) dx, \\ E[x] &= \int_a^b x dF = \int_a^b x F'(x) dx = \int_a^b x f(x) dx. \end{aligned}$$

This is the last time we'll see the RS integral, but the ideas which motivate it will return.

6.9 The Fundamental Theorem Of Calculus and Consequences

Wait...how do we actually calculate these things? So far we've only been able to integrate constant functions (Example 6.1).¹¹³ This section will present *the* theorem of calculus. The Fundamental Theorem of Calculus allows us to easily calculate integrals by the means of derivatives. But is this what we really care about? From the standpoint of analysis, who cares about calculating the area under a curve? Mathematicians

¹¹³If we think of the integral in terms of simple functions, we can integrate those too, but only because we defined the integral explicitly in that case.

have been able to do this quite accurately for millennial. What is *amazing* is that the area under a curve is somehow related to differentiation at all. Despite being able to calculate areas and rates for centuries, mathematicians thought these two tasks were totally unrelated. This is to say, *everyone* takes this theorem for granted. When introduced to integrals in calculus, most are taught that integration is the opposite of differentiation only minutes after the integral is defined, but this is not a very obvious of a relationship.¹¹⁴ Math students are so conditioned to think of integration hand-in-hand with differentiation, that it makes it hard to appreciate the Fundamental Theorem from the standpoint of analysis. **So do not think of this section as a means to calculation, think of it as a beautiful insight into two seemingly disparate topics.** Pontification over.

For such an important pair of results, The Fundamental Theorem of Calculus is not overwhelmingly difficult to prove. The second part is just an application of the Mean Value Theorem,¹¹⁵

Definition 6.11. Let F be a real differentiable function. If $F' = f$, then we say F is the **antiderivative** of f .

Theorem 6.5 (Fundamental Theorem of Calculus I). Let f be integrable on $[a, b]$. For $a \leq x \leq b$, let

$$F(x) = \int_a^x f(t) dt.$$

Then F is continuous on $[a, b]$; furthermore, if f is continuous at a point $x_0 \in [a, b]$, then F is differentiable, and $F'(x_0) = f(x_0)$.

Proof. We will first show that F is continuous. Because f is integrable, it is bounded, so there exists an M for which $|f(t)| \leq M$ on $[a, b]$. If $a \leq x < y \leq b$, Theorem 6.2 Parts 3 and 5 give

$$|F(y) - F(x)| = \left| \int_x^y f(t) dt \right| \leq M(y - x). \quad (18)$$

If we have $|y - x| < \varepsilon/M$, then

$$|F(y) - F(x)| < \varepsilon,$$

so F is continuous.¹¹⁶

Now suppose f is continuous at x_0 . For $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|f(t) - f(x_0)| < \varepsilon$$

for $t \in [a, b]$ which satisfy $|t - x_0| < \delta$. Therefore if we have

$$x_0 - \delta < s \leq x_0 \leq t < x_0 + \delta,$$

and $a \leq s < t \leq b$, by Theorem 6.2 Part 5 we have

$$\left| \int_s^t \underbrace{f(u) - f(x_0)}_{<\varepsilon} du \right| < \varepsilon(t - s)$$

¹¹⁴You're probably thinking "yes it is", but keep in mind that you have known this for years, and have most likely seen several interpretations of the relationship.

¹¹⁵This is one reason that the Mean Value Theorem is so important.

¹¹⁶It actually ended up being uniformly continuous, so happy day.

Not only does this inequality take the form of (18),¹¹⁷ but we satisfy the conditions under which (18) holds, so we have

$$\begin{aligned} |F(t) - F(s) - f(x_0)| &= \left| \int_s^t f(u) du - f(x_0) \right| < \varepsilon(t-s) \\ \left| \frac{F(t) - F(s)}{s-t} - f(x_0) \right| &= \left| \frac{1}{s-t} \int_s^t f(u) du - f(x_0) \right| < \varepsilon \\ \left| \frac{F(t) - F(s)}{s-t} - f(x_0) \right| &< \varepsilon. \end{aligned}$$

That is, for all $\varepsilon > 0$ this final inequality holds for any $s \in (x_0 - \delta, x_0 + \delta)$, so if we take $t = x_0$ this becomes the definition of the following limit:

$$F'(x_0) = \lim_{s \rightarrow x} \frac{F(t) - F(s)}{s - x_0} = f(x_0).$$

□

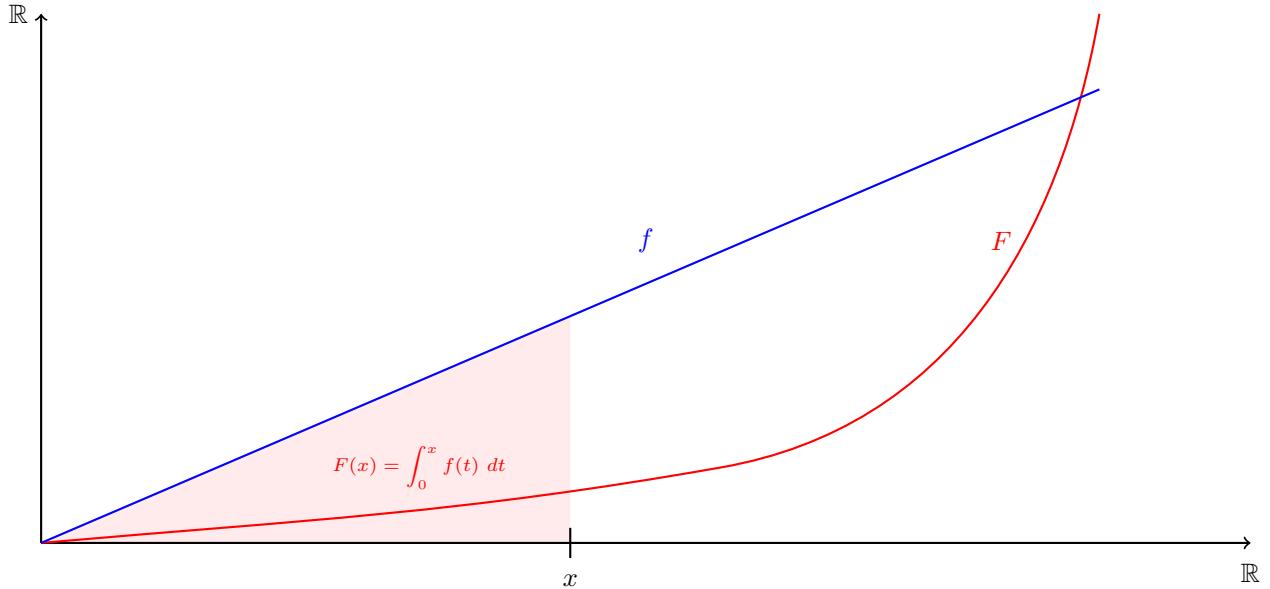


Figure 68: The first part of The Fundamental Theorem of Calculus says that $F' = f$.

Theorem 6.6 (Fundamental Theorem of Calculus II). Let f be integrable on $[a, b]$. If there exists a differentiable function F on $[a, b]$ such that $F' = f$, then

$$\int_a^b f(x) dx = F(b) - F(a).$$

Proof. Let $\varepsilon > 0$. By Riemann's Criterion, there exists some $P + \{x_0, \dots, x_n\}$ of $[a, b]$ such that $U(P, f) - L(P, f) < \varepsilon$. By the Mean Value Theorem, for each i we have a $t_i \in [x_{i-1}, x_i]$ such that

$$\begin{aligned} F(x_i) - F(x_{i-1}) &= F'(t_i)(x_i - x_{i-1}), \\ F(x_i) - F(x_{i-1}) &= f(t_i)\Delta x_i. \end{aligned}$$

¹¹⁷Let $M = \varepsilon$, $y = t$, $x = s$, and $f(t) = f(u)$.

Taking the sum over all i gives

$$\sum_{i=1}^n f(t_i) \Delta x_i = F(b) - F(a).$$

By Proposition 6.1 Part 3,

$$\left| F(b) - F(a) - \int_a^b f(x) dx \right| = \left| \sum_{i=1}^n f(t_i) - \int_a^b f(x) dx \right| < \varepsilon.$$

This holds for all ε , so

$$\int_a^b f(x) dx = F(b) - F(a).$$

□

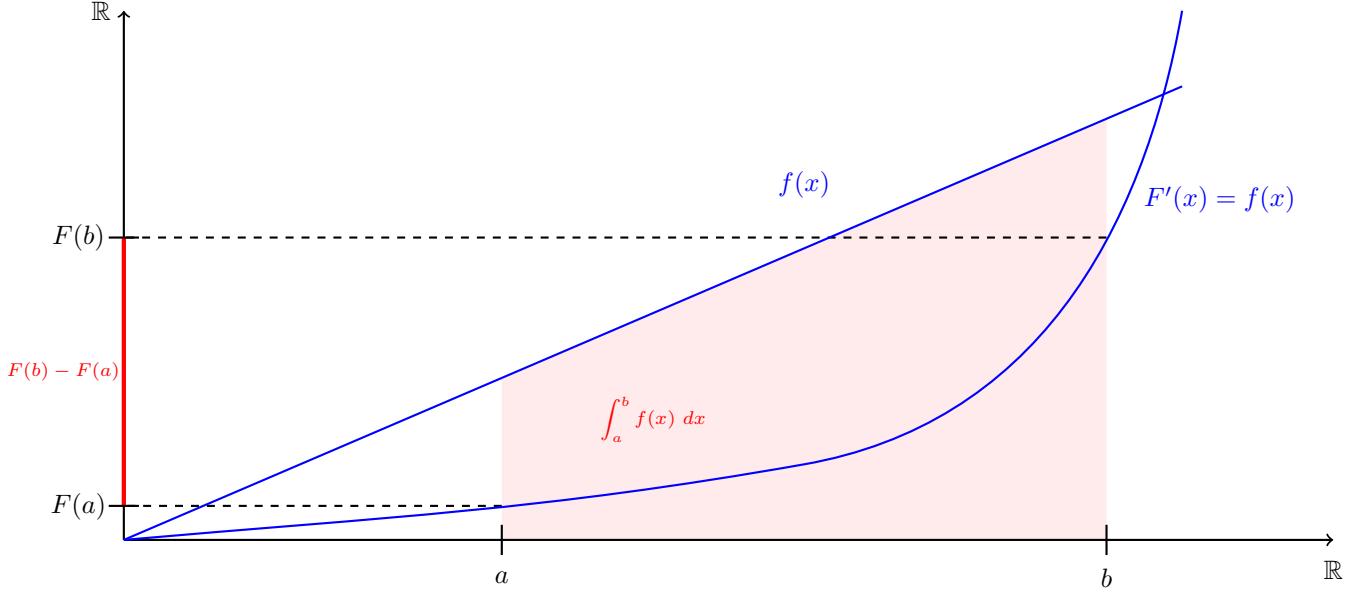


Figure 69: The second part of The Fundamental Theorem of Calculus says that if $F' = f$, then we have $\int_a^b f(x) dt = F(b) - F(a)$.

In essence, the Fundamental Theorem of Calculus (FTC) tells us that integration and differentiation are inverse operations.

A great application of the FTC is to find some integration technique that is analogous to the Product Rule of differentiation.

Proposition 6.9 (Integration by Parts). Suppose F and G are differentiable functions on $[a, b]$, and $F' = f$ and $G' = g$ are integrable as well. Then

$$\int_a^b F(x)g(x) dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x) dx.$$

Proof. Let $H(x) = F(x)G(x)$. We have

$$H'(x) = F'(x)G(x) + G'(x)F(x) = f(x)G(x) + g(x)F(x),$$

so H' is integrable by Proposition 6.5. The FTC gives

$$\begin{aligned} \int_a^b H'(x) \, dx &= H(b) - H(a) \\ \int_a^b f(x)G(x) + g(x)F(x) \, dx &= F(b)G(b) - F(a)G(a) \\ \int_a^b f(x)G(x) \, dx + \int_a^b g(x)F(x) \, dx &= F(b)G(b) - F(a)G(a) \\ \int_a^b F(x)g(x) \, dx &= F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x) \, dx \end{aligned}$$

Remark 6.6 (But What About Discontinuous Functions?). Subsection 6.7 showed that the set of Riemann integrable functions includes many more functions than just those which are continuous. Unfortunately, the FTC requires continuity. We still have no clear way of treating discontinuities when integrating. Example 6.20 and Example 6.23 will discuss this more.

□

6.10 Change of Variables

Change of variables may be something you first saw explicitly in multivariable calculus when working with polar or spherical coordinates, but you actually learned a basic form of it earlier. In the case of functions of a single variable it is often called u -substitution. For the sake of generalization later on, we'll refer to it as change of variables, but just keep in mind that all it is, is u -substitution.

The general idea of change of variables is that it is the opposite of the Chain Rule.

$$\int_{\phi(a)}^{\phi(b)} f(x) \, dx = \int_a^b f(\phi(y))\phi'(y) \, dy$$

We can justify this equality with reasoning similar to that which accompanied Theorem 6.6. This is best done with an example, as we need to deal with two sets of Riemann sums.

Let $f(x) = 5$, $\phi(y) = 2y$, and $[a, b] = [0, 10]$. We have

$$\int_{\phi(a)}^{\phi(b)} f(x) \, dx = \int_0^{20} 5 \, dx = 100, \tag{19}$$

$$\int_a^b \phi(y) \, dy = \int_0^{10} 2y \, dy = 25. \tag{20}$$

Equation (19) corresponds to a square of area 100, while Equation (20) corresponds to a square of area 25. The goal of change of variables is to manipulate the area of the triangle formed by ϕ so it has an area equal to the square formed by f . This allows us to integrate f without even considering its domain. We instead work in ϕ 's domain. First, we want to change the height of the Riemann sums which form the triangle so they correspond to the height of the sums corresponding to f and the rectangle (see Figure 68). This is achieved by composing ϕ with f . Our updated area is the integral of $f(\phi(y)) = 5$ over $[0, 10]$.

$$\int_0^{10} 5 \, dx = 50. \tag{21}$$

But we aren't done yet. The difference between (21) and (19) is the width of the corresponding rectangles. We cannot change the width of the prior though, because the whole point of this is to integrate in the domain

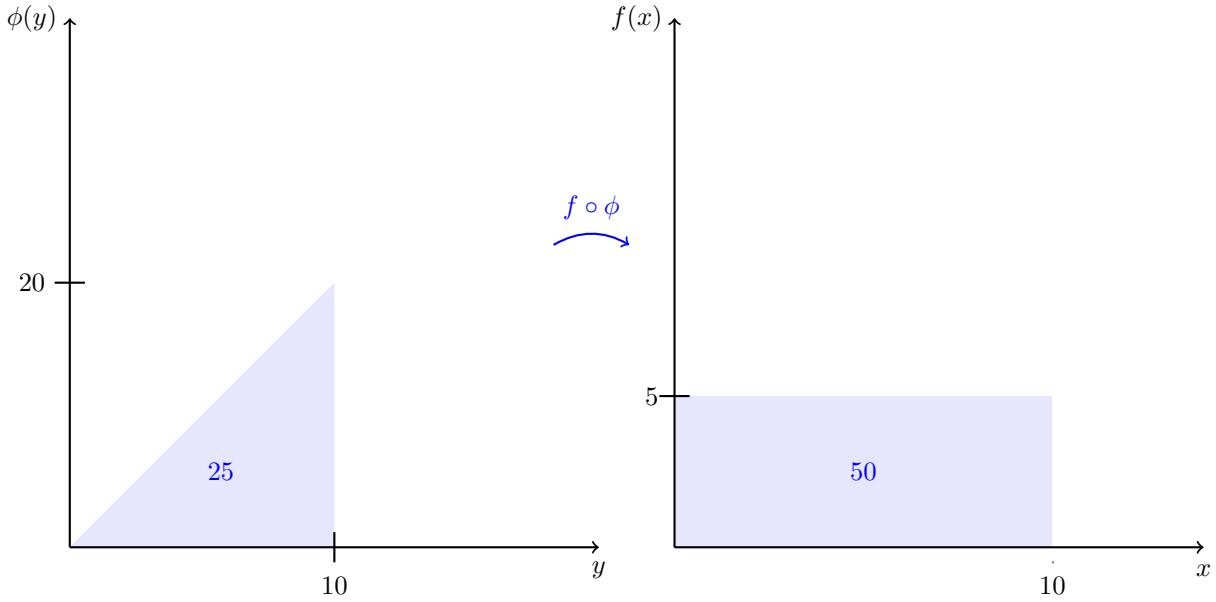


Figure 70: By composing ϕ with f , we modify the height of the integral of ϕ .

of φ . The bounds of integration should not be touched. We can instead scale the height of this rectangle by $\varphi'(y)$, which is the rate at which the width would be scaled if we decided to move to the domain of f (Figure 68). This gives us

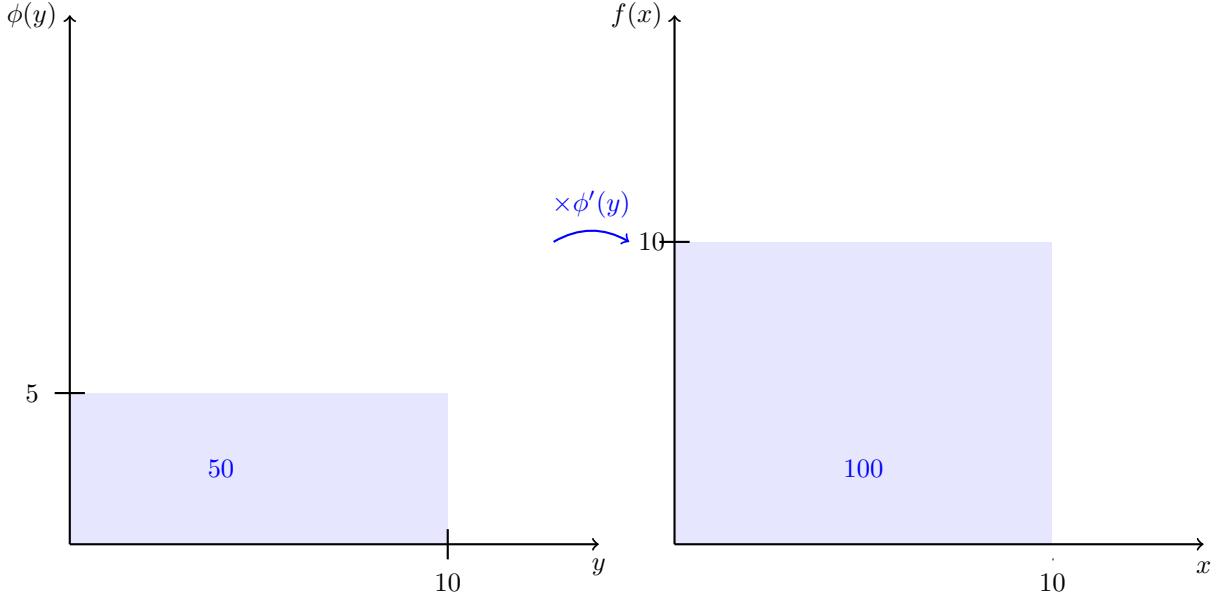


Figure 71: By scaling the height of the square from Figure 67 by a factor of φ' , we get the desired area.

$$\int_0^{10} 5\varphi'(y) dy = \int_0^{10} 5 \cdot 2 dy = 100.$$

Therefore we have scaled the area in (16) to that in (15) by geometrically transforming the function over

which we are integrating. In one step this is:

$$\int_{\phi(a)}^{\phi(b)} f(x) dx = \int_0^{20} 5 dx = 100 = \int_0^{10} 5 \times 2 dy = \int_a^b f(\phi(y))\phi'(y) dy$$

To convince yourself this works, it may be helpful to draw the pictures for the reverse direction. That is, how would we manipulate the square of area 100, to end up with a triangle of area 25. This particular direction is actually the useful direction, as you start with the simpler integral! It is also the direction you proceed in when performing u -substitution.

Remark 6.7 (Why Did I Go to All This Trouble?). This very simple example may seem a little redundant, but change of variables becomes very complicated when working with multivariable functions. It will be really important to have a strong intuition about this process when we move to the general case. It also helps in this simplified case because change of variables doesn't have this nice geometric representation when working with nontrivial functions. Even picking f to be linear would have prevented us from using well known formulas for the area of shapes.

Theorem 6.7 (Change of Variables). Let $[a, b]$ be a closed interval in \mathbb{R} , and $\phi : [a, b] \rightarrow [\phi(a), \phi(b)]$ be a differentiable strictly monotone increasing function such that ϕ' is integrable. If $f : [\phi(a), \phi(b)] \rightarrow \mathbb{R}$ is continuous on $[\phi(a), \phi(b)]$, then $(f \circ \phi)\phi' : [a, b] \rightarrow \mathbb{R}$ is integrable and

$$\int_{\phi(a)}^{\phi(b)} f(u) du = \int_a^b f(\phi(x))\phi'(x) dx = \int_a^b (f \circ \phi)(x) \cdot \phi'(x) dx.$$

Proof. First, we'll show that $(f \circ \phi)\phi'$ is integrable. The function ϕ is differentiable, so it is integrable, and the continuous composition of an integrable function is integrable (Proposition 6.4), so $(f \circ \phi)$ is integrable. We've assumed that ϕ' is integrable, so $(f \circ \phi)\phi'$ is the product of integrable functions, making it integrable. Since f is continuous, it has an antiderivative F .

$$F' = f$$

The chain rule and differentiability of ϕ give

$$(F \circ \phi)' = (F' \circ \phi)\phi' = (f \circ \phi)\phi'.$$

If we apply the fundamental theorem of calculus, we have

$$\int_a^b (f \circ \phi)(x) \cdot \phi'(x) dx = \int_a^b (F \circ \phi)'(x) dx = F(\phi(b)) - F(\phi(a)) = \int_{\phi(a)}^{\phi(b)} f(u) du.$$

But how do we know that $[\phi(a), \phi(b)]$ is a valid interval in \mathbb{R} ? Because ϕ is strictly monotone increasing, $a > b$ implies that $\phi(a) > \phi(b)$. Therefore, $[\phi(a), \phi(b)]$ is a valid interval over which we can integrate. \square

Example 6.21. Suppose $f(x) = (2x^3 + 1)^7 x^2$. We would like to integrate

$$\int_0^1 (2x^3 + 1)^7 x^2 dx,$$

but it is not clear how to integrate with respect to x in this case. We will instead change the variable we are integrating with respect to and appeal to Theorem 6.5. Let $\phi(x) = 2x^3 + 1$, and $f(u) = u^7$. Integration with respect to u gives

$$\int_0^1 (2x^3 + 1)^7 x^2 dx = \frac{1}{6} \int_0^1 f(\phi(x)) 6x^2 dx = \frac{1}{6} \int_0^1 f(\phi(x))\phi'(x) dx = \frac{1}{6} \int_{\phi(0)}^{\phi(1)} f(u) du = \frac{1}{6} \int_1^3 u^7 du = \frac{410}{3}.$$

This makes *no sense* though, because if we calculate the integral directly using the fundamental theorem of calculus, we have

$$\int_{-1}^1 x^2 \, dx = \frac{1}{3}$$

Remark 6.8 (Monotonicity of ϕ). What happens if ϕ is not monotonically increasing? Suppose $f(u) = \sqrt{u}$, and $\phi(x) = x^2$.

$$\int_{-1}^1 x^2 \, dx = \frac{1}{2} \int_{-1}^1 x \cdot 2x \, dx = \int_{-1}^1 \underbrace{\sqrt{x^2}}_{f \circ \phi} \cdot \underbrace{2x}_{\phi'(x)} \, dx \stackrel{?}{=} \frac{1}{2} \int_{(-1)^2}^{1^2} f(u) \, du = \frac{1}{2} \int_1^1 \sqrt{u} \, du = 0$$

This makes *no sense*, because the fundamental theorem of calculus gives

$$\int_{-1}^1 x^2 \, dx = \left(\frac{x^3}{3} \right)_{-1}^1 = \frac{2}{3}.$$

The reason that change of variables won't work in this situation is because ϕ is not monotonically increasing on $[-1, 1]$. Exploring this requires we address one aspect of integration ignored up until this point – orientation. If you look back at Theorem 6.2, you may notice that one familiar property taught in calculus courses is missing:

$$\int_a^b f(x) \, dx = - \int_b^a f(x) \, dx.$$

As far as we are concerned, the right side of this equation is nonsense. We've defined the Riemann integral over a closed interval $[a, b]$, implicitly assuming $a < b$. Even if we let this pass, why would the area under the curve change just because we reoriented the interval? If we're approximating the area under a curve using Riemann sums, the answer is going to be the same regardless of the order in which we count the rectangles. How we make sense of this, and how we formalize the Riemann integral's relationship with orientation, will be the subject of Section 12.

It is worth admitting that the notation we have used for the integral of f on $[a, b]$ does seem to hint at the fact that orientation plays some role in integration. For instance you can read $\int_a^b f(x) \, dx$ as “the integral of f from a to b .” If we want our notation to reflect the fact that the Riemann integral, as defined up until now, measures the area under a function, we could write

$$\int_{[a,b]} f(x) \, dx.$$

This is the notation used by [Tao \(2016a\)](#).

6.11 Null Sets and Lebesgue's Criterion

Riemann's criterion (Theorem 6.1) gives a necessary and sufficient condition for f to be Riemann integrable, making it *very* important. Unfortunately, its difficulty to use is comparable to its importance. Explicitly finding a partition of $[a, b]$ such that $U(P, f) - L(P, f) < \varepsilon$ isn't the most straightforward task. It would be much easier to find a necessary and sufficient condition for Riemann integrability which is related to properties of f we are more comfortable with. Lebesgue's criterion achieves this by defining Riemann integrability in terms of continuity.

We start by defining a special type of set in \mathbb{R} that is so “small” in length. These sets will be so small that they're negligible as far as integration is concerned.

Definition 6.12. Let A be a subset of \mathbb{R} . We refer to A as a *null set (in \mathbb{R})* if for all $\varepsilon > 0$, there exists an open cover $\{U_i\} \subseteq \mathbb{R}$, where $U_i = (a_i, b_i)$, of A such that:

$$\sum_{i=1}^{\infty} (b_i - a_i) < \varepsilon.$$

In other words, if we can find an open cover (Definition 2.17) of A with an arbitrarily small total length, then A is a null set.

Example 6.22 (Finite Sets are Null). Any finite subset of \mathbb{R} is a null set. Suppose $A = \{a_1, \dots, a_n\} \subset \mathbb{R}$ for some $n \in \mathbb{N}$. For all $\varepsilon > 0$ we can cover A with the open cover $\{U_i\}$ where

$$U_i = \left(a_i - \frac{\varepsilon}{4n}, a_i + \frac{\varepsilon}{4n}\right)$$

for $i = 1, \dots, n$. By construction, $a_i \subset U_i$, so $A \subseteq \bigcup_{i=1}^n U_i$. We have

$$\sum_{i=1}^{\infty} \left(\left(a_i + \frac{\varepsilon}{4n}\right) - \left(a_i - \frac{\varepsilon}{4n}\right) \right] = \sum_{i=1}^n \frac{\varepsilon}{2n} = n \cdot \frac{\varepsilon}{2n} = \frac{\varepsilon}{2} < \varepsilon,$$

for any $\varepsilon > 0$. This makes A a null set.

Example 6.23 (Countably Infinite Sets are Null). Suppose $A = \{a_1, a_2, \dots\}$ is a countably infinite set. Define an open cover $\{U_i\}$ as

$$U_i = \left(a_i - \frac{\varepsilon}{2^{i+2}}, a_i + \frac{\varepsilon}{2^{i+2}}\right)$$

for all i . We have

$$\sum_{i=1}^{\infty} \left(\left(a_i + \frac{\varepsilon}{2^{i+2}}\right) - \left(a_i - \frac{\varepsilon}{2^{i+2}}\right) \right] = \sum_{i=1}^{\infty} \frac{\varepsilon}{2^{i+1}} = \frac{1}{2} \sum_{i=1}^{\infty} \frac{\varepsilon}{2^i} = \frac{\varepsilon}{2} < \varepsilon$$

for any $\varepsilon > 0$. Therefore any countably infinite set is a null set.

These previous two examples help develop some notation about null sets. There real line has so many elements, that in the grand scheme of thing, any countable or finite subset is negligible. There are however null sets which are unaccountably infinite, one of which may be the most celebrated pathological example in real analysis.

Example 6.24 (The Cantor Set). We will construct a set known as the Cantor Set (named after Georg Cantor). Begin with the closed unit interval $[0, 1]$. If we remove the inner third $(1/3, 2/3)$ from this interval, we have $[0, 1/3] \cup [2/3, 1]$. Now remove the inner third interval from $[0, 1/3]$ and $[2/3, 1]$, giving us the union

$[0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$. If we keep repeating this process, we have:

$$\begin{aligned}
C_0 &= [0, 1] \\
C_1 &= \left(0, \frac{1}{3}\right] \cup \left(\frac{2}{3}, 1\right] \\
&= \frac{1}{3}C_0 \cup \left(\frac{1}{3}C_0 + \frac{2}{3}\right) \\
C_2 &= \left(\frac{0}{9}, \frac{1}{9}\right] \cup \left(\frac{2}{9}, \frac{3}{9}\right] \cup \left(\frac{6}{9}, \frac{7}{9}\right] \cup \left(\frac{8}{9}, \frac{9}{9}\right] \\
&= \frac{1}{3}C_1 \cup \left(\frac{1}{3}C_1 + \frac{2}{3}\right) \\
C_3 &= \left(\frac{0}{27}, \frac{1}{27}\right] \cup \left(\frac{2}{27}, \frac{3}{27}\right] \cup \left(\frac{6}{27}, \frac{7}{27}\right] \cup \left(\frac{8}{27}, \frac{9}{27}\right] \cup \left(\frac{18}{27}, \frac{19}{27}\right] \cup \left(\frac{20}{27}, \frac{21}{27}\right] \cup \left(\frac{24}{27}, \frac{25}{27}\right] \cup \left(\frac{26}{27}, \frac{27}{27}\right] \\
&= \frac{1}{3}C_2 \cup \left(\frac{1}{3}C_2 + \frac{2}{3}\right) \\
&\vdots \\
C_n &= \frac{1}{3}C_{n-1} \cup \left(\frac{1}{3}C_{n-1} + \frac{2}{3}\right)
\end{aligned}$$

Figure 72 illustrates C_0 through C_6 . The Cantor set as the set of points remaining after repeating this



Figure 72:

process *ad infinitum*, which can also be written as the intersection of all C_n .

$$C = \lim_{n \rightarrow \infty} C_n = \bigcap_{n=0}^{\infty} C_n$$

This set is a null set. At any fixed step n of constructing n we have 2^n disjoint closed intervals $[a_k, b_k]$,

$$C_n = \bigcup_{k=1}^{2^n} [a_k, b_k].$$

By construction each of these 2^n closed intervals has length $1/3^n - (b_k - a_k) = 1/3^n$. For any $\varepsilon > 0$ we can cover C_n with open intervals $(a_k - \varepsilon/2^{n+2}, b_k + \varepsilon/2^{n+2})$.

$$C_n = \bigcup_{k=1}^{2^n} [a_k, b_k] \subset \bigcup_{k=1}^{2^n} \left(a_k - \frac{\varepsilon}{2^{n+2}}, b_k + \frac{\varepsilon}{2^{n+2}}\right).$$

Each element of this open cover has length:

$$\left(b_k + \frac{\varepsilon}{2^{n+2}}\right) - \left(a_k - \frac{\varepsilon}{2^{n+2}}\right) = (b_k - a_k) + 2 \cdot \frac{\varepsilon}{2^{n+1}} = \frac{1}{3^n} + \frac{1}{2} \frac{\varepsilon}{2^n}$$

The total “size” of the open cover is this length multiplied by 2^n :

$$2^n \left(\frac{1}{3^n} + \frac{1}{2} \frac{\varepsilon}{2^n} \right) = \left(\frac{2}{3} \right)^n + \frac{\varepsilon}{2}.$$

The Cantor set C is given as the limit of C_n as $n \rightarrow \infty$, in which case the “size” of our open cover is

$$\lim_{n \rightarrow \infty} \left(\left(\frac{2}{3} \right)^n + \frac{\varepsilon}{2} \right] = \frac{\varepsilon}{2} < \varepsilon,$$

so C is a null set.

Paradoxically, C is also uncountably infinite. Verifying this requires expressing C in a more compact fashion. Ordinarily, elements of \mathbb{R} are expressed according to the decimal (base-10) system. For example, we write

$$234.56 = 2 \cdot 10^2 + 3 \cdot 10^1 + 4 \cdot 10^0 + 5 \cdot 10^{-1} + 6 \cdot 10^{-2}.$$

In general any real number. There exist alternate numerical systems. For example, the binary numeral system writes elements of \mathbb{R} using two digits (traditionally denoted as 0 and 1). The ternary numerical system writes elements using three digits $\{0, 1, 2\}$, and can be used to define the Cantor set. Any decimal element $x \in [0, 1]$ can be expressed as a ternary sequence (d_1, d_2, \dots) where

$$x = d_1 \cdot 3^{-1} + d_2 \cdot 3^{-2} + \dots = \sum_{n=1}^{\infty} d_n \cdot 3^{-n},$$

and $d_n \in \{0, 1, 2\}$ for all n . For example, $1/3$ corresponds to the sequence $(1, 0, 0, \dots)$. It also corresponds to the ternary sequence $(0, 2, 2, 2, 2, 2, \dots)$, as

$$0 \cdot 3^{-1} + 2 \cdot 3^{-2} + 3 \cdot 3^{-4} + \dots = \sum_{n=2}^{\infty} 2 \cdot 3^{-n} = 2 \cdot \frac{1}{6} = \frac{1}{3}.$$

This shows that ternary expansions needn’t be unique. Suppose $x \in [0, 1]$ is given as two ternary expansions: (d_1, d_2, \dots) and (e_1, e_2, \dots) . Under what conditions does

$$\sum_{n=1}^{\infty} d_n \cdot 3^{-n} = \sum_{n=1}^{\infty} e_n \cdot 3^{-n}$$

hold when $(d_1, d_2, \dots) \neq (e_1, e_2, \dots)$? Suppose that m is the first position where these expansions differ, i.e $d_m \neq e_m$. Without loss of generality, assume $d_m > e_m$.

$$\begin{aligned} \sum_{n=1}^{\infty} d_n \cdot 3^{-n} &= \sum_{n=1}^{\infty} e_n \cdot 3^{-n} \\ \implies \sum_{n=1}^{m-1} d_n \cdot 3^{-n} + d_m \cdot 3^{-m} + \sum_{n=m+1}^{\infty} d_n \cdot 3^{-n} &= \sum_{n=1}^{m-1} e_n \cdot 3^{-n} + e_m \cdot 3^{-m} + \sum_{n=m+1}^{\infty} e_n \cdot 3^{-n} \\ \implies d_m \cdot 3^{-m} + \sum_{n=m+1}^{\infty} d_n \cdot 3^{-n} &= e_m \cdot 3^{-m} + \sum_{n=m+1}^{\infty} e_n \cdot 3^{-n} && (e_n = d_n \ \forall n < m) \\ \implies \underbrace{(d_m - e_m)3^{-m}}_{\geq 3^{-m}} + \sum_{n=m+1}^{\infty} d_n \cdot 3^{-n} &= \underbrace{\sum_{n=m+1}^{\infty} e_n \cdot 3^{-n}}_{\leq \sum_{n=m+1}^{\infty} 2 \cdot 3^{-n} = 3^{-m}} && (d_m > e_m \text{ and } e_n \leq 2) \end{aligned}$$

This only holds if $d_m - e_m = 1$, and $d_n = 0$ and $e_n = 2$ for all $n \geq m + 1$.

The Cantor set is comprised of all the elements of $x \in [0, 1]$ with a ternary expansion comprised only of 0 and 2.

$$C = \left\{ \sum_{n=1}^{\infty} d_n \cdot 3^{-n} \mid d_n \in \{0, 2\} \right\}$$

In this case, all ternary expansions are unique. Most importantly, C is uncountable by the same argument used to prove Theorem 1.6.

Lemma 6.3. Suppose $A = \{A_1, A_2, \dots\}$ is a countable union of null sets in \mathbb{R} . Then the union $\cup_n A_n$ is a null set

Proof. For all $\varepsilon > 0$ there exists an open cover $\{U_{n,i}\}$ of A_n such that $U_{n,i} = (a_{n,i}, b_{n,i})$ and

$$\sum_{i=1}^{\infty} (b_{n,i} - a_{n,i}) < \varepsilon/2^n$$

for each n . The union of these open covers is itself an open cover of A ,

$$A \subseteq \bigcup_{n=1}^{\infty} \bigcup_{i=1}^{\infty} U_{n,i} = \bigcup_{n=1}^{\infty} \bigcup_{i=1}^{\infty} (a_i, b_i).$$

This open cover satisfies

$$\sum_{n=1}^{\infty} \sum_{i=1}^{\infty} (b_{n,i} - a_{n,i}) < \sum_{n=1}^{\infty} \varepsilon/2^n = \varepsilon$$

so A is a null set. \square

Next we introduce the concept of oscillation. On its own, this is not especially important, but it will allow us to prove a necessary and sufficient condition for Riemann integrability with relative ease.

Definition 6.13. Suppose f is a bounded real function on $[a, b]$. If $S \subseteq [a, b]$, the *oscillation of f on S* is given as

$$\omega_f(S) = \sup_{x,y \in S} |f(x) - f(y)| = \sup_{x \in S} f(x) - \inf_{y \in S} f(y).$$

The *oscillation of f at a point $x \in [a, b]$* is defined as

$$\omega_f(x) = \inf_{r>0} \omega_f(B_r(x)).$$

Lemma 6.4 (Zero Oscillation \iff Continuous). Suppose f is a real function defined on $[a, b]$. Then f is continuous at $x_0 \in [a, b]$ if and only if $\omega_f(x_0) = 0$.

Proof.

(\implies) Suppose $\omega_f(x_0) = 0$. Then there exists some $\delta > 0$ such that $\omega_f(B_\delta(x_0)) < \varepsilon$ for all $\varepsilon > 0$. In other words, for all $\varepsilon > 0$,

$$\omega_f(B_\delta(x_0)) = \sup_{x,y \in B_\delta(x_0)} |f(x) - f(y)| < \varepsilon.$$

Therefore ε is an upper bound on $|f(x) - f(y)|$ for any two points $x, y \in B_r(x_0)$. This is another way of saying that

$$|f(x) - f(y)| < \varepsilon$$

whenever $|x - y| < \delta$. If we take y to be x_0 we have

$$|f(x) - f(x_0)| < \varepsilon$$

whenever $|x - x_0| < \delta$, which is the definition of f being continuous at x_0 .

(\Leftarrow) Suppose f is continuous at x_0 . For all $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x) - f(x_0)| < \frac{\varepsilon}{2}$$

whenever $|x - x_0| < \delta$. Therefore

$$|f(x) - f(y)| \leq |f(x) - f(x_0)| + |f(x_0) - f(y)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

when $|x - x_0| < \delta$ and $|y - x_0| < \delta$. For any $x, y \in B_\delta(x_0)$,

$$\begin{aligned} & |f(x) - f(y)| < \varepsilon \\ \implies & \sup_{x, y \in B_\delta(x_0)} |f(x) - f(y)| < \varepsilon \\ \implies & w_f(B_\delta(x_0)) < \varepsilon \end{aligned}$$

If $w_f(B_\delta(x_0)) < \varepsilon$ for some δ for all ε , then $w_f(x_0) = \inf_{\delta > 0} w_f(B_\delta(x_0)) = 0$.

□

Intuitively, the oscillation of a function at a point gives us some measure of just how discontinuous it is. If a function has a removable discontinuity (Definition 4.8),

$$\omega_f(x) = |f(x) - f(x_-)| = |f(x) - f(x_+)|.$$

If we have a jump discontinuity (Definition 4.9),

$$\omega_f(x) = |f(x_-) - f(x_+)|.$$

In the event of an essential discontinuity (Definition 4.10) the oscillation corresponds to the degree to which the limits fail to exist. It may be helpful to go back to Section 4.6, consider examples of each type of discontinuity, and then calculate the oscillation at those discontinuities.

Before stating our main result, we need one more lemma related to oscillation.

Lemma 6.5. Suppose f is a real bounded function defined on $[a, b]$. For every $\alpha > 0$, the set $\{x \mid \omega_f(x) < \alpha\}$ is open in $[a, b]$, and $\{x \mid \omega_f(x) \geq \alpha\}$ is closed in \mathbb{R} .

Proof. To show $\{x \mid \omega_f(x) < \alpha\}$ is open, we will show that an arbitrary point of the set is an interior point. We have $\omega_f(a) < \alpha$. By the definition of $\omega_f(a)$, there exists some $r > 0$ such that $\omega_f(B_r(a)) < \alpha$. If $b \in B_r(a)$ and $x \in U \subset B_r(a)$, then $w_f(U) < \alpha$, implying $w_f(b) \leq w_f(U) < \alpha$. Therefore $b \in \{x \mid \omega_f(x) < \alpha\}$, and a is an interior point.

If $\{x \mid \omega_f(x) < \alpha\}$ is open in $[a, b]$, then $\{x \mid \omega_f(x) < \alpha\}^c = \{x \mid \omega_f(x) \geq \alpha\}$ is closed in $[a, b]$. □

Theorem 6.8 (Lebesgue's Criterion for Riemann-Integrability). Suppose f is a real bounded function defined on $[a, b]$. Then, f is Riemann integrable if and only if the set of discontinuities of f on $[a, b]$ is a null set.

Proof.

(\Rightarrow) Suppose f is Riemann integrable. Define the set D to be the points at which f is discontinuous on $[a, b]$. By Lemma 6.4,

$$D = \{x \in [a, b] \mid \omega_f(x) > 0\}.$$

We need to show that D is a null set. For the sake of simplicity, let's restrict our initial attention to the set

$$D_\alpha = \{x \in [a, b] \mid \omega_f(x) \geq \alpha\} \subset D$$

for some fixed $\alpha > 0$. By Riemann's criterion (Theorem 6.1), for all $\varepsilon > 0$, there exists some partition $P = \{x_0, \dots, x_n\}$ such that

$$\begin{aligned} U(P, f) - L(P, f) &< \frac{\alpha\varepsilon}{2} \\ \implies \sum_{i=1}^n M_i \Delta x_i - \sum_{i=1}^n m_i \Delta x_i &< \frac{\alpha\varepsilon}{2} && (\text{def. of } U(P, f) \text{ and } L(P, f)) \\ \implies \sum_{i=1}^n \left(\sup_{x \in [x_{i-1}, x_i]} f(x) - \inf_{x \in [x_{i-1}, x_i]} f(x) \right) \Delta x_i &< \frac{\alpha\varepsilon}{2} && (\text{def. of } M_i \text{ and } m_i) \\ \implies \sum_{i=1}^n \omega_f([x_{i-1}, x_i]) \Delta x_i &< \frac{\alpha\varepsilon}{2} && (\text{def. of } \omega_f(\Delta x_i)). \end{aligned}$$

Now define the set $F = \{i \mid D_\alpha \cap [x_{i-1}, x_i]\}$. This is the set of partition intervals which contain a discontinuity of f . For any $i \in F$, there is some $y \in [x_{i-1}, x_i]$ such that $\omega_f(y) \geq \alpha$, so $\omega_f([x_{i-1}, x_i]) \geq \alpha$. Therefore we only sum over $i \in F$, we have

$$\begin{aligned} \sum_{i \in F} \omega_f([x_{i-1}, x_i]) \Delta x_i &\leq \sum_{i=1}^n \omega_f([x_{i-1}, x_i]) \Delta x_i \\ \implies \sum_{i \in F} \omega_f([x_{i-1}, x_i]) &< \frac{\alpha\varepsilon}{2} \\ \implies \sum_{i \in F} \alpha \Delta x_i &< \frac{\alpha\varepsilon}{2} && (\omega_f([x_{i-1}, x_i]) \geq \alpha \ \forall i \in F) \\ \implies \sum_{i \in F} \Delta x_i &< \frac{\varepsilon}{2} \\ \implies \sum_{i \in F} (x_{i-1} - x_i) &< \varepsilon, \end{aligned}$$

so the collection of open sets $(x_{i-1} - x_i)$ has arbitrary small length when $i \in F$. As it turns out, this collection of sets is also an open cover of D_α by construction,

$$D_\alpha \subseteq \bigcup_{i \in F} (x_{i-1} - x_i).$$

This makes D_α a null set for any $\alpha > 0$.

This is progress, but we want to show that D is a null set, not an arbitrary subset D_α . Fortunately, we can write D as the countable union of D_α as follows:

$$D = \bigcup_{k=1}^{\infty} D_{1/k}.$$

The countable union of null sets is also a null set, so D is a null set.

(\Leftarrow) This direction of the proof is a bit more involved, so we'll break it into the following steps:

- For any $\varepsilon > 0$, find an open cover of points at which f is discontinuous and has an oscillation greater than or equal to $\varepsilon/2(b-a)$. Call this finite open cover $\{U_j\}_{j=1}^m$

2. Find a finite open cover of the points of $[a, b]$ which do not intersect $\{U_j\}_{j=1}^m$.
3. Combine the open covers from Step 1 and Step 2 to form a partition of $[a, b]$.
4. Show that by construction, and clever choices of arguments involving ε , $U(P, f) - L(P, f) < \varepsilon$.

Step 1. Suppose the set of discontinuities of f on $[a, b]$, call it D , is a null set. By Lemma 6.4,

$$D = \{x \in [a, b] \mid \omega_f(x) > 0\}.$$

Given some $\varepsilon > 0$, define

$$E = \left\{ x \mid \omega_f(x) \geq \frac{\varepsilon}{2(b-a)} \right\},$$

where $E \subseteq D$. The set E is the subset of a null set, making itself a null set.¹¹⁸ By the definition of a null set (Definition 6.12), we can find an open cover $\{U_i\}$ of E such that $U_i = (\alpha_i, \beta_i)$ and

$$\sum_{i=1}^{\infty} (\beta_i - \alpha_i) < \frac{\varepsilon}{2 \left(\sup_{x \in [a,b]} f(x) - \inf_{x \in [a,b]} f(x) \right)}.$$

The set E is bounded, as it is a subset of $[a, b]$. In addition, Lemma 6.5 tells us that E is closed. A closed and bounded set in \mathbb{R} is compact, so any open cover of E has a finite subcover. This includes the open cover $\{U_i\}$! There exists finite indices $j = 1, \dots, m$ such that $E \subseteq \cup_{j=1}^m U_j$.

Step 2. Let's look at the set of points in $[a, b]$ that are not in this finite subcover $\{U_j\}$, call it W .

$$W = [a, b] \setminus \cup_{j=1}^m U_j.$$

Because $E \subseteq \cup_{j=1}^m U_j$, $W \subset [a, b] \setminus E$. In other words, for a fixed $x \in W$, $\omega_f(x) < \frac{\varepsilon}{2(b-a)}$ (by the definition of E). In terms of Definition 6.13,

$$\omega_f(x) = \inf_{r>0} \omega_f(B_r(x)) = \inf_{r>0} \sup_{y,z \in (x-r,x+r)} |f(y) - f(z)| < \frac{\varepsilon}{2(b-a)}.$$

This definition tells us for our fixed $x \in W$, there exists some corresponding $r_x > 0$ such that whenever $y, z \in (x - r_x, x + r_x)$, we have

$$\begin{aligned} |f(y) - f(z)| &\leq \sup_{y,z \in (x-r,x+r)} |f(y) - f(z)| < \frac{\varepsilon}{2(b-a)} \\ \implies |f(y) - f(z)| &< \frac{\varepsilon}{2(b-a)} \end{aligned}$$

If we do this for all x , we have an open cover of W in the form of

$$\{(x - r_x, x + r_x) \mid x \in W\}.$$

Does this have a finite open subcover? It would if W is compact, which turns out to be true. Clearly W is bounded as $W \subseteq [a, b]$. Fortunately, W also happens to be closed as we can write it as

$$W = [a, b] \cap (\cup_{j=1}^m U_j)^c = [a, b] \cap (\cap_{j=1}^m U_j^c).$$

Each U_j is open, so U_j^c is closed. The finite intersection of closed sets is closed, so W is closed. If W is closed and bounded $V = \{x_1, \dots, x_N\} \subset W$ such that

$$W \subseteq \bigcup_{x \in V} (x - r_x, x + r_x).$$

¹¹⁸We did not formally state this lemma, but it's quite immediate. Any open cover of D is an open cover of E , so the result holds.

Step 3. We can combine the finite open covers from Step 1 and Step 2 to form a finite cover of $[a, b]$:

$$[a, b] \subseteq \left(\bigcup_{x \in V} (x - r_x, x + r_x) \right) \cup \left(\bigcup_{j=1}^m U_j \right).$$

We can define a partition $P = [y_0, \dots, y_n]$ of $[a, b]$ such that each interval $[y_{\ell-1}, y_\ell]$ is contained entirely in one interval of our finite open cover of $[a, b]$. This means $[y_{\ell-1}, y_\ell]$ will either be a subset of $(x - r_x, x + r_x)$ for some $x \in V$, or $[y_{\ell-1}, y_\ell]$ will be a subset of U_j for some $j = 1, \dots, m$.

We can split P into two using this fact:

$$\begin{aligned} P_1 &= \{y_\ell \in P \mid \exists x \in V \text{ s.t } [y_{\ell-1}, y_\ell] \subseteq (x - r_x, x + r_x)\} \\ P_2 &= \{y_\ell \in P \mid \exists j \text{ s.t } [y_{\ell-1}, y_\ell] \subseteq U_j\} \end{aligned}$$

Step 4.

$$\begin{aligned} U(P, f) - L(P, f) &= \sum_{\ell=1}^n M_\ell \Delta y_\ell - \sum_{\ell=1}^n m_i \Delta y_\ell \\ &= \sum_{\ell=1}^n (M_\ell - m_\ell) \Delta y_\ell \\ &= \sum_{y_\ell \in P_1} (M_\ell - m_\ell) \Delta y_\ell + \sum_{y_\ell \in P_2} (M_\ell - m_\ell) \Delta y_\ell && (P = P_1 \cup P_2 \text{ and } P_1 \cap P_2 = \emptyset) \\ &= \sum_{y_\ell \in P_1} (M_\ell - m_\ell) \Delta y_\ell + \left(\sup_{x \in [a, b]} f(x) - \inf_{x \in [a, b]} f(x) \right) \sum_{y_\ell \in P_2} \Delta y_\ell \\ &= \sum_{y_\ell \in P_1} (M_\ell - m_\ell) \Delta y_\ell + \left(\sup_{x \in [a, b]} f(x) - \inf_{x \in [a, b]} f(x) \right) \sum_{j=1}^m (\beta_j - \alpha_j) && ([y_{\ell-1}, y_\ell] \subset U_j = (\alpha_j, \beta_j) \forall y_\ell \in P_2) \\ &< \sum_{y_\ell \in P_1} (M_\ell - m_\ell) \Delta y_\ell + \left(\sup_{x \in [a, b]} f(x) - \inf_{x \in [a, b]} f(x) \right) \frac{\varepsilon}{2 \left(\sup_{x \in [a, b]} f(x) - \inf_{x \in [a, b]} f(x) \right)} && (\cup_j U_j \text{ is a null set}) \\ &= \sum_{y_\ell \in P_1} (M_\ell - m_\ell) \Delta y_\ell + \frac{\varepsilon}{2} \\ &< \frac{\varepsilon}{2(b-a)} \sum_{y_\ell \in P_1} \Delta y_\ell + \frac{\varepsilon}{2} && (\text{definition of } P_1 \text{ and } r_x) \\ &\leq \frac{\varepsilon}{2(b-a)} (b-a) + \frac{\varepsilon}{2} && \left(\sum_{y_\ell \in P_1} (M_\ell - m_\ell) y_\ell \leq [a, b] \right) \\ &= \varepsilon \end{aligned}$$

□

We know from Remark 6.4 that Riemann integrability is a weaker condition than continuity. Just because a function is “nice” enough to be integrable, does not mean it is “nice” enough to be continuous. However, Lebesgue’s criterion tells us that if a function is “nice” enough to be continuous, except on a null set, then this is equivalent to being “nice” enough to be integrable. Lebesgue’s criterion also immediately gives Proposition 6.7 and Proposition 6.8 as corollaries.

Corollary 6.1. Suppose f is a real bounded function on $[a, b]$. The function f is Riemann integrable if:

1. f has finite discontinuities on $[a, b]$;
2. f is monotonic on $[a, b]$.

Proof. A finite set is a null set. By Proposition 4.2, a monotonic function has at most countable discontinuities. □

Example 6.25 (Thomae's function). Recall Example 4.8 where $f : [0, 1] \rightarrow [0, 1]$ was defined as

$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ for } p, q \in \mathbb{Z} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases},$$

taking $p/q \in \mathbb{Q}$ to be in simplest terms. The set of discontinuities of f on $[0, 1]$ is $[0, 1] \cap \mathbb{Q}$. While this set is infinite, it is countable, making it a null set. By Lebesgue's criterion f is Riemann integrable on $[0, 1]$. Let's calculate the integral of f over $[0, 1]$. Suppose P is an arbitrary partition of $[0, 1]$. Any interval $[x_{i-1}, x_i]$ of this partition contains a rational number because \mathbb{Q} is dense in \mathbb{R} . As such

$$\inf_{x \in [x_{i-1}, x_i]} f(x) = 0$$

for all i . This means $L(P, f) = 0$ is for all 0, so the lower integral is 0. Because f is Riemann integrable, the integral over $[0, 1]$ is the lower integral:

$$\int_0^1 f(x) dx = \underline{\int}_0^1 f(x) dx = 0.$$

6.12 Shortcomings of Riemann Integration

We conclude this section by giving several examples of functions that are either not integrable, or we have no means of evaluating the integral of. From now until section 13, **when we write “integral”, we will mean “Riemann integral”**, but you should keep these examples in the back of your head. They allude to the fact that we have only scratched the surface of integration.

Example 6.26 (Dirichlet Function). Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

As we say in Example 4.26, this function is nowhere continuous. This function also fails to be integrable on any interval. Without loss of generality, we can show that f is not integrable on $[0, 1]$. By the density of \mathbb{Q} in \mathbb{R} , any $[x_{i-1}, x_i]$ will contain both a rational and irrational number for all $[x_{i-1}, x_i] \subset P$. This gives $M_i = 1$ and $m_i = 0$ for all i . Therefore

$$U(P, f) - L(P, f) = \sum_{i=1}^n (M_i - m_i) \Delta x_i = 1 - 0 = 1,$$

for any P . The Riemann Criterion does not hold, so f is not integrable on $[0, 1]$. While it is hard to argue one would ever need to integrate a nowhere continuous function, this is still dissatisfying. We know that \mathbb{Q} is countable, whereas \mathbb{R} is not. This means that an uncountably infinite number of points of $f(x) = 0$ an uncountable number of times, whereas $f(x) = 1$ a countable number of times. Is it crazy to think that the integral on $[0, 1]$ should just be zero then? Regardless of what we think it should be, it still is not Riemann integrable.

Example 6.27 (Unbounded Function). Let $f(x) = 1/x$. This function cannot be integrated on $(0, 1]$, as it is unbounded on this interval, because $f(x-) = \infty$. If we instead tried to use the interval $[0, 1]$, then f would not be defined as $0 \in [0, 1]$.

Example 6.28 (Integral Over All of \mathbb{R}). Suppose we want to integrate some function over all of \mathbb{R} . The whole real line is unbounded, so how do we measure the length of an infinite set and find upper and lower Riemann sums? Integration this way is not possible in this case.

Remark 6.9 (But What About Improper Integrals?). You've likely seen improper integrals that allow us to evaluate integrals of functions on unbounded sets, but these technically are not Riemann integrals. They are limits of Riemann integrals. This is another topic we will take up in Section 14.

Example 6.29 (Discontinuities). While the Fundamental Theorem of Calculus allows us to evaluate integrals for continuous functions, but how do we evaluate discontinuous functions? We know a function with a finite number of discontinuous can be integrated, but we never specified how those discontinuities contribute to the area. If we follow the reasoning from Example 6.11, we would hope that they contribute nothing to area, as there are an uncountably infinite number of continuous points on $[a, b]$ that greatly out number the discontinuous ones.

7 Sequences and Series of Functions

Section 3 looked at sequences and series of numbers. We want to extend these ideas to real functions. It's tempting to think that we have already seen sequences of functions in the context of continuity. If f is continuous at x , then for any sequence $\{x_n\}$ which converges to x , we have $f(x_n) \rightarrow f(x)$. This looks like a sequence of functions, but it's still a numeric sequence. Each $f(x_n)$ is a number. This sequence is just the result of applying a function to a set of numbers. This preliminary example provides a bonafide example of the types of sequences we are turning our attention to.

Example 7.1. Define $f_n : \mathbb{R} \rightarrow \mathbb{R}$ as $f_n(x) = x^n$. We have a sequence of functions in $\{f_n\}$.¹¹⁹ The first several terms of this sequence look like:

$$x, x^2, x^3, x^4, x^5, \dots$$

We can easily discuss a sequence of functions, or its possible limit, in the context of continuity, differentiation, and integration. Determining how sequence and series interact with these concepts will be a major goal. Then we will focus on approximating and/or expressing a function as a limit of a sequence or series.

7.1 Metric Spaces of Functions

Before we can discuss sequences of functions, we need to think about a metric space of functions. Until now, we've only look at metric spaces of numbers, but we never said a metric space couldn't be comprised of more abstract objects. We'll start by giving several definitions of sets of functions that have already been covered. This will amount to just introducing some convenient notation.

Definition 7.1. The *set of continuous real functions* defined on a set $X \subset \mathbb{R}$ is denoted $C(X)$.

Definition 7.2. The *set of continuous bounded real functions* defined on a set $X \subset \mathbb{R}$ is denoted $C_b(X)$. In the event that X is compact, $C_b(X) = C(X)$ (Corollary 4.5).

Definition 7.3. The *set of n -times differentiable (real) functions* defined on a set $X \subset \mathbb{R}$ is denoted $C^n(X)$.

Example 7.2. We have $C_b(X) \subset C(X)$, and $C^\infty(X) \subset \dots C^2(X) \subset C^1(X) \subset C(X)$.

Example 7.3. If we define $\mathcal{R}(X)$ to be the set of all Riemann integrable functions on X , then

$$C^\infty(X) \subset \dots C^2(X) \subset C^1(X) \subset C(X) \subset \mathcal{R}(X).$$

We also have $C_b(X) \subset \mathcal{R}(X)$. For now, we should hold off on working with the set \mathcal{R} , because we'll define a larger set of integrable functions later.

How would we measure the distance between two functions? On what sets of functions is this even possible? Well, if a function is unbounded, then that complicates matters. We cannot exactly measure the distance between two objects if one of them goes off to infinity.¹²⁰ For that reason, we will define a metric on $C_b(X)$. We would also be able to do this for the set of all bounded functions, but for now we'll just assume continuity.

¹¹⁹The functions now have a subscript n , indicating that each n garners a different function.

¹²⁰The Euclidean metric in \mathbb{R} cannot measure the distance between ∞ and $x \in \mathbb{R}$ (although this is also because $\infty \notin \mathbb{R}$).

For $g, f \in C_b(X)$, what is a reasonable choice $d(f, g)$? At some point $x_0 \in X$, we know the distance between the two functions is given by the Euclidean metric.

$$d(f(x_0), g(x_0)) = |f(x_0) - g(x_0)|$$

This isn't really the distance between the two functions though. It's the distance between the *real numbers* the functions take on at the single point x_0 . This doesn't account for all the other points on which the set is defined, $X \setminus \{x_0\}$. Perhaps we could just take the maximum distance between the functions over X ? This would provide us with a conservative estimate, that says "okay well the distance is *at most* this." Just to be careful, we should take the supremum instead of the maximum.¹²¹ Our candidate distance function is

$$d(f, g) = \sup_{x \in X} |f(x) - g(x)|.$$

But does this choice of d satisfy definition 2.1?

We will have $d(f, g) = 0$ if and only $f = g$ in $C_b(X)$. We also have $d(f, g) = d(g, f)$, as $|f(x) - g(x)| = |g(x) - f(x)|$. It only remains to be shown that this choice of metric verifies the triangle inequality. For $f, g, h \in C_b(X)$ the triangle inequality in the Euclidean metric gives

$$\begin{aligned} d(f, g) + d(g, h) &= \sup_{x \in X} |f(x) - g(x)| + \sup_{x \in X} |g(x) - h(x)| \\ &\geq |f(x) - g(x)| + |g(x) - h(x)| \\ &\geq |f(x) - h(x)|. \end{aligned}$$

This shows that $\sup_{x \in X} |f(x) - g(x)| + \sup_{x \in X} |g(x) - h(x)|$ is an upper bound of $|f(x) - h(x)|$, so it is weakly greater than the supremum of $|f(x) - h(x)|$.

$$d(f, g) + d(g, h) \geq \sup_{x \in X} |f(x) - g(x)| + \sup_{x \in X} |g(x) - h(x)| \geq \sup_{x \in X} |f(x) - h(x)| = d(f, h)$$

Definition 7.4. The *supremum/sup metric* on $C_b(X)$ is defined as

$$d(f, g) = \|f - g\|_\infty = \sup_{x \in X} |f - g|.$$

For now, take the sup metric's special notation as given. We will explain where this notation comes from in the beginning of Section 8. We should note that other valid metric spaces are the set of bounded real functions $B(X)$, and $C(X)$ as long as X is compact.¹²²

Example 7.4. Let $f(x) = \sin x$ and $g(x) = \cos x$. Both of these functions are in the set $BC(\mathbb{R})$. We have

$$\|f - g\|_\infty = \sup_{x \in \mathbb{R}} |\cos x - \sin x| = \sqrt{2}.$$

In this case $|\cos x - \sin x| = \sqrt{2}$ happens for all $3k\pi/4 \in \mathbb{R}$ where $k \in \mathbb{Z}$.

In Subsection 7.3, the sup metric on spaces of functions will prove useful in our discussion of sequences of functions.

¹²¹If X is closed than each element of $C_b(X)$ achieves a max by the Extreme Value Theorem, so using the supremum in this case would not be strictly required.

¹²²The function being bounded follows from the compact domain, as these functions are continuous.

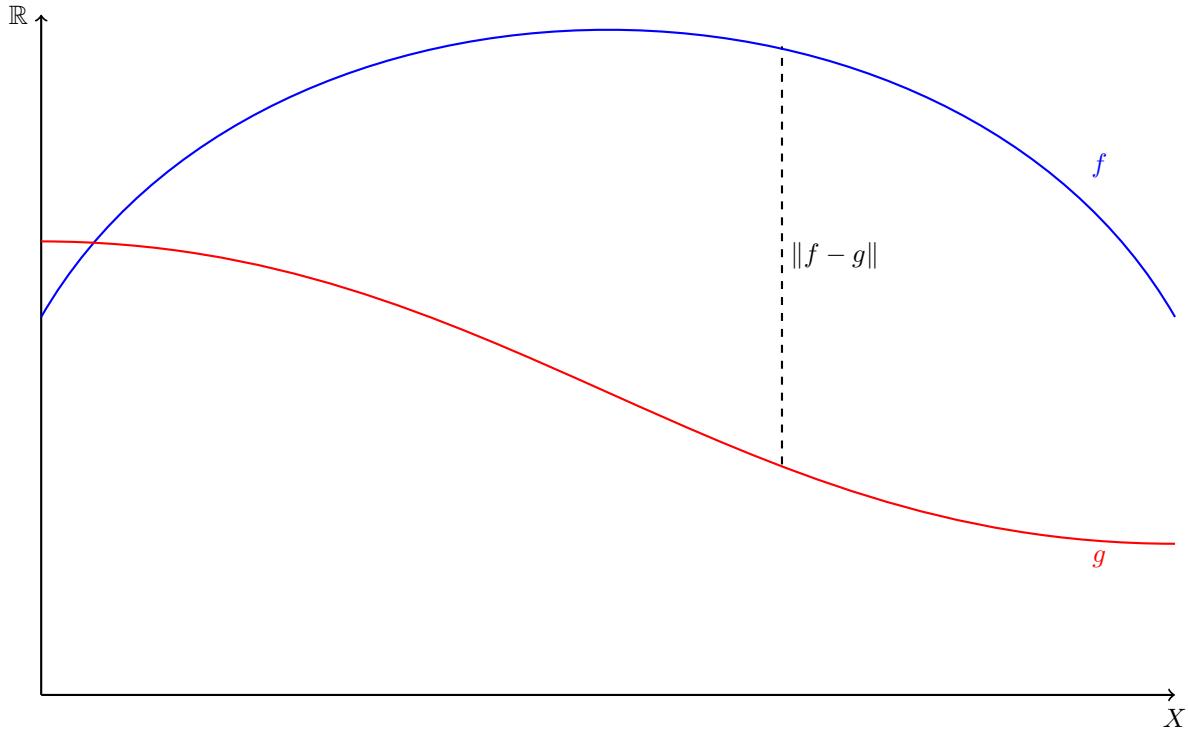


Figure 73: The distance between f and g as given by the sup metric is simply the supremum of the distance between $f(x)$ and $g(x)$ for a fixed $x \in X$.

7.2 Pointwise Convergence

We begin by defining one type of limit of a sequence of functions.

Definition 7.5. Let $E \subset \mathbb{R}$, and $f_n : E \rightarrow \mathbb{R}$ for all n . The sequence of functions $\{f_n\}$ *converges (pointwise)* to $f : E \rightarrow \mathbb{R}$ if for every $x \in E$ and $\varepsilon > 0$, there exists a *corresponding* $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon$ whenever $n \geq N$. In this case we write

$$f(x) = \lim_{n \rightarrow \infty} f_n(x)$$

for $x \in E$.

If $\{f_n\}$ is a sequence of functions on E , then for any fixed $x \in E$, $\{f_n(x)\}$ is a sequence of numbers. If this numerical sequence $\{f_n(x)\}$ converges to $f(x)$ for all $x \in E$, then we have pointwise convergence. This means in practice we may use the same limit rules developed for a sequence of numbers on a sequence of functions. An illustration of this can be seen in Figure 74. The sequence of functions in this figure converges pointwise to x_0 . It very well could converge pointwise to x_1 as well, it may just be that for some ε the value of N that gives $|f_n(x_0) - f(x_0)| < \varepsilon$ for all $n \geq N$ will not work for $|f_n(x_1) - f(x_1)|$. That is to say, **just because $\{f_n\}$ converges pointwise to f on E , that does not mean that one value of N will give $|f_n(x) - f(x)| < \varepsilon$ for all $x \in E$!** This is why in Definition 7.5, we have a *corresponding* N for each $x \in E$ and $\varepsilon > 0$. As the next examples will show, the particular value of N will depend on x and ε .

Example 7.5. Let $f_n = 1/nx$ be a sequence of functions. We can show that $f_n \rightarrow f = 0$ on the interval

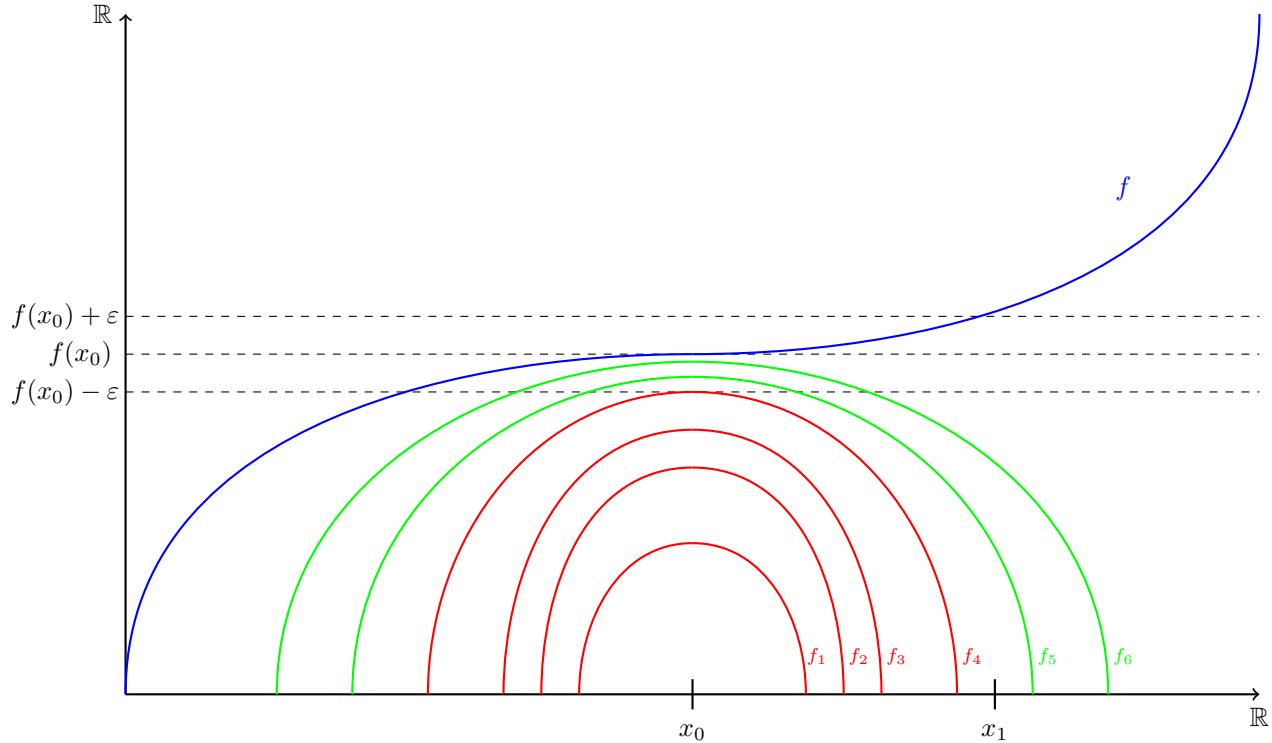


Figure 74: The sequence of functions $\{f_n\}$ converges pointwise to f at x_0 . For this particular choice of ε , we have $|f_n(x_0) - f(x_0)| < \varepsilon$ for all $n \geq 5$.

$(0, 1)$. If we let $N = 1/\varepsilon x$, then

$$|f_n(x) - f(x)| = \left| \frac{1}{nx} - 0 \right| = \frac{1}{nx} < \frac{1}{Nx} = \frac{1}{(1/\varepsilon x)x} = \varepsilon$$

for all $n \geq N$. In any real application we would calculate the limit in the following way:

$$\lim_{n \rightarrow \infty} f_n = \lim_{n \rightarrow \infty} \frac{1}{nx} = \lim_{n \rightarrow \infty} \left(\frac{1}{n} \right) x = 0 \cdot x = 0$$

Example 7.6. If we let $f_n = (nx)/(1+nx^2)$, we have $f_n \rightarrow f(x) = 1/x$ on $(0, \infty)$. If we let $N = 1/\varepsilon x$, then

$$|f_n(x) - f(x)| = \left| \frac{nx}{1+nx^2} - \frac{1}{x} \right| = \left| \frac{nx^2 - 1 + nx^2}{x + nx^3} \right| = \left| -\frac{1}{x + nx^3} \right| = \frac{1}{x + nx^3} < \frac{1}{x + Nx^3} < \frac{1}{x + \left(\frac{1-x\varepsilon}{\varepsilon x^3}\right) x^3} = \varepsilon$$

for all $n \geq N$.

One reason that sequences of functions are so useful is because we may be able to express some complicated function as a limit of relatively “nice” functions.¹²³ Perhaps if we need to take the limit of, differentiate, or integrate some complicated function, we can instead work with a sequence which converges to that function.

¹²³We discussed this type of situation in the passage after Corollary 3.1

If $f_n \rightarrow f$, can we just use the following relationships:

$$\begin{aligned}\int_a^b f(x) dx &= \int_a^b \lim_{n \rightarrow \infty} f_n(x) dx \stackrel{?}{=} \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx \\ f'(x) &= \left(\lim_{n \rightarrow \infty} f_n(x) \right)' \stackrel{?}{=} \lim_{n \rightarrow \infty} f'_n(x) \\ \lim_{t \rightarrow x} f(t) &= \lim_{t \rightarrow x} \left(\lim_{n \rightarrow \infty} f_n(x) \right) \stackrel{?}{=} \lim_{n \rightarrow \infty} \lim_{t \rightarrow x} f_n(t)\end{aligned}$$

Is the limit of the integral the integral of the limits? Is the limit of the derivative the derivative of the limits? If f_n is continuous for all n , then the final proposed relationship asks if the limit of continuous functions is continuous, as we have $\lim_{t \rightarrow x} f_n(t) = f_n(x)$.

$$\lim_{n \rightarrow \infty} \lim_{t \rightarrow x} f_n(t) = \lim_{n \rightarrow \infty} f_n(x) = f(x) \stackrel{?}{=} \lim_{t \rightarrow x} f(t)$$

These are all related to the question first mentioned in Remark 4.5. When are we allowed to move limits in and out of operations? These particular cases happen to be of theoretical interest, as integration, differentiation, and limits are all limiting processes. In effect, we want to know when we can interchange these limiting processes with a limit. Can we do this with pointwise convergence. As the next three examples show, the answer is, in general, an emphatic **NO!**

Example 7.7 (Discontinuous Limit of Continuous Functions). We will restrict attention to $[0, 1] \subset \mathbb{R}$. Let $f_n(x) = x^n$. We have

$$\lim_{n \rightarrow \infty} f_n(x) = f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}.$$

One way to see this is consider the numerical sequence of $\{f_n(a)\}$ for a choice of $a \in [0, 1]$. If $a \in [0, 1)$, then $f_n(a) = a^n$ goes to zero. If $a = 1$, then $f_n(a) = 1$ which makes $\{f_n(a)\}$ a constant sequence. This explains why f is zero on $[0, 1)$ and 1 on $\{1\}$. Figure 75 shows this sequence of functions. Despite f_n being continuous on $[0, 1]$ for all n , we have the limit of $\{f_n\}$ is discontinuous on $[0, 1]$. In particular,

$$\lim_{t \rightarrow 1} f(t) = 0 \neq 1 = \lim_{n \rightarrow \infty} 1 = \lim_{n \rightarrow \infty} \lim_{t \rightarrow 1} f_n(t)$$

Example 7.8 (Integral of Limit \neq Limit of Integral). Consider the sequence of functions given by

$$f_n(x) = \begin{cases} n^2 x & \text{if } 0 \leq x < \frac{1}{n} \\ 2n - n^2 x & \text{if } \frac{1}{n} \leq x < \frac{2}{n} \\ 0 & \text{if } \frac{2}{n} \leq x \leq 1 \end{cases}.$$

We have $f_n \rightarrow f = 0$ on $[0, 1]$. Figure 76 shows what f_n looks like for three values of n . Calculating the integral of f_n for any n amounts to finding the area of a triangle of length $2/n$ and height $n^2(1/n)$.

$$\int_0^1 f_n(x) dx = \frac{1}{2} \left(\frac{2}{n} \left(n^2 \frac{1}{n} \right) \right) = 1$$

Therefore we have

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = \lim_{n \rightarrow \infty} 1 = 1 \neq 0 = \int_0^1 0 dx = \int_0^1 f(x) dx = \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx.$$

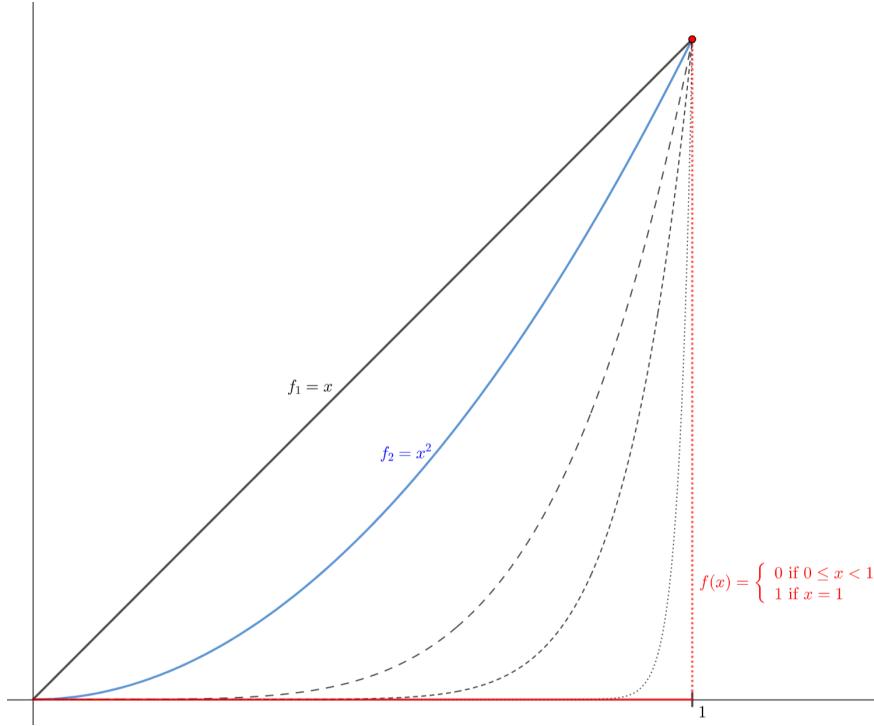


Figure 75: A sequence of continuous functions which converges to a discontinuous function.

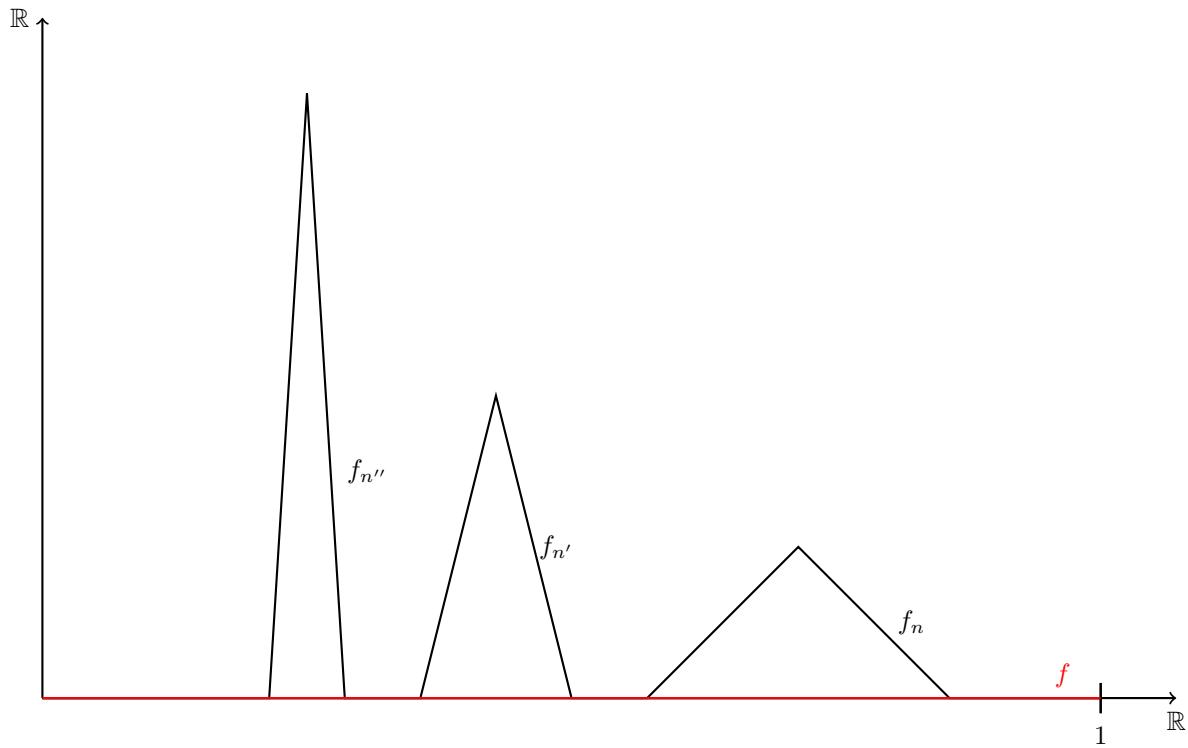


Figure 76: The sequence defined in Example 7.8 for whole numbers $n < n' < n''$.

Example 7.9 (Derivative of Limit \neq Limit of Derivative). Define $f_n(x) = (\sin nx)/x$. We have $f_n \rightarrow f$ where $f(x) = 0$. We also have

$$\lim_{n \rightarrow \infty} f'_n = \lim_{n \rightarrow \infty} \cos nx \neq 0 = f' = \left(\lim_{n \rightarrow \infty} f_n \right)'$$

If we want to apply the topics of the previous three sections to sequences of functions, we'll need a stronger notion of convergence.

7.3 Uniform Convergence

If you think return to the paragraph following Definition 7.5, you may wonder if certain functions converge such that one value of N only corresponds to ε and works for all $x \in E$. We saw this type of reasoning in the definition of uniform continuity. If a function is uniformly continuous on an interval, than for any $\varepsilon > 0$, there exists a single δ that gives $|f(x) - f(y)| < \varepsilon$ when $|x - y| < \delta$, regardless of the choice of x, y . Similarly, if we find that the N in Definition 7.5 only corresponds to ε and not x , i.e it works for all $x \in E$, we say f converges uniformly.

Definition 7.6. Let $E \subset \mathbb{R}$, and $f_n : E \rightarrow \mathbb{R}$ for all n . The sequence of functions $\{f_n\}$ *converges uniformly* to $f : E \rightarrow \mathbb{R}$ if for every $\varepsilon > 0$, there exists a $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon$ whenever $n \geq N$ for all $x \in E$.

Notation 7.1. There is no real standardized notation for uniform convergence. I'll opt to write $f_n \xrightarrow{\text{uni}} f$ to distinguish between the case of pointwise convergence which is notated as $f_n \rightarrow f$.

Just like how uniform continuity is a *much* stronger condition than continuity, uniform convergence is far stronger than pointwise convergence. One way to see this is to compare Figure 77 and Figure 74. If a sequence converges pointwise, it doesn't matter when f_n becomes arbitrarily close to f for different points $x \in E$. With uniform convergence, f_n must get arbitrarily close to f for all $x \in E$ simultaneously. The next example shows explicitly that N is not a function of x if $f_n \xrightarrow{\text{uni}} f$.

Example 7.10. Suppose $f_n = 1/n(1+x^2)$, and $f = 0$. We have $f_n \xrightarrow{\text{uni}} f$ on any interval $[a, b]$. Let $N = 1/\varepsilon$. Because $1/(1+x^2) < 1$, we have

$$|f_n(x) - f(x)| = \left| \frac{1}{n(1+x^2)} - 0 \right| = \frac{1}{n(1+x^2)} < \frac{1}{n} < \frac{1}{N} = \varepsilon$$

for all $n \geq N$. Our choice of N did not depend on x , hence f_n converges to f uniformly.

If $\{f_n\}$ converges uniformly and $\{g_n\}$ converges pointwise, but not uniformly,¹²⁴ then in theory, showing $\{g_n\}$ converges should be easier. In Example 7.10 we needed to apply the inequality $1/(1+x^2) < 1$ to justify our choice of N . We never needed to know any inequalities that would eliminate any reference to x when proving pointwise convergence. Nevertheless, uniform convergence becomes easier to verify due to our next results. The first is familiar, and eliminates the need for us to know the limit of f_n .

Theorem 7.1 (Cauchy Criterion for Uniform Convergence). Let $\{f_n\}$ be a sequence of real functions defined on E . The sequence converges *if and only if* for every $\varepsilon > 0$ there exists a $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \varepsilon$$

for all $x \in E$ when $m, n \geq N$.

¹²⁴We specify that $\{g_n\}$ does not converge uniformly, as every uniformly convergent sequence converges pointwise as well.

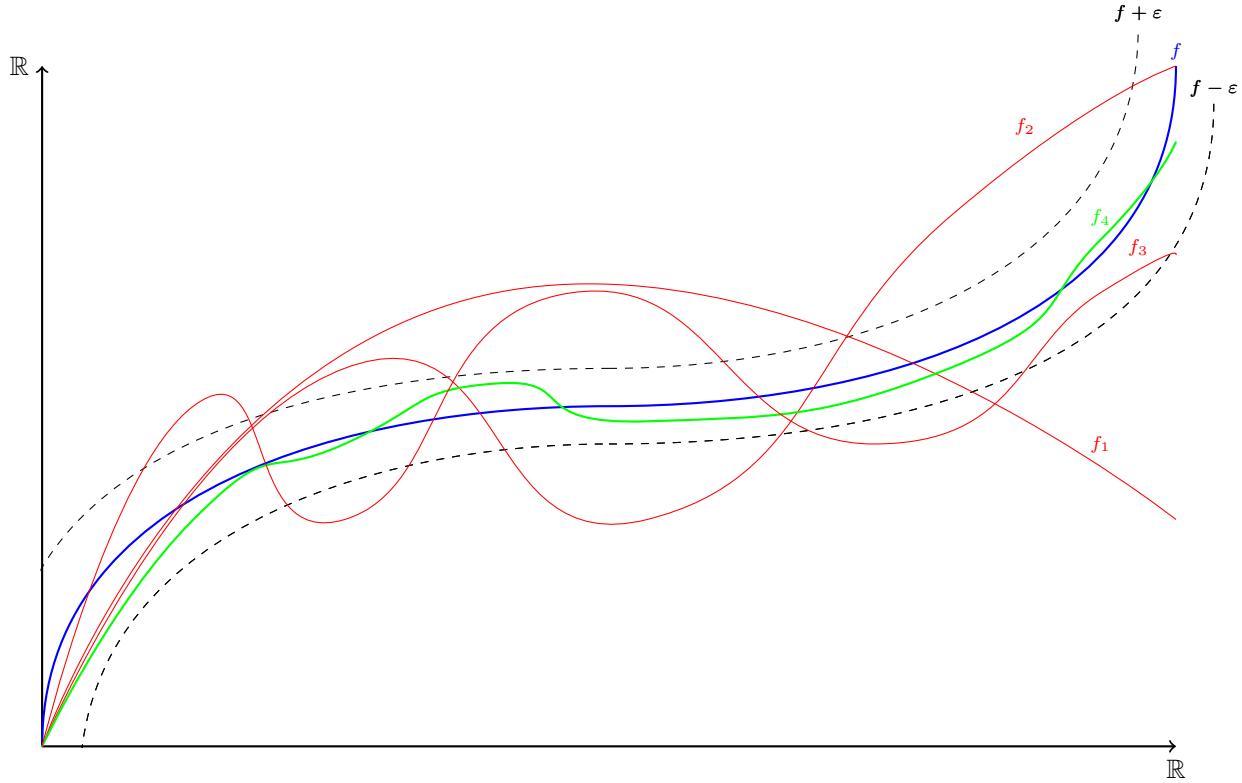


Figure 77: The sequence $\{f_n\}$ converges uniformly to f .

Proof.

(\Rightarrow) Suppose $f_n \xrightarrow{\text{uni}} f$ on E . There exists a $N \in \mathbb{N}$ such that

$$\begin{aligned}|f_n(x) - f(x)| &< \frac{\varepsilon}{2}, \\ |f_m(x) - f(x)| &< \frac{\varepsilon}{2},\end{aligned}$$

for all $x \in E$ whenever $n, m \geq N$. Therefore

$$|f_n(x) - f_m(x)| = |f_n(x) - f_m(x) + 0| = |f_n(x) - f_m(x) + f(x) - f(x)| \leq |f_n(x) - f(x)| + |f(x) - f_m(x)| < \varepsilon$$

for all $x \in E$ whenever $n, m \geq N$.

(\Leftarrow) Suppose for every $\varepsilon > 0$ there exists a $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \varepsilon$$

for all $x \in E$ when $m, n \geq N$. By our assumption, the Cauchy Criterion holds for the numerical sequence $\{f_n(x)\}$ for all $x \in E$, so for each x , $\{f_n(x)\}$ converges. Let $\{f(x)\}$ be the limit of $\{f_n(x)\}$ for each $x \in E$. But this gives that $f_n \rightarrow f$. We just need to show that this convergence is uniform.

For our assumed inequality, fix n and let $m \rightarrow \infty$, noting that $\lim_{m \rightarrow \infty} f_m(x) = f(x)$. This gives

$$\begin{aligned}\lim_{m \rightarrow \infty} |f_n(x) - f_m(x)| &\leq \lim_{m \rightarrow \infty} \varepsilon \\ |f_n(x) - f(x)| &\leq \varepsilon\end{aligned}$$

for every $n \geq N$ and all $x \in E$.¹²⁵

□

There exists a second criterion for uniform convergence that is of more interest, as it will lead to conclusions about the space $C_b(X)$ with the sup metric (Definition 7.4). Suppose we know that $f_n \rightarrow f$. Can we glean any additional information about f_n that would allow us to conclude $f_n \xrightarrow{\text{uni}} f$? The answer is found in the supremum of the distance between f_n and $f(x)$,

$$M_n = \sup_{x \in E} |f_n(x) - f(x)|.$$

This value M_n is shown in Figure 78. In this particular illustration, it appears that $M_n \rightarrow 0$ as $n \rightarrow \infty$. If

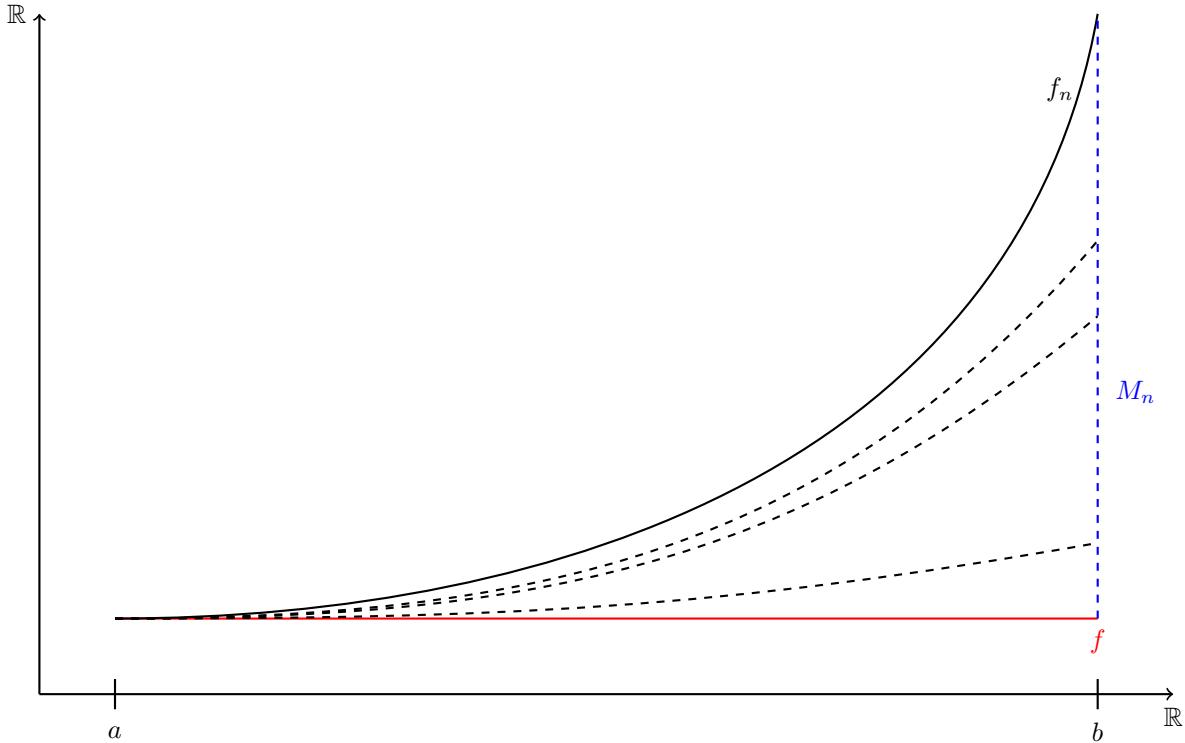


Figure 78: A sequence of functions $f_n \rightarrow f$ on $[a, b]$ where $M_n = \sup_{x \in [a, b]} |f_n(x) - f(x)|$.

this is the case, can we relate this to uniform convergence? As Figure 77 shows, the geometric interpretation of uniform convergence for a fixed ε is that we can always find some N such that the functions f_N, f_{N+1}, \dots , fall within a distance of ε from f on the entirety of E . As it turns out, if $M_n \rightarrow 0$, then we can do this! If $\sup_{x \in [a, b]} |f_n(x) - f(x)| < \varepsilon$, then it follows that $|f_n(x) - f(x)| < \varepsilon$ for all of $x \in E$, as $|f_n(x) - f(x)| \leq \sup_{x \in [a, b]} |f_n(x) - f(x)|$. As long as $M_n \rightarrow 0$, we will be able to do this for all ε , as we can find some large enough N where $M_n < \varepsilon$. This is the entire idea of the proof of the result, which will now be stated and proven formally.

¹²⁵How do we know taking the limit as $m \rightarrow \infty$ works? Well if $m \geq N$, then as we let it grow the inequality will always be satisfied.

Theorem 7.2. Suppose $f_n \rightarrow f$ on E .¹²⁶ Let

$$M_n = \sup_{x \in E} |f_n(x) - f(x)|.$$

We have $f_n \xrightarrow{\text{uni}} f$ if and only if $\lim_{n \rightarrow \infty} M_n = 0$.

Proof.

(\Rightarrow) If $f_n \xrightarrow{\text{uni}} f$ on E there exists an N such that

$$|f_n(x) - f(x)| < \varepsilon$$

for all $x \in E$ whenever $n \geq N$, which includes the value of x on which $\sup |f_n(x) - f(x)|$ is achieved.

Therefore for all ε there exists an $N \in \mathbb{N}$ such that

$$\sup_{x \in E} |f_n(x) - f(x)| < \varepsilon,$$

$$M_n < \varepsilon,$$

giving $M_n \rightarrow 0$.

(\Leftarrow) If $M_n \rightarrow 0$, then for all $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that

$$\begin{aligned} M_n &< \varepsilon \\ \implies \sup_{x \in E} |f_n(x) - f(x)| &< \varepsilon \\ \implies |f_n(x) - f(x)| &< \varepsilon \text{ (for all } x \in E\text{).} \end{aligned}$$

This is the definition of $f_n \xrightarrow{\text{uni}} f$.

□

Example 7.11. Let $f_n(x) = x/n$. We have $f_n \xrightarrow{\text{uni}} 0$ on $[a, b]$, as

$$\lim_{n \rightarrow \infty} \sup_{x \in [a, b]} |f_n(x) - f(x)| = \lim_{n \rightarrow \infty} \left| \frac{\max\{|a|, |b|\}}{n} \right| = 0.$$

Example 7.12. Let $f_n = 1/nx$. While $f_n \rightarrow 0$, the sequence $\{f_n\}$ does not converge uniformly on $(0, 1)$, as

$$\lim_{n \rightarrow \infty} \sup_{x \in (0, 1)} |f_n(x) - f(x)| = \lim_{n \rightarrow \infty} \sup_{x \in (0, 1)} \frac{1}{nx} = \lim_{n \rightarrow \infty} \left(\lim_{x \rightarrow 0} \frac{1}{nx} \right) \neq 0.$$

Corollary 7.1. Let f_n be a sequence of continuous real valued functions. We have $f_n \xrightarrow{\text{uni}} f$ in the Euclidean metric if and only if $f_n \rightarrow f$ in the metric space $C_b(X)$ equipped with the supremum norm.

Proof. If $f_n \rightarrow f$ in $C_b(X)$, then for all $\varepsilon > 0$ and $x \in X$ there exists some $N \in \mathbb{N}$ such that

$$d(f_n, f) = \sup_{x \in X} |f_n(x) - f(x)| = M_n < \varepsilon$$

for all $n \geq N$. This holds if and only if $\lim_{n \rightarrow \infty} M_n = 0$, which in turn is equivalent to $f_n \xrightarrow{\text{uni}} f$ in the Euclidean metric by Theorem 7.2. □

¹²⁶You actually don't need to assume that $f_n \rightarrow f$, but it doesn't really make sense to try and show uniform convergence with pointwise convergence.

Remark 7.1 (The Metric of Uniform Convergence). Corollary 7.1 tells us that convergence (as defined for generic metric spaces in Definition 3.2) in $C_b(X)$ with the sup metric is uniform convergence of functions in \mathbb{R} . This is a really cool result, as it shows that our conservative estimate of distance between elements in $C_b(X)$ given by the supremum metric is linked to the idea of a stronger form of convergence of functions in \mathbb{R} . For this reason, you sometimes will hear the supremum metric called the *uniform metric*, which may be written as

$$\|f - g\|_\infty = \|f - g\|_u.$$

7.4 Properties of Uniform Convergence

As it turns out, uniform convergence is just what we need for the three proposed equalities at the end of Page 144 to hold.¹²⁷

Theorem 7.3. Suppose $f_n \xrightarrow{\text{uni}} f$ on a metric space $E \subset \mathbb{R}$. Let x be a limit point of E , and suppose

$$\lim_{t \rightarrow x} f(t) = A_n.$$

Then $\{A_n\}$ converges, and

$$\lim_{t \rightarrow x} \lim_{n \rightarrow \infty} f_n(t) = \lim_{n \rightarrow \infty} A_n = \lim_{n \rightarrow \infty} \lim_{t \rightarrow x} f_n(t).$$

Proof. Fix $\varepsilon > 0$, and suppose $f_n \xrightarrow{\text{uni}} f$ on E . There exists an $N \in \mathbb{N}$ such that

$$|f_n(t) - f_m(t)| < \varepsilon$$

for all $t \in E$ when $n, m \geq N$. Taking the limit of both sides of this inequality as $t \rightarrow x$, we have

$$\begin{aligned} \lim_{t \rightarrow x} |f_n(t) - f_m(t)| &< \lim_{t \rightarrow x} \varepsilon \\ |A_n - A_m| &< \varepsilon. \end{aligned}$$

Therefore $\{A_n\}$ is a Cauchy sequence in \mathbb{R} , so it converges to some $A \in E$. Now we need to show the limits may be interchanged.

Using the uniform convergence of f_n , we can choose a n such that

$$|f_n(t) - f(t)| < \frac{\varepsilon}{3}, \tag{22}$$

for all $t \in E$, and such that

$$|A_n - A| < \frac{\varepsilon}{3} \tag{23}$$

by the convergence of $\{A_n\}$. For this same n we may choose some open ball $B_\delta(x)$ such that

$$|f_n(t) - A_n| < \frac{\varepsilon}{3} \tag{24}$$

¹²⁷...kind of. We will need to add a single assumption to uniform convergence for the limit of the derivatives to be the derivative of the limit.

if $t \in B_\delta(x) \cap E$, and $t \neq x$.¹²⁸ If we combine (22), (23), and (24), then we can conclude that for all ε there exists a $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\begin{aligned} |f(t) - A| &= |f(t) - A + 0 + 0| \\ &= |f(t) - A + (f_n(t) - f_n(t)) + (A_n - A_n)| \\ &= |(f_n(t) - f(t)) - (A_n - A) - (f_n(t) - A)| \\ &< |(f_n(t) - f(t))| + |(A_n - A)| + |(f_n(t) - A)| \\ &= \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon. \end{aligned}$$

This shows the desired result. \square

Corollary 7.2 (Uniform Convergence Preserves Continuity). If $\{f_n\}$ is a sequence of continuous functions on $E \subset \mathbb{R}$, and if $f_n \xrightarrow{\text{uni}} f$ on E , then f is continuous on E .

Proof. This follows from the fact the alternate definition of continuity given by Corollary 4.2. Each function f_n is continuous, so we can find the limit as $t \rightarrow x$ by evaluating at f at x . Combining this with the equality given in Theorem 7.3 gives the result.

$$\begin{aligned} \lim_{t \rightarrow x} \lim_{n \rightarrow \infty} f_n(t) &= \lim_{n \rightarrow \infty} \lim_{t \rightarrow x} f_n(t) \\ \lim_{t \rightarrow x} f(t) &= \lim_{n \rightarrow \infty} f_n \left(\lim_{t \rightarrow x} t \right) \\ \lim_{t \rightarrow x} f(t) &= \lim_{n \rightarrow \infty} f_n(x) \\ \lim_{t \rightarrow x} f(t) &= f(x) \end{aligned}$$

\square

Example 7.13. By Corollary 7.2, if the limit of a sequence of continuous functions is not continuous, then that sequence must not converge uniformly. This means that $f_n(x) = x^n$ does not converge uniformly on $[0, 1]$ by Example 7.7.

Example 7.14 (The Converse is False). If a sequence of continuous functions converges to a continuous function, this does not imply the sequence converges uniformly. Let $f_n(x) = n^2 x (1 - x^2)^n$ on the interval $(0, 1]$. For all values of $n \in \mathbb{N}$, f_n is continuous on $(0, 1]$. We also have a continuous limit, as $\lim_{n \rightarrow \infty} f_n(x) = 0$ on $(0, 1]$. Nevertheless, applying Theorem 7.2 shows that this sequence does not converge uniformly.

$$\lim_{n \rightarrow \infty} \sup_{x \in (0, 1]} |f_n(x) - f(x)| = \lim_{n \rightarrow \infty} \left(\frac{1}{\sqrt{2n-1}} \right) \not\rightarrow 0.$$

Remark 7.2. There are certain conditions where the converse of Corollary 7.2 does hold. These conditions are given as Theorem 7.13 in [Rudin \(1976\)](#). As it turns out, the previous example works because we restricted our attention to $(0, 1]$, an interval which is not compact.

We now have the information necessary to prove a very nice property of the set of functions $C_b(X)$. Recall that a metric space is complete if every Cauchy sequence converges. Until now, we've only known one

¹²⁸Instead of using an open ball around x , we could say there exists some δ such that $|f_n(t) - A_n| < \varepsilon/3$ for all $t \in E$ which satisfy $|t - x| < \delta$. Either way, all we're doing is using the definition of $\lim_{t \rightarrow x} f_n(t) = A$.

complete metric space, namely \mathbb{R} . Example 3.28 claimed that the set of function $C([a, b])$ was complete, but at the time this could not be proved. Now we will prove this.¹²⁹

Theorem 7.4. ($C_b(X)$ is Complete) Suppose $X \subset \mathbb{R}$. The set of functions $C_b(X)$ is complete.

Proof. Let $\{f_n\}$ be a Cauchy sequence in $C_b(X)$. By Theorem 7.1, the function converges uniformly to some limit f with domain X . All that is left to show is that $f \in C_b(X)$, but this is a consequence of Corollary 7.2. Each f_n is continuous, and the sequence converges uniformly, so f is also continuous. By the definition of uniform convergence, there exists an $n \in \mathbb{N}$ such that $|f(x) - f_n(x)| < 1$ for all $x \in X$. Since $f_n(x)$ is bounded for all n , f must be bounded, giving $f \in C_b(X)$. \square

Putting continuity aside, we can continue to uniform convergence's behavior with respect to integration.

Theorem 7.5 (Uniform Convergence Preserves Integration). Suppose f_n is Riemann integrable on $[a, b]$ for all n , and $f_n \xrightarrow{\text{uni}} f$ on $[a, b]$. Then f is Riemann integrable on $[a, b]$, and

$$\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \int_a^b f(x) dx.$$

Proof. First, we will show that f is integrable on $[a, b]$. Define

$$\varepsilon_n = \sup_{x \in [a, b]} |f_n(x) - f(x)|.$$

This definition gives

$$f_n - \varepsilon_n \leq f \leq f_n + \varepsilon_n,$$

as ε_n is the least upper bound of the distance between f_n and f . Applying Theorem 6.2 to this inequality gives

$$\int_a^b (f_n - \varepsilon_n) dx \leq \int_a^b f dx \int_a^b f dx \leq \int_a^b (f_n + \varepsilon_n) dx. \quad (25)$$

This implies

$$\begin{aligned} \int_a^b f dx - \int_a^b f dx &\leq \int_a^b (f_n + \varepsilon_n) dx - \int_a^b (f_n - \varepsilon_n) dx \\ &= \int_a^b (f_n + \varepsilon_n) - (f_n - \varepsilon_n) dx \\ &= \int_a^b 2\varepsilon_n dx \\ &= 2\varepsilon_n(b - a). \end{aligned}$$

By uniform converge, $\varepsilon_n \rightarrow 0$ (Theorem 7.2), so

$$\int_a^b f dx - \int_a^b f dx \leq 0$$

for all n . The upper and lower integrals must be equal, which is the definition of Riemann integrability.

¹²⁹ $C([a, b]) = BC([a, b])$, as every continuous function on a compact domain is bounded.

Equation (25) also gives the desired equality.

$$\begin{aligned}
\int_a^b f \, dx &\leq \int_a^b (f_n + \varepsilon_n) \, dx \\
&\leq \int_a^b f_n \, dx + \int_a^b \varepsilon_n \, dx \\
\left| \int_a^b f \, dx - \int_a^b f_n \, dx \right| &\leq \varepsilon_n(b-a) \\
\lim_{n \rightarrow \infty} \left| \int_a^b f \, dx - \int_a^b f_n \, dx \right| &\leq \lim_{n \rightarrow \infty} \varepsilon_n(b-a) \\
\left| \int_a^b f \, dx - \lim_{n \rightarrow \infty} \int_a^b f_n \, dx \right| &\leq 0.
\end{aligned}$$

Therefore we have

$$\lim_{n \rightarrow \infty} \int_a^b f_n(x) \, dx = \int_a^b f(x) \, dx.$$

□

Having seen that continuity and integration play well with uniform convergence, it is tempting to assume the same holds for differentiation. Unfortunately, as the next example shows, this is not true! As far as properties of functions go, differentiability is fairly demanding. This means we will need to make assumptions to reach any result about uniform convergence and derivatives.

Example 7.15 (Uniform Convergence Does Not Preserve Differentiation). Let

$$f_n(x) = \frac{\sin nx}{\sqrt{n}}.$$

We have that $f_n(x) \xrightarrow{\text{uni}} 0$ as

$$\lim_{n \rightarrow \infty} \sup_{x \in [a, b]} |f_n(x) - f(x)| = \lim_{n \rightarrow \infty} \sup_{x \in [a, b]} \left| \frac{\sin nx}{\sqrt{n}} \right| = \lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} = 0.$$

Despite having uniform convergence, we have that

$$\lim_{n \rightarrow \infty} f'_n(0) = \lim_{n \rightarrow \infty} \sqrt{n} \cos(n \cdot 0) \neq 0 = \lim_{n \rightarrow 0} 0 = \lim_{n \rightarrow \infty} f'(0).$$

It is also possible that the derivative of a uniform limit may not exist. Let

$$f_n(x) = \sqrt{\frac{1}{n^2} + x^2}.$$

We have that $f_n(x) \xrightarrow{\text{uni}} |x|$, so $f'(0)$ does not exist. This happens despite the fact that $f'_n(0)$ exists for all n .

The next theorem will give the conditions under which uniform convergence will preserve differentiation, but they are fairly strong. For one, the sequence $\{f'_n\}$ must converge uniformly. Not only that, but the sequence $\{f_n\}$ will need to converge pointwise for at least one point.

Theorem 7.6 (Uniform Convergence and Differentiation). Suppose $\{f_n\}$ is a sequence of functions, differentiable on $[a, b]$ and such that $\{f(x_0)\}$ converges pointwise for some $x_0 \in [a, b]$. If $\{f'_n\}$ converges uniformly on $[a, b]$, then $\{f_n\}$ converges uniformly to some function f on $[a, b]$, and

$$f'(x) = \lim_{n \rightarrow \infty} f'_n(x)$$

on $[a, b]$.

Proof. First, we will show that $\{f_n\}$ converges uniformly to f . The sequence $\{f_n(x_0)\}$ converges, so it must be a Cauchy sequence. Similarly, $\{f'_n\}$ converges uniformly, so it is also a Cauchy sequence. This means for all $\varepsilon > 0$, there exists an N such that for all $n, m \geq N$ we have

$$|f_n(x_0) - f_m(x_0)| \leq \frac{\varepsilon}{2}, \quad (26)$$

$$|f_n(t)' - f_m'(t)| \leq \frac{\varepsilon}{2(b-a)}. \quad (27)$$

Applying the Mean Value Theorem to the function $f_n - f_m$, there exists an element $t_0 \in [x, t]$ such that

$$|f'_n(t_0) - f'_m(t_0)| = \frac{|f_n(x) - f_m(x) - (f_n(x) - f_m(x))|}{|x - t|}.$$

This combined with (27) gives

$$\begin{aligned} |f_n(x) - f_m(x) - f_n(t) + f_m(t)| &= |x - t| \cdot |f'_n(t_0) - f'_m(t_0)|, \\ &\leq |x - t| \cdot \frac{\varepsilon}{2(b-a)} \end{aligned} \quad (28)$$

$$\begin{aligned} &= \frac{|x - t|}{b - a} \cdot \frac{\varepsilon}{2}, \\ &\leq \frac{\varepsilon}{2}, \end{aligned} \quad (29)$$

for any $x, t \in [a, b]$ if $n, m \geq N$.¹³⁰ Combing (27) and (29) while invoking the triangle inequality gives

$$\begin{aligned} |f_n(x) - f_m(x)| &\leq |f_n(x) - f_m(x) + 0 + 0| \\ &= |f_n(x) - f_m(x) + (f_n(x_0) - f_n(x_0)) + (f_m(x_0) - f_m(x_0))| \\ &\leq |f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0)| + |f_n - f_m(x_0)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon \end{aligned}$$

for all $x \in [a, b]$ when $n, m \geq N$. The sequence f satisfies the Cauchy Criterion for uniform convergence, so $f_n \xrightarrow{\text{uni}} f$ on $[a, b]$.

Now we will show that under our assumptions, the derivative of the limit is the limit of the derivatives. Fix a point $x \in [a, b]$ and define

$$\begin{aligned} \phi_n(t) &= \frac{f_n(t) - f_n(x)}{t - x}, \\ \phi(t) &= \frac{f(t) - f(x)}{t - x}, \end{aligned}$$

for $x \in [a, b]$ where $x \neq t$. By (28),

$$|\phi_n(t) - \phi_m(t)| \leq \frac{\varepsilon}{2(b-a)}$$

for all $x \in [a, b]$ when $n, m \geq N$. By the Cauchy Criterion for uniform convergence, ϕ_n converges uniformly for $t \neq x$. Since $f_n \rightarrow f$, the uniform limit of $\{\phi_n\}$ must be ϕ .

$$\lim_{n \rightarrow \infty} \phi_n(t) = \phi(t) \quad (a \leq t \leq b, t \neq x)$$

¹³⁰The final line of this inequality follows from the fact that $|x - t| \leq b - a$, because $x, t \in [a, b]$.

Because $\phi_n \xrightarrow{\text{uni}} \phi$ we can interchange limits using Theorem 7.3 to reach the desired equality.

$$f'(x) = \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} = \lim_{t \rightarrow x} \phi(t) = \lim_{t \rightarrow x} \lim_{n \rightarrow \infty} \phi_n(t) = \lim_{n \rightarrow \infty} \lim_{t \rightarrow x} \phi_n(t) = \lim_{n \rightarrow \infty} \lim_{t \rightarrow x} \frac{f_n(t) - f_n(x)}{t - x} = \lim_{n \rightarrow \infty} f'_n(x)$$

□

in

7.5 Approximation with Polynomials

As stated earlier, uniform convergence is useful because we may want to take the limit of integral of a sequence of functions' limit, but not know the limit. With uniform convergence, this is no problem, as we can interchange limits freely. It still could be the case that f_n is difficult to work with. It's worth asking, given some function f can we find a sequence of "easy to work with" functions which converges uniformly to f ? In this case, we know the desired limit of the sequence, but are looking for the sequence.

Some of the easiest functions to work with our polynomials. As it turns out, the answer to this question becomes yes if we look for a sequence of polynomials. This was first shown by Weierstrass, and was briefly introduced in Example 2.29.

Theorem 7.7 (Weierstrass Approximation Theorem). If f is a real continuous function on the interval $[a, b]$, then there exists a sequence of real polynomials $\{P_n\}$ such that $P_n(x) \xrightarrow{\text{uni}} f(x)$ on $[a, b]$.

Proof. First, restrict attention to the case where $[a, b] = [0, 1]$ and $f(0) = f(1) = 0$. We will construct the desired sequence $\{P_n\}$.

Define the function $Q_n(x)$ as

$$Q_n(x) = c_n(1 - x^2)^n$$

where

$$c_n = \left(\int_{-1}^1 (1 - t^2)^n dt \right)^{-1}.$$

This choice of c_n is made such that the integral of $Q_n(x)$ on $[-1, 1]$ is 1 for all values of n . ¹³¹

$$\int_{-1}^1 Q_n(x) dx = c_n \int_{-1}^1 (1 - x^2)^n dx = \frac{\int_{-1}^1 (1 - x^2)^n dx}{\int_{-1}^1 (1 - t^2)^n dt} = 1 \quad (30)$$

Figure 79 shows $Q_n(x)$ for a few values of n . This seemingly random function $Q_n(x)$ will be used to construct $\{P_n\}$, so we may want to know some information about c_n 's magnitude that can be used to show $P_n \xrightarrow{\text{uni}} f$. Using the fact that $(1 - x^2)^n \geq 1 - nx^2$, $1 \leq 1/\sqrt{n}$, the symmetry of $Q_n(x)$, and integration, we have

$$\begin{aligned} c_n &= \left(\int_{-1}^1 (1 - t^2)^n dt \right)^{-1} = \left(2 \int_0^1 (1 - t^2)^n dt \right)^{-1} \leq \left(2 \int_0^{1/\sqrt{n}} (1 - t^2)^n dt \right)^{-1} \leq \left(2 \int_0^{1/\sqrt{n}} 1 - nt^2 dt \right)^{-1} \\ &= \frac{3}{4}\sqrt{n} \\ &< \sqrt{n}. \end{aligned}$$

This bound on c_n gives

$$Q_n(x) \leq \sqrt{n}(1 - \delta^2)^n \quad (31)$$

¹³¹In this sense, c_n is just a scaling factor which standardizes $Q_n(x)$ so this integral is always 1.

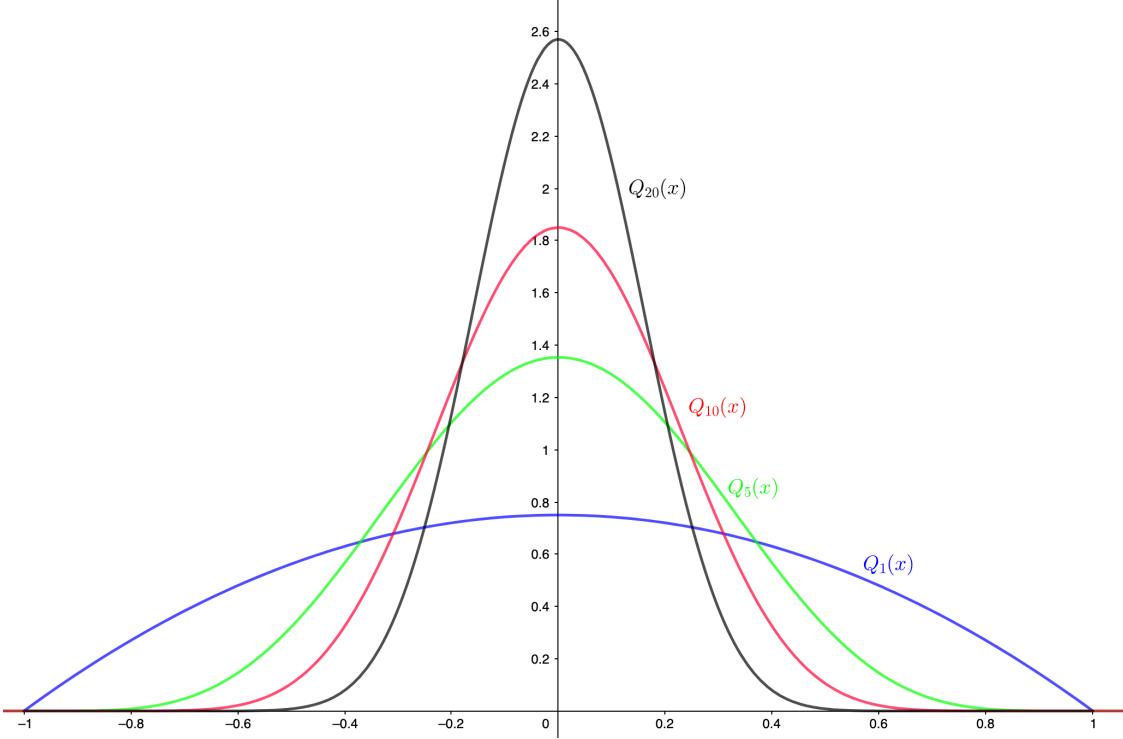


Figure 79: The function $Q_n(x)$ on the interval $[-1, 1]$ for several values of n .

for any $\delta \leq |x|$.

Now we can define our sequence of polynomials to be

$$P_n(x) = \int_{-1}^1 f(x+t)Q_n(t) dt.$$

Due to the fact that f is zero outside the interval $[0, 1]$ we know $f(x+t)$ is zero for t outside the interval $[-x, 1-x]$, so we can rewrite the bounds of integration in the definition of $P_n(x)$.

$$P_n(x) = \int_{-x}^{1-x} f(x+t)Q_n(t) dt$$

If we apply a change of variables (Theorem 6.7) using $u = x + t$ we get

$$P_n(x) = \int_{-x}^{1-x} f(x+t)Q_n(t) dt = \int_0^1 f(u)Q_n(u-x) du.$$

At this point it's worth asking how we know $P_n(x)$ is actually a polynomial. It becomes a bit clearer when writing

$$P_n(x) = \int_0^1 f(u)c_n[1-(t-x)^2]^n dt.$$

The Binomial theorem could be used to expand the integrand, and any term with a t will be integrated and becomes a constant. What remains will be a polynomial in x .

We are now ready to show that $P_n \xrightarrow{\text{uni}} f$. By the continuity of f , for all $\varepsilon > 0$, there exists a corresponding $\delta > 0$ such that $|y - x| < \delta$ implies $|f(y) - f(x)| < \varepsilon/2$. Let $M = \sup |f(x)|$. By (30), (31), and the fact that

$Q_n(x) \geq 0$, we see that for any $x \in [0, 1]$,

$$\begin{aligned}
|P_n(x) - f(x)| &= \left| \int_{-1}^1 f(x+t)Q_n(t) dt - f(x) \right| \\
&= \left| \int_{-1}^1 f(x+t)Q_n(t) dt - f(x) \cdot 1 \right| \\
&= \left| \int_{-1}^1 f(x+t)Q_n(t) dt - f(x) \int_{-1}^1 Q_n(t) dt \right| \\
&= \left| \int_{-1}^1 [f(x+t) - f(x)]Q_n(t) dt \right| \\
&\leq \int_{-1}^1 |f(x+t) - f(x)|Q_n(t) dt \\
&= \int_{-1}^{-\delta} \underbrace{|f(x+t) - f(x)|}_{\leq \sup |f(x)| + \sup |f(x)|} Q_n(t) dt + \int_{-\delta}^{\delta} \underbrace{|f(x+t) - f(x)|}_{\leq \varepsilon/2} Q_n(t) dt + \int_{\delta}^1 \underbrace{|f(x+t) - f(x)|}_{\leq \sup |f(x)| + \sup |f(x)|} Q_n(t) dt \\
&\leq 2M \int_{-1}^{-\delta} Q_n(t) dt + \frac{\varepsilon}{2} \int_{-\delta}^{\delta} Q_n(t) dt + 2M \int_{\delta}^1 Q_n(t) dt \\
&\leq 4M\sqrt{n}(1 - \delta^2)^n + \frac{\varepsilon}{2} \\
&< \varepsilon
\end{aligned}$$

for a sufficiently large n . This means that $P_n \xrightarrow{\text{uni}} f$ on $[0, 1]$.

Finally, suppose f is any real valued continuous function on $[a, b]$. We can extend our result by writing

$$g(x) = f(a + (b - a)x)$$

so that g is defined on $[0, 1]$. We know the Theorem holds for this simpler case, so there exists a sequence of P_n such that $P_n \xrightarrow{\text{uni}} g$ on $[0, 1]$. Since

$$f(x) = g\left(\frac{x-a}{b-a}\right),$$

we have a sequence of polynomials which converge uniformly to f in the form of $P_n\left(\frac{x-a}{b-a}\right)$. □

At first glance, Weierstrass' Approximation Theorem may not seem useful at all. While polynomials are easy to take limits of, integrate, and differentiate, we know f to begin with. This means that it would be easier to just take the integral of f instead of the limit of the integrals of the more complicated P_n 's if we were interested in this value. Nevertheless this result can be useful when proving other results, as we know polynomials satisfy a myriad of desirable properties.

Example 7.16. Define f to be

$$f(x) = \begin{cases} \frac{1}{2} - |\frac{1}{2} - x| & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

If we define

$$P_n(x) = \int_0^1 f(t)Q_n(t-x) dt,$$

we should have that $P_n(x) \xrightarrow{\text{uni}} f$ on $[0, 1]$. Figure 80 shows this convergence. If we want to be more explicit,

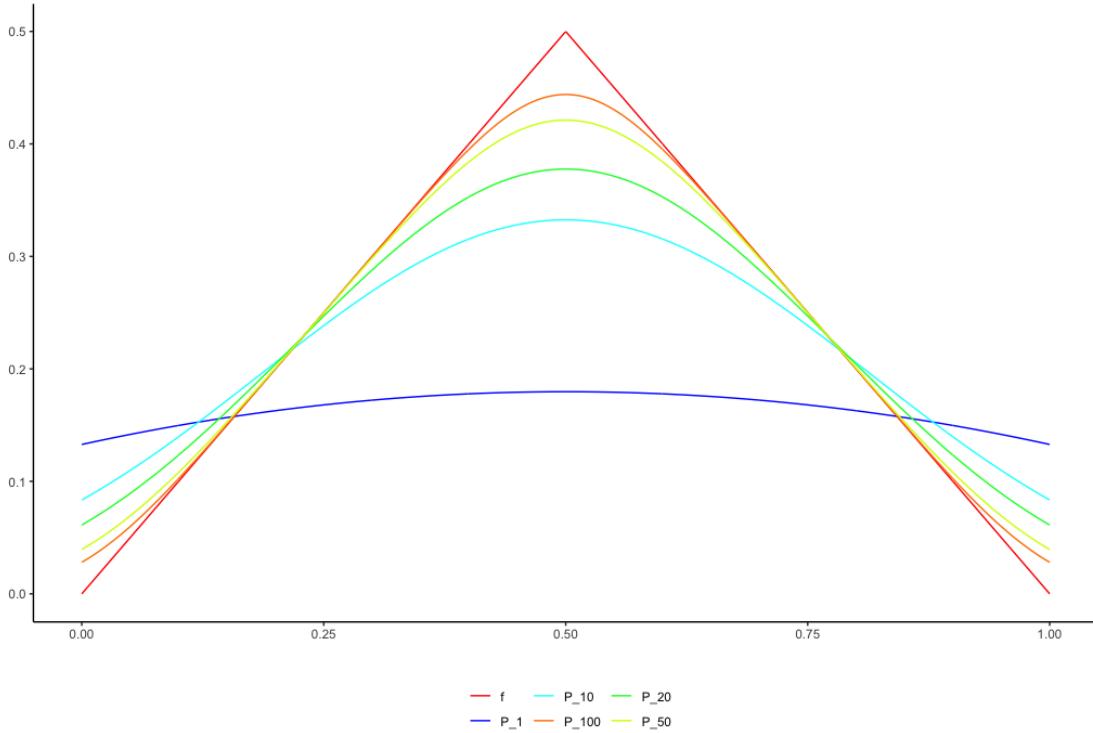


Figure 80: The Weierstrass Approximation Theorem applied to f in Example 7.17.

we can write out P_n to a limited extent.

$$\begin{aligned}
 P_n(x) &= \int_0^1 f(t)Q_n(t-x) dt \\
 &= \int_0^1 \left(\frac{1}{2} - \left| \frac{1}{2} - x \right| \right) c_n [1 - (t-x)^2]^n dt \\
 &= c_n \int_0^1 \left(\frac{1}{2} - \left| \frac{1}{2} - t \right| \right) \sum_{k=0}^n (-t^2 - 2xt - x^2)^k dt
 \end{aligned}$$

At this point we could use the multinomial theorem to expand things further, but it gets fairly ugly. If you use Mathematica, Python with SymPy, or any other symbolic computation platform, you can see the coefficients of $P_n(x)$ for any n . For example, for $n = 1$ we have

$$P_1(x) = -\frac{3}{16}x^2 + \frac{3}{16}x + \frac{17}{128}$$

Remark 7.3 (Stone-Weierstrass). This approximation theorem would go on to be generalized several times. Not only can this result be generalized to complex functions, but it can be generalized beyond real intervals. A more general version is due to Marshall Stone and is known as the Stone-Weierstrass Theorem, and works in a vastly more general topological setting than that of the real line.

7.6 Series

Each sequence of functions $\{f_n\}$ gives rise to a series of functions which results from summing f_n over n . All the properties related to uniform convergence and sequences will follow for series as well.

Definition 7.7. A series of functions $\sum f_n(x)$ converges (pointwise) on $E \subset \mathbb{R}$ if the numerical series $\sum f_n(x_0)$ converges for all $x_0 \in E$. Alternatively, $\sum f_n(x)$ converges on $E \subset \mathbb{R}$ if the sequence of partial sums defined as

$$s_n(x) = \sum_{k=1}^n f_k(x)$$

converges on E . In this case we write $\sum f_n(x) = f$, and call f the sum of the series.¹³²

Definition 7.8. A series of functions $\sum f_n(x)$ converges uniformly on $E \subset \mathbb{R}$ if the sequence of partial sums $\{s_n\}$ converges uniformly on E .

Verifying that a series converges uniformly is slightly different than verifying a sequence converges uniformly. We will turn to a test that is analogous to the comparison test for numerical series.

Theorem 7.8 (Weierstrass M-Test). Suppose $\{f_n\}$ is a sequence of functions defined on $E \subset \mathbb{R}$, and suppose $|f_n(x)| \leq M_n$ for all $x \in E$ and $n \in \mathbb{N}$. Then $\sum f_n$ converges uniformly on E if $\sum M_n$ converges.

Proof. Suppose that $\sum M_n$ converges. The Cauchy Criterion holds for the partial sums of M_n , so for any $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$\left| \sum_{k=1}^n M_k - \sum_{k=1}^m M_k \right| = \sum_{k=n}^m M_k < \varepsilon.$$

Because $|f_n(x)| \leq M_n$ for all $x \in E$, we have

$$|s_n(x) - s_m(x)| = \left| \sum_{k=1}^n f_k(x) - \sum_{k=1}^m f_k(x) \right| = \left| \sum_{k=n}^m f_k(x) \right| \leq \sum_{k=n}^m M_k < \varepsilon$$

for $n, m \geq N$. This makes $s_n(x)$ a Cauchy sequence, so it converges uniformly by Theorem 7.1. This gives the uniform convergence of $\sum f_n(x)$. \square

Much like the Cauchy Criterion, the Weierstrass M-Test does not necessitate knowledge of the actual limit.

Example 7.17. Define f_n to be

$$f_n(x) = \frac{1}{x^2 + n^2}.$$

We have that

$$f_n(x) = \frac{1}{x^2 + n^2} \leq \frac{1}{n^2} = M_n$$

for each n and all $x \in \mathbb{R}$. Because the sequence $\sum M_n = \sum 1/n^2$ converges, the series $\sum f_n(x)$ converges by the Weierstrass M-Test.

Example 7.18 (Converse of Weierstrass M-Test Fails). The converse of Theorem 7.8 need not hold. A somewhat trivial example is found in the constant function $f_n(x) = (-1)^n/n$. The constant series of functions given by $\sum f_n$ converges uniformly to $\ln 2$. Despite this, there exists no bound on $|f_n|$ which forms a convergent numerical series. The smallest possible value of M_n is $1/n$, but $\sum 1/n$ diverges. Any example of a numerical series which converges but fails to converge absolutely would also work.

Theorem 7.9 (Uniform Convergence and Series). Let $\sum f_n$ be a series of functions on the set $E \subset \mathbb{R}$.

¹³²You can assume that $\sum f_n(x) = \sum_{n=1}^{\infty} f_n(x)$.

1. If f_n is continuous for all n , and $\sum_n f_n$ converges uniformly, then $\sum_n f_n$ is continuous.
2. If f_n is Riemann integrable for all n on $[a, b] \subset E$, and $\sum_n f_n$ converges uniformly on $[a, b]$, then $\sum_n f_n$ is Riemann integrable on $[a, b]$, and

$$\int_a^b \sum_{n=1}^{\infty} f_n(x) dx = \sum_{n=1}^{\infty} \int_a^b f_n(x) dx$$

3. Suppose f_n is differentiable on E for all n , $\sum f_n(x_0)$ converges pointwise for some $x_0 \in E$, and $\sum f'_n$ converges uniformly. Then $\sum f_n$ converges uniformly, and

$$\frac{d}{dx} \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \frac{df_n}{dx}$$

Proof. The proofs for these follow from applying Theorem 7.3, Theorem 7.5, and Theorem 7.6 to the sequence of partial sums $\{s_n\}$. \square

Example 7.19. (Weierstrass Function) We can now construct one of the most famous pathological functions in math. Using series of functions, we can construct a function which is continuous on all of \mathbb{R} but fails to be differentiable at a single point. Begin by defining $\varphi(x) = |x|$ on $[-1, 1]$, and extend it to the real line by requiring that $\varphi(x+2) = \varphi(x)$.¹³³

Define the function

$$f(x) = \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n \varphi(4^n x).$$

Figure 81 illustrates f on the interval $[-10, 10]$. We have that $|\varphi(4^n x)| \in [0, 1]$ implying that

$$\left(\frac{3}{4}\right)^n \varphi(4^n x) \leq \left(\frac{3}{4}\right)^n.$$

The series $\sum(3/4)^n$ converges, so by The Weierstrass M-Test, f converges uniformly. Furthermore f is continuous as it is a uniformly convergent series of continuous functions (Theorem 7.9).

Now fix an arbitrary $x \in \mathbb{R}$, and consider whether f is differentiable at x . For an arbitrary $m \in \mathbb{Z}^+$ define

$$\delta_m = \begin{cases} (1/2)4^{-m} & \text{if } \nexists k \in \mathbb{Z} \text{ s.t } 4^m x \leq k \leq 4^m(x + (1/2)4^{-m}) \\ -(1/2)4^{-m} & \text{if } \nexists k \in \mathbb{Z} \text{ s.t } 4^m x \leq k \leq 4^m(x - (1/2)4^{-m}) \end{cases}.$$

For example, if $x = 1/3$ and $m = 1$, then $\delta_m \in \{-1/8, 1/8\}$. To determine the sign we look at the following intervals:

$$\begin{aligned} [4^m x, 4^m(x + (1/2)4^{-m})] &= [4/3, 4(1/3 + 1/8)] = [8/6, 11/6] \\ [4^m x, 4^m(x - (1/2)4^{-m})] &= [4/3, 4(1/3 - 1/8)] = [8/6, 5/6] \end{aligned}$$

The first interval contains no integer, so we take $\delta_m = 1/8$. In general, at least one of these intervals will contain an integer, as the length of the intervals is

$$|4^m x - 4^m(x + (1/2)4^{-m})| = |4^m x - 4^m(x - (1/2)4^{-m})| = 4^m(1/2)4^{-m} = 1/2,$$

¹³³Formally, $\varphi(x) = \begin{cases} 2 - [(x+2) \bmod 2] & \text{if } [(x+2) \bmod 2] > 1 \\ [(x+2) \bmod 2] & \text{otherwise} \end{cases}$

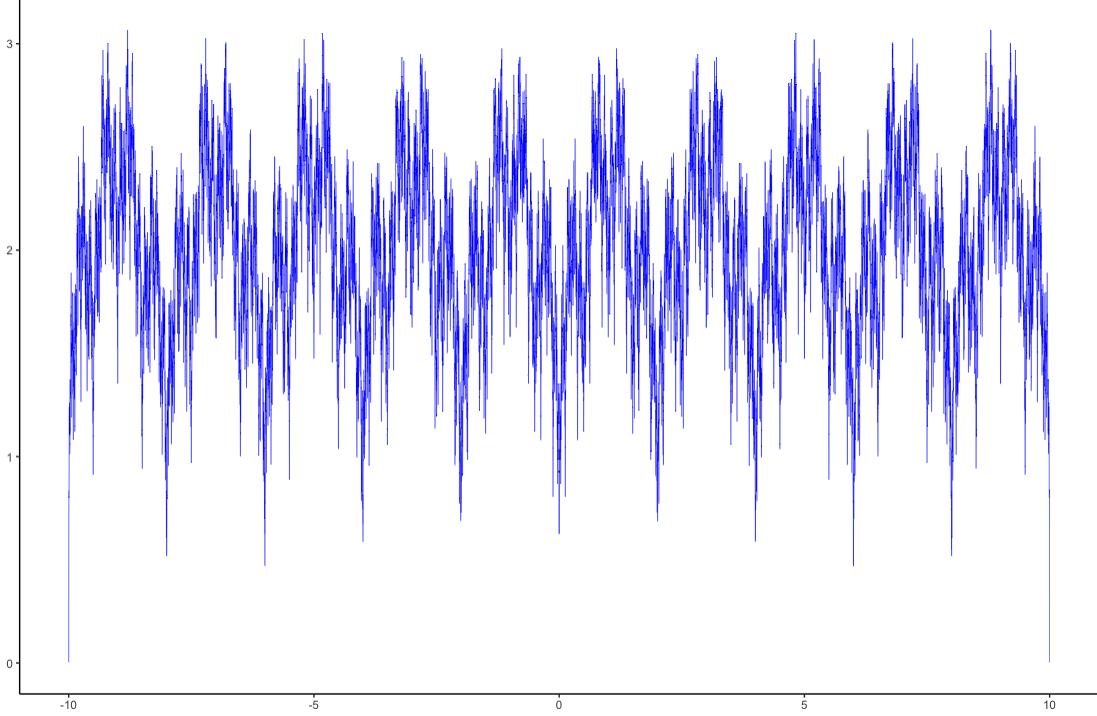


Figure 81:

so our intervals are $[4^m x - 1/2, 4^m x]$ and $[4^m x, 4^m x + 1/2]$. At least one of these intervals must contain an integer, as there is an integer with a distance of $1/2$ from all real numbers.

Define the quotient

$$\gamma_n = \frac{\varphi(4^n(x + \delta_m)) - \varphi(4^n x)}{\delta_m},$$

where n is the index used in the summation which defines f . Consider the cases where $n > m$ and $0 \leq n \leq m$ with concrete numbers. If $n = 2$, $m = 1$, and $x = 1/3$ then

$$\gamma_n = \frac{\varphi(4^2(1/3 + 1/8)) - \varphi(4^2(1/3))}{1/8} = \frac{\varphi(22/3) - \varphi(16/3)}{1/8} = \frac{2/3 - 2/3}{1/8} = 0.$$

By construction, ,

$$4^n(x + \delta_m) - 4^n x = 4^n \delta_m = 4^n(\pm(1/2)4^{-m}) = \pm 2^{2(n-m)-1}$$

is divisible by 2 whenever $n > m$. One of the defining properties of φ is that $\varphi(x) = \varphi(x+2)$ for all $x \in \mathbb{R}$, so $\varphi(4^n(x + \delta_m)) = \varphi(4^n x)$ in this case, giving $\gamma_n = 0$. Consider the other case where $0 \leq n \leq m$. For any $s, t \in \mathbb{R}$,

$$|\varphi(s) - \varphi(t)| \leq |s - t|.$$

Applying this to γ_n gives

$$\begin{aligned} |\gamma_n| &= \left| \frac{\varphi(4^n(x + \delta_m)) - \varphi(4^n x)}{\delta_m} \right| \\ &\leq \left| \frac{4^n(x + \delta_m) - 4^n x}{\delta_m} \right| \\ &\leq 4^n. \end{aligned}$$

This holds with equality when $m = n$.

Using all the work up until this point gives:

$$\begin{aligned}
\lim_{\delta_m \rightarrow 0} \left| \frac{f(x + \delta_m) - f(x)}{\delta_m} \right| &= \lim_{m \rightarrow \infty} \left| \frac{f(x + \delta_m) - f(x)}{\delta_m} \right| \\
&= \lim_{m \rightarrow \infty} \left| \frac{\sum \left(\frac{3}{4}\right)^n \varphi(4^n x + \delta_m) - \sum \left(\frac{3}{4}\right)^n \varphi(4^n x)}{\delta_m} \right| \\
&= \lim_{m \rightarrow \infty} \left| \sum_{n=0}^m \left(\frac{3}{4}\right)^n \frac{\varphi(4^n(x + \delta_m)) - \varphi(4^n x)}{\delta_m} \right| \\
&= \lim_{m \rightarrow \infty} \left| \sum_{n=0}^m \left(\frac{3}{4}\right)^n \gamma_n \right| \\
&\geq \lim_{m \rightarrow \infty} \left| \left(\frac{3}{4}\right)^m \gamma_m \right| - \lim_{m \rightarrow \infty} \left| \sum_{n=0}^{m-1} \left(\frac{3}{4}\right)^n \gamma_n \right| \\
&\geq \lim_{m \rightarrow \infty} \left(\left(\frac{3}{4}\right)^m 4^m - \sum_{n=0}^{m-1} \left(\frac{3}{4}\right)^n 4^n \right) \\
&= \lim_{m \rightarrow \infty} \left(3^m - \sum_{n=0}^{m-1} 3^n \right) \\
&= \lim_{m \rightarrow \infty} \left(3^m - \frac{3^m - 1}{3 - 1} \right) \\
&= \lim_{m \rightarrow \infty} \left(\frac{3^m + 1}{2} \right)
\end{aligned}$$

This value approaches infinity, so the limit does not exist, therefore f is not differentiable at x .

7.7 Power Series

Definition 7.9. A *power series (centered about a)* is an infinite series of the form

$$\sum_{n=0}^{\infty} c_n (x - a)^n$$

Definition 7.10. A real function $f : E \rightarrow \mathbb{R}$, where $E \subset \mathbb{R}$, is *(real) analytic at a* if it can be written as the sum of a power series centered about $a \in E$.

Remark 7.4 (Analyticity). Definition 7.10 hints at the fact that there are functions which are not real but are analytic. A complex function $f : \mathbb{C} \rightarrow \mathbb{C}$ can also be analytic, in which case the radius of convergence is a disc of convergence. In the context of complex functions, analyticity is especially important and gives rise to many of magic results of complex analysis.

Example 7.20. Every polynomial defined on the real line, the set of which is denoted by $\mathcal{P}(\mathbb{R})$, is analytic (on \mathbb{R}). Suppose $f \in \mathcal{P}(\mathbb{R})$ has degree m .

$$f(x) = a_0 + a_1 x + \cdots + a_m x^m = \sum_{n=0}^m a_n x^n$$

This function can be written as a power series centered about 0,

$$f(x) = \sum_{n=0}^{\infty} c_n (x - 0)^n$$

where $c_n = a_n$ for all $n = 1, \dots, m$ and $c_n = 0$ otherwise. Furthermore, f can be written as a polynomial centered around *any* real number $a \in \mathbb{R}$. In a certain sense, power series are sort of like polynomials of an infinite degree.

Example 7.21. Suppose a function f has the form

$$f(x) = \frac{cx^b}{1+x^d},$$

for real numbers $c, b, d \in \mathbb{R}$. This function is analytic on $(-1, 1)$, and can be written as a power series using the familiar limit of a geometric series.

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad (|x| < 1)$$

We have

$$f(x) = \frac{cx^b}{1+x^d} = cx^b \left(\frac{1}{1-(-x^d)} \right) = cx^b \sum_{n=0}^{\infty} (-x^d)^n = cx^b \sum_{n=0}^{\infty} (-1)^n x^{dn} = \sum_{n=0}^{\infty} (-c)^n x^{dn+b}.$$

Definition 7.11. Let $\sum c_n(x-a)^n$ be a power series. We define *radius of convergence R* of this series to be the quantity

$$R = \frac{1}{\limsup_{n \rightarrow \infty} |c_n|^{1/n}}$$

where we adopt the convention that $1/0 = \infty$ and $1/\infty = 0$.

As the name suggest, the radius of convergence of a power series dictates the set on which a power series will converge.

Theorem 7.10 (Cauchy-Hadamard). Let $\sum c_n(x-a)^n$ be a power series. The series will converge on the set $\{x \in \mathbb{R} \mid |x-a| < R\}$ and diverge on the set $\{x \in \mathbb{R} \mid |x-a| > R\}$.

Proof. Without loss of generality, assume $a = 0$.¹³⁴ Suppose that $|x-a| = |x| < R$, where R^{-1} is neither 0 nor ∞ , and fix an $r > R$. By the definition of R ,

$$\limsup_{n \rightarrow \infty} |c_n|^{1/n} > \frac{1}{r},$$

which implies that $\limsup |c_n|r^n > 1$. This means there are infinitely many indices n for which $|c_n|r^n > 1$, so c_nx^n does not approach 0 for any x with $|x| = r > R$, and the partial sum of $\sum c_nx^n$ will never converge. This means that $\sum c_nx^n$ diverges for any $r > R$.

Instead suppose $0 < r < R$. This time there $|c_n|r^n > 1$ for all except finitely many indices n , allowing us to take the maximum value of the terms over these indices.

$$M = \max \{|c_n|r^n \mid |c_n|r^n > 1\}$$

For all n and $|x| < r$ we have

$$|c_nx^n| = |c_n||x|^n = |c_n|r^n \left| \frac{x}{r} \right|^n \leq M \left| \frac{x}{r} \right|^n.$$

The series $\sum M|x/r|^n$ converges as $|x/r| < 1$, so the power series converges by the Weierstrass M-Test (Theorem 7.8). \square

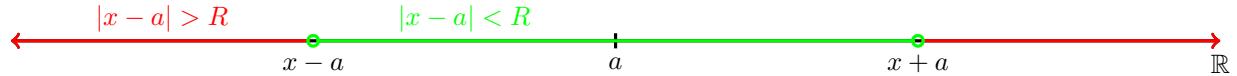


Figure 82: The radius of convergence for a power series. The series converges for all values in green, and diverges for all values in red. It is not clear what happens when $|x - a| = R$.

There's a blaring omission from Theorem 7.10 – what the hell happens when $|x - a| = R$? It's not possible *a priori* to determine whether a power series converges or diverges at this boundary. This can be illustrated with a handful of examples.

Example 7.22 (Power Series Converges at Boundary). The power series $\sum x^n/n^2$ has a radius of convergence of

$$R = \frac{1}{\limsup_{n \rightarrow \infty} |1/n^2|^{1/n}} = \limsup_{n \rightarrow \infty} 1 = 1,$$

but on $\{-1, 1\}$ by the Weierstrass M-Test and the convergence of $\sum 1/n^2$ and $\sum -1/n^2$.

Example 7.23 (Power Series Diverges at Boundary). The power series $\sum x^n$ has a radius of convergence of $R = 1$, but fails to converge on $\{-1, 1\}$.

Example 7.24 (Power Series Converges and Diverges at Boundary). The power series $\sum (-1)^{n+1}(x^n/n)$ has a radius of convergence of $R = 1$. At the point $x = 1$, the power series equals the convergent alternating series $\sum (-1)^n/n$. At the point $x = -1$, the power series becomes the divergent harmonic series $\sum 1/n$.

While a power series' convergence when $|x - a| = R$ is ambiguous, if the series does converge at one of the boundary points, then the convergence “is nice” in the sense that the series is continuous at that boundary point. This result is due to the Norwegian mathematician Abel.

Proposition 7.1 (Abel's Theorem). Let $f(x) = \sum c_n(x-a)^n$ be a power series with a radius of convergence $0 < R < \infty$. If the series converges at $a + R$, then f is continuous at $a + R$, i.e for $x \in (a - R, a + R)$

$$\lim_{x \rightarrow a+R} \sum_{n=0}^{\infty} c_n(x-a)^n = \sum_{n=0}^{\infty} c_n R^n.$$

The analogous result holds for $a - R$.

Proof. We will show the result for a series which converges at $a + R$. The case where the series converges at $a - R$ is nearly identical.¹³⁴

First, define $d_n = c_n R^n$ and $y = (x - a)/R$. Assuming $|x - a| < R$, it is necessarily the case that $|y| < 1$. Abel's Theorem can now be written as

$$\lim_{y \rightarrow 1} \sum_{n=0}^{\infty} d_n y^n = \sum_{n=0}^{\infty} d_n,$$

when $\sum_{n=0}^{\infty} d_n$ converges.¹³⁵ The proof will proceed using this alternate formulation.

¹³⁴We can always just replace x with $y = x - a$ where y is centered around 0.

¹³⁵Alternatively, the second case can be shown via the first and replacing c_n with $(-1)^n c_n$.

¹³⁶ $\sum d_n$ converging is the same as $\sum c_n(x - a)^n$ converging at $R + a$.

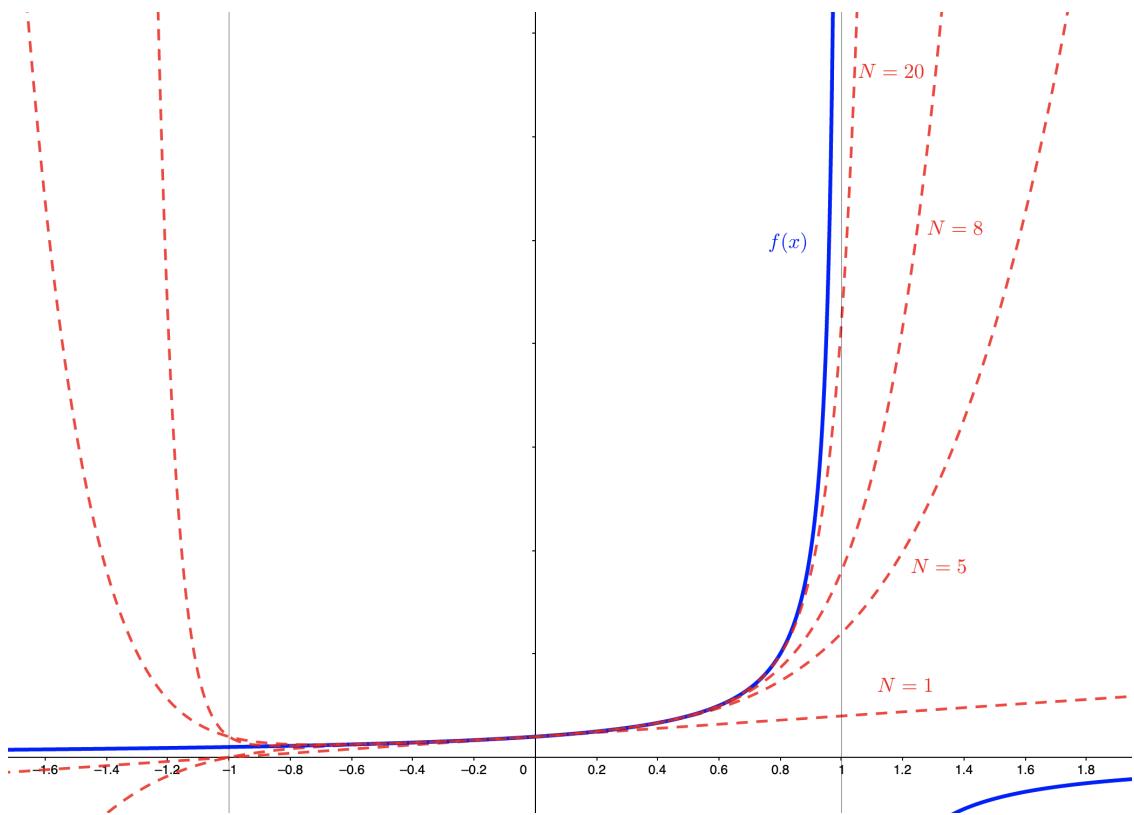


Figure 83: The analytic function $f(x) = 1/(1-x)$ along with the series $\sum_{n=0}^N x^n$ for various values of N . The sum converges to f as $N \rightarrow \infty$ as long as $|x| < 1$.

Let $S_n = \sum_{k=0}^n d_k$, also adopting the convention $S_{-1} = 0$. We have

$$\begin{aligned}
\sum_{n=0}^m d_n y^n &= \sum_{n=0}^m (d_n + 0 + \cdots + 0) y^n \\
&= \sum_{n=0}^m (d_n + (d_1 - d_1) + \cdots + (d_{n-1} - d_{n-1})) y^n \\
&= \sum_{n=0}^m ((d_1 + \cdots + d_{n+1} + d_n) - (d_1 + \cdots + d_{n+1} + d_n)) y^n \\
&= \sum_{n=0}^m \left(\sum_{k=0}^n d_k - \sum_{k=0}^{n-1} d_k \right) y^n \\
&= \sum_{n=0}^m (S_n - S_{n-1}) y^n \\
&= \sum_{n=0}^m S_n y^n - \sum_{n=0}^m S_{n-1} y^n \\
&= (S_0 y^0 + \cdots + S_m y^m) - (\underbrace{S_{-1} y^0 + \cdots + S_{m-1} y^m}_0) \\
&= (S_0 y^0 + S_1 y^1 + \cdots + S_{m-1} y^{m-1}) - (S_0 y^1 + \cdots + S_{m-1} y^m) + S_m y^m \\
&= \sum_{n=0}^{m-1} S_n y^n - \sum_{n=0}^{m-1} S_n y^{n+1} + S_m y^m \\
&= \sum_{n=0}^{m-1} S_n y^n - y \sum_{n=0}^{m-1} S_n y^n + S_m y^m \\
&= (1-y) \sum_{n=0}^{m-1} S_n y^n + S_m y^m.
\end{aligned}$$

recalling that $|y| < 1$, we can take the limit of this series as $m \rightarrow \infty$ to get

$$\begin{aligned}
\lim_{m \rightarrow \infty} \sum_{n=0}^m d_n y^n &= \lim_{m \rightarrow \infty} (1-y) \sum_{n=0}^{m-1} S_n y^n + \underbrace{\lim_{m \rightarrow \infty} S_m y^m}_{|y| < 1 \implies \rightarrow 0} \\
\sum_{n=0}^{\infty} d_n y^n &= (1-y) \sum_{n=0}^{\infty} S_n y^n.
\end{aligned}$$

We defined S_n such that

$$\lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} \sum_{k=0}^n d_k = \sum_{n=0}^{\infty} d_n.$$

By the definition of a limit, for all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that

$$\left| \sum_{n=0}^{\infty} d_n - S_n \right| < \varepsilon \leq \varepsilon \cdot \frac{1-|y|}{2(1-y)}$$

for all $n > N$. Combine this with the fact that

$$(1-y) \sum_{n=0}^{\infty} y^n = (1-y) \frac{1}{1-y} = 1$$

to get

$$\begin{aligned}
\left| \sum_{n=0}^{\infty} d_n y^n - \sum_{n=0}^{\infty} d_n \right| &= \left| (1-y) \sum_{n=0}^{\infty} S_n y^n - \sum_{n=0}^{\infty} d_n \right| \\
&= \left| (1-y) \sum_{n=0}^{\infty} S_n y^n - \sum_{n=0}^{\infty} d_n \right| \\
&= \left| (1-y) \sum_{n=0}^{\infty} S_n y^n - 1 \cdot \left(\sum_{n=0}^{\infty} d_n \right) \right| \\
&= \left| (1-y) \sum_{n=0}^{\infty} S_n y^n - (1-y) \sum_{n=0}^{\infty} y^n \cdot \left(\sum_{n=0}^{\infty} d_n \right) \right| \\
&= \left| (1-y) \sum_{n=0}^{\infty} \left(S_n - \sum_{n=0}^{\infty} d_n \right) y^n \right| \\
&\leq (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + (1-y) \sum_{n=N+1}^{\infty} \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n \\
&\leq (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + (1-y) \sum_{n=N+1}^{\infty} \varepsilon |y|^n \\
&\leq (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + (1-y)\varepsilon \cdot \frac{1-|y|}{2(1-y)} \sum_{n=0}^{\infty} |y|^n \\
&= (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + (1-y)\varepsilon \cdot \frac{1-|y|}{2(1-y)} \frac{1}{1-|y|} \\
&= (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + \frac{\varepsilon}{2}.
\end{aligned}$$

If we take $\delta > 0$ to be

$$\delta = 1 + \frac{\varepsilon}{2} \left(\sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n \right)^{-1},$$

for any $x > 1 - \delta$ (that is within a distance of δ from the boundary point of interest 1), then

$$\begin{aligned}
\left| \sum_{n=0}^{\infty} d_n y^n - \sum_{n=0}^{\infty} d_n \right| &= (1-y) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + \frac{\varepsilon}{2} \\
&< (1-\delta) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + \frac{\varepsilon}{2} \\
&= \left(1 - 1 + \frac{\varepsilon}{2} \left(\sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n \right)^{-1} \right) \sum_{n=0}^N \left| S_n - \sum_{n=0}^{\infty} d_n \right| \cdot |y|^n + \frac{\varepsilon}{2} \\
&= \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon.
\end{aligned}$$

Therefore,

$$\lim_{y \rightarrow 1} \sum_{n=0}^{\infty} d_n y^n = \sum_{n=0}^{\infty} d_n,$$

□

Example 7.25. Take it as given that the function $\log(1 + x)$ is analytic, and

$$\log(1 + x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$$

for $|x| < 1$. For $x = 1$, the series $\sum_{n=0}^{\infty} \frac{(-1)^n}{n+1}$ converges, so by Abel's Theorem, $\log(1 + x)$ is continuous at $x = 1$. Needless to say, this is not the most efficient way of showing that $\log(1 + x)$ is continuous at $x = 1$.

In the previous example, the punchline was a bit obvious. The continuity of $\log(1 + x)$ at $x = 1$ is not news. In fact, we're often more familiar with analytic functions in their “normal” form, instead of when represented as power series. Perhaps we can work backwards to conclude a power series is converges at the boundary of its radius of convergence if its limit is continuous at that point. Unfortunately, this is not the case, which is to say that the converse of Abel's Theorem *does not hold!*

Example 7.26. The function $f(x) = 1/(1 + x^2)$ on $(0, 1)$ can be written as the power series $\sum (-1)^n x^{2n}$ with radius of convergence of $R = 1$. While f is continuous at $x = 1$, the series does not converge at the point $x = 1$.

At some point you probably took a calculus course which introduced you to power series, and discussed how to integrate and differentiate power series. If you want to perform these operations on a power series, you can simply apply them to each term in the series. This was something special that could be done with power series, but not in general. Your calculus teacher may have emphasized this, but forgone any formal explanation. Now that we know all about uniform convergence and have Theorem 7.9, we can formally prove that power series exhibit these nice properties.

Theorem 7.11. Suppose the power series $\sum c_n(x - a)^n$ has a radius of convergence of R , and define $f(x) = \sum c_n(x - a)^n$ on the set of values such that $|x - a| < R$. Then $\sum c_n(x - a)^n$ converges uniformly to f on $[a - R + \varepsilon, a + R - \varepsilon]$ for all $0 < \varepsilon < R$. Furthermore, f is continuous and differentiable in $(a - R, a + R)$ and

$$f'(x) = \sum_{n=1}^{\infty} n c_n (x + a)^{n-1}.$$

Proof. Without loss of generality, take $a = 0$. Suppose $0 < \varepsilon < R$. For $|x| \leq R - \varepsilon$,

$$|c_n x^n| \leq |c_n(R - \varepsilon)^n|,$$

where $\sum_{n=0}^{\infty} c_n(R - \varepsilon)^n$ converges because $R - \varepsilon$ is in the interval of convergence (Theorem 7.10). By the Weierstrass M-Test (Theorem 7.8), the power series $\sum c_n x^n$ converges uniformly on $[-R + \varepsilon, R - \varepsilon]$. Unfortunately, uniform convergence does not preserve differentiation (Example 7.15), so we still need to show the second portion of the claim using Theorem 7.9. That is, we need to show that f converges pointwise for some $x \in (-R, R)$ and f' is uniformly convergent on $(-R, R)$.

We have already shown that f converges uniformly on $[-R + \varepsilon, R - \varepsilon]$ for all $0 < \varepsilon < R$, so f converges pointwise not only for some $x \in (-R, R)$, but for all $x \in (-R, R)$! The series $\sum_{n=1}^{\infty} n c_n (x + a)^{n-1}$ is itself a power series with coefficients $d_n = n c_n$, and we can denote its radius of convergence as R' . This radius can

be written as

$$\begin{aligned}
R' &= \frac{1}{\limsup_{n \rightarrow \infty} |d_n|^{1/n}} \\
&= \frac{1}{\limsup_{n \rightarrow \infty} |nc_n|^{1/n}} \\
&= \frac{1}{\limsup_{n \rightarrow \infty} (n|c_n|)^{1/n}} \\
&= \frac{1}{\limsup_{n \rightarrow \infty} (n)^{1/n} (c_n)^{1/n}} \\
&= \frac{1}{\limsup_{n \rightarrow \infty} n^{1/n} \limsup_{n \rightarrow \infty} c_n^{1/n}} \\
&= \frac{1}{1 \cdot \limsup_{n \rightarrow \infty} c_n^{1/n}} \\
&= R.
\end{aligned}$$

We get the same radius of convergence as the original series, so by the first part of this proof f' converges uniformly on $(-R, R)$. Part 3 of Theorem 7.9 is satisfied, and f is differentiable (and theorem continuous). \square

Theorem 7.11 allows us to integrate analytic functions by integrating each term of a power series.

Corollary 7.3. Under the hypotheses of Theorem 7.10, f is Riemann integrable on $(a - R, a + R)$, and for $[\alpha, \beta] \subset [-R + \varepsilon, R - \varepsilon]$ we have

$$\int_\alpha^\beta f(x) dx = \sum_{n=1}^{\infty} c_n \frac{(\alpha - a)^{n+1} - (\beta - a)^{n+1}}{n + 1}.$$

An important insight from Theorem 7.11 is that the derivative of a power series is itself a power series. This means we can apply Theorem 7.11 to $f'(x)$ to get $f''(x)$, and then apply it to $f^{(3)}(x)$, and then apply it to $f^{(4)}$, etc. This can be done *ad infinitum*, which is stated by the next corollary.

Corollary 7.4. Under the hypotheses of Theorem 7.10, f is infinitely differentiable on $(a - R, a + R)$,¹³⁷ and

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)(n-2) \cdots (n-k) c_n (x-a)^{n-k}.$$

Example 7.27. The function $f(x) = |x|^k$ is only k times differentiable at $x = 0$, so it is not analytic on any interval containing $x = 0$, otherwise it would be infinitely differentiable.

Corollary 7.4 tells us that all (real) analytic functions are infinitely differentiable, but it is not the case that every infinitely differentiable real function is analytic. Before a counterexample can be provided we need to consider a very important type of power series.

¹³⁷We could simply write $f \in C^\infty((-R, R))$

7.8 Taylor Series

Until now, we've focused mostly on the theoretical properties of power series, while only providing two examples of analytic functions where a power series was explicitly written down (Example 7.20 and Example 7.21). Both of these examples were relatively straight forward. It remains unclear how to represent an analytic function $f(x)$ as a power series. How do we even begin to consider which type of infinite series will equal $f(x)!$? The seeds for this problem's solution were planted back in Section 5.7.

Recall that Taylor's Theorem (Theorem 5.6) gives an explicit formula for approximating a function f at a point x_0 using its derivatives.

$$f(x) \approx P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(0)}{k!}(x - x_0)^k$$

In this particular case, we assumed f can be differentiated n times, but what happens if f is analytic? By Corollary 7.4, all analytic functions are infinitely differentiable, so can take Taylor's Theorem and run with it by letting the degree of P_n go to infinity? The answer is yes. If we do this we get a special type of power series which will converge (uniformly) to an analytic function.

Theorem 7.12 (Taylor's Theorem II). Suppose $f(x)$ is a real analytic function, i.e it can be written as

$$f(x) = \sum_{n=0}^{\infty} c_n x^n$$

for all $|x| < R$. Then we have

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x - a)^n,$$

where this series is referred to as the *Taylor series for f about a* . If $a = 0$ we call this series the *Maclaurin series for f* .

Proof. Suppose $f(x)$ is analytic and can be represented by some power series $\sum_{n=0}^{\infty} c_n x^n$ for all $|x| < r$.

$$\begin{aligned}
f(x) &= \sum_{n=0}^{\infty} c_n x^n \\
&= \sum_{n=0}^{\infty} c_n [x + (a - a)]^n \\
&= \sum_{n=0}^{\infty} c_n \underbrace{[(x - a) + a]}_{{\text{apply binomial thm}}}^n \\
&= \sum_{n=0}^{\infty} c_n \sum_{k=0}^n \binom{n}{k} a^{n-k} (x - a)^k \\
&= c_0 \sum_{k=0}^0 \binom{0}{k} a^{0-k} (x - a)^k + c_1 \sum_{k=0}^1 \binom{1}{k} a^{1-k} (x - a)^k + c_2 \sum_{k=0}^2 \binom{2}{k} a^{2-k} (x - a)^k + \dots \\
&= c_0 + \sum_{k=0}^1 \binom{1}{k} c_1 a^{1-k} (x - a)^k + \sum_{k=0}^2 \binom{2}{k} c_2 a^{2-k} (x - a)^k + \dots \\
&= c_0 + \left(\binom{1}{0} c_1 a^{1-0} (x - a)^0 + \binom{1}{1} c_1 a^{1-1} (x - a)^1 \right] + \left(\binom{2}{0} c_2 a^{2-0} (x - a)^0 + \binom{2}{1} c_2 a^{2-1} (x - a)^1 + \binom{2}{2} c_2 a^{2-2} (x - a)^2 \right] + \dots \\
&= \left(\left(\binom{0}{0} c_0 a^{0-0} + \binom{1}{0} c_1 a^{1-0} + \binom{2}{0} c_2 a^{2-0} + \dots \right] (x - a)^0 + \left(\binom{1}{1} c_1 a^{1-1} + \binom{2}{1} c_2 a^{2-1} + \dots \right] (x - a)^1 + \dots \right. \\
&= \left. \left(\sum_{n=0}^{\infty} \binom{n}{0} c_n a^{n-0} \right] (x - a)^0 + \left(\sum_{n=1}^{\infty} \binom{n}{1} c_n a^{n-1} \right] (x - a)^1 + \left(\sum_{n=2}^{\infty} \binom{n}{2} c_n a^{n-2} \right] (x - a)^2 + \dots \right. \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \binom{n}{k} c_n a^{n-k} \right] (x - a)^k \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \binom{n}{k} c_n a^{n-k} \right] (x - a)^k
\end{aligned}$$

Recalling the formula for $f^{(k)}$ given by Corollary 7.4, we have

$$\begin{aligned}
f(x) &\sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \binom{n}{k} c_n a^{n-k} \right] (x - a)^k \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} c_n a^{n-k} \right] (x - a)^k \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \frac{n(n-1)(n-2)\cdots(n-k+1)(n-k)!}{k!(n-k)!} c_n a^{n-k} \right] (x - a)^k \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \frac{n(n-1)(n-2)\cdots(n-k+1)c_n a^{n-k}}{k!} \right] (x - a)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{\left(\sum_{n=k}^{\infty} n(n-1)(n-2)\cdots(n-k+1)c_n a^{n-k} \right]}_{f^{(k)}(a) \text{ by Corollary 7.4}} (x - a)^k \\
&= \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x - a)^k.
\end{aligned}$$

This is our desired equality. \square

With this part of Taylor's Theorem in hand, we can now derive a power series representation of any

analytic function.

Example 7.28. The function $f(x) = e^x$ is analytic, so we can find its Taylor series about $a = 0$.

$$e^x = \sum_{n=0}^{\infty} \frac{\frac{d^n}{dx^n}(e^x)|_{x=0}}{n!} (x-0)^n = \sum_{n=0}^{\infty} \frac{e^0 x^n}{n!} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Finding the Taylor series or functions is straightforward, but is it possible there are other power series representations of an analytic function? Perhaps e^x can be written as some other power series. Are Taylor series not only a power series representation of an analytic function, but also *the* power series representation? Fortunately – yes! This is an immediate corollary of Taylor's Theorem.

Corollary 7.5 (Uniqueness of Power Series). Suppose $f : E \rightarrow \mathbb{R}$, where $E \subset \mathbb{R}$, is analytic at some point $a \in E$. If f has two power series expansions centered about a :

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} c_n (x-a)^n; \\ f(x) &= \sum_{n=0}^{\infty} d_n (x-a)^n; \end{aligned}$$

each with non-zero radii of convergences, then $c_n = d_n$ for all $n \in \mathbb{N}$.

Proof. By Taylor's Theorem,

$$c_n = \frac{f^{(n)}(a)}{n!} = d_n.$$

□

Remark 7.5 (Power Series at a Point). It's often much easier to refer to a “function's power/Taylor series”, but this is technically incorrect. A power series for a function is always defined at a point of the function's domain. It's wholly possible that a function has different power series for different points of its domain.

In light of Theorem 7.12 and Corollary 7.5, showing a function is not analytic at a point is a matter of showing it does not have a valid Taylor expansion at that point. This means we can now show that the converse of Corollary 7.4 does not hold.

Example 7.29 (An Infinitely Differentiable Non-Analytic Function). Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}.$$

This function is infinitely differentiable at the point $x = 0$, the proof of which can be found [here](#). Suppose f is analytic. For all $n \in \mathbb{N}$, $f^{(n)}(0) = 0$, so the Taylor series of the function at $x = 0$ is

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \sum_{n=0}^{\infty} \frac{0}{n!} x^n = 0.$$

The power series converges to zero on the real line, instead of converging uniformly to $f(x)$. This is a contradiction, so f cannot be analytic.

Example 7.29 is particularly important when comparing properties of real functions to properties of complex functions. One of the hallmark properties of complex functions is that analyticity is in fact equivalent to being infinitely differentiable (such complex functions are called *holomorphic*). A more surprising fact is that if any complex function is differentiable, it must be infinitely differentiable. Being differentiable is synonymous with being holomorphic, which is in turn synonymous with being analytic. This discrepancy between real functions and complex functions tells us something very important – differentiability is a *much* stronger condition when dealing with complex functions.

7.9 Arzelà–Ascoli Theorem

One of the hallmark results in real analysis is the Bolzano-Weierstrass theorem (Corollary 3.3) which told us that all bounded sequences in \mathbb{R}^k have a convergent subsequence. This followed from the Heine-Borel theorem and the fact that any bounded sequence in \mathbb{R}^k can easily be interpreted as a bounded sequence in a closed subset of \mathbb{R}^k . Does this result hold for more abstract spaces? Recall from Remark 7.1 that uniform convergence of real functions is just a special case of convergence in a metric space. The metric space just happens to be a set of functions. Could it be possible to generalize the Bolzano-Weierstrass theorem to a space of functions?

Example 7.30. Consider the (closed) unit ball in the set $C([0, 1])$.

$$\bar{B}_1(0) = \{f \in C([0, 1]) \mid \|f\|_\infty \leq 1\}$$

This set is bounded as

$$\begin{aligned} d(f, g) &\leq d(f, 0) + d(0, g) && \text{(triangle inequality)} \\ &\leq \|f\|_\infty + \|g\|_\infty && (d(f, 0) = \|f\|_\infty) \\ &\leq 1 + 1 && (f, g \in \bar{B}_1(0)) \\ &= 2 \end{aligned}$$

for any two $f, g \in \bar{B}_1(0)$. We can also verify this set is closed by showing that $(\bar{B}_1(0))^c$ is open. Let g be an arbitrary element of $(\bar{B}_1(0))^c$, and define $r = \|g\|_\infty - 1$. Now let $h \in B_r(g) \subset C([0, 1])$. We have

$$\begin{aligned} \|h\|_\infty &= \|h + (g - g)\|_\infty \\ &= \|g - (g - h)\|_\infty \\ &\geq \|g\|_\infty - \|g - h\|_\infty \\ &> \|g\|_\infty - r && (h \in B_r(g)) \\ &= \|g\|_\infty - (\|g\|_\infty - 1) && (r = \|g\|_\infty - 1) \\ &= 1 \end{aligned}$$

Because $\|h\|_\infty > 1$, $h \notin \bar{B}_1(0)$, so $h \in (\bar{B}_1(0))^c$ for an arbitrary $h \in B_r(g) \subset C([0, 1])$. This means $B_r(g) \subset (\bar{B}_1(0))^c$ for any $g \in (\bar{B}_1(0))^c$, making $(\bar{B}_1(0))^c$ open.

Now consider the sequence $f_n(x) = x^n$ for $f_n : [0, 1] \rightarrow \mathbb{R}$. We have

$$\|f_n\|_\infty = \sup_{x \in [0, 1]} |x^n| = 1^n = 1 \leq 1$$

for all n , so $\{f_n\}$ is a sequence in the closed and bounded space $C([0, 1])$. From Example 7.7 and Corollary 7.2 we know that this sequence does not converge in the metric space $(C([0, 1]), \|\cdot\|_\infty)$, as convergence of a

sequence in this space amounts to uniform convergence of a sequence of real functions. Perhaps we can find a convergent subsequence though, after all $C([0, 1])$ is closed and bounded, so perhaps it's compact like a closed and bounded set in \mathbb{R} ! Unfortunately, this cannot be the case.

Example 7.7 established that

$$\lim_{n \rightarrow \infty} f_n(x) = f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}.$$

Keep in mind that pointwise convergence is really about sequences in \mathbb{R} , as we're considering the sequence $\{f_n(x)\} \subset \mathbb{R}$ for each x . If we apply Theorem 3.3 to the sequences $\{f_n(x)\}$ for each x , we have that *every* subsequence of $\{f_n(x)\}$ will converge to $f(x)$ for all $x \in [0, 1]$. This means any subsequence of $\{f_n\}$ would have to converge pointwise to the discontinuous f . This rules out there being a uniformly convergent subsequence of functions, because any uniform limit would agree with the pointwise limit. This cannot be the case though, as this would contradict Corollary 7.2. Therefore, $\{f_n\} \subset C([0, 1])$ has no convergent subsequence.

So no – we cannot generalize the Bolzano-Weierstrass theorem beyond \mathbb{R}^k . This particular example also tells us something interesting about compactness. Theorem 3.4 tells us that a necessary condition for compactness is that every sequence has a convergent subsequence, so Example 7.30 provides an example of a closed and bounded set that is not compact. We already saw such a set in Example 2.41. One reason this may be surprising in this particular context is that $C([0, 1])$ is defined using the compact interval $[0, 1]$, so one might conjecture that the Heine-Borel theorem “carries over”, but alas it does not. The next question becomes, what conditions could we impose on $C(X)$ for $X \subseteq \mathbb{R}$ such that $C(X)$ has a convergent subsequence given X is compact? Maybe uniform continuity is the missing puzzle piece here? Again – no. In Example 7.30, every element in $C([0, 1])$ is uniformly continuous, as each element is a continuous function defined on a compact set (Theorem 4.8). What we need is uniform continuity “on steroids” across all elements of a sequence $\{f_n\}$.

Definition 7.12. Given two metric spaces X and Y , define $C(X, Y)$ as the set of all continuous functions $f : X \rightarrow Y$. A subset $\mathcal{F} \subset C(X)$ is *(uniformly) equicontinuous (on X)* if for all $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, p) < \delta$ implies $d_Y(f(x), f(p)) < \varepsilon$ for all $x, p \in X$ and for all $f \in \mathcal{F}$.

The key phrase in the definition of equicontinuity is “for all $f \in \mathcal{F}$ ”. Uniform continuity strengthened continuity by ensuring that δ did not depend on the point $x \in X$. Equicontinuity goes even further. Not only does δ not depend on $x \in X$, but it also doesn't depend on $f \in \mathcal{F}$! The same δ will work for all $x \in X$, and across all the functions in \mathcal{F} . In a sense, an equicontinuous collection of functions \mathcal{F} is uniformly uniformly continuous. In the event that \mathcal{F} is some sequence of functions $\{f_n\}$, equicontinuity means that δ will not be a function of x or n .

Example 7.31. The sequence of functions from Example 7.30 is not equicontinuous. Once again, we know that f_n is uniformly continuous for all n , because each function is a continuous function on a compact set. Unfortunately, it cannot be equicontinuous. The trouble arises at the point $x = 1$. Suppose $\varepsilon \in (0, 1/2)$, and pick some arbitrary $\delta > 0$. Let $p \in (1 - \delta, 1)$ such that $|p - 1| < \delta$. We know that $\lim_{n \rightarrow \infty} p^n = 0$ because $p < 1$, so we can find some N such that

$$|p^n - 0| = |p^n| = p^n < \varepsilon < \frac{1}{2}$$

for all $n > N$. Therefore, when $n > N$, we have

$$\begin{aligned} |f_n(1) - f_n(p)| &= |1 - p^n| \\ &> |1 - 1/2| && (p^n < \varepsilon < 1/2) \\ &= 1/2 \\ &\geq \varepsilon \end{aligned}$$

Regardless of the choice of δ , we've found some f_n (in particular all f_n such that $n > N$ for a specific N), and some $\varepsilon > 0$, such that $|f_n(1) - f_n(p)| > \varepsilon$, violating the definition of equicontinuity.

Example 7.32. Suppose we have a sequence of real functions $\{f_n\}$ defined on $[a, b]$ which satisfy $|f'_n(x)| \leq M$ for all $x \in [a, b]$ and all n . In other words $\|f'_n\|_\infty \leq M$ for all n . An example of such a sequence is $f_n(x) = x + 1/n$, as $f'_n(x) = 1$ for all x and all n . It turns out that any such sequence is equicontinuous. Let x and y be arbitrary points in $[a, b]$. The mean value theorem (Corollary 5.1) tells us for any $c \in (x, y)$ we have

$$\begin{aligned} |f_n(x) - f_n(y)| &= |f'_n(c)| \cdot |x - y|, \\ &\leq M |x - y| && (|f'_n(x)| \leq M \ \forall x \in [a, b]). \end{aligned}$$

This inequality holds for each n . Let $\varepsilon > 0$ and let $\delta = \varepsilon/M$. Suppose $|x - y| < \delta$ for any $x, y \in [a, b]$. We have

$$\begin{aligned} |f_n(x) - f_n(y)| &= |f'_n(c)| \cdot |x - y|, \\ &\leq M |x - y|, \\ &< M \cdot \frac{\varepsilon}{M}, && (|x - y| < \delta) \\ &= \varepsilon. \end{aligned}$$

This holds regardless of our choice of points $x, y \in [a, b]$ and the element of the sequence (corresponding to the choice of n), so $\{f_n\}$ is equicontinuous on $[a, b]$.

We will now prove that the key sufficient condition for a bounded sequence in $C(X)$ to have a convergent subsequence is equicontinuity. This result is known as the Arzelà–Ascoli theorem, which for now will only concern real functions. Later on in Section 13, we'll present a much more general version of the Arzelà–Ascoli theorem, but for now we'll (mostly) stick to the presentation in Rudin (1976).

Theorem 7.13 (Arzelà–Ascoli Theorem I). Let X be a compact subset of \mathbb{R} , and $\{f_n\}$ be a sequence of functions in the metric space $(C(X), \|\cdot\|_\infty)$.

Part II

Higher Dimensions

8 Real Functions of Several Variables

Now we'll take an aside to review linear algebra and multivariable functions. Most of this will look familiar, but some of the notation may differ slightly from what is seen in a linear algebra course. Come Section 17,

some of these definitions will be repeated with different notation. Right now the notation will be more tailored to Euclidean space. Many results related to vector spaces will also be stated as fact or given in examples. The proofs of them can be found in any linear algebra textbook.

8.1 Euclidean Space

Before we can define Euclidean space, we need to introduce the concept of a vector space.

Definition 8.1. A *vector space* is a set of vectors V over a field of scalars F equipped with two operations: $+ : V \times V \rightarrow V$, and $\cdot : F \times V \rightarrow V$. These operations, called addition and scalar multiplication, must satisfy the following properties for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $a, b \in F$:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$,
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$,
3. there is a unique $\mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x}$,
4. $\mathbf{x} + (-1) \cdot \mathbf{x} = \mathbf{0}$ for the identity element $1 \in F$,
5. $1 \cdot \mathbf{x} = \mathbf{x}$ for the identity element $1 \in F$,
6. $a \cdot (b \cdot \mathbf{x}) = (ab) \cdot \mathbf{x}$,
7. $(c + d) \cdot \mathbf{x} = c \cdot \mathbf{x} + d \cdot \mathbf{x}$,
8. $c \cdot (\mathbf{x} + \mathbf{y}) = c \cdot \mathbf{x} + c \cdot \mathbf{y}$.

Many times we drop the “.” when performing scalar multiplication, as it is often clear from the context what operation is being performed.

Certain vector spaces are equipped with a function that measures the “length” of a vector in V . Norms can be seen as a generalization of the absolute value of a number.

Definition 8.2. A *normed vector space* is a vector space V over a field of scalars F equipped with a function $\|\cdot\| : V \rightarrow [0, \infty)$ called a *norm* which for all $\mathbf{x}, \mathbf{y} \in V$ and $c \in F$ satisfies:

1. $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$,
2. $\|c\mathbf{x}\| = |c| \cdot \|\mathbf{x}\|$,
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (Triangle Inequality).

Example 8.1. A trivial normed vector space is the set \mathbb{R} over \mathbb{R} . In this case, a vector is just a real number.

The next two examples of normed vector spaces are fairly abstract. The goal with them is to get a sense of just how general vector spaces can be.

Example 8.2 (Vector Space of Real Sequences). Let \mathbb{R}^∞ be the space of all real sequences $\{x_n\}$, and \mathbb{R} be the set of scalars. An element $x \in \mathbb{R}^\infty$ can be written as $x = \{x_n\}$ for $n = 1, \dots, \mathbb{N}$. If we define

$$\begin{aligned} x + y &= \{x_n\} + \{y_n\} = \{x_n + y_n\} \\ ax &= a\{x_n\} = \{ax_n\} \end{aligned}$$

for $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^\infty$, then we have a vector space. To make this a normed vector space, we can restrict our attention to sequence for which $\|x\|_2 < \infty$ where

$$\|x\|_2 = \|\{x_n\}\|_2 = (x_1^2 + x_2^2 + \dots)^{\frac{1}{2}} = \left(\sum_{n \in \mathbb{N}} x_n^2 \right)^{\frac{1}{2}}.$$

We will call this normed vector space ℓ^2 . For $x_n = 1/n$ in ℓ^2 , then

$$\|\{1/n\}\| = \left(\sum_{n=1}^{\infty} \frac{1}{n^2} \right)^{\frac{1}{2}} = \left(\frac{\pi^2}{6} \right)^{\frac{1}{2}} = \frac{\pi}{\sqrt{6}}.$$

Example 8.3 (Vector Space of Continuous Functions). Let $C([a, b])$ be the set of real continuous functions on $[a, b]$. This set forms a vector space over the set of scalars \mathbb{R} . This is also a normed vector space equipped with the sup norm, which is written as

$$\|f\|_\infty = \sup_{x \in [a, b]} |f(x)|.$$

Because any continuous on a bounded interval is bounded, we have that the elements of $C([a, b])$ are bounded functions, and $C_b([a, b]) = C([a, b])$.

Remark 8.1 (Norm vs. Metric). If you think all the way back to Example 2.2, we mentioned that norms and metrics are related. This is because every normed vector space is a metric space if we define

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

The converse however is not true. A metric space needn't be a vector space, as a metric space isn't necessarily endowed with the algebraic structure of a vector space. In this sense, metric spaces are actually generalizations of normed vector spaces! When we have a normed vector space and want to discuss it in the context of a metric space, we will assume $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. This is why I said that “supremum metric” is not a common term. If you're interested in the set $C([a, b])$ with $d(f, g) = \sup_{x \in [a, b]} |f - g|$, then we simply specify that it has the *sup norm* (pronounced “soup norm”), and infer the corresponding metric.

The most useful consequence of this is that all our definitions related to convergence and continuity in metric spaces carry over to arbitrary normed vector spaces. This allows us to unify quite a bit of the material we've covered. We already saw some of this in Section 7 with uniform convergence. Recall from Remark 7.1 that uniform convergence of real functions is equivalent to convergence in the space $C_b(X)$ where $X \subseteq \mathbb{R}$.

Despite having worked in Euclidean space for the majority of the previous sections, having assumed familiarity, we will formally define it now.

Definition 8.3. We define the normed vector space comprised of vectors $\mathbf{x} \in \mathbb{R}^n$, scalers $c \in \mathbb{R}$, a norm $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$ given by

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_n^2},$$

and an *dot/inner product* $\cdot : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i \cdot y_i,$$

as *Euclidean space*.

Example 8.4 (A Trivial Example). Let \mathbb{R} be the vector space over a field of scalars \mathbb{R} . The norm on this vector space is $|x|$. In this case, the dot product just becomes multiplication of real numbers. This happens to also be how scalar multiplication behaves, as each scalar is also a real number.

Example 8.5 (Column Vectors). We represent an element $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ with a column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Notation 8.1 ($\|\cdot\|$ vs. $\|\cdot\|_2$). In Example 2.2, we talked about the Euclidean norm in relation to the Euclidean metric (that relationship follows from Remark 8.1). We denoted it by $\|\cdot\|_2$. This is the most technical possible notation (for reasons we will see in the next example). That being said, if you are ever working with a normed vector space and it is obvious how the norm would be defined, feel free to use $\|\cdot\|$. If you're working in \mathbb{R}^n , everyone will assume it is with the Euclidean norm. It never hurts to use $\|\cdot\|_2$, but **you can assume this is the norm being used with \mathbb{R}^n from here on out unless otherwise specified.**

Example 8.6 (But Where Does the 2 Come From?). But why do we even denote the Euclidean metric with $\|\cdot\|_2$ in the first place? This comes from a more general norm on \mathbb{R}^n called the p -norm, written as

$$\|\mathbf{x}\|_p = (|x_1|^p + \dots + |x_n|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

for $p \in \mathbb{N}$. We take the absolute value of each component of \mathbf{x} before exponentiating by p to ensure $\|\mathbf{x}\|_p > 0$.¹³⁸ If we let $p = 2$, then we have the Euclidean norm as given in Definition 8.3. If we take $p = 1$, then we have

$$\|\mathbf{x}\|_1 = |x_1| + \dots + |x_n|.$$

This particular choice would give rise to the taxi-cab metric from Example 2.3. If we take our vector space to be \mathbb{R} , then

$$\|x\|_p = (|x|^p)^{\frac{1}{p}} = |x|$$

for any choice of p .

A useful exercise is picking different choices of p for $\|\mathbf{x}\|_p$ and a fixed $\mathbf{x} \in \mathbb{R}^n$. This will give us a better

¹³⁸When we defined a norm, we gave its domain as nonnegative numbers.

sense of how the idea of “length” changes as p changes. Let $\mathbf{x} = (1/2, 1, 2)$.

$$\begin{aligned}\|\mathbf{x}\|_1 &= \frac{1}{2} + 1 + 2 = 3.5 \\ \|\mathbf{x}\|_2 &= \left(\left(\frac{1}{2}\right)^2 + 1^2 + 2^2 \right)^{\frac{1}{2}} = \left(\frac{1}{4} + 1 + 4 \right)^{\frac{1}{2}} = 2.29 \\ \|\mathbf{x}\|_5 &= \left(\left(\frac{1}{2}\right)^5 + 1^5 + 2^5 \right)^{\frac{1}{5}} = \left(\frac{1}{32} + 1 + 32 \right)^{\frac{1}{5}} = 2.01 \\ \|\mathbf{x}\|_{10} &= \left(\left(\frac{1}{2}\right)^{10} + 1^{10} + 2^{10} \right)^{\frac{1}{10}} = \left(\frac{1}{4} + 1 + 1024 \right)^{\frac{1}{10}} = 2.0002 \\ \|\mathbf{x}\|_{15} &= \left(\left(\frac{1}{2}\right)^{15} + 1^{15} + 2^{15} \right)^{\frac{1}{15}} = \left(\frac{1}{2^{15}} + 1 + 2^{15} \right)^{\frac{1}{15}} = 2.000004\end{aligned}$$

What is happening here? When we let $p = 1$, the length of \mathbf{x} is the sum of its components lengths. All the components contribute equally. When we take $p = 2$, instead of summing the lengths, we first square them. Small components like $1/2$ get smaller, and large components like 2 get bigger.¹³⁹ After doing this, we sum them, and then take the square root so the length is not artificially inflated by use squaring each component. By squaring each number, we let the larger components “contribute” more to the length. We do so even more if we take p to be larger. In fact, if we let $p \rightarrow \infty$, the largest component of \mathbf{x} will begin to dominate the others, being the only one that contributes to the norm. We actually end up with

$$\lim_{p \rightarrow \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty = \sup\{|x_1|, \dots, |x_n|\} = \max\{|x_1|, \dots, |x_n|\}.$$

This is just the sup norm, but for a finite set of points!

We can also use the p -norm for sequence in \mathbb{R}^∞ . We write the space of sequences in \mathbb{R}^∞ for which $\|\{x_n\}\| < \infty$ as ℓ^p space. We already saw ℓ^2 in Example 8.1. The only real difference between ℓ^p and \mathbb{R}^n equipped with $\|\cdot\|_p$, is that the vectors in ℓ^p are infinite sequences. In \mathbb{R}^n each vector is finite. This is why for all $\mathbf{x} \in \mathbb{R}^n$ we have $\|\mathbf{x}\|_p < \infty$. We don’t need to eliminate any elements of \mathbb{R}^n to get a normed vector space like we need to for \mathbb{R}^∞ . This difference is so subtle, that you may even see the Euclidean metric on \mathbb{R}^n associated with the notation ℓ^2 , and the taxi-cab metric on \mathbb{R}^n associated with ℓ^1 .

Remark 8.2 (The Missing Link). We now know that as we let $p \rightarrow \infty$ for $\|\cdot\|_p$ we end up with a discrete version of the sup norm that we saw in Subsection 7.1. Is it possible to define $\|f\|_p$ for $f \in C([a, b])$, and conclude that $\lim_{n \rightarrow \infty} \|f\|_p = \|f\|_\infty$. The answer is yes.¹⁴⁰ We will see this for the first time in Section 14, and dedicate an entire section to it with Section 18.

Example 8.7 (Euclidean Norm and Dot Products). The dot product and the norm are related via the following equation:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_n^2} = \sqrt{x_1 x_1 + \dots + x_n x_n} = \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

Proposition 8.1 (Cauchy-Schwarz Inequality). For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x} \cdot \mathbf{y}\| \leq \|\mathbf{x}\| \|\mathbf{y}\|$. Equivalently,

$$\left(\sum_{i=1}^n x_i y_i \right)^2 \leq \left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i^2 \right).$$

¹³⁹“small” means less than 1, “large” means greater than 1.

¹⁴⁰Well, *very* technically no. The limit will turn out to *essentially* by the sup norm.

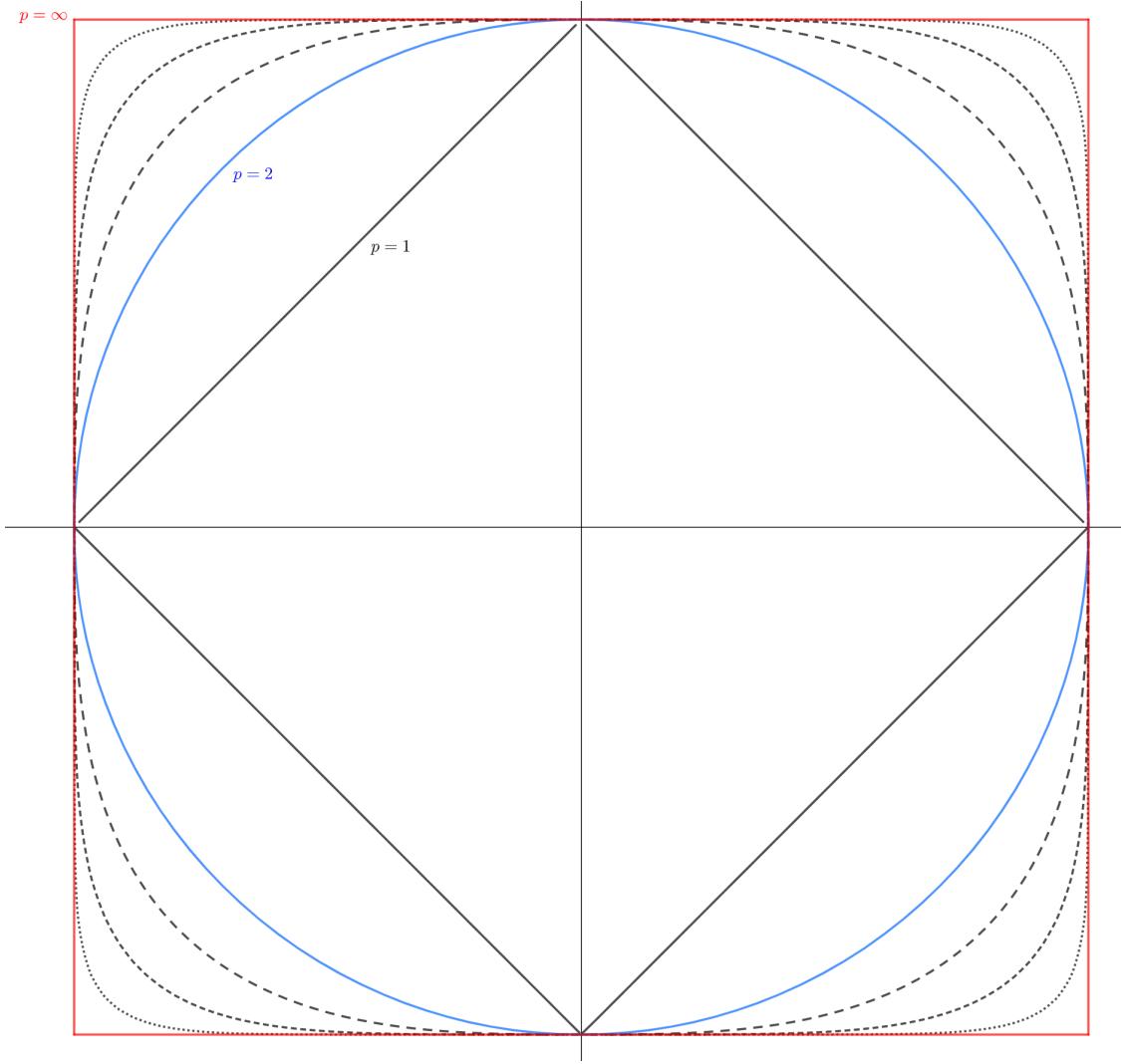


Figure 84:

Proof. Define the function

$$f(t) = \sum_{i=1}^n (x_i t - y_i)^2.$$

The function f cannot be negative, as it is a sum of squares.

$$\begin{aligned} f(t) &\geq 0 \\ \sum_{i=1}^n (x_i t - y_i)^2 &\geq 0 \\ \sum_{i=1}^n (x_i^2 t^2 - 2x_i y_i t + y_i^2) &\geq 0 \\ \left(\sum_{i=1}^n x_i^2 \right) t^2 - 2 \left(\sum_{i=1}^n x_i y_i \right) t + \left(\sum_{i=1}^n y_i^2 \right) &\geq 0 \end{aligned}$$

This quadratic function corresponds to a non-negative paraboloid mapping $\mathbb{R}^n \mapsto \mathbb{R}^+$. It is either a perfect

square with a repeated root when $f(t) = 0$, or it has no real roots, so the discriminant is non-positive.¹⁴¹

$$\begin{aligned} 4 \left(\sum_{i=1}^n x_i y_i \right)^2 - 4 \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) &\leq 0 \\ 4 \|\mathbf{x} \cdot \mathbf{y}\| - 4 \|\mathbf{x}\| \|\mathbf{y}\| &\leq 0 \\ \|\mathbf{x} \cdot \mathbf{y}\| &\leq \|\mathbf{x}\| \|\mathbf{y}\| \end{aligned}$$

The Cauchy-Schwarz inequality holds when $\mathbf{x} = \mathbf{y}$ (Example 8.7). In this case f is a perfect square and has a repeated root, so the discriminant is 0. \square

Now we list some familiar definitions from linear algebra. Familiarity with properties related to these definitions will be assumed.

Definition 8.4. Suppose a set of vectors V forms a vector field over a field of scalars F . For a set of vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in V$ and scalars $c_1, \dots, c_n \in F$, the *linear combination of these vectors and scalars* is

$$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_n \mathbf{x}_n = \sum_{i=1}^n c_i \mathbf{x}_i$$

Definition 8.5. Suppose a set of vectors V forms a vector field over a field of scalars F . The *span* of vectors $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset V$ is the smallest linear subspace which contains S . Alternatively, it is given as

$$\text{Span}(S) = \left\{ \sum_{i=1}^n c_i \mathbf{x}_i \mid n \in \mathbb{N}, \mathbf{x}_i \in S, c_i \in F \right\}.$$

Definition 8.6. Suppose a set of vectors V forms a vector field over a field of scalars F . A set of vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in V$ are *linear independent* if there exists no non-trivial (not all equaling 0) set of scalars $c_1, \dots, c_n \in F$ such that

$$c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \cdots + c_n \mathbf{x}_n = \mathbf{0}$$

Definition 8.7. Suppose a set of vectors V forms a vector field over a field of scalars F . If the set of $\mathbf{x}_1, \dots, \mathbf{x}_n \in V$ spans V and is independent, it is a *basis* for V . The *dimension* of V is number of vectors which form a basis.

Example 8.8. We denote the basis of a Euclidean space \mathbb{R}^n as $\mathbf{e}_1, \dots, \mathbf{e}_n$ where \mathbf{e}_i is a vector comprised of all zeroes, except the i th entry which is 1.

Remark 8.3 (Generalizing Euclidean Space). If you look at the definition of Euclidean space, you should note it has an additional feature that is not required of a normed vector space. We defined a special function which we called the “dot product”. In Section 17, we will study a type of vector spaces called Hilbert Spaces. These are generalizations of Euclidean space, as not only are they equipped with a norm, but they also have a function analogous to the dot product.

8.2 Linear Transformations

Now we review a special type of mapping between vector spaces.

¹⁴¹Recall that for a quadratic equations $ax^2 + bx + c$, the discriminant $b^2 - 4ac$ determines whether the roots of the equation are real, repeated, or complex.

Definition 8.8. A *linear transformation/linear mapping* between vector spaces V and W is a function $T : V \rightarrow W$ which satisfies the following properties for all $\mathbf{x}, \mathbf{y} \in V$ and scalars c :

1. $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$
2. $T(c\mathbf{x}) = cT(\mathbf{x})$

If $V = W$, then T is sometimes referred to as a *linear operator*.

When considering linear transformations in Euclidean space, matrices prove especially useful. For a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$, there exists a unique $m \times n$ matrix such that $T(\mathbf{x}) = A\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$.

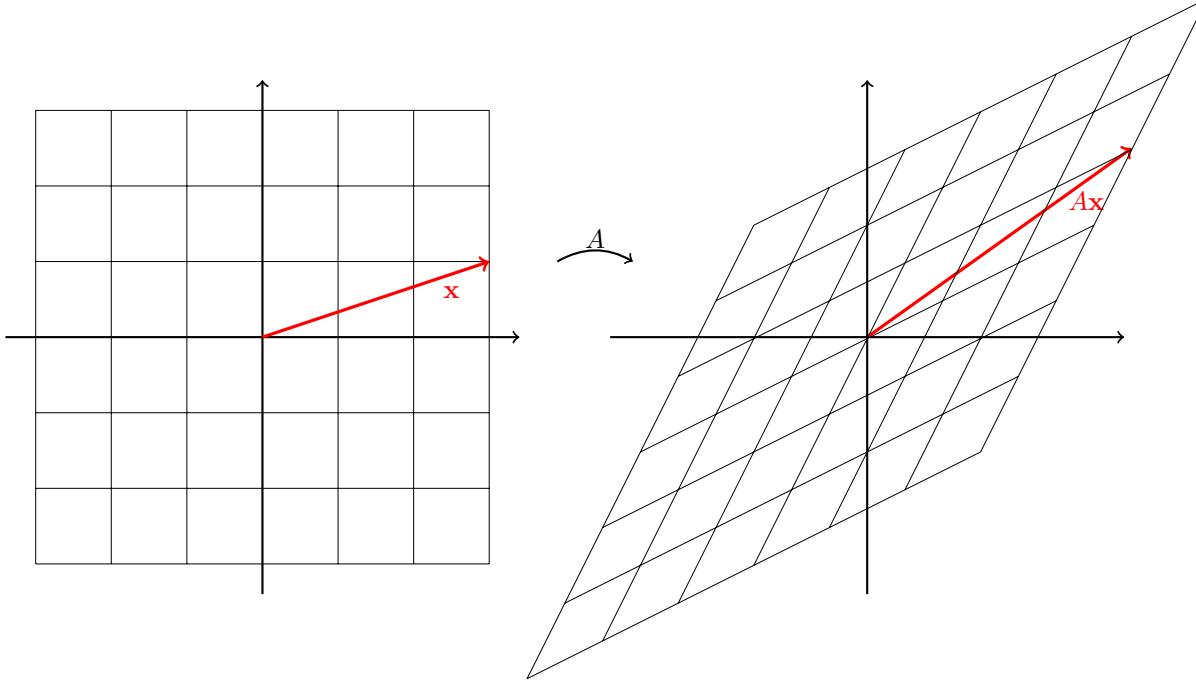


Figure 85: A linear transformation $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given as $T(\mathbf{x}) = A\mathbf{x}$ for a 2×2 matrix A .

Example 8.9. The function $f(x) = ax$ is a linear transformation from \mathbb{R} to \mathbb{R} .

$$\begin{aligned} f(x+z) &= a(x+z) = ax + az = f(x) + f(z) \\ f(cx) &= a(cx) = c(ax) = cf(x) \end{aligned}$$

In this case, the matrix corresponding to $f(x)$ is simply the scalar $a \in \mathbb{R}$. This will come to be an *important fact!* A 1×1 matrix is simply a scalar.

Example 8.10. The rotation operator $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ which “rotates” \mathbb{R}^2 clockwise about the origin by θ degrees is represented by the matrix

$$A = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}.$$

Example 8.11. The projection operator $T : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ which projects $(x, y, z) \in \mathbb{R}^3$ to $(x, y) \in \mathbb{R}^2$ is represented by the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

If a linear transformation T has an inverse T^{-1} , then it is applied via matrix multiplication with A^{-1} . Linear transformations, like any function, is not necessarily invertible. An extensive list of equivalent conditions under which a transformation is invertible is given by the [invertible matrix theorem](#).

Example 8.12. We can revisit the previous three examples.

1. The inverse of the linear transformation $f(x) = ax$ is $f^{-1}(y) = y/a$. In this case $A^{-1} = \begin{bmatrix} 1/a \\ 0 \end{bmatrix}$.
2. To invert this particular rotation operator, we “rotate” \mathbb{R}^2 counter-clockwise about the origin by θ degrees.

$$A^{-1} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

3. The projection transformation from Example 8.11 is not invertible as the number of rows and columns of A equal. Equivalently, the dimension of T ’s domain and codomain differ.

Remark 8.4 (The Geometry of Linear Transformations). Suppose $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear transformation. If we illustrate how T maps $\mathbf{x} \mapsto A\mathbf{x}$ like in Figure 85, it will necessarily be the case that the origin remains fixed and all “lines remain lines”. As such, transformations have at least one of the following geometric effects on \mathbb{R}^n : shearing, scaling/dilating, reflecting, projection (onto \mathbb{R}^m). It may be the case that the geometric effect is captured by a composition of these effects, as the composition of linear transformations is a linear transformation.

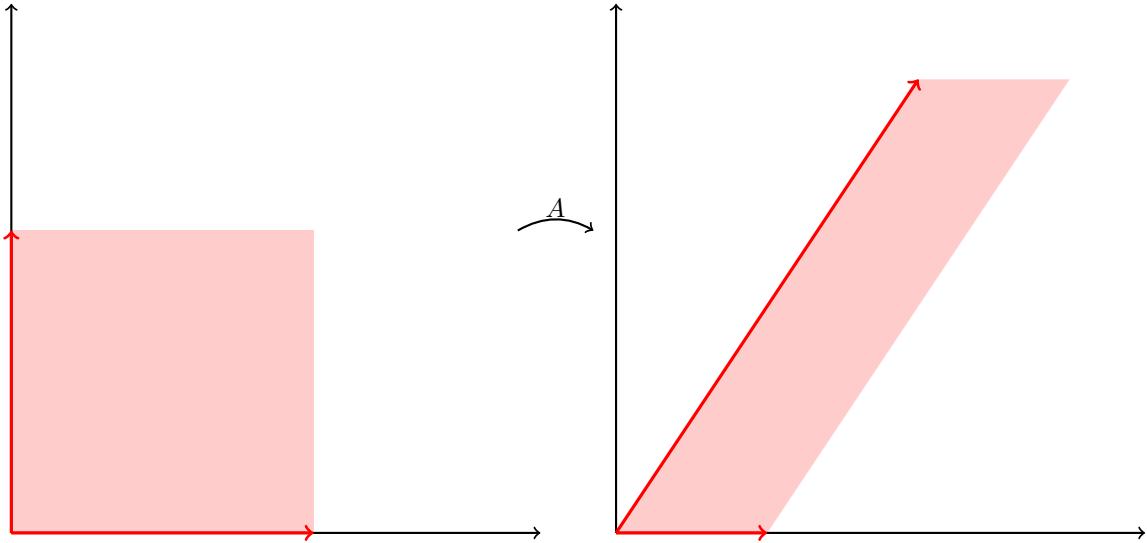


Figure 86:

In the special case where $n = 2$, we can define the set $P(\mathbf{x}, \mathbf{y}) = \{a\mathbf{x} + b\mathbf{y} \mid a, b \in [0, 1]\} \subset \mathbb{R}^2$. This set is a parallelogram with one vertex fixed at the origin. The image of this set under T , $T(P(\mathbf{x}, \mathbf{y}))$, is also a parallelogram because T is linear. A useful way to visualize transformations between \mathbb{R}^2 and \mathbb{R}^2 is to look at the image of the unit square (see Figure 86). When looking at nonlinear transformations between \mathbb{R}^2 and \mathbb{R}^2 the image of the unit square will not be a parallelogram, a fact which proves useful when comparing nonlinear transformations to their linear counterparts.

Finally, we can consider the area of $T(P(\mathbf{x}, \mathbf{y}))$ relative to $P(\mathbf{x}, \mathbf{y})$. The familiar *determinant* of A , $\det A$, is the *signed* factor by which the unit squares area is scaled when transformed by T . The factor by which the unit square is signed is $|\det A|$.

Thus far, all examples of linear transformations have been between Euclidean spaces, but this need not be the case. The next several sections will be focused on Euclidean space, but it is still informative to look at linear transformations between more abstract vector spaces. Some of these examples will become particularly relevant when discussing functional analysis in later sections.

Example 8.13. Let $C([a, b])$ be the set of real continuous functions on $[a, b]$. Define $T : C([a, b]) \rightarrow \mathbb{R}$ as

$$T(f) = \int_a^b f(x) \, dx.$$

For all $f, g \in C([a, b])$ and $c \in \mathbb{R}$,

$$\begin{aligned} T(f + g) &= \int_a^b f(x) + g(x) \, dx = \int_a^b f(x) \, dx + \int_a^b g(x) \, dx = T(f) + T(g) \\ T(cf) &= \int_a^b cf(x) \, dx = c \int_a^b f(x) \, dx = cT(f) \end{aligned}$$

Example 8.14. Let $C^1([0, 1])$ be the set of real differentiable functions on $[0, 1]$. Define $T : C^1([0, 1]) \rightarrow C([0, 1])$ as

$$T(f) = \frac{df}{dx}.$$

For all $f, g \in C^1([0, 1])$ and $c \in \mathbb{R}$,

$$\begin{aligned} T(f + g) &= \frac{d}{dx}(f + g) = \frac{df}{dx} + \frac{dg}{dx} = T(f) + T(g) \\ T(cf) &= \frac{d}{dx}(cf) = c \frac{df}{dx} = cT(f) \end{aligned}$$

8.3 Nonlinear Transformations and Functions

In general, transformations between \mathbb{R}^n and \mathbb{R}^m need not be linear. A function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ takes the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \dots, x_n) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}.$$

The function \mathbf{f} evaluated at some point $\mathbf{x} \in \mathbb{R}^n$ is itself a vector in \mathbb{R}^m , hence the boldface notation.

When $n = m = 1$, \mathbf{f} becomes $f : \mathbb{R} \rightarrow \mathbb{R}$. These were the functions which we restricted our attention to up until this section. The special case where $m = 1$ may also be familiar, and is often the subject of a multivariable calculus course. In this case, $f(\mathbf{x})$ maps a vector $\mathbf{x} \in \mathbb{R}^n$ to a scalar $f(\mathbf{x}) \in \mathbb{R}^m$ allowing us to illustrate f as a “surface” defined on \mathbb{R}^2 . These functions are often called *functions of several variables*. In the event a function of several variables is linear, the resulting surface is a plane which passes through $\mathbf{0}$.

Example 8.15 (Function of Several Variables). Define the function $f(\mathbf{x}) = x_1^3 - 3x_1x_2^2$. For each $\mathbf{x} \in \mathbb{R}^2$, f returns a scalar value. These values form a surface (Figure 89) over the domain of f . Visualizing f this way is not quite the same as the method used in Figure 87. If we were to recreate Figure 87 with $f(\mathbf{x})$, we would show \mathbb{R}^2 being mapped onto the real line \mathbb{R} .

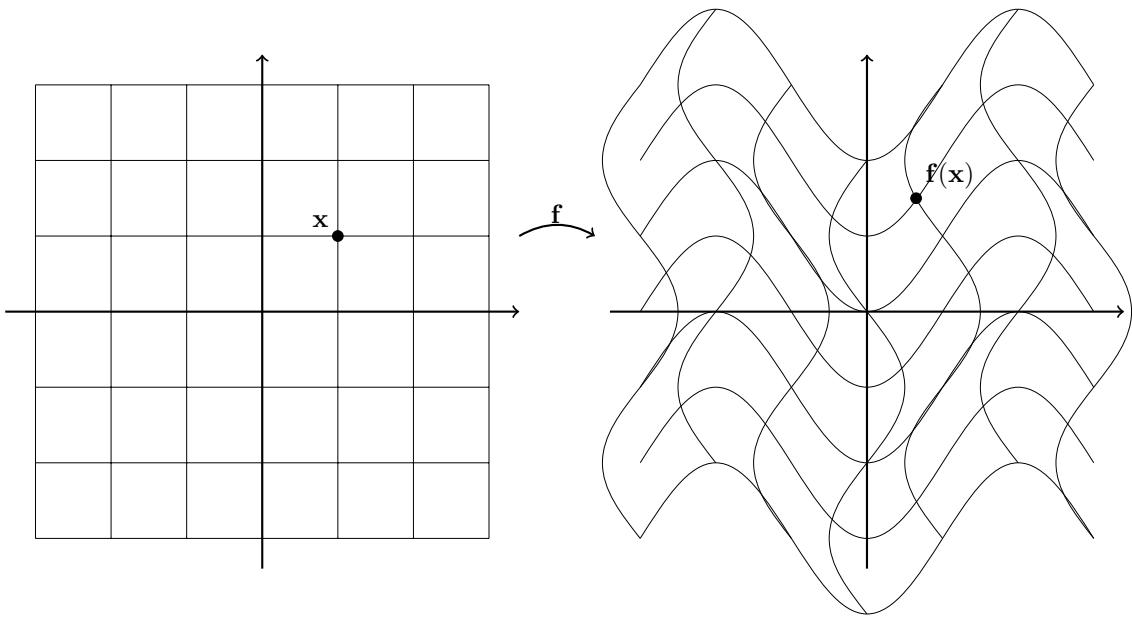


Figure 87: A nonlinear function $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$.

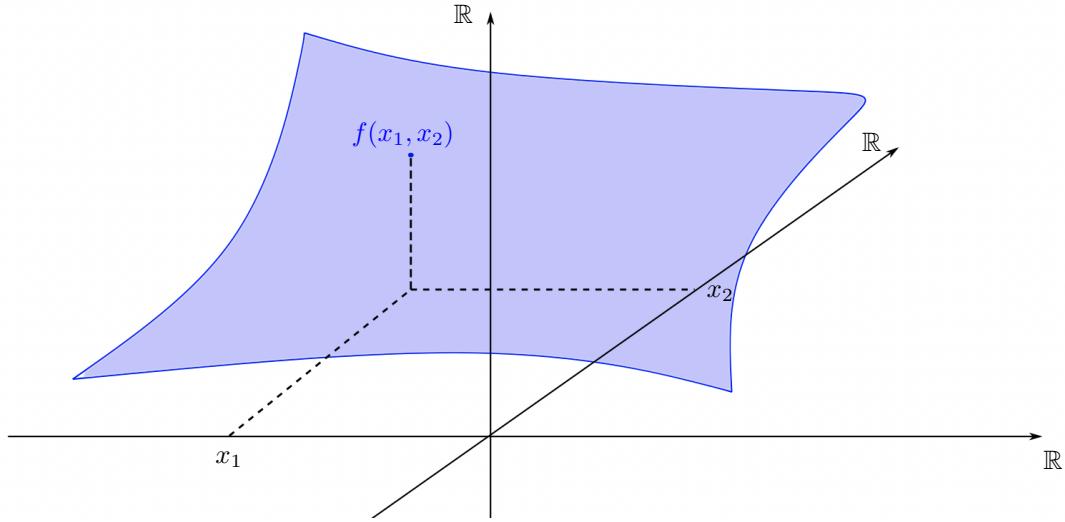


Figure 88: The surface corresponding to $f : \mathbb{R}^2 \rightarrow \mathbb{R}$.

When $n = 1$, we have a *vector valued function* $\mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}^m$. These functions can be visualized as a curve in \mathbb{R}^m which is traced out by the vector $\mathbf{f}(t)$ which varies over “time” $t \in \mathbb{R}$ (Figure 90). If a vector valued function is linear, then $\mathbf{f}(t)$ traces out a line which passes through $\mathbf{0}$ in \mathbb{R}^m .

Example 8.16 (Vector Valued Functions). Define $\mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}^3$ as

$$\mathbf{f}(t) = \begin{bmatrix} \cos t \\ \sin t \\ t \end{bmatrix}.$$

This function traces out a helix in \mathbb{R}^3 .

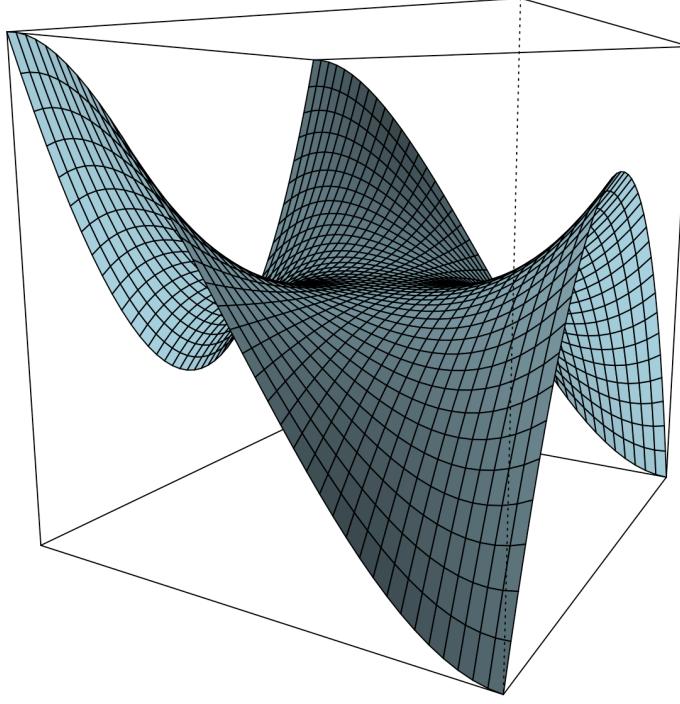


Figure 89: The function $f(\mathbf{x}) = x_1^3 - 3x_1x_2^2$.

Now consider two examples where $n = m = 2$.

Example 8.17. Define $\mathbf{f}(\mathbf{x})$ as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_1^2 - x_2^2 \\ 2x_1x_2 \end{bmatrix}.$$

This transformation is shown in Figure 91. The red region in the domain is the boundary of the unit square ∂S where

$$S = \left\{ a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix} \mid a, b \in [0, 1] \right\}.$$

This boundary can be written as the union of four line segments parameterized by t .

$$\partial S = \underbrace{\{(0, t) \mid t \in [0, 1]\}}_{x_2=0 \text{ on } [0,1]} \cup \underbrace{\{(t, 0) \mid t \in [0, 1]\}}_{x_1=0 \text{ on } [0,1]} \cup \underbrace{\{(1, t) \mid t \in [0, 1]\}}_{x_2=1 \text{ on } [0,1]} \cup \underbrace{\{(t, 1) \mid t \in [0, 1]\}}_{x_1=1 \text{ on } [0,1]}$$

When applying \mathbf{f} to ∂S , we have

$$\begin{aligned} \mathbf{f}(\partial S) &= \{(f_1(0, t), f_2(0, t)) \mid t \in [0, 1]\} \cup \{(f_1(t, 0), f_2(t, 0)) \mid t \in [0, 1]\} \\ &\quad \cup \{(f_1(1, t), f_2(1, t)) \mid t \in [0, 1]\} \cup \{(f_1(t, 1), f_2(t, 1)) \mid t \in [0, 1]\} \\ &= \{(0 - t^2, (2)(0)t) \mid t \in [0, 1]\} \cup \{(t^2 - 0^2, (2)(0)t) \mid t \in [0, 1]\} \\ &\quad \cup \{(1 - t^2, (2)(1)t) \mid t \in [0, 1]\} \cup \{(t^2 - 1^2, (2)(1)t) \mid t \in [0, 1]\} \\ &= \underbrace{\{(-t^2, 0) \mid t \in [0, 1]\}}_{x_2=0 \text{ on } [-1,0]} \cup \underbrace{\{(t^2, 0) \mid t \in [0, 1]\}}_{x_2=0 \text{ on } [0,1]} \cup \underbrace{\{(1 - t^2, 2t) \mid t \in [0, 1]\}}_{x_2=2\sqrt{x_1-1} \text{ on } [0,1]} \cup \underbrace{\{(t^2 - 1, 2t) \mid t \in [0, 1]\}}_{2\sqrt{x_1+1} \text{ on } [-1,0]} \end{aligned}$$

This calculation gives rise to the red boundary post-transformation in Figure 91. Recalling Remark 8.4, it's clear that \mathbf{f} is not linear, as the set bounded by this calculated boundary, $\mathbf{f}(S)$, is not a parallelogram.

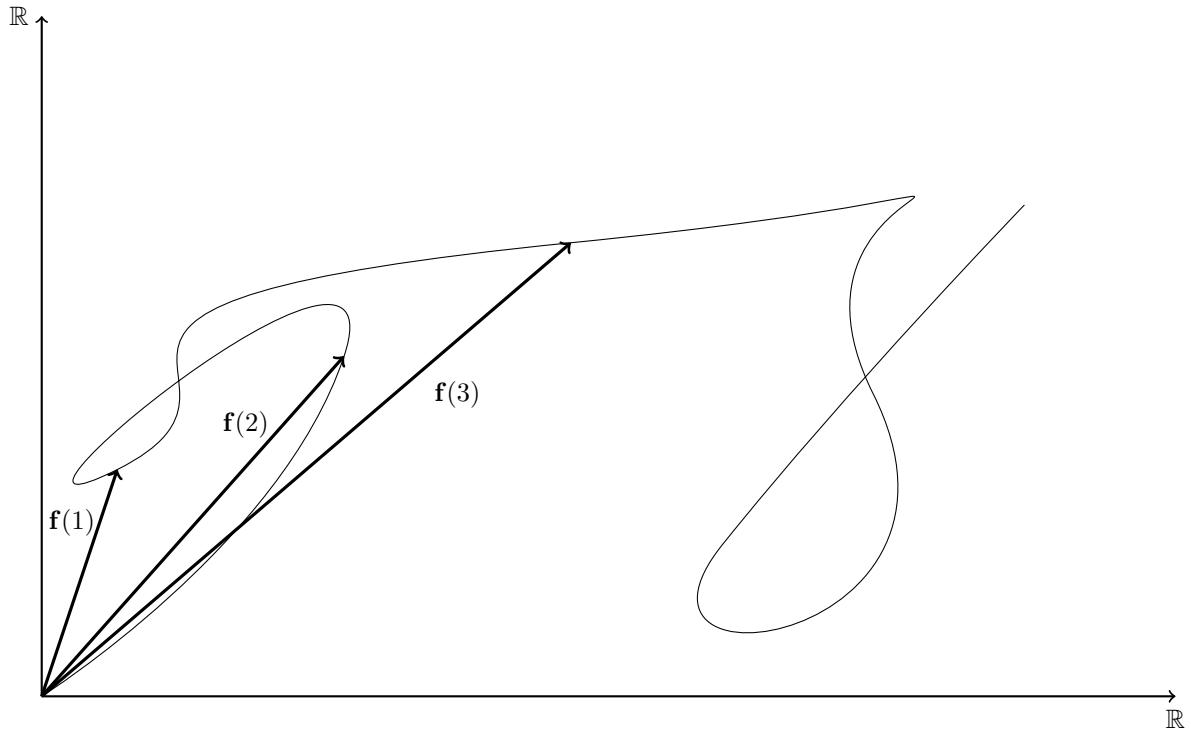


Figure 90: A vector valued function $\mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}^2$.

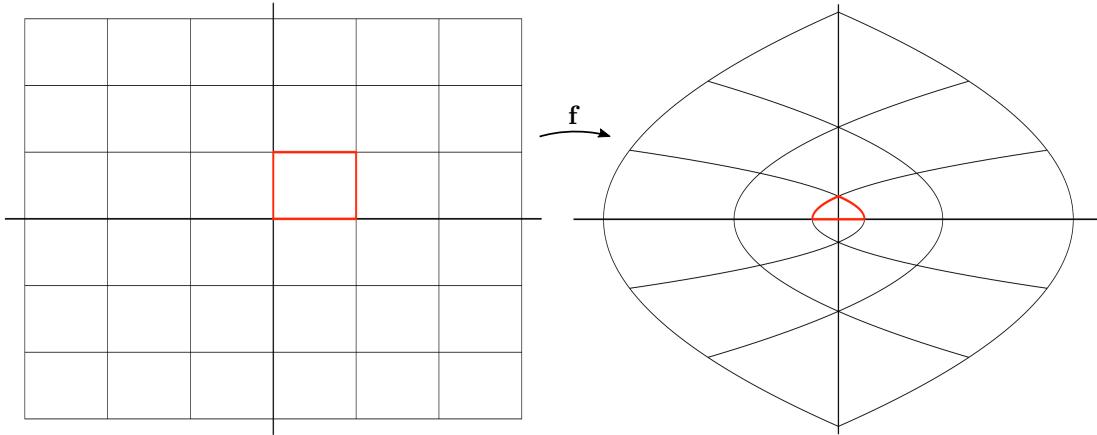


Figure 91: The transformation $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. The boundary of the unit square, ∂S is shown before the transformation is shown on the left. These two images are not drawn to scale relative to each other.

Example 8.18 (Polar Coordinates). Define the function $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as

$$\mathbf{f}(\theta, r) = \begin{bmatrix} r \cos \theta \\ r \sin \theta \end{bmatrix}.$$

For every $(\theta, r) \in \mathbb{R}^2$, $\mathbf{f}(\theta, r)$ is the vector in \mathbb{R}^2 with length r which forms an angle of θ with the positive

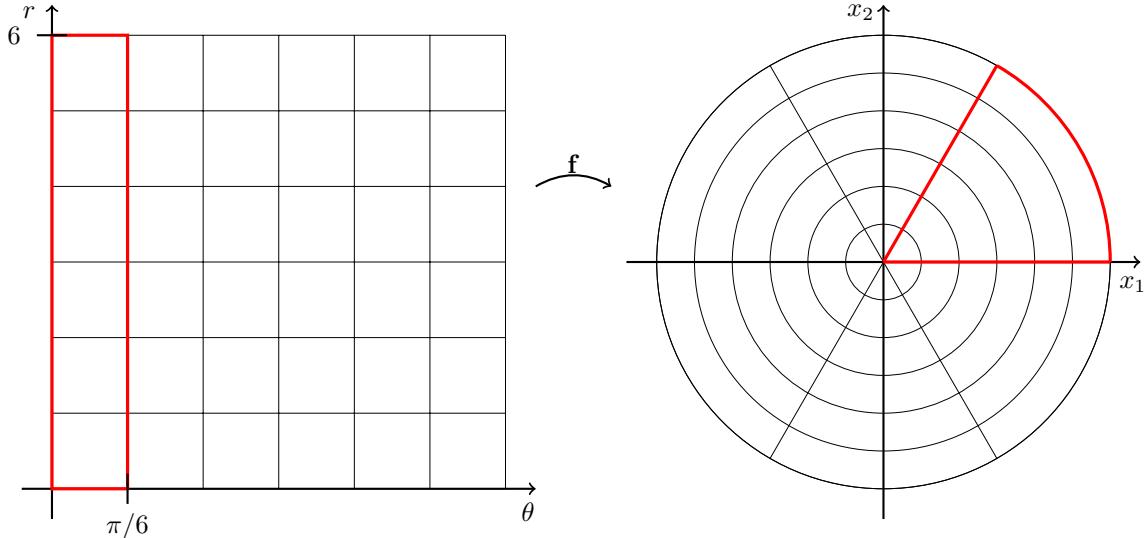


Figure 92: The transformation $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ corresponding to changing to polar coordinates.

x_1 axis. This transformation also has an inverse in the form

$$\mathbf{f}^{-1}(x_1, x_2) = \begin{bmatrix} \sqrt{x_1^2 + x_2^2} \\ \text{atan2}(x_2, x_1) \end{bmatrix}$$

where

$$\text{atan2}(x_2, x_1) = \begin{cases} \arctan(x_2/x_1) & \text{if } x_1 > 0 \\ \arctan(x_2/x_1) + \pi & \text{if } x_1 > 0 \text{ and } x_2 \geq 0 \\ \arctan(x_2/x_1) - \pi & \text{if } x_1 > 0 \text{ and } x_2 < 0 \\ \pi/2 & \text{if } x_1 = 0 \text{ and } x_2 > 0 \\ -\pi/2 & \text{if } x_1 = 0 \text{ and } x_2 < 0 \\ \text{undefined} & \text{if } x_1 = x_2 = 0 \end{cases}$$

Example 8.19 (Spherical Coordinates). Polar coordinates can be generalized to \mathbb{R}^3 . Define $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ as

$$\mathbf{f}(\theta, \varphi, r) = \begin{bmatrix} r \sin \theta \cos \varphi \\ r \sin \theta \sin \varphi \\ r \cos \theta \end{bmatrix}.$$

8.4 Continuity

In Section 4, Definition 4.2 defined continuity for a function between general metric spaces. Despite the generality of this definition, all examples considered were of functions mapping $\mathbb{R} \mapsto \mathbb{R}$. While \mathbb{R} does form a vector space (over the field of scalars \mathbb{R}), it is not a particularly interesting vector space as $\dim(\mathbb{R}) = 1$.

Example 8.20. Consider $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined as $f(\mathbf{x}) = x_1^2 + x_2^2$. This function is continuous at $\mathbf{0} \in \mathbb{R}^2$. For all $\varepsilon > 0$, define $\delta = \sqrt{\varepsilon}$. Whenever

$$d(\mathbf{x}, \mathbf{0}) = \sqrt{(x_1 - 0)^2 + (x_2 - 0)^2} = \sqrt{x_1^2 + x_2^2} < \sqrt{\varepsilon},$$

which can be rewritten as $x_1^2 + x_2^2 < \varepsilon$, we have

$$d(f(\mathbf{x}), f(\mathbf{0})) = x_1^2 + x_2^2 - 0 = x_1^2 + x_2^2 < \varepsilon.$$

Example 8.21. Suppose $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is defined as $\mathbf{f}(\mathbf{x}) = \mathbf{x}$. For all $\varepsilon > 0$, define $\delta = \varepsilon$. Whenever

$$d(\mathbf{x}, \mathbf{0}) < \delta = \varepsilon,$$

we have

$$d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{0})) = d(\mathbf{x}, \mathbf{0}) < \varepsilon,$$

therefore \mathbf{f} is continuous at $\mathbf{0}$.

These two examples are relatively straight forward, as verifying the continuity of a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can become troublesome using the definition of continuity directly. Fortunately, for most $f : \mathbb{R}^n \rightarrow \mathbb{R}$ we can verify continuity by using Corollary 4.3 and Theorem 4.4.

Example 8.22. Define $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}) = x_3 \sin(x_1 x_2).$$

If we define $g(x_3) = x_3$, $h(x_1) = x_1$, and $l(x_2) = x_2$, then we can rewrite f as

$$f(\mathbf{x}) = g(x_3) \sin(h(x_1)l(x_2)).$$

The functions g , h , and l are all single variable continuous functions on the entire set \mathbb{R} , so we the composition and product of them is continuous on all of \mathbb{R} . This means that f is continuous on \mathbb{R}^3 .

In the general case where $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, it turns out that determining continuity is simply a matter of considering each component $f_1(\mathbf{x}), \dots, f_m(\mathbf{x})$ separately.

Proposition 8.2. Define a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}$$

where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i = 1, \dots, m$. The function \mathbf{f} is continuous at $\mathbf{p} \in \mathbb{R}^n$ if and only if $f_j(\mathbf{x})$ is continuous at \mathbf{p} for all $i = 1, \dots, m$.

Proof.

(\Rightarrow) Suppose \mathbf{f} is continuous at \mathbf{p} . For all $\varepsilon > 0$, there exists some $\delta > 0$ such that

$$d_{\mathbb{R}^m}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{p})) = ((f_1(\mathbf{x}) - f_1(\mathbf{p}))^2 + \dots + (f_m(\mathbf{x}) - f_m(\mathbf{p}))^2)^{1/2} < \varepsilon$$

whenever $d_{\mathbb{R}^n}(\mathbf{x}, \mathbf{p}) < \delta$. But

$$\begin{aligned} d_{\mathbb{R}}(f_j(\mathbf{x}), f_j(\mathbf{p})) &= |f_j(\mathbf{x}) - f_j(\mathbf{p})| \\ &= [f_j(\mathbf{x}) - f_j(\mathbf{p})]^2^{1/2} \\ &\leq ((f_1(\mathbf{x}) - f_1(\mathbf{p}))^2 + \dots + (f_m(\mathbf{x}) - f_m(\mathbf{p}))^2)^{1/2} < \varepsilon \end{aligned}$$

will also hold for an arbitrary component f_j whenever $d_{\mathbb{R}^m}(\mathbf{x}, \mathbf{p}) < \delta$, so f_j is continuous at \mathbf{p} for all i .

(\Leftarrow) Suppose f_j is continuous at \mathbf{p} for all i . For all $\varepsilon > 0$ and i , there exists some δ_i such that

$$d_{\mathbb{R}^m}(f_j(\mathbf{x}), f_j(\mathbf{p})) = |f_j(\mathbf{x}) - f_j(\mathbf{p})| < \frac{\varepsilon}{\sqrt{m}}$$

whenever $d_{\mathbb{R}^m}(\mathbf{x}, \mathbf{p}) < \delta_i$. If we take $\delta = \min_i \{\delta_i\}$, then

$$|f_j(\mathbf{x}) - f_j(\mathbf{p})| < \frac{\varepsilon}{\sqrt{m}}$$

will hold for all i simultaneously. Therefore, whenever $d_{\mathbb{R}^m}(\mathbf{x}, \mathbf{p}) < \delta$,

$$\begin{aligned} d_{\mathbb{R}^m}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{p})) &= ((f_1(\mathbf{x}) - f_1(\mathbf{p}))^2 + \dots + (f_m(\mathbf{x}) - f_m(\mathbf{p}))^2)^{1/2} \\ &< ((\varepsilon/\sqrt{m})^2 + \dots + (\varepsilon/\sqrt{m})^2)^{1/2} \\ &= [m(\varepsilon^2/m)]^{1/2} \\ &= \varepsilon \end{aligned}$$

so \mathbf{f} is continuous at \mathbf{p} .

□

8.5 The Space of Bounded Linear Transformations

Until now, we have mostly considered Euclidean space. One interesting exception to this was Example 8.3 where the set of continuous functions mapping \mathbb{R} to \mathbb{R} formed a vector space over the set of scalars \mathbb{R} . We will now look at another vector space where vectors are a special type of function.

Definition 8.9. Let V and W be normed vector spaces. A linear transformation $T : V \rightarrow W$ is *bounded* if for all $v \in V$, there exists some $M \in \mathbb{R}_+$ such that

$$\|T(v)\|_W \leq M \|v\|_V.$$

Notation 8.2. When working across multiple vector spaces, each with their own norm, just writing $\|\cdot\|$ can lead to ambiguity and confusion. To prevent this, it is helpful to write $\|\cdot\|_V$ where V is the vector space in question.

A bounded linear transformation simply maps bounded subsets in one vector space into bounded sets in another vector space (or possibly the same one).

Example 8.23. Let $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the transformation corresponding to the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 8 & 3 \end{bmatrix}.$$

For any $\mathbf{x} \in \mathbb{R}^2$,

$$\|A\mathbf{x}\| = \left\| \begin{bmatrix} 2x_1 + x_2 \\ 8x_1 + 3x_2 \end{bmatrix} \right\| = ((2x_1 + x_2)^2 + (8x_1 + 3x_2)^2)^{1/2} = (68x_1^2 + 52x_1x_2 + 10x_2^2)^{1/2}.$$

If we let

$$M = \frac{(68x_1^2 + 52x_1x_2 + 10x_2^2)^{1/2}}{(x_1^2 + x_2^2)^{1/2}},$$

then

$$\|A\mathbf{x}\| \leq M(x_1^2 + x_2^2)^{1/2} = M \|\mathbf{x}\|.$$

Example 8.24. Recall the linear transformation $T : C^1([0, 1]) \rightarrow C([0, 1])$ from Example 8.14.

$$T(f) = \frac{df}{dx}$$

When endowed with the supremum norm $\|\cdot\|_\infty$ (Example 8.3), this linear transformation is unbounded. It suffices to show that for some $f_n \in C^1([0, 1])$ where $f_n \rightarrow f$, $\|T(f_n)\|_\infty$ “blows up” as $n \rightarrow \infty$ and f_n approaches f . Let $f_n(x) = \sin(2\pi nx)$. We have

$$T(f_n) = \frac{df_n}{dx} = (2\pi n) \cos(2\pi nx),$$

which gives

$$\|T(f_n)\|_\infty = \sup_{x \in [0, 1]} (2\pi n) \cos(2\pi nx) = 2\pi n,$$

so $\|T(f_n)\|_\infty \rightarrow \infty$.

Fortunately, the transformations we care the most about at the moment (those between Euclidean spaces), are guaranteed to be bounded.

Proposition 8.3. Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a linear transformation. Then T is bounded.

Proof. Let A be the $m \times n$ matrix corresponding to T where $A = (a_{ij})$. For all $\mathbf{x} \in \mathbb{R}^n$,

$$T(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{j=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{mj}x_j \end{bmatrix}.$$

The norm of $T(\mathbf{x})$ is

$$\|T(\mathbf{x})\| = \|A\mathbf{x}\| = \left(\sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}x_j \right)^2 \right)^{1/2}. \quad (32)$$

We can apply the Cauchy-Schwarz inequality (Proposition 8.1) to the inner summation, giving

$$\left(\sum_{j=1}^n a_{ij}x_j \right)^2 \leq \left(\sum_{j=1}^n a_{ij}^2 \right) \left(\sum_{j=1}^n x_j^2 \right) = \left(\sum_{j=1}^n a_{ij}^2 \right) \|\mathbf{x}\|^2.$$

When combined with (32), we have

$$\begin{aligned} \|T(\mathbf{x})\|^2 &\leq \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}^2 \right) \|\mathbf{x}\|^2 \\ \implies \|T(\mathbf{x})\| &\leq \underbrace{\left(\sum_{i=1}^m \left(\sum_{j=1}^n a_{ij}^2 \right) \right)^{1/2}}_M \|\mathbf{x}\|. \end{aligned}$$

□

The reason bounded linear transformations are of particular interest is because they themselves form a vector space!

Definition 8.10. Let V and W be vector spaces defined over the field F . Define the set $L(V, W)$ to be the *vector space of all bounded linear transformations* $T : V \rightarrow W$ (over F). In the event that $V = W$, we may write $L(V)$.

The vector space $L(V, W)$ can be a bit confusing at first. We have two separate vector spaces V and W , and by looking at all the linear transformations between these two vector spaces, we get yet another vector space. This can be verified by confirming that $L(V, W)$ satisfies the conditions outlined in Definition 8.1. Let $T, T' \in L(V, W)$ and $c \in F$. Define addition and scalar multiplication as

$$(T + T')(\mathbf{v}) = T(\mathbf{v}) + T'(\mathbf{v})$$

$$(c \cdot T)(\mathbf{v}) = c(T(\mathbf{v}))$$

where $\mathbf{v} \in V$ and $T(\mathbf{v}) \in W$. To confirm these operations are well defined, we must verify $L(V, W)$ is closed under addition and scalar multiplication. For $T, T' \in L(V, W)$, $\mathbf{x}, \mathbf{y} \in V$, and $c \in F$,

$$\begin{aligned} (T + T')(\mathbf{x} + \mathbf{y}) &= T(\mathbf{x} + \mathbf{y}) + T'(\mathbf{x} + \mathbf{y}) \\ &= T(\mathbf{x}) + T(\mathbf{y}) + T'(\mathbf{x}) + T'(\mathbf{y}) \\ &= [T(\mathbf{x}) + T'(\mathbf{x})] + [T(\mathbf{y}) + T'(\mathbf{y})] \\ &= (T + T')(\mathbf{x}) + (T + T')(\mathbf{y}) \\ (T + T')(c\mathbf{x}) &= T(c\mathbf{x}) + T'(c\mathbf{x}) \\ &= cT(\mathbf{x}) + cT'(\mathbf{x}) \\ &= c[T(\mathbf{x}) + T'(\mathbf{x})] = c(T + T')(\mathbf{x}) \end{aligned}$$

so $T + T' \in L(V, W)$. Next we verify that $c \cdot T \in L(V, W)$. For $\mathbf{x}, \mathbf{y} \in V$ and $s \in F$,

$$\begin{aligned} (cT)(\mathbf{x} + \mathbf{y}) &= c(T(\mathbf{x} + \mathbf{y})) \\ &= c(T(\mathbf{x}) + T(\mathbf{y})) \\ &= cT(\mathbf{x}) + cT(\mathbf{y}) \\ &= (cT)(\mathbf{x}) + (cT)(\mathbf{y}) \\ (cT)(s\mathbf{x}) &= c(T(s\mathbf{x})) \\ &= c(sT(\mathbf{x})) \\ &= s(cT(\mathbf{x})) \\ &= s(cT)(\mathbf{x}) \end{aligned}$$

Now we show that $L(V, W)$ satisfies the properties of a vector space. Recall that for $\mathbf{x} \in V$, $T(\mathbf{x}) \in W$, so it behaves like any other vector in W . For all $T, T', T'' \in L(V, W)$, $\mathbf{x} \in V$, and $a, b \in F$:

1. $(T + T')(\mathbf{x}) = T(\mathbf{x}) + T'(\mathbf{x}) = T'(\mathbf{x}) + T(\mathbf{x}) = (T' + T)(\mathbf{x})$ for all $\mathbf{x} \in V$
2. $T(\mathbf{x}) + (T' + T'')(\mathbf{x}) = T(\mathbf{x}) + T'(\mathbf{x}) + T''(\mathbf{x}) = (T + T')(\mathbf{x}) + T''(\mathbf{x})$ for all $\mathbf{x} \in V$
3. Define $T_0 \in L(V, W)$ to be $T_0(\mathbf{x}) = \mathbf{0}$, where $\mathbf{0} \in W$. This transformation is unique, as $\mathbf{0} \in W$ is unique. We have

$$T_0(\mathbf{x}) + T(\mathbf{x}) = \mathbf{0} + T(\mathbf{x}) = T(\mathbf{x})$$

for all $\mathbf{x} \in V$.

4. For $1 \in F$,

$$(T + (-1)T)(\mathbf{x}) = T(\mathbf{x}) + (-1)T(\mathbf{x}) = \mathbf{0}$$

for all $\mathbf{x} \in V$.

5. For $1 \in F$,

$$(1 \cdot T)(\mathbf{x}) = 1 \cdot T(\mathbf{x}) = T(\mathbf{x})$$

for all $\mathbf{x} \in V$.

6. $(a \cdot (b \cdot T))(\mathbf{x}) = a \cdot ((b \cdot T)(\mathbf{x})) = a \cdot (b \cdot (T)(\mathbf{x})) = ab \cdot T(\mathbf{x}) = (ab \cdot T)(\mathbf{x})$ for all $\mathbf{x} \in V$.

7. $((c+d)T)(\mathbf{x}) = (c+d)T(\mathbf{x}) = cT(\mathbf{x}) + dT(\mathbf{x}) = (c \cdot T)(\mathbf{x}) + (d \cdot T)(\mathbf{x}) = (cT + dT)(\mathbf{x})$ for all $\mathbf{x} \in V$.

8. $(c \cdot (T + T'))(\mathbf{x}) = c(T + T')(\mathbf{x}) = c(T(\mathbf{x}) + T'(\mathbf{x})) = cT(\mathbf{x}) + cT'(\mathbf{x}) = (cT + cT')(\mathbf{x})$ for all $\mathbf{x} \in V$.

We showed each property is satisfied for an arbitrary $\mathbf{x} \in V$, so they hold for all vectors in V .

For the remainder of this section we'll restrict our attention to $L(\mathbb{R}^n, \mathbb{R}^m)$, as this is the only space of bounded linear transformations we will work with when generalizing differentiation. In Section 16 we will return to the general case of $L(V, W)$.

Remark 8.5 (Dimension of $L(\mathbb{R}^n, \mathbb{R}^m)$). The vector space $L(\mathbb{R}^n, \mathbb{R}^m)$ has dimension $n \times m$. Each element of $L(\mathbb{R}^n, \mathbb{R}^m)$ can be written as a unique $m \times n$ matrix,¹⁴² so we can consider $L(\mathbb{R}^n, \mathbb{R}^m)$ in terms of matrices. For any $T \in L(\mathbb{R}^n, \mathbb{R}^m)$, we have some $A = (a_{ij})$ which can be written as

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = a_{11} \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} + \cdots + a_{1n} \begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} + \cdots + a_{m1} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \end{bmatrix} + \cdots + a_{mn} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

The transformations corresponding to this collection of matrices with all zero entries except for the (i, j) th element form the basis for $L(\mathbb{R}^n, \mathbb{R}^m)$. There are $m \times n$ matrices, so the dimension of $L(\mathbb{R}^n, \mathbb{R}^m)$ is $\dim(\mathbb{R}^n) \times \dim(\mathbb{R}^m)$.

After defining vector spaces, the notion of a normed vector space was introduced. While $L(\mathbb{R}^n, \mathbb{R}^m)$ is a vector space, can we equip it with a valid norm and get a normed vector space? Norms quantify a vector's distance from the origin in a vector space, but what would this even mean when the vectors in question are linear transformations in $L(\mathbb{R}^n, \mathbb{R}^m)$? In the case of $L(\mathbb{R}^n, \mathbb{R}^m)$, the origin is the transformation $T_0(\mathbf{x}) = \mathbf{0}$. How would we measure the distance between some $T(\mathbf{x}) = A\mathbf{x}$ and T_0 ? One idea would be to measure the distance between $A\mathbf{x}$ and $\mathbf{0}$ in \mathbb{R}^m , which is just the Euclidean norm of $A\mathbf{x}$.

$$\|A\mathbf{x} - \mathbf{0}\|_{\mathbb{R}^m} = \|A\mathbf{x}\|_{\mathbb{R}^m} = \left(\left(\sum_{j=1}^n a_{1j}x_j \right)^2 + \cdots + \left(\sum_{j=1}^n a_{mj}x_j \right)^2 \right)^{1/2}$$

This approach seems tractable, but we still have to address one question – what vector $\mathbf{x} \in \mathbb{R}^n$ do we evaluate $\|A\mathbf{x}\|_{\mathbb{R}^m}$ at? We need some standard \mathbf{x} to “test” the magnitude of $T(\mathbf{x}) = A\mathbf{x}$. One great choice is some $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x}\|_{\mathbb{R}^n} = 1$, as this \mathbf{x} has some standard unitary distance from $\mathbf{0} \in \mathbb{R}^n$. Unfortunately, there are *many* $\mathbf{x} \in \mathbb{R}^n$ for which $\|\mathbf{x}\| = 1$, so we would need to calculate $\|A\mathbf{x}\|_{\mathbb{R}^m}$ for all of these $\mathbf{x} \in \mathbb{R}^n$. At this point, we have the set

$$\{\|A\mathbf{x}\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} = 1\}.$$

¹⁴²To be pedantic, $L(\mathbb{R}^n, \mathbb{R}^m)$ is isomorphic to the set of $m \times n$ real matrices.

To get a definitive magnitude of $T(\mathbf{x})$, we will just take the maximum value of $\|A\mathbf{x}\|_{\mathbb{R}^m}$ in this set, in effect being as liberal as possible when calculating $\|T(\mathbf{x})\|$. This norm is well established and given in the following definition.

Definition 8.11. The *operator norm* on $L(\mathbb{R}^n, \mathbb{R}^m)$, $\|\cdot\|_{\text{op}} : L(\mathbb{R}^n, \mathbb{R}^m) \rightarrow [0, \infty)$, is given as

$$\|T\|_{\text{op}} = \|A\|_{\text{op}} = \sup \{\|T(\mathbf{x})\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} = 1\} = \sup \{\|A\mathbf{x}\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} = 1\}$$

for $T \in L(\mathbb{R}^n, \mathbb{R}^m)$ with a matrix representation of A .

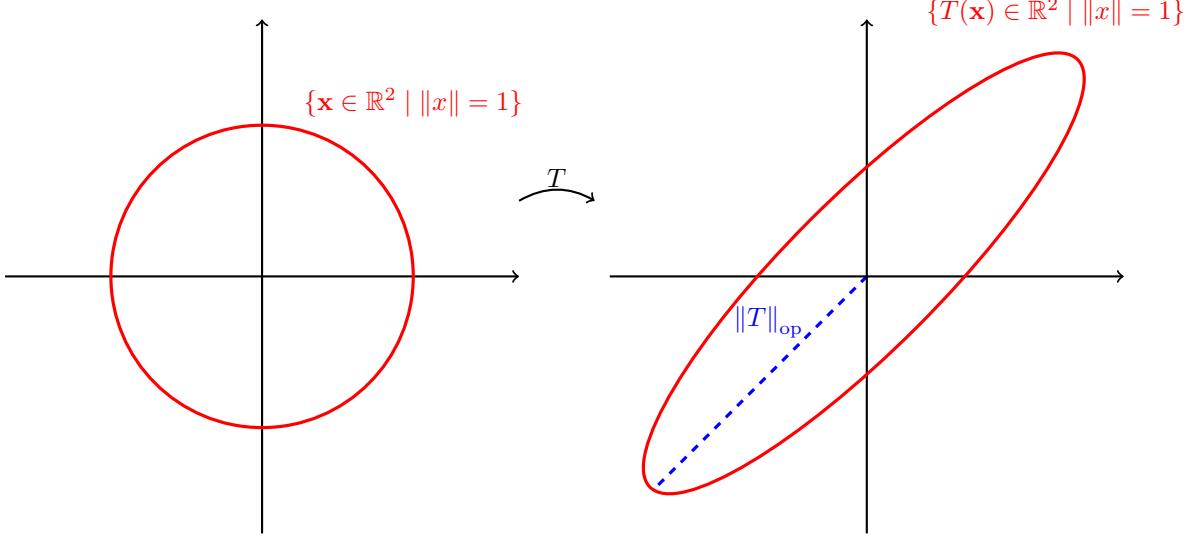


Figure 93: The operator norm visualized for a $T \in L(\mathbb{R}^n, \mathbb{R}^m)$.

Remark 8.6 (Operator Norm, Equivalent Definitions). Definition 8.11 gave one formulation for $\|T\|_{\text{op}}$, but there are a handful of alternate formulations that can be used.

$$\begin{aligned} \|T\|_{\text{op}} &= \sup \{\|T(\mathbf{x})\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} = 1\} \\ &= \sup \{\|T(\mathbf{x})\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} \leq 1\} \\ &= \sup \{\|T(\mathbf{x})\|_{\mathbb{R}^m} \mid \|\mathbf{x}\|_{\mathbb{R}^n} < 1\} \\ &= \sup \left\{ \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} \\ &= \inf \{c \geq 0 \mid \|T(\mathbf{x})\|_{\mathbb{R}^m} \leq c \|\mathbf{x}\|_{\mathbb{R}^n}\} \end{aligned}$$

Just to be sure that $\|T\|_{\text{op}}$ is a valid norm, we should confirm it satisfies the properties of a norm as delineated in Definition 8.2. The properties will directly follow from those of the Euclidean norm $\|T(\mathbf{x})\|_{\mathbb{R}^m}$.

1. $\|T\|_{\text{op}}$ will be 0 if and only if $T(\mathbf{x}) = \mathbf{0}$, as this is equivalent to $\|T(\mathbf{x})\|_{\mathbb{R}^m} = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, which is in turn equivalent to

$$\|T\|_{\text{op}} = \sup \left\{ \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} = \sup \{0\} = 0.$$

2. Let $c \in \mathbb{R}$.

$$\|cT\|_{\text{op}} = \sup \left\{ \frac{\|cT(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} = \sup \left\{ |c| \cdot \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} = |c| \sup \left\{ \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} = |c| \cdot \|T\|_{\text{op}}$$

3. Let $T, T' \in L(\mathbb{R}^n, \mathbb{R}^m)$.

$$\begin{aligned}
\|T + T'\|_{\text{op}} &= \sup \left\{ \frac{\|(T + T')(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} \\
&= \sup \left\{ \frac{\|T(\mathbf{x}) + T'(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} \\
&\leq \sup \left\{ \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m} + \|T'(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} \\
&= \sup \left\{ \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} + \sup \left\{ \frac{\|T'(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} \\
&= \|T\|_{\text{op}} + \|T'\|_{\text{op}}
\end{aligned}$$

The operator norm also enables us to define a metric on $L(\mathbb{R}^n, \mathbb{R}^m)$ as $d(T, T') = \|T - T'\|_{\text{op}}$.

Example 8.25. Consider the linear transformation $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(\mathbf{x}) = A\mathbf{x}$ where

$$A = \begin{bmatrix} 1 & 4 \\ 5 & 6 \end{bmatrix}.$$

We have

$$\|T(\mathbf{x})\| = \|A\mathbf{x}\| = \left\| \begin{bmatrix} x_1 + 4x_2 \\ 5x_1 + 6x_2 \end{bmatrix} \right\| = [(x_1 + 4x_2)^2 + (5x_1 + 6x_2)^2]^{1/2} = (26x_1^2 + 68x_1x_2 + 52x_2^2)^{1/2}.$$

The maximum of this subject to $\|\mathbf{x}\| = 1$ is $[39 + 5(53)^{1/2}]^{1/2}$,¹⁴³ so

$$\|T\|_{\text{op}} = [39 + 5(53)^{1/2}]^{1/2}.$$

The normed vector space of $L(\mathbb{R}^n, \mathbb{R}^m)$ has a handful of important properties that we will use when considering differentiation in more general settings.

Lemma 8.1. If $T \in L(\mathbb{R}^n, \mathbb{R}^m)$, then

$$\|T(\mathbf{x})\|_{\mathbb{R}^m} \leq \|T\|_{\text{op}} \cdot \|\mathbf{x}\|_{\mathbb{R}^n}$$

Proof. By an alternate definition of the operator norm and the definition of the supremum,

$$\|T\|_{\text{op}} \geq \frac{\|T(\mathbf{x})\|_{\mathbb{R}^m}}{\|\mathbf{x}\|_{\mathbb{R}^n}}.$$

Multiplying both sides of this inequality by $\|\mathbf{x}\|_{\mathbb{R}^n}$ gives

$$\|T(\mathbf{x})\|_{\mathbb{R}^m} \leq \|T\|_{\text{op}} \cdot \|\mathbf{x}\|_{\mathbb{R}^n}.$$

□

Proposition 8.4. If $T \in L(\mathbb{R}^n, \mathbb{R}^m)$, then $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is uniformly continuous on \mathbb{R}^n .

¹⁴³This can be calculated via Lagrange multipliers. Alternatively the operator norm of $T \in L(\mathbb{R}^n, \mathbb{R}^m)$ coincides with the largest eigenvalue of $A^T A$. The details of this are not particular important.

Proof. Let \mathbf{p} be an arbitrary element of \mathbb{R}^n . For $\varepsilon > 0$, let $\delta = \varepsilon / \|T\|_{\text{op}}$. Whenever

$$d_{\mathbb{R}^n}(\mathbf{x}, \mathbf{p}) = \|\mathbf{x} - \mathbf{p}\|_{\mathbb{R}^n} < \delta = \varepsilon / \|T\|_{\text{op}},$$

we have

$$\begin{aligned} d_{\mathbb{R}^m}(T(\mathbf{x}), T(\mathbf{p})) &= \|T(\mathbf{x}) - T(\mathbf{p})\|_{\mathbb{R}^m} \\ &= \|T(\mathbf{x} - \mathbf{p})\|_{\mathbb{R}^m} && (T \text{ is linear}) \\ &\leq \|T\|_{\text{op}} \cdot \|\mathbf{x} - \mathbf{p}\|_{\mathbb{R}^n} && (\text{Lemma 8.1}) \\ &< \|T\|_{\text{op}} \cdot \frac{\varepsilon}{\|T\|_{\text{op}}} \\ &= \varepsilon. \end{aligned}$$

This not only shows that T is continuous on \mathbb{R}^n , but also that it is uniformly continuous, as δ does not depend on \mathbf{p} . \square

Lemma 8.2. Suppose $T \in L(\mathbb{R}^n, \mathbb{R}^m)$ and $T' \in L(\mathbb{R}^m, \mathbb{R}^k)$ are represented by matrices A and B , respectively. Then

$$\|T'(T)\|_{\text{op}} = \|T' \circ T\|_{\text{op}} \leq \|T'\|_{\text{op}} \cdot \|T\|_{\text{op}}$$

Proof. First recall from linear algebra that the composition of linear functions is linear, and that $T(T') \in L(\mathbb{R}^n, \mathbb{R}^k)$. For all $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} \|(BA)\mathbf{x}\|_{\mathbb{R}^k} &= \|B(A\mathbf{x})\|_{\mathbb{R}^k} \\ &\leq \|B\|_{\text{op}} \|A\mathbf{x}\|_{\mathbb{R}^k} && (\text{Lemma 8.1}) \\ &\leq \|B\|_{\text{op}} \|A\|_{\text{op}} \|\mathbf{x}\|_{\mathbb{R}^n} && (\text{Lemma 8.1}) \end{aligned}$$

Dividing this inequality by $\|\mathbf{x}\|_{\mathbb{R}^n}$ gives

$$\begin{aligned} \frac{\|(BA)\mathbf{x}\|_{\mathbb{R}^k}}{\|\mathbf{x}\|_{\mathbb{R}^n}} &\leq \|B\|_{\text{op}} \|A\|_{\text{op}} \quad (\forall \mathbf{x} \in \mathbb{R}^n) \\ \implies \sup \left\{ \frac{\|(BA)\mathbf{x}\|_{\mathbb{R}^k}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \right\} &\leq \|B\|_{\text{op}} \|A\|_{\text{op}} \\ \implies \|BA\|_{\text{op}} &\leq \|B\|_{\text{op}} \|A\|_{\text{op}} \\ \implies \|T'(T)\|_{\text{op}} &\leq \|T'\|_{\text{op}} \|T\|_{\text{op}} \end{aligned}$$

\square

Our final lemma that will be of use in the next section concern linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ which are invertible. For the sake of motivation, suppose $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is defined as $T(\mathbf{x}) = A\mathbf{x}$ where

$$A = \begin{bmatrix} 2 & 1 \\ 8 & 3 \end{bmatrix}.$$

The transformation T is invertible, as $\det(A) = -2$. Now suppose $T' \in L(\mathbb{R}^2)$ is given as $T'(\mathbf{x}) = A'\mathbf{x}$ where

$$A' = \begin{bmatrix} 2 + \varepsilon & 1 + \varepsilon \\ 8 + \varepsilon & 3 + \varepsilon \end{bmatrix}$$

for some arbitrarily small ε . This second transformation will be invertible as long as $\det(A') = -2 - 4\varepsilon \neq 0$, which is the case when $|\varepsilon| < 1/2$. This should not be a surprise considering T' is relatively “close” to T , and small changes to the entries of A will have small net changes to $\det(A)$, thereby maintaining a nonzero determinant (and invertibility). But how close are T and T' ? This is quantified by the metric $d_{L(\mathbb{R}^2)}(T, T) = \|T - T'\|_{\text{op}}$.

$$d_{L(\mathbb{R}^2)}(T, T) = \|A - A'\|_{\text{op}} = \left\| \begin{bmatrix} \varepsilon & \varepsilon \\ \varepsilon & \varepsilon \end{bmatrix} \right\|_{\text{op}} = \sup \left\{ \frac{\left\| \begin{bmatrix} \varepsilon x_1 + \varepsilon x_2 \\ \varepsilon x_1 + \varepsilon x_2 \end{bmatrix} \right\|_{\mathbb{R}^2}}{\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|_{\mathbb{R}^2}} \right\} = \sup \left\{ \frac{\sqrt{2}(\varepsilon x_1 + \varepsilon x_2)}{\sqrt{x_1^2 + x_2^2}} \right\}$$

This supremum turns out to be 2ε . It’s reasonable to conclude that any transformation within a distance of 2ε of T is invertible. In other words, any transformation in the open ball $B_{2\varepsilon}(T) \subset L(\mathbb{R}^2)$ is invertible. This raises the question, just how large can the radius of this open ball be? The bound on this radius is given in Lemma 8.3!

Lemma 8.3. Suppose $T, T' \in L(\mathbb{R}^n)$ where T is invertible. If

$$\|T - T'\|_{\text{op}} < \|T^{-1}\|_{\text{op}}^{-1},$$

then T' is invertible.

Proof. For all $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} \|T^{-1}\|_{\text{op}}^{-1} \|\mathbf{x}\|_{\mathbb{R}^n} &= \|T^{-1}\|_{\text{op}}^{-1} \|T^{-1}(T(\mathbf{x}))\|_{\mathbb{R}^n} && \text{(Definition of Inverse)} \\ &= \|T^{-1}\|_{\text{op}}^{-1} \|T^{-1}\|_{\text{op}} \|T(\mathbf{x})\|_{\mathbb{R}^n} && \text{(Lemma 8.1)} \\ &= \|T(\mathbf{x})\|_{\mathbb{R}^n} \\ &= \|T(\mathbf{x}) - T'(\mathbf{x}) + T'(\mathbf{x})\|_{\mathbb{R}^n} && (-T'(\mathbf{x}) + T'(\mathbf{x}) = \mathbf{0}) \\ &= \|(T - T')(\mathbf{x}) + T'(\mathbf{x})\|_{\mathbb{R}^n} && \text{(definition of } + \text{ on } L(\mathbb{R}^n)) \\ &\leq \|(T - T')(\mathbf{x})\|_{\mathbb{R}^n} + \|T'(\mathbf{x})\|_{\mathbb{R}^n} && \text{(triangle inequality)} \\ &\leq \|T - T'\|_{\text{op}} \|\mathbf{x}\|_{\mathbb{R}^n} + \|T'(\mathbf{x})\|_{\mathbb{R}^n} && \text{(Lemma 8.1)} \end{aligned}$$

In particular,

$$\begin{aligned} \|T^{-1}\|_{\text{op}}^{-1} \|\mathbf{x}\|_{\mathbb{R}^n} &\leq \|T - T'\|_{\text{op}} \|\mathbf{x}\|_{\mathbb{R}^n} + \|T'(\mathbf{x})\|_{\mathbb{R}^n} \\ \implies \|T^{-1}\|_{\text{op}}^{-1} \|\mathbf{x}\|_{\mathbb{R}^n} - \|T - T'\|_{\text{op}} \|\mathbf{x}\|_{\mathbb{R}^n} &\leq \|T'(\mathbf{x})\|_{\mathbb{R}^n} \\ \implies \underbrace{\left(\|T^{-1}\|_{\text{op}}^{-1} - \|T - T'\|_{\text{op}} \right)}_{>0} \|\mathbf{x}\|_{\mathbb{R}^n} &\leq \|T'(\mathbf{x})\|_{\mathbb{R}^n} \end{aligned}$$

where the value in parentheses is greater than zero because $\|T - T'\|_{\text{op}} < \|T^{-1}\|_{\text{op}}^{-1}$. This final inequality implies that for any $\mathbf{x} \neq \mathbf{0}$, $\|T'(\mathbf{x})\|_{\mathbb{R}^n} > 0$, meaning $T'(\mathbf{x}) \neq \mathbf{0}$. In other words, $T'(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. This is one of the sufficient conditions for T' to be invertible which is given by [invertible matrix theorem](#). \square

Example 8.26. Any transformation $T \in L(\mathbb{R})$ takes the form $T(x) = ax$ for $a \in \mathbb{R}$. A transformation in this space is invertible if and only if $a \neq 0$. We can use this fact to confirm Lemma 8.3 holds. Suppose

$T(x) = 10x$. According to Lemma 8.3, if

$$\|T - T'\|_{\text{op}} < \|T^{-1}\|_{\text{op}}^{-1},$$

then $T'(x) = a'x$ is invertible. For T and T' we have

$$\begin{aligned} \|T - T'\|_{\text{op}} &= \|a' - 10\| \\ &= \sup \{ \|(a' - x)(\mathbf{x})\|_{\mathbb{R}^1} \mid \|\mathbf{x}\|_{\mathbb{R}^1} = 1 \} \\ &= \sup \{ |(a' - 10)x| \mid |x| = 1 \} \\ &= |a' - 10| \\ \|T^{-1}\|_{\text{op}}^{-1} &= \|10^{-1}\|_{\text{op}}^{-1} \\ &= \|1/10\|_{\text{op}}^{-1} \\ &= (1/10)^{-1} \\ &= |10|. \end{aligned}$$

In our case Lemma 8.3 tells us that T' is invertible if

$$|a' - 10| < 10,$$

a fact which was already implied by T' being invertible if and only if $a' \neq 0$. This example also illuminates the fact that the converse of Lemma 8.3 **does not hold**. It's entirely possible (and likely) that there are invertible transformations outside the open ball $B_{\|T^{-1}\|_{\text{op}}^{-1}}(T)$. In the case of $L(\mathbb{R})$, $T'(x) = a'x$ is invertible for all $a' \in \mathbb{R} \setminus \{0\}$, but Lemma 8.3 only identifies $B_{10}(10) \subset \mathbb{R} \setminus \{0\}$.

Remark 8.7 (Transformation vs. Matrix). Until now, I've been careful to distinguish a linear transformation $T(\mathbf{x}) = A\mathbf{x}$ from the matrix A . Sometimes, you'll see people use them interchangeably. It's very common to simply write $T\mathbf{x}$ instead of $T(\mathbf{x})$ just to make things reminiscent of matrix multiplication, or write $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ even though A is a matrix (this is what Rudin (1976) does). Going forward we will adopt the notation of Rudin (1976) by writing $A \in L(\mathbb{R}^n, \mathbb{R}^m)$, eliminating the need to specify $T(\mathbf{x}) = A\mathbf{x}$ where $T \in L(\mathbb{R}^n, \mathbb{R}^m)$.

9 Differentiation with Several Variables

We can now revisit differentiation in the context of functions $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. While this will require us to use the linear algebra discussed in the previous section, it will also give deeper insights into the nature of the derivative as defined on \mathbb{R} .

9.1 The Derivative as a Linear Map

Recall in Section 5.7 where the the derivative of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ was recast in the light of linear approximation with Theorem 5.5. Remark 5.6 defined the derivative of f at x , $f'(x)$, as the value such that

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x) - f'(x) \cdot h}{h} = 0.$$

In other words, $f(x + h) \approx f(x) + f'(x) \cdot h$ (Figure 94), or

$$f(x + h) - f(x) \approx f'(x) \cdot h.$$

This approximation is a linear transformation in terms of h , as $f'(x)$ is a scalar in \mathbb{R} (see Example 8.9). Considering the derivative from this perspective helps us generalize differentiation to functions from \mathbb{R}^n to \mathbb{R}^m .

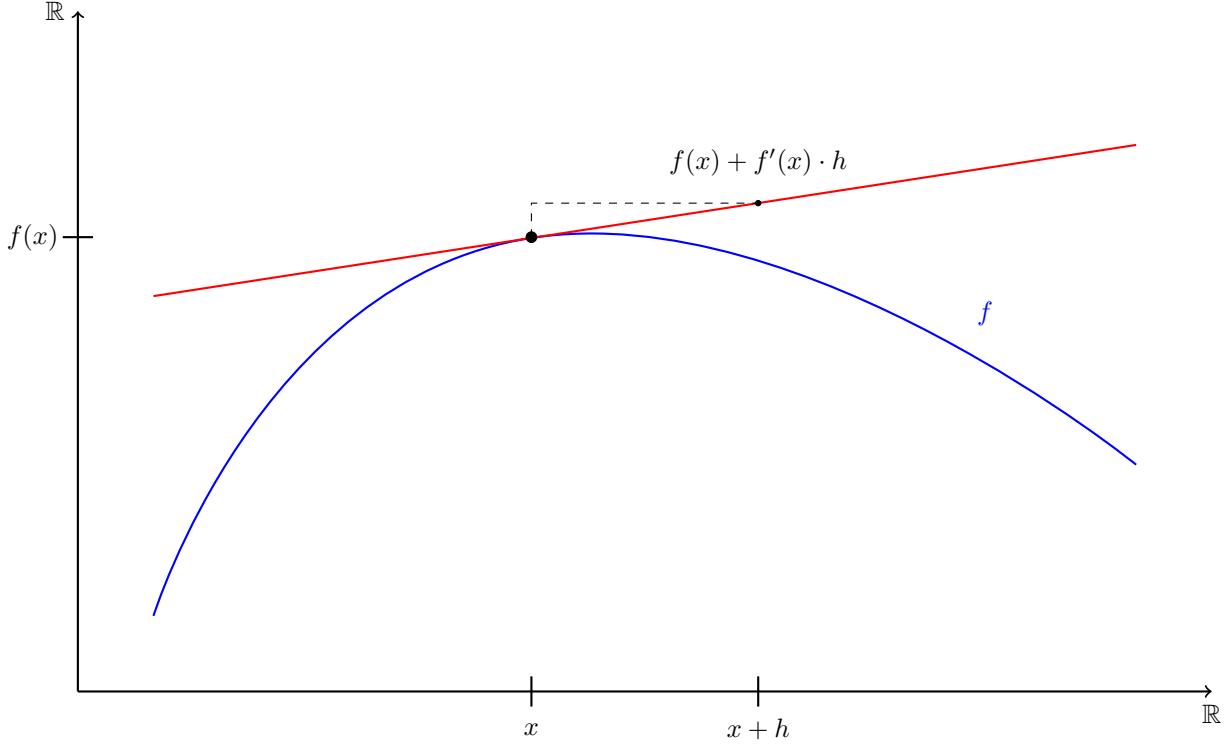


Figure 94: We can approximate $f(x + h)$ by adding $f'(x) \cdot h$ to $f(x)$.

Definition 9.1. Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If there exists a linear transformation $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0, \quad (33)$$

then we say \mathbf{f} is *differentiable at \mathbf{x}* , and refer to $\mathbf{f}'(\mathbf{x}) = A$ as the *(total) derivative of \mathbf{f} at \mathbf{x}* .

The derivative $\mathbf{f}'(\mathbf{x})$ gives rise to the best linear approximation of \mathbf{f} at the point \mathbf{x} , precisely in the same way that $f'(x)$ is the best linear approximation of $f : \mathbb{R} \rightarrow \mathbb{R}$ at x . This fact is illustrated in Figure 95.

Example 9.1. Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$. In this case we have

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x) - A \cdot h}{h} = 0,$$

where $A = f'(x) \in \mathbb{R}$ is a scalar (which is a 1×1 matrix).

Example 9.2 ($f : \mathbb{R}^n \rightarrow \mathbb{R}$). Suppose $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$. In this case (33) becomes

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_{\mathbb{R}}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0,$$

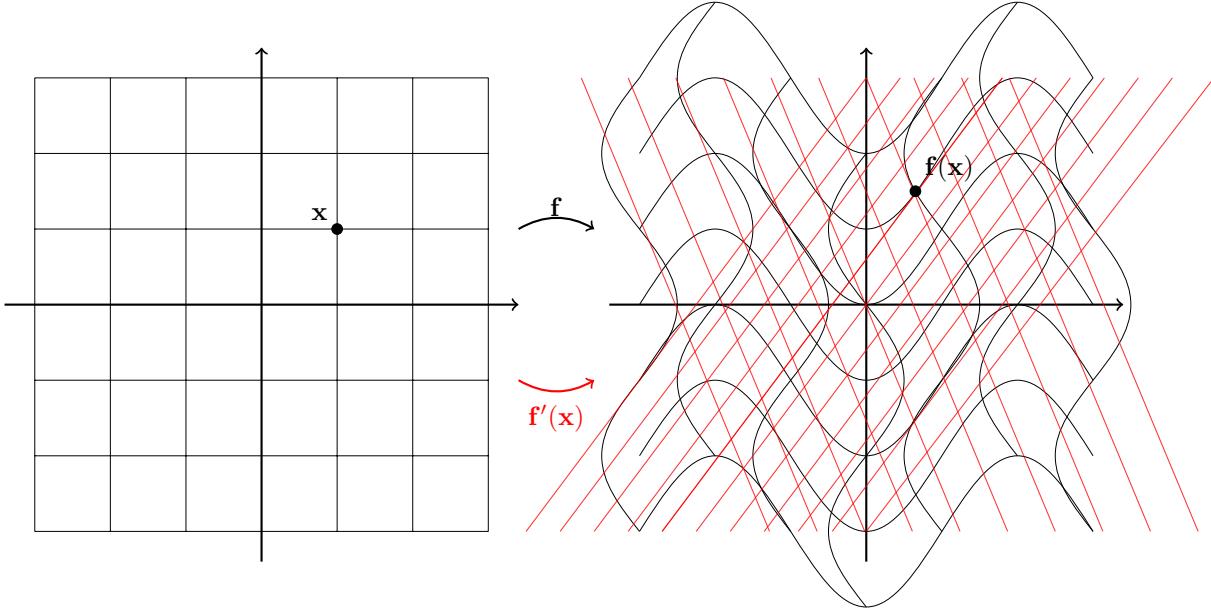


Figure 95: The linear transformation corresponding to the matrix $f'(\mathbf{x})$ is the best linear approximation of f near the point \mathbf{x} .

where A is a $1 \times n$ matrix. The transformation $A = f'(\mathbf{x}) \in L(\mathbb{R}^n, \mathbb{R})$ gives rise to the approximation

$$f(\mathbf{x} + \mathbf{h}) \approx f(\mathbf{x}) + f'(\mathbf{x})\mathbf{h}$$

where $f(\mathbf{x}) + f'(\mathbf{x})\mathbf{h}$ is a plane tangent to f at the point \mathbf{x} .

Example 9.3 (Derivative of a Linear Transformation). Suppose $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear transformation defined as $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$, where A is an $m \times n$ matrix. In this case what is $f'(\mathbf{x})$? Well, \mathbf{f} is already a linear transformation, so it stands to reason that the linear transformation which approximates \mathbf{f} best at \mathbf{x} is A itself. We can confirm this with Definition 9.1:

$$\begin{aligned} \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &= \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|A(\mathbf{x} + \mathbf{h}) - A\mathbf{x} - A\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\ &= \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|A\mathbf{x} + A\mathbf{h} - A\mathbf{x} - A\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\ &= \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{0}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\ &= 0. \end{aligned}$$

We therefore have $f'(\mathbf{x}) = A$ when $f'(\mathbf{x}) = A\mathbf{x}$. Is this surprising though? Think of the case where $n = m = 1$ and $f(x) = ax$ for some $a \in \mathbb{R}$ (Example 8.9). In this case, $f'(x) = a$, which is consistent with $f'(\mathbf{x}) = A$.

Remark 9.1 ($f'(\mathbf{x})$, Remainder Term). The derivative $f'(\mathbf{x})$ underlies the approximation

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) \approx \mathbf{f}(\mathbf{x}) + f'(\mathbf{x})\mathbf{h}.$$

This approximation can be written as an equality if we account for the approximation's error/remainder, which we will write as $\mathbf{r}(\mathbf{h})$.

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + f'(\mathbf{x})\mathbf{h} + \mathbf{r}(\mathbf{h})$$

We of course need $\mathbf{r}(\mathbf{h})$ to satisfy some properties to be compatible with Definition 9.1.

$$\begin{aligned}
& \mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{h} + \mathbf{r}(\mathbf{h}) \\
\implies & \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h} = \mathbf{r}(\mathbf{h}) \\
\implies & \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} = \|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m} \\
\implies & \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = \frac{\|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
\implies & \underbrace{\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}}}_{0} = \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
\implies & \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0
\end{aligned}$$

This gives us an alternate definition for $\mathbf{f}'(\mathbf{x})$, namely it is the linear transformation satisfying

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{h} + \mathbf{r}(\mathbf{h})$$

where $\mathbf{r}(\mathbf{h})$ satisfies $\lim_{\mathbf{h} \rightarrow \mathbf{0}} \|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m} / \|\mathbf{h}\|_{\mathbb{R}^n} = 0$.

Okay all of this is cool, but how the heck do we actually calculate $\mathbf{f}'(\mathbf{x})$? We were able to intuit $\mathbf{f}'(\mathbf{x})$ in Example 9.3 and verify our educated guess, but we didn't really calculate anything. Before we discuss how to calculate $\mathbf{f}'(\mathbf{x})$, we should consider whether there is *a* $\mathbf{f}'(\mathbf{x})$ or *many* $\mathbf{f}'(\mathbf{x})$. Is the derivative $\mathbf{f}'(\mathbf{x})$ even unique? The space $L(\mathbb{R}^n, \mathbb{R}^m)$ contains so many possible linear transformations, so who is to say that there aren't two elements of $L(\mathbb{R}^n, \mathbb{R}^m)$ satisfying Definition 9.1 for $\mathbf{f}(\mathbf{x})$? This is precisely what our next result tells us.

Theorem 9.1 (Uniqueness of $\mathbf{f}'(\mathbf{x})$). Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$ where $E \subset \mathbb{R}^n$, $\mathbf{x} \in E$, and

$$\begin{aligned}
& \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_1\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0 \\
& \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_2\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0
\end{aligned}$$

for $A_1, A_2 \in L(\mathbb{R}^n, \mathbb{R}^m)$. Then $\mathbf{f}'(\mathbf{x}) = A_1 = A_2$.

Proof. Define $B = A_1 - A_2$, where $B \in L(\mathbb{R}^n, \mathbb{R}^m)$. It suffices to show that $B = 0$,¹⁴⁴ that is $B\mathbf{h} = \mathbf{0}$ for all

¹⁴⁴That is, B is the zero matrix which corresponds to the additive identity $T_0(\mathbf{x}) = \mathbf{0}$ in the space $L(\mathbb{R}^n, \mathbb{R}^m)$.

$\mathbf{h} \neq \mathbf{0} \in \mathbb{R}^n$. We have

$$\begin{aligned}
\|B\mathbf{h}\|_{\mathbb{R}^m} &= \|(A_1 - A_2)\mathbf{h}\|_{\mathbb{R}^m} \\
&= \|A_1\mathbf{h} - A_2\mathbf{h}\|_{\mathbb{R}^m} \\
&= \|\mathbf{0} + A_1\mathbf{h} - A_2\mathbf{h}\|_{\mathbb{R}^m} \\
&= \|(\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x} + \mathbf{h})) + (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x})) + A_1\mathbf{h} - A_2\mathbf{h}\|_{\mathbb{R}^m} \\
&= \|(\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_1\mathbf{h}) + (\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_2\mathbf{h})\|_{\mathbb{R}^m} \\
&\leq \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_1\mathbf{h}\|_{\mathbb{R}^m} + \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_2\mathbf{h}\|_{\mathbb{R}^m} \quad (\text{Triangle Inequality}) \\
\frac{\|B\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &\leq \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_1\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} + \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_2\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|B\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &\leq \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_1\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} + \lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A_2\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
&\leq 0 + 0 \\
\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|B\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &= 0
\end{aligned}$$

Now fix $\mathbf{h}_0 \neq \mathbf{0} \in \mathbb{R}^n$. We know

$$\lim_{\mathbf{h}_0 \rightarrow \mathbf{0}} \frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = 0,$$

but this does *not imply* that $\|B\mathbf{h}_0\|_{\mathbb{R}^m} = 0$, as we are taking the limit as $\mathbf{h}_0 \rightarrow \mathbf{0}$.¹⁴⁵ If we could express this limit as

$$\lim_{? \rightarrow \mathbf{0}} \frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = 0,$$

for some $? \neq \mathbf{0}$ independent of $\|B\mathbf{h}_0\|_{\mathbb{R}^m}/\|\mathbf{h}_0\|_{\mathbb{R}^n}$, then $\|B\mathbf{h}_0\|_{\mathbb{R}^m}/\|\mathbf{h}_0\|_{\mathbb{R}^n}$ would be a constant unrelated to the limit, which would simplify our problem. To do this, we can rewrite this limit as

$$\lim_{\mathbf{h}_0 \rightarrow \mathbf{0}} \frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = \lim_{t \rightarrow 0} \frac{\|B(t\mathbf{h}_0)\|_{\mathbb{R}^m}}{\|t\mathbf{h}_0\|_{\mathbb{R}^n}} = 0$$

where we let $\mathbf{h}_0 \rightarrow \mathbf{0}$ by scaling it by t and letting $t \rightarrow 0$. We can use the linearity of t and the definition of a norm to write

$$0 = \lim_{t \rightarrow 0} \frac{\|B(t\mathbf{h}_0)\|_{\mathbb{R}^m}}{\|t\mathbf{h}_0\|_{\mathbb{R}^n}} = \lim_{t \rightarrow 0} \frac{\|tB\mathbf{h}_0\|_{\mathbb{R}^m}}{\|t\mathbf{h}_0\|_{\mathbb{R}^n}} = \lim_{t \rightarrow 0} \frac{|t|}{|t|} \frac{\|B(\mathbf{h}_0)\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = \lim_{t \rightarrow 0} \frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}}.$$

We know have the limit of a constant which equals zero in

$$\lim_{t \rightarrow 0} \frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = 0,$$

so $\frac{\|B\mathbf{h}_0\|_{\mathbb{R}^m}}{\|\mathbf{h}_0\|_{\mathbb{R}^n}} = 0$. This, along with $\mathbf{h}_0 \neq \mathbf{0}$, implies that $\|B(\mathbf{h}_0)\|_{\mathbb{R}^m} = 0$. Therefore $B\mathbf{h}_0 = \mathbf{0}$ for all $\mathbf{h}_0 \neq \mathbf{0}$, which is the desired result. \square

The crucial step in this proof is rewriting our limit as the limit of a scalar, and then eliminating that scalar from the function in question using the linearity of B , hence this presentation being belabored and much more pedantic than that of Rudin (1976).

Now that we know $\mathbf{f}'(\mathbf{x})$ is unique, we can cover how to calculate this transformation. To do so we will turn to a familiar concept from multivariable calculus.

¹⁴⁵ $\lim_{x \rightarrow 0} \frac{3x^2}{x} = 0$, but this does not imply that $3x^2 = 0$ for all x .

9.2 Partial Derivatives and the Jacobian

At first differentiating $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ may seem daunting. The function is comprised of m components $f_j(\mathbf{x})$, each of which is a function of n variables x_j . We came across a similar problem when determining whether \mathbf{f} was continuous in Section 8.4. Proposition 8.2 allowed us to decompose this problem by looking at each component separately. This is exactly how we will handle differentiation of \mathbf{f} .

Consider $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$. While we've eliminated one dimension of complexity in the form of multiple components, f_j is still a function of n variables, so which variable(s) do we differentiate with respect to? How do we treat the other variables? The answer comes in the form of partial derivatives.

Definition 9.2. Suppose $f : E \rightarrow \mathbb{R}$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If the limit

$$\frac{\partial f}{\partial x_j}(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h}$$

exists, we refer to $\frac{\partial f}{\partial x_j}(\mathbf{x})$ as the *partial derivative of f with respect to x_j evaluated at \mathbf{x}* .

The partial derivative $\frac{\partial f}{\partial x_j}$ only considers the rate at which f changes with respect to x_j , holding all the other variables fixed. As such, calculating a partial derivative is nearly identical to calculating a derivative for a single variable function, and all the properties of Section 5 extend to partial derivatives when considered in isolation. It's only a matter of treating all variables other than x_j as constants. We can also interpret the partial derivative in terms of linear approximation if we hold (Figure 96).

$$f(\mathbf{x} + h\mathbf{e}_j) = f(\mathbf{x}) + \left(\frac{\partial f}{\partial x_j}(\mathbf{x}) \cdot h \right) \mathbf{e}_j$$

For $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, each f_i has n partial derivatives. If we consider all f_i , we have $m \times n$ partial derivatives associated with \mathbf{f} (evaluated at a point \mathbf{x}):

$$\frac{\partial f_1}{\partial x_1}(\mathbf{x}) \quad \frac{\partial f_1}{\partial x_2}(\mathbf{x}) \quad \cdots \quad \frac{\partial f_1}{\partial x_n}(\mathbf{x})$$

$$\frac{\partial f_2}{\partial x_1}(\mathbf{x}) \quad \frac{\partial f_2}{\partial x_2}(\mathbf{x}) \quad \cdots \quad \frac{\partial f_2}{\partial x_n}(\mathbf{x})$$

$$\vdots \quad \vdots \quad \ddots \quad \vdots$$

$$\frac{\partial f_m}{\partial x_1}(\mathbf{x}) \quad \frac{\partial f_m}{\partial x_2}(\mathbf{x}) \quad \cdots \quad \frac{\partial f_m}{\partial x_n}(\mathbf{x})$$

Each of these partial derivatives approximate the change in one component in response to changing one variable. Is it possible to combine these isolated linear approximations into $\mathbf{f}'(\mathbf{x})$? Fortunately the answer is yes! It is no coincidence at all that $\mathbf{f}'(\mathbf{x})$ is an $m \times n$ matrix, and we have $m \times n$ scalar values $\frac{\partial f_i}{\partial x_j}$.

Theorem 9.2 (Differentiability \implies Partial Differentiability). Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If \mathbf{f} is differentiable at \mathbf{x} , then $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ exist for all x_j and f_i , and

$$\mathbf{f}'(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}.$$

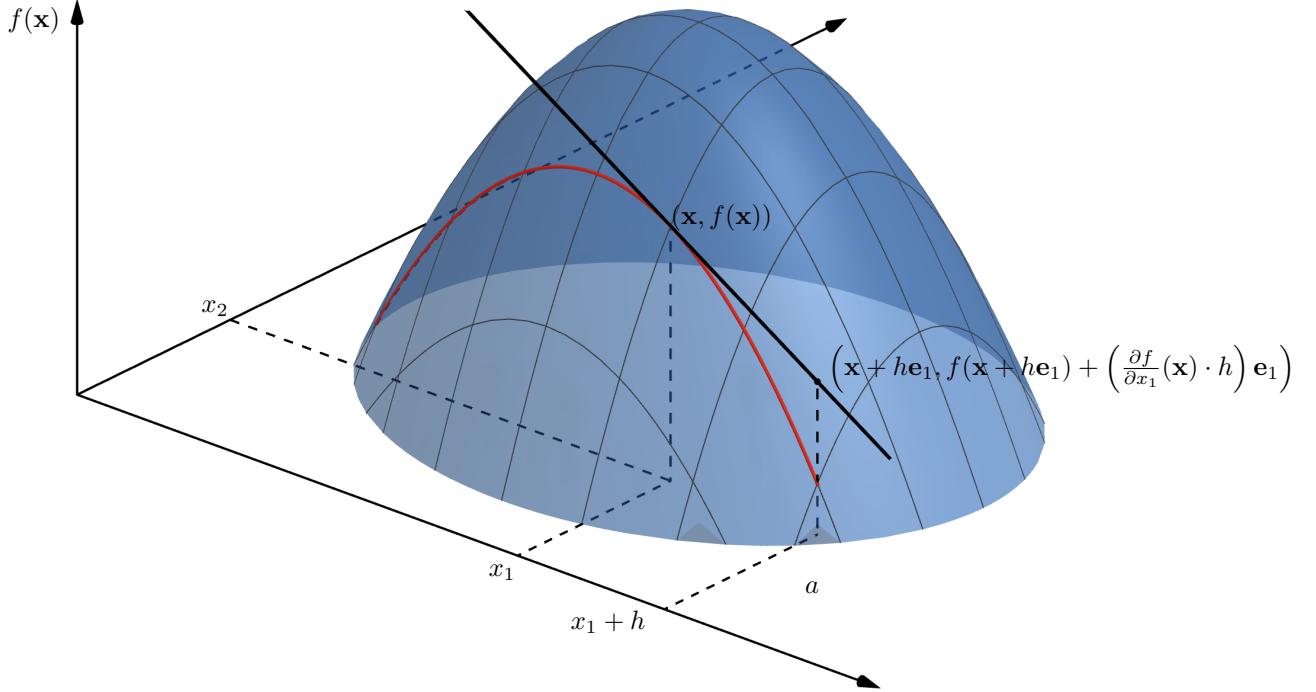


Figure 96:

Proof. Since \mathbf{f} is differentiable at \mathbf{x} we have

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{h} + \mathbf{r}(\mathbf{h})$$

where $\lim_{\mathbf{h} \rightarrow 0} \|\mathbf{r}(\mathbf{h})\|_{\mathbb{R}^m} / \|\mathbf{h}\|_{\mathbb{R}^n} = 0$ (Remark 9.1). For a fixed $j \in 1, \dots, n$, take $\mathbf{h} = h\mathbf{e}_j$, giving

$$\mathbf{f}(\mathbf{x} + h\mathbf{e}_j) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})(h\mathbf{e}_j) + \mathbf{r}(h\mathbf{e}_j)$$

where $\lim_{h \rightarrow 0} \|\mathbf{r}(h\mathbf{e}_j)\|_{\mathbb{R}^m} / \|h\mathbf{e}_j\|_{\mathbb{R}^n} = 0$. Since $\mathbf{f}'(\mathbf{x})$ is a linear transformation, $\mathbf{f}'(\mathbf{x})(h\mathbf{e}_j) = h\mathbf{f}'(\mathbf{x})\mathbf{e}_j$.

$$\begin{aligned} & \mathbf{f}(\mathbf{x} + h\mathbf{e}_j) = \mathbf{f}(\mathbf{x}) + \mathbf{f}'(\mathbf{x})(h\mathbf{e}_j) + \mathbf{r}(h\mathbf{e}_j) \\ \implies & \mathbf{f}(\mathbf{x} + h\mathbf{e}_j) = \mathbf{f}(\mathbf{x}) + h\mathbf{f}'(\mathbf{x})\mathbf{e}_j + \mathbf{r}(h\mathbf{e}_j) \\ \implies & \frac{\mathbf{f}(\mathbf{x} + h\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{h} + \frac{\mathbf{r}(h\mathbf{e}_j)}{h} = \mathbf{f}'(\mathbf{x})\mathbf{e}_j \\ \implies & \lim_{h \rightarrow 0} \frac{\mathbf{f}(\mathbf{x} + h\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{h} + \underbrace{\lim_{h \rightarrow 0} \frac{\mathbf{r}(h\mathbf{e}_j)}{h}}_0 = \lim_{h \rightarrow 0} \mathbf{f}'(\mathbf{x})\mathbf{e}_j \\ \implies & \lim_{h \rightarrow 0} \frac{\mathbf{f}(\mathbf{x} + h\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{h} = \mathbf{f}'(\mathbf{x})\mathbf{e}_j \end{aligned}$$

If we write this out in terms of vectors we have

$$\lim_{h \rightarrow 0} \begin{bmatrix} \frac{f_1(\mathbf{x} + h\mathbf{e}_j) - f_1(\mathbf{x})}{h} \\ \vdots \\ \frac{f_m(\mathbf{x} + h\mathbf{e}_j) - f_m(\mathbf{x})}{h} \end{bmatrix} = \begin{bmatrix} f'_{1j}(\mathbf{x}) \\ \vdots \\ f'_{mj}(\mathbf{x}) \end{bmatrix}$$

where $f'_{ij}(\mathbf{x})$ is the (i, j) th entry of $\mathbf{f}'(\mathbf{x})$. A limit of a vector function exists if and only if the limit of the corresponding components exist,¹⁴⁶, so we have

$$\lim_{h \rightarrow 0} \frac{f_i(\mathbf{x} + h\mathbf{e}_j) - f_i(\mathbf{x})}{h} = f'_{ij}(\mathbf{x}) \quad \forall i.$$

By Definition 9.2 this equality is

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) = f'_{ij}(\mathbf{x}).$$

□

The matrix given in Theorem 9.2 is so important, it is often referred to by a specific name and notation.

Definition 9.3. Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$ is differentiable at an interior point $\mathbf{x} \in E$. The *Jacobian (matrix) of \mathbf{f} (evaluated at \mathbf{x})* is the $m \times n$ matrix

$$\mathbf{J}_{\mathbf{f}}(\mathbf{x}) = \mathbf{f}'(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \quad (34)$$

At times it will be necessary to reference $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ for all f_i comprising \mathbf{f} . To do this in a succinct fashion, we will define the vector

$$\frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\mathbf{x}) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(\mathbf{x}) \end{bmatrix}.$$

With this notation we can write the Jacobian as

$$\mathbf{J}_{\mathbf{f}}(\mathbf{x}) = \left[\frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) \quad \cdots \quad \frac{\partial \mathbf{f}}{\partial x_n}(\mathbf{x}) \right]$$

Example 9.4. Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$. In this case

$$f'(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}) \quad \frac{\partial f}{\partial x_2}(\mathbf{x}) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\mathbf{x}) \right].$$

The matrix $f'(\mathbf{x})$ is a linear transformation mapping $\mathbb{R}^n \mapsto \mathbb{R}$, forming a hyperplane given by the equation

$$f(\mathbf{x}) + f'(\mathbf{x})\mathbf{h} = f(\mathbf{x}) + \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{x})h_j.$$

This hyperplane is tangent to the surface $f(\mathbf{x})$ at the point \mathbf{x} . If $n = 2$, then the plane is given by

$$f(\mathbf{x}) + \frac{\partial f}{\partial x_1}(\mathbf{x})h_1 + \frac{\partial f}{\partial x_2}(\mathbf{x})h_2.$$

¹⁴⁶This has never been explicitly proved, but the result can be shown in a similar fashion to Proposition 8.2.

Example 9.5. Recall Example 8.17 where $\mathbf{f}(\mathbf{x})$ as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_1^2 - x_2^2 \\ 2x_1 x_2 \end{bmatrix}.$$

For some $\mathbf{x} \in \mathbb{R}^2$, we have

$$\mathbf{f}'(\mathbf{x}) = \begin{bmatrix} 2x_1 & -2x_2 \\ 2x_2 & 2x_1 \end{bmatrix}.$$

Let's use $\mathbf{f}'(\mathbf{x})$ to approximate \mathbf{f} near the point $\mathbf{x} = (2, 1)$. At this point we have

$$\mathbf{f}'(2, 1) = \begin{bmatrix} 2(2) & -2(1) \\ 2(1) & 2(2) \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ 2 & 4 \end{bmatrix}.$$

The linear transformation $\mathbf{f}'(2, 1)$, superimposed over the nonlinear \mathbf{f} , is shown in Figure 97.

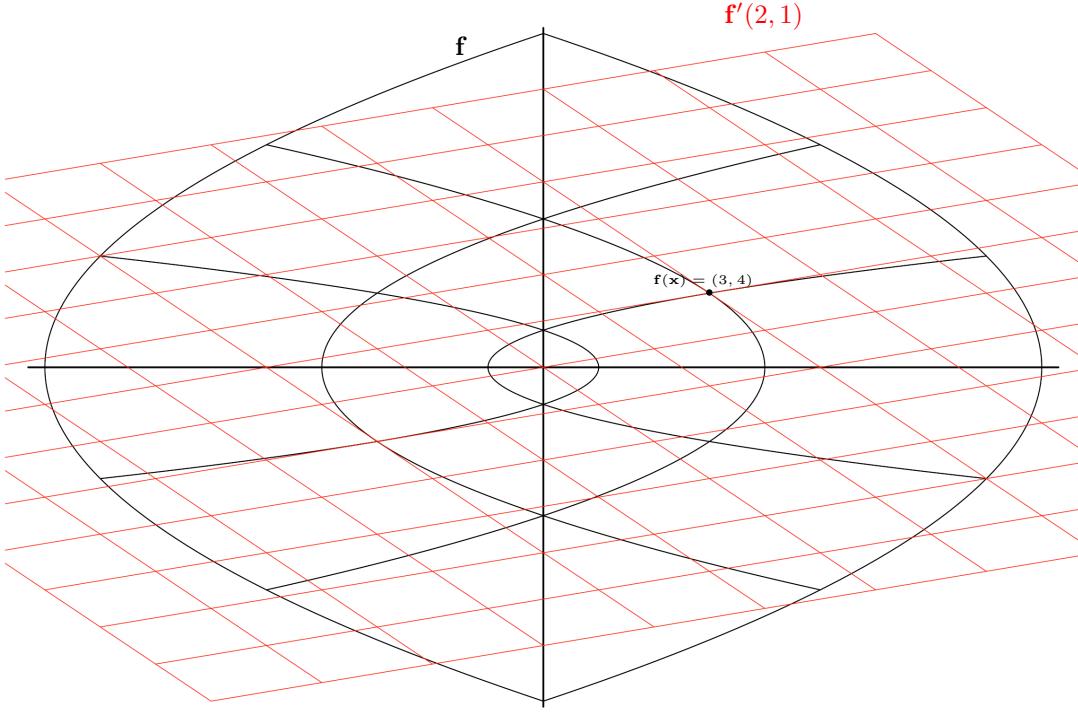


Figure 97:

We have

$$\mathbf{f} \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \mathbf{h} \right) \approx \mathbf{f} \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} \right) + \mathbf{f}' \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} \right) \mathbf{h}$$

for some small displacement vector \mathbf{h} . As $\mathbf{h} \rightarrow \mathbf{0}$, this approximation becomes better and better. Figure 97 illustrates this, as the linear approximation coincides with \mathbf{f} fairly well for points near $\mathbf{f}(2, 1) = (3, 4)$. This becomes even clearer if we zoom in on $\mathbf{f}(\mathbf{x}) = (3, 4)$ and include more grid lines (Figure 98 and Figure 99).

Example 9.6 (Polar Coordinates). Define the function $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as

$$\mathbf{f}(\theta, r) = \begin{bmatrix} r \cos \theta \\ r \sin \theta \end{bmatrix}.$$

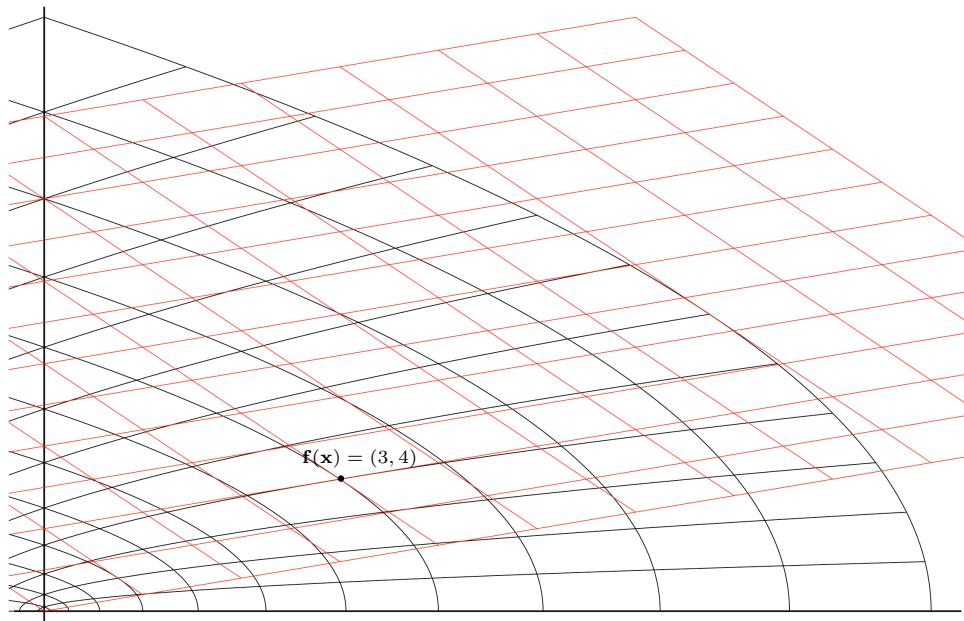


Figure 98:

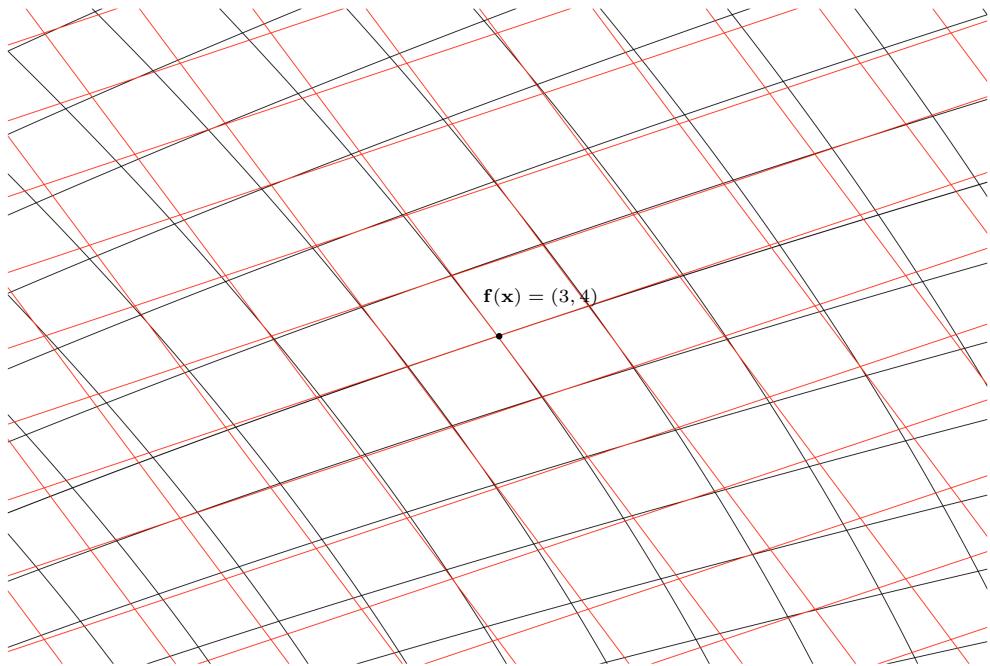


Figure 99:

For some $(\theta, r) \in \mathbb{R}^2$,

$$\mathbf{f}'(\theta, r) = \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & -r \cos \theta \end{bmatrix}.$$

Example 9.7 (Spherical Coordinates). Define $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ as

$$\mathbf{f}(r, \varphi, \theta) = \begin{bmatrix} r \sin \theta \cos \varphi \\ r \sin \theta \sin \varphi \\ r \cos \theta \end{bmatrix}.$$

For some $(r, \varphi, \theta) \in \mathbb{R}^3$,

$$\mathbf{f}'(\theta, r) = \begin{bmatrix} \sin \varphi \cos \theta & r \cos \varphi \cos \theta & -r \sin \varphi \sin \theta \\ \sin \varphi \cos \theta & r \cos \varphi \sin \theta & r \sin \varphi \cos \theta \\ \cos \varphi & -r \sin \varphi & 0 \end{bmatrix}.$$

Theorem 9.3 give us the means of calculating the total derivative $\mathbf{f}'(\mathbf{x})$, and in doing so verifies that $\frac{\partial f_i}{\partial x_j}$ exists for all (i, j) . The converse of this second statement need not hold.

Example 9.8. Define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ as

$$f(\mathbf{x}) = \begin{cases} (x_1^2 x_2) / (x_1^4 + x_2^2) & \text{if } \mathbf{x} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{x} = \mathbf{0} \end{cases}.$$

Let's calculate the partial derivatives at the point $\mathbf{0}$, thereby verifying their existence.

$$\begin{aligned} \frac{\partial f}{\partial x_1}(\mathbf{0}) &= \lim_{h \rightarrow 0} \frac{f(\mathbf{0} + h\mathbf{e}_1) - f(\mathbf{0})}{h} \\ &= \lim_{h \rightarrow 0} \frac{(h^2 \cdot 0) / (h^4 + 0^2)}{h} \\ &= 0 \\ \frac{\partial f}{\partial x_2}(\mathbf{0}) &= \lim_{h \rightarrow 0} \frac{f(\mathbf{0} + h\mathbf{e}_2) - f(\mathbf{0})}{h} \\ &= \lim_{h \rightarrow 0} \frac{(0 \cdot h^2) / (0^4 + h^2)}{h} \\ &= 0 \end{aligned}$$

If \mathbf{f} were differentiable at $\mathbf{0}$, then there would exist some $A = [a_1 \ a_2]$ such that

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{|\mathbf{f}(\mathbf{0} + \mathbf{h}) - \mathbf{f}(\mathbf{0}) - A\mathbf{h}|}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0$$

for $\mathbf{h} = (h_1, h_2)$. We have

$$\begin{aligned} \frac{|\mathbf{f}(\mathbf{0} + \mathbf{h}) - \mathbf{f}(\mathbf{0}) - A\mathbf{h}|}{\|\mathbf{h}\|_{\mathbb{R}^n}} &= \frac{\frac{h_1^2 h_2}{h_1^4 + h_2^2} - 0 - (a_1 h_1 + a_2 h_2)}{\sqrt{h_1^2 + h_2^2}} \\ &\rightarrow \infty \end{aligned}$$

So \mathbf{f} is not differentiable.

It turns out that the existence of the partial derivatives in the previous example are not sufficient in establishing \mathbf{f} 's differentiability at \mathbf{x} because the partial derivatives fail to be continuous at the point \mathbf{x} . With the additional caveat of all the partial derivatives associated with \mathbf{f} being continuous at \mathbf{x} , we are able to formulate a converse for Theorem 9.2

Theorem 9.3 ((Continuous) Partial Differentiability \implies Differentiability). Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ exist for all (i, j) , and $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ are continuous at \mathbf{x} , then \mathbf{f} is differentiable at \mathbf{x} .

Proof. Suppose $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ exist for all (i, j) , and $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ are continuous at \mathbf{x} .¹⁴⁷ We will verify that

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0,$$

for some $m \times n$ matrix A , namely the Jacobian $A = \mathbf{J}_f(\mathbf{x})$ (Definition 9.3). Appealing directly to the definition of a limit (Definition 4.1), it suffices to show that for all $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{J}_f(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} < \varepsilon$$

when $\|\mathbf{h} - \mathbf{0}\| < \delta$. This is equivalent to

$$\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{J}_f(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} < \varepsilon \|\mathbf{h}\|_{\mathbb{R}^n} \quad (35)$$

whenever $\|\mathbf{h}\| < \delta$.

Define

$$\frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\mathbf{x}) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(\mathbf{x}) \end{bmatrix}.$$

This vector valued function $\frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})$ is continuous at \mathbf{x} by the continuity of its components (Proposition 8.2), so there exists a $\delta_j > 0$ such that

$$\left\| \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{h}) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x} + \mathbf{h}) \right\| < \frac{\varepsilon}{nm}$$

whenever $\|(\mathbf{x} + \mathbf{h}) - \mathbf{x}\| = \|\mathbf{h}\| < \delta$. If we define $\delta = \min\{\delta_1, \dots, \delta_n\}$, then whenever $\|\mathbf{h}\|_{\mathbb{R}^n} < \delta$,

$$\left\| \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{h}) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x} + \mathbf{h}) \right\|_{\mathbb{R}^m} < \frac{\varepsilon}{nm} \quad (\text{for all } j = 1, \dots, n). \quad (36)$$

Now that we listed what we want to show, and what we know, let's jump in!

If $\mathbf{h} = (h_1, \dots, h_n)$, then (35) becomes

$$\left\| \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \sum_{j=1}^n h_j \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) \right\|_{\mathbb{R}^m} < \varepsilon \|\mathbf{h}\|_{\mathbb{R}^n},$$

where the matrix $\mathbf{J}_f(\mathbf{x})\mathbf{h}$ is written in terms of the sum of vectors. Fix $j = 1$. By the MVT (Corollary 5.1),

$$\frac{f_i(\mathbf{x} + h_1 \mathbf{e}_1) - f_i(\mathbf{x})}{h_1} = \frac{\partial f_i}{\partial x_1}(\mathbf{x} + t_1 \mathbf{e}_1) \quad (37)$$

for some $t_2 \in (0, h_1)$, for all f_i and x_1 fixed. Now because $t_i < h_1$,

$$\|\mathbf{x} + t_i \mathbf{e}_1 - \mathbf{x}\|_{\mathbb{R}^n} = \|t_i \mathbf{e}_1\|_{\mathbb{R}^n} \leq \|\mathbf{h}\|_{\mathbb{R}^n} < \delta$$

so the conditions under which (36) hold are satisfied for $\mathbf{x} + t_i \mathbf{e}_1$ and \mathbf{x} :

$$\left\| \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x} + t_i \mathbf{e}_1) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) \right\|_{\mathbb{R}^m} < \frac{\varepsilon}{nm},$$

¹⁴⁷Keep an eye out for when we use this continuity in the proof!

which gives,

$$\left| \frac{\partial f_i}{\partial x_1}(\mathbf{x} + t\mathbf{e}_1) - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right| < \left\| \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x} + t\mathbf{e}_1) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) \right\|_{\mathbb{R}^m} < \frac{\varepsilon}{nm} \quad (38)$$

for all f_i . Combing (37) and (38) gives

$$\begin{aligned} & \left| \frac{f_i(\mathbf{x} + h_1\mathbf{e}_1) - f_i(\mathbf{x})}{h_1} - \frac{\partial f_i}{\partial x_1}(\mathbf{x}) \right| < \frac{\varepsilon}{nm} \\ \implies & \left| f_i(\mathbf{x} + h_1\mathbf{e}_1) - f_i(\mathbf{x}) - \frac{\partial f_i}{\partial x_1}(\mathbf{x})h_1 \right| < \frac{\varepsilon|h_1|}{nm} \end{aligned}$$

for all i . Now if we sum over all the i , we can conclude that

$$\begin{aligned} \left\| \mathbf{f}(\mathbf{x} + h_1\mathbf{e}_1) - \mathbf{f}(\mathbf{x}) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x})h_1 \right\|_{\mathbb{R}^m} & \leq \sum_{i=1}^m \left| f_i(\mathbf{x} + h_1\mathbf{e}_1) - f_i(\mathbf{x}) - \frac{\partial f_i}{\partial x_1}(\mathbf{x})h_1 \right| \quad (\text{Triangle Inequality, Definition 8.2}) \\ & < \sum_{i=1}^m \frac{\varepsilon|h_1|}{nm} \\ & = m \cdot \frac{\varepsilon|h_1|}{nm} \\ & = \frac{\varepsilon|h_1|}{n} \\ & \leq \frac{\varepsilon \|h\|_{\mathbb{R}^n}}{n} \\ \left\| \mathbf{f}(\mathbf{x} + h_1\mathbf{e}_1) - \mathbf{f}(\mathbf{x}) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x})h_1 \right\|_{\mathbb{R}^m} & < \frac{\varepsilon \|h\|_{\mathbb{R}^n}}{n} \end{aligned} \quad (39)$$

Now consider x_2 . By the MVT (Corollary 5.1),

$$\frac{f_i(\mathbf{x} + h_1\mathbf{e}_1 + h_2\mathbf{e}_2) - f_i(\mathbf{x} + h_1\mathbf{e}_1)}{h_2} = \frac{\partial f_i}{\partial x_2}(\mathbf{x} + t_2\mathbf{e}_2)$$

for some $t_2 \in (0, h_2)$, for all f_i and x_2 fixed. By the same line of reason used to conclude (40), we have

$$\left\| \mathbf{f}(\mathbf{x} + h_1\mathbf{e}_1 + h_2\mathbf{e}_2) - \mathbf{f}(\mathbf{x} + h_1\mathbf{e}_1) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x})h_2 \right\|_{\mathbb{R}^m} < \frac{\varepsilon \|h\|_{\mathbb{R}^n}}{n}. \quad (40)$$

This process of applying the MVT for $\frac{\partial f_i}{\partial x_j}$ for all i , and then establishing an inequality with the same form of (39) and (40) can be repeated for x_3, \dots, x_n . After doing this for $j = 1, \dots, n$, we have n inequalities of the form

$$\left\| \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^j h_\ell \mathbf{e}_\ell) - \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^{j-1} h_\ell \mathbf{e}_\ell) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})h_j \right\|_{\mathbb{R}^m} < \frac{\varepsilon \|h\|_{\mathbb{R}^n}}{n}.$$

Summing these inequalities gives

$$\sum_{j=1}^n \left\| \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^j h_\ell \mathbf{e}_\ell) - \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^{j-1} h_\ell \mathbf{e}_\ell) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})h_j \right\|_{\mathbb{R}^m} < \varepsilon \|h\|_{\mathbb{R}^m}. \quad (41)$$

If we apply the triangle equality to the left hand side, we have a telescoping series which reduces to the left

hand side of (35):

$$\begin{aligned}
\sum_{j=1}^n \left\| \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^j h_\ell \mathbf{e}_\ell) - \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^{j-1} h_\ell \mathbf{e}_\ell) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) h_j \right\|_{\mathbb{R}^m} &\geq \left\| \sum_{j=1}^n \left(\mathbf{f}(\mathbf{x} + \sum_{\ell=1}^j h_\ell \mathbf{e}_\ell) - \mathbf{f}(\mathbf{x} + \sum_{\ell=1}^{j-1} h_\ell \mathbf{e}_\ell) - \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) h_j \right) \right\|_{\mathbb{R}^m} \\
&= \left\| \left(\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1) - \mathbf{f}(\mathbf{x}) - \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) h_1 \right) \right. \\
&\quad + \left. \left(\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + h_2 \mathbf{e}_2) - \mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1) - \frac{\partial \mathbf{f}}{\partial x_2}(\mathbf{x}) h_2 \right) + \dots \right. \\
&\quad + \left. \left(\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_{n-1} \mathbf{e}_{n-1}) - \mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_{n-2} \mathbf{e}_{n-2}) - \frac{\partial \mathbf{f}}{\partial x_{n-1}}(\mathbf{x}) h_{n-1} \right) \right. \\
&\quad + \left. \left[\underbrace{\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_n \mathbf{e}_n)}_{\mathbf{f}(\mathbf{x} + \mathbf{h})} - \mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_{n-1} \mathbf{e}_{n-1}) - \frac{\partial \mathbf{f}}{\partial x_n}(\mathbf{x}) h_n \right] \right\|_{\mathbb{R}^m} \\
&= \left\| \mathbf{f}(\mathbf{x} + \mathbf{h}) + \underbrace{[\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_{n-1} \mathbf{e}_{n-1}) - \mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1 + \dots + h_{n-1} \mathbf{e}_{n-1})]}_0 + \dots \right. \\
&\quad + \left. \underbrace{[\mathbf{f}(\mathbf{x} + h_1 \mathbf{e}_1) - \mathbf{f}(\mathbf{x})]}_0 - \mathbf{f}(\mathbf{x}) - \left(\frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) h_1 + \dots + \frac{\partial \mathbf{f}}{\partial x_n}(\mathbf{x}) h_n \right) \right\|_{\mathbb{R}^m} \\
&= \left\| \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \sum_{j=1}^n h_j \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) \right\|_{\mathbb{R}^m}
\end{aligned}$$

Combined with (41), we have

$$\left\| \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \sum_{j=1}^n h_j \frac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}) \right\|_{\mathbb{R}^m} < \varepsilon \|\mathbf{h}\|_{\mathbb{R}^n},$$

which is the desired result! \square

Remark 9.2 (Directional Derivatives). The partial derivatives of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ happen to be a special case of directional derivatives. For some $\mathbf{u} \in \mathbb{R}^n$, the directional derivative of f is given as

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}.$$

This measures the instantaneous rate of change of f at the point \mathbf{x} while moving in the direction of \mathbf{u} . We have

$$\frac{\partial f_i}{\partial x_j} = D_{\mathbf{e}_j} f_i(\mathbf{x}).$$

Munkres (1999), Apostol (1974), and Tao (2016b) define directional derivatives and present partial derivatives as a special case, while Rudin (1976) simply eschews their presentation all together.

9.3 The Chain Rule

For differentiable functions $f : \mathbb{R} \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow \mathbb{R}$, the chain rule (Theorem 5.3) allowed us to conclude that $g \circ f$ was differentiable and

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

The intuition behind this result was that $f'(x)$ acts as a sort of “exchange rate” between a change in x and a change in the input of g (which is $f(x)$). This theorem is readily extended to the total derivative.

Theorem 9.4 (The Chain Rule). Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$, where $E \subset \mathbb{R}^n$, is differentiable at an interior point $x \in E$. If $\phi : D \rightarrow \mathbb{R}^k$, where $\mathbf{f}(E) \subset D \subset \mathbb{R}^m$, and ϕ is differentiable at $\mathbf{f}(\mathbf{x})$, then the function

$$\mathbf{F} = \phi \circ \mathbf{f}$$

is differentiable at \mathbf{x} . Furthermore,

$$\mathbf{F}'(\mathbf{x}) = \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x}).$$

Proof. The function \mathbf{f} is differentiable at \mathbf{x} , and ϕ at $\mathbf{f}(\mathbf{x})$, so

$$\begin{aligned}\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &= 0, \\ \lim_{\mathbf{k} \rightarrow \mathbf{0}} \frac{\|\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}\|_{\mathbb{R}^k}}{\|\mathbf{k}\|_{\mathbb{R}^m}} &= 0.\end{aligned}$$

These equalities can be expressed as

$$\begin{aligned}\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &= 0 \\ \implies \lim_{\mathbf{h} \rightarrow \mathbf{0}} \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} &= \lim_{\mathbf{h} \rightarrow \mathbf{0}} r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} &= r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^n}\end{aligned}\tag{42}$$

$$\|\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}\|_{\mathbb{R}^k} = q(\mathbf{k}) \|\mathbf{k}\|_{\mathbb{R}^m}\tag{43}$$

where $q(\mathbf{k})$ and $r(\mathbf{h})$ are remainder terms which go to 0 as $\mathbf{k} \rightarrow \mathbf{0}$ and $\mathbf{h} \rightarrow \mathbf{0}$.¹⁴⁸ We must show

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^k}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0.$$

The apparent difficulty in proving this result is that the desired equality takes $\mathbf{h} \rightarrow \mathbf{0}$, while also including ϕ' , which is defined for a displacement vector $\mathbf{k} \rightarrow \mathbf{0}$. To resolve this set $\mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})$. This way $\lim_{\mathbf{h} \rightarrow \mathbf{0}} \mathbf{k} = \mathbf{0}$, and we can use (43) when $\mathbf{h} \rightarrow \mathbf{0}$. We have

$$\begin{aligned}\|\mathbf{k}\|_{\mathbb{R}^m} &= \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})\|_{\mathbb{R}^m} \\ &= \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) + [\mathbf{f}'(\mathbf{x})\mathbf{h} - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^m} \quad (\text{add } \mathbf{0}) \\ &= \|\mathbf{f}'(\mathbf{x})\mathbf{h} + [\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^m} \\ &\leq \|\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} + \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} \quad (\text{Triangle Inequality}) \\ &= \|\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m} + r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^n} \quad (\text{Equation (42)}) \\ &\leq \|\mathbf{h}\|_{\mathbb{R}^n} \|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^n} \quad (\text{Lemma 8.1}) \\ &= \left(\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + r(\mathbf{h}) \right) \|\mathbf{h}\|_{\mathbb{R}^n}.\end{aligned}\tag{44}$$

We also have

$$\begin{aligned}\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} &= \phi(\mathbf{f}(\mathbf{x} + \mathbf{h})) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} \\ &= \phi(\mathbf{f}(\mathbf{x} + \mathbf{h}) - [\mathbf{f}(\mathbf{x}) + \mathbf{f}(\mathbf{x})]) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} \quad (\text{add } \mathbf{0}) \\ &= \phi(\mathbf{f}(\mathbf{x}) + \underbrace{[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})]}_{\mathbf{k}}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} \\ &= \phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} \quad (\text{definition of } \mathbf{k}) \\ &= \phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h} + [\phi'(\mathbf{f}(\mathbf{x}))\mathbf{k} - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{h}] \quad (\text{add } \mathbf{0}) \\ &= (\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}) + [\phi'(\mathbf{f}(\mathbf{x}))\mathbf{k} - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h}] \\ &= (\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}) + \phi'(\mathbf{f}(\mathbf{x}))[\mathbf{k} - \mathbf{f}'(\mathbf{x})\mathbf{h}] \quad (\phi'(\mathbf{f}(\mathbf{x})) \text{ linear}) \\ &= (\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}) + \phi'(\mathbf{f}(\mathbf{x}))[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}] \quad (\text{definition of } \mathbf{k})\end{aligned}$$

¹⁴⁸Recall in Remark 9.1 that we can write the definition of the total derivative using a remainder terms which tends to zero. We did something similar here, although in this case our remainder terms are scalar functions because we define it as the norm of a small distance.

If we take the norm of both sides then,

$$\begin{aligned}
\|\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^k} &= \|(\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}) + \phi'(\mathbf{f}(\mathbf{x}))[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^k} \\
&\leq \|\phi(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \phi(\mathbf{f}(\mathbf{x})) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{k}\|_{\mathbb{R}^k} + \|\phi'(\mathbf{f}(\mathbf{x}))[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^k} \quad (\text{Triangle Inequality}) \\
&= q(\mathbf{k}) \|\mathbf{k}\|_{\mathbb{R}^k} + \|\phi'(\mathbf{f}(\mathbf{x}))[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^k} \quad (\text{Equation (43)}) \\
&\leq q(\mathbf{k}) \|\mathbf{k}\|_{\mathbb{R}^k} + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot \|[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^m} \quad (\text{Lemma 8.1}) \\
&= q(\mathbf{k}) \|\mathbf{k}\|_{\mathbb{R}^k} + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^m} \quad (\text{Equation (42)}) \\
&\leq q(\mathbf{k}) \left(\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + r(\mathbf{h}) \right) \|\mathbf{h}\|_{\mathbb{R}^n} + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot r(\mathbf{h}) \|\mathbf{h}\|_{\mathbb{R}^m} \quad (\text{Equation (44)}) \\
&= \left(q(\mathbf{k}) \left(\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + r(\mathbf{h}) \right) + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot r(\mathbf{h}) \right) \|\mathbf{h}\|_{\mathbb{R}^m}
\end{aligned}$$

If we divide this final inequality by $\|\mathbf{h}\|_{\mathbb{R}^n}$ and let $\mathbf{h} \rightarrow \mathbf{0}$ we have:

$$\begin{aligned}
\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^k}}{\|\mathbf{h}\|_{\mathbb{R}^n}} &\leq \lim_{\mathbf{h} \rightarrow \mathbf{0}} \left(q(\mathbf{k}) \left(\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + r(\mathbf{h}) \right) + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot r(\mathbf{h}) \right) \\
&= \lim_{\mathbf{h} \rightarrow \mathbf{0}} q(\mathbf{k}) \|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + \lim_{\mathbf{h} \rightarrow \mathbf{0}} q(\mathbf{k})r(\mathbf{h}) + \lim_{\mathbf{h} \rightarrow \mathbf{0}} \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot r(\mathbf{h}) \\
&= \lim_{\mathbf{h} \rightarrow \mathbf{0}} q(\mathbf{k}) \|\mathbf{f}'(\mathbf{x})\|_{\text{op}} + q(\mathbf{k}) \cdot 0 + \|\phi'(\mathbf{f}(\mathbf{x}))\|_{\text{op}} \cdot 0 \quad (r(\mathbf{h}) \rightarrow 0) \\
&= \lim_{\mathbf{h} \rightarrow \mathbf{0}} q(\mathbf{k}) \|\mathbf{f}'(\mathbf{x})\|_{\text{op}} \\
&= 0 \cdot \|\mathbf{f}'(\mathbf{x})\|_{\text{op}} \quad (\lim_{\mathbf{h} \rightarrow \mathbf{0}} \mathbf{k} = \mathbf{0} \text{ and } q(\mathbf{k}) \rightarrow 0) \\
&= 0
\end{aligned}$$

Therefore

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - \phi'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^k}}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0$$

so $\mathbf{F} = \phi \circ \mathbf{f}$ is differentiable at \mathbf{x} , and the $k \times n$ matrix $g'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x})$ is the total derivative at \mathbf{x} . \square

Example 9.9 (“Multivariable” Chain Rule). Suppose $\mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}^2$ and $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, where \mathbf{f} is differentiable at t and g at $\mathbf{f}(t)$. Define $h = g \circ \mathbf{f}$. The chain rule gives the existence of $h'(x)$, and the equality

$$h'(x) = \begin{bmatrix} \frac{\partial g}{\partial f_1}(\mathbf{f}(x)) & \frac{\partial g}{\partial f_2}(\mathbf{f}(x)) \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x}(x) \\ \frac{\partial f_2}{\partial x}(x) \end{bmatrix} = \frac{\partial g}{\partial f_1}(\mathbf{f}(x)) \cdot \frac{\partial f_1}{\partial x}(x) + \frac{\partial g}{\partial f_2}(\mathbf{f}(x)) \cdot \frac{\partial f_2}{\partial x}(x).$$

This is likely the version of the chain rule you saw in multivariable calculus. In general if $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as $h = g \circ f$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we have

$$\frac{\partial h}{\partial x_i} = \sum_{j=1}^m \frac{\partial g}{\partial f_j} \frac{\partial f_j}{\partial x_i}.$$

Example 9.10. Suppose $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ where $f(\mathbf{0}) = (1, 2)$ and

$$\mathbf{f}'(\mathbf{0}) = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix}.$$

If $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is defined as $\phi(x_1, x_2) = (x_1 + 2x_2 + 1, 3x_1x_2)$, then ϕ is differentiable at $f(\mathbf{0})$ by Theorem 9.3. We have

$$\phi'(\mathbf{x}) = \begin{bmatrix} 1 & 3x_1 \\ 2 & 3x_2 \end{bmatrix}.$$

The chain rule gives

$$(\phi \circ \mathbf{f})'(\mathbf{0}) = \phi'(\mathbf{f}(\mathbf{0}))\mathbf{f}'(\mathbf{0}) = \begin{bmatrix} 1 & 3(1) \\ 2 & 3(2) \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 6 \\ 2 & 4 & 12 \end{bmatrix}$$

9.4 Mean Value Theorems

In Section 5 we proved one of the hallmark results from calculus – the mean value theorem (Corollary 5.2). Since then, we've used the result to prove the second part of the fundamental theorem of calculus (Theorem 6.6), along with several other results. It may be wise then to extend the mean value theorem so we can use it in more general settings, but is this possible? What would the mean value theorem look like for $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$?

The mean value theorem concerns three points in the domain of f : a , b , and c . Crucially $c \in (a, b)$. The first challenge of generalizing the mean value theorem is determining what it means for a point \mathbf{c} to be “between” \mathbf{a} and \mathbf{c} in \mathbb{R}^n . An example may make this problem a bit more concrete.

Example 9.11. Suppose $f : S \rightarrow \mathbb{R}^m$ where $S \subset \mathbb{R}^2$ is shown in Figure 100. What does it mean for some

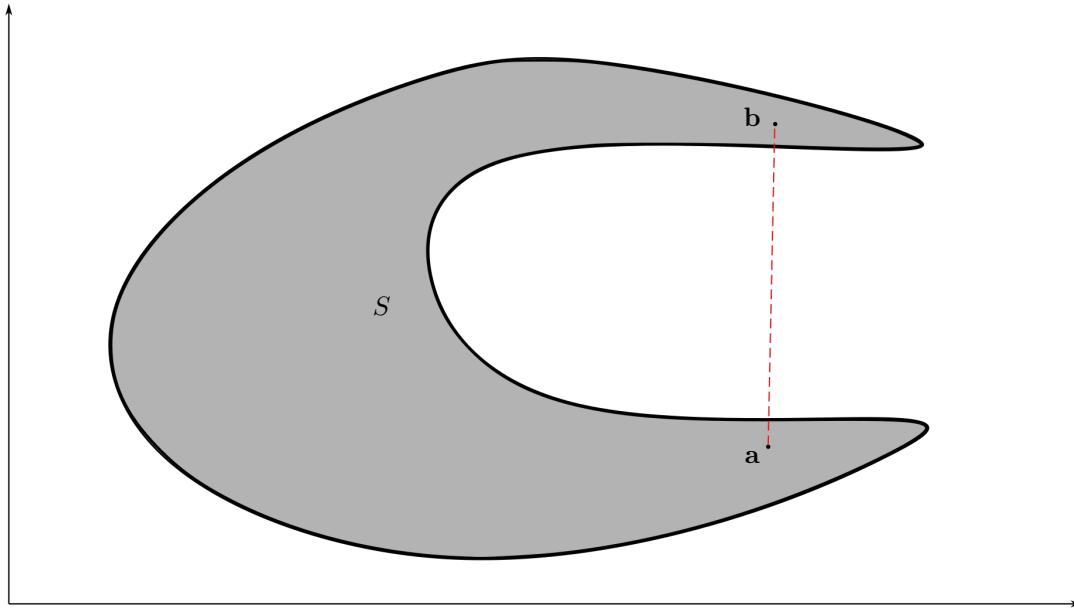


Figure 100: The domain of the set f .

point \mathbf{c} to be “between” \mathbf{a} and \mathbf{b} ? One reasonable definition which the figure hints at, is a point \mathbf{c} would lie on some line between \mathbf{a} and \mathbf{b} . This raises a bit of an issue in the context of the set S though, as there are points that lie on the line between \mathbf{a} and \mathbf{b} , but do not belong to S . We will want to rule out these types of sets.

Definition 9.4. A set $S \subset \mathbb{R}^n$ is *convex* if for all $\mathbf{a}, \mathbf{b} \in S$ the line segment connecting \mathbf{a} and \mathbf{b} is entirely contained in S . That is

$$\{(1 - t)\mathbf{a} + t\mathbf{b} \mid t \in [0, 1]\} \subset S.$$

A point in the set $\{(1 - t)\mathbf{a} + t\mathbf{b} \mid t \in [0, 1]\}$ is sometimes called a *convex combination of \mathbf{a} and \mathbf{b}* .

Example 9.12. The real line \mathbb{R} is trivially a convex set, which is why we do not explicitly address convexity when working with the mean value theorem for $f : \mathbb{R} \rightarrow \mathbb{R}$.

As the name implies, a convex set takes the form of a convex shape when illustrated.

Remark 9.3 (Convex Analysis). Convex analysis is a branch of mathematics concerned with the study of convex functions and convex sets, and plays an integral role in the theory of optimization and economic theory.

If we want to generalize the mean value theorem to higher dimensions, we should focus on functions defined on convex sets, otherwise it may be the case that the function is undefined at the \mathbf{c} in between \mathbf{b} and \mathbf{a} of interest.

Before jumping to the general case of $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, let's consider the intermediary case of $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Proposition 9.1 (Mean Value Theorem in Several Variables). Let $f : E \rightarrow \mathbb{R}$ where $E \subset \mathbb{R}^n$ is a convex set. If f is differentiable on the set of convex combinations of \mathbf{a} and \mathbf{b} , then there exists a convex combination of \mathbf{a} and \mathbf{b} , call it \mathbf{c} , such that

$$f(\mathbf{b}) - f(\mathbf{a}) = f'(\mathbf{c})(\mathbf{b} - \mathbf{a}).$$

Proof. Define the function $\mathbf{x} : [0, 1] \rightarrow E$ as

$$\mathbf{x}(t) = (1-t)\mathbf{a} + t\mathbf{b}.$$

This function corresponds to the line segment connecting \mathbf{a} and \mathbf{b} . Using \mathbf{x} define the function $\phi : [0, 1] \rightarrow \mathbb{R}$ as the composition $\phi(t) = (\mathbf{x}(t))$. For all intents and purposes, $\phi(t)$ is f reparameterized to be a function of a single variable on the restricted domain that is the line connecting \mathbf{a} and \mathbf{b} .¹⁴⁹ We have

$$\begin{aligned}\phi(0) &= f(\mathbf{a}), \\ \phi(1) &= f(\mathbf{b}).\end{aligned}$$

We also have assumed that f is differentiable, so we can apply the chain rule to get

$$\phi'(t) = f'(\mathbf{x}(t))\mathbf{x}'(t) = f'(\mathbf{x}(t)) \cdot \frac{d}{dt}[(1-t)\mathbf{a} + t\mathbf{b}] = f'(\mathbf{x}(t))(\mathbf{b} - \mathbf{a}).$$

By the mean value theorem (Corollary 5.2), there exists some $t^* \in (0, 1)$ such that

$$\begin{aligned}\phi(1) - \phi(0) &= \phi(t^*) \\ \implies f(\mathbf{b}) - f(\mathbf{a}) &= f'(\mathbf{x}(t^*))(\mathbf{b} - \mathbf{a}).\end{aligned}$$

Therefore, we have our \mathbf{c} in the form of $\mathbf{x}(t^*)$!, and we know that $\mathbf{c} \in E$ because E is convex. □

Example 9.13. Define $f(x_1, x_2) = x_2^2 x_1$ on all of \mathbb{R}^2 . We have

$$f'(\mathbf{x}) = \begin{bmatrix} 2x_2 x_1 \\ x_2^2 \end{bmatrix}.$$

For $\mathbf{a} = (0, 0)$ and $\mathbf{b} = (5, 5)$, we have

$$f(5, 5) - f(0, 0) = f'\left(5\sqrt{3}/3, 5\sqrt{3}/3\right) \begin{bmatrix} 5 \\ 5 \end{bmatrix},$$

where $(5\sqrt{3}/3, 5\sqrt{3}/3)$ is on the line with endpoints \mathbf{b} and \mathbf{a} . If we were to modify the domain of f to make it $\mathbb{R}^2 \setminus \{(5\sqrt{3}/3, 5\sqrt{3}/3)\}$, Proposition 9.1 fails as the domain is no longer convex.

¹⁴⁹But restricting the domain like this doesn't matter, we're only concerned with this line segment anyway.

Now consider $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Do we have

$$\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a}) \stackrel{?}{=} \mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a}).$$

Example 9.14. Define $\mathbf{f} : [0, 2\pi] \rightarrow \mathbb{R}^2$ as the familiar $\mathbf{f}(t) = (\cos t, \sin t)$. If we let $a = 0$ and $b = 2\pi$, is there some $c \in (0, 2\pi)$ such that

$$\begin{aligned} \mathbf{f}(b) - \mathbf{f}(a) &\stackrel{?}{=} \mathbf{f}'(c)(b - a) \\ \implies \mathbf{f}(2\pi) - \mathbf{f}(0) &\stackrel{?}{=} \begin{bmatrix} -\sin c & \cos c \end{bmatrix} (2\pi - 0) \\ \implies \begin{bmatrix} 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \end{bmatrix} &\stackrel{?}{=} \begin{bmatrix} -\sin c & \cos c \end{bmatrix} (2\pi - 0) \\ \implies \begin{bmatrix} 0 & 0 \end{bmatrix} &\stackrel{?}{=} \begin{bmatrix} -\sin c & \cos c \end{bmatrix}. \end{aligned}$$

There exists no such c satisfying this equality, as $-\sin c$ and $\sin c$ will never be 0 at the same time.

Instead of a full fledged mean value theorem for $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we instead have a quasi-mean value theorem in the form of an inequality

Theorem 9.5. Suppose $\mathbf{f} : E \rightarrow \mathbb{R}^m$, where $E \subset \mathbb{R}^n$ is a convex set, is differentiable on E . If there exists a real number M such that

$$\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} \leq M$$

for all $\mathbf{x} \in E$, then

$$\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_{\mathbb{R}^m} \leq M \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n}$$

for all $\mathbf{a}, \mathbf{b} \in E$.

Proof. Fix $\mathbf{a}, \mathbf{b} \in E$, and define $\mathbf{x} : [0, 1] \rightarrow E$ as

$$\mathbf{x}(t) = (1-t)\mathbf{a} + t\mathbf{b}.$$

Then set E is convex, so $\mathbf{x}(t) \in E$ for all $t \in [0, 1]$. Define $\phi(t) : [0, 1] \rightarrow \mathbb{R}^m$ as $\phi(t) = \mathbf{f}(\mathbf{x}(t))$. If we differentiate this we have

$$\begin{aligned} \phi'(t) &= \mathbf{f}'(\mathbf{x}(t))\mathbf{x}'(t) && \text{(chain rule)} \\ \implies \phi'(t) &= \mathbf{f}'(\mathbf{x}(t))(\mathbf{b} - \mathbf{a}) && \text{(differentiate } \mathbf{x}(t)\text{)} \\ \implies \|\phi'(t)\|_{\mathbb{R}^m} &= \|\mathbf{f}'(\mathbf{x}(t))(\mathbf{b} - \mathbf{a})\|_{\mathbb{R}^m} \\ \implies \|\phi'(t)\|_{\mathbb{R}^m} &\leq \|\mathbf{f}'(\mathbf{x}(t))\|_{\text{op}} \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n} && \text{(Lemma 8.1)} \\ \implies \|\phi'(t)\|_{\mathbb{R}^m} &\leq M \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n} && \left(\|\mathbf{f}'(\mathbf{x})\|_{\text{op}} \leq M \right) \end{aligned}$$

for all $t \in [0, 1]$. By Proposition 9.1, there exists a $t^* \in (0, 1)$ such that

$$\begin{aligned} \phi(1) - \phi(0) &= \phi'(t^*)(1 - 0) \\ \implies \|\phi(1) - \phi(0)\|_{\mathbb{R}^m} &= \|\phi'(t^*)\|_{\mathbb{R}^m} \\ \implies \|\phi(1) - \phi(0)\|_{\mathbb{R}^m} &\leq M \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n} \\ \implies \|\mathbf{f}(\mathbf{x}(1)) - \mathbf{f}(\mathbf{x}(0))\|_{\mathbb{R}^m} &\leq M \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n} && \text{(definition of } \phi\text{)} \\ \implies \|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_{\mathbb{R}^m} &\leq M \|\mathbf{b} - \mathbf{a}\|_{\mathbb{R}^n} && \text{(definition of } \mathbf{x}\text{)} \end{aligned}$$

□

Example 9.15. Reconsider Example 9.14. For $\mathbf{f}'(t) = (-\sin t, \cos t)$,

$$\|\mathbf{f}'(t)\|_{\text{op}} = \sup \left\| \begin{bmatrix} -\sin t \\ \cos t \end{bmatrix} \mid t = 1 \right\|_{\mathbb{R}^2} = (\sin^2(1) + \cos^2(1))^{1/2} = 1 \leq 1.$$

Theorem 9.5 tells us that

$$\begin{aligned} \|\mathbf{f}(b) - \mathbf{f}(a)\|_{\mathbb{R}^m} &\leq 1 \cdot |b - a|, \\ \implies \|\mathbf{f}(b) - \mathbf{f}(a)\|_{\mathbb{R}^m} &\leq |b - a|. \end{aligned}$$

This inequality can be interpreted geometrically – the length of the cord between $\mathbf{f}(b)$ and $\mathbf{f}(a)$ on the unit circle is bounded by $|b - a|$.

9.5 Clairaut's Theorem

Much like differentiation for $f : \mathbb{R} \rightarrow \mathbb{R}$, we can take partial derivatives of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ several times over. This is novel compared to the prior case, as we may want to differentiate f with respect to one variable, and then with respect to another variable. For example we may be interested in

$$\frac{\partial f}{\partial x_1 \partial x_2}(\mathbf{x}) = \frac{\partial}{\partial x_1} \left(\frac{\partial f}{\partial x_2}(\mathbf{x}) \right).$$

A geometric interpretation of this partial derivative is given in Figure 101.

One of the big takeaways from analysis is that order *really matters* when operations involve limits. Is partial differentiation twice with respect to different variables any different? That is to say

$$\frac{\partial f}{\partial x_1 \partial x_2}(\mathbf{x}) \stackrel{?}{=} \frac{\partial f}{\partial x_2 \partial x_1}(\mathbf{x}).$$

If we compare Figure 101 and Figure 102, the equality seems holds. This equality is known as Clairaut's Theorem, and is one of the highlights of multivariable calculus. When you were first introduced to it, you likely were not too concerned with the conditions under which it held, as the goal at the time was getting comfortable with computation of partial derivatives. Now we need to be a bit more careful, as this equality need not hold in general.

Theorem 9.6 (Clairaut's Theorem). Suppose $f : E \rightarrow \mathbb{R}$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If f is twice continuously differentiable at \mathbf{x} ,¹⁵⁰ then

$$\frac{\partial f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{x})$$

for all (i, j) .

Proof. Assume $i \neq j$, otherwise we have nothing to prove. We will show the result for $\mathbf{x} = \mathbf{0}$. We have assumed f is twice continuously differentiable at $\mathbf{0}$, so we can find a $\delta_1 > 0$ such that

$$\left| \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{x}) - \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right| < \frac{\varepsilon}{2} \tag{45}$$

for all $\varepsilon > 0$ when $\|\mathbf{x} - \mathbf{0}\|_{\mathbb{R}^n} = \|\mathbf{x}\|_{\mathbb{R}^n} < \delta$, and a $\delta_2 > 0$ such that

$$\left| \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{x}) - \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \frac{\varepsilon}{2} \tag{46}$$

¹⁵⁰Meaning that $\frac{\partial f}{\partial x_j \partial x_i}(\mathbf{x})$ exists and is continuous for all (i, j) .

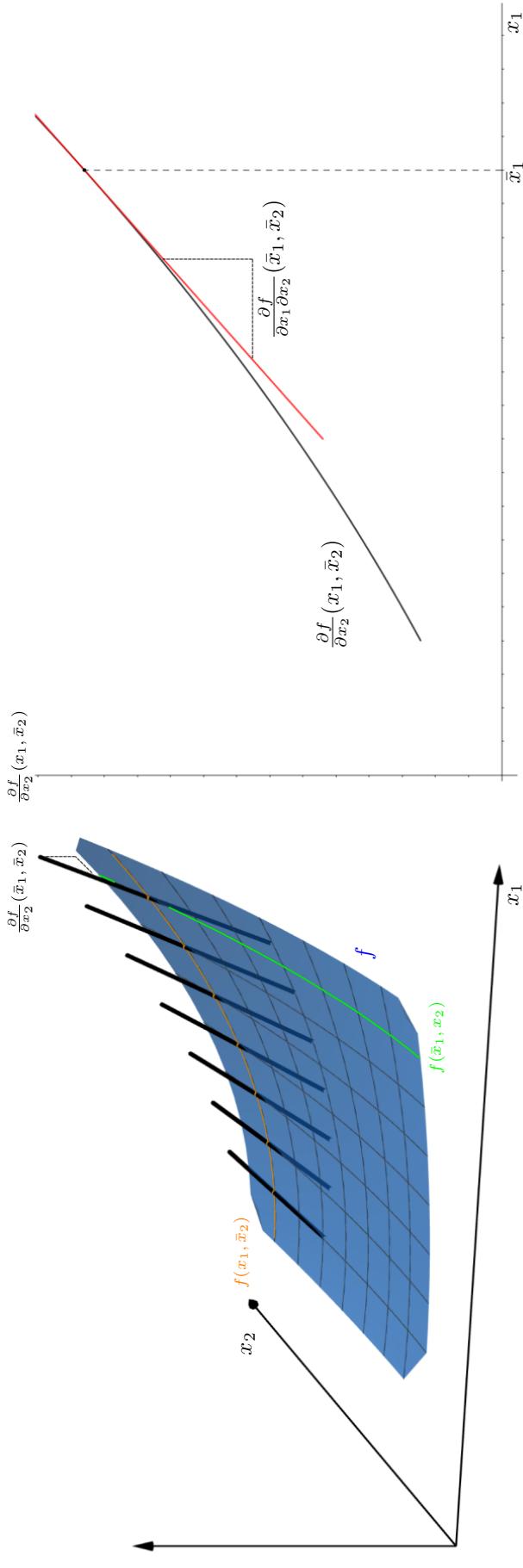


Figure 101: The derivative $\frac{\partial f}{\partial x_2}(\bar{x}_1, \bar{x}_2)$ corresponds to the slope of the line tangent to f at the point (\bar{x}_1, \bar{x}_2) and pointing in the direction of \mathbf{e}_2 . The slope of these lines changes as we move the point of tangency by varying x_1 . We can graph the slope of this tangent line for values of x_1 , giving the graph on the right. If we draw a line tangent to this new function at the point \bar{x}_1 , then the line's slope is given as $\frac{\partial}{\partial x_1} \left(\frac{\partial f}{\partial x_2}(\bar{x}_1, \bar{x}_2) \right)$

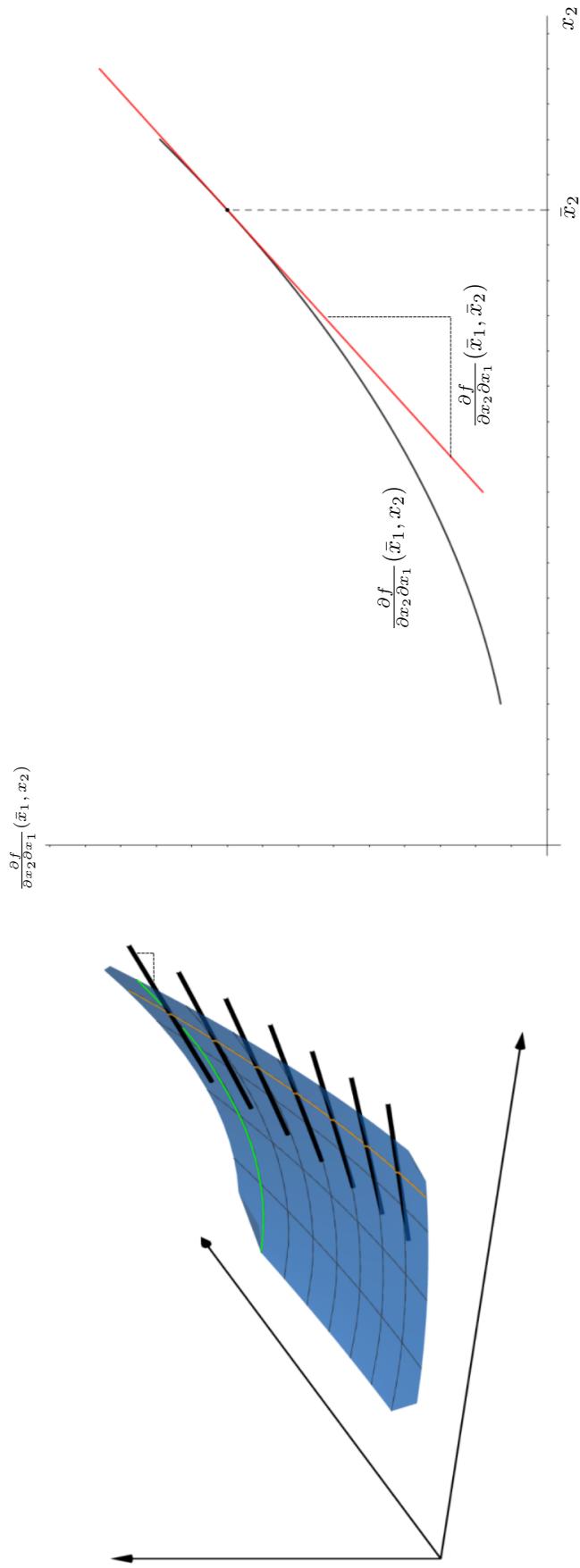


Figure 102: The same exact illustration as Figure 101, except for the partial derivative $\frac{\partial f}{\partial x_2 \partial x_1}(\mathbf{x})$

for all $\varepsilon > 0$ when $\|\mathbf{x}\|_{\mathbb{R}^n} = \|\mathbf{x}\|_{\mathbb{R}^n} < \delta_2$. If we take $\delta = \min\{\delta_1, \delta_2\}/2$, then these inequalities will both hold for all $\varepsilon > 0$ when $\|\mathbf{x}\|_{\mathbb{R}^n} < 2\delta$.

The bulk of the proof takes place in two symmetric steps:¹⁵¹

Step 1: Define the single variable function $\psi : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\psi(t) = \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + t \mathbf{e}_j).$$

ψ is a function in the \mathbf{e}_j variable. We defined ψ such that:

$$\begin{aligned} \psi'(t) &= \lim_{h \rightarrow 0} \frac{\psi(t+h) - \psi(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + (t+h) \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + t \mathbf{e}_j)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial x_i}((x_i \mathbf{e}_i + t \mathbf{e}_j) + h \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + t \mathbf{e}_j)}{h} \\ &= \frac{\partial}{\partial x_j} \left(\frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + t \mathbf{e}_j) \right) \\ &= \frac{\partial f}{\partial x_j \partial x_i}(x_i \mathbf{e}_i + t \mathbf{e}_j) \end{aligned}$$

We know that $\phi'(t)$ exists because f is twice continuously differentiable. We can apply the MVT (Corollary 5.1) to this ϕ . There exists a $x_j \in (0, \delta)$ such that

$$\begin{aligned} \frac{\psi(\delta) - \psi(0)}{\delta - 0} &= \psi'(x_j) \\ \implies \psi(\delta) - \psi(0) &= \delta \psi'(x_j) \\ \implies \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) &= \delta \frac{\partial f}{\partial x_j \partial x_i}(x_i \mathbf{e}_i + x_j \mathbf{e}_j) && \text{(definition of } \psi \text{ and } \psi') \\ \implies \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) &= \delta \frac{\partial f}{\partial x_j \partial x_i}(x_i \mathbf{e}_i + x_j \mathbf{e}_j) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \\ \implies \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) &= \delta \left(\frac{\partial f}{\partial x_j \partial x_i}(x_i \mathbf{e}_i + x_j \mathbf{e}_j) - \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right) && (47) \end{aligned}$$

Now recall that (46) holds for $\|\mathbf{x}\|_{\mathbb{R}^n} < 2\delta$. In the event that $x_i < \delta$, we have

$$\|x_i \mathbf{e}_i + x_j \mathbf{e}_j\| \leq |x_j| + |x_i| < 2\delta,$$

as the MVT gave $x_j \in (0, \delta)$. We can therefore combine (46) and (47) and conclude

$$\left| \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \frac{\varepsilon \delta}{2} \quad (x_i < \delta)$$

This inequality will hold for all $x_i < \delta$, so the inequality is preserved when integrating over the interval

¹⁵¹Which is quite apropos considering the symmetry being proved

$[0, \delta]$ with respect to x_i .

$$\begin{aligned}
& \int_0^\delta \left| \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| dx_i < \int_0^\delta \frac{\varepsilon \delta}{2} dx_i && \text{(Theorem 6.2)} \\
& \Rightarrow \left| \int_0^\delta \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) - \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) - \delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) dx_i \right| < \int_0^\delta \frac{\varepsilon \delta}{2} dx_i && \text{(Proposition 6.6)} \\
& \Rightarrow \left| \int_0^\delta \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i + \delta \mathbf{e}_j) dx_i - \int_0^\delta \frac{\partial f}{\partial x_i}(x_i \mathbf{e}_i) dx_i - \delta \int_0^\delta \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) dx_i \right| < \frac{\varepsilon \delta^2}{2} && \text{(Theorem 6.2)} \\
& \Rightarrow \left| [f(x_i \mathbf{e}_i + \delta \mathbf{e}_j)]_0^\delta - [f(x_i \mathbf{e}_i)]_0^\delta - \delta^2 \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \frac{\varepsilon \delta^2}{2} && \text{(Fund. Thm. Calc)}
\end{aligned}$$

Our final inequality is

$$\left| f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) + f(0) - \delta^2 \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \frac{\varepsilon \delta^2}{2} \quad (48)$$

Step 2: Define the single variable function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\sigma(t) = \frac{\partial f}{\partial x_j}(t \mathbf{e}_i + x_j \mathbf{e}_j).$$

σ is a function in the \mathbf{e}_i variable for which

$$\sigma'(t) = \frac{\partial f}{\partial x_i \partial x_j}(t \mathbf{e}_i + x_j \mathbf{e}_j).$$

At this point, we can emulate step 1 by applying the mean value theorem to σ , and integrating the resulting inequality with respect to x_j to get

$$\left| f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_j) + f(0) - \delta^2 \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right| < \frac{\varepsilon \delta^2}{2} \quad (49)$$

Finally, we add inequalities (48) and (49) from part 1 and part 2, respectively.

$$\begin{aligned}
& \left| f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_j) + f(0) - \delta^2 \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right| + \left| f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_j) + f(0) - \delta^2 \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \frac{\varepsilon \delta^2}{2} + \frac{\varepsilon \delta^2}{2} \\
& \Rightarrow \left| f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_j) + f(0) - \delta^2 \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) - \left(f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_j) + f(0) - \delta^2 \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right) \right| < \varepsilon \delta^2 && \text{(Tri. Ineq.)} \\
& \Rightarrow \left| \underbrace{[f(\delta \mathbf{e}_i + \delta \mathbf{e}_j) - f(\delta \mathbf{e}_i + \delta \mathbf{e}_j)]_0^\delta}_{0} - \underbrace{[f(\delta \mathbf{e}_i) - f(\delta \mathbf{e}_i)]_0^\delta} - \underbrace{[f(\delta \mathbf{e}_j) - f(\delta \mathbf{e}_j)]_0^\delta} + \underbrace{[f(0) - f(0)]_0^\delta}_{0} - \delta^2 \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) + \delta^2 \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) \right| < \varepsilon \delta^2 \\
& \Rightarrow \left| \delta^2 \left(\frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) - \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right) \right| < \varepsilon \delta^2 \\
& \Rightarrow \delta^2 \left| \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) - \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right| < \varepsilon \delta^2 \\
& \Rightarrow \left| \frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) - \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}) \right| < \varepsilon
\end{aligned} \quad (\delta > 0)$$

This holds for all $\varepsilon > 0$, so it must be the case that

$$\frac{\partial f}{\partial x_j \partial x_i}(\mathbf{0}) = \frac{\partial f}{\partial x_i \partial x_j}(\mathbf{0}).$$

□

Corollary 9.1. Suppose $\mathbf{f} : E \rightarrow \mathbb{R}$ where $E \subset \mathbb{R}^n$, and \mathbf{x} is an interior point of E . If f is twice continuously differentiable at \mathbf{x} , then

$$\frac{\partial \mathbf{f}}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial x_j \partial x_i}(\mathbf{x})$$

for all (i, j) .

Proof. Clairaut's Theorem holds for all the components of $\frac{\partial \mathbf{f}}{\partial x_i \partial x_j}(\mathbf{x})$ and $\frac{\partial \mathbf{f}}{\partial x_i \partial x_j}(\mathbf{x})$. \square

Example 9.16. The assumption that f is twice continuously differentiable is essential. Define

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq \mathbf{0} \\ 0 & \text{if } (x_1, x_2) = \mathbf{0} \end{cases}.$$

Computations show that

$$\frac{\partial f}{\partial x_1 \partial x_2}(\mathbf{0}) = 1 \neq -1 = \frac{\partial f}{\partial x_2 \partial x_1}(\mathbf{0}).$$

9.6 The Inverse Function Theorem

Proposition 5.1 specified the conditions under which the differentiability of f implies the differentiability of f^{-1} , and in the case that f^{-1} is differentiable then

$$(f^{-1})'(f(x)) = \frac{1}{f'(x)}.$$

We now take up generalizing this result for $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$.

First, suppose that $\mathbf{f}^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ exists. If we know that \mathbf{f} is invertible, then calculating the $(\mathbf{f}^{-1})'(\mathbf{y})$ for some $\mathbf{y} \in \mathbb{R}^m$, where $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ for $\mathbf{x} \in \mathbb{R}^n$, is an application of the chain rule. For an $n \times n$ identity matrix I ,

$$\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) = I\mathbf{x}.$$

Differentiating both sides gives

$$(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x}) = I,$$

so $(\mathbf{f}^{-1})'(\mathbf{y}) = [\mathbf{f}'(\mathbf{x})]^{-1}$. In other words, the differentiability of $\mathbf{f}^{-1}(\mathbf{y})$ implies that $\mathbf{f}'(\mathbf{x})$ is invertible. But is the converse true? Is $\mathbf{f}'(\mathbf{x})$ being invertible a sufficient condition for $\mathbf{f}^{-1}(\mathbf{y})$ to exist and be differentiable? The answer is yes (somewhat), and this will be established by the inverse function theorem. Before we state and prove the result, we need to do some prep work, as the proof contains several steps.¹⁵²

The first lemma that we will use to prove the inverse function theorem is really a theorem unto itself, and concerns a special type of map on a metric space.

Definition 9.5. Let (X, d) be a metric space where $f : X \rightarrow X$. The function f is a *contraction* if we have

$$d(f(x), f(y)) \leq \kappa \cdot d(x, y)$$

for a *contraction constant* $\kappa \in [0, 1)$ and all $x, y \in X$.

¹⁵²During my second semester course in undergraduate real analysis (which covered the material in Tao (2016b)), the proof of the inverse function theorem was by far the most complicated and “difficult” proof.

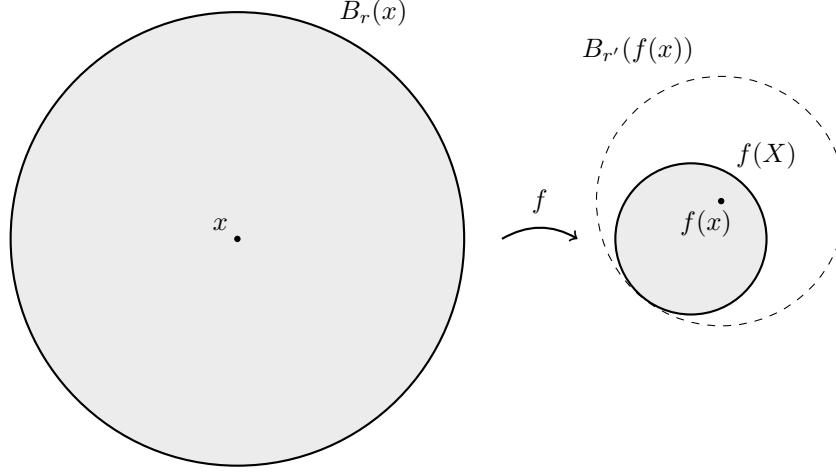


Figure 103: The function f is a contraction. For all $x \in X$ all points in the ball $B_r(x)$ are mapped into $B_{r'}(f(x))$, i.e. $B_{r'}(f(x)) \subset f(X)$, where $r' < r$. This figure illustrates this for one particular $x \in X$.

Example 9.17. Consider the \mathbb{R} with the standard metric. The function

$$f(x) = \frac{x}{a} + b$$

is a strict contraction for $a > 1$.

$$d(f(x), f(y)) = \left| \frac{x}{a} + b - \left(\frac{y}{a} + b \right) \right| = \left| \frac{x-y}{a} \right| = \frac{1}{|a|} d(x, y) < d(x, y).$$

Example 9.18 (Contractions and Continuity). Any contraction $f : X \rightarrow X$ is continuous. For all $\varepsilon > 0$, define $\delta = \varepsilon/\kappa$. Whenever $|x - y| < \delta = \varepsilon/\kappa$, we have

$$d(f(x), f(y)) = \kappa \cdot d(x, y) < \kappa \cdot (\varepsilon/\kappa) = \varepsilon.$$

Furthermore, δ is independent of ε , so contractions are uniformly continuous.

Example 9.19. Consider the open unit ball $B_1(\mathbf{0}) \in \mathbb{R}^2$. Define $f : B_1(\mathbf{0}) \rightarrow B_1(\mathbf{0})$ as

$$\mathbf{f}(\mathbf{x}) = d(\mathbf{x}, \mathbf{0}) \cdot \mathbf{x} = \|\mathbf{x}\|_{\mathbb{R}^2} \cdot \mathbf{x}.$$

This function is a contraction. When applied to a point $\mathbf{x} \in B_1(\mathbf{0})$, \mathbf{x} is “pulled to” the origin by a factor equal to its distance from the origin. This factor is always less than 1 as $d(\mathbf{x}, \mathbf{0}) < 1$ by the definition $B_1(\mathbf{0})$. Without loss of generality, assume $\|\mathbf{x}\|_{\mathbb{R}^2} < \|\mathbf{y}\|_{\mathbb{R}^2}$.

$$\begin{aligned} (x_1 - y_1) + (x_2 - y_2) &\geq (\|\mathbf{x}\|_{\mathbb{R}^2} x_1 - \|\mathbf{y}\|_{\mathbb{R}^2} y_1) + (\|\mathbf{x}\|_{\mathbb{R}^2} x_2 - \|\mathbf{y}\|_{\mathbb{R}^2} y_2) \\ \implies (x_1 - y_1)^2 + (x_2 - y_2)^2 &\geq (\|\mathbf{x}\|_{\mathbb{R}^2} x_1 - \|\mathbf{y}\|_{\mathbb{R}^2} y_1)^2 + (\|\mathbf{x}\|_{\mathbb{R}^2} x_2 - \|\mathbf{y}\|_{\mathbb{R}^2} y_2)^2 \\ \implies [(x_1 - y_1)^2 + (x_2 - y_2)^2]^{1/2} &\geq [(\|\mathbf{x}\|_{\mathbb{R}^2} x_1 - \|\mathbf{y}\|_{\mathbb{R}^2} y_1)^2 + (\|\mathbf{x}\|_{\mathbb{R}^2} x_2 - \|\mathbf{y}\|_{\mathbb{R}^2} y_2)^2]^{1/2} \\ \implies d(\mathbf{x}, \mathbf{y}) &\geq d(\|\mathbf{x}\|_{\mathbb{R}^2} \mathbf{x}, \|\mathbf{y}\|_{\mathbb{R}^2} \mathbf{y}) \\ \implies d(\mathbf{x}, \mathbf{y}) &\geq d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})) \end{aligned}$$

If we apply \mathbf{f} successively to a point \mathbf{x} , we have

$$f^n(\mathbf{x}) = f(f(\cdots(f(\mathbf{x})))) = \|\mathbf{x}\|_{\mathbb{R}^2}^n \mathbf{x} \rightarrow \mathbf{0}.$$

Interestingly enough we have $\mathbf{f}(\mathbf{0}) = \mathbf{0}$. This type of point turns out to be quite special.

Definition 9.6. Define the function $f : X \rightarrow X$. The point $x \in X$ is called a *fixed point of f* if $f(x) = x$.

As Example 9.18 hints at, there is a special relationship between contractions and fixed points. Intuitively, we'd think that as we apply a contraction successively, all the points in X get closer and closer to some common fixed point. This should ring a bell if you were ever board in class in elementary school and This is formalized in an exceptionally important and useful theorem, which is only classified as a lemma as our primary use for it will be proving the inverse function theorem.

Lemma 9.1 (Banach Fixed Point Theorem/Contraction Mapping Theorem). Let (X, d) be a non-empty complete metric space and $f : X \rightarrow X$ be a contraction. The function f has a unique fixed point x^* .

Proof. The proof is constructive in the sense that we will find an explicit formula for $x^* \in X$. The only real piece of information we have is that X is complete, so the result must hinge on a Cauchy sequence converging. But what does this result have to do with sequences? Recall in Example 9.18, we formed a numerical sequence which converged to a fixed point by successively applying a contraction. Let's generalize this approach

Define the sequence $x_n = f(x_{n-1})$. The function f is a contraction, so we have

$$\begin{aligned} d(x_2, x_1) &= d(f(x_1), f(x_0)) \geq \kappa \cdot d(x_1, x_0) \\ \implies d(x_3, x_2) &= d(f(x_2), f(x_1)) \geq \kappa \cdot d(x_2, x_1) \geq \kappa(\kappa \cdot d(x_1, x_0)) \\ \implies d(x_4, x_3) &= d(f(x_3), f(x_2)) \geq \kappa \cdot d(x_3, x_2) \geq \kappa(\kappa \cdot d(x_2, x_1)) \geq \kappa(\kappa(\kappa \cdot d(x_1, x_0))) \\ &\vdots \\ \implies d(x_{n+1}, x_n) &\geq \kappa^n d(x_1, x_0). \end{aligned} \tag{50}$$

We want to show $\{x_n\}$ converges to some point. Since X is complete, it suffices to show that $\{x_n\}$ is Cauchy. Let $m, n \in \mathbb{N}$ where $m > n$.

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_n) && \text{(Triangle Inequality)} \\ &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + d(x_{m-2}, x_n) && \text{(Triangle Inequality)} \\ &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + d(x_{m-2}, x_{m-3}) + d(x_{m-3}, x_n) && \text{(Triangle Inequality)} \\ &\vdots \\ &\leq d(x_{m-1}, x_{m-2}) + d(x_{m-2}, x_{m-3}) + \cdots + d(x_{n+1}, x_n) \\ &\leq \kappa^{m-1} d(x_1, x_0) + \kappa^{m-2} d(x_1, x_0) + \cdots + \kappa^n d(x_1, x_0) && \text{(Equation (50))} \\ &= d(x_1, x_0) \cdot \sum_{j=n}^{m-1} \kappa^j \\ &= d(x_1, x_0) \cdot \sum_{j=0}^{m-n-1} \kappa^n \cdot \kappa^j \\ &= \kappa^n d(x_1, x_0) \cdot \sum_{j=0}^{m-n-1} \kappa^j \\ &\leq \kappa^n d(x_1, x_0) \cdot \sum_{j=0}^{\infty} \kappa^j && (0 \leq \kappa < 1) \\ &= \kappa^n d(x_1, x_0) \left(\frac{1}{1-\kappa} \right) \end{aligned}$$

For all $\varepsilon > 0$, define $N = \log_{\kappa} \left(\frac{\varepsilon(1-\kappa)}{d(x_1, x_0)} \right)$. For all $m, n > N$, we have

$$d(x_n, x_m) < \kappa^N d(x_1, x_0) \left(\frac{1}{1-\kappa} \right) = \kappa^{\log_{\kappa} \left(\frac{\varepsilon(1-\kappa)}{d(x_1, x_0)} \right)} d(x_1, x_0) \left(\frac{1}{1-\kappa} \right) = \left(\frac{\varepsilon(1-\kappa)}{d(x_1, x_0)} \right) d(x_1, x_0) \left(\frac{1}{1-\kappa} \right) = \varepsilon,$$

so $\{x_n\}$ is Cauchy and converges to some $x^* \in X$. Using the continuity of f (Example 9.18) to bring a limit inside f (Corollary 4.2), we have

$$x^* = \lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} f(x_{n-1}) = \lim_{n \rightarrow \infty} f\left(\lim_{n \rightarrow \infty} x_{n-1}\right) = T(x^*).$$

We have found a fixed point in x^* . This point is also unique. For the sake of contradiction, suppose there is a second fixed point y^* . In this case

$$d(f(x^*), f(y^*)) = d(x^*, y^*) > \kappa \cdot d(x^*, y^*),$$

which contradicts f being a contraction. \square

The proof for the Banach fixed point theorem is particularly informative, as it gives the fixed point as the limit of the sequence generated from successive applications of the contraction. Given a contraction we can calculate fixed points by taking this limit. This process may seem familiar if you ever got bored in elementary school and had a calculator handy to play with and pass time.¹⁵³ If you type in *any* number greater than zero, and then repeatedly take square-roots, what happens? The results will eventually converge to the fixed point 1. This is because $f(x) = \sqrt{x}$ is a contraction on $(0, \infty)$.

Remark 9.4 (Fixed Point Theorems). The Banach fixed point theorem (whose namesake was the Polish Mathematician Stefan Banach), is one of many fixed point theorems. They play a central role in numerical analysis, differential equations, and in economic theory. Mathematician John Nash's application of Kakutani's fixed point theorem to prove the existence of equilibria in games (see [Nash \(1950\)](#)) is a landmark result that earned him a Nobel Prize.

Now we can state and prove our theorem.

Theorem 9.7. Let $E \subset \mathbb{R}^n$, and $\mathbf{f} : E \rightarrow \mathbb{R}^n$ be continuously differentiable on E . Suppose the linear transformation $\mathbf{f}'(\mathbf{x}_0) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is invertible for an interior point $\mathbf{x}_0 \in E$. Then there exists open sets $U, V \subset \mathbb{R}^n$ such that $\mathbf{x}_0 \in U \subset E$, $\mathbf{f}(\mathbf{x}_0) \in V$, \mathbf{f} is invertible on U , and $\mathbf{f}(U) = V$. In addition, $\mathbf{f}^{-1} : V \rightarrow W$ is continuously differentiable for all $\mathbf{f}(\mathbf{x}) \in V$ where, and

$$(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x})) = [\mathbf{f}'(\mathbf{x})]^{-1}.$$

Proof. This theorem has *a lot* of moving parts, so we'll break the proof into a handful of steps:

1. There exists an open set $U \subset E$ containing \mathbf{x}_0 such that $\mathbf{f}'(\mathbf{x})$ is invertible for all $\mathbf{x} \in U$.
2. The function $\mathbf{f} : U \rightarrow V$ is invertible.
3. The set $V = \mathbf{f}(U)$ is open.
4. The function ϕ is differentiable for all $\mathbf{f}(\mathbf{x}) \in V$, and $\phi'(\mathbf{f}(\mathbf{x})) = [\mathbf{f}'(\mathbf{x})]^{-1}$.
5. The function ϕ is *continuously* for all $\mathbf{f}(\mathbf{x}) \in V$.

¹⁵³As I write this, I'm beginning to suspect this is not a common childhood memory and that I was just a very boring/nerdy child.

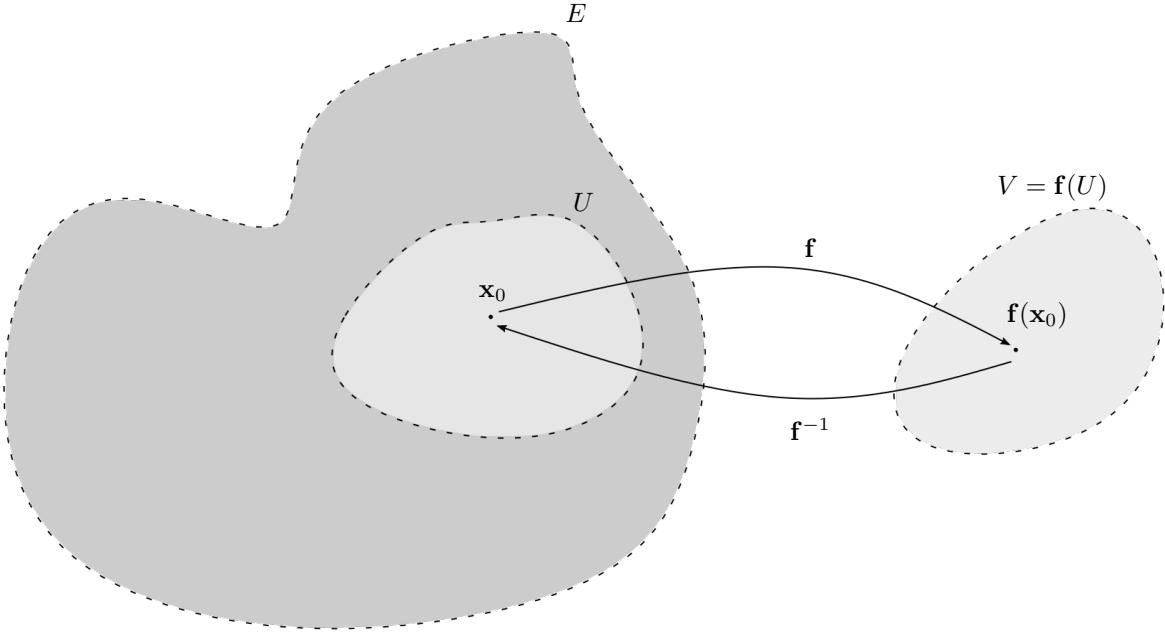


Figure 104: Under the conditions of the inverse function theorem, f is invertible on the set U and $(f^{-1})'(f(\mathbf{x})) = (f'(\mathbf{x}))^{-1}$ for all $\mathbf{x} \in U$.

Step 1: The function f is continuously differentiable on E , so for all $\varepsilon > 0$ there exists some $\delta > 0$ such that

$$\|f'(\mathbf{x}) - f'(\mathbf{x}_0)\|_{\text{op}} < \varepsilon$$

whenever all $\|\mathbf{x} - \mathbf{x}_0\|_{\mathbb{R}^n} < \delta$. If we take $\varepsilon = 1/\left(2\|[f'(\mathbf{x}_0)]^{-1}\|_{\text{op}}\right)$,¹⁵⁴ then there is some $\delta > 0$ such that we have

$$\|f'(\mathbf{x}) - f'(\mathbf{x}_0)\|_{\text{op}} < \frac{1}{2\|[f'(\mathbf{x}_0)]^{-1}\|_{\text{op}}} = \frac{\|[f'(\mathbf{x}_0)]^{-1}\|_{\text{op}}^{-1}}{2} < \|[f'(\mathbf{x}_0)]^{-1}\|_{\text{op}}^{-1} \quad (51)$$

whenever $\|\mathbf{x} - \mathbf{x}_0\| < \delta$. Now recall Lemma 8.3 which said that if $T \in L(\mathbb{R}^n)$ is invertible and T' is “close enough” to T , then T' is invertible. Equation (51) satisfies the conditions of Lemma 8.3 for all $\|\mathbf{x} - \mathbf{x}_0\| < \delta$. Therefore, $f'(\mathbf{x})$ is invertible for all $\|\mathbf{x} - \mathbf{x}_0\| < \delta$, i.e for all \mathbf{x} in the open set $U = B_\delta(\mathbf{x}_0)$ where δ is a function of $\|[f'(\mathbf{x}_0)]^{-1}\|_{\text{op}}^{-1}$.

Step 2: It suffices to show that $f : U \rightarrow V$ is a bijection. For each $\mathbf{y} \in \mathbb{R}^n$, we can define a function $\varphi_{\mathbf{y}} : E \rightarrow \mathbb{R}^n$ as

$$\varphi_{\mathbf{y}}(\mathbf{x}) = \mathbf{x} + [f'(\mathbf{x}_0)]^{-1}(\mathbf{y} - f(\mathbf{x})).$$

Note that

$$\varphi_{\mathbf{y}}(\mathbf{x}) = \mathbf{x} \iff [f'(\mathbf{x}_0)]^{-1}(\mathbf{y} - f(\mathbf{x})) = \mathbf{0} \iff \mathbf{y} = f(\mathbf{x}). \quad (52)$$

¹⁵⁴Remember that we are given that $[f'(\mathbf{x}_0)]^{-1}$ exists.

In other words, \mathbf{x} is a fixed point of $\varphi_{\mathbf{y}}$ if and only if $\mathbf{y} = \mathbf{f}(\mathbf{x})$.¹⁵⁵ If we apply the chain rule to $\varphi_{\mathbf{y}}(\mathbf{x})$ we have

$$\begin{aligned}\varphi'_{\mathbf{y}}(\mathbf{x}) &= I - [\mathbf{f}'(\mathbf{x}_0)]^{-1} \mathbf{f}(\mathbf{x}) \\ &= [\mathbf{f}'(\mathbf{x}_0)]^{-1} \mathbf{f}'(\mathbf{x}_0) - [\mathbf{f}'(\mathbf{x}_0)]^{-1} \mathbf{f}(\mathbf{x}) \\ &= [\mathbf{f}'(\mathbf{x}_0)]^{-1} (\mathbf{f}'(\mathbf{x}_0) - \mathbf{f}(\mathbf{x})).\end{aligned}\tag{53}$$

If we combine this with (51), we have

$$\begin{aligned}\|\varphi'_{\mathbf{y}}(\mathbf{x})\|_{\text{op}} &= \|[\mathbf{f}'(\mathbf{x}_0)]^{-1} (\mathbf{f}'(\mathbf{x}_0) - \mathbf{f}(\mathbf{x}))\|_{\text{op}} \\ &\leq \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \|\mathbf{f}'(\mathbf{x}_0) - \mathbf{f}(\mathbf{x})\|_{\text{op}} \quad (\text{Lemma 8.2}) \\ &< \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \frac{\|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}}}{2} \quad (\text{Equation (51)}) \\ &= \frac{1}{2}\end{aligned}$$

for all $\mathbf{x} \in U$.¹⁵⁶ If we apply Theorem 9.5 to $\varphi_{\mathbf{y}}(\mathbf{x})$ and the bound $\|\varphi'_{\mathbf{y}}(\mathbf{x})\|_{\text{op}} < \frac{1}{2}$, we have

$$\|\varphi_{\mathbf{y}}(\mathbf{x}) - \varphi_{\mathbf{y}}(\mathbf{x}')\| \leq \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n}\tag{54}$$

for all $\mathbf{x}, \mathbf{x}' \in U$. This makes $\varphi_{\mathbf{y}}(\mathbf{x})$ a contraction.¹⁵⁷ This does however show that $\varphi_{\mathbf{y}}$ has at most one fixed point in U .¹⁵⁸ For the sake of contradiction, suppose $\varphi_{\mathbf{y}}(\mathbf{x}) = \mathbf{x}$ and $\varphi_{\mathbf{y}}(\mathbf{x}') = \mathbf{x}'$. This would contradict (54):

$$\|\varphi_{\mathbf{y}}(\mathbf{x}) - \varphi_{\mathbf{y}}(\mathbf{x}')\| = \|\mathbf{x} - \mathbf{x}'\| \not\leq \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n}.$$

For all \mathbf{y} , $\varphi_{\mathbf{y}}(\mathbf{x})$ has at most one fixed point. By (52), we have $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for at most one $\mathbf{x} \in U$, making \mathbf{f} injective on U . We also are given that $V = \mathbf{f}(U)$, so \mathbf{f} is also surjective on V by the definition of $\mathbf{f}(U)$. If $\mathbf{f} : U \rightarrow V$ is injective and surjective, it is invertible.

Step 3. Let \mathbf{y}' be an arbitrary point in V . We must show there exists an open ball containing \mathbf{y}' which lies entirely in V , making any arbitrary point of V an interior point. There exists a unique $\mathbf{x}' \in U$ such that $\mathbf{f}(\mathbf{x}') = \mathbf{y}'$ because \mathbf{f} is injective. The set U is open, so $\mathbf{x}' \in U$ is an interior point, which is to say there exists some $B_r(\mathbf{x}')$ so small that the closure of $B_r(\mathbf{x}')$ is in U , $\overline{B_r(\mathbf{x}') \subset U}$.¹⁵⁹ Note that for any $\mathbf{x} \in \overline{B_r(\mathbf{x}')}$,

$$\|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} \leq r.\tag{55}$$

For this radius r , we will show that

$$B_{\frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|_{\text{op}}}}(\mathbf{y}') \subset V,$$

which makes our arbitrary \mathbf{y}' an interior point of V , thereby showing V is open. Equivalently, for any $\mathbf{y} \in \mathbb{R}^n$ satisfying

$$\|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}^n} < \frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|_{\text{op}}},\tag{56}$$

$\mathbf{y} \in V$. Figure 105 attempts to illustrate this step of our proof.

¹⁵⁵We know that $[\mathbf{f}'(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x})) = \mathbf{0} \implies (\mathbf{y} - \mathbf{f}(\mathbf{x}))$ because $[\mathbf{f}'(\mathbf{x}_0)]^{-1}$ cannot be the zero matrix, otherwise $[\mathbf{f}'(\mathbf{x}_0)]^{-1}(\mathbf{f}'(\mathbf{x}_0)) \neq I$, which violates $\mathbf{f}'(\mathbf{x}_0)$ be invertible.

¹⁵⁶For all $\mathbf{x} \in U$, as these are the set of \mathbf{x} for which $\|\mathbf{x} - \mathbf{x}_0\|_{\mathbb{R}^n} < \delta$ and (51) holds.

¹⁵⁷But we cannot use the Banach fixed point theorem yet. We don't know the range of $\varphi_{\mathbf{y}}(\mathbf{x})$.

¹⁵⁸Actually, any contraction has at most one fixed point.

¹⁵⁹Why do we take the closure of this set as to have a closed subset? This will become clear in a bit.

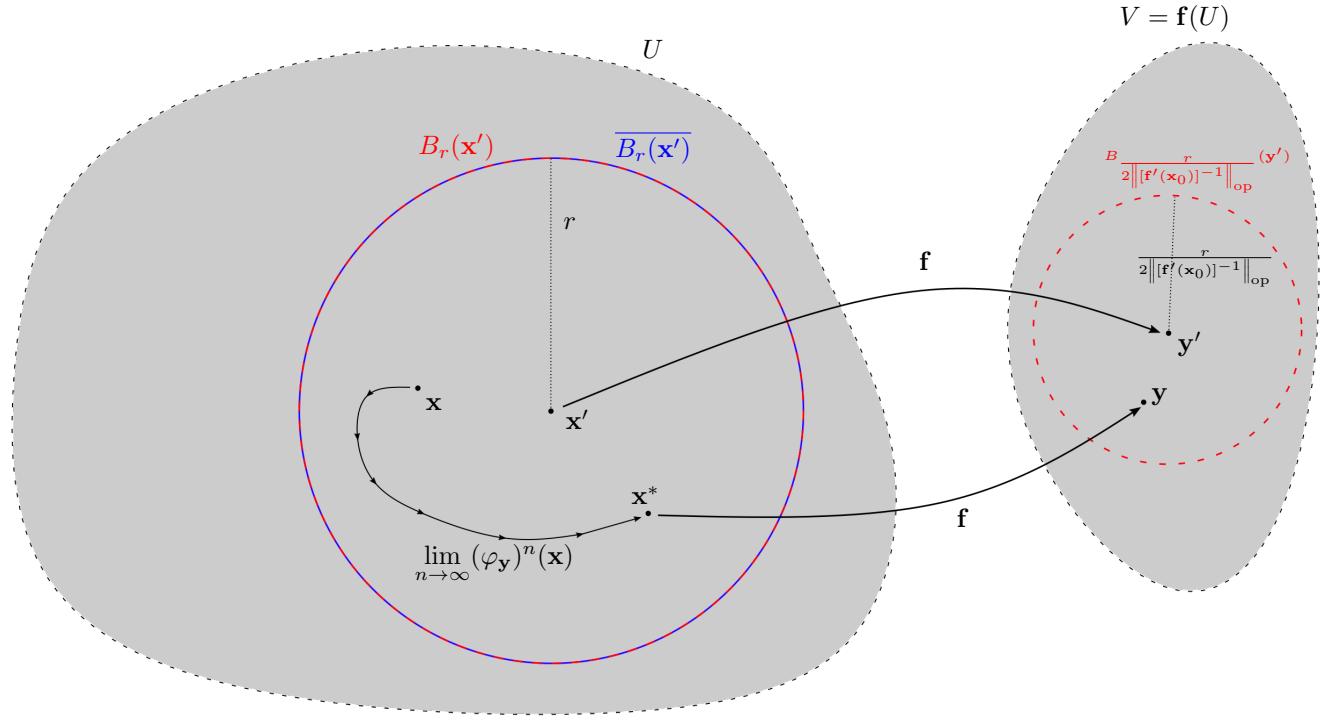


Figure 105: In Step 3, we were show that all V is open by showing any $\mathbf{y} \in V$ is an interior point. This is done by showing all $\mathbf{y} \in V$ for all \mathbf{y} satisfying $\|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}}^n < \frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|_{\text{op}}^{-1}}$. We do this by using applying the Banach fixed point theorem to $\varphi_{\mathbf{y}}(\mathbf{x}) : \overline{B_r(\mathbf{x}') \rightarrow \overline{B_r(\mathbf{x}')}}$ to identify a fixed point \mathbf{x}^* which results from successively applying $\varphi_{\mathbf{y}}$. By the construction of $\varphi_{\mathbf{y}}$, $\mathbf{f}(\mathbf{x}^*) = \mathbf{y}$

Suppose $\mathbf{y} \in \mathbb{R}^n$ satisfies (55). For all $\mathbf{x} \in \overline{B_r(\mathbf{x}')}$, define $\varphi_{\mathbf{y}}(\mathbf{x})$ as we did in Step 2. We have

$$\begin{aligned}
\|\varphi_{\mathbf{y}}(\mathbf{x}) - \mathbf{x}'\|_{\mathbb{R}^n} &= \|\varphi_{\mathbf{y}}(\mathbf{x}) + (\varphi_{\mathbf{y}}(\mathbf{x}') - \varphi_{\mathbf{y}}(\mathbf{x}')) - \mathbf{x}'\|_{\mathbb{R}^n} && \text{(add } \mathbf{0}) \\
&\leq \|\varphi_{\mathbf{y}}(\mathbf{x}) - \varphi_{\mathbf{y}}(\mathbf{x}')\|_{\mathbb{R}^n} + \|\varphi_{\mathbf{y}}(\mathbf{x}') - \mathbf{x}'\|_{\mathbb{R}^n} && \text{(Triangle Inequality)} \\
&\leq \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|\varphi_{\mathbf{y}}(\mathbf{x}') - \mathbf{x}'\|_{\mathbb{R}^n} && \text{(Inequality (54))} \\
&\leq \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|(\mathbf{x}' + [\mathbf{f}'(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x}'))) - \mathbf{x}'\|_{\mathbb{R}^n} && \text{(Definition of } \varphi_{\mathbf{y}}(\mathbf{x}')) \\
&= \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x}'))\|_{\mathbb{R}^n} \\
&\leq \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \|\mathbf{y} - \mathbf{f}(\mathbf{x}')\|_{\mathbb{R}^n} && \text{(Lemma 8.1)} \\
&= \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}^n} && \text{(\mathbf{f}(\mathbf{x}') = \mathbf{y}')} \\
&< \frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|_{\mathbb{R}^n} + \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|_{\text{op}}^{-1}} && \text{(\mathbf{y} \text{ satisfies (56)})} \\
&\leq \frac{1}{2} r + \|[\mathbf{f}'(\mathbf{x}_0)]^{-1}\|_{\text{op}} \frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|_{\text{op}}^{-1}} && \text{(\mathbf{x} \text{ satisfies (55)})} \\
&= r
\end{aligned}$$

If $\|\varphi_{\mathbf{y}}(\mathbf{x}) - \mathbf{x}'\|_{\mathbb{R}^n} < r$, then $\varphi_{\mathbf{y}}(\mathbf{x})$ satisfies (55), so $\varphi_{\mathbf{y}}(\mathbf{x}) \in B_r(\mathbf{x}') \subset \overline{B_r(\mathbf{x}')}$ for all $\mathbf{x} \in \overline{B_r(\mathbf{x}')}$. Now that we know $\varphi_{\mathbf{y}}(\mathbf{x})$ has the same domain and codomain, and satisfies (54), we have that $\varphi_{\mathbf{y}} : \overline{B_r(\mathbf{x}')} \rightarrow \overline{B_r(\mathbf{x}')}$ is a contraction on $\overline{B_r(\mathbf{x}')}$. The set $\overline{B_r(\mathbf{x}')}$ is a closed subset of \mathbb{R}^n so it is compact (Theorem 2.5),¹⁶⁰ thereby making it complete (Theorem 3.6). We can now apply the Banach fixed point theorem (Lemma 9.1) to obtain a fixed point $\varphi_{\mathbf{y}}(\mathbf{x}^*) = \mathbf{x}^*$. By (52), $\mathbf{f}(\mathbf{x}^*) = \mathbf{y}$, so

$$\mathbf{y} \in \underbrace{\mathbf{f}\left(\overline{B_r(\mathbf{x}')}\right)}_{B \frac{r}{2\|\mathbf{f}'(\mathbf{x}_0)\|^{-1}}(\mathbf{y}')} \subset f(U) = V.$$

This makes V open.

Step 4. Let $\mathbf{x} \in U$ and $\mathbf{h} \in \mathbb{R}^n$ such that $\mathbf{x} + \mathbf{h} \in U$. By the injectivity of \mathbf{f} , there exists unique $\mathbf{y} = \mathbf{f}(\mathbf{x}) \in V$ and $\mathbf{f}(\mathbf{x} + \mathbf{h}) \in V$. Define the displacement vector $\mathbf{k} \in V$ as $\mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{y}$. Define $\varphi_{\mathbf{y}}(\mathbf{x})$ the same way we did in Step 2 and Step 3.

$$\begin{aligned} \varphi_{\mathbf{y}}(\mathbf{x} + \mathbf{h}) - \varphi_{\mathbf{y}}(\mathbf{x}) &= \left[(\mathbf{x} + \mathbf{h}) + (\mathbf{f}'(\mathbf{x}_0))^{-1} (\mathbf{y} - \mathbf{f}(\mathbf{x} + \mathbf{h})) \right] - \left[\mathbf{x} + (\mathbf{f}'(\mathbf{x}_0))^{-1} (\mathbf{y} - \mathbf{f}(\mathbf{x})) \right] \\ &= \mathbf{h} + (\mathbf{f}'(\mathbf{x}_0))^{-1} (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x} + \mathbf{h})) \\ &= \mathbf{h} + (\mathbf{f}'(\mathbf{x}_0))^{-1} (\mathbf{y} - \mathbf{f}(\mathbf{x} + \mathbf{h})) \\ &= \mathbf{h} + (\mathbf{f}'(\mathbf{x}_0))^{-1} [-\mathbf{k}] \\ &= \mathbf{h} - (\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k} \end{aligned} \quad \begin{array}{l} (\mathbf{f}(\mathbf{x}) = \mathbf{y}) \\ (\mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{y}) \end{array}$$

Inequality (54) holds for $\mathbf{x} + \mathbf{h}$ and \mathbf{x} as they are elements of U , so

$$\begin{aligned} \|\varphi_{\mathbf{y}}(\mathbf{x} + \mathbf{h}) - \varphi_{\mathbf{y}}(\mathbf{x})\|_{\mathbb{R}^n} &\leq \frac{1}{2} \|(\mathbf{x} + \mathbf{h}) - \mathbf{x}\|_{\mathbb{R}^n} \\ \implies \|\mathbf{h} - (\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k}\|_{\mathbb{R}^n} &\leq \frac{1}{2} \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies \|\mathbf{h} - (\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k}\|_{\mathbb{R}^n} &\geq -\frac{1}{2} \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies \|\mathbf{h}\|_{\mathbb{R}^n} + \|(\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k}\|_{\mathbb{R}^n} &\geq -\frac{1}{2} \|\mathbf{h}\|_{\mathbb{R}^n} \quad \text{(Triangle Inequality)} \\ \implies \|(\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k}\|_{\mathbb{R}^n} &\geq \frac{1}{2} \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies 2 \|(\mathbf{f}'(\mathbf{x}_0))^{-1} \mathbf{k}\|_{\mathbb{R}^n} &\geq \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies 2 \|(\mathbf{f}'(\mathbf{x}_0))^{-1}\|_{\text{op}} \|\mathbf{k}\|_{\mathbb{R}^n} &\geq \|\mathbf{h}\|_{\mathbb{R}^n} \\ \implies 2 &\geq \|(\mathbf{f}'(\mathbf{x}_0))^{-1}\|_{\text{op}} \frac{\|\mathbf{h}\|_{\mathbb{R}^n}}{\|\mathbf{k}\|_{\mathbb{R}^n}} \end{aligned} \quad (57)$$

Finally, note that $\mathbf{f}^{-1}(\mathbf{x} + \mathbf{h})$ exists according to Step 1, and \mathbf{k} is defined such that $\mathbf{k} \rightarrow \mathbf{0}$ is equivalent to $\mathbf{h} \rightarrow \mathbf{0}$. We now have enough information to show that $(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x}))$ exists, and $(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x})) = [\mathbf{f}'(\mathbf{x})]^{-1}$

¹⁶⁰This is why we wanted a closed subset of U ! We needed to make sure the subset was complete.

using Definition 9.1

$$\begin{aligned}
\lim_{\mathbf{k} \rightarrow 0} \frac{\|\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x}) + \mathbf{k}) - \mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) - [\mathbf{f}'(\mathbf{x})]^{-1}\mathbf{k}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} &= \lim_{\mathbf{k} \rightarrow 0} \frac{\|\mathbf{f}^{-1}(\mathbf{y} + \mathbf{k}) - \mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) - [\mathbf{f}'(\mathbf{x})]^{-1}\mathbf{k}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\mathbf{y} = \mathbf{f}(\mathbf{x})) \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x} + \mathbf{h})) - \mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) - [\mathbf{f}'(\mathbf{x})]^{-1}\mathbf{k}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\mathbf{y} + \mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h})) \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|(\mathbf{x} + \mathbf{h}) - \mathbf{x} - [\mathbf{f}'(\mathbf{x})]^{-1}\mathbf{k}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) = \mathbf{x}) \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|\mathbf{h} - [\mathbf{f}'(\mathbf{x})]^{-1}\mathbf{k}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|\mathbf{h} - [\mathbf{f}'(\mathbf{x})]^{-1}[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})]\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})) \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|[f'(\mathbf{x})]^{-1}\mathbf{f}'(\mathbf{x})\mathbf{h} - [f'(\mathbf{x})]^{-1}[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x})]\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\mathbf{h} = [f'(\mathbf{x})]^{-1}\mathbf{f}'(\mathbf{x})\mathbf{h}) \\
&= \lim_{\mathbf{k} \rightarrow 0} \frac{\|[f'(\mathbf{x})]^{-1}[\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}]\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} \\
&\leq \lim_{\mathbf{k} \rightarrow 0} \|(\mathbf{f}'(\mathbf{x}_0))^{-1}\|_{\text{op}} \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} && (\text{Lemma 8.1}) \\
&\leq \lim_{\mathbf{k} \rightarrow 0} \|(\mathbf{f}'(\mathbf{x}_0))^{-1}\|_{\text{op}} \frac{\|\mathbf{h}\|_{\mathbb{R}^n} \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{k}\|_{\mathbb{R}^n}} \\
&\leq \lim_{\mathbf{k} \rightarrow 0} \|(\mathbf{f}'(\mathbf{x}_0))^{-1}\|_{\text{op}} \frac{\|\mathbf{h}\|_{\mathbb{R}^n} \|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
&\leq \lim_{\mathbf{k} \rightarrow 0} 2 \cdot \frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}} \\
&= 2 \lim_{\mathbf{h} \rightarrow 0} \underbrace{\frac{\|\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\|_{\mathbb{R}^m}}{\|\mathbf{h}\|_{\mathbb{R}^n}}}_{0} && (\mathbf{h} \rightarrow 0 \implies \mathbf{k} \rightarrow 0) \\
&= 0
\end{aligned}$$

Step 6. The function $\mathbf{f}^{-1} : V \rightarrow U$ is differentiable, so it is continuous. We can think of differentiation of \mathbf{f} as an operator between U and $L(\mathbb{R}^n)$, specifically between U and the set of invertible transformations in $L(\mathbb{R}^n)$, call it Ω . This operator $\mathbf{f}' : U \rightarrow L(\mathbb{R}^n)$ is continuous because \mathbf{f} is continuously differentiable on E . Finally the inversion function $I : \Omega \rightarrow \Omega$ is continuous.¹⁶¹ We have three continuous functions, the composition of which is continuous. One such composition is:

$$I(\mathbf{f}'(\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})))) = I(\mathbf{f}'(\mathbf{x})) = [\mathbf{f}'(\mathbf{x})]^{-1} = (\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x})),$$

so \mathbf{f}^{-1} is continuously differentiable for all $\mathbf{f}(\mathbf{x}) \in V$.

□

This was a *lot*, so a quick recap about what the inverse function theorem tell us:

1. A continuously differentiable function \mathbf{f} is invertible *in an open ball* of any point \mathbf{x} at which $[\mathbf{f}'(\mathbf{x})]^{-1}$ is invertible. Sometimes we say \mathbf{f} is *locally invertible near \mathbf{x}* .
2. In an open ball of \mathbf{x} where \mathbf{f} is invertible, \mathbf{f}^{-1} is continuously differentiable at $\mathbf{f}(\mathbf{x})$, and

$$(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x})) = [\mathbf{f}'(\mathbf{x})]^{-1}$$

¹⁶¹Take this as given

These facts can also be cast in the light of linear approximation. The best linear approximation of \mathbf{f} at a point \mathbf{x} is given by the linear transformation $\mathbf{f}'(\mathbf{x})$. This means $\mathbf{f}(\mathbf{x}) \approx \mathbf{f}'(\mathbf{x})$ for all \mathbf{x} in some small open ball. If the linear transformation $\mathbf{f}'(\mathbf{x})$ is invertible, then \mathbf{f} should be invertible in the small open ball where $\mathbf{f}(\mathbf{x}) \approx \mathbf{f}'(\mathbf{x})$. Furthermore, this inverse is best approximated at the point \mathbf{x} (linearly) by the inverse of the derivative of \mathbf{f} .

Example 9.20. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) = \begin{cases} x + 2x^2 \sin(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}.$$

This function is shown in Figure 106.

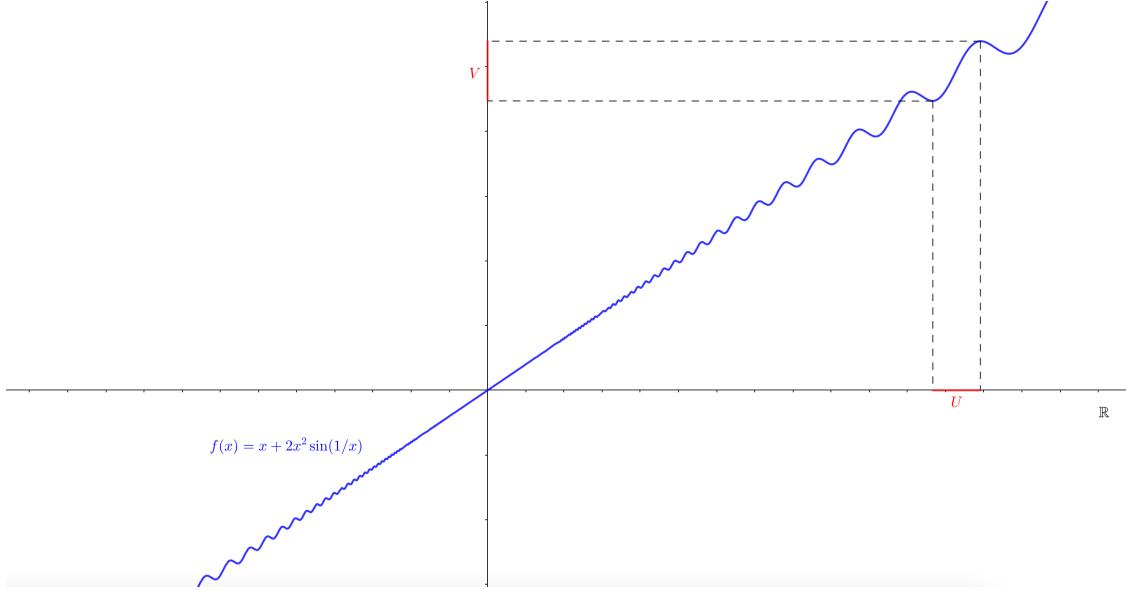


Figure 106: The inverse function theorem cannot be applied at $x = 0$, but can be for any $\mathbf{x} \neq 0$. For any $\mathbf{x} \neq 0$ we can find an open ball U for which f is injective.

The derivative of f is given as

$$f'(x) = \begin{cases} 1 - 2\cos(1/x) + 4x\sin(1/x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}.$$

We have $f'(0) = 1$, as

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{[(x+h) + 2(x+h)^2 \sin(1/(x+h))] - [x + 2x^2 \sin(1/x)]}{h} &= \lim_{h \rightarrow 0} \frac{[2(x+h)^2 \sin(1/(x+h))] - [x + 2x^2 \sin(1/x)]}{h} + \frac{h}{h} \\ &= \lim_{h \rightarrow 0} \frac{[2(x+h)^2 \sin(1/(x+h))] - [x + 2x^2 \sin(1/x)]}{h} + 1 \\ &= 1. \end{aligned}$$

The function f is not continuously differentiable at 0, so the inverse function theorem does not hold at $x = 0$. There exists no open ball U containing 0 for which f is injective, which can be seen in Figure 106. As f

approaches 0 from either direction, it begins to oscillate more and more rapidly. This means f cannot be inverted in any open ball around 0.

We can however verify part of the inverse function theorem holds for all $x \neq 0$. Suppose $x = 1/10$. Let

$$U = (0.088216\dots, 0.141311\dots)$$

such that

$$f(U) = V = (0.073544\dots, 0.169776\dots).$$

The endpoints of U are local minima and maxima, and $f'(x) > 0$ for all $x \in U$, so f is monotonically increasing on U . This means $f : U \rightarrow V$ is injective on U , and invertible (it is trivially surjective on $V = f(U)$).

Example 9.21 (Locally Invertible Everywhere but Not Globally Invertible). Suppose $\mathbf{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is defined as $\mathbf{f}(\mathbf{x}) = (e^{x_1} \cos(x_2), e^{x_1} \sin(x_2))$. This function is not invertible on \mathbb{R}^2 as it is periodic (thus not injective):

$$\mathbf{f}(x_1, x_2) = \mathbf{f}(x_1, x_2 + 2\pi n) \quad \forall n \in \mathbb{Z}.$$

Despite this, we can use the inverse function theorem to show that \mathbf{f} is locally invertible at any point $\mathbf{x} \in \mathbb{R}$. For some fixed $\mathbf{x} \in \mathbb{R}^2$ we have

$$\mathbf{f}'(\mathbf{x}) = \begin{bmatrix} e^{x_1} \cos x_2 & e^{x_1} \sin x_2 \\ -e^{x_1} \sin x_2 & e^{x_1} \cos x_2 \end{bmatrix}.$$

For all $\mathbf{x} \in \mathbb{R}$,

$$\det(\mathbf{f}'(\mathbf{x})) = e^{2x_1}(\cos^2 x_2 + \sin^2 x_2) = e^{2x_1},$$

which is non-zero for all $\mathbf{x} \in \mathbb{R}^2$, meaning $\mathbf{f}'(\mathbf{x})$ is invertible on all of \mathbb{R}^2 . By the inverse function theorem, \mathbf{f} is locally invertible at any point $\mathbf{x} \in \mathbb{R}$ in an open ball U . In any such open ball, the derivative of $\mathbf{f} : V \rightarrow U$, where $V = \mathbf{f}(U)$, we have

$$\begin{aligned} (\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x})) &= [\mathbf{f}'(\mathbf{x})]^{-1} \\ &= \frac{1}{\det(\mathbf{f}'(\mathbf{x}))} \cdot \mathbf{f}'(\mathbf{x}) \\ &= e^{-2x_1} \begin{bmatrix} e^{x_1} \cos x_2 & e^{x_1} \sin x_2 \\ -e^{x_1} \sin x_2 & e^{x_1} \cos x_2 \end{bmatrix} \\ &= \begin{bmatrix} e^{-x_1} \cos x_2 & e^{-x_1} \sin x_2 \\ -e^{-x_1} \sin x_2 & e^{-x_1} \cos x_2 \end{bmatrix}. \end{aligned}$$

Remark 9.5 (How Local is “Locally Invertible”). The inverse function theorem only stipulates that \mathbf{f} is locally invertible on some open set U . It says *nothing* about the size of U . [Lang \(2012\)](#) gives a result related to the size of U .

9.7 The Implicit Function Theorem

The inverse function theorem says that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ can be solved locally in terms of \mathbf{x} (i.e inverted) if $\mathbf{f}'(\mathbf{x})$ is invertible. Alternatively, $\mathbf{f}(\mathbf{x}) - \mathbf{y} = \mathbf{0}$ can be solved locally in terms of \mathbf{x} if $\mathbf{f}'(\mathbf{x})$ is invertible. Written out

in full, the system

$$\begin{aligned} f_1(x_1, \dots, x_n) - y_1 &= 0 \\ f_2(x_1, \dots, x_n) - y_2 &= 0 \\ &\vdots \\ f_n(x_1, \dots, x_n) - y_n &= 0 \end{aligned}$$

can be written as

$$\begin{aligned} f_1^{-1}(y_1, \dots, y_n) &= x_1 \\ f_2^{-1}(y_1, \dots, y_n) &= x_2 \\ &\vdots \\ f_n^{-1}(y_1, \dots, y_n) &= x_3 \end{aligned}$$

if $\mathbf{f}'(\mathbf{x})$ is invertible, where $\mathbf{f}^{-1} : U \rightarrow V$ is defined locally on some open $U \subset \mathbb{R}^n$ containing \mathbf{x} and $V = \mathbf{f}(U)$. Can this result be extended to arbitrary systems of \mathbf{x} and \mathbf{y} , not just systems of the form $\mathbf{f}(\mathbf{x}) - \mathbf{y} = \mathbf{0}$? Let's investigate this with a simple example.

Example 9.22. Suppose we have the equation corresponding to the unit circle, $x^2 + y^2 = 1$. Can we solve for one variable in terms of the other? Yes and no. If we attempt to solve for y , we get

$$y = \pm\sqrt{1 - x^2},$$

which is not well defined as it takes on multiple values. We can however solve for y locally, as illustrated in Figure 107. To be specific, we have

$$\begin{aligned} y &= \sqrt{x^2 - 1} \text{ if } (x, y) \in (-1, 1) \times (0, 1], \\ y &= -\sqrt{x^2 - 1} \text{ if } (x, y) \in (-1, 1) \times [-1, 0). \end{aligned}$$

By restricting our attention to a region of \mathbb{R}^2 , we can ensure that $x^2 + y^2 = 1$ is a valid function and solve for y . But can we do this for some open ball around any part of $x^2 + y^2 = 1$? Unfortunately, no. In the event $y = 0$, we do not get a well defined function in terms of x . For example, at $(x, y) = (1, 0)$:

$$\begin{aligned} \sqrt{x^2 - 1} &= \sqrt{1^2 - 1} = \sqrt{0} = 0, \\ -\sqrt{x^2 - 1} &= -\sqrt{1^2 - 1} = -\sqrt{0} = 0. \end{aligned}$$

The inverse function theorem relates the differentiability of a function to whether it is invertible, so perhaps differentiation $x^2 + y^2 = 1$ will provide insight as to why we cannot solve for y locally when $y = 0$. Recall from calculus that the derivative of this curve at a point is given by *implicit differentiation*.

$$\begin{aligned} \frac{d}{dx}[x^2 + y^2] &= \frac{d}{dx}[1] \\ \implies \frac{d}{dx}[x^2] + \frac{d}{dx}[y^2] &= 0 \\ \implies 2x + 2y \frac{dy}{dx} &= 0 \\ \implies \frac{dy}{dx} &= -\frac{x}{y} \end{aligned}$$

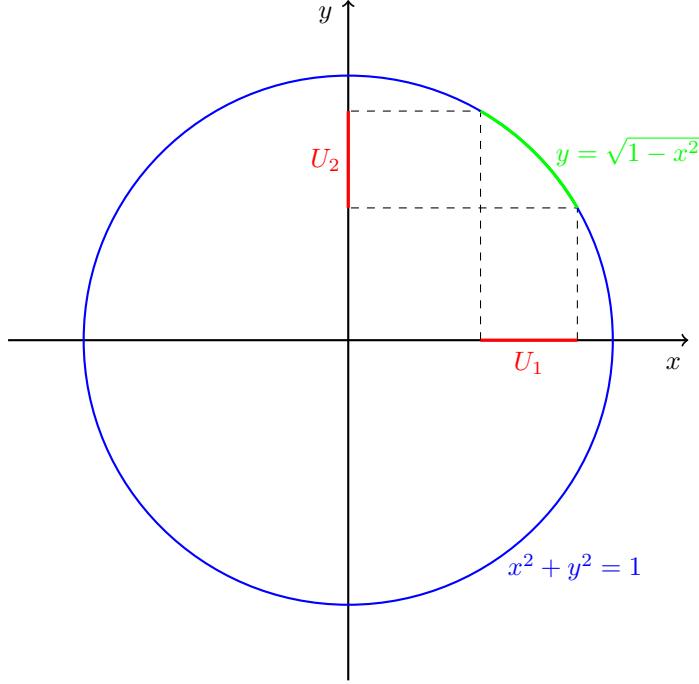


Figure 107: Given the equation $x^2 + y^2 = 1$, we have $y = \sqrt{1 - x^2}$ for all $(x, y) \in U_1 \times U_2$.

The slope of the line tangent to $x^2 + y^2 = 1$ is given by $-x/y$. This derivative is defined everywhere *except* when $y = 0$, which is precisely when we are not able to solve for y locally! Not only is this not a coincidence, but also this is reminiscent of the inverse function theorem which determines whether we can solve $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for \mathbf{x} using the existence of a derivative.

Before stating and proving the implicit function theorem, we should properly outline what an implicit function is. Until now we have exclusively dealt with *explicit functions* $f : X \rightarrow Y$ if it is written as $y = f(x)$. An *implicit function* is instead given by the relation $f(x, y) = 0$. Given an implicit function, solving one variable in terms of another amounts to rewriting the relation $f(x, y) = 0$ as an explicit function. As we saw for $f(x, y) = x^2 + y^2 - 1 = 0$, we cannot always find an explicit function corresponding $f(x, y)$ for all (x, y) , but we can do it locally for certain values of (x, y) sometimes.

Notation 9.1. To present the implicit function theorem in a succinct fashion, we will introduce some specific notation. Suppose $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. We can concatenate these two vectors and write

$$(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_n, y_1, \dots, y_m) \in \mathbb{R}^{n+m}.$$

If we have some function $\mathbf{F} : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$, then we can decompose $\mathbf{F}'(\mathbf{x}, \mathbf{y})$ into two matrices:

$$\mathbf{F}'(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_1}{\partial x_n}(\mathbf{x}, \mathbf{y}) & \left| \begin{array}{ccc} \frac{\partial F_1}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_1}{\partial y_m}(\mathbf{x}, \mathbf{y}) \end{array} \right. \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial F_m}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_m}{\partial x_n}(\mathbf{x}, \mathbf{y}) & \left| \begin{array}{ccc} \frac{\partial F_m}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_m}{\partial y_m}(\mathbf{x}, \mathbf{y}) \end{array} \right. \end{pmatrix} = \begin{bmatrix} \mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{y}) & \mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) \end{bmatrix},$$

where $\mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{y}) \in L(\mathbb{R}^n, \mathbb{R}^m)$ and $\mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) \in L(\mathbb{R}^m)$. For some $\mathbf{h} \in \mathbb{R}^n$ and $\mathbf{k} \in \mathbb{R}^m$,

$$\mathbf{F}'(\mathbf{x}, \mathbf{y})(\mathbf{h}, \mathbf{k}) = \mathbf{F}'(\mathbf{x}, \mathbf{y}) \begin{bmatrix} \mathbf{h} \\ \mathbf{k} \end{bmatrix} = \mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{y})\mathbf{h} + \mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y})\mathbf{k}$$

In general for any linear transformation $A \in L(\mathbb{R}^{n+m}, \mathbb{R}^m)$ we can always write it as $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ where $A_1 \in L(\mathbb{R}^n, \mathbb{R}^m)$ and $A_2 \in L(\mathbb{R}^m)$.

Lemma 9.2. Suppose $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \in L(\mathbb{R}^{n+m}, \mathbb{R}^m)$ and $A_2 \in L(\mathbb{R}^m)$ is invertible. For all $\mathbf{h} \in \mathbb{R}^n$, there is a unique $\mathbf{k} \in \mathbb{R}^m$ such that $A(\mathbf{h}, \mathbf{k}) = \mathbf{0}$.

Proof. Define $\mathbf{k} = -(A_2)^{-1}A_1\mathbf{h}$. We have

$$A(\mathbf{h}, \mathbf{k}) = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{k} \end{bmatrix} = A_1\mathbf{h} + A_2\mathbf{k} = A_1\mathbf{h} + A_2[-(A_2)^{-1}A_1\mathbf{h}] = A_1\mathbf{h} - A_1\mathbf{h} = \mathbf{0}.$$

□

Theorem 9.8 (Implicit Function Theorem). Let $\mathbf{F} : E \rightarrow \mathbb{R}^m$, where $E \subset \mathbb{R}^{n+m}$ is open, be a continuously differentiable function such that $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ for some $(\mathbf{x}_0, \mathbf{y}_0) \in E$. If $\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)$ is invertible, then there exists open sets $U \subset \mathbb{R}^{n+m}$ and $V \subset \mathbb{R}^n$, with $(\mathbf{x}_0, \mathbf{y}_0) \in U$ and $\mathbf{x}_0 \in V$ such that:

1. For all $\mathbf{x} \in V$ there is a corresponding unique \mathbf{y} such that $(\mathbf{x}, \mathbf{y}) \in U$ and $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$;
2. If this unique \mathbf{y} is defined to be $\mathbf{y} = \phi(\mathbf{x})$, then $\mathbf{g}(\mathbf{x}_0) = \mathbf{y}_0$, $\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in W$, $\phi : W \rightarrow \mathbb{R}^m$ is continuously differentiable and,

$$\mathbf{g}'(\mathbf{x}_0) = -[\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)]^{-1}\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0).$$

Proof. This proof will be broken into the following steps:

1. Define a function $\mathbf{H} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+n}$ such that we can apply the inverse function theorem to \mathbf{H} at $(\mathbf{x}_0, \mathbf{y}_0)$
2. Define W and verify that: $\mathbf{x}_0 \in W$, W is open, and for all $\mathbf{x} \in W$ there exists a unique \mathbf{y} such that $(\mathbf{x}, \mathbf{y}) \in U$ and $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.
3. Define $\mathbf{g}(\mathbf{x}) = \mathbf{y}$ using Step 2 and verify that: $\mathbf{g}(\mathbf{x}_0) = \mathbf{y}_0$, $\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in W$, $\phi : W \rightarrow \mathbb{R}^m$ is continuously differentiable, and $\mathbf{g}'(\mathbf{x}_0) = -[\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)]^{-1}\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$.

Step 1. In order to prove this result, we will rewrite our problem in terms of a function $\mathbf{H} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+n}$ and then apply the inverse function theorem at the point $(\mathbf{x}_0, \mathbf{y}_0)$.

Define

$$\mathbf{H}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{F}(\mathbf{x}, \mathbf{y}) \end{bmatrix}.$$

In order to apply the inverse function theorem to $\mathbf{H}(\mathbf{x}, \mathbf{y})$, we need to verify that \mathbf{H} is continuously differentiable on $E \subset \mathbb{R}^{n+m}$ and $\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)$ is an invertible element of $L(\mathbb{R}^{n+m})$. The components of

\mathbf{H} , \mathbf{x} and \mathbf{F} , are continuously differentiable, so \mathbf{H} is. Calculating $\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)$ gives

$$\begin{aligned}
\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0) &= \begin{bmatrix} \frac{\partial H_1}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_1}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial H_1}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_1}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial H_n}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_n}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial H_n}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_n}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \frac{\partial H_{n+1}}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_{n+1}}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial H_{n+1}}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_{n+1}}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial H_{n+m}}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_{n+m}}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial H_{n+m}}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial H_{n+m}}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial x_1}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial x_1}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial x_1}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial x_1}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial x_n}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial x_n}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial x_n}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \frac{\partial F_1}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_1}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial F_1}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_1}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_m}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial F_m}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_m}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \end{bmatrix} \\
&= \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_1}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial F_1}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_1}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_m}{\partial x_n}(\mathbf{x}_0, \mathbf{y}_0) & \frac{\partial F_m}{\partial y_1}(\mathbf{x}_0, \mathbf{y}_0) & \dots & \frac{\partial F_m}{\partial y_m}(\mathbf{x}_0, \mathbf{y}_0) \end{bmatrix} \\
&= \left(\begin{array}{c|c} I & 0 \\ \hline \mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) & \mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) \end{array} \right).
\end{aligned}$$

The determinant of $\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)$ is given as:¹⁶²

$$\det(\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)) = \det(I) \det(\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) - I \mathbf{0} \mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)) = 1 \cdot \det(\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)) = \det(\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)) \neq 0,$$

where $\det(\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)) \neq 0$ as a result of $\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)$ being invertible. Therefore, we have that $\det(\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)) \neq 0$, so $\mathbf{H}'(\mathbf{x}_0, \mathbf{y}_0)$ is invertible where \mathbf{H} is continuously differentiable on $E \subset \mathbb{R}^{n+m}$.

By the inverse function theorem (Theorem 9.7), there exists open sets $U, V \subset \mathbb{R}^{n+m}$ with $(\mathbf{x}_0, \mathbf{y}_0) \in U$,

¹⁶²The determinant of any block matrix $\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right)$ is $\det(A) \det(D - CA^{-1}B)$, which is a natural extension of

$$\det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = ad - bc = a(d - ca^{-1}b).$$

$\mathbf{H}(\mathbf{x}_0, \mathbf{y}_0) \in V$, such that \mathbf{H} is invertible on U .¹⁶³ Because $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$,

$$\mathbf{H}(\mathbf{x}_0, \mathbf{y}_0) = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) \end{bmatrix} = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{0} \end{bmatrix} = (\mathbf{x}_0, \mathbf{0}) \in V \quad (58)$$

Step 2. Define the set $W = \{\mathbf{x} \in \mathbb{R}^n \mid (\mathbf{x}, \mathbf{0})\} \in V$. The three desired properties of W are shown below:

- (a) By (58), $\mathbf{x}_0 \in W$.
- (b) Suppose $\mathbf{x} \in W$. Then $(\mathbf{x}, \mathbf{0}) \in V$, where V is open. Thus, there exists an open ball $B_r((\mathbf{x}, \mathbf{0})) \subset V$. For this radius r and this \mathbf{x} , we have an open ball

$$\{(\mathbf{x}', \mathbf{0}) \in V \mid \|\mathbf{x} - \mathbf{x}'\| < r\} \subset W.$$

This makes W open.

- (c) For $\mathbf{x} \in W$, $(\mathbf{x}, \mathbf{0}) \in V$. Since $\mathbf{H} : U \rightarrow V$ is invertible on U it is surjective, and there exists some $(\mathbf{x}, \mathbf{y}) \in U$ such that

$$\begin{aligned} \mathbf{H}(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}, \mathbf{0}) && (\mathbf{H} \text{ is surjective}) \\ \implies \begin{bmatrix} \mathbf{x} \\ \mathbf{F}(\mathbf{x}, \mathbf{y}) \end{bmatrix} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} && (\text{Definition of } \mathbf{H}) \\ \implies \mathbf{F}(\mathbf{x}, \mathbf{y}) &= \mathbf{0} \end{aligned}$$

for this \mathbf{y} . Now suppose there exists some other \mathbf{y}' such that $\mathbf{F}(\mathbf{x}, \mathbf{y}') = \mathbf{0}$. We have

$$\begin{aligned} \mathbf{f}(\mathbf{x}, \mathbf{y}) &= \mathbf{0} = \mathbf{f}(\mathbf{x}, \mathbf{y}') \\ \implies \begin{bmatrix} \mathbf{x} \\ \mathbf{f}(\mathbf{x}, \mathbf{y}) \end{bmatrix} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{f}(\mathbf{x}, \mathbf{y}') \end{bmatrix} \\ \implies \mathbf{H}(\mathbf{x}, \mathbf{y}) &= \mathbf{H}(\mathbf{x}, \mathbf{y}') && (\text{Definition of } \mathbf{H}) \\ \implies \mathbf{y} &= \mathbf{y}', && (\mathbf{H} \text{ injective on } U) \end{aligned}$$

so \mathbf{y} is unique.

Step 3. Define $\phi : W \rightarrow \mathbb{R}^m$ such that

$$\mathbf{g}(\mathbf{x}) = \{\mathbf{y} \in W \mid (\mathbf{x}, \mathbf{y}) \in U \text{ and } \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}\} \quad (59)$$

This function is well defined, as $\mathbf{g}(\mathbf{x})$ takes on a unique value \mathbf{y} according to Step 2 (so the set in (59) is always a singleton.). The three desired properties of \mathbf{g} are shown below:

- (a) For \mathbf{x}_0 ,

$$\mathbf{g}(\mathbf{x}_0) = \{\mathbf{y} \in W \mid (\mathbf{x}_0, \mathbf{y}) \in U \text{ and } \mathbf{F}(\mathbf{x}_0, \mathbf{y}) = \mathbf{0}\} = \{\mathbf{y}_0\}.$$

- (b) By the definition of ϕ ,

$$\mathbf{H}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \begin{bmatrix} \mathbf{x} \\ \mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} = (\mathbf{x}, \mathbf{0}),$$

so $\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$.

¹⁶³The inverse function theorem also gives $V = \mathbf{H}(U)$.

(c) The inverse function theorem gave that $\mathbf{H} : U \rightarrow V$ is invertible, so we have

$$\begin{aligned}\mathbf{H}^{-1}(\mathbf{H}(\mathbf{x}, \mathbf{g}(\mathbf{x}))) &= \mathbf{H}^{-1}(\mathbf{x}, \mathbf{0}) \\ \implies (\mathbf{x}, \mathbf{g}(\mathbf{x})) &= \mathbf{H}^{-1}(\mathbf{x}, \mathbf{0}).\end{aligned}$$

The inverse function theorem also gives that \mathbf{H}^{-1} is continuously differentiable, so its components must be. This makes \mathbf{g} continuously differentiable.

(d) Define $\Phi(\mathbf{x}) = (\mathbf{x}, \mathbf{g}(\mathbf{x}))$. By part (b), $\mathbf{F}(\Phi(\mathbf{x})) = \mathbf{0}$. Applying the chain rule to this gives

$$\begin{aligned}\mathbf{F}'(\Phi(\mathbf{x}))\Phi'(\mathbf{x}) &= 0 \\ \implies \mathbf{F}'(\mathbf{x}, \mathbf{g}(\mathbf{x})) \begin{bmatrix} I \\ \mathbf{g}'(\mathbf{x}) \end{bmatrix} &= \mathbf{0} \\ \implies \begin{bmatrix} \mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{g}(\mathbf{x})) & \mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{g}(\mathbf{x})) \end{bmatrix} \begin{bmatrix} I \\ \mathbf{g}'(\mathbf{x}) \end{bmatrix} &= \mathbf{0} \\ \implies \mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{g}(\mathbf{x})) + \mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{g}(\mathbf{x}))\mathbf{g}'(\mathbf{x}) &= \mathbf{0}\end{aligned}$$

Applying this linear transformation at the point $(\mathbf{x}_0, \mathbf{y}_0)$ gives

$$\begin{aligned}\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)) + \mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))\mathbf{g}'(\mathbf{x}_0) &= \mathbf{0} \\ \implies \mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) + \mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)\mathbf{g}'(\mathbf{x}_0) &= \mathbf{0} \quad (\mathbf{g}(\mathbf{x}_0) = \mathbf{y}_0) \\ \implies \mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)\mathbf{g}'(\mathbf{x}_0) &= -\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0) \\ \implies \mathbf{g}'(\mathbf{x}_0) &= -[\mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)]^{-1}\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)\end{aligned}$$

□

To make the proof of the implicit function theorem more concrete, we can replicate the proof with an actual example.

Example 9.23. Define $\mathbf{F} : \mathbb{R}^4 \rightarrow \mathbb{R}^2$ as

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{F}(x_1, x_2, y_1, y_2) = \begin{bmatrix} 3x_1 + y_1 \\ x_2 + 2y_2 \end{bmatrix}.$$

For $\mathbf{x}_0 = (1, -3)$, and $\mathbf{y}_0 = (-3, 1)$, we have $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$ and

$$\mathbf{F}'(\mathbf{x}_0, \mathbf{y}_0) = \left(\begin{array}{cc|cc} 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \end{array} \right) = \begin{bmatrix} \mathbf{F}'_{\mathbf{x}}(\mathbf{x}, \mathbf{y}) & \mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) \end{bmatrix}.$$

The matrix $\mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y})$ is invertible, and

$$[\mathbf{F}'_{\mathbf{y}}(\mathbf{x}, \mathbf{y})]^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Now define $\mathbf{H} : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ as

$$\mathbf{H}(x_1, x_2, y_1, y_2) = \begin{bmatrix} x_1 \\ x_2 \\ 3x_1 + y_1 \\ x_2 + 2y_2 \end{bmatrix}.$$

FINISH

Example 9.24 (The Special Case of Inverting \mathbf{f}). As stated earlier, the inverse function theorem is a special case of the implicit function theorem. Define $\mathbf{F} : \mathbb{R}^{2n} \rightarrow \mathbb{R}^n$ as $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x}) - \mathbf{y}$ for some $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$. If $\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0}$, and $\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$ is invertible, we can solve express $\mathbf{f}(\mathbf{x}_0) - \mathbf{y}_0 = \mathbf{0}$ in terms of \mathbf{x}_0 . This is of course the same as solving $\mathbf{f}(\mathbf{x}_0) = \mathbf{y}_0$ for \mathbf{x}_0 , i.e inverting \mathbf{f} . Also note that

$$\mathbf{F}'(\mathbf{x}_0, \mathbf{y}_0) = \begin{bmatrix} \mathbf{f}'(\mathbf{x}_0) & -I \end{bmatrix},$$

so $\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$ being invertible is equivalent to $\mathbf{f}'(\mathbf{x}_0)$ being invertible. In this case we have some \mathbf{g} such that $\mathbf{g}(\mathbf{y}) = \mathbf{x}$ in an open ball around \mathbf{f} . In this case $\mathbf{g} = \mathbf{f}^{-1}$. Furthermore

$$(\mathbf{f}^{-1})'(\mathbf{f}(\mathbf{x}_0)) = (\mathbf{f}^{-1})'(\mathbf{y}_0) = \mathbf{g}'(\mathbf{y}_0) = -[\mathbf{F}'_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)]^{-1} \mathbf{F}'_{\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = -[\mathbf{f}'(\mathbf{x}_0)]^{-1}(-I) = [\mathbf{f}'(\mathbf{x}_0)]^{-1}.$$

Example 9.25. Suppose we have a system of equations

$$\begin{aligned} 3x + y - z + u^2 &= 0 \\ x - y + 2z + u &= 0 \\ 2x + 2y - 3z + 2u &= 0 \end{aligned}$$

Note that

$$\begin{aligned} 0 + 0 - 0 &= 0 \\ \implies (x - y + 2z + u) + (2x + 2y - 3z + 2u) - (3x + y - z + u^2) &= 0 \\ \implies 3u - u^2 &= 0 \\ \implies u &\in \{0, 3\}. \end{aligned}$$

We can use the implicit function theorem to determine which variables we can solve this system in terms of. This system can be written as the implicit function

$$\mathbf{F}(x, y, z, u) = \begin{bmatrix} 3x + y - z + u^2 \\ x - y + 2z + u \\ 2x + 2y - 3z + 2u \end{bmatrix}.$$

For any $(x, y, z, u) \in \mathbb{R}^4$ we have

$$\mathbf{F}'(x, y, z, u) = \begin{bmatrix} 3 & 1 & -1 & 2u \\ 1 & -1 & 2 & 1 \\ 2 & 2 & -3 & 2 \end{bmatrix}.$$

We have:

$$\begin{aligned}\det\left(\mathbf{F}'_{(x,y,z)}(x,y,z,u)\right) &= \begin{vmatrix} 3 & 1 & -1 \\ 1 & -1 & 2 \\ 2 & 2 & -3 \end{vmatrix} = 0 \\ \det\left(\mathbf{F}'_{(y,z,u)}(x,y,z,u)\right) &= \begin{vmatrix} 1 & -1 & 2u \\ -1 & 2 & 1 \\ 2 & -3 & 2 \end{vmatrix} = 3 - 2u \neq 0 & (u \in \{0, 3\}) \\ \det\left(\mathbf{F}'_{(x,z,u)}(x,y,z,u)\right) &= \begin{vmatrix} 3 & -1 & 2u \\ 1 & 2 & 1 \\ 2 & -3 & 2 \end{vmatrix} = 21 - 14u \neq 0 & (u \in \{0, 3\}) \\ \det\left(\mathbf{F}'_{(x,y,u)}(x,y,z,u)\right) &= \begin{vmatrix} 3 & 1 & 2u \\ 1 & -1 & 1 \\ 2 & 2 & 2 \end{vmatrix} = 12u - 18 \neq 0 & (u \in \{0, 3\})\end{aligned}$$

The matrix $\mathbf{F}'_{(x,y,z)}(x,y,z,u)$ is not invertible, so we cannot solve for u in terms of x, y, z . On the other hand $\mathbf{F}'_{(y,z,u)}$, $\mathbf{F}'_{(x,z,u)}$, and $\mathbf{F}'_{(x,y,u)}$ are all invertible, so we can solve for x, y , and z in terms of the other variables.

10 Riemann Integration with Several Variables

Section 6 developed the theory of Riemann integration for real bound functions, mostly following the direction of [Rudin \(1976\)](#). We know extend this theory to functions of several variables. Neither [Rudin \(1976\)](#) nor [Tao \(2016a,b\)](#) treat this material, instead choosing to refine the notion of integration on \mathbb{R} with a concept we will take up in Section 13. Riemann integration with functions of several variables is covered by [Apostol \(1974\)](#), [Munkres \(1999\)](#), and [Spivak \(1965\)](#).

First, we will extend the definition of Riemann integrability to a bounded function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. This proves to be relatively straightforward and the existence of the integral is given by Riemann's criterion and Lebesgue's criterion, two results introduced in Section 6. Then we will consider two issues that arise when integrating functions of several variables – how do we calculate these integrals, and how do we integrate over regions that are not rectangles. Finally, we properly generalize u -substitution and introduce change of variables in earnest.

10.1 Integration over a Rectangle

We will quickly generalize the key definitions and results of Section 6.1 and 6.2. Recall that Definition 6.1 outlined how to partition an interval of \mathbb{R} . We can extend this definition to \mathbb{R}^n and partition some rectangle in \mathbb{R}^n . In this context, a *rectangle* in \mathbb{R}^n is simply the Cartesian product of n intervals in \mathbb{R} .

Definition 10.1. Suppose $Q \subset \mathbb{R}^n$ is a rectangle in \mathbb{R}^n where

$$Q = [a_1, b_1] \times \cdots \times [a_n, b_n].$$

A *partition* of Q is a tuple $P = P_1, \dots, P_n$ such that P_j is a partition of $[a_j, b_j]$ for all j . If I_j is one of the subintervals of the partition P_j for all j , then we say $R = I_1 \times \dots \times I_n$ is a *subrectangle determined by P* . If

R is a rectangle determined by P , we will write $R \in P$.¹⁶⁴

Definition 10.2. A partition $P^* = (P_1^*, \dots, P_n^*)$ is a *refinement* of $P = (P_1, \dots, P_n)$ if P_j^* is a refinement of P_j for at least one j . P^* is a *common refinement* of P and P' if P_j^* is a refinement for P_j and P'_j for all J :

$$P^* = (P_1 \cup P'_1, \dots, P_n \cup P'_n).$$

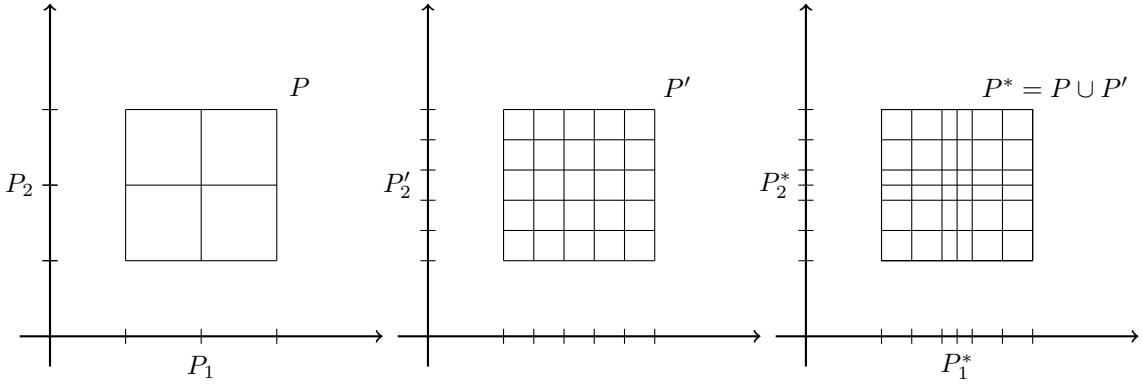


Figure 108: Two partitions of a common rectangle in \mathbb{R}^2 , along with a common refinement.

Suppose $P = (P_1, \dots, P_n)$ partitions $Q = [a_1, b_1] \times \dots \times [a_n, b_n] \subset \mathbb{R}^n$, where $P_j = \{x_{0,j}, \dots, x_{k_j,j}\}$. That is, we have n one-dimensional partitions P_j indexed by j , each of which is comprised of k_j disjoint intervals $[x_{i-1,j}, x_{i,j}]$ such that

$$[a_j, b_j] = \bigcup_{i=1}^{k_j} [x_{i-1,j}, x_{i,j}].$$

A subrectangle is a rectangle $R \subseteq Q$ of the form

$$R = [x_{\ell_1-1,1}, x_{\ell_1,1}] \times [x_{\ell_2-1,2}, x_{\ell_2,2}] \times \dots \times [x_{\ell_n-1,n}, x_{\ell_n,n}].$$

We simply take the Cartesian product of the ℓ_1 th interval of P_1 , the ℓ_2 th interval of P_2 , etc. Each P_j has k_j intervals to choose from, giving k_j ways to choose ℓ_j . This means there are $k_1 \times k_2 \times \dots \times k_n$ subrectangles determined by P , the union of which is Q .

$$Q = \bigcup_{\ell_1=1}^{k_1} \bigcup_{\ell_2=1}^{k_2} \dots \bigcup_{\ell_n=1}^{k_n} \left(\prod_{j=1}^n (x_{\ell_j-1,j}, x_{\ell_j,j}) \right).$$

When $n = 1$ in Section 6, our subrectangles were simply the subintervals $[x_{i-1}, x_i]$ of the one dimensional partition P . The length of each of these subrectangles were written as $\Delta x_i = x_i - x_{i-1}$. For $n > 1$, the analogous concept of length is volume (or area if $n = 2$). For $R = [x_{\ell_1-1,1}, x_{\ell_1,1}] \times \dots \times [x_{\ell_n-1,n}, x_{\ell_n,n}]$, we have

$$v(R) = (x_{\ell_1,1} - x_{\ell_1-1,1}) \times \dots \times (x_{\ell_n,n} - x_{\ell_n-1,n}) = \Delta x_{\ell_1,1} \times \dots \times \Delta x_{\ell_n,n}.$$

For the sake of convenience, we will always denote the volume of R as $v(R)$.

If some bounded function f is defined over $Q \subset \mathbb{R}^n$, we can approximate the volume “underneath” f with Riemann sums.

¹⁶⁴This is a slight abuse of notation as P is a n -tuple of sets.

Definition 10.3. Suppose Q is a rectangle in \mathbb{R}^n and $f : Q \rightarrow \mathbb{R}$ is a bounded function. Let P be a partition of Q . For each subrectangle $R \in P$, define

$$M_R = \sup_{\mathbf{x} \in R} f(\mathbf{x})$$

$$m_R = \inf_{\mathbf{x} \in R} f(\mathbf{x}).$$

The *upper Riemann sum* and *lower Riemann sum* are given as

$$U(P, f) = \sum_{R \in P} M_r v(R),$$

$$L(P, f) = \sum_{R \in P} m_r v(R),$$

respectively.

Figure 109 and Figure 110 illustrate $L(P, f)$ and $U(P, f)$ for a given example. Note that by construction $L(P, f) \leq U(P, f)$, as

$$\begin{aligned} \inf_{\mathbf{x} \in R} f(\mathbf{x}) &\leq \sup_{\mathbf{x} \in R} f(\mathbf{x}) \quad (\forall R \in P), \\ \implies m_R &\leq M_r \quad (\forall R \in P), \\ \implies \sum_{R \in P} m_R &\leq \sum_{R \in P} M_r, \\ \implies \sum_{R \in P} m_R v(R) &\leq \sum_{R \in P} M_r v(R), \\ \implies L(P, f) &\leq U(P, f). \end{aligned}$$

We also know that $U(P, f)$ and $L(P, f)$ must exist. The set $\{f(\mathbf{x}) \mid \mathbf{x} \in R\} \subset \mathbb{R}$ is bounded for all $R \in P$, as f is bounded on Q . By the completeness of \mathbb{R} , each of these sets must have an infimum and supremum, so M_R and m_R exists for each R . If $L(P, f)$ and $U(P, f)$ are well defined for any P , we can take the supremum and infimum of them, respectively.

Definition 10.4. Suppose $\mathbf{P}(Q)$ is the set of all partitions of a rectangle $Q \subset \mathbb{R}^n$. If $f : Q \rightarrow \mathbb{R}$ is a bounded function, we define the *upper Riemann integral of f over Q* and *lower Riemann integral of f over Q* as:

$$\begin{aligned} \bar{\int}_Q f(\mathbf{x}) d\mathbf{x} &= \inf_{P \in \mathbf{P}(Q)} U(P, f), \\ \underline{\int}_Q f(\mathbf{x}) d\mathbf{x} &= \sup_{P \in \mathbf{P}(Q)} L(P, f). \end{aligned}$$

When these bounds coincide, we have a Riemann integral for $f : Q \rightarrow \mathbb{R}$.

Definition 10.5. Suppose f is a bounded real function on the rectangle $Q \subset \mathbb{R}^n$. If

$$\int_Q f(\mathbf{x}) d\mathbf{x} = \bar{\int}_Q f(\mathbf{x}) d\mathbf{x},$$

then f is *Riemann integrable (on Q)*, and we write the common value of the upper and lower Riemann integral as

$$\int_Q f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_Q f(x_1, \dots, x_n) d(x_1, \dots, x_n).$$

This common value is the *(multiple) Riemann integral of f (over Q)*.

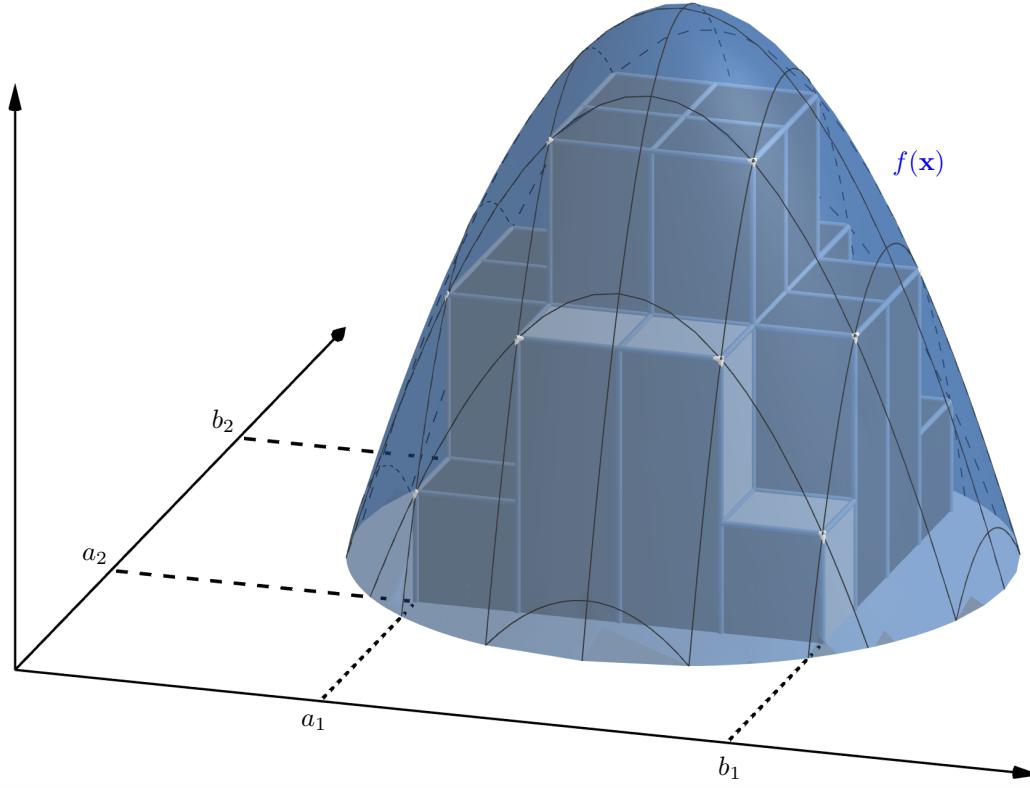


Figure 109: Suppose a partition of $Q = [a_1, b_1] \times [a_2, b_2]$ gives 16 subrectangles of equal area. The sum of the volume of the 16 prisms in gray is given by the lower Riemann sum $L(P, f)$.

In the event that $n = 2$, we call this integral the *double (Riemann) integral* of f and can write

$$\iint_Q f(\mathbf{x}) \, d\mathbf{x}.$$

Similarly, if $n = 3$ we call it the *triple (Riemann) integral* of f and can write

$$\iiint_Q f(\mathbf{x}) \, d\mathbf{x}.$$

We now turn to the question of when precisely some $f : Q \rightarrow \mathbb{R}$ is Riemann integrable for $Q \subset \mathbb{R}^n$? When we faced this question for $n = 1$, our first answer came in the form of Riemann's criterion (Theorem 6.1) which gave a necessary and sufficient condition for integrability in terms of partitions and an arbitrary $\varepsilon > 0$. We then circled back to this question and were introduced to Lebesgue's criterion (Theorem 6.8) which says a function is integrable if its discontinuities are “negligible”. Both of these criteria generalize to multiple Riemann integrals. Before presenting them, we will verify a function is Riemann integrable “from scratch” by generalizing Example 6.1.

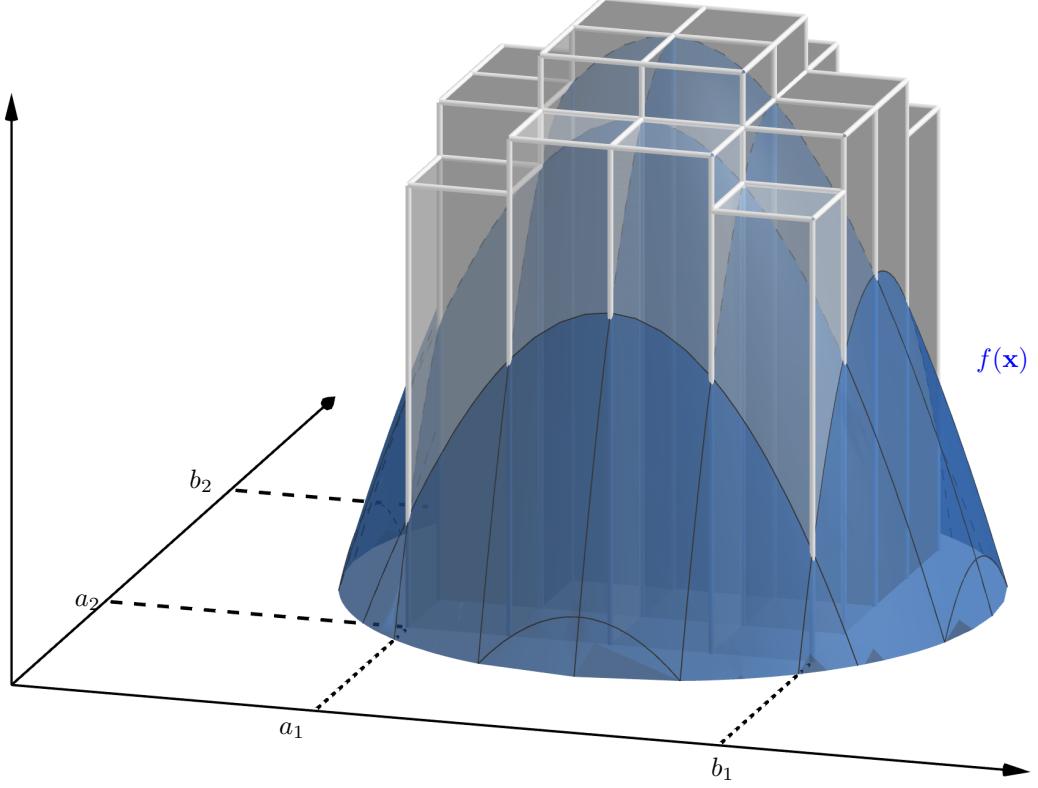


Figure 110: The upper Riemann sum counterpart to Figure 109.

Example 10.1. Suppose $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as $f(\mathbf{x}) = c$ for some constant $c \in \mathbb{R}$. Let $Q = [0, 1]^n$ be the unit hypercube in \mathbb{R}^n . Let's calculate the integral of f over Q . An arbitrary partition of Q is comprised of n one-dimensional partitions of the form $P_j = \{0, x_{1,j}, \dots, x_{k_j-1,j}, 1\}$ (where $x_{0,j} = 0$ and $x_{k_j,j} = 1$). Regardless of the partition P , f is constant over all of Q , so it will be constant over any $R \in P$. Thus,

$$M_R = \sup_{\mathbf{x} \in R} f(\mathbf{x}) = \sup\{c\} = c,$$

$$m_R = \inf_{\mathbf{x} \in R} f(\mathbf{x}) = \inf\{c\} = c.$$

Therefore,

$$\begin{aligned} M_R &= m_r \quad (\forall R \in P) \\ \implies \sum_{R \in P} M_r v(R) &= \sum_{R \in P} m_r v(R) \\ \implies U(P, f) &= L(P, f) \end{aligned}$$

We took P to be arbitrary so this holds for all $P \in \mathbf{P}(Q)$ and we have

$$\begin{aligned} \inf_{P \in \mathbf{P}(Q)} U(P, f) &= \sup_{P \in \mathbf{P}(Q)} L(P, f), \\ \implies \bar{\int}_Q f(\mathbf{x}) d\mathbf{x} &= \underline{\int}_Q f(\mathbf{x}) d\mathbf{x}, \end{aligned}$$

so f is Riemann integrable over Q .

We still need to calculate the integral. For our arbitrary $P \in \mathbf{P}(Q)$,¹⁶⁵ let's calculate the common value

$$U(P, f) = L(P, f) = \sum_{R \in P} c \cdot v(R) = c \sum_{R \in P} v(R).$$

This is a matter of calculating the sum of volumes $v(R)$, where

$$R = [x_{\ell_1-1,1}, x_{\ell_1,1}] \times [x_{\ell_2-1,2}, x_{\ell_2,2}] \times \cdots \times [x_{\ell_n-1,n}, x_{\ell_n,n}]$$

for $\ell_j \in \{1, \dots, k_j\}$ for $j = 1, \dots, n$. For a fixed $R \in P$ we have

$$v(R) = (x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n}) = \Delta x_{\ell_1,1} \times \cdots \times \Delta x_{\ell_n,n}.$$

The sum of all such $v(R)$ is

$$\sum_{R \in P} v(R) = \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \cdots \sum_{\ell_n=1}^{k_n} ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})).$$

If we focus on the innermost summation we have

$$\begin{aligned} \sum_{\ell_n=1}^{k_n} ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})) &= ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{1,n} - x_{0,n})) \\ &\quad + ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{2,n} - x_{1,n})) + \cdots \\ &\quad + ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{k_j,n} - x_{k_j-1,n})) \\ &= ((x_{1,n} - x_{0,n}) + (x_{2,n} - x_{1,n}) + \cdots + (x_{k_n,n} - x_{k_n-1,n})) \\ &\quad \times ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})) \\ &= (x_{k_n,n} - x_{0,n}) ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})). \end{aligned}$$

But $x_{k_j,n} = 1$ and $x_{0,n} = 0$, so

$$\sum_{\ell_n=1}^{k_n} ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})) = ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})),$$

and we've eliminated any reference to the index ℓ_n by showing its inclusion amounts to multiplication by 1.

We can repeat this simplification $n - 1$ more times:¹⁶⁶

¹⁶⁵It's very tempting to assert that the sum of $v(R)$ must be $v(Q) = (1 - 0)^n = 1$, as Q is the disjoint union of subrectangles R . Do we know this is how volume behaves though? We have not formally explored the properties of length/area/volume yet, and jumping to this conclusion is fairly extreme. We will return to this and related issues in Section 12.

¹⁶⁶Just looking at the algebra, this rearranging may not seem intuitive, but it actually has a very nice geometric intuition. We essentially are summing the subrectangles across a fixed dimension taking advantage of the fact that the length in any direction is 1. Try drawing a picture for $n = 2$ and working through the steps with an example!

$$\begin{aligned}
\sum_{R \in P} v(R) &= \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \cdots \sum_{\ell_n=1}^{k_n} ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})) \\
&= \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \cdots \sum_{\ell_{n-1}=1}^{k_{n-1}} 1 \cdot ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})) \\
&= \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \cdots \sum_{\ell_{n-2}=1}^{k_{n-2}} 1^2 \cdot ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-2} - x_{\ell_{n-2}-1,n-2})) \\
&\vdots \\
&= \sum_{\ell_1=1}^{k_1} 1^{n-1} (x_{\ell_1} - x_{\ell_1-1,1}) \\
&= 1^n \\
&= 1.
\end{aligned}$$

Therefore for any $P \in \mathbf{P}(Q)$,

$$U(P, f) = L(P, f) = c \sum_{R \in P} v(R) = c \cdot 1 = c.$$

The supremum and infimum of $\{c\}$ coincide, and we have

$$\int_Q f(\mathbf{x}) d\mathbf{x} = \int_{[0,1]^n} c d\mathbf{x} = c.$$

10.2 Riemann and Lebesgue's Criteria

We'll begin by building towards Riemann's criterion.

Lemma 10.1. Suppose P is a partition of the rectangle $Q \subset \mathbb{R}^n$, and $f : Q \rightarrow \mathbb{R}$ is a bounded function. If P' is a refinement of P , then $L(P, f) \leq L(P', f)$ and $U(P', f) \leq U(P, f)$.

Proof. Suppose $Q = [a_1, b_1] \times \cdots \times [a_n, b_n]$, and $P = (P_1, \dots, P_j)$ where $P_j = \{x_{0,j}, \dots, x_{k_j,j}\}$ for $j = 1, \dots, n$. It suffices to show the result when P' is a refinement of P which results from adding a single point to one P_j .¹⁶⁷ Suppose we add the point x^* to P_1 , such that $x^* \in [x_{i-1,1}, x_{i,1}]$.

$$P'_1 = \{x_{0,1}, \dots, x_{i-1,1}, x^*, x_{i,1}, \dots, x_{k_j,1}\}$$

We now have a refinement $P' = (P'_1, \dots, P_n)$.

Define

$$S = [x_{\ell_2-1,2}, x_{\ell_2,2}] \times \cdots \times [x_{\ell_n-1,n}, x_{\ell_n,n}]$$

for some arbitrary choice of intervals of (P_2, \dots, P_n) given by the indices ℓ_2, \dots, ℓ_n . A subrectangle of P or P' takes the form

$$\begin{aligned}
R &= [x_{\ell_1-1,1}, x_{\ell_1,1}] \times [x_{\ell_2-1,2}, x_{\ell_2,2}] \times \cdots \times [x_{\ell_n-1,n}, x_{\ell_n,n}], \\
&= [x_{\ell_1-1,1}, x_{\ell_1,1}] \times S.
\end{aligned}$$

¹⁶⁷If it holds in this case, then we simply apply the result iteratively until we have the desired refinement.

Because P and P' differ only with respect to P_1 , the choice of S (i.e the choice of the intervals from P_2, \dots, P_n which define R) can be fixed for now. While nearly all the subrectangles of P coincide with those of P' , we have a notable exception involving x^* . For a fixed S , P has a subrectangle of the form $R_S = [x_{i-1,1}, x_{i,1}] \times S$, whereas P' has two subrectangles $R'_S = [x_{i-1,1}x^*] \times S$ and $R''_S = [x^*, x_{i,1}] \times S$. The inclusion of x^* in P_1 splits the subrectangle R_S into two disjoint subrectangles,

$$R_S = R'_S \cup R''_S.$$

Figure 111 illustrates this in the case where $Q \subset \mathbb{R}^2$.

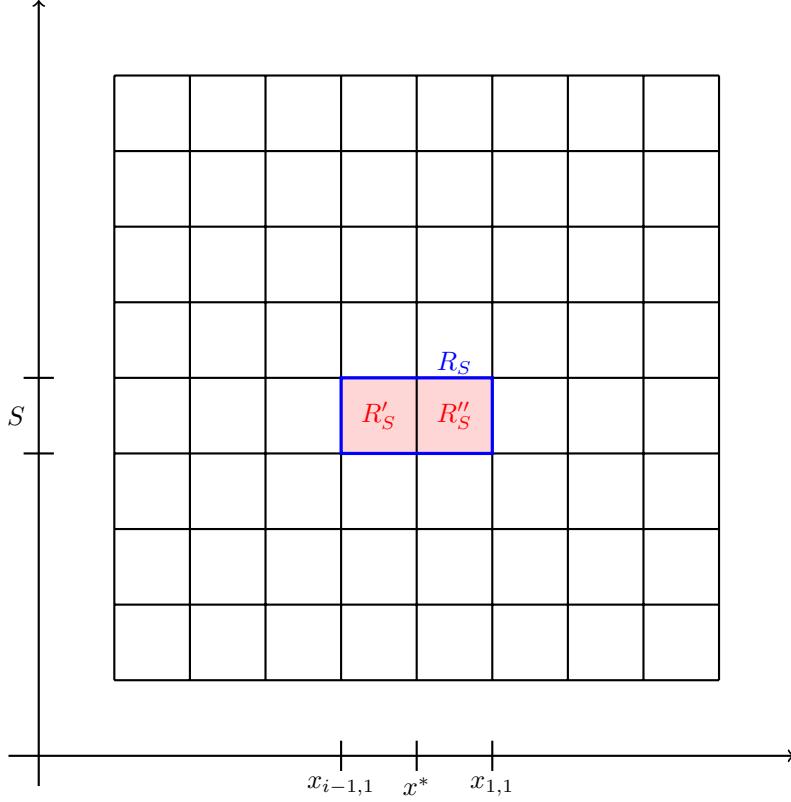


Figure 111: For a fixed subrectangle $S \subset \mathbb{R}^{n-1}$, the subrectangle $R_s = [x_{i-1,1}, x_{i,1}] \times S \in P$ is partitioned into $R'_S = [x_{i-1,1}x^*] \times S$ and $R''_S = [x^*, x_{i,1}] \times S$ when refining P with the addition of x^* to P_1 .

Since $R_s = R'_s \cup R''_s$,

$$\inf_{\mathbf{x} \in R_S} f(\mathbf{x}) \leq \inf_{\mathbf{x} \in R'_S} f(\mathbf{x}) \implies m_{R_S} \leq m_{R'_S},$$

$$\inf_{\mathbf{x} \in R_S} f(\mathbf{x}) \leq \inf_{\mathbf{x} \in R''_S} f(\mathbf{x}) \implies m_{R_S} \leq m_{R''_S},$$

$$\sup_{\mathbf{x} \in R_S} f(\mathbf{x}) \leq \sup_{\mathbf{x} \in R'_S} f(\mathbf{x}) \implies M_{R_S} \geq M_{R'_S},$$

$$\sup_{\mathbf{x} \in R_S} f(\mathbf{x}) \leq \sup_{\mathbf{x} \in R''_S} f(\mathbf{x}) \implies M_{R_S} \geq M_{R''_S}.$$

Using these inequalities, we get

$$\begin{aligned}
m_{R_S} v(R_s) &= m_{R_S} ((x_{i,1} - x_{i-1,1})v(S)] \\
&= m_{R_S} ((x_{i,1} - x_{i-1,1} + (x^* - x^*))v(S)] \quad (x^* - x^* = 0) \\
&= m_{R_S} (((x_{i,1} - x^*) + (x^* - x_{i-1,1}))v(S)] \\
&= m_{R_S} ((x_{i,1} - x^*)v(S) + (x^* - x_{i-1,1})v(S)] \\
&= m_{R_S} (v(R'_S) + v(R''_S)) \\
&= m_{R_S} v(R'_S) + m_{R_S} v(R''_S) \\
&\leq m_{R'_S} v(R'_S) + m_{R''_S} v(R''_S) \quad (m_{R_S} \leq m_{R'_S} \text{ and } m_{R_S} \leq m_{R''_S}), \\
M_{R_S} v(R_s) &= M_{R_S} ((x_{i,1} - x_{i-1,1})v(S)] \\
&\vdots \\
M_{R_S} v(R_s) &\geq M_{R'_S} v(R'_S) + M_{R''_S} v(R''_S)
\end{aligned}$$

These inequalities hold for all S , and summing over the subrectangles in each partition gives $L(P, f) \leq L(P', f)$ and $U(P', f) \leq U(P, f)$. \square

Lemma 10.2. Suppose f is a bounded real function on the rectangle $Q \subset \mathbb{R}^n$. Then

$$\underline{\int}_Q f(\mathbf{x}) d\mathbf{x} \leq \bar{\int}_Q f(\mathbf{x}) d\mathbf{x}.$$

Proof. Suppose P_1 and P_2 partition Q , and let P^* be a common refinement of P_1 and P_2 . By Lemma 10.1 and $L(P^*, f) \leq U(P^*, f)$,

$$L(P_1, f) \leq L(P^*, f) \leq U(P^*, f) \leq U(P_2, f).$$

If we fix P_1 and P_2 we can take the supremum and infimum over $\mathbf{P}(Q)$.

$$\begin{aligned}
\sup_{P_1 \in \mathbf{P}(Q)} L(P_1, f) &\leq \sup_{P_2 \in \mathbf{P}(Q)} U(P_2, f) \\
\implies \underline{\int}_Q f(\mathbf{x}) d\mathbf{x} &\leq \bar{\int}_Q f(\mathbf{x}) d\mathbf{x}
\end{aligned}$$

\square

Theorem 10.1 (Riemann's Criterion). Suppose f is a bounded real function on the rectangle $Q \subset \mathbb{R}^n$. Then f is Riemann integrable on Q if and only if for all $\varepsilon > 0$, there exists a partition $P \in \mathbf{P}(Q)$ such that

$$U(P, f) - L(P, f) < \varepsilon.$$

Proof.

(\Rightarrow) Suppose f is Riemann integrable, and let $\varepsilon > 0$. By the completeness of the real numbers and properties of the supremum, there exists a partition P_1 such that

$$\begin{aligned}
U(P_1, f) - \sup_{P \in \mathbf{P}([a,b])} U(P, f) &< \frac{\varepsilon}{2} \\
\implies U(P_1, f) - \bar{\int}_Q f(\mathbf{x}) d\mathbf{x} &< \frac{\varepsilon}{2} \quad (\text{definition 10.4}) \\
\implies U(P_1, f) - \underline{\int}_Q f(\mathbf{x}) d\mathbf{x} &< \frac{\varepsilon}{2} \quad \left(\bar{\int}_Q f(\mathbf{x}) d\mathbf{x} \geq \underline{\int}_Q f(\mathbf{x}) d\mathbf{x} \right)
\end{aligned}$$

Similarly, there exists a partition P_2 such that

$$\int_Q f(\mathbf{x}) \, d\mathbf{x} - L(P_2, f) < \frac{\varepsilon}{2}.$$

If we choose P^* to be the common refinement of P_1 and P_2 , then by Lemma 6.1,

$$\begin{aligned} U(P^*, f) - \int_Q f(\mathbf{x}) \, d\mathbf{x} &< U(P_1, f) - \int_Q f(\mathbf{x}) \, d\mathbf{x} < \frac{\varepsilon}{2}, \\ \int_Q f(\mathbf{x}) \, d\mathbf{x} - L(P^*, f) &< \int_Q f(\mathbf{x}) \, d\mathbf{x} - L(P_2, f) < \frac{\varepsilon}{2}. \end{aligned}$$

Combining these two inequalities for P^* yields

$$\begin{aligned} U(P^*, f) - L(P^*, f) &= U(P^*, f) - L(P^*, f) + 0 \\ &= U(P^*, f) - L(P^*, f) + \left(- \int_Q f(\mathbf{x}) \, d\mathbf{x} + \int_Q f(\mathbf{x}) \, d\mathbf{x} \right) \\ &= \left(U(P^*, f) - \int_Q f(\mathbf{x}) \, d\mathbf{x} \right) + \left(\int_Q f(\mathbf{x}) \, d\mathbf{x} - L(P^*, f) \right) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon. \end{aligned}$$

(\Leftarrow) Suppose for all $\varepsilon > 0$ there exists a partition such that $U(P, f) - L(P, f) < \varepsilon$. By the definition of supremum and infimum we have

$$\begin{aligned} L(P, f) &\leq \sup_{P \in \mathbf{P}(Q)} U(P, f) = \int_Q f(\mathbf{x}) \, d\mathbf{x}, \\ \int_a^b f(\mathbf{x}) \, d\mathbf{x} &\leq \inf_{P \in \mathbf{P}(Q)} U(P, f) \leq U(P, f), \end{aligned}$$

for all $P \in \mathbf{P}(Q)$. Combining this inequalities with Lemma 6.2 gives

$$L(P, f) \leq \int_Q f(\mathbf{x}) \, d\mathbf{x} \leq \bar{\int}_Q f(\mathbf{x}) \, d\mathbf{x} \leq U(P, f).$$

This implies that

$$0 \leq \int_Q f(\mathbf{x}) \, d\mathbf{x} - \bar{\int}_Q f(\mathbf{x}) \, d\mathbf{x} \leq \varepsilon,$$

but if this holds for all $\varepsilon > 0$,

$$\int_Q f(\mathbf{x}) \, d\mathbf{x} = \bar{\int}_Q f(\mathbf{x}) \, d\mathbf{x}.$$

Therefore f is Riemann integrable. □

The proof of Riemann's condition for multiple Riemann integrals is virtually the same as the proof for Riemann integrals in \mathbb{R} . The big difference comes in the amount of effort we need to put in to use Riemann's criterion. Working with partitions in \mathbb{R}^n is exponentially more difficult than working with partitions in \mathbb{R} . Instead it is much easier to work with the general version of Lebesgue's criterion, which is stated in terms of discontinuities instead of partitions.

Definition 10.6. Suppose $A \subset \mathbb{R}^n$. The set A is a *null set (in \mathbb{R}^n)* if for every $\varepsilon > 0$ there exists an open cover $\{U_i\} \subset \mathbb{R}$, where $U_i = (a_{1,i}, b_{1,i}) \times \cdots \times (a_{n,i}, b_{n,i})$ is an open rectangle, such that

$$\sum_{i=1}^{\infty} (b_{1,i} - a_{1,i}) \times \cdots \times (b_{n,i} - a_{n,i}) = \sum_{i=1}^{\infty} v(U_i) < \varepsilon.$$

Much like Lemma 10.2 and Theorem 10.1, the proof for Lebesgue's criterion for multiple integrals is nearly line for line the same as it was for integrals of single variable functions. As such, I'll just state the theorem, and omit the proof.

Theorem 10.2 (Lebesgue's Criterion). Suppose f is a real bounded function defined on a rectangle $Q \subset \mathbb{R}^n$. Then, f is Riemann integrable *if and only if* the set of discontinuities of f on Q is a null set.

Lebesgue's criterion allows us to immediately conclude that most of the functions of several variables we encounter, namely continuous ones, in practice are in fact Riemann integrable.

Corollary 10.1 (Continuous \implies Riemann Integrable). Suppose f is a real bounded function defined on a rectangle $Q \subset \mathbb{R}^n$. If f is continuous on Q , then it is Riemann integrable.

10.3 Iterated Integrals and Fubini's Theorem

When we first saw Riemann integrals, spent quite a bit of time exploring the properties of integrals before addressing the elephant in the room that was calculating them. This of course is done quite readily using the fundamental theorem of calculus (Theorem 6.6). Alas, there is no direct generalization of this theorem that would allow us to calculate

$$\int_Q f(\mathbf{x}) \, d\mathbf{x}.$$

Perhaps we can still find someway to make the fundamental theorem of calculus work for us. Let's return to the example of a constant function on $[0, 1]^n$ (Example 10.1).

Example 10.2. Define $f : Q \rightarrow \mathbb{R}$, where $Q = [0, 1]^n$, as $f(\mathbf{x}) = c$. Recall from Example 6.1, that if $n = 1$,

$$\int_Q f(\mathbf{x}) \, d\mathbf{x} = \int_0^1 c \, dx = \sum_{i=1}^n c \cdot \Delta x_i = c,$$

where $P = \{x_0, \dots, x_n\}$ is a partition of $[0, 1]$. In general, we saw that

$$\int_Q f(\mathbf{x}) \, d\mathbf{x} = \sum_{R \in P} c \cdot v(R) = \sum_{\ell_1=1}^{k_1} \sum_{\ell_2=1}^{k_2} \cdots \sum_{\ell_n=1}^{k_n} c((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})) = c.$$

We simplified these nested summations by simplifying each one starting with the innermost one and working our way back. Let's revisit this process. First let's reverse the order of the summations,

$$\int_Q f(\mathbf{x}) \, d\mathbf{x} = \sum_{\ell_n=1}^{k_n} \sum_{\ell_{n-1}=1}^{k_{n-1}} \cdots \sum_{\ell_1=1}^{k_1} c((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_n-1,n})),$$

as the order of the summations will have no effect on the value. Focusing on the inner most summation, we

have

$$\begin{aligned}
\sum_{\ell_1=1}^{k_1} c((x_{1,1} - x_{0,1}) \times \cdots \times (x_{k_1,n} - x_{k_1-1,n})) &= c((x_{1,1} - x_{0,1}) + (x_{2,1} - x_{1,1}) + \cdots + (x_{k_1,1} - x_{k_1-1,1})) \\
&\quad \times ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})) \\
&= \left(\sum_{i=1}^{k_1} c \cdot \Delta x_{i,1} \right) ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})) \\
&= \left(\int_0^1 c \, dx_1 \right) ((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_{n-1},n-1} - x_{\ell_{n-1}-1,n-1})).
\end{aligned}$$

By focusing on the inner summation first, we restrict our attention to the one dimensional partition P_1 (an element of P), and can express the value in terms of a Riemann integral in one dimension. This eliminates the index ℓ_1 , while simultaneously relating this step to an integral in the dimension which ℓ_1 indexes over. Like before we can repeat this step for the remaining summations, working our way backwards.

$$\begin{aligned}
\int_Q f(\mathbf{x}) \, d\mathbf{x} &= \sum_{\ell_n=1}^{k_n} \sum_{\ell_{n-1}=1}^{k_{n-1}} \cdots \sum_{\ell_1=1}^{k_1} c((x_{\ell_1,1} - x_{\ell_1-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_{n-1},n})) \\
&= \sum_{\ell_n=1}^{k_n} \sum_{\ell_{n-1}=1}^{k_{n-1}} \cdots \sum_{\ell_2=1}^{k_1} \left(\int_0^1 c \, dx_1 \right) ((x_{\ell_2,1} - x_{\ell_2-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_{n-1},n})) \\
&= \sum_{\ell_n=1}^{k_n} \sum_{\ell_{n-1}=1}^{k_{n-1}} \cdots \sum_{\ell_3=1}^{k_1} \left(\int_0^1 \left(\int_0^1 c \, dx_1 \right) \, dx_2 \right) ((x_{\ell_3,1} - x_{\ell_3-1,1}) \times \cdots \times (x_{\ell_n,n} - x_{\ell_{n-1},n})) \\
&\quad \vdots \\
&= \int_0^1 \int_0^1 \cdots \int_0^1 c \, dx_1 \, dx_2 \, \cdots \, dx_n
\end{aligned}$$

Remarkably, we have

$$\int_{[0,1]^n} f(\mathbf{x}) \, d\mathbf{x} = \int_0^1 \int_0^1 \cdots \int_0^1 f(\mathbf{x}) \, dx_1 \, dx_2 \, \cdots \, dx_n.$$

Integrating f over Q with respect to \mathbf{x} is a matter of integrating in one dimension iteratively over Q projected to the i -th dimension with respect to x_i ! Furthermore, the order of our summations when calculating $v(R)$ was moot, so we can integrate over each x_i separately in any order we want. In other words, if $\mathbf{y} = (x_1, \dots, x_{i-1}, x_{i+1}, x_n)$, then

$$\int_{[0,1]^n} f(\mathbf{x}) \, d\mathbf{x} = \int_{[0,1]^{n-1}} \int_0^1 f(\mathbf{x}) \, dx_i \, d\mathbf{y}$$

for all i . These two equalities are reminiscent of calculating volume. If we have a cube in \mathbb{R}^3 , we have

$$V = \text{length} \times \text{width} \times \text{height}.$$

Length, width, and height are all just measures of volume in one dimension. We also can find their product in any order, as multiplication is commutative. The integral $\int_Q f(\mathbf{x}) \, d\mathbf{x}$ is a volume in \mathbb{R}^n .

A fair criticism of this example is that $f(\mathbf{x})$ is a constant. This makes the example trivial in a sense, as $f(\mathbf{x})$ does not depend on any elements of \mathbf{x} . Perhaps our findings from Example 10.1 do not generalize to non-trivial functions. Our next theorem addresses this, and specifies under what conditions we can decompose a multiple integral into separate integrals which can be calculated using the fundamental theorem of calculus.

Theorem 10.3 (Fubini's Theorem for Riemann Integrals, General Version). Suppose $Q = A \times B$ is a rectangle in \mathbb{R}^{n+m} , where A is a rectangle in \mathbb{R}^n and B is a rectangle in \mathbb{R}^m . Let $f : Q \rightarrow \mathbb{R}$ be a bounded function written in the form $f(\mathbf{x}, \mathbf{y})$ for $\mathbf{x} \in A$ and $\mathbf{y} \in B$. If f is integrable over Q , then $\underline{\int}_B f(\mathbf{x}, \mathbf{y})$ and $\bar{\int}_B f(\mathbf{x}, \mathbf{y})$ (which are functions of \mathbf{x}) are integrable over A with respect to \mathbf{x} . Furthermore the integral of f over Q can be expressed as two *iterated integrals*:

$$\int_Q f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) = \int_A \underline{\int}_B f(\mathbf{x}, \mathbf{y}) dy d\mathbf{x} = \int_A \bar{\int}_B f(\mathbf{x}, \mathbf{y}) dy d\mathbf{x}$$

Proof. Assume that f is integrable over Q . For the sake of convenience, we will write:

$$\begin{aligned}\underline{I}(\mathbf{x}) &= \underline{\int}_B f(\mathbf{x}, \mathbf{y}) dy, \\ \bar{I}(\mathbf{x}) &= \bar{\int}_B f(\mathbf{x}, \mathbf{y}) dy.\end{aligned}$$

Let P be an arbitrary partition of Q . This partition can be written as $P = (P_A, P_B)$ where P_A and P_B are partitions of A and B , respectively. Note that any subrectangle of P takes the form $R_A \times R_B$ for subrectangles $R_A \in P_A$ and $R_B \in P_B$.

1. Show that $L(P, f) \leq L(P_A, \underline{I})$.
2. Show that $U(P, f) \geq U(P_A, \bar{I})$.
3. Combine inequalities from Step 1 and Step 2 to reach two inequalities.
4. Use the two derived inequalities from Step 3 to prove that \bar{I} and \underline{I} are integrable.
5. Verify that

$$\int_Q f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) = \int_A \underline{I}(\mathbf{x}) d\mathbf{x} = \int_A \bar{I}(\mathbf{x}) d\mathbf{x}.$$

Step 1. Consider an arbitrary subrectangle $R = R_A \times R_B \in P$. For some $\mathbf{x}_0 \in R_A$,

$$m_R = \inf_{(\mathbf{x}, \mathbf{y}) \in R_A \times R_B} f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}_0, \mathbf{y}) \quad (\forall \mathbf{y} \in R_B).$$

If this inequality holds for all $\mathbf{y} \in R_B$, then it must hold for the value of \mathbf{y} which gives the “smallest” value of $f(\mathbf{x}_0, \mathbf{y})$.

$$\begin{aligned}m_R &\leq f(\mathbf{x}_0, \mathbf{y}) \quad (\forall \mathbf{y} \in R_B) \\ \implies m_R &\leq \inf_{\mathbf{y} \in R_B} f(\mathbf{x}_0, \mathbf{y}) \\ \implies m_R(f) &\leq m_{R_B}(f(\mathbf{x}_0, \mathbf{y})).\end{aligned}$$

The bound $m_R(f)$ and $m_{R_B}(f(\mathbf{x}_0, \mathbf{y}))$ are expressed as functions to eliminate any confusion as to which function they refer to. Keep in mind that \mathbf{x}_0 is fixed, so $f(\mathbf{x}_0, \mathbf{y})$ is a function of \mathbf{y} , whereas f depends

on (\mathbf{x}, \mathbf{y}) . We can multiple this inequality by $v(R_B)$, and then sum over all $R_B \in P_B$.

$$\begin{aligned}
& m_R(f) \leq m_{R_B}(f(\mathbf{x}_0, \mathbf{y})) \\
\implies & m_R(f)v(R_B) \leq m_{R_B}(f(\mathbf{x}_0, \mathbf{y}))v(R_B) \quad (\forall v(R_B) \in P_B) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \underbrace{\sum_{R_B \in P_B} m_{R_B}(f(\mathbf{x}_0, \mathbf{y}))v(R_B)}_{L(P_B, f(\mathbf{x}_0, \mathbf{y}))} \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq L(P_B, f(\mathbf{x}_0, \mathbf{y})) \quad (\text{Definition 10.3})
\end{aligned}$$

We took $P = (P_A, P_B)$ to be an arbitrary partition of Q , so this inequality holds for all $P_B \in \mathbf{P}(B)$.

$$\begin{aligned}
& \sum_{R_B \in P_B} m_R(f)v(R_B) \leq L(P_B, f(\mathbf{x}_0, \mathbf{y})) \quad (\forall P_B \in \mathbf{P}(B)) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \sup_{P_B \in \mathbf{P}(B)} L(P_B, f(\mathbf{x}_0, \mathbf{y})) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \int_B f(\mathbf{x}_0, \mathbf{y}) d\mathbf{y} \quad (\text{Definition 10.4}) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \underline{I}(\mathbf{x}_0) \quad (\text{Definition of } \underline{I}(\mathbf{x}_0))
\end{aligned}$$

We took $\mathbf{x}_0 \in \mathbb{R}_A$ to be completely arbitrary here, so this inequality will hold for every $\mathbf{x}_0 \in R_A$.

$$\begin{aligned}
& \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \underline{I}(\mathbf{x}) \quad (\forall \mathbf{x} \in R_A) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq \underbrace{\inf_{\mathbf{x} \in R_A} \underline{I}(\mathbf{x})}_{m_{R_A}(\underline{I})} \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B) \leq m_{R_A}(\underline{I}) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(R_B)v(R_A) \leq m_{R_A}(\underline{I})v(R_A) \\
\implies & \sum_{R_B \in P_B} m_R(f)v(Q) \leq m_{R_A}(\underline{I})v(R_A) \quad (v(R_A)v(R_B) = V(R_A \times R_B) = V(Q))
\end{aligned}$$

The subrectangle R_A is completely arbitrary, so this holds for all R_A .

$$\begin{aligned}
& \sum_{R_B \in P_B} m_R(f)v(Q) \leq m_{R_A}(\underline{I})v(R_A) \quad (\forall R_A \in P_A) \\
\implies & \sum_{R_A \in P_A} \sum_{R_B \in P_B} m_R(f)v(Q) \leq \sum_{R_A \in P_A} m_{R_A}(\underline{I})v(R_A) \\
\implies & \sum_{R \in Q} m_R(f)v(Q) \leq \sum_{R_A \in P_A} m_{R_A}(\underline{I})v(R_A) \quad (P = P_A \times P_B) \\
\implies & L(P, f) \leq L(P_A, \underline{I}) \quad (\text{Definition 10.3})
\end{aligned}$$

Step 2. We now repeat Step 1, but with upper limits and the inequalities reversed.

$$\begin{aligned}
& M_R(f) \geq M_{R_B}(f(\mathbf{x}_0, \mathbf{y})) \\
\implies & M_r(f)v(R_B) \geq M_{R_B}(f(\mathbf{x}_0, \mathbf{y}))v(R_B) \quad (\forall R_B \in P_B) \\
\implies & \sum_{R_B \in P_B} M_r(f)v(R_B) \geq \sum_{R_B \in P_B} M_{R_B}(f(\mathbf{x}_0, \mathbf{y}))v(R_B) \quad (\forall P_B \in \mathbf{P}(B)) \\
\implies & \sum_{R_B \in P_B} M_r(f)v(R_B) \geq \bar{I}(\mathbf{x}) \quad (\forall \mathbf{x} \in R_A) \\
\implies & \sum_{R_B \in P_B} M_r(f)v(R_B) \geq M_{R_A}(\bar{I}) \\
\implies & \sum_{R_B \in P_B} M_r(f)v(Q) \geq M_{R_A}(\bar{I})v(R_A) \quad (\forall R_A \in P_A) \\
\implies & \sum_{R \in P} M_r(f)v(Q) \geq \sum_{R_A \in P_A} M_{R_A}(\bar{I})v(R_A) \\
\implies & U(P, f) \geq U(P_A, \bar{I})
\end{aligned}$$

Step 3. Lemma 10.2 asserts that:

$$\begin{aligned}
& \underline{I}(\mathbf{x}) \leq \bar{I}(\mathbf{x}) \quad (\forall \mathbf{x} \in A) \\
\implies & \inf_{\mathbf{x} \in R_A} \underline{I}(\mathbf{x}) \leq \inf_{\mathbf{x} \in R_A} \bar{I}(\mathbf{x}) \quad (\forall R_A \in P_A) \\
\implies & m_{R_A}(\underline{I}) \leq m_{R_A}(\bar{I}) \quad (\text{Definition 10.3}) \\
\implies & \sum_{R_A \in P_A} m_{R_A}(\underline{I}) \leq \sum_{R_A \in P_A} m_{R_A}(\bar{I}) \\
\implies & L(P_A, \underline{I}) \leq L(P_A, \bar{I}) \quad (\text{Definition 10.3}) \\
\implies & L(P, f) \leq L(P_A, \underline{I}) \leq L(P_A, \bar{I}) \quad (\text{Step 1}) \\
\implies & L(P, f) \leq L(P_A, \underline{I}) \leq L(P_A, \bar{I}) \leq U(P_A, \bar{I}) \quad (m_{R_A} \leq M_{R_A}) \\
\implies & L(P, f) \leq L(P_A, \underline{I}) \leq L(P_A, \bar{I}) \leq U(P_A, \bar{I}) \leq U(P_A, I) \quad (\text{Step 2}).
\end{aligned}$$

This same train of thought can be applied to conclude $U(P_A, \underline{I}) \leq U(P_A, \bar{I})$, which then gives

$$L(P, f) \leq L(P_A, \underline{I}) \leq U(P_A, \underline{I}) \leq U(P_A, \bar{I}) \leq U(P_A, I).$$

We therefore have

$$L(P, f) \leq L(P_A, \underline{I}) \leq L(P_A, \bar{I}) \leq U(P_A, \bar{I}) \leq U(P, f) \quad (60)$$

$$L(P, f) \leq L(P_A, \underline{I}) \leq U(P_A, \underline{I}) \leq U(P_A, \bar{I}) \leq U(P, f). \quad (61)$$

Note that (60) and (61) do not compare $L(P_A, \bar{I})$ and $U(P_A, \underline{I})$. It's ambiguous as to which of these values is larger.

Step 4. We have assumed f is integrable over Q , so Riemann's condition tells us that for all $\varepsilon > 0$ (Theorem 10.1) there exists some $P = (P_A, P_B)$ such that $U(P, f) - L(P, f) < \varepsilon$. But (60) implies that

$$U(P_A, \bar{I}) - U(P_A, \underline{I}) \leq U(P, f) - L(P, f),$$

so $U(P_A, \bar{I}) - U(P_A, \underline{I}) < \varepsilon$. Thus for all ε , Riemann's condition is satisfied by P_A , and $\bar{I}(\mathbf{x})$ is integrable. Similarly we can use (61) to show that $U(P_A, \underline{I}) - U(P_A, \bar{I}) < \varepsilon$, implying that $\underline{I}(\mathbf{x})$ is integrable.

Step 5. By Definition 10.4 and Definition 10.5:

$$L(P_A, \bar{I}) \leq \int_A \bar{I} \, d\mathbf{x} \leq U(P_A, \bar{I}), \quad (62)$$

$$L(P_A, \underline{I}) \leq \int_A \underline{I} \, d\mathbf{x} \leq U(P_A, \underline{I}), \quad (63)$$

$$L(P, f) \leq \int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \leq U(P, f), \quad (64)$$

for all $P = (P_A, P_B)$. Combining (62) and (60) give

$$\begin{aligned} L(P, f) &\leq L(P_A, \bar{I}) \leq \int_A \bar{I} \, d\mathbf{x} \leq U(P_A, \bar{I}) \leq U(P, f) \\ \implies L(P, f) &\leq \int_A \bar{I} \, d\mathbf{x} \leq U(P, f). \end{aligned} \quad (65)$$

If we combine (65) and (64), we have

$$\left| \int_A \bar{I} \, d\mathbf{x} - \int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \right| \leq U(P, f) - L(P, f).$$

For all ε , we can always find a P such that $U(P, f) - L(P, f) < \varepsilon$, so

$$\left| \int_A \bar{I} \, d\mathbf{x} - \int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \right| < \varepsilon$$

for all ε . This means that

$$\int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} = \int_A \bar{I} \, d\mathbf{x},$$

as desired. Similarly, combining (63) and (61) allows us to conclude

$$\int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} = \int_A \underline{I} \, d\mathbf{x}.$$

By the definition of \underline{I} and \bar{I} , we have

$$\int_Q f(\mathbf{x}, \mathbf{y}) \, d(\mathbf{x}, \mathbf{y}) = \int_A \int_B f(\mathbf{x}, \mathbf{y}) \, dy \, dx = \int_A \int_B \bar{f}(\mathbf{x}, \mathbf{y}) \, dy \, dx.$$

□

Example 10.3 (Interchanging Order of Integration). The decision to label the variable \mathbf{y} and \mathbf{x} in Theorem 10.3 is an arbitrary decision, so it also applies when integrating with respect to \mathbf{x} first. That is, if \mathbf{f} is integrable over Q , then $\int_Q f(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}$ and $\int_Q \bar{f}(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}$ exist, and

$$\int_Q f(\mathbf{x}, \mathbf{y}) \, d(\mathbf{x}, \mathbf{y}) = \int_B \int_A f(\mathbf{x}, \mathbf{y}) \, dx \, dy = \int_B \int_A \bar{f}(\mathbf{x}, \mathbf{y}) \, dx \, dy.$$

In sum, Fubini's theorem gives

$$\int_Q f(\mathbf{x}, \mathbf{y}) \, d(\mathbf{x}, \mathbf{y}) = \int_B \int_{\underline{A}}^{\bar{A}} f(\mathbf{x}, \mathbf{y}) \, dx \, dy = \int_B \int_A \bar{f}(\mathbf{x}, \mathbf{y}) \, dx \, dy = \int_A \int_{\underline{B}}^{\bar{B}} f(\mathbf{x}, \mathbf{y}) \, dy \, dx = \int_A \int_B \bar{f}(\mathbf{x}, \mathbf{y}) \, dy \, dx.$$

Example 10.4. Let $Q = [0, 1] \times [0, 1]$ and define $f : Q \rightarrow \mathbb{R}$ as

$$f(x, y) = \begin{cases} 1 & x \in \mathbb{R} \setminus \mathbb{Q} \\ 1 & y \in \mathbb{R} \setminus \mathbb{Q} \\ 1 - 1/q & x = p/q, y \in \mathbb{Q} \end{cases}.$$

If x and/or y are irrational, then $f(x, y) = 0$. If both x and y are rational then $f(x, y) = 1 - 1/q$ where $x = p/q$ is in lowest terms. The function is constant everywhere except on the countable set \mathbb{Q}^2 , so the set of discontinuities of f on \mathbb{Q} is a null set. By Lebesgue's criterion (Theorem 10.2), f is integrable. Before using Fubini's theorem, let's integrate f over Q using Definition 10.4.

Let P be a partition of $Q = [0, 1] \times [0, 1]$. Any subrectangle $R \in P$ contains an irrational number, so

$$M_R = \sup_{\mathbf{x} \in R} f(\mathbf{x}) = 1,$$

giving $U(P, f) = 1$ for any partition P . Definition 10.4 gives

$$\int_{[0,1]^2} f(x, y) d(x, y) = \bar{\int}_{[0,1]^2} f(x, y) d(x, y) = \inf_{P \in \mathbf{P}(Q)} U(P, f) = 1.$$

Now let's verify Fubini's theorem. In the event that $\mathbf{x} \in \mathbb{R} \setminus \mathbb{Q}$, $f(x, y) = 1$ for all $y \in \mathbb{R}$, so

$$\bar{\int}_0^1 f(x, y) dy = \underline{\int}_0^1 f(x, y) dy = 1 \quad (x \in \mathbb{R} \setminus \mathbb{Q}).$$

Instead, suppose $x = p/q \in \mathbb{Q}$. Any partition contains an irrational number and a rational number (\mathbb{Q} is dense in \mathbb{R}), so

$$\begin{aligned} \bar{\int}_0^1 f(x, y) dy &= 1 && (x \in \mathbb{Q}) \\ \underline{\int}_0^1 f(x, y) dy &= 1 - 1/q && (x \in \mathbb{Q}) \end{aligned}$$

We have

$$\begin{aligned} \bar{\int}_0^1 f(x, y) dy &= 1 \\ \underline{\int}_0^1 f(x, y) dy &= \begin{cases} 1 & x \in \mathbb{R} \setminus \mathbb{Q} \\ 1 - 1/q & x = p/q \in \mathbb{Q} \end{cases} \end{aligned}$$

Both of which are indeed integrable. Note that Fubini's theorem only says these upper and lower integrals are integrable functions. It does not say that $\int_B f(x, y) d(x, y)$ exists! In fact, for our example $\int_B f(x, y) d(x, y)$ fails to exist, as the upper and lower integrals do not agree when $x \in \mathbb{Q}$. Lastly, let's verify that the iterated integrals equal the integral of f over Q . Integrating the upper integral with respect to x gives

$$\int_{[0,1]^2} f(\mathbf{x}, \mathbf{y}) d(x, y) = \int_0^1 \bar{\int}_0^1 f(x, y) dy dx = \int_0^1 1 dx = 1.$$

Integrating the lower integral with respect to x is similar to our original calculation of $\int_Q f(x, y) d(x, y)$. Any interval of a partition will contain an irrational number, so

$$M_R = \sup_{x \in R} f(x) = \sup_{x \in \mathbb{R}} \int_0^1 f(x, y) dy = 1,$$

and $U(P, f) = 1$ for all P . Thus

$$\int_0^1 \int_0^1 f(x, y) \, dy \, dx = 1,$$

which implies

$$\int_0^1 \int_0^1 f(x, y) \, dy \, dx = 1$$

because we already verified $\underline{\int} f(x, y) \, dy$ is integrable.

Remark 10.1 ($\int g = \int h \not\Rightarrow g = h$). Example 10.3 highlights one of the more subtle results from Fubini's theorem. Suppose g and h are two integrable functions on $[a, b]$. If $g = h$ on $[a, b]$, then

$$\int_a^b g(x) \, dx = \int_a^b h(x) \, dx.$$

Example 10.3 shows that the converse is not true. Define

$$g(x) = \int_0^1 f(x, y) \, dy = 1$$

$$h(x) = \int_0^1 f(x, y) \, dy = \begin{cases} 1 & x \in \mathbb{R} \setminus \mathbb{Q} \\ 1 - 1/q & x = p/q \in \mathbb{Q} \end{cases}$$

Despite $g(x) \neq h(x)$ on $[0, 1]$, we have

$$\int_0^1 g(x) \, dx = \int_0^1 h(x) \, dx$$

by Fubini's theorem. In general, Fubini's theorem says that the upper and lower integral of $f(\mathbf{x}, \mathbf{y})$ with respect to \mathbf{y} are “close enough” (but not necessarily equal) that integrating them over \mathbf{x} gives the same quantity:

$$\int_A \underline{\int}_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \, d\mathbf{x} = \int_A \bar{\int}_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \, d\mathbf{x}.$$

This raises the question, how do we know if two functions are “close enough” to each other on an interval to establish the equality of their integrals over that interval? Example 10.3 actually gives us a bit of a hint about this. The functions $g(x)$ and $h(x)$ are equal as long as $x \notin \mathbb{Q}$. The set of points at which they differ is

$$\{x \in [0, 1] \mid g(x) \neq h(x)\} = \mathbb{Q} \cap [0, 1].$$

The set $\mathbb{Q} \cap [0, 1]$ is a null set, so it is “negligible” as far as integration is concerned. This is no coincidence, and we will formalize this in Section 13.

A special case of Fubini's theorem arises when

$$\int_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} = \bar{\int}_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y}.$$

If two functions are equal, then they have the same integral, so

$$\int_A \underline{\int}_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \, d\mathbf{x} = \int_A \bar{\int}_B f(\mathbf{x}, \mathbf{y}) \, d\mathbf{y} \, d\mathbf{x}.$$

But $\underline{\int}_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \bar{\int}_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ is the definition of $f(\mathbf{x}, \mathbf{y})$ being integrable over B for all $\mathbf{x} \in A$, so

$$\begin{aligned} \int_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} &= \underline{\int}_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \bar{\int}_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} \\ \implies \int_Q f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) &= \int_A \int_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x}. \end{aligned}$$

The addition of the assumption that $f(\mathbf{x}, \mathbf{y})$ is integrable over B for all $\mathbf{x} \in A$ simplifies Fubini's theorem considerably, and eliminates many knife edge cases such as Example 10.3 (albeit at the cost of some generality).

Corollary 10.2 (Fubini's Theorem for Riemann Integrals, “Nicer” Version). Suppose $Q = A \times B$ is a rectangle in \mathbb{R}^{n+m} , where A is a rectangle in \mathbb{R}^n and B is a rectangle in \mathbb{R}^m . Let $f : Q \rightarrow \mathbb{R}$ be a bounded function written in the form $f(\mathbf{x}, \mathbf{y})$ for $\mathbf{x} \in A$ and $\mathbf{y} \in B$. If f is integrable over Q and $f(\mathbf{x}, \mathbf{y})$ is integrable over B for all $\mathbf{x} \in A$, then

$$\int_Q f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) = \int_A \int_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x}.$$

If we also have that $f(\mathbf{x}, \mathbf{y})$ is integrable over A for all $\mathbf{y} \in B$, then

$$\int_Q f(\mathbf{x}, \mathbf{y}) d(\mathbf{x}, \mathbf{y}) = \int_B \int_A f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \int_A \int_B f(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x}$$

Example 10.5 (Thomae's Function Generalized). Let $Q = [0, 1]^2$, and define $f : Q \rightarrow \mathbb{R}$ as

$$f(x, y) = \begin{cases} (1/q)(1/s) & x = p/q \text{ and } y = r/s \text{ for } p, q, r, s \in \mathbb{Z} \\ 0 & (x, y) \notin \mathbb{Q}^2 \end{cases}.$$

The function f is constant everywhere on Q except on the null set \mathbb{Q}^2 , so f is integrable on Q . The function $f(\mathbf{x}, \mathbf{y})$ is integrable with respect to y over $[0, 1]$ for all x , and

$$\int_0^1 f(x, y) dy = 0.$$

Similarly, The function $f(\mathbf{x}, \mathbf{y})$ is integrable with respect to x over $[0, 1]$ for all y , and

$$\int_0^1 f(x, y) dx = 0.$$

By Fubini's theorem,

$$\begin{aligned} \int_{[0,1]^2} f(x, y) d(x, y) &= \int_0^1 \int_0^1 f(x, y) dy dx = \int_0^1 0 dx = 0, \\ \int_{[0,1]^2} f(x, y) d(x, y) &= \int_0^1 \int_0^1 f(x, y) dx dy = \int_0^1 0 dy = 0. \end{aligned}$$

Example 10.6 (Converse of Fubini's Theorem Fails). Suppose $Q = [0, 1]^2$, and $f : Q \rightarrow \mathbb{R}$ is defined as

$$f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2}.$$

The function f is integrable over $[0, 1]$ with respect to x for all y , and integrable over $[0, 1]$ with respect to y for all x .

$$\begin{aligned} \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy &= -\frac{1}{1+x^2} \\ \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx &= \frac{1}{1+y^2} \end{aligned}$$

Integrating each of these with respect to the other variable gives

$$\begin{aligned}\int_0^1 \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy dx &= \int_0^1 -\frac{1}{1+x^2} dx = -\frac{\pi}{4} \\ \int_0^1 \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx dy &= \int_0^1 \frac{1}{1+y^2} dy = \frac{\pi}{4}.\end{aligned}$$

The iterated integrals disagree, so what happened? It turns out that $f(x, y)$ was never integrable over Q to begin with, as it is unbounded! While Fubini's theorem states that the integrability of f over Q ensures the upper and lower integrals with respect to x and y exists, the existence of these upper and lower integrals do not guarantee that f is integrable over Q . This is still the case when the upper and lower integrals coincide for x and y :

$$\begin{aligned}\int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy &= \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy = \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dy = -\frac{1}{1+x^2}, \\ \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx &= \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx = \int_0^1 \frac{x^2 - y^2}{(x^2 + y^2)^2} dx = \frac{1}{1+y^2}.\end{aligned}$$

Even with Corollary 10.2, we still need to calculate the iterated integrals from scratch. If we assume that f is continuous, then we are able to use the fundamental theorem of calculus. This is the version of Fubini's theorem that is most often used and introduced in a multivariable calculus class.

Corollary 10.3 (Fubini's Theorem for Riemann Integrals, “Nicest” Version). Suppose $Q = [a_1, b_1] \times \cdots \times [a_n, b_n]$ is a rectangle in \mathbb{R}^n . If $f : Q \rightarrow \mathbb{R}$ is continuous on Q , then

$$\int_Q f(\mathbf{x}) d\mathbf{x} = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(\mathbf{x}) dx_1 \cdots dx_n.$$

Proof. If $f(\mathbf{x})$ is continuous on Q , then $f(\mathbf{x})$ is continuous with respect to x_i for all i . Therefore $f(\mathbf{x})$ is integrable with respect to all x_i , and we can apply Corollary 10.2. \square

Example 10.7. Suppose $f(x, y) = 6xy^2$, and we want to integrate it on the rectangle $[2, 4] \times [1, 2]$. We have

$$\begin{aligned}\int_{[2,4] \times [1,2]} 6xy^2 d(x, y) &= \int_2^4 \int_1^2 6xy^2 dy dx = \int_2^4 (2xy^3)_1^2 dx = \int_2^4 14x dx = [7x^2]_2^4 = 84, \\ \int_{[2,4] \times [1,2]} 6xy^2 d(x, y) &= \int_1^2 \int_2^4 6xy^2 dx dy = \int_1^2 [3x^2y^2]_1^2 dy = \int_1^2 36y^2 dy = [12y^3]_1^2 = 84.\end{aligned}$$

Example 10.8. Suppose $f(x, y)$ is a continuous function on a subrectangle $Q = [a_1, b_1] \times [a_2, b_2]$. If $f(x)$ can be written as $f(x) = g(x)h(y)$, then Fubini's theorem gives

$$\int_Q f(x, y) d(x, y) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} g(x) \underbrace{h(y)}_{\text{constant}} dx dy = \int_{a_2}^{b_2} h(y) \underbrace{\int_{a_1}^{b_1} g(x) dx}_{\text{constant}} dy = \int_{a_2}^{b_2} h(y) dy \int_{a_1}^{b_1} g(x) dx$$

10.4 Jordan Content

Until now, we've only defined the integral of $f(\mathbf{x})$ over some rectangle $Q \subset \mathbb{R}^n$. When $n = 1$, this is sufficient, as any closed and bounded (compact) set in \mathbb{R} is a trivial rectangle. For $n > 1$, this is no longer the case. For example we may desire to integrate a function $f(\mathbf{x})$ over the unit circle. We want to generalize Riemann integration such that we can integrate over a much more general collection of subsets of \mathbb{R}^n .

Before we can integrate over general (bounded) regions in \mathbb{R}^n , we need to determine how to measure the area/volume of such regions. For the sake of exposition, set $n = 2$, and suppose $S \subset \mathbb{R}^2$. Until now, S has taken the form $[a_1, b_1] \times [a_2, b_2]$ and has area $v(S) = (b_1 - a_1)(b_2 - a_2)$. But how do we measure the area if S is not a rectangle? This problem is reminiscent of finding the area under a curve, which gave the definition of the Riemann integral. In that case, we approximated the area underneath f with upper and lower Riemann sums, and were able to calculate the area underneath f when the infimum and supremum of these sums agreed. We can do something similar for general sets $S \subset \mathbb{R}^2$ using our definition of partitions.

We have assumed S is bounded, so it is the subset of some rectangle $Q \subset \mathbb{R}^2$. We can approximate the area of $S \subset Q$ by summing the areas of subrectangles corresponding to a partition of P . One way to do this is looking at the set of subrectangles of P which are contained *within* S , in other words subrectangles which are subsets of S° (recall the definition of interior as given in Section 2.4). We can approximate the area of S by summing over all such subrectangles:

$$\sum_{R \in P, R \subseteq S^\circ} v(R).$$

We could also estimate the area of S by looking at subrectangles which contain elements of S , or limit points of S . In other words, looking at subrectangles which intersect \bar{S} (recall the definition of closure as given in Section 2.4).

$$\sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R).$$

Figure 112 illustrates these approximations for various partitions of Q .

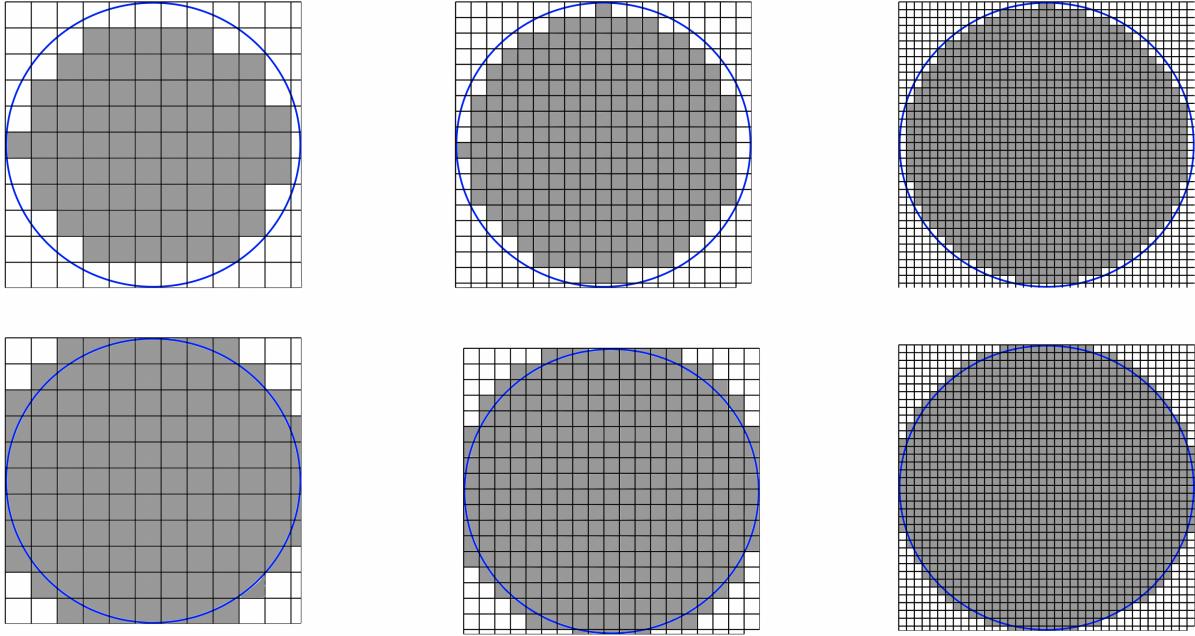


Figure 112: The area of S can be approximated as $\sum_{R \in P, R \subseteq S^\circ} v(R)$ (top row), or as $\sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R)$ (bottom row).

It appears that as we refine P , our two approximations for converge, much like upper and lower Riemann sums. This observation inspires the following definition which resembles the definition of the Riemann integral.

Definition 10.7. Let $S \subset \mathbb{R}^n$ be a bounded set, and Q be a rectangle containing S . Given some $P \in \mathbf{P}(Q)$, consider:

$$\begin{aligned}\underline{J}(P, S) &= \sum_{R \in P, R \subseteq S^\circ} v(R), \\ \bar{J}(P, S) &= \sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R).\end{aligned}$$

The numbers

$$\begin{aligned}\underline{c}(S) &= \sup_{P \in \mathbf{P}(Q)} \underline{J}(P, S), \\ \bar{c}(S) &= \inf_{P \in \mathbf{P}(Q)} \bar{J}(P, S),\end{aligned}$$

are defined to be the *inner Jordan content of S* and *outer Jordan content of S* , respectively. In the event that $\underline{c}(S) = \bar{c}(S)$, we say S is *Jordan-measurable*, denote this common value as $c(S)$, and refer to it as the *Jordan content of S* .

While the math behind the definitions make sense, what is with the very specific name? Why call this “Jordan content” and not just “volume”? For one, part of its name is due to the mathematician who developed it (Camille Jordan), but this isn’t the curious part. What the hell does “content” mean? While this is an excellent question, don’t worry too much about it at the moment. We’ll revisit it in a bit with an interesting example.

Remark 10.2 (Null Set vs. Zero Jordan Content). Jordan content gives a formal definition of area/volume, something that we did not have in earnest until now. The operative phrase here is “in earnest”, as we did give a definition for a set that has “negligible” area/volume – null sets (Definition 10.6). How does Jordan content relate to null sets? Null sets are defined using countable open covers. Jordan content is defined using the subrectangles of partitions, of which there are a finite number. In this sense, null sets are defined using finite coverings, while Jordan content is defined using finite coverings.

As a consequence, any set with Jordan content zero is necessarily a null set. Suppose S is a bounded set in \mathbb{R}^n which is Jordan measurable, and $c(S) = 0$. By the definition of Jordan content,

$$c(S) = \bar{c}(S) = \inf_{P \in \mathbf{P}(Q)} \bar{J}(P, S) = \inf_{P \in \mathbf{P}(Q)} \sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R) = 0.$$

Much like how it is not easy to calculate a Riemann integral directly using its definition, it is not easy to calculate the Jordan content of a set using Definition 10.7 (when possible).

Lemma 10.3. Let $S \subset \mathbb{R}^n$ be a bounded set contained in a rectangle Q . For any partition $P \in \mathbf{P}(Q)$,

$$\bar{J}(P, S) - \underline{J}(P, S) = \bar{J}(P, \partial S).$$

Proof. As illustrated in figure 112, any of the subrectangles contains in S° will satisfy $R \cap \bar{S} \neq \emptyset$. This means many of the rectangles $R \in P$ are used to calculate both $\bar{J}(P, S)$ and $\underline{J}(P, S)$. To be specific, the only rectangles that contribute to the difference in $\bar{J}(P, S)$ and $\underline{J}(P, S)$ are those contained in the set

$$\bar{S} \setminus S^\circ = (S^\circ \cup \partial S) \setminus S^\circ = \partial S.$$

Put more explicitly,

$$\begin{aligned}
\bar{J}(P, S) - \underline{J}(P, S) &= \sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R) - \sum_{R \in P, R \subseteq S^\circ} v(R) \\
&= \sum_{R \in P, R \cap (S^\circ \cup \partial S) \neq \emptyset} v(R) - \sum_{R \in P, R \subseteq S^\circ} v(R) \\
&= \left(\sum_{R \in P, R \cap S^\circ \neq \emptyset} v(R) + \sum_{R \in P, R \cap \partial S \neq \emptyset} v(R) \right) - \sum_{R \in P, R \subseteq S^\circ} v(R) \\
&= \sum_{R \in P, R \cap \partial S \neq \emptyset} v(R).
\end{aligned}$$

Because the closure of a boundary is the boundary itself, $\overline{\partial S} = \partial S$, we have

$$\bar{J}(P, S) - \underline{J}(P, S) = \sum_{R \in P, R \cap \partial S \neq \emptyset} v(R) = \sum_{R \in P, R \cap \overline{\partial S} \neq \emptyset} v(R) = \bar{J}(P, \partial S).$$

Each rectangle R in a partition P (of Q) is a compact set, so \square

Proposition 10.1 (Properties of Jordan Content). Let $S \subset \mathbb{R}^n$ be a bounded set. Then,

1. $\bar{c}(\partial S) = \bar{c}(S) - \underline{c}(S)$;
2. S is Jordan measurable if and only if ∂S has Jordan content zero.

Proof.

1. The set S is bounded, so it is contained in some rectangle Q . For an arbitrary partition $P \in \mathbf{P}(Q)$, Lemma 10.3 gives

$$\bar{J}(P, \partial S) = \bar{J}(P, S) - \underline{J}(P, S).$$

By the definition of supremum and infimum, for any $P \in \mathbf{P}(Q)$, $\bar{J}(P, S) \geq \bar{c}(S)$ and $\underline{c}(S) \geq \underline{J}(P, S)$, so Lemma 10.3 becomes

$$\begin{aligned}
&\bar{J}(P, \partial S) \geq \bar{c}(S) - \underline{c}(S) && (\forall P \in \mathbf{P}(Q)) \\
&\implies \underbrace{\inf_{P \in \mathbf{P}(Q)} \bar{J}(P, \partial S)}_{\bar{c}(S)} \geq \bar{c}(S) - \underline{c}(S) \\
&\implies \bar{c}(\partial S) \geq \bar{c}(S) - \underline{c}(S).
\end{aligned}$$

Now consider $\bar{J}(P, S)$ and $\underline{J}(P, S)$ in relation to $\bar{c}(S)$ and $\underline{c}(S)$, respectively. By the completeness of the real numbers and the definition of supremum/infimum, for all $\varepsilon > 0$ we can pick partitions $P_1, P_2 \in \mathbf{P}(Q)$ such that

$$\begin{aligned}
\bar{J}(P_1, S) &< \bar{c}(S) + \varepsilon/2 \\
\underline{J}(P_2, S) &> \underline{c}(S) - \varepsilon/2
\end{aligned}$$

If we define the common refinement $P = P_1 \cup P_2$, then $\bar{J}(P, S) \leq J(P_1, S)$ and $\underline{J}(P, S) \geq \underline{J}(P_2, S)$, so

$$\begin{aligned}
& \bar{J}(P, S) - \underline{J}(P, S) \leq \underbrace{\bar{J}(P_1, S)}_{<\bar{c}(S)+\varepsilon/2} - \underbrace{\underline{J}(P_2, S)}_{>\underline{c}(S)-\varepsilon/2} && (\forall \varepsilon > 0, P = P_1 \cup P_2 \in \mathbf{P}(Q)) \\
\implies & \underbrace{\bar{J}(P, S) - \underline{J}(P, S)}_{\bar{J}(P, \partial S)} < \bar{c}(S) - \underline{c}(S) + \varepsilon && (\forall \varepsilon > 0, P = P_1 \cup P_2 \in \mathbf{P}(Q)) \\
\implies & \bar{J}(P, \partial S) < \bar{c}(S) - \underline{c}(S) + \varepsilon && (\forall \varepsilon > 0, P = P_1 \cup P_2 \in \mathbf{P}(Q)) \\
\implies & \bar{J}(P, \partial S) \leq \bar{c}(S) - \underline{c}(S) && (\forall P = P_1 \cup P_2 \in \mathbf{P}(Q)) \\
\implies & \underbrace{\inf_{P \in \mathbf{P}(Q)} \bar{J}(P, \partial S)}_{\bar{c}(\partial S)} \leq \bar{c}(\partial S) - \underline{c}(S) && (\forall P = P_1 \cup P_2 \in \mathbf{P}(Q)) \\
\implies & \bar{c}(\partial S) \leq \bar{c}(S) - \underline{c}(S).
\end{aligned}$$

Therefore, $\bar{c}(\partial S) \leq \bar{c}(S) - \underline{c}(S)$ and $\bar{c}(\partial S) \geq \bar{c}(S) - \underline{c}(S)$, so

$$\bar{c}(S) = \bar{c}(\partial S) - \underline{c}(S).$$

2. The set S is Jordan measurable if $\bar{c}(S) = \underline{c}(S)$, which occurs if and only if $\bar{c}(\partial S) = 0$ using the first part of the proposition.

□

In the words of [Apostol \(1974\)](#), this proposition tells us that a bounded set is Jordan-measurable *if and only if* its boundary is not “too thick”. Let’s look at some examples.

Example 10.9 (Jordan Content of Rectangle). Define a rectangle $S = [a_1, b_1] \times [a_2, b_2]$ in \mathbb{R}^2 . The boundary of this rectangle is given as the union of disjoint sets (the “edges” of the rectangle)

$$\partial S = \{(a_1, y) \mid y \in [a_2, b_2]\} \cup \{(b_1, y) \mid y \in [a_2, b_2]\} \cup \{(x, a_2) \mid x \in [a_1, b_1]\} \cup \{(x, b_2) \mid x \in [a_1, b_1]\}.$$

I claim that this boundary is a null set because each “edge” is a null set, making it the countable union of null sets. Let’s take the edge $\{(a_1, y) \mid y \in [a_2, b_2]\}$ as a representative. It a closed and bounded subset of Euclidean space, so it is compact. We can cover this edge with the following open rectangle:

$$\left(a_1 - \frac{\varepsilon}{4(b_1 - a_2 + \varepsilon)}, a_1 + \frac{\varepsilon}{4(b_1 - a_2 + \varepsilon)} \right) \times (a_2 - \varepsilon/2, b_2 + \varepsilon/2),$$

where $\varepsilon > 0$ is some arbitrary number. When we calculate the “area” of this open rectangle covering one edge of S , we have

$$\begin{aligned}
& \left[a_1 + \frac{\varepsilon}{4(b_1 - a_2 + \varepsilon)} - \left(a_1 - \frac{\varepsilon}{4(b_1 - a_2 + \varepsilon)} \right) \right] \times [b_2 + \varepsilon/2 - (a_2 - \varepsilon/2)] = \frac{\varepsilon}{2(b_1 - a_2 + \varepsilon)} \times (b_1 - a_2 + \varepsilon) \\
& = \varepsilon/2 \\
& < \varepsilon.
\end{aligned}$$

This means that this edge is a null set. Similar arguments hold for the other three edges, so ∂S is a null set. By Proposition 10.1, S is Jordan measurable. To calculate the Jordan content of S , we can focus on $\bar{c}(S)$ (or $\underline{c}(S)$), because $c(S) = \bar{c}(S) = \underline{c}(S)$. If S is contained in a larger rectangle Q , then we can partition Q such that

S is contained in a single rectangle of the partition, where this rectangle is $R = [a_1 - \varepsilon, b_1 + \varepsilon] \times [a_2 - \varepsilon, b_2 + \varepsilon]$ for an arbitrary $\varepsilon > 0$. This gives

$$\begin{aligned}\bar{J}(P, S) &= v(R) \\ &= (b_2 - a_2 + 2\varepsilon)(b_1 - a_1 + 2\varepsilon) \\ &= (b_2 - a_2)(b_1 - a_1) + 2\varepsilon[(b_1 - a_1) + (b_2 - a_2) + 2\varepsilon]\end{aligned}$$

This rectangle covers the closure of S for all $\varepsilon > 0$, so the infimum is

$$c(S) = \inf \bar{J}(P, S) = (b_2 - a_2)(b_1 - a_1).$$

This argument is readily generalized to $S \subset \mathbb{R}^n$.

Example 10.10 (Jordan Content of Finite Set). Suppose $S \subset \mathbb{R}$ is a set of finite elements.

$$S = \{a_1, \dots, a_n\}.$$

We have $\partial S = S$, so the boundary is a finite set. Finite sets are null sets, so by Proposition 10.1, S is Jordan measurable. The closure of S is itslef, $\bar{S} = S$, so we can cover the closure with a partition comprised of intervals $(a_i - \varepsilon/2n, a_i + \varepsilon/2n)$ for $i = 1, \dots, n$ and an arbitrary $\varepsilon > 0$. In this case

$$\bar{J}(P, S) = \sum_{i=1}^n \varepsilon/n = \varepsilon.$$

The Jordan content is the infimum of $\bar{J}(P, S)$, which in this case is given as $\varepsilon \rightarrow 0$,

$$c(S) = 0.$$

Interestingly enough, S is also a null set because it is finite.

This example highlights an important question – **how are null sets and Jordan measurable sets with content zero related?** These concepts attempt to capture similar characterisites of a set, that being arbitrarily small “volume”. Null sets are defined in terms of *countable* coverings, whereas Jordan content is defined in terms of *finite* coverings. This immeaditly implies a useful proposition.

Proposition 10.2 (Zero Jordan Content \implies Null Set). Suppose $S \subset \mathbb{R}^n$ is a bounded null set. Then it is Jordan measurable and has a Jordan content of zero.

Example 10.11 (Null Set $\not\Rightarrow$ Jordan Measurable). Consider the set of rational numbers \mathbb{Q} . This set is countable, so it is a null set. In spite of this, $\partial\mathbb{Q} = \mathbb{R}$. The real numbers are not a null set, so Proposition 10.1 tells us that the set of rational numbers is not Jordan measurable. This example is also useful because it highlights an intuitive property that Jordan content fails to exhibit. The set \mathbb{Q} is countable, so we can write it as the disjoint union of singeltons. If we have $\mathbb{Q} = \{q_1, \dots\}$, then

$$\mathbb{Q} = \bigcup_{i=1}^{\infty} \{q_i\}.$$

Each of these singeltons is Jordan measurable and has Jordan content zero (see Example 10.10),

$$c(\{q_i\}) = 0 \quad \forall q_i \in \mathbb{Q}.$$

If Jordan content attempts to formalize the area/volume of a set, we would hope that the sum of the Jordan content of disjoint sets would be the Jordan content of their disjoint union, but this does not hold:

$$c(\mathbb{Q}) = c\left(\bigcup_{i=1}^{\infty} \{q_i\}\right) \neq \sum_{i=1}^{\infty} c(\{q_i\}) = \sum_{i=1}^{\infty} 0 = 0.$$

Unfortunately, this example shows this equality does not hold. In fact we shouldn't even be writing $c(\mathbb{Q})$, because \mathbb{Q} is not Jordan measurable!

It appears that for sets with negligible volume, the definition of a null set is more robust than a Jordan measurable set with content zero.¹⁶⁸ At the same time, the definition of null sets provides no guidance to measuring the volume of a set with non-negligible volume, hence the introduction of Jordan content. However, this discrepancy does raise another question – **why not define the volume of sets in a fashion similar to and consistent with the definition of null sets?** This is the right question to be asking! In fact, this is precisely what we will do in Section 14 when laying down the foundation of measure theory.

10.5 Integration over General Regions

Before establishing the connection between Jordan measurable sets and Riemann integrals over general regions, let's think about how we may want to define the integral of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ for any bounded set $S \subset \mathbb{R}^n$. Since S is bounded, we can enclose it in some rectangle $Q \subset \mathbb{R}^n$. We already know how to integrate f over Q , so is it possible to somehow “exclude” the portions of f defined on $Q \setminus S$? This can be done simply taking the value of f to be zero on this set.

Definition 10.8. Let $S \subset \mathbb{R}^n$ be a bounded set contained in a rectangle Q , and $f : S \rightarrow \mathbb{R}$ be a bounded function. Define $f_S : \mathbb{R}^n \rightarrow \mathbb{R}$ to be the function

$$f_S(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in S \\ 0 & \mathbf{x} \notin S \end{cases}$$

The (*multiple*) *Riemann integral* of f (over S) is defined as

$$\int_S f(\mathbf{x}) \, d\mathbf{x} = \int_Q f_S(\mathbf{x}) \, d\mathbf{x},$$

and will exist as long as f_S is integrable over Q .¹⁶⁹

Lebesgue's criterion (Theorem 10.2) gave necessary and sufficient conditions for the existence of the integral of $f(\mathbf{x})$ over some rectangle $Q \subset \mathbb{R}^n$. Perhaps we can just extend it to integrals over general regions by requiring the set of discontinuities of f on S to be a null set, afterall integrability of f on S is defined as integrability of f_S on a rectangle Q . This is exactly what we'll do, but it requires us to address one complication. First note that because the set of discontinuities of f is a null set on S , then the set of discontinuities of f_S on S will be a null set. We also know that f_S is constant on $Q \setminus S$, so it is continuous on this entire set. The only place that gives us trouble is the boundary ∂S .

By the definition of f_S , f_S will likely have many discontinuities on the boundary ∂S (see Figure 113). The boundary is where f_S “jumps” from taking on the values of f on S , to taking on the value zero on $Q \setminus S$. To ensure that these discontinuities are a null set, we need to impose the assumption that ∂S is a

¹⁶⁸Are there conditions we could impose such that the reverse of Proposition 10.2 holds? It turns out there is. If $S \subset \mathbb{R}^n$ is closed, then null sets will be Jordan measurable and have content zero.

¹⁶⁹It shouldn't come as a shock that it doesn't matter how we choose Q . This fact is formalized by Lemma 13.1 in Munkres (1999).

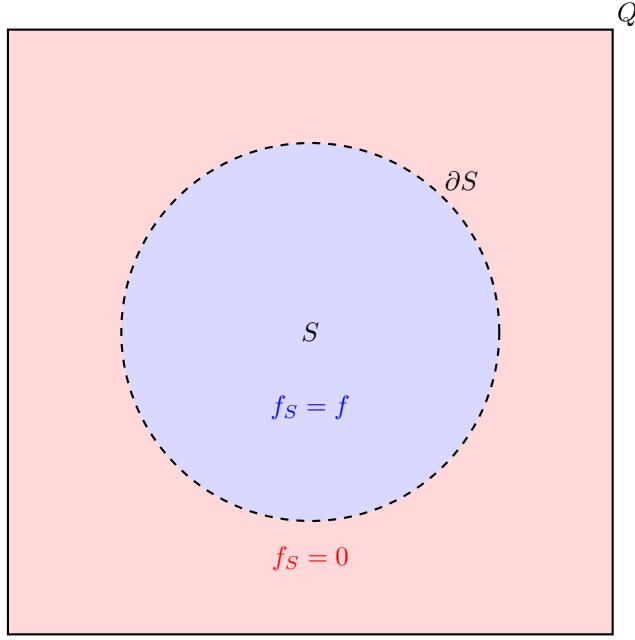


Figure 113: Unless $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in S$, then there is at least one discontinuity on ∂S . In fact, if $f(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in S$, then f_S is discontinuous on all of ∂S .

null set. But what do we know about sets whose boundaries are null sets? By Proposition 10.1, these are Jordan measurable sets! Therefore Lebesgue's theorem is readily extended to general regions, as long as those regions are given by Jordan measurable sets.

Theorem 10.4. Suppose $S \subset \mathbb{R}^n$ is Jordan measurable set, and $f : S \rightarrow \mathbb{R}$ is a bounded function. The function f is integrable over S if and only if the set of discontinuities of f on S is a null set.

Proof.

(\Rightarrow) Suppose f is integrable on S . This means for some rectangle $Q \subset \mathbb{R}^n$ which contains S , the function f_S is integrable on Q . By Theorem 10.2, the set of discontinuities of f_S on Q are a null set. This means that the discontinuities of f_S restricted to the set S is a null set, but this is simply the set of discontinuities of f on S .

(\Leftarrow) See the argument preceding the theorem. □

Example 10.12 (Integrating a Constant). Suppose we want to integrate the function $f(\mathbf{x}) = 1$ over some Jordan measurable set $S \subset \mathbb{R}^n$. We can find some rectangle Q such that $S \subset Q$, and define the function

$$\chi_S(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S \\ 0 & \mathbf{x} \notin S \end{cases}.$$

The function χ_S is discontinuous on the set ∂S , which is a null set because S is Jordan measurable, so Theorem 10.4 tells us that the following integral exists:

$$\int_S 1 d\mathbf{x}.$$

What is the value of this integral? For all partitions $P \in \mathbf{P}(Q)$, we have

$$M_R = \sup\{\chi_S(\mathbf{x}) \mid x \in R\} = \begin{cases} 1 & R \cap S \neq \emptyset \\ 0 & R \cap S = \emptyset \end{cases} \quad (\forall R \in P).$$

If we calculate the upper Riemann sums using M_R , we have

$$\begin{aligned} U(P, \chi_S) &= \sum_{R \in P} M_R v(R) \\ &= \sum_{R \in P, R \cap S \neq \emptyset} v(R) + 0 \\ &= \sum_{R \in P, R \cap S \neq \emptyset} v(R) + \sum_{R \in P, R \cap \partial S} v(R) \quad (\partial S \text{ null set}) \\ &= \sum_{R \in P, R \cap \bar{S} \neq \emptyset} v(R) \\ &= \bar{J}(P, S) \quad (\forall P \in \mathbf{P}(Q)). \end{aligned}$$

Since this holds for all partitions of Q , we have

$$\begin{aligned} \int_S 1 \, d\mathbf{x} &= \int_Q \chi_S \, d\mathbf{x} \\ &= \int_Q \chi_S \, d\mathbf{x} \\ &= \inf_{P \in \mathbf{P}(Q)} U(P, \chi_S) \\ &= \bar{J}(P, S) \quad (\forall \bar{J}(P, S), \text{ including sup}) \\ &= c(S). \end{aligned}$$

As you may have expected, the integral of 1 over the set S is simply the Jordan content of S !

This last example is interesting, as it gives us a second way to define Jordan content. The Jordan content can be defined as

$$c(S) = \int_S 1 \, d\mathbf{x}$$

for a Jordan measurable set. This happens to be how [Munkres \(1999\)](#) approaches integration over general regions, but he uses the term “recifiable sets” instead of Jordan measurable.

10.6 Change of Variables

In Section 6.10 we showed that

$$\int_{\phi(a)}^{\phi(b)} f(x) \, dx = \int_a^b f(\phi(y))\phi'(y) \, dy,$$

for a differentiable and monotonically increasing function φ such that φ' is integrable. One interpretation of this result is that we are able to “invert” the chain rule. Another interpretation is that we’ve transformed the interval we are integrating over using the mapping $[a, b] \mapsto [\varphi(a), \varphi(b)]$. How does this generalize to the case where $f : \mathbb{R}^n \rightarrow \mathbb{R}$?

First, we’ll define a special type of function that the transformation $\varphi : [a, b] \rightarrow [\varphi(a), \varphi(b)]$ falls into the category of.¹⁷⁰

¹⁷⁰Subtle category theory pun

Definition 10.9. A function $\mathbf{f} : A \rightarrow B$, where A and B are subspaces of Euclidean space, is a *diffeomorphism* if it is differentiable, and has a differentiable inverse.

Diffeomorphism is synonymous with “a change of variables”. It also sounds like some other scary sounding math definitions you may have come across: homomorphism, isomorphism, homeomorphism, endomorphism, etc. The common suffix “morphism” comes from the Greek for “form”, and all these terms refer to special types of functions which preserve some important characteristic/form of a space. In the case where a diffeomorphism ϕ is applied to some space A , then we’ve preserved the differentiability of A such that we can always apply a smooth inverse ϕ^{-1} . The definition presented here is simplified for the given context. A more general definition arises in differential topology.

Example 10.13. Suppose ϕ is a differentiable strictly monotone increasing function on $[a, b]$ such that ϕ' is integrable on $[a, b]$. These are the assumed conditions of Theorem 6.7. The function is strictly monotone, so it is injective. If we make the stronger assumption that ϕ' is not only integrable, but continuous, then the intermediate value theorem tells us that ϕ is surjective. Therefore, ϕ is a bijection and invertible. We also can apply the inverse function theorem (Proposition 5.1) to conclude that ϕ' is differentiable. This makes ϕ a diffeomorphism!

Lemma 10.4. Suppose $\mathbf{f} : A \rightarrow B$ is a diffeomorphism. Then $\mathbf{f}^{-1} : B \rightarrow A$ is a diffeomorphism.

Proof. By the definition of a diffeomorphism, \mathbf{f}^{-1} exists, and \mathbf{f} □

If we’re integrating $f(\mathbf{x})$ over A , what happens if we apply a diffeomorphism $\phi : A \rightarrow B$? Is it possible to relate the integral of f over A to an integral over B ? Perhaps we can turn to the generalized chain rule for guidance, as the change of variables in \mathbb{R} was an “inverted” chain rule. The chain rule tells us that

$$(f \circ \phi)' = f'(\phi(\mathbf{x}))\phi'(\mathbf{x}),$$

so perhaps we can just follow the same intuition used to justify Theorem 6.7.

$$\int_{\phi(A)} f(\mathbf{x}) \, d\mathbf{x} \stackrel{?}{=} \int_A f(\phi(\mathbf{y}))\phi'(\mathbf{y}) \, d\mathbf{y}.$$

This proposed equality has a major problem. In the single variable formulation of change of variables, φ' scales the function $f(\varphi(y))$ to offset the change in the bounds of integration. In the proposed equality, $\phi'(\mathbf{y})$ isn’t a single scalar adjusting $f(\phi(\mathbf{y}))$, it’s a matrix. An immediate consequence of this is that $f(\phi(\mathbf{y}))\phi'(\mathbf{y})$ isn’t even a function defined on \mathbb{R} , so the integral in question isn’t even defined. So how do we get a single scalar to adjust $f(\phi(\mathbf{y}))$ from a matrix $\phi'(\mathbf{y})$? We just use the absolute value of the determinant associated with it. Not only does this determinant exist under the assumption that ϕ is differentiable, but it also is an intuitive choice of scalar.

Theorem 10.5 (Change of Variables). Let $g : A \rightarrow B$ be a diffeomorphism of open sets in \mathbb{R}^n . If $f : B \rightarrow \mathbb{R}$ is continuous function on B , then $(f \circ g)\det(g') : A \rightarrow \mathbb{R}$ is integrable over A and

$$\int_{\phi(A)} f(\mathbf{x}) \, d\mathbf{x} = \int_A f(\phi(\mathbf{y})) |\det(\phi'(\mathbf{y}))| \, d\mathbf{y}.$$

The proof of this major result is so involved, it gets its own subsection. For now, let’s look at some classic examples to build intuition.

Example 10.14 (*u*-Substitution). Suppose we have a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and a diffeomorphism $\phi : \mathbb{R} \rightarrow \mathbb{R}$.

$$\det(\phi'(\mathbf{y})) = \det(\phi'(y)) = \phi'(y),$$

so Theorem 10.5 tells us that

$$\int_{g(a)}^{g(b)} f(x) dx = \int_a^b f(\phi(y))\phi'(y) dy.$$

This is Theorem 6.7, which is usually referred to as “*u*-substitution” in calculus courses.

Example 10.15 (Integration in Polar Coordinates). Suppose we want to integrate a function $f(\mathbf{x})$ over the portion of a circle of radius a that lies in the first quadrant of the \mathbb{R}^2 plane. This integral is expressed as

$$\int_B f(\mathbf{x}) d\mathbf{x},$$

$$B = \{\mathbf{x} \in \mathbb{R}^2 \mid x_1, x_2 > 0, \|\mathbf{x}\| < a\}.$$

If we didn’t want to integrate over the general region B , we could transform it such that its image is a rectangle in \mathbb{R}^2 . Let’s do this using the diffeomorphism $\phi(r, \theta) = (r \cos \theta, r \sin \theta)$ which expresses B in polar coordinates (illustrated in Figure 113). In this case,

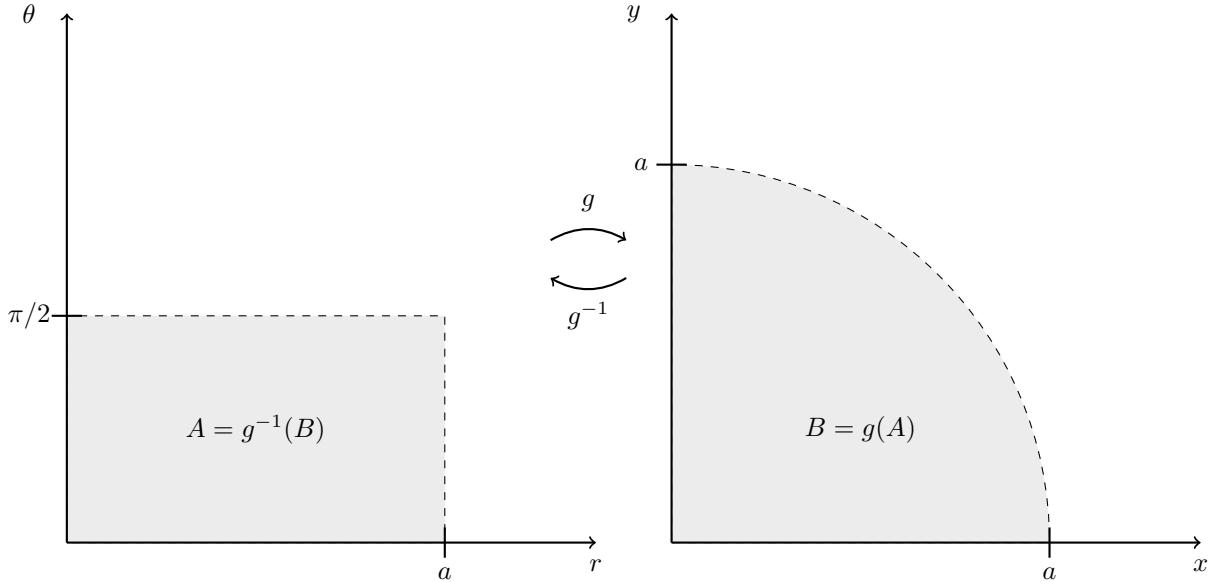


Figure 114:

$$A = \phi^{-1}(B) = \{\mathbf{y} = (r, \theta) \mid 0 < r < a, 0 < \theta < \pi/2\},$$

and

$$\det(\phi') = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r(\cos^2 \theta + \sin^2 \theta) = r.$$

The change of variables theorem tells us that

$$\int_B f(\mathbf{x}) d\mathbf{x} = \int_A f(\phi(\mathbf{y})) y_1 d\mathbf{y},$$

where $y_1 = r$. If we apply Fubini's theorem to write this as an iterated integral, we have

$$\int_B f(\mathbf{x}) d\mathbf{x} = \int_0^{\pi/2} \int_0^a f(\phi(r, \theta))r dr d\theta.$$

To make this more concrete, let's integrate $f(\mathbf{x}) = 4x_1x_2 - 7$ over the region B where the radius is $a = \sqrt{2}$.

$$\begin{aligned} \int_B f(\mathbf{x}) d\mathbf{x} &= \int_0^{\pi/2} \int_0^a f(\phi(r, \theta))r dr d\theta \\ &= \int_0^{\pi/2} \int_0^{\sqrt{2}} [4(r \cos \theta)(r \sin \theta) - 7]r dr d\theta \\ &= \int_0^{\pi/2} \int_0^{\sqrt{2}} 4r^3 \cos \theta \sin \theta - 7r dr d\theta \\ &= \int_0^{\pi/2} (r^4 \cos \theta \sin \theta - 7r^2/2)_0^{\sqrt{2}} d\theta \\ &= \int_0^{\pi/2} 4 \cos \theta \sin \theta - 7d\theta \\ &= \int_0^{\pi/2} 2 \sin 2\theta - 7d\theta \quad (\sin 2\theta = 2 \sin \theta \cos \theta) \\ &= (-\cos 2\theta - 7\theta)_0^{\pi/2} \\ &= 2 - 7\pi/2 \end{aligned}$$

Example 10.16 (Spherical Coordinates). Similar to polar coordinates in \mathbb{R}^2 , we can integrate with respect to spherical coordinates in \mathbb{R}^3 . If ϕ is the diffeomorphism corresponding to a transformation into spherical coordinates (see Example 8.19), then

$$\int_B f(\mathbf{x}) d\mathbf{x} = \int_A f(r \sin \theta \cos \varphi, r \sin \theta \sin \varphi, r \cos \theta) r^2 \sin \theta dr d\varphi d\theta.$$

Example 10.17 (Geometry of the Determinant). We can use change of variables to be more explicit in showing the determinant is the factor by which a linear transformation scales sets. Suppose we have a square matrix $A \in L(\mathbb{R}^n)$ corresponding to a linear transformation $T(\mathbf{x}) = A\mathbf{x}$, and some Jordan measurable set $S \in \mathbb{R}^n$. If we apply T to S , then what is the Jordan content of $T(S)$? Example 10.12 established that

$$c(S) = \int_S 1 d\mathbf{x}.$$

If we apply the transformation $T(\mathbf{x}) = A\mathbf{x}$ on the set S , we have

$$c(S) = \int_S 1 d\mathbf{x} = \int_{T(S)} 1 \cdot |\det(T')| d\mathbf{x} = \int_{T(S)} |\det A| d\mathbf{x}.$$

The determinant of a matrix of constants A is not a function of \mathbf{x} , so we can factor it out

$$\begin{aligned} c(S) &= \int_{T(S)} |\det A| d\mathbf{x} \\ \implies c(S) &= |\det A| \int_{T(S)} 1 d\mathbf{x} \\ \implies c(S) &= |\det A| c(T(S)) \quad \left(\int_{T(S)} 1 d\mathbf{x} = c(T(S)) \right) \\ \implies c(T(S)) &= C(S) \cdot |\det A|. \end{aligned}$$

The determinant's familiar geometric properties are still relevant for Jordan content!

10.7 Change of Variables, Proof

Proving Theorem 10.5 not only requires a handful of lemmas, but also the introduction of a few definitions. The first definition that will be relevant for the proof is a special type of function.

Definition 10.10. Let $f : A \rightarrow B$ be a function where $A, B \subset \mathbb{R}^n$ where $n \geq 2$.

$$\mathbf{f}(\mathbf{x}) = (\mathbf{f}_1(\mathbf{x}), \dots, \mathbf{f}_n(\mathbf{x})) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}$$

We say f is *preserves the i -th coordinate* if $f_i(\mathbf{x}) = x_i$ for all $\mathbf{x} \in A$. The function f is *primitive* if it preserves at least one coordinate $i = 1, \dots, n$.

Example 10.18. Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be $\mathbf{f}(\mathbf{x}) = (x_1^2, 3x_2, \sin(x_1 x_4), x_4)$. This function is primitive because it preserves the coordinate $i = 4$.

Our first lemma tells us that, roughly speaking, we can decompose a diffeomorphism as the composition of primitive diffeomorphisms.

Lemma 10.5. Let $\mathbf{f} : A \rightarrow B$ be a diffeomorphism of open sets in \mathbb{R}^n where $n \geq 2$. For all $\mathbf{x}_0 \in A$, there exists some open ball $U_0 = B_r(\mathbf{x}_0)$ contained in A , and a sequence of primitive diffeomorphisms of open sets in \mathbb{R}^n ,

$$U_0 \xrightarrow{\mathbf{g}_1} U_1 \xrightarrow{\mathbf{g}_2} U_2 \longrightarrow \cdots \longrightarrow U_{k-1} \xrightarrow{\mathbf{g}_k} U_k,$$

such that $\mathbf{g}_k \circ \mathbf{g}_{k-1} \circ \cdots \circ \mathbf{g}_2 \circ \mathbf{g}_1 = \mathbf{f}$ on the set U_0 .

Proof. We will prove the result by showing that the general case can, without loss of generality, actually be expressed as increasingly simple special cases. This process gives us four steps:

1. Showing that the general case
- 2.
- 3.

Step 1. Suppose $\mathbf{f} : A \rightarrow B$ is a diffeomorphism of open sets in \mathbb{R}^n , and let \mathbf{x}_0 be some arbitrary element of the domain A . The function f is a diffeomorphism on A , so it is differentiable at \mathbf{x}_0 and the linear transformation $\mathbf{f}'(\mathbf{x}_0)$ exists, and has a well defined inverse $[\mathbf{f}'(\mathbf{x}_0)]^{-1}$. Define the following functions from \mathbb{R}^n to \mathbb{R}^n :

$$\begin{aligned} \mathbf{g}_1(\mathbf{x}) &= \mathbf{x} + \mathbf{x}_0, \\ \mathbf{g}_2(\mathbf{x}) &= \mathbf{x} - \mathbf{f}(\mathbf{x}_0), \\ \mathbf{G}(\mathbf{x}) &= [\mathbf{f}'(\mathbf{x}_0)]^{-1} \mathbf{x}. \end{aligned}$$

The function \mathbf{g}_1 is clearly a diffeomorphism, while \mathbf{g}_2 and \mathbf{G} are diffeomorphisms because \mathbf{f} is a diffeomorphism. Define the function

$$\tilde{\mathbf{f}} = \mathbf{G} \circ \mathbf{g}_2 \circ \mathbf{f} \circ \mathbf{g}_1.$$

Figure 115 shows illustrates these defined functions. The composition of diffeomorphisms is itself a

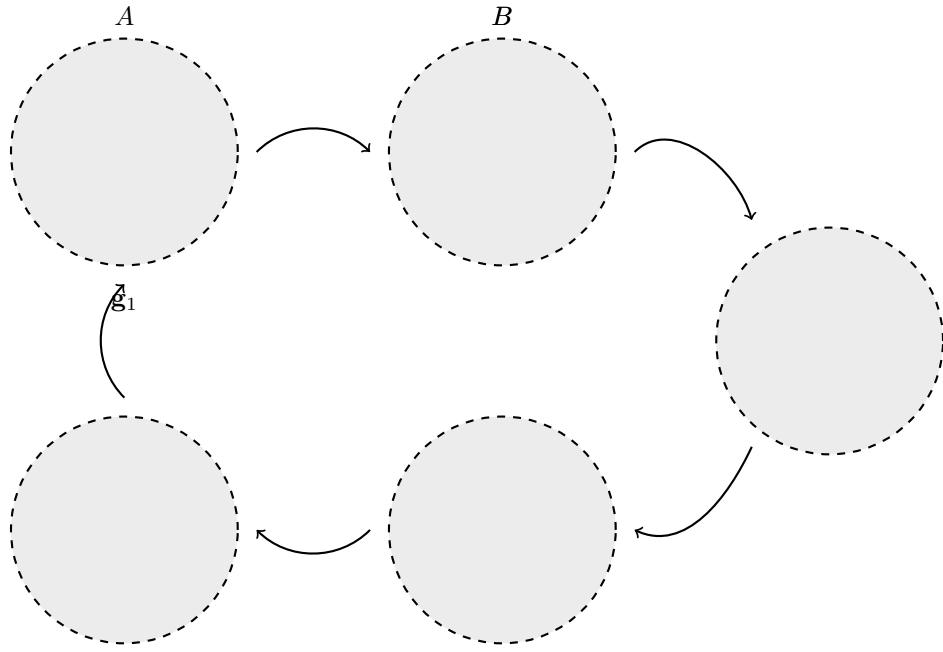


Figure 115:

diffeomorphism, so \tilde{f} is a diffeomorphism. We've defined \tilde{f} such that $\tilde{f}(\mathbf{0}) = \mathbf{0}$ and $\tilde{f}'(\mathbf{0}) = \mathbf{I}$.

$$\begin{aligned}
 \tilde{f}(\mathbf{0}) &= (\mathbf{G} \circ \mathbf{g}_2 \circ \mathbf{f} \circ \mathbf{g}_1)(\mathbf{0}) \\
 &= \mathbf{G}(\mathbf{g}_2(\mathbf{f}(\mathbf{g}_1(\mathbf{0})))) \\
 &= \mathbf{G}(\mathbf{g}_2(\mathbf{f}(\mathbf{x}))) \\
 &= \mathbf{G}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0))
 \end{aligned}$$

Step 2.

Step 3.

Step 4.

□

11 Manifolds in Euclidean Space

11.1 Motivation – Smooth Curves

11.2 Smooth Manifolds

11.3 Tangent Spaces

12 Multilinear Algebra and Differential Forms

12.1 Familiar Examples

12.2 Manifolds

12.3 Tensors

12.4 Wedge Product

12.5 Tangent Vectors and Differential Forms

12.6 Stokes' Theorem

Part III

Measure Theory

13 Point-Set Topology Revisited

All the way back in Section 2 we discussed metric spaces. This material laid the groundwork for nearly all the material up until this point. We're about to venture into the wild, but very fun, world of measure theory and functional analysis, and this will require a more abstract understanding of topology and the structure of sets. As we'll see, *everything* in Section 2 can be interpreted as a special case of a more general type of set. This section is by no means exhaustive, and only highlights material that is relevant for analysis. The best introductory reference for topological spaces is [Munkres \(2000\)](#), who gives a thorough foundation in topology as it pertains to analysis, geometry, and algebra.

It's also worth noting that the material from this section onward tends to be more abstract and is the “jumping off” point for many references aimed at students in an introductory graduate real analysis course. One of the best payoffs of a more abstract approach is it will put the material covered up until this point in a broader context, and give us the tools required to see the forest through the trees.^{[171](#)}

13.1 A Beautiful Day in the Neighborhood

At the very beginning of Section 2 I stated “One of the main goals of calculus is to study rates of change and limiting behavior. Both of these concepts require some notion of distance”. Some of the most important definitions and results we've covered involved inequalities that were written in terms of a metric d and involved arbitrarily small distances $\varepsilon > 0$. Many of these could alternitally be written in the form of open balls B_r . Here are just a few examples:

	Concept	Inequality	Open Ball Form
Definition 3.2	Convergence of $\{x_n\} \subset X$	$d(x_n, x) < \varepsilon$	$x_n \in B_\varepsilon(x)$
Definition 3.5	Cauchy Sequence $\{x_n\} \subset X$	$d(x_n, x_m) < \varepsilon$	$x_n \in B_\varepsilon(x_m)$
Definition 4.1	$\lim_{x \rightarrow x_0} f(x) = L$	$d_X(x, p) < \delta, d_Y(f(x), L) < \varepsilon$	$x \in B_\delta(p), f(x) \in B_\varepsilon(L)$
Definition 4.2	Continuity of $f : X \rightarrow Y$	$d_X(x, p) < \varepsilon, d_Y(f(x), f(p)) < \varepsilon$	$x \in B_\delta(p), f(x) \in B_\varepsilon(f(p))$
Theorem 6.1	Riemann's Criterion	$U(P, f) - L(P, f) < \varepsilon$	$U(P, f) \in B_\varepsilon(L(P, f))$
Definition 6.12	Null sets in \mathbb{R}	$\sum_{i=1}^{\infty} (b_i - a_i) < \varepsilon$	$\sum_{i=1}^{\infty} (b_i - a_i) \in B_\varepsilon(0)$
Defintion 7.5	Pointwise convergence	$ f_n(x) - f(x) < \varepsilon$	$f_n(x) \in B_\varepsilon(f(x))$

There is a subtle difference between writing these definitions in terms of inequalities and writing them in terms of open balls. For reference, let's focus on the definition of a continuous function between two metric spaces $f : X \rightarrow Y$. In inequality form we explicitly reference the distances $d_X(x, p)$ and $d_Y(f(x), f(p))$. In the open ball version all we have $x \in B_\delta(p)$ and $f(x) \in B_\varepsilon(f(p))$, which tell us “these objects are close to each other in the sense that they belong to the same open ball”, but it doesn't *directly* make any statements about the exact distance. Yes, the underlying definition of $B_\delta(p)$ and $B_\varepsilon(f(p))$ are based on the metrics d_X and d_Y , but we could interpret this as a means to an end. Maybe all we need for continuity is the concept

¹⁷¹This will be especially true in Section 17.

open sets which deal with proximity instead of exact distances, and metric spaces gave us a concrete way of defining these open sets.

Let's think about what a space would look like if it had a "built in" notion of proximity instead of distances. Given a set X , we would want to determine if a subset $A \subseteq X$ contains points that are "close" to x , for all $x \in X$. In other words, for any $x \in X$, we want to find a collection of subsets of X that contain points close to x (such a collection will be a subset of the powerset of X). We'll now define such a mapping, and present the axioms that make it consistent with the concept of proximity

Definition 13.1. Give a set of points X , a *topological space* is an ordered pair (X, \mathcal{N}) where \mathcal{N} maps elements of X to collections of subsets of X , $x \mapsto \mathcal{N}(x) \subset \mathcal{P}(X)$,¹⁷² which satisfies the following for all $x \in X$:

1. $\mathcal{N}(x) \neq \emptyset$.
2. For all $A \in \mathcal{N}(x)$, $x \in A$
3. For all $A \in \mathcal{N}(x)$ and $B \subseteq X$, if $A \subseteq B$ then $B \in \mathcal{N}(x)$
4. For all $A, B \in \mathcal{N}(x)$, $A \cap B \in \mathcal{N}(x)$
5. For all $A \in \mathcal{N}(x)$, there exists some $B \in \mathcal{N}(x)$ such that $A \in \mathcal{N}(y)$ for all $y \in B$.

The collection of sets $\mathcal{N}(x)$ are the *neighborhoods of x* .

In a topological space, a neighborhood of x can be thought of as the set of points that are "close" to x (hence the name neighborhood). The first property two properties are a bit trivial – every point has a neighborhood, and any neighborhood of x must contain x . All points $x \in X$ are trivially close to themselves, so it wouldn't make sense for a neighborhood of x to not contain x . Next we have that if any neighborhood of x belongs to some larger set, that larger set is also a neighborhood of x . One way to think about this is if someone lives in a neighborhood of a city, then the city they live in can also be considered a "larger" neighborhood. The third property tells us that if we look have two neighborhoods of points that are close to x , then the elements common to each neighborhood is itself a neighborhood of points close to x . The final property is the most abstract. It tells us the, roughly speaking, a neighborhood of x is shared by all its neighbors (given by B) who are sufficiently close.

Example 13.1 (Standard Topology on \mathbb{R}). Let $X = \mathbb{R}$ and define a neighborhood of x to be any open interval containing x .

$$\mathcal{N}(x) = \{(a, b) \in \mathbb{R} \mid x \in (a, b)\}.$$

Let's verify \mathcal{N} satisfies the properties given in Definition 13.1, and use this as a chance to get a feel for what the fourth property means.

1. $x \in (-\infty, \infty) = \mathbb{R}$ for all x , so $\mathbb{R} \in \mathcal{N}(x)$ for all x , therefore $\mathcal{N}(x) \neq \emptyset$
2. By definition $x \in (a, b)$ for each (a, b) which contains x .
3. If $x \in (a, b)$ where $(a, b) \subseteq (c, d)$, then $(c, d) \in \mathcal{N}(x)$ because it is an open interval containing x .
4. If $x \in (a, b)$ and $x \in (c, d)$, then $(a, b) \cap (c, d) = (\max\{a, c\}, \min\{b, d\})$ which is an open set containing x . This makes it a neighborhood of x .

¹⁷² $\mathcal{P}(X)$ is the power set of X .

5. Let (a, b) be an arbitrary neighborhood of x . There exists some ε such that $(x - \varepsilon, x + \varepsilon) \in (a, b)$. By property 2, (a, b) is a neighborhood of all $y \in (x - \varepsilon, x + \varepsilon)$.

Definition 13.1 is *great* for building intuition (hopefully?), but it's actually not the most prominent definition of a topological space (of which there are [several](#)). Nearly all references present a definition that takes the “starting point” to be a type of set we’re quite used to working with in metric spaces.¹⁷³

Definition 13.2. Let X be a nonempty set. A *topology on X* , $\mathcal{T} \subset \mathcal{P}(X)$, is a collection of subsets of X such that:

1. $\emptyset, X \in \mathcal{T}$.
2. If $\{U_\alpha\}_{\alpha \in A} \subseteq \mathcal{T}$, then $\cup_{\alpha \in A} U_\alpha \in \mathcal{T}$ (closed under arbitrary unions).
3. If $\{U_1, \dots, U_n\} \subseteq \mathcal{T}$, then $\cap_{i=1}^n U_i \in \mathcal{T}$ (closed under finite intersections).

The subsets that belong to \mathcal{T} are known as *open sets*. The pair (X, \mathcal{T}) is a *topological space*.¹⁷⁴

When mathematicians were developing the basics of topology, different definitions were suggested, but over time Definition 13.2 became the standard. While the intuition isn’t as clear as Definition 13.1, it’s *much* easier to work with. It also is consistent with the properties of open sets we established in Section 2 (Theorem 2.2, Example 2.17, and Example 2.22). We’ll now show the two definitions are interchangeable. The proof of this result is informative (and a bit fun) because we’ll explicitly define a mapping between neighborhoods and open sets.

Theorem 13.1. Definition 13.2 and Definition 13.1 present equivalent formulations of a topological space.

Proof.

(\Rightarrow) Let (X, \mathcal{T}) be a topological space as defined in Definition 13.2. Define a neighborhood of x as a subset of X which contains an open set containing x .

$$\mathcal{N}(x) = \{A \subseteq X \mid \text{there exists a } U \in \mathcal{T} \text{ s.t } x \in U \subseteq A\}$$

We will verify that $\mathcal{N}(x)$ satisfies the axioms given in Definition 13.1.

1. By the first axiom of open sets, $X \in \mathcal{T}$. This means for each $x \in X$, we have $x \in X \subseteq X$ for $X \in \mathcal{T}$. By the definition of $\mathcal{N}(x)$, $X \in \mathcal{N}(x)$ for each x , so $\mathcal{N}(x) \neq \emptyset$.
2. By the definition of $\mathcal{N}(x)$, if $A \in \mathcal{N}(x)$, then $x \in U \subseteq A$ for some open set $U \in \mathcal{T}$. If $x \in U \subseteq A$, then $x \in A$. Therefore, all neighborhoods of x contain x .
3. Suppose $A \in \mathcal{N}(x)$ and $A \subseteq B \subseteq X$. There exists an open set $U \in \mathcal{T}$ such that $x \in U \subseteq A$, but since $A \subseteq B$, then U also satisfies $x \in U \subseteq B$ for this open set U . Therefore $B \in \mathcal{N}(x)$.
4. Let $A, B \in \mathcal{N}(x)$. There exists open sets $U, V \in \mathcal{T}$ such that $x \in U \subseteq A$ and $x \in V \subseteq B$. By third axiom of open sets, $U \cap V$ is an open set. This open set also satisfies $x \in U \cap V \subseteq A \cap B$, so $A \cap B \in \mathcal{N}(x)$.

¹⁷³I've chosen to follow [Brown \(2006\)](#) by starting with the neighborhood definition.

¹⁷⁴In many situations, the topology \mathcal{T} can usually be inferred by the space X , so we just refer to X as a topological space. This is just like how we would refer to X as a metric space instead of (X, d) with the metric d if it was obvious what metric was being used.

5. Let $A \in \mathcal{N}(x)$. There exists an open set $U \in \mathcal{T}$ such that $x \in U \subseteq A$. We have $U \in \mathcal{N}(x)$ because $x \in U \subseteq U$ where $U \in \mathcal{T}$. For each $y \in U$, $y \in U \subseteq U$ where $U \in \mathcal{T}$, so $U \in \mathcal{N}(y)$. Since $U \subseteq A$, we know $A \in \mathcal{N}(y)$ for all $y \in U$ by the third axiom of neighborhoods (which we've already shown is satisfied). We've just shown the fifth axiom holds by taking $U = B$ where B is given in Definition 13.1.

(\Leftarrow) Let (X, \mathcal{N}) as defined in Definition 13.1. Define an open set $U \subset X$ as a set which is a neighborhood of each of its points.

$$\mathcal{T} = \{U \subseteq X \mid U \in \mathcal{N}(x) \text{ for all } x \in U\}$$

We will verify that \mathcal{T} satisfies the axioms given in Definition 13.2.

1. It is vacuously true that $\emptyset \in \mathcal{T}$. Suppose $x \in X$. We know $\mathcal{N}(x) \neq \emptyset$ so there exists some $A \in \mathcal{N}(x)$. Since $\mathcal{N}(X) \subset \mathcal{P}(X)$, $A \in \mathcal{P}(X)$, and by the definition of the power set, $A \subset X$. By the second axiom of neighborhoods, $X \in \mathcal{N}(x)$ for all $x \in X$, so X is open by the definition of \mathcal{T} .
2. Let $\{U_\alpha\} \subset \mathcal{T}$ be an arbitrary collection of open sets. For all α , $U_\alpha \in \mathcal{N}(x)$ for all $x \in U_\alpha$. Define $U = \cup_\alpha U_\alpha$. If $U = \emptyset$, we're done, as we've already shown that $\emptyset \in \mathcal{T}$. If U is not empty, then let $x \in U$. There exists some specific α' such that $x \in U_{\alpha'}$. By the definition of \mathcal{T} , $U_{\alpha'} \in \mathcal{N}(x)$. By the third axiom of neighborhoods, $U \in \mathcal{N}(x)$ because $U_{\alpha'} \subseteq U$. This holds for all x , so $U \in \mathcal{T}$.
3. It suffices to show that $U \cap V \in \mathcal{T}$ for $U, V \in \mathcal{T}$ because we could simply apply the result to U_1 and U_2 , and then to $U_1 \cap U_2$ and U_3 , etc. If $U \cap V = \emptyset$ then we have already established that $U \cap V \in \mathcal{T}$. If the intersection is not empty, let $x \in U \cap V$ be an arbitrary element. The point x belongs to U and V , so $U, V \in \mathcal{N}(x)$ by the definition of \mathcal{T} . By the fourth axiom of neighborhoods, $U \cap V \in \mathcal{N}(x)$. This holds for all x , so $U \cap V \in \mathcal{T}$.

□

An immediate consequence of this is we can define a neighborhood for a topological space as given by (X, \mathcal{T}) .

Definition 13.3. If (X, \mathcal{T}) is a topological space, then $V \subseteq X$ is an *neighborhood of x* if there exists some $U \in \mathcal{T}$ such that $x \in U \subseteq V$.

Proving Theorem 13.1 may not have been too painful, but it's still a fairly abstract result. To make things a bit more clear, let's look at the interplay between Definition 13.1, Definition 13.2, and Theorem 13.1 in the context of a metric space.

Example 13.2 (The Metric Topology). Let (X, d) be any metric space. We'll define a neighborhood of x as any set which contains an open ball around x .¹⁷⁵

$$\mathcal{N}(x) = \{A \subseteq X \mid \text{there exists a } r > 0 \text{ s.t } B_r(x) \subseteq A\}.$$

1. If we let $r = \infty$ then $B_r(x) = X$, so $x \in B_r(X)$ and $\mathcal{N}(x) \neq \emptyset$.
2. By the definition of a metric, $d(x, x) = 0$, so $x \in B_r(x)$. If $A \in \mathcal{N}(x)$, then $B_r(x) \subseteq A$ for some r , and $x \in B_r(x) \in A$.

¹⁷⁵This definition of a neighborhood in a metric space is consistent with general point set topology, opposed to the definition given by Rudin (1976).

3. If $A \in \mathcal{N}(x)$ then there exists some $B_r(x)$ such that $B_r(x) \subseteq A$. If $A \subseteq B$, then $B_r(x) \subseteq B$, giving $B \in \mathcal{N}(x)$.
4. If $A, B \in \mathcal{N}(x)$, then we have $r, q > 0$ such that $B_r(x) \subseteq A$ and $B_q(x) \subseteq B$. If we define

$$r' = \max_{y \in A \cap B} d(x, y)$$

then $B_{r'} \subseteq A \cap B$ so $A \cap B \in \mathcal{N}(x)$.

5. If $A \in \mathcal{N}(x)$ then there exists some $B_r(x) \subseteq A$. For all $y \in B = B_r(x)$ we have $y \in B_r(x) \subset A$, so $A \in \mathcal{N}(y)$ for each y by the third axiom of neighborhoods.

Given this definition of $\mathcal{N}(x)$, the proof of Theorem 13.1 tells us that we can define an open set as any set $U \subseteq X$ such that U is a neighborhood of all of its points $x \in U$. In other words, there exists some $r > 0$ such that $B_r(x) \subseteq U$ for all $x \in U$. This is exactly the same definition as that given in Definition 2.10, although Definition 2.10 was stated in terms of the intermediate Definition 2.8.

Given a topological space (X, \mathcal{T}) , we can give other definitions that should seem familiar.

Definition 13.4. Let (X, \mathcal{T}) be a topological space, and $U \in \mathcal{T}$. We say $U^c \subset X$ is a *closed set*.

This definition is consistent with Theorem 2.1 and Corollary 2.2. We can generalize Theorem 2.2.

Proposition 13.1. Let $\{E\}_\alpha$ and $\{E_1, \dots, E_n\}$ be an arbitrary collection of closed sets and finite collection of closed sets respectively. Then,

1. $\bigcap_\alpha E_\alpha$ is closed.
2. $\bigcup_{i=1}^n E_i$ is closed.

Proof. These follow directly from DeMorgan's laws and the definition of open and closed sets. The sets E_α^c are open by the definition of a closed set, and the arbitrary union of open sets is open, so

$$\bigcup_\alpha E_\alpha^c$$

is open. The compliment of this is closed, and this compliment is

$$\left(\bigcup_\alpha E_\alpha^c \right)^c = \bigcap_\alpha E_\alpha.$$

The second result follows from the same reasoning applied to the finite intersection of open sets. \square

We can also give more general definitions of key concepts introduced in Section 2.

Definition 13.5. Let $E \subseteq X$ for some topological space (X, \mathcal{T}) . We say $x \in X$ is an *interior point of S* if there exists some $U \in \mathcal{T}$ such that $x \in U \subseteq S$.

Definition 13.6. Let $E \subseteq X$ for some topological space (X, \mathcal{T}) . We say $x \in X$ is an *limit point of S* if every neighborhood of x contains a point $y \in E$ where $y \neq x$.

Definition 13.7. Let A be a subset of some topological space (X, \mathcal{T}) . The *interior* of A , written as A° , is the union of all the open sets contained in A .

$$A^\circ = \bigcup \{U \in \mathcal{T} \mid U \subseteq A\}$$

Definition 13.8. Let A be a subset of some topological space (X, \mathcal{T}) . The *closure* of A , written as \bar{A} , is the intersection of all the closed sets contained in A .

$$\bar{A} = \bigcap\{E \in \mathcal{P}(X) \mid E^c \in \mathcal{T}, E \subseteq A\}$$

Definition 13.9. Let A be a subset of some topological space (X, \mathcal{T}) . The *boundary* of A , written as ∂A , is defined as

$$\partial A = \bar{A} \setminus A^\circ$$

Definition 13.10. Let A be a subset of some topological space (X, \mathcal{T}) . The set A is *dense in X* if $\bar{A} = X$.

Example 13.3 (Metric Spaces). Each of these last definitions were originally given in the context of metric spaces in Section 2, but only Definition 13.8 is identical to its metric space counterpart (Definition 2.15). Let's verify that for a metric space (X, d) , the previous definitions are equivalent to the general definitions for a topological space (X, \mathcal{T}) where \mathcal{T} is induced by the metric d (as in Example 13.2).

The motivation for defining a topological space (X, \mathcal{T}) was to generalize metric spaces, so let's look at some examples of topologies that aren't defined via some metric d associated with X .

Example 13.4 (Discrete Topology). Let X be any nonempty set. The power set of X is a valid topology, $\mathcal{T} = \mathcal{P}(X)$.

Example 13.5 (Trivial Topology). Let X be any nonempty set, and define the topology $\mathcal{T} = \{\emptyset, X\}$.

Example 13.6 (Relative Topology). If (X, \mathcal{T}) is a topological space and $Y \subset X$, then the set

$$\mathcal{T}_Y = \{U \cap Y \mid U \in \mathcal{T}\}$$

is a topology on Y . Consider the real line \mathbb{R} equipped with the standard topology (Example 13.1). We can form a topological space on the unit interval $(0, 1)$ by equipping it with the relative topology with respect to \mathbb{R} .

$$\mathcal{T}_{(0,1)} = \{U \cap (0, 1) \mid U \in \mathcal{T}\} = \{(a, b) \subset \mathbb{R} \mid a, b \in (0, 1), a < b\}$$

Example 13.7 (Cofinite Topology). Let X be a nonempty set. Define an open set as the emptyset and any set with a finite complement.

$$\mathcal{T} = \{U \subseteq X \mid U = \emptyset \text{ or } |U^c| < \infty\}$$

In the event that X is finite, then all subsets of X of a finite complement, so the cofinite topology becomes the discrete topology.

Remark 13.1 (To Metric or Not To Metric?). We've now seen a few examples of topological spaces that aren't metric spaces...or have we? Could it be the case that some of these topological spaces are actually metric spaces in disguise? For instance the standard topology on \mathbb{R} almost trivially coincides with the topology induced by the Euclidean metric in \mathbb{R} . The set of intervals (a, b) are precisely the open sets associated with the metric $d(x, y) = |x - y|$, hence the name open intervals. Now consider the discrete topology of Example 13.4 in relation to the discrete metric (Example 2.11). Are these names purely a coincidence? Let $U \subseteq X$. If

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

then $B_{1/2}(x) \subseteq U$ for all $x \in U$, so U is open (which we've shown using the definition for metric spaces). This holds for all $U \subseteq X$, so all subsets of X are open. In other words the topology induced by the discrete metric is $\mathcal{T} = \mathcal{P}(X)$, which is the discrete topology, so there is no coincidence here. It doesn't seem probable that any topology \mathcal{T} on some X can be induced by some metric, otherwise we wouldn't have introduced topologies as a generalization of metric spaces. So the question becomes, given a topological space (X, \mathcal{T}) , under what conditions can the open sets in \mathcal{T} be given as the sets that are open with respect to a metric space (X, d) ? We'll find one answer to this question in Section 13.2.

If we can define multiple topologies on X (Example 13.4 and Example 13.5), then how do we compare these topologies? One way is via the subset relation.

Definition 13.11. Given two topologies \mathcal{T}_1 and \mathcal{T}_2 defined on the set X , we say \mathcal{T}_1 is *weaker* than \mathcal{T}_2 if $\mathcal{T}_1 \subseteq \mathcal{T}_2$. We say \mathcal{T}_1 is *stronger* than \mathcal{T}_2 if $\mathcal{T}_1 \supseteq \mathcal{T}_2$.

Example 13.8. Let \mathcal{T}_1 be the discrete topology on X , and \mathcal{T}_2 be the trivial topology on X . The trivial topology is weaker than the discrete topology as

$$\mathcal{T}_2 = \{\emptyset, X\} \subseteq \mathcal{P}(X) = \mathcal{T}_1.$$

We can actually go a step further. Let \mathcal{T}_3 be an arbitrary topology on X . The collection \mathcal{T}_3 must contain $\{\emptyset, X\}$, so the trivial topology is weaker than \mathcal{T}_3 . Any topology is a subset of $\mathcal{P}(X)$, so the discrete topology is stronger than \mathcal{T}_3

$$\mathcal{T}_2 = \{\emptyset, X\} \subseteq \mathcal{T}_3 \subseteq \mathcal{P}(X) = \mathcal{T}_1.$$

It turns out that the trivial and discrete topologies are the weakest and strongest topologies respectively.

If topologies don't necessarily "come from" metrics, then where do they come from? There must be some easier way to define topologies than by explicitly defining all the open sets in \mathcal{T} . It turns out there is.

Definition 13.12. Let X be a set and $\mathcal{B} \subseteq \mathcal{P}(X)$ be a collection of subsets of X . We say \mathcal{B} is a *basis for a topology on X* if:

1. For all $x \in X$ there exists a $B \in \mathcal{B}$ such that $x \in B$.
2. If $x \in B_1 \cap B_2$ for $B_1, B_2 \in \mathcal{B}$, then there exists some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$.

We refer to the elements of \mathcal{B} as *basis elements*.

As the name implies, a basis \mathcal{B} can be used to define a topology on X .

Definition 13.13. Given a basis \mathcal{B} for a topology on X , the *topology generated by \mathcal{B}* is defined as

$$\mathcal{T}_{\mathcal{B}} = \{U \subseteq X \mid \text{for all } x \in U \text{ there exists a } B \in \mathcal{B} \text{ s.t } x \in B \subseteq U\}$$

Before we develop some of the machinery required to work with bases for topologies, we'll prove two fairly immediate results.

Proposition 13.2. Given a basis \mathcal{B} for a topology on X , $\mathcal{T}_{\mathcal{B}}$ is a topology.

Proof. We'll show that $\mathcal{T}_{\mathcal{B}}$ satisfies the three axioms given in Definition 13.2.

1. It's vacuously true that $\emptyset \in \mathcal{T}_{\mathcal{B}}$. For all $x \in X$ there exists a $B \in \mathcal{B}$ such that $x \in B \subseteq X$ by Definition 13.11, so $X \in \mathcal{T}_{\mathcal{B}}$.

2. Suppose $\{U_\alpha\}_{\alpha \in A} \subseteq \mathcal{T}_\mathcal{B}$. For $x \in \cup_{\alpha \in A} U_\alpha$, there exists some U_α such that $x \in U_\alpha$. Since $U_\alpha \in \mathcal{T}_\mathcal{B}$, there exists some $B \in \mathcal{B}$ such that $x \in B \subseteq U_\alpha$, so

$$x \in B \subseteq U_\alpha \subseteq \bigcup_{\alpha \in A} U_\alpha,$$

which gives $\cup_{\alpha \in A} U_\alpha \in \mathcal{T}_\mathcal{B}$.

3. Let $U, V \in \mathcal{T}_\mathcal{B}$. For some $x \in U \cap V$, there exists $B_1, B_2 \in \mathcal{B}$ such that $x \in B_1 \subseteq U$ and $x \in B_2 \subseteq V$. The second property of a basis gives the existence of a $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. Since $B_1 \subseteq U$ and $B_2 \subseteq V$, $B_3 \subseteq U \cap V$. Therefore for all $x \in U \cap V$ there exists a B_3 such that $x \in B_3 \in U \cap V$, giving $U \cap V \in \mathcal{T}_\mathcal{B}$ (See Figure 116). We can now repeat this argument for a finite intersection of sets by applying it to two at a time (i.e via induction).

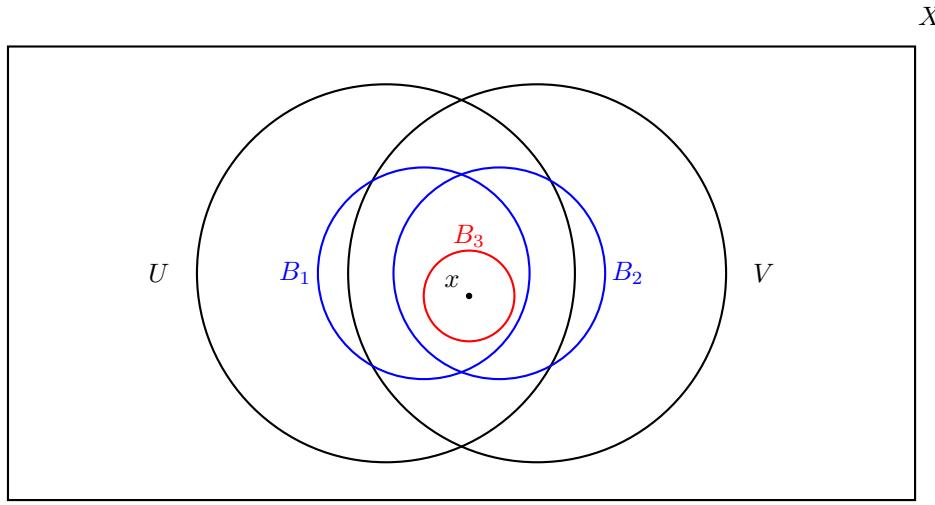


Figure 116:

□

Proposition 13.3. Given a basis \mathcal{B} for a topology on X , $B \in \mathcal{T}_\mathcal{B}$ for all basis elements $B \in \mathcal{B}$. In other words, all basis elements are open sets in $(X, \mathcal{T}_\mathcal{B})$.

Proof. For all $x \in B$, $x \in B \subseteq B$, so B is an element of $\mathcal{T}_\mathcal{B}$ by Definition 13.13. □

Now we can consider a result that lets us leverage topological bases to characterize topological spaces without having to define $\mathcal{T}_\mathcal{B}$ from scratch.

Proposition 13.4. Let \mathcal{B} be a basis on X . The topology generated by \mathcal{B} , $\mathcal{T}_\mathcal{B}$, is the collection of all unions of basis elements $B \in \mathcal{B}$.

$$\mathcal{T}_\mathcal{B} = \left\{ \bigcup_{\beta} B_\beta \mid \{B_\beta\} \subseteq \mathcal{P}(\mathcal{B}) \right\}$$

In other words, any open set $U \in \mathcal{T}_\mathcal{B}$ can be written as a union of basis elements B .

Proof.

(\subseteq) Suppose $U \in \mathcal{T}_{\mathcal{B}}$. For each $x \in U$, there exists some B_x such that $x \in B_x \subseteq U$ by [Definition 13.13](#). If we take the union of all these basis elements, we end up with U .

$$\bigcup_{x \in U} B_x = U$$

Therefore we're able to write all open sets as a union of basis elements, and $\mathcal{T}_{\mathcal{B}} \subseteq \{\cup_{\beta} B_{\beta} \mid \{B_{\beta}\} \subseteq \mathcal{P}(\mathcal{B})\}$

(\supseteq) Suppose $\{B_{\beta}\}$ is a collection of basis elements in \mathcal{B} . Proposition 13.3 established that $B_{\beta} \in \mathcal{T}_{\mathcal{B}}$ for all B_{β} , so the arbitrary union of all $\{B_{\beta}\}$ is also in $\mathcal{T}_{\mathcal{B}}$ by the second property of open sets ([Definition 13.2](#)).

$$\bigcup_{\beta} B_{\beta} \in \mathcal{T}_{\mathcal{B}}$$

Therefore we have $\mathcal{T}_{\mathcal{B}} \supseteq \{\cup_{\beta} B_{\beta} \mid \{B_{\beta}\} \subseteq \mathcal{P}(\mathcal{B})\}$

□

The magic of Proposition 13.4 is that all we need to do if verify a collection of sets is a basis, and we can construct a topology on X by looking at the unions of all the basis elements. This gives us another lens through which we can view familiar examples.

Example 13.9 (Metric Spaces). Suppose (X, d) is a metric space. We can verify that the set of all open balls in X happens to be a basis for the metric topology using [Definition 13.12](#). For each $x \in X$ we have $x \in B_r(x)$ for all $r \geq 0$. Suppose $B_1 = B_r(y)$ and $B_2 = B_q(z)$, and let $x \in B_1 \cap B_2$. Define

$$r^* = \min\{r - d(x, y), q - d(x, z)\}.$$

We can verify that $B_{r^*}(x) \subseteq B_1 \cap B_2$ (drawing a picture here can help). Let $p \in B_{r^*}(x)$.

$$\begin{aligned} d(y, p) &\leq d(y, x) + d(x, p) && \text{(triangle inequality)} \\ &\leq d(x, y) + r^* && (p \in B_{r^*}(x)) \\ &< d(x, y) + (r - d(x, y)) && (\text{def of } r^*) \\ &= r \\ d(z, p) &\leq d(z, x) + d(x, p) && \text{(triangle inequality)} \\ &\leq d(x, z) + r^* && (p \in B_{r^*}(x)) \\ &< d(x, z) + (q - d(x, z)) && (\text{def of } r^*) \\ &= q \end{aligned}$$

Therefore, for an arbitrary $p \in B_{r^*}(x)$, $p \in B_r(y)$ and $p \in B_q(z)$, so

$$p \in B_r(y) \cap B_q(z) = B_1 \cap B_2,$$

giving $B_{r^*}(x) \subseteq B_1 \cap B_2$. This makes the set $\mathcal{B} = \{B_r(x) \mid x \in X, r > 0\}$ a valid basis. By [Definition 13.13](#), the topology generated by \mathcal{B} is

$$\mathcal{T}_{\mathcal{B}} = \{U \subseteq X \mid \text{for all } x \in U \text{ there exists a } B_r(x) \text{ s.t } x \in B_r(x) \subseteq U\}.$$

In other words a set is open if all points in it are interior points. This is the exact same topology as that given by the metric d (Example 13.2, [Definition 2.15](#)). If we combine this with Proposition 13.4, we have that any open set in a metric space can be written as the union of open balls.

Example 13.10 (Real Line). A special case of Example 13.8 is given by $X = \mathbb{R}$ equipped with the Euclidean metric. In this case the basis is the set of all open intervals of the form (a, b) .

Definition 13.14. *subbase*

Definition 13.15. *neighborhood (local) basis*

Example 13.11 (Product Topology). If $(X_1, \mathcal{T}_1), (X_2, \mathcal{T}_2), \dots, (X_n, \mathcal{T}_n)$ are topological spaces then the product topology on $X = \prod_{i=1}^n X_i$ is given by the basis consisting of all the cartesian products of open sets from X_1, \dots, X_n .

$$\mathcal{B} = \{U_1 \times U_2 \times \dots \times U_n \mid U_1 \in \mathcal{T}_1, \dots, U_n \in \mathcal{T}_n\}$$

Is this a valid basis? Since $X_i \in \mathcal{T}_i$ for all i , $X = \prod_{i=1}^n X_i \in \mathcal{B}$. Therefore, $x \in X$ for $X \in \mathcal{B}$ for all $x \in X$. For $U, V \in \mathcal{B}$, we have

$$\begin{aligned} U \cap V &= \left(\prod_{i=1}^n U_i \right) \cap \left(\prod_{i=1}^n V_i \right) && (U_i, V_i \in \mathcal{T}_i) \\ &= \prod_{i=1}^n (U_i \cap V_i) \\ &\in \mathcal{B} && (U_i \cap V_i \in \mathcal{T}_i). \end{aligned}$$

If we let $B_1 = U$, $B_2 = V$, and $B_3 = U \cap V$, then we satisfy the second property of bases. In this case we refer to $\mathcal{T}_{\mathcal{B}}$ as the product topology.

The definition of $\mathcal{T}_{\mathcal{B}}$ made no mention of uniqueness. As the next example shows, we can find two bases \mathcal{B} and \mathcal{B}' such that $\mathcal{T}_{\mathcal{B}'} = \mathcal{T}_{\mathcal{B}}$.

Example 13.12. Let $X = \mathbb{R}^2$. The standard topology on \mathbb{R}^2 is a special case of the product topology.

Example 13.13 (Order Topology). Suppose X is a set equipped with a binary relation \triangleleft such that for all $x, y, z \in X$: $x \triangleleft x$ (reflexivity), if $x \triangleleft y$ and $y \triangleleft z$ then $x \triangleleft z$ (transitivity), if $x \triangleleft y$ and $y \triangleleft x$ then $x = y$ (antisymmetry), and $x \triangleleft y$ or $y \triangleleft x$ (totality).

Example 13.14 (Lexicographic Order Topology on \mathbb{R}). content...

Remark 13.2 (More About Bases). We've shown that for any given basis \mathcal{B} , we can generate a topology $\mathcal{T}_{\mathcal{B}}$. Can we work in the reverse direction? Given some topology \mathcal{T} , can we find some basis \mathcal{B} such that $\mathcal{T} = \mathcal{T}_{\mathcal{B}}$? It turns out we can. For more details on this, see Lemma 13.2 of [Munkres \(2000\)](#).

13.2 Countability and Separation Axioms

The definition of a topological space, along with the accompanying definitions involving bases, can be a bit daunting because the concepts involved are so general. **Definition 13.2** is wide open (pun intended) as far as the properties of \mathcal{T} are concerned. In many settings it happens to be *too* broad. To establish meaningful results about a topological space (X, \mathcal{T}) , we can introduce a handful of additional properties which come in two flavors:

Countability: There exists some basis for a topology that meets certain requirements with respect to countability.

Separation: Points in a topological set X can be “separated” from each other using some type of sets.

13.3 Continuity

13.4 Nets

13.5 Filters

13.6 Various Notions of Compactness

13.7 The Product Topology

13.8 Pointwise and Uniform Convergence

14 Measures

In Section 6 we developed the theory of the Riemann integral for real bounded functions. This was generalized to multivariable functions in higher dimensions in Section 10. In both contexts, we interpret the integral of a function to be some area/volume associated with the function on a set (opposed to the interpretation from Section 12). In doing this, we needed to make some assumptions about length and volume in \mathbb{R}^n . Namely, we assumed that the length of an interval $[a, b] \subseteq \mathbb{R}$ was $b - a$, and the volume of a rectangle $Q = [a_1, b_1] \times \dots \times [a_n, b_n]$ in \mathbb{R}^n is $(b_1 - a_1) \times \dots \times (b_n - a_n)$. A more sophisticated definition of volume in \mathbb{R}^n was defined as the Jordan content ([Definition 10.7](#)) of a set. Finally, as early as Section 6, we introduced the concept of null sets in \mathbb{R}^n ([Definition 6.12](#), [Definition 10.6](#)). These were sets that we could for all intents and purposes “ignore” when integrating (Theorem 6.8, Theorem 10.2). Unfortunately, we saw that our definition of a null set can be incompatible with Jordan content ([Example 10.11](#)). This is all to say that the area/volume of a set is *really* important when it comes to integration, and developing a rigorous definition of area/volume may be even more complicated than Jordan content. This entire section will be dedicated to developing such a definition.

14.1 Motivating Example and Problem Statement

Suppose want to assign some *measure* to the subsets of \mathbb{R} . For instance we may want to know the measure of $[a, b] \subseteq \mathbb{R}$. The term measure will be used as a catch-all generalization for “length”, “area”, and “volume”. We want to develop some function $\mu : \mathcal{P}(\mathbb{R}) \rightarrow [0, \infty]$ that systematically assigns the each subset of \mathbb{R} to some positive real-valued measurement. How would we want μ to behave such that it’s properties are consistent with concepts like length, area, and volume?

Firstly, the measure of a set $E \subseteq \mathbb{R}$ shouldn’t change if E “moves”. For instance, the area of a rectangle shouldn’t depend on it’s location or orientation in space. In geometric terms, we want μ to be *translation invariant*. If we translate, rotate, or reflect E to get some new set E' , then $\mu(E) = \mu(E')$. We also want μ to “conserve measure” in the sense that if we “cut” E up in separate sets $\{E_1, E_2, \dots\}$, then the sum of the measures of E_i should be the measure of E . Measure shouldn’t magically disappear during the process of partitioning E into a collection of sets. We also want $\mu(E)$ to be consistent with our rudimentary definitions of length and area. If $E = [a, b] \subseteq \mathbb{R}$ then $\mu(E) = b - a$.

Unfortunately, such a function defined on all of $\mathcal{P}(X)$ for $X \subseteq \mathbb{R}$ does not exist. We can illustrate this by constructing a set first introduced by [Vitali \(1905\)](#).

Example 14.1 (Vitalli Set). Start by defining the following equivalence relation on the set $[0, 1]$:

$$x \sim y := x - y \in \mathbb{Q} \quad (x, y \in [0, 1])$$

We can think of y as x “shifted” to the left or right by some rational number when $x \sim y$. For instance if $x = e$ and $y = e + 1/2$, then $x \sim y$ because y can be interpreted as x shifted by some rational number (in this case $1/2$). On the other hand, if $x = e$ and $y = \pi$, then $x \not\sim y$ because $\pi \notin \mathbb{Q}$. The equivalence classes associated with \sim are given as

$$[x] = \{y \in [0, 1] \mid x \sim y\} = \{y \in [0, 1] \mid x - y \in \mathbb{Q}\} = \{q + x \mid q \in \mathbb{Q}, q + x \in [0, 1]\}.$$

Like any collection of equivalence classes, the set of classes $\{[x] \mid x \in [0, 1]\}$ partition the set $[0, 1]$.

$$\begin{aligned} \bigcup_{x \in [0, 1]} [x] &= [0, 1] \\ [x] \cap [x'] &= \emptyset \quad (\text{for all } x, x' \in [0, 1]) \end{aligned}$$

Let’s pick one element from each equivalence class to form a set $V \subset [0, 1]$.¹⁷⁶ By construction we have a bijection between elements of V and equivalence classes $[x]$. The set V is referred to as a *Vitalli set*.

We will show the desired properties of μ are not mutually consistent. We’ll do this by writing $\mu(V)$ as the sum of μ evaluated at two disjoint sets. Define a “shifted” version of the V by some rational $q \in \mathbb{Q} \cap [0, 1]$ in two steps. In the first step we’ll shift the elements of V to the right by q , which we can express as $V + q$. The shifted set $V + q$ is no longer a subset of $[0, 1]$, so we’ll “cut off” the portion of the set which “sticks out” to the right and shift it back to the right one unit. After this second step we have a set $V_q \subset [0, 1]$.

$$V_q = \underbrace{\{x + q \mid x \in V \cap [0, 1 - q]\}}_{\text{Shifted to the right } q} \cup \underbrace{\{x + q - 1 \mid q \in V \cap [1 - q, 1]\}}_{\text{Shifted to the right } q \text{ then to left } 1}$$

The two sets that make up V_q are disjoint because the elements of V each correspond to different equivalence classes. We could also write V_q in terms of modular arithmetic.

$$V_q = \{x + q \pmod{1} \mid x \in V\}$$

For any two $q, r \in \mathbb{Q}$, we have $V_q \cap V_r = \emptyset$. If these sets weren’t disjoint, then there would exist some $x \in V_q \cap V_r$ such that either $x - q$ and $x - r$ are in the same equivalence class, or $x - q + 1$ and $x - r + 1$ are in the same equivalence class. This cannot be the case though because we constructed V such that no two elements are in the same equivalence class. We also have that

$$\bigcup_{q \in \mathbb{Q} \cap [0, 1]} V_q = [0, 1],$$

so we can write $[0, 1]$ as a disjoint union of $\{V_q\}$.

Under the desired properties of μ we have

$$\mu([0, 1]) = \mu \left(\bigcup_{q \in \mathbb{Q} \cap [0, 1]} V_q \right) = \sum_{q \in \mathbb{Q} \cap [0, 1]} \mu(V_q),$$

¹⁷⁶Are we allowed to do this? As long as we assume the [axiom of choice](#) holds then yes. If we don’t make this assumption, then [Solovay \(1970\)](#) shows all sets can be measured. If you’re interested in the axiomatic foundation of mathematics, then this is a rabbit hole worth going down.

as the measure of $[0, 1]$ is “preserved” when partitioned by $\{V_q\}$. This property, along with the translation invariance of μ , also give

$$\mu(V) = \mu(V \cap [0, 1 - q]) + \mu(V \cap [0, 1 - q]) = \mu(V_q).$$

Finally, it should be the case that $\mu([0, 1]) = 1$, otherwise μ wouldn’t be consistent with any of our previous intuition about length. Therefore,

$$\sum_{q \in \mathbb{Q} \cap [0, 1]} \mu(V) \stackrel{?}{=} \mu([0, 1]) = 1.$$

Unfortunately, this cannot be the case. A infinite sum of a number $\mu(v) \in [0, \infty]$ will equal 0 if $\mu(V) = 0$, otherwise it will tend to infinity. It isn’t possible for it to equal 1!

This contradiction leaves us with three possible options:

1. Just keep using the Riemann integral without worrying about the measure-related underpinnings, but in doing so commit to a life of blissful mathematical ignorance where we can only integrate real functions.
2. Define μ to have a weaker set of properties, even if that may violate the intuition about length and area that we’ve developed since birth.
3. Concede that we cannot define μ on the set $\mathcal{P}(X)$ for $X \subseteq \mathbb{R}$, and instead focus on some subset of $\mathcal{M} \subseteq \mathcal{P}(X)$ which are “measurable”. If $X = [0, 1]$ then it appears we’ll have $\mathcal{M} \subset \mathcal{P}(X)$, but perhaps there are situations where $\mathcal{M} = \mathcal{P}(X)$.

In this case, the lesser of three evils seems to be the third course of action. Our contradiction was the product of a comically pathological construction, so hopefully it will be the case that $\mathcal{M} \subseteq \mathcal{P}(X)$ will include any set would ever want to measure.

Up until this point, we’ve proposed desirable properties of μ in the context of \mathbb{R} . Before formally developing the basics of measure theory, let’s broaden our perspective. Let’s consider measuring subsets of an arbitrary set X . We won’t require X to be a metric space, vector space, or even a topological space. Will the desired properties of μ differ when $X \neq \mathbb{R}$? They certainly should be, considering one of the properties we discussed was that $\mu([a, b]) = b - a$ when $[a, b] \subset \mathbb{R}$. In general, X may not have “intervals”, because the elements of X may not be ordered. Instead we’ll want μ to satisfy $\mu(\emptyset) = 0$. We may not have have some precise prior idea of what measure looks like in X (like we do with real intervals and length), but we would think that the measure of the trivial subset $\emptyset \subseteq X$ which is common to all X is associated with the “trivial” number 0. We also cannot require that μ be translation invariant, as this is property requires defined operations on subsets of X , which we may not have. We’ll come back to translation invariance later, but in general we won’t require μ to satisfy it in an effort to be as general as possible with our treatment of X .¹⁷⁷

In sum, our goal is to find some set $\mathcal{M} \subseteq \mathcal{P}(X)$ that contains as many sets as possible, and then define $\mu : \mathcal{M} \rightarrow [0, \infty]$ such that:

¹⁷⁷Wait, but wasn’t translation invariance used to get our contradiction in Example 14.1?! If we get ride of this, then perhaps we can define μ on all of $\mathcal{P}(X)$. As it turns out, we can find some elements of $\mathcal{P}(X)$ that μ is not well defined on for non-trivial μ , and do so without using any additional properties of X that may limit our scope. For more details about this see Chapter 5 of [Oxtoby \(2013\)](#) and [Ulam \(1930\)](#). A better way of thinking of Example 14.1 is as a demonstration that measuring sets is complicated, not a *bona fide* proof of the fact that we need to define μ on some domain $\mathcal{M} \subseteq \mathcal{P}(X)$.

1. $\mu(\emptyset) = 0$.
2. $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for disjoint $A_i \in \mathcal{M}$.

14.2 Constructing Measures

At this point in the introduction of measure theory, it's very convenient to "jump ahead" and introduce formal for the function μ and the collection of sets which will serve as its domain, and put off justifying the latter definition (via a result that is known as Carathéodory's extension theorem). I'm going to take a more constructive approach and use Carathéodory's extension theorem to motivate definitions.

We can start by defining a prototype for the function μ that is well defined on $\mathcal{P}(X)$, but doesn't have all the properties of length/area that we're used to. To do this we'll weaken one of the desired properties of μ . Instead of having $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for disjoint $A_i \in \mathcal{M}$, we will only assume

$$\mu^*(\cup_{i=1}^{\infty} B_i) \leq \sum_{i=1}^{\infty} \mu^*(B_i)$$

for any $B_i \in \mathcal{P}(X)$ (even if the sets are not disjoint). In other words, μ^* evaluated at a union cannot exceed the sum of μ^* at each set. This property shouldn't be too controversial if we consider it in the context of \mathbb{R} . For example, suppose $A = [0, 1]$ and $B = [0, 2]$. The length of $A \cup B = [0, 2]$ is 2, while the sum of the lengths of A and B is $1 + 2 = 3$. The only thing we lose with this change involves the case where sets are disjoint. If $A = [0, 1]$ and $B = [1, 2]$, then our property requires $A \cup B = [0, 2] \leq 1 + 1 = 2$. In a perfect world, we'd be able to conclude that $\mu^*(\cup_{i=1}^{\infty} B_i) \leq \sum_{i=1}^{\infty} \mu^*(B_i)$ holds with equality in the special case of disjoint sets, but we already saw this is too much to ask for a function defined on all of $\mathcal{P}(X)$.

Definition 14.1. An *outer measure μ^* on a set X* is a function $\mu^* : \mathcal{P}(X) \rightarrow [0, \infty]$ that satisfies:

1. $\mu^*(\emptyset) = 0$.
2. $\mu^*(A) \leq \mu^*(B)$ when $A \subseteq B \subseteq X$ (*monotonicity*).
3. $\mu^*(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$ where $A_i \subseteq X$ (*subadditivity*).

Example 14.2. Define a premeasure on X as

$$\mu^*(A) = \begin{cases} 0 & A = \emptyset \\ 1 & A \neq \emptyset \end{cases}$$

Clearly $\mu^*(\emptyset) = 0$. To verify the second property of μ^* we need to consider three cases. If $A \subseteq B$ where $A \neq \emptyset$ and $B = \emptyset$, then

$$\mu^*(A) = 1 \leq 1 = \mu^*(B).$$

If $A = \emptyset$, then

$$\mu^*(A) = 0 \leq 1 = \mu^*(B).$$

If $B = \emptyset$ then $A = \emptyset$, giving

$$\mu^*(A) = 0 \leq 0 = \mu^*(B)$$

Therefore the second property holds. Finally let $\{A_i\}$ be a collection of subsets of X . In the event $A_i = \emptyset$ for all i , then

$$\mu^*(\cup_{i=1}^{\infty} A_i) = \mu^*(\emptyset) = 0 \leq 0 = \sum_{i=1}^{\infty} 0 = \sum_{i=1}^{\infty} \mu^*(A_i).$$

If there exists a single i such that $A_i \neq \emptyset$, then

$$\mu^*(\cup_{i=1}^{\infty} A_i) = 1 \leq \sum_{i=1}^{\infty} \mu^*(A_i),$$

since $\mu^*(A_i) = 1$ for at least one i . This makes μ^* a valid premeasure.

So where do premeasures come from? Example 14.2 seems a bit artificial as μ^* was defined with the express purpose of satisfy the required properties of a premeasure, not the purpose of actually measuring subsets of X . Put another way, when we were considering measuring subsets of \mathbb{R} , we wanted intervals $[a, b]$ to have a measure of $b - a$ because this is how we've always thought about length. In general, suppose we have some prior idea of what values the measure of set should take on for some $\mathcal{E} \subseteq \mathcal{P}(X)$ given by a function $\rho : \mathcal{E} \rightarrow \mathcal{P}(X)$. In the case of $X = \mathbb{R}$ we had $\rho([a, b]) = b - a$ for subsets of the form $[a, b]$. Is it possible to go from ρ to some premeasure μ^* that satisfies the properties of [Definition 14.1](#), and if it is then how is this done? This is the subject of the first theorem in this section.

Theorem 14.1. Let $\mathcal{E} \subseteq \mathcal{P}(X)$ be a collection of subsets of X and $\rho : \mathcal{E} \rightarrow [0, \infty]$ be *any* set function such that $\emptyset, X \in \mathcal{E}$ and $\rho(\emptyset) = 0$. For any $A \subseteq X$,

$$\mu^*(A) = \inf \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\}$$

is a premeasure on X , and will be referred to as *the outer measure induced by ρ* .

Proof. First we must verify the function μ^* is well defined. We've assumed $X \in \mathcal{E}$, so we always have $A \subseteq \cup_{i=1}^{\infty} E_i$ when $E_i \in \mathcal{E}$ for all i . This means the set over which the infimum is taken is never empty. The infimum is also guaranteed to exist because the codomain of ρ is $[0, \infty] \subseteq \bar{\mathbb{R}}$, and the extended real numbers are complete. Now we will check that $\mu^*(A)$ satisfies the properties given by [Definition 14.1](#).

1. If $A = \emptyset$, then let $E_i = \emptyset$ for all j . We're able to do this because we've assumed $\emptyset \in \mathcal{E}$. We have

$$A = \emptyset \subseteq \emptyset = \bigcup_{i=1}^{\infty} \emptyset = \bigcup_{i=1}^{\infty} E_i.$$

If we sum the values of ρ calculated on this collection $\{E_i\}$ we have

$$\sum_{i=1}^{\infty} \rho(E_i) = \sum_{i=1}^{\infty} \rho(\emptyset) = \sum_{i=1}^{\infty} 0 = 0.$$

Since the domain of ρ is $[0, \infty]$, this is the infimum of interest. This gives $\mu^*(\emptyset) = 0$.

2. Suppose $A \subseteq B$. If $B \subseteq \cup_{i=1}^{\infty} E_i$, then $A \subseteq \cup_{i=1}^{\infty} E_i$, so

$$\left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } B \subseteq \bigcup_{i=1}^{\infty} E_i \right\} \subseteq \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\}.$$

The infimum of a set is always weakly greater than the infimum of its subsets.

$$\begin{aligned} & \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } B \subseteq \bigcup_{i=1}^{\infty} E_i \right\} \subseteq \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\} \\ \implies & \inf \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } B \subseteq \bigcup_{i=1}^{\infty} E_i \right\} \geq \inf \left\{ \sum_{i=1}^{\infty} \rho(E_i) \mid E_i \in \mathcal{E} \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\} \\ \implies & \mu^*(B) \geq \mu^*(A) \end{aligned}$$

Thus we have the desired inequality $\mu^*(A) \leq \mu^*(B)$.

3. Suppose $\{A_j\} \subset \mathcal{P}(X)$, and let $\varepsilon > 0$ be arbitrary. For each j there exists some $\{E_i^j\} \subseteq \mathcal{E}$ such that $A_j \subseteq \bigcup_{i=1}^{\infty} E_i^j$ and

$$\begin{aligned} \sum_{i=1}^{\infty} \rho(E_i^j) &\leq \inf \left\{ \sum_{k=1}^{\infty} \rho(E_k^j) \mid E_k^j \in \mathcal{E} \text{ and } A_j \subseteq \bigcup_{k=1}^{\infty} E_k^j \right\} + \frac{\varepsilon}{2^j}, & (\forall j \text{ and } \forall \varepsilon > 0) \\ \implies \sum_{i=1}^{\infty} \rho(E_i^j) &\leq \mu^*(A_j) + \frac{\varepsilon}{2^j} & (\forall j \text{ and } \forall \varepsilon > 0) \\ \implies \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \rho(E_i^j) &\leq \sum_{j=1}^{\infty} \left(\mu^*(A_j) + \frac{\varepsilon}{2^j} \right) & (\text{sum over } j) \\ \implies \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \rho(E_i^j) &\leq \sum_{j=1}^{\infty} \mu^*(A_j) + \varepsilon & (\sum_{j=1}^{\infty} 1/2^j = 1) \\ \implies \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \rho(E_i^j) &\leq \sum_{j=1}^{\infty} \mu^*(A_j) & (\text{previous inequality held } \forall \varepsilon > 0). \end{aligned}$$

Define $A = \bigcup_{j=1}^{\infty} A_j$. We have

$$A \subseteq \bigcup_{j=1}^{\infty} \bigcup_{i=1}^{\infty} E_i^j,$$

so it must be the case that

$$\mu^*(A) \leq \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \rho(E_i^j)$$

as $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \rho(E_i^j)$ is a member of the set which $\mu^*(A)$ is the infimum of. Therefore

$$\mu^*(A) \leq \sum_{j=1}^{\infty} \mu^*(A_j).$$

□

The premeasure μ^* given by Theorem 14.1 should actually seem a little familiar once we think about it at a higher level. We start with some function ρ that gives the measure of sets on a specific type of subset of X , the collection of which are \mathcal{E} . We then determine $\mu^*(A)$ for $A \in \mathcal{P}(X)$ by covering A with a collection of sets $\{E_i\} \subseteq \mathcal{E}$. We calculate the measure of the cover as given by ρ for each cover, $\sum_{i=1}^{\infty} \rho(E_i)$,¹⁷⁸ and then we find the infimum of $\sum_{i=1}^{\infty} \rho(E_i)$ over all covers. This infimum is $\mu^*(A)$. We're effectively approximating the size of A with elements of \mathcal{E} and taking μ^* to be the limit of this approximation process. It's the same idea we used to define the upper Riemann integral (Definition 6.4) and the outer Jordan content (Definition 10.7). Figure 117 illustrates this approximation process.

Example 14.3 (Outer Measure on \mathbb{R}). Let $X = \mathbb{R}$, $\mathcal{E} = \{[a, b] \mid a, b \in \mathbb{R}, a < b\}$, and $\rho([a, b]) = b - a$. Clearly,

$$\mu^*([a, b]) = b - a.$$

Since μ^* is defined on the entire collection $\mathcal{P}(X)$, it can even be defined for sets constructed like the Vitali set in Example 14.1.

¹⁷⁸We're not assuming anything about the sets E_i being disjoint here, so $\sum_{i=1}^{\infty} \rho(E_i)$ may "double count" some measure.

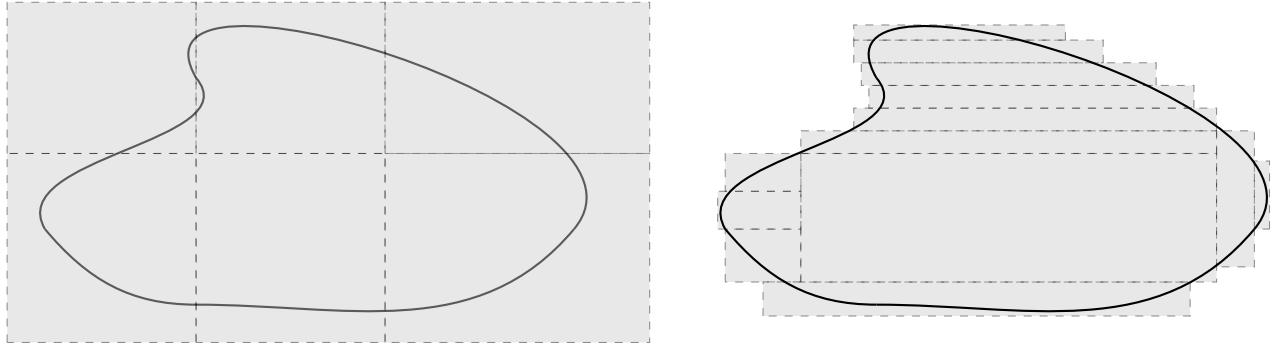


Figure 117:

Theorem 14.2 (Carathéodory's Extension Theorem). If μ^* is an outer measure on X we can define the collection of sets

$$\mathcal{M} = \{A \subseteq X \mid \mu^*(E) = v \text{ for all } E \subseteq X\} \subseteq \mathcal{P}(X).$$

We refer to elements of \mathcal{M} as *μ^* -measurable* and particular choices of $E \subset X$ *test sets*. The collection \mathcal{M} has two properties :

1. If $A \in \mathcal{M}$, then $A^c \in \mathcal{M}$.
2. If $\{A_j\} \subseteq \mathcal{M}$, then $\cup_{j=1}^{\infty} A_j \in \mathcal{M}$.

Furthermore, if we define $\mu = \mu^*|_{\mathcal{M}}$ (the outer measure μ^* restricted to the domain \mathcal{M}), then μ satisfies:

1. $\mu(\emptyset) = 0$.
2. $\mu(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j)$ for disjoint $A_i \in \mathcal{M}$.

Proof. First we will verify the properties of \mathcal{M} .

1. Suppose $A \in \mathcal{M}$. By the definition of \mathcal{M} we have:

$$\begin{aligned} \mu^*(E) &= \mu^*(E \cap A) + \mu^*(E \cap A^c) && (\forall E \subseteq X) \\ \implies \mu^*(E) &= \mu^*(E \cap (A^c)^c) + \mu^*(E \cap A^c) && (\forall E \subseteq X) \\ \implies \mu^*(E) &= \mu^*(E \cap A^c) + \mu^*(E \cap (A^c)^c) && (\forall E \subseteq X) \\ \implies A^c &\in \mathcal{M}. \end{aligned}$$

2. Proving this property is a bit trickier. We'll first show that \mathcal{M} is closed under finite unions. It suffices to show that $A \cup B \in \mathcal{M}$ for $A, B \in \mathcal{M}$, and then apply the argument inductively to reach the conclusion for any finite collection of sets in \mathcal{M} . By the definition of \mathcal{M} , for all $E \subseteq X$ we have

$$\begin{aligned} \mu^*(E) &= \mu^*(E \cap A) + \mu^*(E \cap A^c) && (A \in \mathcal{M}) \\ &= \mu^*(E \cap A) + \mu^*[(E \cap A^c) \cap B] + \mu^*[(E \cap A^c) \cap B^c] && (B \in \mathcal{M}, E \cap A^c \subseteq X) \\ &= \mu^*(E \cap A) + \mu^*[(E \cap A^c) \cap B] + \mu^*[E \cap (A \cup B)^c] && ((E \cap A^c) \cap B^c = E \cap (A \cup B)^c) \\ &\geq \mu^*[(E \cap A) \cup (E \cap A^c) \cap B] + \mu^*[E \cap (A \cup B)^c] && (\mu^* \text{ subadditive}) \\ &= \mu^*[E \cap (A \cup B)] + \mu^*[E \cap (A \cup B)^c] && ((E \cap A) \cup (E \cap A^c) \cap B = E \cap (A \cup B)) \end{aligned}$$

But we also have that $\mu^*(E) \leq \mu^*[E \cap (A \cup B)] + \mu^*[E \cap (A \cup B)^c]$ because μ^* is subadditive and $[E \cap (A \cup B)] \cup [E \cap (A \cup B)^c] = E$. Therefore

$$\begin{aligned} \mu^*[E \cap (A \cup B)] + \mu^*[E \cap (A \cup B)^c] &\leq \mu^*(E) \leq \mu^*[E \cap (A \cup B)] + \mu^*[E \cap (A \cup B)^c], \quad (\forall E \subseteq X) \\ \implies \mu^*(E) &= \mu^*[E \cap (A \cup B)] + \mu^*[E \cap (A \cup B)^c], \quad (\forall E \subseteq X) \\ \implies A \cup B &\in \mathcal{M} \end{aligned}$$

Now consider the actual case we're interested in, that being a countably infinite collection $\{A_j\} \subseteq \mathcal{M}$. Without loss of generality, we will assume that the collection of sets are all disjoint.¹⁷⁹ We'll start by working with a finite subset $\{A_1, \dots, A_n\} \subseteq \{A_j\} \subseteq \mathcal{M}$, and eventually arrive at our result by letting $n \rightarrow \infty$. Applying the definition of $A_n \in \mathcal{M}$ to the test set $(\cup_{j=1}^n A_j) \subseteq X$ gives

$$\begin{aligned} \mu^*(E \cap (\cup_{j=1}^n A_j)) &= \mu^*(E \cap (\cup_{j=1}^n A_j) \cap A_n) + \mu^*(E \cap (\cup_{j=1}^n A_j) \cap A_n^c), \\ &= \mu^*(E \cap A_n) + \mu^*(E \cap (\cup_{j=1}^{n-1} A_j)). \end{aligned}$$

Now take $(\cup_{j=1}^{n-1} A_j) \subseteq X$ to be the test set and the μ^* -measurable set to be A_{n-1} , allowing us to rewrite $\mu^*(E \cap (\cup_{j=1}^{n-1} A_j))$.

$$\begin{aligned} \mu^*(E \cap (\cup_{j=1}^n A_j)) &= \mu^*(E \cap A_n) + \mu^*(E \cap (\cup_{j=1}^{n-1} A_j)) \\ &= \mu^*(E \cap A_n) + [\mu^*(E \cap A_{n-1}) + \mu^*(E \cap (\cup_{j=1}^{n-2} A_j))] \\ &= [\mu^*(E \cap A_n) + \mu^*(E \cap A_{n-1})] + \mu^*(E \cap (\cup_{j=1}^{n-2} A_j)) \\ &= \sum_{j=n-1}^n \mu^*(E \cap A_j) + \mu^*(E \cap (\cup_{j=1}^{n-2} A_j)) \end{aligned}$$

We can do this $n - 2$ more times, at each step k taking $\cup_{j=n-k+1}^n A_j$ to be the test set in the definition of $A_{n-k+1} \in \mathcal{M}$, which gives

$$\begin{aligned} \mu^*(E \cap (\cup_{j=1}^n A_j)) &= \sum_{j=1}^n \mu^*(E \cap A_j) + \underbrace{\mu^*(E \cap A_1 \cap A_1^c)}_{\mu^*(\emptyset)} \\ \implies \mu^*(E \cap (\cup_{j=1}^n A_j)) &= \sum_{j=1}^n \mu^*(E \cap A_j) \end{aligned} \tag{66}$$

If we apply the subadditivity of μ^* to $E \cap (\cup_{j=1}^n A_j)$ and $E \cap (\cup_{j=1}^n A_j)^c$ for the arbitrary test set $E \subseteq X$

¹⁷⁹Wait, why can we do this? Suppose we have an arbitrary union of sets $U \cup V$. We can always write this union as $[U \cap V^c] \cup [U \cap V] \cup [U^c \cap V]$. This type of argument generalizes to any case where we have a countable number of sets.

we have

$$\begin{aligned}
& \mu^*([E \cap (\cup_{j=1}^n A_j)] \cup [E \cap (\cup_{j=1}^n A_j)^c]) \geq \mu^*(E \cap (\cup_{j=1}^n A_j)) + \mu^*(E \cap (\cup_{j=1}^n A_j)^c), \\
\implies & \mu^*(E) = \mu^*(E \cap (\cup_{j=1}^n A_j)) + \mu^*(E \cap (\cup_{j=1}^n A_j)^c), \\
\implies & \mu^*(E) \geq \sum_{j=1}^n \mu^*(E \cap A_j) + \mu^*(E \cap (\cup_{j=1}^n A_j)^c), \\
\implies & \lim_{n \rightarrow \infty} \mu^*(E) \geq \lim_{n \rightarrow \infty} \sum_{j=1}^n \mu^*(E \cap A_j) + \lim_{n \rightarrow \infty} \mu^*(E \cap (\cup_{j=1}^n A_j)^c), \\
\implies & \mu^*(E) \geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c), \\
\implies & \mu^*(E) \geq \mu^*(\cup_{j=1}^{\infty} (E \cap A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c), \quad (\mu^* \text{ subadditive}) \\
\implies & \mu^*(E) \geq \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c).
\end{aligned} \tag{Equation 66}$$

But by the subadditivity of μ^* we also have $\mu^*(E) \geq \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c)$, so just like in the case with the finite union we have

$$\begin{aligned}
& \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c) \leq \mu^*(E) \leq \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c), \quad (\forall E \subseteq X) \\
\implies & \mu^*(E) = \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c), \quad (\forall E \subseteq X) \\
\implies & (\cup_{j=1}^{\infty} A_j) \in \mathcal{M}.
\end{aligned}$$

Now we'll verify that $\mu = \mu^*|_{\mathcal{M}}$ satisfies the claimed properties.

1. For all $E \subseteq X$ we have

$$\begin{aligned}
& \mu^*(E) = \mu^*(E \cap X), \\
\implies & \mu^*(E) = 0 + \mu^*(E \cap \emptyset^c), \\
\implies & \mu^*(E) = \mu^*(\emptyset) + \mu^*(E \cap \emptyset^c), \quad (\mu^*(\emptyset) = 0) \\
\implies & \mu^*(E) = \mu^*(E \cap \emptyset) + \mu^*(E \cap \emptyset^c), \\
\implies & \emptyset \in \mathcal{M}, \\
\implies & \mu(\emptyset) = \mu^*(\emptyset), \quad (\mu = \mu^*|_{\mathcal{M}}) \\
\implies & \mu(\emptyset) = 0.
\end{aligned}$$

2. Suppose $\{A_j\} \subseteq \mathcal{M}$. We already showed that $\cup_{j=1}^{\infty} A_j \in \mathcal{M}$, so μ is well defined for this union and will take on the value given by μ^* . Recall that when we proved that $\cup_{j=1}^{\infty} A_j \in \mathcal{M}$, we established

$$\mu^*(E) \geq \sum_{j=1}^{\infty} \mu^*(E \cap A_j) + \mu^*(E \cap (\cup_{j=1}^{\infty} A_j)^c) \quad (\forall E \subseteq X).$$

Let's take the test set to be $E = \cup_{j=1}^{\infty} A_j \subseteq X$ giving

$$\mu^*(\cup_{j=1}^{\infty} A_j) \geq \sum_{j=1}^{\infty} \mu^*(A_j).$$

The reverse of this weak inequality holds because μ^* is subadditive, so we have

$$\begin{aligned} \sum_{j=1}^{\infty} \mu^*(A_j) &\leq \mu^*\left(\cup_{j=1}^{\infty} A_j\right) \leq \sum_{j=1}^{\infty} \mu^*(A_j), \\ \implies \mu^*(A_j) &= \mu^*\left(\cup_{j=1}^{\infty} A_j\right), \\ \implies \mu(A_j) &= \mu\left(\cup_{j=1}^{\infty} A_j\right). \quad (\mu = \mu^*|_{\mathcal{M}}, \cup_{j=1}^{\infty} A_j \in \mathcal{M}) \end{aligned}$$

□

Theorem 14.2 tells us exactly how we can go from an outer measure to a function $\mu : \mathcal{M} \rightarrow [0, \infty]$ that satisfies all the properties we want of μ ! We define \mathcal{M} to be the collection of all sets that are “well-behaved” in the sense that if we intersect it with *any* “test” set $E \subseteq X$, then the outer measure of E should be preserved in the sense that

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

One way to think about this condition, known as *Carathéodory’s criterion*, is that it’s just a special case of countable additivity since $E = (E \cap A) \cup (E \cap A^c)$ where $(E \cap A) \cap (E \cap A^c) = \emptyset$. In this context though it’s *extremely* special because it happens to be a sufficient condition for the subadditivty of μ^* to “become” additivity of $\mu = \mu^*|_{\mathcal{M}}$.

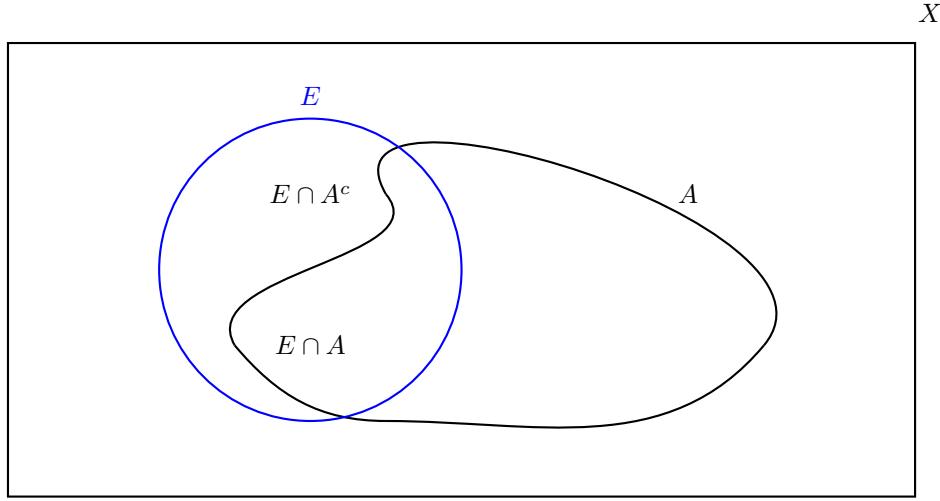


Figure 118: To determine whether or not $A \in \mathcal{M}$, we verify that $\mu^*(E)$ can be given as the sum of the disjoint sets formed by intersection E and $A - E \cap A$ and $E \cap A^c$ – for all $E \subseteq X$.

So we’re done right? Well not really. Theorem 14.1 and Theorem 14.2 provide a blueprint of how to create a specific μ based on some set function ρ . Royden and Fitzpatrick (2010) refer to this particular μ as the *Carathéodory measure induced by ρ* . In general, we may not really care where μ comes from. For this reason we’ll provide general definitions for μ and \mathcal{M} that are completely agnostic about construction (like how Definition 14.1 didn’t tell us where outer measures came from), but satisfy the properties given in Theorem 14.1. The second of these definitions will *finally* be the rigorous definition of “measure” that we’ve been flirting with up until now.

Definition 14.2. A *σ -algebra on X* (read as “sigma-algebra”) is a collection of sets $\mathcal{E} \subseteq \mathcal{P}(X)$ which satisfy the following properties:

1. If $E \in \mathcal{E}$, then $E^c \in \mathcal{E}$ (\mathcal{E} is closed under complements).
2. If $\{E_\alpha\} \subset \mathcal{E}$ is a countable collection of sets, then $\cup_\alpha E_\alpha \in \mathcal{E}$ (\mathcal{E} is closed under countable unions).

Sometimes a third property is given as

3. If $\{E_\alpha\} \subset \mathcal{E}$ is a countable collection of sets, then $\cap_\alpha E_\alpha \in \mathcal{E}$ (\mathcal{E} is closed under countable intersections).

But this is implied by the first two properties and De Morgan's laws.

Definition 14.3. Suppose \mathcal{M} is a σ -algebra on X . A *measure μ on (X, \mathcal{M})* (or on \mathcal{M} or X if the context is clear) is a function $\mu : \mathcal{M} \rightarrow [0, \infty]$ such that

1. $\mu(\emptyset) = 0$.
2. $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for disjoint $A_i \in \mathcal{M}$.

The elements of \mathcal{M} are referred to as *measurable sets*, while the pair (X, \mathcal{M}) is called a *measurable space*. If μ is a measure on (X, \mathcal{M}) then the triple (X, \mathcal{M}, μ) is a *measure space*.

Most treatments of measure theory (such as that given by Folland (1999)) start with **Definition 14.2**, consider properties of σ -algebras, introduce **Definition 14.3**, consider the properties of measures, *and then* talk about constructing measures. I've moved things around here for the sake of motivating the definition of σ -algebras. Now we'll look at properties of σ -algebras and measure spaces in general before introducing what is considered *the* measure on \mathbb{R} that we've been using since gradeschool (we just didn't know that it was justified with rigorous mathematics at the time).

14.3 σ -Algebras

The first question we need to consider when thinking about σ -algebras is whether they exist.

14.4 Measures

dirac, counting,

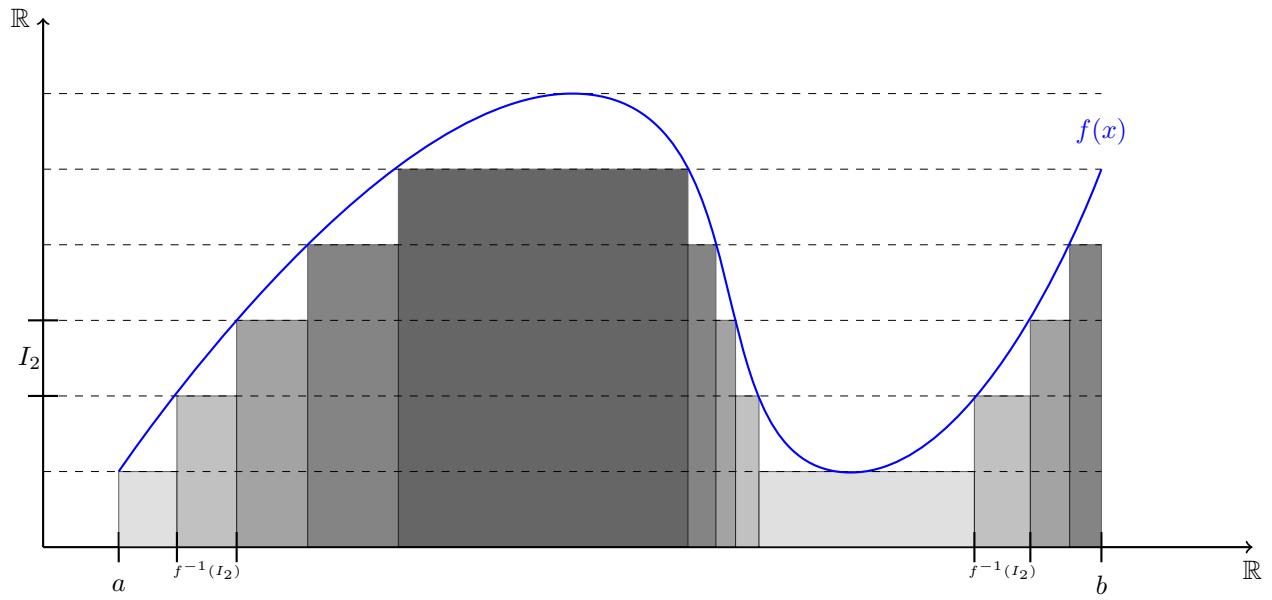


Figure 119: Lebesgue Integral.

14.5 Measures on \mathbb{R} , Lebesgue Measure

14.6 Translation Invariance

15 Integration Revisited

15.1 Measurable Functions

15.2 Integration of Simple Functions

15.3 Integration of Nonnegative Functions

15.4 Integration of Real Functions

15.5 Convergence Revisited

15.6 Product Measures

15.7 Lebesgue Integration in n -Dimensions

16 Differentiation with Measures

16.1 Absolute Continuity in \mathbb{R}

16.2 Signed Measures

16.3 Radon-Nikodym Derivative

Part IV

Functional Analysis

17 Foundations

17.1 Normed Vector Spaces

ℓ^p, L^p

17.2 Important Examples

17.3 Dual Spaces and Hahn-Banach

unbounded in one norm, bounded in another

continuity iff finite dimension iff bounded

17.4 The Baire Category Theorem

17.5 Topological Vector Spaces

17.6 Hilbert Spaces

18 L^p Spaces

18.1 Basic Theory

18.2 The Dual of L^p

18.3 Inequalities

19 Riesz Representation Theorem

19.1 Continuous Functions with Compact Support

19.2 Radon Measures

19.3 Main Theorem

19.4 Related Results and Corollaries

20 Foundations of Fourier Analysis

20.1 Convolutions

21 Operator Theory

22 Distribution Theory

Part V

Probability Theory

23 Introduction to Probability

23.1 Probability Spaces

23.2 Random Variables

23.3 Independence

23.4 Distributions Functions and Densities

23.5 Convergence

23.6 Expectation and Moments

23.7 Characteristic Functions 332

24 Laws of Large Numbers

References

- Apostol, T. M. (1974). *Mathematical Analysis* (2 ed.). Addison Wesley.
- Brown, R. (2006). Topology and groupoids.
- Dummit, D. S. and R. M. Foote (2004). *Abstract Algebra* (3 ed.). John Wiley & Sons.
- Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications* (2 ed.). John Wiley & Sons.
- Lang, S. (2012). *Real and functional analysis*, Volume 142. Springer Science & Business Media.
- Munkres, J. R. (1999). *Analysis on Manifolds* (1 ed.). Addison-Wesley.
- Munkres, J. R. (2000). *Topology* (2 ed.). Addison-Wesley.
- Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences* 36(1), 48–49.
- Oxtoby, J. C. (2013). *Measure and category: A survey of the analogies between topological and measure spaces*, Volume 2. Springer Science & Business Media.
- Royden, H. L. and P. Fitzpatrick (2010). *Real Analysis* (4 ed.). Pearson.
- Rudin, W. (1976). *Principles of Mathematical Analysis* (3 ed.). McGraw-Hill.
- Rudin, W. (1986). *Real and Complex Analysis* (3 ed.).
- Solovay, R. M. (1970). A model of set-theory in which every set of reals is lebesgue measurable. *Annals of Mathematics*, 1–56.
- Spivak, M. (1965). *Calculus on manifolds: a modern approach to classical theorems of advanced calculus*. Addison-Wesley.
- Tao, T. (2016a). *Analysis I* (1 ed.). Hindustan Book Agency.
- Tao, T. (2016b). *Analysis II* (1 ed.). Hindustan Book Agency.
- Ulam, S. M. (1930). *Zur Masstheorie in der allgemeinen Mengenlehre*. Uniwersytet, seminarjum matematyczne.
- Vitali, G. (1905). *Sul problema della misura dei Gruppi di punti di una retta: Nota*. Tip. Gamberini e Parmeggiani.