ANALYSIS

REACHING FOR INFINITY

Steve Trettel

Table of contents

Preface 1

About This Book 3

I. Paradox 9

1. ⋆ Infinite Processes 11
- 1.1. The Diagonal of a Square . . . 11
- 1.2. Quadrature of the Parabola . . . 14
- 1.3. The Circle Constant . . . 19

2. ⋆ Dubious Computations 29
- 2.1. Convergence, Concern and Contradiction . . . 29
- 2.2. Infinite Expressions for $\sin(x)$ . . . 33
- 2.3. The Infinitesimal Calculus . . . 41

II. Numbers 45

3. Operations 47

4. Order 53
- 4.1. Definitions Requiring an Order . . . 55
- 4.2. Working with Inequalities . . . 55
- 4.3. Working with Absolute Values . . . 57
- 4.4. ⋆ Topology . . . 59

5. Completeness 61
- 5.1. Dreaming of Infinity . . . 61
- 5.2. Suprema and Infima . . . 63
- 5.3. Completeness . . . 66
- 5.4. Working with inf and sup . . . 68

6. The Real Numbers 73
- 6.1. Infinites and Infinitesimals . . . 75
- 6.2. Irrationals . . . 78
- 6.3. Uncountability . . . 81

6.4. ⋆ Topology 82

7. ⋆ Infinity **85**
7.1. Order 85
7.2. Arithmetic 86

III. Sequences 87

8. Definitions **89**
8.1. Convergence 91
8.2. Divergence 93
8.3. Uniqueness 94
8.4. ⋆ Topology 96

9. Calculation **97**
9.1. Limits and Inequalities 97
9.2. Limits and Field Operations 101
9.3. Applications 106

10. Monotone Convergence **109**
10.1. Infinite Recursion 110
10.2. ⋆ The Number e 113
10.3. Application: Defining Irrational Powers 116

11. Subsequences **119**
11.1. Continued Fractions 120
11.2. Subsequences and Convergence 126
11.3. The Bolzano Weierstrass Theorem 128

12. Cauchy Sequences **131**
12.1. ⋆ Properties of Cauchy Sequences 133
12.2. Cauchy $\iff$ Convergent 135

IV. Functions 137

13. Definition **139**
13.1. Definition and Examples 141
13.2. Composition and Inverses 143
13.3. Increasing, Decreasing & Convexity 144
13.4. ⋆ Functions of a Real Variable 145

14. Continuity **147**
14.1. Using the $\epsilon - \delta$ Definition 148
14.2. Continuity With Sequences 149

14.3. Analogs of the Limit Theorems . . . 152
14.4. Useful Examples . . . 153
14.5. Continuity and the Rationals . . . 154

15. Transcendental Functions 157
15.1. Functional Equations . . . 157
15.2. Exponentials . . . 160
15.3. Logarithms . . . 162
15.4. Trigonometric . . . 164
15.5. ⋆ Elementary Functions . . . 165

16. Theory 169
16.1. Extreme Values . . . 169
16.2. Intermediate Values . . . 171
16.3. Uniform Continuity . . . 174

17. Limits 177
17.1. Limits from Above and Below . . . 180

V. Series & Products 183

18. Definition 185
18.1. Telescoping . . . 189
18.2. Geometric Series . . . 190

19. Convergence 195
19.1. The Cauchy Criterion . . . 195
19.2. Comparison . . . 197
19.3. The Ratio Test . . . 198
19.4. ⋆ Other Convergence Tests . . . 200
19.5. ⋆ Conditionally Convergent Series . . . 200

20. Power Series 203
20.1. Example Power Series . . . 204
20.2. Power Series for Functions . . . 205

21. Switching Limits 207
21.1. Dominated Convergence (Tannery's Theorem) . . . 208
21.2. Application: Continuity of Power Series . . . 212
21.3. Application: Double Sums . . . 213

VI. Derivatives 217

22. Definition 219
22.1. Using the Definition . . . 220
22.2. The Derivative as a Function . . . 222

23. Properties 223
23.1. Differentiation and Field Operations . . . 224
23.2. The Chain Rule . . . 226
23.3. Exponentials and Logs . . . 228
23.4. The Power Rule . . . 231

24. Theory 235
24.1. Derivatives and Extrema . . . 235
24.2. The Mean Value Theorem . . . 238
24.3. MVT Applications . . . 241

25. Power Series 249
25.1. Differentiating Power Series . . . 249
25.2. Power Series Representations . . . 253
25.3. ⋆ Smoothness & Analyticity . . . 258

26. The Exponential 261
26.1. Prior Work . . . 261
26.2. Finding a Power Series . . . 262

27. ⋆ Exponentials PDEs and ODEs 269
27.1. Generalizing the Exponential . . . 269
27.2. Solving Differential Equations . . . 275

VII. Integrals 285

28. Definition 287
28.1. Properties of Area . . . 288
28.2. Axioms . . . 289
28.3. The Fundamental Theorem . . . 294
28.4. The Work to Come . . . 299

29. Construction 301
29.1. Definition . . . 301
29.2. Partitions . . . 302
29.3. Integrability Criteria . . . 304
29.4. Riemann Sums . . . 307

30. ⋆ Examples **309**
30.1. Some Polynomials . . . 309
30.2. Exponentials . . . 312
30.3. A Logarithm . . . 316
30.4. Using the Freedom of Partition . . . 318

31. Theory **321**
31.1. Verifying the Axioms . . . 321
31.2. Integrability . . . 325
31.3. Linearity . . . 328
31.4. Power Series . . . 330
31.5. ⋆ Dominated Convergence . . . 332

32. π **337**
32.1. π and the Circle . . . 337
32.2. Integrals and Inverse Trigonometry . . . 339
32.3. Series with ArcTan . . . 343

Project **347**
Trigonometry . . . 347
Euler's Identity . . . 353
The Basel Problem . . . 355
Hints . . . 361

Preface

This is the beginnings of a textbook for a course on real analysis. In its current form, it covers a bit more than 1-semester's worth of material (with optional, skippable topics starred).

About This Book

In my opinion, analysis is perhaps the *best* undergraduate course we offer students for several reasons (and I'm a topologist, so this isn't just preference for my own field shining through!):

- **Its Rigorous:** Precision and clarity of thought - the foundations of mathematical proof - are fundamental skills that one learns in a mathematics degree. And, the material of real analysis naturally lends itself to practicing this material: from limits ($\forall\epsilon\exists N\forall n > N\ldots$) to continuity ($\forall\epsilon\exists\delta\forall x$ s.t.$|x - a| < \delta\ldots$) the definitions of intuitive ideas turn out to be more subtle than one originally imagines.
- **Its Familiar:** Many fundamental results in real analysis show up in calculus courses or before: from knowing that the square root is continuous to using differentiation to find maxima and minima. This prior familiarity allows us to wade deeper into technical arguments (and consequently get farther into the material) than in subjects where *both* the concepts and proof techniques are new.
- **There are Surprises:** While the foundational results may be familiar, real analysis is a far cry from just filling in overly-pedantic details to known results. Indeed, the entire theory is a battle to *preserve our intuition* against harsh reality: from the nonexistence of infinitesimals to the wonderful variety of pathological functions there are surprises lurking just below the surface in almost every topic.
- **There's Drama:** Mathematics is more than just a collection of theorems; its a field of study, developed by many humans over several millennia on Earth. This human side to mathematics is crucially important to its its practice; while *the theorems we've proven* rise above the messy contingent details of life on this planet, *which theorems we've proven* is a function of our history. And the history of analysis is rich; from the methods of the ancients to the feud of Newton and Leibniz to the struggles of formalization, important theorems and historical turning points go hand-in-hand.
- **It Bridges the Pure/Applied Divide:** As mathematics grows it may feel a bit fractured into "pure" and "applied" subdisciplines especially at the undergraduate level (not to mention "statistics" and "theoretical physics"; earlier branches of our common family tree). But analysis is fundamental to both sides, laying the foundations for modern Geometry / Lie Theory / Dynamical Systems, as well as Mathematical Modeling, Numerical Analysis and Optimization.

As one of the cornerstones to an undergraduate mathematics degree, there are already many excellent real analysis texts out there. So what makes it worth it to contribute *yet another tome* to the stockpile? The story of analysis is as broad as it is deep, and there are many narratives one can tell: I hope to tell one that emphasizes the points below.

A Sequences Forward Approach

The definition of sequence convergence is one of the first "nested quantifier definitions" to appear in analysis, and proving many theorems about sequences up front provides a fruitful playground for getting used to such definitions (and proof strategies, like the "ϵ-N game").

Taking advantage of all this work done early on, this book takes the *sequence* as the fundamental object in analysis, and develops tools to study other 'nested quantifier type' definitions in terms of sequences. In particular

- We prove that (the standard ϵ-δ) continuity of a function f at a point a is equivalent to the following: for every sequence $x_n \to a$, the sequence $f(x_n)$ converges to $f(a)$. This allows for simple proofs of many facts about continuity building directly off of limits.
- We prove that the $\epsilon-\delta$ notion of limit of a function is equivalent to the following sequence version: $\lim_{x\to a} f(x) = L$ if for all sequences $x_n \to a$ with $x_n \neq a$, the sequence $f(x_n)$ converges to L. This allows one to translate sequence convergence theorems directly to facts about limits of functions, and provides a natural way to work with left and right hand limits.
- After Darboux integrability we discuss the Riemann integral, and its definition involving *all sequences of shrinking partitions* also allows us to prove several properties of integrals from corresponding statements about sequences.

Discovering the Elementary Functions

Some parts of real analysis can be taught completely abstractly, speaking only of functions f and g, and never specifying *particular* functions at all. But other parts of the field are dedicated specifically to understanding and constructing specific functions, from the familiar exponentials logs and trigonometric functions to more esoteric special functions like the gamma function, bessel functions, and jacobi elliptic functions.

This book attempts to show students a bit of both sides of analysis, by building into the main text a construction of the exponential and logarithmic functions, and allowing them to work out a full construction of the trigonometric functions as a final product. We define these elementary functions via *functional equations*, so we call a function an exponential if it is a continuous nonconstant solution to $E(x+y) = E(x)E(y)$ and L a logarithm if its a (continuous, nonconstant) solution to $L(xy) = L(x) + L(y)$.

The work to understand these functions is spread out over several sections of the text: whenever we learn new material (continuity, differentiability, power series, integration) we illustrate it by making more progress on understanding the elementary functions. One of the highlights of the course is the construction of the exponential as a power series, and explorations of this power series in further mathematics.

Axiomatic Integration

In contrast to many real analysis texts, we introduce the integral *axiomatically*, by proposing three axioms that anything worthy of being called 'an integral' ought to satisfy (this approach is based on that carried out out in Serge Lang's book, as well as in the lecture notes of Pete Clark). These axioms specify only that (1) the integral of a constant function is the area of a rectangle (2) if $f \leq g$ on an interval then their integrals inherit the same inequality, and (3) the subdivision rule: the integral from a to b is the same as the sum of the integrals over $[a, c]$ and $[c, b]$ for $c \in [a, b]$.

From these axioms alone, we prove that *if* $\int$ is any integral satisfying these axioms and f is a continuous integrable function, then the fundamental theorem of calculus holds. From here, we can prove many things (contingent on an integral existing) in a way that does not depend on the messy details of any particular construction. Indeed, we prove for integrable continuous functions

- The integral is linear, when restricted to continuous functions.
- You can integrate power series term by term
- U-substitution and integration by parts are valid integration techniques, when restricted to continuous functions.

The rationale for this approach is twofold. One, working with any particular integral (Riemann, Darboux, Lebesgue, etc) involves complicated arguments where the spirit can be lost in the details. But working axiomatically forces an argument to rely *only* on simple geometric premises. Second, the existence of so many different integrals (with different advantages/disadvantages) can be rather confusing to a beginning student, the axiomatic approach clearly separates out facts that are true for *any possible integral* (things you can prove from the axioms) from those that are *true of a particular integral* (things you can only prove using a particular construction).

Of course, there certainly remains an important place for showcasing at least one construction here: namely to prove that this entire theory isn't vacuous! But the importance is lessened, and students can treat the (sometime daunting) theory of the Darboux integral as more of a 'covering all our bases' than as a fundamental topic that needs to be deeply understood before moving on. To be sure, there is still some payout from the construction: we prove that the Darboux integral really is linear (on its entire space of integrable functions, not just the continuous ones), we generalize the integrability of power series term by term to a version of Dominated Convergence for the Darboux integral, and we use the ability to calculate integrals with Riemann sums to provide often-unseen infinite series converging to the natural logarithm and π.

Historically Important Problems

We make sure not just to cover the classic theorems required in a first analysis course, but also *use* them to solve problems of historical significance. In particular we rigorously discuss the following:

- The babylonian approximation to the square root of 2, as an introduction to monotone sequences, recursive sequences, and later continued fractions.
- Archimedes method of measuring the circle by exhaustion as a motivation for the nested interval theorem and for developing a theory of subsequences.
- Archimedes quadrature of the parabola as an introduction to the geometric series.
- The construction of the Koch snowflake fractal.
- Euler's formula and the relationship of the trigonometric functions to the complex exponential.
- As a final project: a solution to the Basel problem, proving $\sum_{n\geq 1}\frac{1}{n^2}=\frac{\pi^2}{6}$ using the material from the course.

A Unified Theory of Limit Switching

To determine when it is possible to permute a limit with an infinite sum, we prove Tannery's theorem (which we call Dominated Convergence, as it is a special case of the Lebesgue Dominated Convergence theorem applied to the measure space ℓ^1). As the course continues, any time we are faced with needing to exchange some limiting process with a sum, we start from this theorem and prove a relevant generalization. In all, we collect the following:

- Dominated Convergence for series
- Dominated Convergence for limits of series of functions
- Dominated Convergence for derivatives of series
- Dominated Convergence for integrals of series.

This gives a set of easily memorable conditions (because they are the same, or very similar in all cases) on when you can pull a limit inside of one of these operations.

To keep all limit-sum-exchanges in this family of similar theorems we depart from several usual topics in a first analysis course: we do not discuss uniform convergence, nor its implications for differentiation/integration of sequences of functions. Longer term, this book will be expanded into a one year analysis course, and the entire focus of the second half will be functional analysis; where a thorough treatment of these topics will be undertaken.

A Foundation for Functional Analysis

This book is aimed to provide undergraduates a foundation that will make it possible to study some aspects of functional analysis (convergence in function spaces, the linear algebra of differential equations, fourier series, and the calculus of variations). Some choices that were made specifically to help with this transition to more advanced analysis are

- Basing all limit switching off dominated convergence theorems. This makes the transition to the Lebesgue integral natural, and makes it clear why such an upgrade is needed (as our Dominated convergence theorem for the Darboux integral is slightly weaker than our other dominated convergence theorems; and the theorem we would have expected is exactly the one which is true for Lebesgue).
- Defining things axiomatically (the real numbers via the completeness axiom, the elementary functions by functional equations, and the integral by area properties) models many of the definitions of higher mathematics, where the focus is on *what an object is for* versus on any specific construction.
- We spend an extended amount of time discussing the exponential function, and how to use such a definition to extend the domain of a function from its original home (the real numbers) to more abstract spaces (complex numbers, matrices, linear maps, differential operators). The brief introduction to differential equations foreshadows some of the material coming in the sequel.

Part I.

Paradox

1. ⋆ Infinite Processes

1.1. The Diagonal of a Square

Around 3700 years ago, a babylonian student was assigned a homework problem, and their work (in clay) fortuitously survived until the modern day.

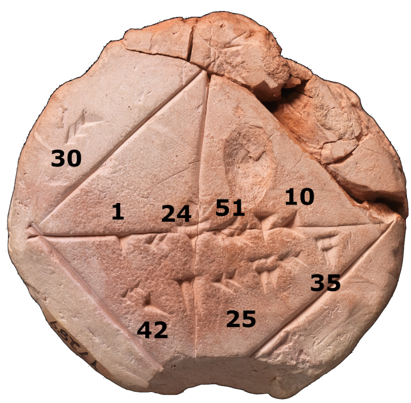


Figure 1.1.: Tablet YBC-7289

The problem involved measuring the length of the diagonal of a square of side length 1/2, which involves the square root of 2. The tablet records a babylonian approximation to $\sqrt{2}$ (Though it does so in base 60, where the 'decimal' expression is 1.(24)(51)(10))

$$\sqrt{2} \approx \frac{577}{408} \approx 1.414215686\cdots$$

Definition 1.1 (Base Systems for Numerals)**.** If $b > 1$ is a positive integer, *base-b* refers to expressing a number in terms of powers of b. In base 10 we write 432 to mean $4 \cdot 10^2 + 3 \cdot 10^1 + 2 \cdot 10^0$, whereas in base 5 the string of digits 432 would denote $4 \cdot 5^2 + 3 \cdot 5^1 + 2 \cdot 5^0$.

Numbers between 0 and 1 can also be expressed in a base system, using *negative powers of the base*. In base 10, 0.231 means $2 \cdot 10^{-1} + 3 \cdot 10^{-2} + 1 \cdot 10^{-3}$, whereas in base 5 the same string of digits would denote $2 \cdot 10^{-1} + 3 \cdot 5^{-2} + 1 \cdot 5^{-3}$.

The babylonians used base 60, meaning all numbers were written as a series in 60^n for n ranging over the integers. This tablet records the approximate square root of 2 as

$$1.(24)(51)(10)$$

Which, in base 60 denotes

$$\begin{aligned}\sqrt{2} &\approx 1\cdot 60^0 + 24\cdot 60^{-1} + 51\cdot 60^{-2} + 10\cdot 60^{-3}\\ &= 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3}\\ &= 1 + \frac{24}{60} + \frac{17}{1200} + \frac{1}{21600}\\ &= \frac{577}{408}\end{aligned}$$

Exercise 1.1. By inscribing a regular hexagon in a circle, the Babylonians approximated π to be $25/8$. Compute the base 60 'decimal' form of this number.

The tablet itself does not record how the babylonians came up with so accurate an approximation, but we have been able to reconstruct their reasoning in modern times

Example 1.1 (Babylonian Algorithm Computing $\sqrt{2}$)**.** Starting with a rectangle of area 2, call one of its sides x. If the rectangle is a square, then $x = \sqrt{2}$ exactly. And the closer our rectangle is to a square, the closer x is to $\sqrt{2}$. Thus, starting from this rectangle, we can build an *even better approximation* by making it more square. Precisely, the side lengths of this rectangle are x and $2/x$, and a rectangle with one side the *average* of these two numbers, will be closer to a square than this one.

Starting from a rectangle with side lengths 1 and 2, applying this procedure once improves our estimate from 1 to $3/2$, and then applying it again improves it to $577/408$. This Babylonian approximation is just the third element in an *infinite sequence* of approximations to $\sqrt{2}$

Exercise 1.2 (Babylonian Algorithm Computing $\sqrt{2}$)**.** Carry out this process, and show you get $577/408$ as the third approximation to $\sqrt{2}$. What's the next term in the sequence? How many decimal places is this accurate to in base 10? (Feel free to use a calculator of course!)

Exercise 1.3 (Computing Cube Roots)**.** Can you modify the babylonians procedure which found approximates of $\sqrt{2}$ to instead find rational approximates of $\sqrt[3]{2}$?

Here, instead of starting with a rectangle of sides x, y let's start with a three dimensional brick with a square base (sides x and x), height y, and area 2. Our goal is to find a "closer to cube" shaped brick than this one, and then to iterate. Propose a method of getting "closer to cube-shaped" and carry it out: what are the side lengths of the next shape in terms of x and y?

Start with a simple rectangular prism of volume 2 and iterate this procedure a couple times to get an approximate value of $\sqrt[3]{2}$. How close is your approximation?

It is clear from other Babylonian writings that they knew this was merely an approximation, but it took over a thousand years before we had more clarity on the nature of $\sqrt{2}$ itself.

1.1.1. Pythagoras

We often remember the Pythagoreans for the theorem bearing their name. But while they did prove this, the result (likely without proof) was known for millennia before them. The truly new, and shocking contribution to mathematics was the discovery that there must be numbers beyond the rationals, if we wish to do geometry.

Theorem 1.1 ($\sqrt{2}$ is irrational)**.** *There is no fraction p/q which squares to* 2.

To give a proof of this fact we need one elementary result of number theory, known as Euclid's Lemma (which says that if a prime p divides a product ab, then p must divide either a or b).

Proof. (Sketch) Assume p/q is in lowest terms, and squares to 2. Then $p^2/q^2 = 2$ so $p^2 = 2q^2$. Thus 2 divides p^2, so in fact 2 divides p (Euclid's lemma), meaning p is even.

Thus, we can write $p = 2k$ for some other integer k, which gives $(2k)^2 = 2q^2$, or $4k^2 = 2q^2$. Dividing out one factor of 2 yields $2k^2 = q^2 <$ so 2 divides q^2, and thus (Euclid's lemma, again) 2 divides q.

But now we've found that both p and q are divisible by 2, which means p/q is not in lowest terms after all, a contradiction! Thus there can not have been any fraction squaring to 2 in the first place. □

Exercise 1.4. Following analogous logic, prove that $\sqrt{3}$ is irrational. Generalize this to prove that $\sqrt{6}$ is irrational. But be careful! Make sure that your proof doesn't also apply to $\sqrt{9}$ (which of course, IS rational).

Knowing now that $\sqrt{2}$ is irrational, it is clear that the Babylonian procedure will never *exactly* return the correct answer, as if it starts with a rationally-sided rectangle, it'll always produce another with rational side lengths. But its a natural question to wonder just *how good* are the babylonian approximations?s

Definition 1.2 (The Babylonian Algorithm and Number Theory)**.** Because $\sqrt{2}$ is irrational, there is no pair of integers p, q with $p^2 = 2q^2$. Good rational approximations to $\sqrt{2}$ will *almost* satisfy this equation, and we will call an approximation *excellent* if it is only off by 1: that is p/q is an excellent approximation if

$$p^2 = 2q^2 + 1$$

Exercise 1.5 (The Babylonian Algorithm and Number Theory)**.** Prove that all approximations produced by the babylonian sequence starting from the rectangle with sides 1 and 2 are excellent, by induction.

To acomodate this discovery, the Greeks had to *add a new number to their number system* - in fact, after really absorbing the argument, they needed to add many. Things like $\sqrt{3}$, but also

$$\sqrt{1+\sqrt{3-\sqrt{2+\frac{\sqrt{3}+\sqrt{2}}{5}}}}$$

are called *constructible numbers*, as they were constructed by the greeks using a compass and straightedge, to extend the rational numbers.

1.2. Quadrature of the Parabola

The idea to compute some seemingly unreachable quantity by a succession of better and better approximations may have begun in babylon, but truly blossomed in the hands of Archimedes.

In his book *The Quadrature of the Parabola*, Archimedes relates the area of a parabolic segment to the area of the largest triangle that can be inscribed within.

Theorem 1.2. *The area of the segment bounded by a parabola and a chord is* $4/3^{rd}s$ *the area of the largest inscribed triangle.*

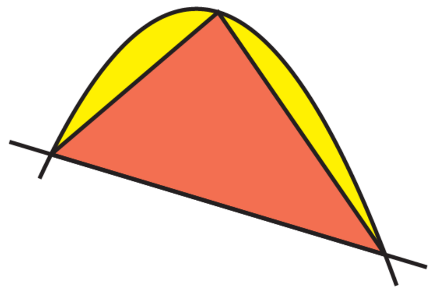


Figure 1.2.: A parabolic region and its largest inscribed triangle

After first describing how to find the largest inscribed triangle (using a calculation of the *tangent lines* to a parabola), Archimedes notes that this triangle divides the remaining region into two more parabolic regions. And, he could fill these with their largest triangles as well!

These two triangles then divide the remaining region of the parabola into *four new parabolic regions*, each of which has their own largest triangle, and so on.

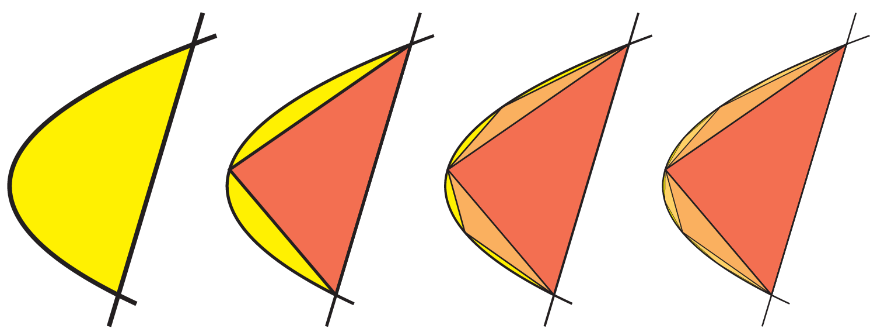


Figure 1.3.: Archimedes' infinite construction of the parabolic segment from triangles

Archimedes proves that in the limit, after doing this infinitely many times, the triangles *completely fill* the parabolic segment, with zero area left over. Thus, the only task remaining is to add up the area of these infinitely many triangles. And here, he discoveries an interesting pattern.

We will call the first triangle in the construction *stage 0* of the process. Then the two triangles we make next comprise *stage 1*, the ensuing four triangles *stage 2*, and the next eight *stage 3*.

Proposition 1.1 (Area of the n^{th} stage)**.** *The total area of the triangles in each stage is* $1/4$ *the total area of triangles in the previous stage.*

If A_n is the area in the n^{th} stage, Archimedes is saying that $A_{n+1} = \frac{1}{4}A_n$. Thus

$$A_0 = T \quad A_1 = \frac{1}{4}T \quad A_2 = \frac{1}{16}T \quad A_3 = \frac{1}{64}T \ldots$$

And the total area A is the infinite sum

$$\begin{aligned} A &= T + \frac{1}{4}T + \frac{1}{16}T + \frac{1}{64}T + \cdots \\ &= \left(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots\right)T \end{aligned}$$

Now Archimedes only has to sum this series. For us moderns this is no trouble: we recognize this immediately as a geometric series

But why is it called *geometric*? Well (this is not the only reason, but...) Archimedes was the first human to sum such a series, and he did so completely geometrically. Ignoring the leading 1, we can interpret all the fractions as proportions of the area of a square. The first term $1/4$ tells us to take a quarter of the square, the next term says to take a quarter of a quarter more, and so on. Repeating this process infinitely, Archimedes ends up with the following figure, where the highlighted squares on the diagonal represent the completed infinite sum.

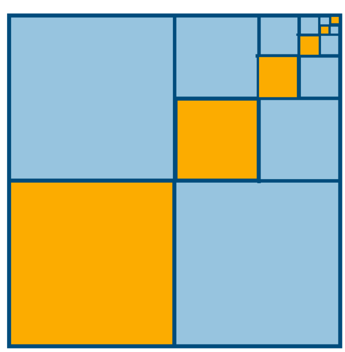


Figure 1.4.: The infinite process: $1/4 + 1/16 + 1/64 + \cdots$

He then notes that this is precisely one third the area of the bounding square, as two more identical copies of this sequence of squares fill it entirely (just slide our squares to the left, or down). Thus, this infinite sum is precisely $1/3$, and so the total area is 1 plus this, or $4/3$.

This tells us an important fact, beyond just the area of the parabola we sought! We were looking to compute the area of a *curved shape*, and the procedure we found could never give us the answer exactly, but only an infinite sequence of better approximations. Being acquainted with the work of Pythagoras and the Babylonians, this might have well led us to conjecture that the area of the parabola must be *irrationally* related to the area of the triangle. But Archimedes showed this is not the case; our infinite sum here evaluates to a *rational number*, 4/3!

Infinite sequences of rational numbers can sometimes produce a wholly new number, and sometimes just converge to another rational.*

How can we tell? This is one motivating reason to develop a rigorous study of such objects. But it gets even more important, if we try to generalize Archimedes' argument.

1.2.1. Icarus

Archimedes' quadrature of the parabola represents a monumental leap forward in human history. This is the first time in the mathematical literature where infinity is not treated as some distant ideal, but rather a real place that *can be reached*. And the argument itself is an absolute classic - involving the first occurrence of an infinite series in mathematics, and a wonderfully geometric summation method (hence the name *geometric series*, which survives until today). The elegance of Archimedes' calculation is almost dangerous - its easy to be blinded by its apparent simplicity, and – like Icarus – fly too close to the sun, falling from these heights of logic directly into contradiction.

Archimedes visualized his argument for the sum $\sum \frac{1}{4^n}$ as though it was occurring *inside* of a larger square, but there's another perspective we could take. Call the total sum S,

$$S = 1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots$$

and note that multiplying S by $1/4$ is the same as removing the first term, as it shifts all the terms down by one space:

$$\frac{1}{4}S = \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \cdots = S - 1$$

Thus, $\frac{1}{4}S = S - 1$, and we can *solve* this algebraic equation directly to find $S = 4/3$. The beauty of this argument is that unlike Archimedes' original, its not tied to the number $1/4$ at all! Imagine we took some number r, and we wanted to add up the infinite sum

$$1 + r + r^2 + r^3 + r^4 + r^5 + r^6 + r^7 + \cdots + r^n + \cdots$$

Call that sum S, and notice that we have the same property, multiplying the sum by r shifts every term down by one, so we get the same result as if we just removed the first term:

$$rS = S - 1$$

We can then solve this for S and get

$$S = \frac{1}{1-r}$$

This gives us what we expect when $r = 1/4$, and trying it for other fractions, like $r = 1/5$ or $r = 23/879$, we can confirm (with the help of a computer) that the infinite sum really does approach the value this formula gives!

Amazingly, it even works for negative numbers, after we think about what this means. If $r = \frac{-1}{2}$ then

$$1 + r + r^2 + r^3 + r^4 + r^5 + \cdots = 1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \cdots$$

Using our formula above we see that this is supposed to converge to

$$S = \frac{1}{1 - \left(\frac{-1}{2}\right)} = \frac{1}{1 + \frac{1}{2}} = \frac{1}{\frac{3}{2}} = \frac{2}{3}$$

And, using a computer to add up the first 100 terms we see

$$S \approx 0.66666666666666666666666666666692962030174033726847057618$$

This is pretty incredible, as our original geometric reasoning doesn't make sense for $r = -1/2$, but the algebra works just fine! We may also wish to investigate what happens when $r = 1$, which would give

$$S = 1 + 1 + 1 + 1 + 1 + 1 + \cdots$$

This is going off to infinity, and our formula gives $S = 1/(1-1) = 1/0$, which could make sense: we could even take this as an indication that we should define $1/0 = \infty$. But things get more interesting with $r = -1$. Here the sum is

$$S = 1 - 1 + 1 - 1 + 1 - 1 + 1 - 1 + 1 - 1 + 1 - \cdots$$

As we add this up term by term, we first have 1, then 0, then 1 then 0, over and over agan as we repeatedly add a 1, and then immediately cancel it out. This isn't getting close to any number at all! But our formula gives

$$S = \frac{1}{1-(-1)} = \frac{1}{2}$$

Now we have a real question - did we just **discover a new, deep fact of mathematics** - that we can sensibly assign values to series like this, that we weren't originally concerned with, or did we **discover a limitation of our theorem**? This is an interesting, and important question to come out of our playing around!

Thus far, we haven't seen any cases where our theorem has output any 'obviously' wrong answers, so we may be inclined to trust it. But this does not hold up to further scrutiny: what about when $r = 2$? Here the sum is

$$1 + 2 + 4 + 8 + 16 + 32 + \cdots$$

which is clearly going to infinity. But our formula disagrees, as it would have you belive the sum is $S = 1/(1-2) = -1$. This raises the more general problem: **when working with infinity, sometimes a formula you derive works, and sometimes it doesn't. How can you tell when to trust it?**

Exercise 1.6. Explain what goes wrong with the argument when $r = 2$...

1.3. The Circle Constant

The curved shape that everyone was *really* interested in was not the parabola, but the circle. Archimedes tackles this in his paper *The Measurement of the Circle*, where he again constructs a finite sequence of approximations built from triangles, and then reasons about the circle *out at infinity*. First, we need a definition:

Definition 1.3 (π and τ)**.** The area of the unit circle is denoted by the constant π. The circumference of the unit circle is denoted by the constant τ.

Archimedes came up with a sequence of overestimates, and underestimates for π by inscribing and circumscribing regular polygons.

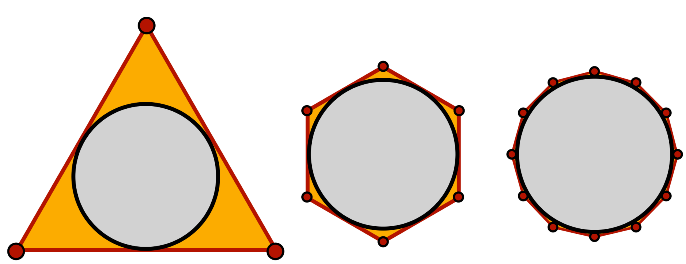


Figure 1.5.: Circumscribed polygons provide an overestimate of the area of the circle.

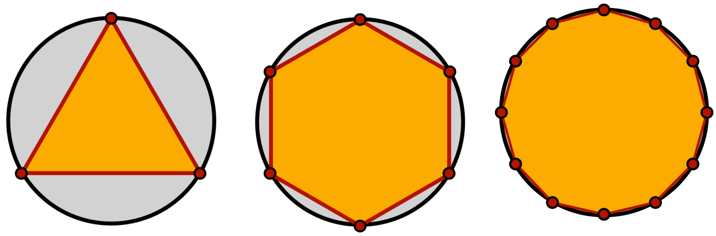


Figure 1.6.: Inscribed polygons provide an underestimate of the area of the circle.

Any polygon inside the unit circle gave an underestimate, and any polygon outside gave an overestimate. The more sides the polygon had, the better the approximations would be.

Calculating the area and perimeter of regular n-gons is (theoretically) straightforward, as they can be decomposed into $2n$ right triangles. Drawing a diagram, we find the relations below;

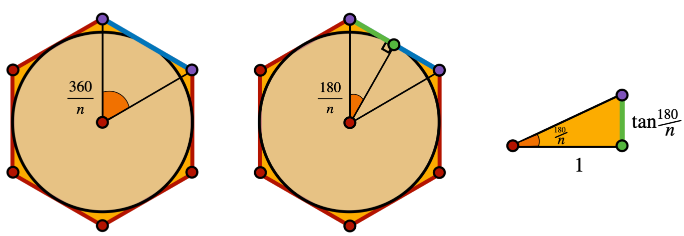


Figure 1.7.: Decomposing a circumscribed polygon into right triangles.

Proposition 1.2 (Area of a Circumscribed Polygon). *The area of a regular n-gon circumscribing the unit circle is given by*

$$\begin{aligned} C_n &= 2n \cdot \left(\frac{1}{2} \cdot 1 \cdot \tan\frac{180}{n}\right) \\ &= n \tan\frac{180}{n} \end{aligned}$$

Proposition 1.3 (Perimeter of a Circumscribed Polygon). *The perimeter of a regular n-gon circumscribing the unit circle is given by*

$$P_n = 2n \cdot \tan\frac{180}{n}$$

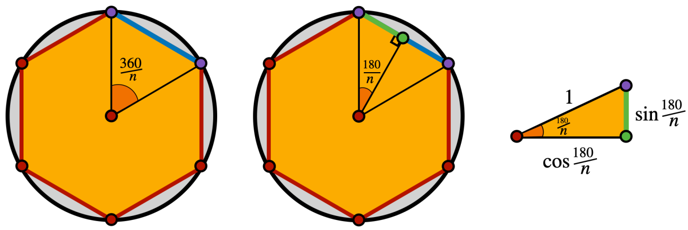


Figure 1.8.: Decomposing an inscribed polygon into right triangles.

Proposition 1.4 (Area of a Inscribed Polygon)**.** *The area of a regular n-gon inscribed in the unit circle is given by*

$$\begin{aligned} a_n &= 2n \cdot \left(\frac{1}{2} \cdot \cos\frac{180}{n} \cdot \sin\frac{180}{n}\right) \\ &= \frac{n}{2} \sin\frac{360}{n} \end{aligned}$$

Where we used the trigonometric identity $\sin(2x) = 2\sin x \cos x$ to simplify a_n above.

Proposition 1.5 (Perimeter of a Inscribed Polygon)**.** *The perimeter of a regular n-gon inscribed in the unit circle is given by*

$$p_n = 2n \cdot \sin\frac{180}{n}$$

Using these, Archimedes calculated away all the way to the 96-gon, which provided him with the estimates

$$\frac{223}{71} < \pi < \frac{22}{7}$$

This was the best estimate of π calculated during the classical period of the Greeks, but the same method was applied by Chinese mathematician Zu Chongzi in the 400s CE to much *much* larger polygons.
Working with the $24,576$-gon, he found

$$\frac{355}{113} < \pi < \frac{22}{7}$$

The lower bound here, $355/113$ is the best possible rational approximation of π with denominator less than four digits, and equals $3.14159292\cdots$, whereas $\pi = 3.14159265\cdots$. This was the most accurate approximate to π calculated anywhere in the world for over 800 years, and was only surpassed in the late 1300s by Indian mathematician Madhava, about whom we'll learn more soon.

Remark 1.1. The next best rational approximation is $\frac{52163}{16604}$, which is a significantly more complicated looking fraction!

1.3.1. Proving $\tau = 2\pi$

While impressive, Archimedes' main goal was not the *approximate* calculation above, but rather an *exact theorem*. He wanted to understand the true relationship between the area and perimeter of the circle, and wished to use these approximations as a guide to what is happening with the real circle, "out at infinity".

To understand this case, Archimedes argues that as n goes to infinity, the sequences of inscribed and circumscribed polygons approach the circle, and so *in the limit*, the sequences of areas must tend to the area of the circle (π) and the sequences of perimeters must tend to the perimeter of the circle (τ).

$$A_n \to \pi \qquad P_n \to \tau$$

But, now look carefully at the form of the expressions we derived for the circumscribing polygons in Proposition 1.2 and Proposition 1.3:

$$A_n = n \cdot \tan \frac{180}{n} \qquad P_n = 2n \cdot \tan \frac{180}{n}$$

Here, we do not need to worry about explicitly calculating A_n or P_n; all we need to notice is that the perimeter is *exactly* twice the area, $P_n = 2A_n$! This makes sense:

- Each polygon is built out of n triangles.
- The area of a triangle is half its base times its height
- The height of each triangle is 1 (the radius of the circle)
- Thus, the area the sum of half all the bases, or half the perimeter!

But since this exact relationship holds for every single value of n, Archimedes argued it must also be true in the limit, so the perimeter is twice the area:

Theorem 1.3 (Archimedes)**.**

$$\tau = 2\pi$$

1.3.2. Icarus, Reprise

Archimedes again leaves us with an argument so elegant and deceptively simple that its easy to under-appreciate its subtlety and immediately fall prey to contradiction. What if we attempt to repeat Archimedes argument, but with a different sequence of polygons approaching the circle?

Remark 1.2. To be fair to the master, Archimedes is *much, much* more careful in his paper than I was above, so part of the apparent simplicity is a consequence of my omission.

For example, what if we start with a square circumscribing the circle, and then at each stage produce a new polygon with the following rule:

- At each corner of the polygon, find the largest square that fits within the polygon, and remains outside the circle. Then remove this square.

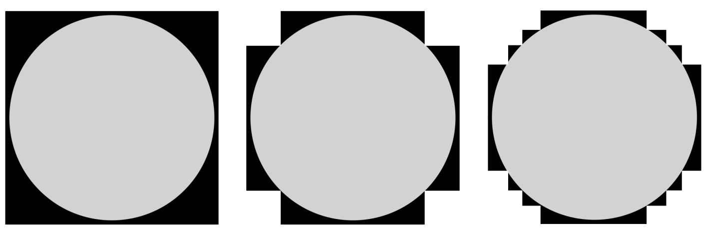


Figure 1.9.: Iteratively removing the largest square outside the circle at each vertex produces a sequence of right angled polygons which converges to the circle.

Exactly like in Archimedes' example this sequence of polygons approaches the circle as we repeat over and over. In fact, in the limit - this sequence *literally becomes the circle* (meaning that after infinitely many steps, there are no points of the resulting shape remaining outside the circle at all). Thus, just as for our original sequence of polygons, we expect that the areas and perimeters of these shapes approach the areas and perimeters of the circle itself. That is,

$$A_n \to \pi, \qquad P_n \to \tau$$

While the behavior of A_n takes a bit of work to understand, this sequence of polygons is constructed to make analyzing the perimeters particularly nice. Look what happens at each stage near a dent: two edges are turned inward to the circle, but do not change in length.

Figure 1.10.: Removing a square at a vertex does not change the perimeter of the polygon, as it replaces two segments with two other segments of the same length.

Since adding a dent does not change the length of the perimeter, each polygon in our sequence has *exactly the same perimeter* as the original! The original perimeter is easy to calculate, each side of the square is a diameter of the unit circle, so its total perimeter is 8. But since this both does not change *and* converges in the limit to the circles circumference, we have just derived the amazing fact that

$$\tau = 8$$

This is inconsistent with what we learn from Archimedes' argument which shows that $\pi < 22/7$ and $\tau = 2\pi$, so $\tau < 44/7 = 6.2857\ldots$. It appears that we have applied the same argument twice, and found a contradiction in comparing the results!

Exercise 1.7 (Convergence to the Diagonal)**.** We can run an argument analogous to the above which proves that $\sqrt{2} = 2$, by looking at a sequence of polygons that converge to a right triangle with legs of length 1. Let T_0 denote the unit square, and T_n

Prove that as n goes to infinity the area of the polygons T_n do converge to the area of the triangle (Hint: can you write down a formula for the *total error* between T_n and the triangle?) Also, prove that the length of the zig-zag diagonal side of the T_n has length 2 always, independent of n. Thus, the limit of the zigzag, which becomes the hypotenuse of the triangle, has length 2!

But the pythagorean theorem tells us that its length must be $\sqrt{1^2 + 1^2} = \sqrt{2}$, so in fact we have proven $\sqrt{2} = 2$, or $2 = 4$, a contradiction in mathematics.

Its quite difficult to pinpoint exactly what goes wrong here, and thus this presents a particularly strong argument for why we need analysis: without a rigorous understanding of infinite processes and limits, we can never be sure if our seemingly reasonable calculations give the right answers, or lies!

1.3.3. ...How did they do it?

With our modern access to calculator technology, the trigonometric formulas above essentially solves the problem: for example, plug in $n = 96$ to a calculator (set to degrees!) to replicate the work of Archimedes in one click.

But this poses a historical problem: of course the ancients did not have a calculator, so how did they compute such accurate approximations millennia ago? And there's also a potential logical problem lurking in the background: inside our calculator there is some algorithm computing the trigonometric functions, and perhaps that algorithm depends on already knowing something about the value of π. If so, using this calculator to give a from-first-principles estimate of π would be circular!

To compute their estimates, both Archimedes and Zu Chongzi landed on an idea similar to the Babylonians and their computation of $\sqrt{2}$: they found an *iterative procedure* that starts with one polygon, and doubles its number of sides. With such a procedure in hand, they could start with any polygon and rapidly scale it up to better and better estimates. Beginning with an hexagon, Archimedes only needed to double four times:

$$6 \to 12 \to 24 \to 48 \to 96$$

Exercise 1.8 (The Doublings of Zu Chongzi)**.** How many times did Zu Chongzi double the sides of a hexagon to reach the 24,576 gon?

Following Archimedes, we'll look at the doubling procedure for the perimeter of inscribed polygons: given p_n we seek a method to compute p_{2n}. By the formula in Proposition 1.4, it is enough to be able to compute $\sin(360/(2n))$ in terms of $\sin(360/n)$, that is, we need to be able to compute the sine of half the angle. The half-angle identities from trigonometry prove helpful here:

Definition 1.4 (Half Angle Identities)**.**

$$\cos\left(\frac{\theta}{2}\right) = \sqrt{\frac{1+\cos\theta}{2}} \qquad \sin\left(\frac{\theta}{2}\right) = \sqrt{\frac{1-\cos\theta}{2}}$$

$$\tan\left(\frac{\theta}{2}\right) = \sqrt{\frac{1-\cos\theta}{1+\cos\theta}} = \frac{\sin\theta}{1-\cos\theta} = \frac{1-\cos\theta}{\sin\theta}$$

Also making use of the pythagorean identity $\sin^2\theta + \cos^2\theta = 1$, we can compute as follows:

$$\begin{aligned}\sin\frac{\theta}{2} &= \sqrt{\frac{1-\cos\theta}{2}}\\ &= \sqrt{\frac{1-\sqrt{\cos^2\theta}}{2}}\\ &= \sqrt{\frac{1-\sqrt{1-\sin^2\theta}}{2}}\end{aligned}$$

Lets write $s_n = \sin(180/n)$ for brevity. Then, the above formula tells us how to compute s_{2n} if we know s_n:

$$s_{2n} = \sqrt{\frac{1-\sqrt{1-s_n^2}}{2}}$$

This sort of relationship is called a *recurrence relation*, or a *recursively defined sequence* as it tells us how to compute the next term in the sequence if we have the previous one. Notice there are no more trigonometric formulas in the recurrence - so if we can find the value s_n for *any polygon*, we can start with that, and iteratively double.

Example 1.2 (A Recurrence for p_n)**.** By Proposition 1.5, we see that $p_n = 2ns_n$. Thus $p_{2n} = 2(2n)s_{2n} = 4s_{2n}$, and using the recurrence for s_{2n} we see

$$\begin{aligned}p_{2n} &= 4ns_{2n}\\ &= 4n\sqrt{\frac{1-\sqrt{1-s_n^2}}{2}}\\ &= 2n\sqrt{2-2\sqrt{1-s_n^2}}\\ &= 2n\sqrt{2-\sqrt{4-4s_n^2}}\end{aligned}$$

But, since $s_n = p_n/(2n)$, substituting this in gives a relation between p_{2n} and p_n directly:

$$\begin{aligned}p_{2n} &= 2n\sqrt{2-\sqrt{4-4s_n^2}}\\ &= 2n\sqrt{2-\sqrt{4-\left(\frac{p_n}{n}\right)^2}}\end{aligned}$$

The incredible fact: even though we used trigonometry to derive this recurrence, we do not need to know how to evaluate any trigonometric functions to actually use it! All we need to be able to do is find the perimeter of *some* inscribed n-gon, and then we can repeatedly double over and over!

But how can we get started? A beautiful observation of Archimedes was that a regular hexagon inscribed in the circle has perimeter *exactly equal to 6*, as it can be decomposed into six equilateral triangles, whose side length is the circle's radius. And with that, we are off!

Example 1.3 (The Perimeter of an Inscribed 96-gon)**.** Since $p_6 = 6$, we begin with a doubling to find p_{12} :

$$\begin{aligned} p_{12} &= 12\sqrt{2-\sqrt{4-\left(\frac{6}{6}\right)^2}} \\ &= 12\sqrt{2-\sqrt{3}} \end{aligned}$$

Using this, we know $\frac{p_{12}}{12} = \sqrt{2-\sqrt{3}}$, and we can double again:

$$\begin{aligned} p_{24} &= 24\sqrt{2-\sqrt{4-\left(2-\sqrt{3}\right)}} \\ &= 24\sqrt{2-\sqrt{2+\sqrt{3}}} \end{aligned}$$

Now doubling to the 48 gon,

$$\begin{aligned} p_{48} &= 48\sqrt{2-\sqrt{4-(2-\sqrt{2+\sqrt{3}})}} \\ &= 48\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{3}}}} \end{aligned}$$

One more doubling brings us to the 96-gon,

$$p_{96} = 96\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{3}}}}}$$

Numerically approximating this gives 6.282063901781019276222, which is more recognizable to us if we compute the half perimeter:

$$\frac{p_{96}}{2} \approx 3.141031950890\ldots$$

Exercise 1.9. Find a recurrence relation for the area a_{2n} of the inscribed polygon, in terms of the area a_n of a polygon with half as many sides.

Exercise 1.10. Let $t_n = \tan(180/n)$. Show that t_n satisfies the recurrence relation

$$t_{2n} = \sqrt{1 + \frac{1}{t_n^2}} - \frac{1}{t_n}$$

Hint: you'll need some trig identities to write everything in terms of tangent! Use this to find a recurrence relation for P_n. Can you use this to find the circumference of an octagon circumscribing the unit circle?

After all of this are still left with a fundamental question: what sort of number *is* π? Archimedes' calculation out at infinity showed the area and circumference of a circle were related, but did not give us an exact value for either. These approximate calculations lead to some pretty scary looking numbers, but we know better than to trust that: we've already seen an infinite series of archimedes that summed to a nice rational number, and soon we will meet a nested sequence of square roots that collapses to a single root at infinity:

$$\sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}} = \frac{1 + \sqrt{5}}{2}$$

2. ★ Dubious Computations

We already saw, when visiting infinite processes from antiquity, that it is very easy to get confused and derive a contradiction when working with infinity. But on the other hand, infinite arguments turn out to be so useful that they are irresistible! Certain objects, like π, seem to be all but out of reach without invoking infinity somewhere, and while the lessons of the ancients implore us to be careful, more than once a good mathematician has thrown caution to the wind, in the hopes of gazing upon startling new truths.

2.1. Convergence, Concern and Contradiction

2.1.1. Madhava, Leibniz & $\pi/4$

Madhava was a Indian mathematician who discovered many infinite expressions for trigonometric functions in the 1300's, results which today are known as Taylor Series after Brook Taylor, who worked with them in 1715. In a particularly important example, Madhava found a formula to calculate the arc length along a circle, in terms of the tangent: or phrased more geometrically, the arc of a circle contained in a triangle with base of length 1.

> The first term is the product of the given sine and radius of the desired arc divided by the cosine of the arc. The succeeding terms are obtained by a process of iteration when the first term is repeatedly multiplied by the square of the sine and divided by the square of the cosine. All the terms are then divided by the odd numbers 1, 3, 5, ... The arc is obtained by adding and subtracting respectively the terms of odd rank and those of even rank.

As an equation, this gives

$$\theta = \frac{\sin\theta}{\cos\theta} - \frac{1}{3}\frac{\sin^2\theta}{\cos^2\theta}\left(\frac{\sin\theta}{\cos\theta}\right) + \frac{1}{5}\frac{\sin^2\theta}{\cos^2\theta}\left(\frac{\sin^2\theta}{\cos^2\theta}\frac{\sin\theta}{\cos\theta}\right) + \cdots$$

$$= \tan\theta - \frac{\tan^3\theta}{3} + \frac{\tan^5\theta}{5} - \frac{\tan^7\theta}{7} + \frac{\tan^9\theta}{9} - \cdots$$

If we take the arclength $\pi/4$ (the diagonal of a square), then both the base and height of our triangle are equal to 1, and this series becomes

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

This result was also derived by Leibniz (one of the founders of modern calcuous), using a method close to something you might see in Calculus II these days. It goes as follows: we know (say from the last chapter) the sum of the geometric series

$$\sum_{n\geq 0} r^n = \frac{1}{1-r}$$

Thus, substituting in $r = -x^2$ gives

$$\sum_{n\geq 0} (-1)^n x^{2n} = \frac{1}{1+x^2}$$

and the right hand side of this is the derivative of arctangent! So, anti-differentiating both sides of the equation yields

$$\begin{aligned}
\arctan x &= \int \sum_{n\geq 0} (-1)^n x^{2n}\, dx \\
&= \sum_{n\geq 0} \int (-1)^n x^{2n}\, dx \\
&= \sum_{n\geq 0} (-1)^n \frac{x^{2n+1}}{2n+1}
\end{aligned}$$

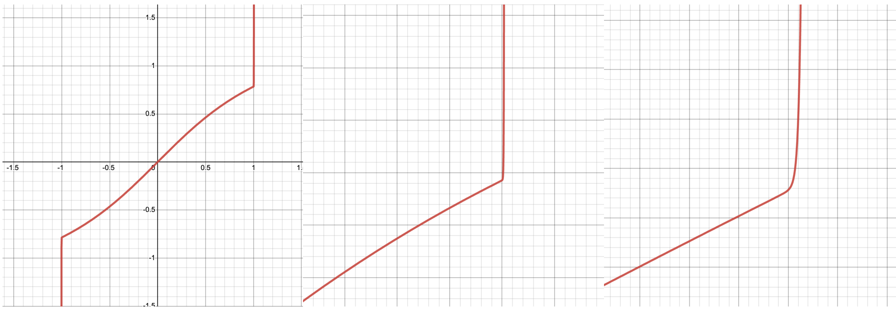
Finaly, we take this result and plug in $x = 1$: since $\arctan(1) = \pi/4$ this gives what we wanted:

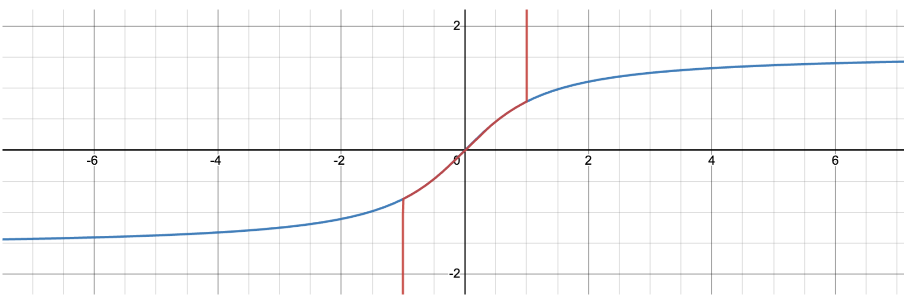
$$\frac{\pi}{4} = \sum_{n\geq 0} (-1)^n \frac{1}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

This argument is *completely full* of steps that should make us worried:

- **Why can we substitute a variable into an infinite expression and ensure it remains valid?**
- **Why is the derivative of arctan a rational function?**
- **Why can we integrate an infinite expression?**
- **Why can we switch the order of taking an infinte sum, and integration?**
- **How do we know which values of x the resulting equation is valid for?**

But beyond all of this, we should be even *more worried* if we try to plot the graphs of the partial sums of this supposed formula for the arctangent.





The infinite series we derived seems to match the arctangent *exactly* for a while, and then abruptly stop, and shoot off to infinity. Where does it stop? *Right at the point we are interested in, $\theta = \pi/4$, so $\tan(\theta) = 1$. So, even a study of which intervals a series converges in will not be enough here, we need a theory that is so precise, it can even tell us exactly what happens at the single point forming the boundary between order and chaos.

And perhaps, before thinking the eventual answer might simply say the series always converges at the endpoints, it turns out at the other endpoint $x = -1$, this series itself *diverges*! So whatever theory we build will have to account for such messy cases.

2.1.2. Dirichlet & $\log 2$

In 1827, Dirichlet was studying the sums of infinitely many terms, thinking about the *alternating harmonic series*

$$\sum_{n\geq 1} \frac{(-1)^n}{n+1}$$

Like the previous example, this series naturally emerges from manipulations in calculus: beginning once more with the geometric series $\sum_{n\geq 0} r^n = \frac{1}{1-r}$. We substitute

$r = -x$ to get a series for $1/(1+x)$ and then integrate term by term to produce a series for the logarithm:

$$\log(1+x) = \int \frac{1}{1+x} dx = \int \sum_{n \geq 0} (-1)^n x^n$$

$$= \sum_{n \geq 0} (-1)^n \frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

Finally, plugging in $x = 1$ yields the sum of interest. It turns out not to be difficult to prove that this series does indeed approach a finite value after the addition of infinitely many terms, and a quick check adding up the first thousand terms gives an approximate value of 0.6926474305598, which is very close to log(2) as expected..

$$\log(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} \cdots$$

What happens if we multiply both sides of this equation by 2?

$$2\log(2) = 2 - 1 + \frac{2}{3} - \frac{1}{2} + \frac{2}{5} - \frac{1}{3} + \frac{2}{7} - \frac{1}{4} + \frac{2}{9} - \frac{1}{5} \cdots$$

We can simplify this expression a bit, by re-ordering the terms to combine similar ones:

$$\begin{aligned} 2\log(2) &= (2-1) - \frac{1}{2} + \left(\frac{2}{3} - \frac{1}{3}\right) - \frac{1}{4} + \left(\frac{2}{5} - \frac{1}{5}\right) - \cdots \\ &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots \end{aligned}$$

After simplifying, we've returned to exactly the same series we started with! That is, we've shown $2\log(2) = \log(2)$, and dividing by $\log(2)$ (which is nonzero!) we see that $2 = 1$, a contradiction!

What does this tell us? Well, the only difference between the two equations is *the order in which we add the terms.* And, we get different results! This reveals perhaps the most shocking discovery of all, in our time spent doing dubious computations: **infinite addition is not always commutative, even though finite addition always is**.

Here's an even more dubious-looking example where we can prove that $0 = \log 2$. First, consider the infinite sum of zeroes:

$$0 = 0 + 0 + 0 + 0 + 0 + \cdots$$

Now, rewrite each of the zeroes as $x - x$ for some specially chosen xs:

$$0 = (1-1) + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{3}\right) + \left(\frac{1}{4} - \frac{1}{4}\right) + \cdots$$

Now, do some re-arranging to this:

$$\left(1 + \frac{1}{2} - 1\right) + \left(\frac{1}{3} + \frac{1}{4} - \frac{1}{2}\right) + \left(\frac{1}{5} + \frac{1}{6} - \frac{1}{3}\right) + \cdots$$

Make sure to convince yourselves that all the same terms appear here after the rearrangement!

Simplifying this a bit shows a pattern:

$$\left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \cdots$$

Which, after removing the parentheses, is the familiar series $\sum \frac{(-1)^n}{n}$. But this series equals $\log(2)$ (or, was it $2\log 2$?) So, if we are to believe that arithmetic with infinite sums is valid, we reach the contradiction

$$0 = \log 2$$

2.2. Infinite Expressions for $\sin(x)$

The sine function (along with the other trigonometric, exponential, and logarithmic functions) differs from the common functions of early mathematics (polynomials, rational functions and roots) in that it is defined not by a formula but *geometrically.*

Such a definition is difficult to work with if one actually wishes to *compute*: for example, Archimedes after much trouble managed to calculate the exact value of $\sin(\pi/96)$ using a recursive doubling procedure, but he would have failed to calculate $\sin(\pi/97)$ - 97 is not a multiple of a power of 2, so his procedure wouldn't apply! The search for a *general formula* that you could plug numbers into and compute their sine, was foundational to the arithmetization of geometry.

2.2.1. Infinite Sum of Madhava

Beyond the series for the arctangent, Madhava also found an infinite series for the sine function. The first thing that needs to be proven is that $\sin(x)$ satisfies the following integral equation: (Check this, using your calculus knowledge!)

$$\sin(\theta) = \theta - \int_0^\theta \int_0^t \sin(u)\, dudt$$

This equation mentions sine on *both sides*, which means we can use it as a *recurrence relation* to find better and better approximations of the sine function.

Definition 2.1 (Integral Recurrence For $\sin(x)$.)**.** We define a sequence of functions $s_n(x)$ recursively as follows:

$$s_{n+1}(\theta) = \theta - \int_0^\theta \int_0^t s_n(u)\, dudt$$

Given any starting function $s_0(x)$, applying the above produces a sequence $s_1(x)$, $s_2(x)$, $s_3(x), \ldots$ which we will use to approximate the sine function.

Example 2.1 (The Series for $\sin(x)$)**.** Like any recursive procedure, we need to start somewhere: so let's begin with the simplest possible (and quite incorrect) "approximation" that $s_0(\theta) = 0$. Integrating this twice still gives zero, so our first approximation is

$$s_1(\theta) = \theta - \int_0^\theta \int_0^t 0\ dudt = \theta - 0 = \theta$$

Now, plugging in $s_1 = \theta$ yields our second approximation:

$$\begin{aligned} s_2(\theta) &= \theta - \int_0^\theta \int_0^t udu dt \\ &= \theta - \int_0^\theta \left. \frac{u^2}{2} \right|_0^t dt \\ &= \theta - \int_0^\theta \frac{t^2}{2}\, dt \\ &= \theta - \left. \frac{t^3}{3 \cdot 2} \right|_0^\theta \\ &= \theta - \frac{\theta^3}{3!} \end{aligned}$$

Repeating gives the third,

$$\begin{aligned} s_3(\theta) &= \theta - \int_0^\theta \int_0^t \left(u - \frac{u^3}{3}\right) dudt \\ &= \theta - \int_0^\theta \left(\frac{t^2}{2} - \frac{t^4}{4\cdot 3!}\right) dt \\ &= \theta - \left(\frac{\theta^3}{3\cdot 2} - \frac{\theta^5}{5\cdot 4\cdot 3!}\right) \\ &= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} \end{aligned}$$

Carrying out this process *infinitely many times* yields a conjectured formula for the sine function as an infinite polynomial:

Proposition 2.1 (Madhava Infinite Sine Series)**.**

$$\sin(\theta) = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \frac{\theta^9}{9!} - \frac{\theta^{11}}{11!} + \cdots$$

Exercise 2.1. Find a similar recursive equation for the cosine function, and use it to derive the first four terms of its series expansion.

One big question about this procedure is **why in the world should this work**? We found a function that sin(x) satisfies, and then we plugged *something else into that function* and started iterating: what justification do we have that this should start to approach the sine? We can check after the fact that it (seems to have) worked, but this leaves us far from any understanding of what is actually going on.

2.2.2. Infinite Product of Euler

Another infinite expression for the sine function arose from thinking about the behavior of polynomials, and the relation of their formulas to their roots. As an example consider a quartic polynomial $p(x)$ with roots at $x = a, b, c, d$. Then we can recover p up to a constant multiple as a product of linear factors with roots at a, b, c, d. If the y−intercept is $p(0) = k$, we can give a fully explicit description

$$p(x) = k\left(1 - \frac{x}{a}\right)\left(1 - \frac{x}{b}\right)\left(1 - \frac{x}{c}\right)\left(1 - \frac{x}{d}\right)$$

In 17334, Euler attempted to apply this same reasoning in the infinite case to the trigonometric function sin(x). This has roots at every integer multiple of π, and so following the finite logic, should factor as a product of linear factors, one for each root. There's a slight technical problem in directly applying the above argument, namely that $\sin(x)$ has a root at $x = 0$, so $k = 0$. One work-around is to consider the function $\frac{\sin x}{x}$. This is not actually defined at $x = 0$, but one can prove $\lim_{x\to 0} \frac{\sin x}{x} = 1$, and attempt to use $k = 1$

GRAPH

Its roots agree with that of $\sin(x)$ except there is no longer one at $x = 0$. That is, the roots are $\dots, -3\pi, -2\pi, -\pi, \pi, 2\pi, 3\pi, \dots$, and the resulting factorization is

$$\frac{\sin x}{x} = \cdots \left(1 + \frac{x}{3\pi}\right)\left(1 + \frac{x}{2\pi}\right)\left(1 + \frac{x}{\pi}\right)\left(1 - \frac{x}{\pi}\right)\left(1 - \frac{x}{2\pi}\right)\left(1 - \frac{x}{3\pi}\right)\cdots$$

Euler noticed all the factors come in pairs, each of which represented a difference of squares.

$$\left(1 - \frac{x}{n\pi}\right)\left(1 + \frac{x}{2n\pi}\right) = \left(1 - \frac{x^2}{n^2\pi^2}\right)$$

Not worrying about the fact that infinite multiplication may not be commutative (a worry we came to appreciate with Dirichlet, but this was after Euler's time!), we may re-group this product pairing off terms like this, to yield

$$\frac{\sin x}{x} = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right)\cdots$$

Finally, we may multiply back through by x and get an infinite product expression for the sine function:

Proposition 2.2 (Euler)**.**

$$\sin x = x\left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\left(1 - \frac{x^2}{9\pi^2}\right)\cdots$$

This incredible identity is actually correct: there's only one problem - the argument itself is wrong!

Exercise 2.2. In his argument, Euler crucially uses that if we know

- all the zeroes of a function
- the value of that function is 1 at $x = 0$

then we can factor the function as an infinite polynomial in terms of its zeroes. This implies that a function is *completely determined* by its value at $x = 0$ and its zeroes (because after all, once you know that information you can just write down a formula like Euler did!) This is absolutely true for all finite polynomials, but it fails spectacularly in general.

Show that this is a serious flaw in Euler's reasoning by finding a *different* function that has all the same zeroes as $\sin(x)/x$ and is equal to 1 at zero (in the limit)!

Exercise 2.3 (The Wallis Product for π)**.** In 1656 John Wallis derived a remarkably beautiful formula for π (though his argumnet was not very rigorous).

$$\frac{\pi}{2} = \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\frac{8}{7}\frac{8}{9}\frac{10}{9}\frac{10}{11}\frac{12}{11}\frac{12}{13}\cdots$$

Using Euler's infinite product for $\sin(x)$ evaluated at $x = \pi/2$, give a derivation of Wallis' formula.

2.2.3. The Basel Problem

The Italian mathematician Pietro Mengoli proposed the following problem in 1650:

Definition 2.2 (The Basel Problem)**.** Find the exact value of the infinite sum

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots$$

By directly computing the first several terms of this sum one can get an estimate of the value, for instance adding up the first 1,000 terms we find $1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots \frac{1}{1{,}000^2} = 1.6439345\ldots$, and ading the first million terms gives

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{1,000^2} + \cdots + \frac{1}{1,000,000^2} = 1.64492406\ldots$$

so we might feel rather confident that the final answer is somewhat close to 1.64. But the interesting math problem isn't to approximate the answer, but rather to figure out something exact, and knowing the first few decimals here isn't of much help.

This problem was attempted by famous mathematicians across Europe over the next 80 years, but all failed. All until a relatively unknown 28 year old Swiss mathematician named Leonhard Euler published a solution in 1734, and immediately shot to fame. (In fact, this problem is named the Basel problem after Euler's hometown.)

Proposition 2.3 (Euler)**.**

$$\sum_{n\geq 1}\frac{1}{n^2} = \frac{\pi^2}{6}$$

Euler's solution begins with two different expressions for the function $\sin(x)/x$, which he gets from the sine's series expansion, and his own work on the infinite product:

$$\begin{aligned}\frac{\sin x}{x} &= 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \frac{x^8}{9!} - \frac{x^{10}}{11!} + \cdots \\ &= \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{2^2\pi^2}\right)\left(1 - \frac{x^2}{3^2\pi^2}\right)\cdots\end{aligned}$$

Because two polynomials are the same if and only if the coefficients of all their terms are equal, Euler attempts to generalize this to infinite expressions, and equate the coefficients for sin. The constant coefficient is easy - we can read it off as 1 from both the series and the product, but the quadratic term already holds a deep and surprising truth.

From the series, we can again simply read off the coefficient as $-1/3!$. But from the product, we need to think - after multiplying everything out, what sort of products will lead to a term with x^2? Since each factor is *already quadratic* this is more straightforward than it sounds at first - the only way to get a quadratic term is to take one of the quadratic terms already present in a factor, and multiply it by 1 from another factor! Thus, the quadratic terms are $-\frac{x^2}{2^2\pi^2} - \frac{x^2}{3^2\pi^2} - \frac{x^2}{4^2\pi^2} - \cdots$. Setting the two coefficients equal (and dividing out the negative from each side) yields

$$\frac{1}{3!} = \frac{1}{\pi^2} + \frac{1}{2^2\pi^2} + \frac{1}{3^2\pi^2} + \cdots$$

Which quickly leads to a solution to the original problem, after multiplying by π^2:

$$\frac{\pi^2}{3!} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$$

Euler had done it! There are of course many dubious steps taken along the way in this argument, but calculating the numerical value,

$$\frac{\pi^2}{3!} = 1.64493406685\ldots$$

We find it to be exactly the number the series is heading towards. This gave Euler the confidence to publish, and the rest is history.

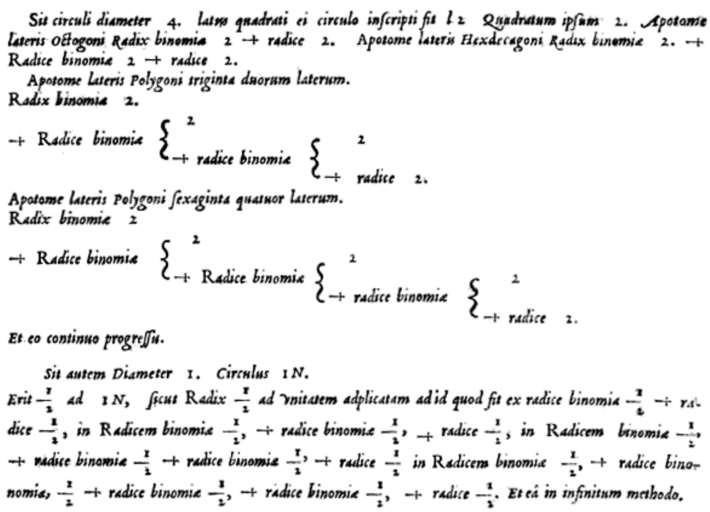
But we analysis students should be looking for potential troubles in this argument. What are some that you see?

2.2.4. Viète's Infinite Trigonometric Identity

Viete was a French mathematician in the mid 1500s, who wrote down for the first time in Europe, an exact expression for π in 1596.

Proposition 2.4 (Viète's formula for π)**.**

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2}\frac{\sqrt{2+\sqrt{2}}}{2}\frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}\frac{\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{2}}}}}{2}\cdots$$



Sit circuli diameter 4. latus quadrati ei circulo inscripti fit l 2 Quadratum ipsum 2. Apotome lateris Octogoni Radix binomia 2 + radice 2. Apotome lateris Hexdecagoni Radix binomia 2. + Radice binomia 2 + radice 2.

Apotome Lateris Polygoni triginta duorum laterum.
Radix binomia 2.
+ Radice binomia { 2 + radice binomia { 2 + radice 2.

Apotome lateris Polygoni sexaginta quatuor laterum.
Radix binomia 2
+ Radice binomia { 2 + Radice binomia { 2 + radice binomia { 2 + radice 2.

Et eo continuo progressu.

Sit autem Diameter 1. Circulus 1 N.
Erit $\frac{1}{2}$ ad 1 N, sicut Radix $\frac{1}{2}$ ad unitatem adplicatam ad id quod fit ex radice binomia $\frac{1}{2}$ + radice $\frac{1}{2}$, in Radicem binomia $\frac{1}{2}$, + radice binomia $\frac{1}{2}$, + radice $\frac{1}{2}$, in Radicem binomia $\frac{1}{2}$, + radice binomia $\frac{1}{2}$ + radice binomia $\frac{1}{2}$, + radice $\frac{1}{2}$ in Radicem binomia $\frac{1}{2}$, + radice binomia, $\frac{1}{2}$ + radice binomia $\frac{1}{2}$, + radice binomia $\frac{1}{2}$, + radice $\frac{1}{2}$. Et eâ in infinitum methodo.

Figure 2.1.: Viete's original publicaiton of this formula - it predates our modern notation for square roots!

How could one derive such an incredible looking expression? One approach uses trigonometric identities...an infinite number of times! Start with the familiar function sin(x). Then we may apply the double angle identity to rewrite this as

$$\sin(x) = 2\sin\left(\frac{x}{2}\right)\cos\left(\frac{x}{2}\right)$$

Now we may apply the double angle identity *once again* to the term $\sin(x/2)$ to get

$$\begin{aligned}\sin(x) &= 2\sin\left(\frac{x}{2}\right)\cos\left(\frac{x}{2}\right)\\ &= 4\sin\left(\frac{x}{4}\right)\cos\left(\frac{x}{4}\right)\cos\left(\frac{x}{2}\right)\end{aligned}$$

and again

$$\sin(x) = 8\sin\left(\frac{x}{8}\right)\cos\left(\frac{x}{8}\right)\cos\left(\frac{x}{4}\right)\cos\left(\frac{x}{2}\right)$$

and again

$$\sin(x) = 16\sin\left(\frac{x}{16}\right)\cos\left(\frac{x}{16}\right)\cos\left(\frac{x}{8}\right)\cos\left(\frac{x}{4}\right)\cos\left(\frac{x}{2}\right)$$

And so on....after the n^{th} stage of this process one can re-arrange the the above into the following (completely legitimate) identity:

$$\frac{\sin x}{2^n \sin\frac{x}{2^n}} = \cos\frac{x}{2}\cos\frac{x}{4}\cos\frac{x}{8}\cos\frac{x}{16}\cdots\cos\frac{x}{2^n}$$

Viete realized that as n gets really large, the function $2^n\sin(x/2^n)$ starts to look a lot like the function x...and making this replacement in the formula as we let n go to infinity yields

Proposition 2.5 (Viète's Trigonometric Identity)**.**

$$\frac{\sin x}{x} = \cos\frac{x}{2}\cos\frac{x}{4}\cos\frac{x}{8}\cos\frac{x}{16}\cdots$$

An incredible, infinite trigonometric identity! Of course, there's a huge question about its derivation: are we *absolutely sure* we are justified in making the denominator there equal to x? But carrying on without fear, we may attempt to plug in $x = \pi/2$ to both sides, yielding

$$\frac{2}{\pi} = \cos\frac{\pi}{4}\cos\frac{\pi}{8}\cos\frac{\pi}{16}\cos\frac{\pi}{32}\cdots$$

Now, we are left just to simplify the right hand side into something computable, using more trigonometric identities! We know $\cos\pi/4$ is $\frac{\sqrt{2}}{2}$, and we can evaluate the other terms iteratively using the half angle identity:

$$\cos\frac{\pi}{8} = \sqrt{\frac{1+\cos\frac{\pi}{4}}{2}} = \sqrt{\frac{1+\frac{\sqrt{2}}{2}}{2}} = \frac{\sqrt{2+\sqrt{2}}}{2}$$

$$\cos\frac{\pi}{16} = \sqrt{\frac{1+\cos\frac{\pi}{8}}{2}} = \sqrt{\frac{1+\frac{\sqrt{2+\sqrt{2}}}{2}}{2}} = \frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}$$

Substituting these all in gives the original product. And, while this derivation has a rather dubious step in it, the end result seems to be correct! Computing the first ten terms of this product on a computer yields $0.63662077105\ldots$, wheras $2/\pi = 0.636619772$. In fact, Viete used his own formula to compute an approximation of π to nine correct decimal digits. This leaves the obvious question, **Why does this argument work?**

2.3. The Infinitesimal Calculus

In trying to formalize many of the above arguments, mathematicians needed to put the calculus steps on a firm footing. And this comes with a whole collection of its own issues. Arguments trying to explain in clear terms what a derivative or integral was really supposed to be often led to nonsensical steps, that cast doubt on the entire procedure. Indeed, the history of calculus is itself so full of confusion that it alone is often taken as the motivation to develop a rigorous study of analysis. Because we have already seen so many other troubles that come from the infinite, we will content ourselves with just one example here: what is a derivative?

The derivative is meant to measure the slope of the tangent line to a function. In words, this is not hard to describe. But like the sine function, this does not provide a means of *computing*, and we are looking for a *formula*. Approximate formulas are not hard to create: if $f(x)$ is our function, and h is some small number the quantity

$$\frac{f(x+h)-f(x)}{h}$$

represents the slope of the secant line to f between x and h. For any finite size in h this is only an approximation, and so thinking of this like Archimedes did his polygons and the circle, we may decide to write down a *sequence of ever better approximations*:

$$D_n = \frac{f\left(x+\frac{1}{n}\right)-f(x)}{\frac{1}{n}}$$

and then define the derivative as the *infiniteth* term in this sequence. But this is just incoherent, taken at face value. If $1/n \to 0$ as $n \to \infty$ this would lead us to

$$\frac{f(x+0)-f(x)}{0} = \frac{0}{0}$$

So, something else must be going on. One way out of this would be if our sequence of approximates *did not actually converge to zero* - maybe there were infinitely small nonzero numbers out there waiting to be discovered. Such hypothetical numbers were called *infinitesimals*.

Definition 2.3 (Infinitesimal)**.** A positive number ϵ is infinitesimal if it is smaller than $1/n$ for all $n \in \mathbb{N}$.

This would resolve the problem as follows: if dx is some infinitesimal number, we could define the derivative as

$$D = \frac{f(x+dx) - f(x)}{dx}$$

But this leads to its own set of difficulties: its easy to see that if ϵ is an infinitesimal, then so is 2ϵ, or $k\epsilon$ for any rational number k.

Exercise 2.4. Prove this: if ϵ is infinitesimal and $k \in \mathbb{Q}$ show $k\epsilon$ is infinitesimal$.

So we can't just say define the derivative by saying "choose some infinitesimal dx" - there are many such infinitesimals and we should be worried about which one we pick! What actually happens if we try this calculation in practice, showcases this.

Let's attempt to differentiate x^2, using some infinitesimal dx. We get

$$\begin{aligned}(x^2)' = \frac{(x+dx)^2 - x^2}{dx} &= \frac{x^2 + 2xdx + dx^2 - x^2}{dx} \\ &= \frac{2xdx + dx^2}{dx} = 2x + dx\end{aligned}$$

Here we see the derivative is not what we expected, but rather is $2x$ plus an infinitesimal! How do we get rid of this? One approach (used very often in the foundational works of calculus) is simply to discard any infinitesimal that remains at the end of a computation. So here, because $2x$ is finite in size and dx is infinitesimal, we would just discard the dx and get $(x^2)' = 2x$ as desired.

But this is not very sensible: when exactly are we allowed to do this? If we can discard an infinitesimal whenever its added to a finite number, shouldn't we already have done so with the $(x + dx)$ that showed up in the numerator? This would have led to

$$\frac{(x+dx)^2 - x^2}{dx} = \frac{x^2 - x^2}{dx} = \frac{0}{dx} = 0$$

So, the when we throw away the infinitesimal matters deeply to the answer we get! This does not seem right. How can we fix this? One approach that was suggested was to say that we *cannot* throw away infinitesimals, but that the square of an infinitesimal is *so small that it is precisely zero*: that way, we keep every infinitesimal but discard any higher powers. A number satisfying this property was called *nilpotent* as *nil* was another word for zero, and *potency* was an old term for powers (x^2 would be the *second potency of x).

Definition 2.4. A number ϵ is nilpotent if $\epsilon \neq 0$ but $\epsilon^2 = 0$.

If our infinitesimals were nilpotent, that would solve the problem we ran into above. Now, the calculation for the derivative of x^2 would proceed as

$$\frac{(x+dx)^2 - x^2}{dx} = \frac{x^2 + 2xdx + dx^2 - x^2}{=} \frac{2xdx + 0}{dx} = 2x$$

But, in trying to justify *just this one calculation* we've had to invent two new types of numbers that had never occurred previously in math: we need positive numbers smaller than any rational, and we also need them (or at least some of those numbers) to square to precisely zero. **Do such numbers exist?**

Part II.

Numbers

3. Operations

Highlights of this Chapter: We begin axiomatizing the real numbers by axiomatizing their operations of addition and multiplication, leading to the field axioms. We give careful definitions of various notations from arithmetic, and do several example calculations (including a proof that $2+2=4$ and $(a+b)^2=a^2+2ab+b^2$) to exhibit that all arithmetical facts are consequences of the field axioms.

The first step to axiomatizing numbers is to give a precise description of addition, subtraction, multiplication and division. These operations naturally group into two pairs (addition/subtraction as well as multiplication/division) of operation/inverse, so first we will formalize the notion of an *invertible operation*. Furthermore, the two operations are related to one another by the *distributive law*. Two invertible operations bonded together by the distributive law form a mathematical structure we call a *field*, which is what we axiomatize in this chapter.

Definition 3.1 (Binary Operation)**.** A binary operation $\star$ on a set S is a rule that takes any two elements of S and combines them to make a new element of S.

Formally, this is a function $\star : S \times S \to S$. Whereas we often write functions $f : S \times S \to S$ as $f(a,b)$ for a binary operation we traditionally write the function name in the *middle* so $a \star b$ instead of $\star(a,b)$.

Example 3.1. Addition is a binary operation on the natural numbers, integers, rationals, and real numbers. Subtraction is a binary operation on the integers, but not on the natural numbers, as $4-7=-3$ gives an element not in the original set.

Definition 3.2 (Commutativity and Associativity)**.** An operation $\star$ is commutative if the order the elements are combined does not affect the outcome: for all elements $a, b \in S$

$$a \star b = b \star a$$

An operation is *associative* if combinations of 3 or more terms can be re-grouped at will (not changing the order), without affecting the outcome: for all $a, b, c \in S$

$$(a \star b) \star c = a \star (b \star c)$$

Example 3.2 (Commutativity and Associativity)**.** The operation of addition is commutative and associative, but the operation of subtraction is neither. The operation of matrix multiplication is associative, but is not commutative in general.

An operation which is commutative but not associative is given by the children's game *rock paper scissors*: if $S = \{r, p, s\}$ we may define the operation $\star$ to select the *winning element* of any pair. Thus, because paper beats rock, we have $r \star p = p$. Explain why this is commutative, and find an example proving it is not associative.

Definition 3.3 (Identity Element)**.** Let S be a set with binary operation $\star$. Then an element $e \in S$ is an *identity* for the operation if it does not change any elements under combination. Formally, for all $s \in S$

$$e \star s = s \star e = s$$

Given a binary operation $\star$ on a set S with identity $e \in S$, an element $x \in S$ is *invertible* if it can be combined with something to produce the identity. That is, if there exists a $y \in S$ with

$$x \star y = y \star x = e$$

This element y is called the *inverse* of x. An operation $\star$ is called *invertible* if every element of S has an inverse.

Example 3.3 (Identity Element)**.** Zero is the identity of the operation of addition, 1 is the identity of multiplication (in any familiar number system you'd like to take as an example). The identity matrix $\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$ is the identity of 2×2 matrix multiplication.

Not all operations have an identity. Can you see why there is no identity operation for exponentiation x^y on the positive integers?

Example 3.4 (Inverse)**.** The operation of addition is invertible, and its inverse is *subtraction.* The operation of multiplication is not invertible, because the number 0 does not have an inverse (you can't divide by zero! We'll prove this soon)

Definition 3.4 (Group)**.** A *group* is a set G with an associative, invertible binary operation e.

The concept of a group is ubiquitous in mathematics, as it formalizes the idea of a *nice* binary operation. But for analysis, we need more than this: numbers come with two binary operations (addition and multiplication) and so we need to describe how they *interact.*

Definition 3.5 (The Distributive Law)**.** Let S be a set with two commutative binary operations $+, \cdot$. Then $\cdot$ distributes over $+$ if for all $a, b, c \in S$ we have

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$$

Definition 3.6 (Field). A Field is a set $\mathbb{F}$ with two binary operations denoted $+$ (addition) and $\cdot$ (multiplication) satisfying the following axioms.

- (Commutativity) If $a, b \in \mathbb{F}$ then $a + b = b + a$ and $a \cdot b = b \cdot a$.
- (Associativity) If $a, b, c \in \mathbb{F}$ then $(a + b) + c = a + (b + c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
- (Identities) There are special elements denoted $0, 1 \in \mathbb{F}$ where for all $a \in \mathbb{F}$, $a + 0 = a$ and $1 \cdot a = a$.
- (Inverses) For every $a \in \mathbb{F}$ there is an element $-a$ such that $a + (-a) = 0$. If $a \neq 0$, then there is also an element a^{-1} such that $a \cdot a^{-1} = 1$.
- (Distributivity) If $a, b, c \in \mathbb{F}$ then $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$

3.0.1. Shorthand Notations

There is a lot of notation that we use to simplify writing out basic arithmetic expressions in fields. I will attempt a list of these shorthands here. First, some relating to the operations themselves.

- We often write the operation of multiplication simply as juxtaposition, without any intervening symbol. That is, we write ab instead of $a \cdot b$.
- We make the convention that multiplication *precedes* addition, so we evaluate the expression $ab + c$ as $(a \cdot b) + c$ not $a \cdot (b + c)$. This allows us to drop a lot of parentheses, making things easier to read.

Next, some notation for certain elements: - We define the symbol 2 to be the result of the operation $1 + 1$. - We define the symbol 3 to be the result of the operation $2 + 1$ - We define the symbol 4 to be the result of the operation $3 + 1$ - We define the symbol 5 to be the result of the operation $4 + 1$ - Etc

This lets us easily write down repeated addition, as we will see below $3x = x + x + x$ for any x. It's also useful to have some notation for repeated multiplication, which we denote with *powers*

- The notation x^2 will mean the product $x \cdot x$.
- The notation x^3 will mean the product $x \cdot x \cdot x$.
- Etc...
- The notation x^0 will denote the multiplicative identity 1.

We also introduce another notation for multiplicative inverse, to make formulas more readable:

- If $c \neq 0$ we write $\frac{a}{c}$ for ac^{-1}.

3.0.2. Computations in Fields

Example 3.5.

$$2x = x + x$$

To prove this for an arbitrary $x \in \mathbb{F}$, we recall the definition $2 = 1 + 1$ and use the distributive property:

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

Finally the last equality follows as 1 is the multiplicative identity, so

Example 3.6.

$$0x = 0$$

To prove this for an arbitrary $x \in \mathbb{F}$, recall that 0 is the additive identity so for any field element c, we have $0 + c = c$. Thus, when $c = 0$ we have $0 + 0 = 0$. We can use this together with the distributive property to get

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

Now, we can take the additive inverse of $0x$ and add it to both sides:

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

This gives the additive identity 0 by definition on the left side, and cancels one of the factors of $0x$ on the right, yielding

$$0 = 0x + 0$$

Finally we use again that 0 is the additive identity to see $0x + 0 = 0x$, which gives us what we want:

$$0x = 0$$

Example 3.7 (The Zero-Product Property)**.** Let a, b be elements of a field and assume that $ab = 0$. Then either $a = 0$ or $b = 0$.

We assume that both a and b are nonzero, and see that we reach a contradiction. Since they're nonzero, they have multiplicative inverses a^{-1} and b^{-1}, so we may multiply both sides of $ab = 0$ by these to get

$$b^{-1}a^{-1}ab = b^{-1}a^{-1}0$$

On the left this simplifies to $b^{-1}1b = b^{-1}b = 1$ by definition, and on the right this becomes $0(b^{-1}a^{-1}) = 0$ by the previous example. Thus, we've proven $0 = 1$! So this could not have been the case, and either a or b must have not been invertible to start with - they must have been zero.

Example 3.8.

$$-x = (-1)x$$

The definition of the symbol $-x$ is the element of $\mathbb{F}$ which, when added to x, gives 0. Thus, to prove that $-x = -1x$ we want to prove that if you add $(-1)x$ to x, you get 0. Since 1 is the additive identity, we know $1x = x$ so we may write

$$x + (-1x) = 1x + (-1x)$$

Using the fact that multiplication is commutative and the distributive law, we may factor out the x:

$$1x + (-1)x = (1 + (-1))x$$

Now, by definition $1 + (-1)$ is the additive identity 0, so this is just equal to $0x$. But by Example 3.6$ we know $0x = 0$! Thus

$$x + (-1x) = 0$$

And so $-1x$ is the additive inverse of x as claimed. Thus we may write $-x = (-1)x$

Example 3.9.

$$(-1)(-1) = 1$$

This is an immediate corollary of the above: we know that $(-1)x$ is the additive inverse of x, and so $(-1)(-1)$ is the additive inverse of -1. But this is just 1 itself, by definition!

Exercise 3.1. For any $x \in \mathbb{F}$ we have

$$-(-x) = x$$

Exercise 3.2. Prove, using only the field axioms and the definitions of the symbols $0, 1, 2, 3, 4$ that the following is true:

$$2 + 2 = 4$$

Example 3.10.

$$2 \cdot 2 = 4$$

This is a corollary of **?@exr-2-plus-2** above, as using the distributive law we see

$$2 \cdot 2 = 2 \cdot (1 + 1) = 2 \cdot 1 + 2 \cdot 1 = 2 + 2$$

And we already know $2 + 2 = 4$

All of the standard arithmetic “rules” learned in grade school are consequences of the field axioms, and so you are welcome to use all of them in this course, without comment. However, to feel justified in doing this, its good to prove a couple of them yourself, to convince yourself that you could in fact trace and any all such manipulations back to the rigorous axioms we laid down.

Exercise 3.3 (The difference of squares)**.** Prove that for any $a, b \in \mathbb{F}$

$$(a + b)(a - b) = a^2 - b^2$$

In your proof you may use the field axioms, the notational shorthands, and any of the example properties proved above in the notes. Anything else you need, you should prove from this.

Exercise 3.4. Prove, using the field axioms and our notational shorthands, for any a, b and $c \neq 0$

$$\frac{a + b}{c} = \frac{a}{c} + \frac{b}{c}$$

Exercise 3.5. Prove that fraction addition works by finding a common denominator: for any a, c and nonzero b, d

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bd}{bd}$$

In your proof you may use the field axioms, the notational shorthands, and any of the example properties proved above in the notes. Anything else you need, you should prove from this.

4. Order

Highlights of this Chapter: We define the notion of inequality in terms of the notion of *positivity* which we axiomatize, leading to the definition of an *ordered field*. We prove this new axiom is required as not all fields can be ordered (by looking at the complex numbers), and then we investigate several important properties and definitions related to order that are essential to real analysis:

- We define absolute value, and give several characterizations
- We prove the triangle inequality
- We define square roots, and n^{th} roots

Field theory (the study of mathematical objects satisfying the field axioms) is a broad subject in mathematics, underpinning large swaths of abstract algebras as well as analysis. The first step in deciding how to order numbers is to axiomatize what it means for a number to be *positive*.

Definition 4.1 (Positive Elements)**.** A subset $P \subset \mathbb{F}$ is called the *positive elements* if

- (Trichotomy) For every $a \in \mathbb{F}$ exactly one of the following is true: $a = 0$, $a \in P$ $-a \in P$.
- (Closure) If $a, b \in P$ then $a + b \in P$ and $ab \in P$.

Definition 4.2 (Ordered Field)**.** A n *ordered field* is a field $\mathbb{F}$ together with a subset $P \subset \mathbb{F}$ of positive elements.

Definition 4.3 (Inequalities)**.** If F is an ordered field and $a, b \in \mathbb{F}$ then we write $a < b$ as a shorthand for the statement that $b - a \in P$, and we write $a \leq b$ if either $a < b$ or $a = 0$.

Analogously, we write $a > b$ if $a - b \in P$ and $a \geq b$ if either $a > b$ or $a = b$.

4.0.1. Properties of Ordered Fields

Being an ordered field requires more structure *above and beyond just being a field.* Not every field can be ordered! In this short section we explore some properties of ordered fields in general.

Proposition 4.1 (1 is a Positive Number). *If (F, P) is any ordered field, then $1 \in P$.*

Proof. Since $1 \neq 0$ we know that either $1 \in P$ or $-1 \in P$. So, to show $1 \in P$ its enough to see $-1 \in P$ leads to contradiction.

If $-1 \in P$ then by closure, $(-1)(-1) = 1 \in P$: so now we have *both* 1 and -1 in P, contradicting trichotomy. □

Exercise 4.1. Let F be an ordered field and $x \neq 0$ an element. Then $x^2 > 0$.

Proposition 4.2 ($\mathbb{C}$ is not ordered). *The complex numbers cannot be made into an ordered field: there is no subset $P \subset \mathbb{C}$ such that P is a positive cone for $\mathbb{C}$.*

Proof. The complex numbers contain an element i with the property that $i^2 = -1$. If they were ordered, since $i \neq 0$ we know either $i \in P$ or $-i \in P$, but both of these lead to contradiction.

If $i \in P$ then $i^2 = -1 \in P$ contradicting the previous theorem that $1 \in P$ always. And $-i \in P$ leads to the same problem: $(-i)^2 = (-i)(-i) = --(i^2) = i^2 = -1$, so $-1 \in P$ again. □

This may seem like a strange example to start with, as the course is about *real* analysis. But its actually quite important: every time we introduce a new concept to the foundations of our theory we should ask ourselves, *is this an axiom, or a theorem*? We don't want to add as axioms things that we can already prove from the existing axioms, as that is redundant! So before adding a new axiom, we should convince ourself that its *necessary*: that it is impossible to prove the existence of this new structure given the previous. And that's what this example does. By exhibiting something that satisfies all the field axioms but *cannot be ordered*, we see that it is logically impossible to prove the existence of an order from the field axioms alone, and thus we must take **?@def-order** as a new axiom.

Exercise 4.2. The rational numbers admit a unique ordering: there is only one set $P \subset \mathbb{Q}$ that satisfies the axioms of a positive cone for $\mathbb{Q}$. Can you prove this? *Hint: once we know that 1 is positive, show this determines that a fraction p/q is positive if and only if p and q have the same sign.*

Some fields admit more than one possible ordering, and so can be turned into an ordered field in more than one way! An example of this is the field $\mathbb{Q}(\tau)$ with $\tau^2 = 2$ admits two possible orderings, one where $\tau \in P$ and the other with $-\tau \in P$.

4.1. Definitions Requiring an Order

Definition 4.4 (Intervals)**.** Let F be an ordered field. We write $[a, b]$ for the set $\{x \mid a \leq x \leq b\}$, and call this set a *closed interval* in F. Similarly we write (a, b) for the set $\{x \mid a < x < b\}$, which we call an *open interval.* Mixed intervals are also possible, such as $[a, b) = \{x \mid a \leq x < b\}$.

An *unbounded interval*, or a *ray* is a set of the form $\{x \mid x > a\}$ or $\{x \mid x \geq a\}$. We call the first an *open ray* and the latter a *closed ray*, and often denote them (a, ∞) or $[a\infty)$ as a shorthand. Similarly with $(-\infty, a)$ and $(-\infty, a]$.

Definition 4.5 (Absolute Value)**.** Let $\mathbb{F}$ be an ordered field. Then the *absolute value* is a function $|\cdot| : \mathbb{F} \to \mathbb{F}$ defined by

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

Definition 4.6 (The $\sqrt{\cdot}$ symbol)**.** Let $\mathbb{F}$ be an ordered field, and $x \in \mathbb{F}$. If there exists a $y \geq 0$ in $\mathbb{F}$ such that $y^2 = x$, we call y the *square root of* x and denote $\sqrt{x}$.

We can generalize this by defining $\sqrt[q]{x}$ to be the number y with $y^p = x$, and similarly to the above, prove that if $a < b$ are positive field elements, then $\sqrt[q]{a} < \sqrt[q]{b}$.

Exercise 4.3 (No Square Roots of Negatives)**.** Let F be any ordered field, and let $x < 0$. Prove that x does not have a square root in F.

Definition 4.7 (Rational Powers)**.** Let $a \in \mathbb{F}$ and $p/q \in \mathbb{Q}$.
Then if the element $a^p \in \mathbb{F}$ has a q^{th} root, we define the fractional power $a^{p/q}$ as

$$a^{p/q} = \sqrt[q]{a^p}$$

4.2. Working with Inequalities

All the standard properties of inequalities from arithmetic hold in an ordered field, and so you will be able to use them without comment throughout the course. However, its good to derive a few of these for yourselves from the definitions at first, to see how it goes.

Example 4.1 (Inequality is antisymmetric)**.** By trichotomy we see that for every $x \neq y$ we have either $x < y$ or $y < x$ (as, $x - y \neq 0$ implies either $x - y \in P$, so $x - y > 0$ and $x > y$ or the reverse).

Proposition 4.3 (Inequality is transitive)**.** *Let F be an ordered field and a, b, c in F. If $a < b$ and $b < c$, then $a < c$.*

Proof. If $a < b$ then $b - a \in P$. Similarly, $b < c$ implies $c - b \in P$. Closure then tells us their sum, $(c - b) + (b - a) \in P$, and so after simplifying,

$$c + (-b + b) - a = c + 0 - a = c - a \in P$$

This is the definition of $c > a$. □

Exercise 4.4 (Adding to an Inequality)**.** Let F be an ordered field and $a, b, c \in F$ with $a < b$. Then

$$a + c < b + c$$

Proposition 4.4 (Multiplying an Inequality)**.** *Let F be an ordered field and $a, b, c \in F$ with $a < b$. Then if $c > 0$, it follows that $ca < cb$, and if $c < 0$ we have instead $ca > cb$.*

Proof. First treat the case $c > 0$. Since $a < b$ we know $b-a \in P$, and $c \in P$ so $c(b-a) \in P$ by the closure axiom. Distributing gives $cb - ca \in P$ which is the definition of $cb > ca$.

Now, if $c < 0$, we know $c \notin P$, so $-c \in P$. Closure then gives $(-c)(b - a) \in P$, and simplifying yields $-cb + ca \in P$ or $ca - cb \in P$, the definition of $ca > cb$. □

4.2.1. Powers and Roots

Some basic inequalities for powers and roots that will prove useful: like other basic properties of inequalities, you do not need to prove or cite these when you use them in this course, but it is good to have a reference seeing *why* they are true from our axioms.

Example 4.2 ($x \mapsto x^2$ is increasing)**.** If F is an ordered field and $a, b \in F$ are elements with $0 < a < b$ then $a^2 < b^2$.

To prove this, we use both Proposition 4.3 and Proposition 4.4. Since $a < b$ and $a > 0$ we see $a^2 < ab$. But since $a < b$ and $b > 0$, we see $ab < b^2$. Putting these together yields $a^2 < ab < b^2$, so $a^2 < b^2$.

Its necessary to assume a, b are positive in the theorem above: for example $-3 < 1$ but $(-3)^2 = 9$ is not less than $1^2 = 1$. In fact this proof works in reverse as well (check this!) to provide the following useful fact:

Proposition 4.5. *If $a, b \in \mathbb{F}$ are positive elements of an ordered field, then*

$$a < b \iff a^2 < b^2$$

This generalizes to arbitrary powers:

Exercise 4.5 ($x \mapsto x^n$ is increasing)**.** Prove that if F is an ordered field containing positive elements a, b, then for all $n \in \mathbb{N}$, $a < b$ if and only if $a^n < b^n$.

In fact, when n is odd, you may wish to prove that you can remove the assumption that $a, b > 0$.

Here's a quick fact about inequalities that will prove useful to us later on in the course:

Exercise 4.6 (Bernoulli's inequality)**.** Let $\mathbb{F}$ be an ordered field and $x > 0$ be a positive element. Prove by induction that for all natural numbers $\mathbb{N}$

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

Exercise 4.7 ($\sqrt{\cdot}$ is increasing)**.** Prove that if $0 < x < y$ in an ordered field F, and F contains the square roots $\sqrt{x}, \sqrt{y}$, then $\sqrt{x} < \sqrt{y}$.

Proposition 4.6. *If $r \in \mathbb{Q}$, $r > 0$ is a positive rational number and $x, y \in F$ are positive field elements*

$$x < y \implies x^r < y^r$$

Proof. Use that $x^r = x^{p/q} = (\sqrt[q]{x^p})$ to break this into two problems: first $x < y$ implies $x^p < y^p$. Now, if $u = x^p$ and $v = y^p$ we have $u < v \implies \sqrt[q]{u} < \sqrt[q]{v}$, completing the proof. ☐

4.3. Working with Absolute Values

Proposition 4.7 (Absolute Values and Maxima)**.** *For all x in an ordered field,*

$$|x| = \max\{x, -x\}$$

Corollary 4.1. *If x, a are in an ordered field, the conditions $-x < a$ and $x < a$ are equivalent to*

$$|x| < a$$

Proof. If $-x < a$ and $x < a$ then $\max\{x, -x\} < a$, so by Proposition 4.7, $|x| < a$. Conversely, if $|x| < a$ then $\max\{x, -x\} < a$ so both $x < a$ and $-x < a$. ☐

Corollary 4.2 (Defining Feature of the Absolute Value)**.** *Let F be an ordered field: then $|x| < a$ if and only if $-a < x < a$.*

Proof. By the above $|x| < a$ means $x < a$ and $-x < a$. Multiplying the second inequality by -1 yields $x > -a$, and stringing them together results in $-a < x < a$. □

Finally, we can get a *formula* for the absolute value in terms of squaring and roots.

Example 4.3. For all x in an ordered field $|x| = \sqrt{x^2}$.

Example 4.4 (Multiplication and the Absolute Value)**.**

$$|xy| = |x||y|$$

$$\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$$

The interaction of the absolute value with *addition* is more subtle, but crucial. One of the most important inequalities in all of analysis is the *triangle inequality* of the absolute value:

Proposition 4.8 (The Triangle Inequality)**.** *For any x, y in an ordered field*

$$|x + y| \leq |x| + |y|$$

Proof. It suffices to prove that we have both

$$x + y \leq |x| + |y| \qquad -(x + y) \leq |x| + |y|$$

For the first, note that as $x \leq |x|$ and $y \leq |y|$,

$$x + y \leq |x| + y \leq |x| + |y|$$

Similar reasoning succeeds for the second as $-x \leq |x|$ and $-y \leq |y|$:

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

□

Exercise 4.8. Let $a_1 + a_2 + \cdots + a_n$ be any finite sum. Prove that

$$\left|\sum_{i=1}^{n} a_i\right| \leq \sum_{i=1}^{n} |a_i|$$

The reverse triangle inequality is another very useful property of absolute values, logically equivalent to the usual triangle inequality, but giving a lower bound for $|a - b|$ instead of an upper bound for $|a + b|$.

Exercise 4.9 (Reverse Triangle Inequality)**.** Prove that for all a, b in an ordered field F

$$||a| - |b|| \leq |a - b|$$

Finally, two corollaries of the triangle inequality and its reverse, by replacing y with $-y$.

Corollary 4.3 (Corollaries of the Triangle Inequality)**.** *For all x, y in an ordered field,*

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

$$|x + y| \geq ||x| - |y||$$

4.4. ⋆ Topology

A final familiar property that arises from ordering a field is the notion of *open sets* and *closed sets*. This in turn is the foundations of the subject of *topology* or the abstract study of shape, which becomes quite important in advanced applications of analysis.

We will not require any deep theory in this course, and stop pause briefly to give a definition of openness and closedness.

Definition 4.8 (Open Set)**.** A set of the form $(a, b) = \{x \mid a < x < b\}$ is called an *open interval*. A set $U \subset \mathbb{F}$ is called open, if for every point $u \in U$ there is some open interval I containing u which is fully contained in U:

$$u \in I \subset U$$

One notable property of this definition: the empty set $\emptyset = \{\}$ is open, as this condition is *vacuously true*: there are no points of $\emptyset$ so this condition doesn't pose any restriction!

Exercise 4.10. Explain why the set $U = \{x \mid x > 0 \text{ and } x \neq 2\}$ is an open set.

Exercise 4.11. Let $\{U_n\}$ be *any* collection of open sets. Prove that the union $\bigcup_n U_n$ is also open.

Hint: his collection doesn't have to be finite, so induction won't help us here. Can you supply a direct proof, using the definition of union and open?

Definition 4.9 (Closed Set)**.** A set is $K \subset \mathbb{F}$ is *closed*, if its complement is an open set.

Exercise 4.12. Show that intervals of the form $[a, b] = \{x \mid a \leq x \leq b\}$ are closed sets. This is why we call them *closed intervals* in calculus courses.

This terminology is rather unfortunate when first learning the subject, as while open and closed are antonyms in english, they are not in mathematics! Being open is a special property that most sets do not have, and so being closed (which is defined relative to an open set) is also a special property. Most sets are neither open nor closed!

Example 4.5 (A set that is neither open nor closed)**.** The set $S = [1, 2)$ is neither open nor closed. Its not open because the point $1 \in S$, but there is no open interval containing 1 which is fully contained in S (every open interval containing 1 contains numbers *smaller than 1* as well).

To see its not closed, we need to show that its complement is not open. Its complement is the set

$$S^c = \{x \mid x < 1\} \cup \{x \mid x \geq 2\}$$

Here we have the same problem at the number two: $2 \in S^c$ but there is no open interval containing 2 which is fully inside S^c, as any such interval would contain points less than 2, and these are not in S^c.

Thus, $[1, 2)$ is neither open nor closed.

But perhaps even stranger, not only can sets be *neither* open nor closed, but they can also be *both* open and closed! Such sets are called *clopen*.

Example 4.6 (A set that is both open and closed)**.** If $\mathbb{F}$ is the entire ordered field, $\mathbb{F}$ is both open and closed.

To see it is open, note for any x we can form the interval $(x - 1, x + 1)$ and this lies inside of $\mathbb{F}$. To see its closed, note that its complement is the empty set and this is vacuously open as commented above.

5. Completeness

> **Highlights of this Chapter:** We look to formalize the notion of limit used by the babylonians and archimedes, and come to the Nested Interval Property. We see that this property does not hold in $\mathbb{Q}$, so we must seek another axiom which implies us. This leads us to bounds, infima, and suprema. We study the properties of this new definition, use it to define completeness, and show completeness does indeed imply the nested interval property, as we wished.

Now that we have axiomatized the notion of a 'number line' as an ordered field, it's time to try and figure out how to describe "completed" infinite processes in a formal way. This is an inherently slippery notion, as it runs into the difficulty of "talking about infinity, without saying infinity" that lies at the heart of analysis.

So, before introducing the abstract tools that end up best suited for this task (the infimum and supremum), we'll begin with some motivational exploration, and think about what sort of theorems we would *want to be true* in a number system that allows one to do infinite constructions.

5.1. Dreaming of Infinity

Archimedes idea for calculating π was to give an upper bound and a lower bound for the area of a circle, in terms of the area a_n of an inscribed polygon and a circumscribed polygon A_n. This provided an interval that archimedes hoped to trap π inside of, each time n grows, a_n grows and A_n shrinks - so the *confidence interval* of Archimedes shrinks!

$$\cdots [a_6, A_6] \supset [a_{12}, A_{12}] \supset [a_{24}, A_{24}] \supset \cdots$$

A collection of intervals like this is called nested:

Definition 5.1 (Nested Intervals). A sequence of intervals $I_1, I_2, I_3, \ldots$ in an ordered field is called *nested* if for all n, $I_{n+1} \subseteq I_n$.

As these nested intervals shrink in size, the hope is that they zero in on π exactly: mathematically we might express this with an *intersection* over all intervals (where the question mark over the equals means we have not proven this, but hope its true)

$$\bigcap_n [a_n, A_n] \stackrel{?}{=} \{\pi\}$$

The babylonian process approximating $\sqrt{2}$ can also be recast in terms of a sequence of nested intervals: where we take the two sides w_n, h_n (width and height) of each approximating rectangle as a confidence interval around $\sqrt{2}$. We of course want, that in the limit this zeroes in directly on the square root,

$$\bigcap_n [w_n, h_n] \stackrel{?}{=} \{\sqrt{2}\}$$

In formulating any of these processes (pre-rigorously, say, in antiquity) mathematicians always assumed without proof that if you had a collection of shrinking intervals, they were shrinking around *some real number* that could be captured after infinitely many steps. One way to formalize this hope would be the following 'dream theorem'.

> **Dream Theorem:** In a *complete number system*, every sequence of nested intervals has a nonempty intersection.

This sort 'dream result' has a strange but important status in mathematics: its not a theorem we can prove right now, but rather a guiding light as we march forward. We should investigate our current axioms and ask if they can prove this - and, if not, we should look for *additional axioms* to improve our notion of number line until we can!

How do we tell if our current axioms imply this dream theorem? In a situation like this, mathematicians may try to ask *what sort of things satisfy the current axioms* and look at these for inspiration. Here - the rational numbers satisfy the axioms of an ordered field, and this provides a big hint: Pythagoras proved that there is no rational square root of 2, which implies the Babylonian process does not zero in on any number at all, but rather at infinity reaches nothing!

$$\bigcap_n [w_n, h_n] = \emptyset$$

Because there is at least one ordered field (the rationals) that does not satisfy the dream theorem, we know that these axioms are not enough.

> The axioms of an ordered field are not enough to deal with completed infinity: there are ordered fields in which the dream theorem is false.

This tells us we must look to *extend our axiom system* and search out a new axiom that will help our number system capture the slippery notion of infinite processes. Happily, it turns out a productive approach to this grows naturally out of our discussion of nested intervals. But, to decrease the complexity instead of focusing on the entire interval $[\ell_n, u_n]$, we will look separately at the sequence of *lower bounds* ℓ_n and *upper bounds* u_n. Understanding the behavior of either of these will turn out to be enough to extend our axiom system appropriately.

5.2. Suprema and Infima

A confidence interval like $[\text{width}_n, \text{height}_n]$ or $[\text{inscribed}_n, \text{circumscribed}_n]$ gives us for each n both an *upper bound* for the number we are after, and a *lower bound*.
It will be useful to describe these concepts more precisely.

Definition 5.2 (Bounds)**.** Let S be a nonempty subset of an ordered field. An *upper bound* for S is an element $u \in \mathbb{F}$ greater than or equal to all the elements of S:

$$\forall s \in S \, s \leq u$$

A *lower bound* for S is an element $\ell \in \mathbb{F}$ which is less than or equal to all the elements of S:

$$\forall s \in S \, \ell \leq u$$

S is said to be *bounded above* if there exists an upper bound, and to be bounded below if there exists a lower bound. If S is both bounded above and below, then S is said to be *bounded*.

Definition 5.3 (Maximum & Minimum)**.** Let S be a nonempty subset of an ordered field. Then S has a \$maximum* if there is an element of $M \in S$ that is also an upper bound for S, and a *minimum* if some element m is also a lower bound for S.

The maximum and minimum elements of a set are the *best possible* upper and lower bounds when they exist: after all, you couldn't hope to find a *smaller* lower bound than the maximum, as the maximum would be greater than it, so it couldn't be an upper bound! While maxima and minima always exist for *finite sets* things get trickier with infinity. For example, the open interval $(0, 1)$ of rational numbers does not have any maximum element.

The correct generalization of *maximum* to cases like this is called the *supremum*: the best possible upper bound.

Definition 5.4 (Supremum)**.** Let S be a set which is bounded above. The *least upper bound* for S is a number σ such that

- σ is an upper bound for S
- If u is any upper bound, then $\sigma \leq u$.

When such a least upper bound exists, we call it the *supremum of S* and denote it $\sigma = \sup S$.

This notion of best possible upper bound allows us to rigorously capture the notion of *endpoint* even for infinite sets that do not have a maximum.

Example 5.1 (A set with no maximum)**.** The set $(0,1) = \{x \in \mathbb{Q} \mid 0 < x < 1\}$ has no maximal element, but it does have a supremum in $\mathbb{Q}$, namely $1 = \sup S$.

Definition 5.5 (Infimum)**.** The infimum of a set S is the *least upper bound*: that is, an element λ where

- λ is a lower bound for S.
- If ℓ is any other lower bound for S, then $\ell \leq \lambda$.

If such an element exists it is denoted $\lambda = \inf S$.

Example 5.2.

- The set $\mathbb{N}$ has no upper bounds at all, so $\sup \mathbb{N}$ does not exist. It has many lower bounds (like 0, and -14), and its infimum is $\inf \mathbb{N} = 1$.

- The rational numbers themselves have no upper nor lower bound, so $\sup \mathbb{Q}$ and $\inf \mathbb{Q}$ do not exist.

Exercise 5.1. Consider the following subsets of the rational numbers. State whether or not they have infima or suprema; when they do, give the inf and sup.

- $[1,3]$
- $[1,3)$
- $\{x \mid x^2 < 1\}$
- $\{x \mid x^3 < 1\}$
- $\{x \mid 1 + \frac{1}{n}, n \in \mathbb{N}\}$
- $\{x \mid 1 + \frac{(-1)^n}{n},\ n \in \mathbb{N}\}$

5.2.1. Infinite Processes

By focusing on one bound at at time, this new terminology lets us rigorously capture the ideas of the babylonians and archimedes, by giving a name to the *idealized endpoint* of their infinite processes.

Example 5.3 (The Babylonian $\sqrt{2}$)**.** Define the set of numbers S by $s_0 = 1$ and $s_n = \frac{s_n + 2/s_n}{2}$, so

$$S = \left\{1, \frac{3}{2}, \frac{577}{408}, \frac{665857}{470832} \dots\right\}$$

These are all lower bounds for $\sqrt{2}$ coming from the widths of a rectangle approximating a square that started with side length 1. With each increasing n, this approximation gets larger (and better), which gives a hint on how to formalize this in our new vocabulary.

Consider this entire collection *as a set* $\{s_0, s_1, s_2, s_3, \dots, s_n, \dots\}$ with each n this set. As the *larger* elements of this set are *better* approximations of $\sqrt{2}$, we can define the result of this infinite process with the *least upper bound* or supremum $\sup S$.

Example 5.4 (Inscribed Perimeters of Archimedes)**.** Archimedes' study of inscribed n-gons give an ever increasing sequence of areas and perimeters, approaching (but never reaching at any finite stage) the area and perimeter of the unit circle. We review these below:

Let $p_6 = 6$, and define p_{2n} recursively as in Example 1.2: $p_{2n} = 2n\sqrt{2 - \sqrt{4 - \left(\frac{p_n}{n}\right)^2}}$. Defining P as the set of numbers produced by this recurrence,

$$P = \{p_6, p_{12}, p_{24}, p_{48}, p_{96}, \dots, p_{24576}, \dots\}$$

Each individual number here gives us a *lower bound* for the circles circumference, as it is computed from a polygon inside the circle. Because this sequence is increasing, larger elements of P represent *better approximations*, so we can rigorously define the result after infinitely many doublings as the least upper bound to P, or $\sup P$.

Example 5.5 (Circumscribed Perimeters of Archimedes)**.** The other side of archimedes' confidence intervals were computed using circumscribed polygons. Following Exercise 1.10, we precisely define the circumscribed perimeters by setting $P_4 = 8$ and using the recurrence $t_{2n} = \sqrt{1 + \frac{1}{t_n^2}} - \frac{1}{t_n}$ for the tangent, to get a recurrence for P_{2n} in terms of P_n.

This sequence is *decreasing* as n grows, so the better estimates for circumference come from the *smaller numbers*. We formalize this by creating the set P of all approximate perimeters

$$P = \{P_4, P_8, P_{16}, P_{32}, P_{64}, \cdots\}$$

and speaking of the final result of the infinite process as the *infimum* or $\inf P$

One reason that suprema and infima are a useful technology to develop is that they are *more general than nested intervals*: we can make sense of them to talk about any infinite processes, even ones that we naturally have only a lower, or upper estimate for.

Exercise 5.2. Give a formal statement of the Basel problem solved by Euler in terms of infima or suprema.

Exercise 5.3. Give a formal statement of the result of the infinite product of Viete in terms of infima or suprema. *Hint: first figure out, as you add new terms to this approximation, is it increasing or decreasing in value?*

5.3. Completeness

Because infima and suprema are such a useful tool to precisely describe the final state of certain infinite processes, they are a natural choice of object to concentrate on when looking for an additional axiom for our number system. Indeed - after some thought you can convince yourself that the statement *every infinite process that should end in some number, does end in some number* is equivalent to the following definition of *completeness*.

Definition 5.6 (Completeness)**.** An ordered set is *complete* if every nonempty subset S that is bounded above has a supremum.

Remark 5.1. One question you might ask yourself is why we chose *supremum* here, and not *infimum* - or better, why not both?! It turns out that all of these options are *logically equivalent*, as you can prove in some exercises below. So, any one of them suffices

We can formalize Pythagoras' observation about the irrationality of $\sqrt{2}$ in this language

Theorem 5.1 ($\mathbb{Q}$ is not complete)**.** *The set* $S = \{s \in \mathbb{Q} \mid s^2 < 2\}$ *does not have a supremum in* $\mathbb{Q}$*.*

Sketch. A rigorous proof can be given by contradiction: assume that a supremum $\sigma = \sup S$ exists, and then show that we must have $\sigma^2 = 2$ by ruling out the possibilities $\sigma^2 < 2$ and $\sigma^2 > 2$. The calculations required for these steps are more relevant to the next chapter, so we postpone until then (specifically, Example 6.1 and Exercise 6.4).

Once its known that the supremum must satisfy $\sigma^2 = 2$, we apply Pythagoras' observation (Theorem 1.1) that there are no rational solutions to this equation, to reach a contradiction. □

Thus, asking a field to be complete is a constraint above and beyond being an ordered field. So, this is a good candidate for an additional axiom! But before we too hastily accept it, we should check that it actually solves our problem:

Theorem 5.2 (Nested Interval Property)**.** *Let $\mathbb{F}$ be an ordered field* ***which is also complete****, and $I_0, I_1, I_2, \ldots, I_n, \ldots$ be a collection of nested closed intervals. Then their intersection is nonempty:*

$$\bigcap_{n\geq 0} I_n \neq \emptyset$$

Proof. Let $I_n = [a_n, b_n]$. We need to use the fact that $\mathbb{F}$ is complete to help us find a number which lies in I_n for every n. One idea - consider the set of lower endpoints

$$A = \{a_1, a_2, a_3, \ldots, a_n, \ldots\}$$

This set is nonempty, and because the intervals are nested any one of the b_n's serves as an upper bound for A.

By completeness the supremum must exist: lets call this $\alpha = \sup A$. Now we just need to see that $\alpha \in I_n = [a_n, b_n]$ for all n. Fix some n: then as $a_n \in A$ and α is an upper bound, we know that $a_n \leq \alpha$. But b_n is an upper bound for A so the *least upper bound* must satisfy $\alpha \leq b_n$. Putting these together

$$a_n \leq \alpha \leq b_n \implies \alpha \in I_n$$

And, since this holds for all natural numbers n, we actually have $\alpha \in \bigcap_n I_n$, so the intersection is nonempty. □

We can even take these tools farther, and see that infima and suprema can tell us exactly to a process that produces confidence intervals $[a_n, b_n]$ after infinitely many steps: if A is the set of lower endpoints and B is the set of upper endpoints, we can prove (1) $\sup A$ and $\inf B$ both exist, and (2) if the lengths of the intervals I_n tend to zero, then the nested intersection actually contains just a single point: the number we are after!

This is exactly the validation we needed: while ordered fields do not have enough structure to formalize the infinite processes undertaken in mathematics, *complete ordered fields* do satisfy the dream theorem! We will study their properties intensively in the next chapter. But for now, we turn to the concept of supremum and infimum themselves, as these seemingly simple ideas will underlie our entire theory of the real numbers.

Exercise 5.4. The proof of the nested interval theorem used the *endpoints* of the intervals crucially in the proof. One might wonder if the same theorem holds for *open intervals* (even though the proof would have to change).

Show the analogous theorem for open intervals is false by finding a counter example: can you find a collection of nested open intervals whose intersection is empty?

Exercise 5.5. Either give an example of each (explaining why your example works) or provide an argument (it doesn't have to be a formal proof) why no such example should exist:

- A sequence of nested closed intervals, whose intersection contains exactly n points, for some finite $n > 1$.
- A sequence of nested closed rays whose intersection is empty. (A closed ray has the form $[a, \infty)$ or $(-\infty, a]$ as in Definition 4.4).

5.4. Working with inf and sup

Proposition 5.1 (Uniqueness of Supremum). *If the supremum of a set exists, it is unique.*

Proof. Let A be a set. To show uniqueness, we will assume that there are two numbers x and y which both satisfy the definition of the supremum of A, and then we will show $x = y$. Thus, any two possibilities for the supremum are equal, so if theres a supremum at all there can only be one.

To prove $x = y$, we will prove $x \leq y$ and $y \leq x$. Once we have these two, we can immediately conclude that since we can't simultaneously have $x < y$ and $y < x$ (what axiom of an ordered field would this violate?) we must have $x = y$.

If x and y both are least upper bounds for A, then they are both in particular upper bounds. So, x is an upper bound and y is a *least upper bound* implies $y \leq x$. But similarly, y being an upper bound while x is a *least upper bound* implies $x \leq y$. Thus $x = y$ and so the supremum is unique. □

Remark 5.2. These are two important proof techniques in analysis.
First, one way to show that something is *unique* is to show that if you had two of them, they have to be equal. Second, to show $x = y$ it is often useful to show both $x \leq y$ and $y \leq x$.

Exercise 5.6. Prove the infimum of a set is unique when it exists.

Proposition 5.2. *Let A be a set which is bounded above. An upper bound α for A is actually the supremum if for every positive $\epsilon > 0$, there exists some element of A greater than $\alpha - \epsilon$.*

Proof. **(In the book, Theorem 1.24, page 26)** Let's prove the *contrapositive*, meaning we assume the *conclusion* is false and prove the *premise* is false. The conclusion would be false if there were *some* positive ϵ where no element of a is larger than $\alpha - \epsilon$. But this means that $\alpha - \epsilon \geq a$ for all $a \in A$, or that $\alpha - \epsilon$ is an upper bound for A. Since this is less than α (remember, ϵ is positive), we found a smaller upper bound, so α cannot be the least upper bound: thus its *false* that $\alpha = \sup A$.

Since anytime our proposed condition doesn't hold, α isnt the supremum, this means if α were the supremum, the condition must hold! And this is what we sought to prove. □

Remark 5.3. The contrapositive is a very useful proof style, especially in situations where the premise is something short, and the conclusion is something complicated. By taking a look at the contrapositive, you get to assume the negation of the conclusion, meaning you get to assume the complicated thing, and then use it to prove the simple thing (the negation of the premise

Exercise 5.7. Prove the corresponding characterization of infima: a lower bound ℓ for a set A is the infimum if for every positive $\epsilon > 0$ there is some element of A less than $\ell + \epsilon$.

Exercise 5.8. Let A, B be nonempty bounded subsets of a complete field, and suppose $A \subset B$. Prove that $\sup A \leq \sup B$.

Exercise 5.9. Let A, B be subsets of a complete ordered field with $\sup A < \sup B$.

- Prove that there is an element $b \in B$ which is an upper bound for A.
- Give an example to show this is not necessarily true if we only assume $\sup A \leq \sup B$.

Example 5.6. Let A be a bounded set with supremum $\sup A$ and c an element of the field. Define the set $S = \{a + c \mid a \in A\}$. Then

$$\sup S = c + \sup A$$

To prove this, we need to show two things: (1) that $c + \sup A$ is an upper bound for S, and (2) that its in fact the least upper bound.

First, we consdier (1). Since $\sup A$ is an upper bound for A, we know $\forall a \in A, a \leq \sup A$. Adding c to both sides, we also have $c + a \leq c + \sup A$ for all a, which implies $c + \sup A$ is an upper bound.

Now, (2). Let u be *any upper bound* for S. This means that $u \geq c + a$ for all $a \in A$, so subtracting c from both sides, that $u - c \geq a$. Thus, $u - c$ is an upper bound for A, and this is real progress because we know $\sup A$ is the *least upper bound.* That implies $\sup A \leq u - c$ and so adding c to both sides, $c + \sup A \leq u$. Putting this all together, we assumed u was *any upper bound* and we proved $c + \sup A$ was a smaller one.

Thus, $c + \sup A$ is the least upper bound to S, and so by definition we have $\sup S = c + \sup A$ as required.

Exercise 5.10. Let $c > 0$ and A be a bounded set with supremum $\sup A$. Define the set $S = \{ca \mid a \in A\}$. Then $\sup S$ exists and

$$\sup S = c \sup A$$

Exercise 5.11. Let A, B be two bounded nonempty sets. Assuming that the suprema and infima of A and B both exist, prove they do for $A \cup B$ as well and

$$\sup A \cup B = \max\{\sup A, \sup B\}$$

$$\inf A \cup B = \min\{\inf A, \inf B\}$$

Exercise 5.12. For each item, compute the supremum and infimum, or explain why they does not exist. (You should explain your answers but you do not need to give a rigorous proof)

- $A = \left\{\frac{(-1)^n}{n} \mid n \in \mathbb{N}\right\}$$
- Fix $\beta \in (0, 1)$, and define $B = \{\beta^n \mid n \in \mathbb{N}\}$
- Fix $\gamma \in (1, \infty)$ and define $C = \{\gamma^n \mid n \in \mathbb{N}\}$.

Exercise 5.13 (Sup and Inf of Intervals)**.** Let A, B be two open intervals in $\mathbb{R}$, and assume that $\sup A = \inf B$.

True or false: it is possible to add a single point to $A \cup B$ so the entire set is an interval. (Explain your reasoning, but you don't have to write a rigorous proof).

5.4.1. ⋆ Equivalents to Completeness

Here we tackle the natural questions about why we chose *suprema* to codify completeness in a series of exercises. Our goal at the end of these is to show that the following three possible completeness axioms are all logically equivalent:

- (1) Any nonempty set thats bounded above has a supremum.
- (2) Any nonempty set thats bounded below has an infimum.
- (3) Any nonempty set thats bounded has a supremum and infimum.

Exercise 5.14. For a set A let $-A$ denote the set of additive inverses: $-A = \{-a \mid a \in A\}$. Prove that in a complete field if A is nonempty and bounded below then

$$\sup(-A) = -\inf(A)$$

Thus, assuming that suprema exist forces infima to exist, so in our list above, (1) implies (2).

Exercise 5.15. Prove the converse of the above: if we instead assume that the infimum of every nonempty set thats bounded below exists, show that the supremum of every nonempty set thats bounded above exists.

This shows (2) implies (1), so all together we know that (1) and (2) are equivalent. But since (3) is just the conditions (1) and (2) together, we can derive (3) from either as

$$(1) \implies (1) \text{ and } (2) = (3)$$
$$(2) \implies (2) \text{ and } (1) = (3)$$

Thus both (1) and (2) imply (3). But since (1) and (2) are themselves special cases of (3), we already know (3) implies each of them! So, both of (1) and (2) are equivalent to (3), and all three conditions are logically equivalent to one another.

6. The Real Numbers

Highlights of this Chapter: We see that there exists a unique complete ordered field, and use this to axiomatically define the real numbers. We then investigate properties of this number system, proving several foundational results related to the archimedean property and nested intervals:

- The real numbers do not contain any infinite numbers or infinitesimals.
- The square root of 2 is a real number.
- The rational numbers are dense in the reals.
- The real numbers are uncountable.

We have now carefully axiomatized the properties that are used in classical mathematics when dealing with the number line, defining a the structure of a *complete ordered field*.

Definition 6.1. A complete ordered field is an ordered field that satisfies the completeness axiom. Precisely, it is a set $\mathbb{F}$ with the following properties

- **Addition:** A commutative associative operation $+$, with identity 0, where ever element has an additive inverse.
- **Multiplication:** A commutative associative operation $\cdot$ with identity $1 \neq 0$, where every nonzero element has a multiplicative inverse.
- **Distributivity:** For all $a, b, c \in \mathbb{F}$ we have $a(b+c) = ab + ac$
- **Order:** A subset $P \subset \mathbb{F}$ called *the positives* containing exactly one of $x, -x$ for every nonzero $x \in \mathbb{F}$, which is closed under addition and multiplication: if $a, b \in P$ then $a + b \in P$ and $ab \in P$.
- **Completeness:** Every nonempty subset $A \subset \mathbb{F}$ which is bounded above has a least upper bound.

The subject of *real analysis* is the study of complete ordered fields and their properties, so everything that follows in this course logically follows from this set of axioms, *and nothing more.* The success and importance of the above definition is best exemplified by the following theorem:

Remark 6.1. This was very important work at the turn of the previous century; as neither step is a priori obvious. It's easy to write down axiom systems that don't describe *anything* because they're inconsistent (for example, add to ordered field axioms that all polynomials have at least one zero, and there is no longer such a structure), and its also common that axioms don't uniquely pick out a single object but rather describe an entire class (the axioms of a group define a whole subject, not a single example).

Theorem 6.1 (Uniqueness of the Reals)**.** *There exists a complete ordered field, and it is unique. We call this field* the real numbers *and denote it by* $\mathbb{R}$.

This theorem represents the culmination of much work at the end of the 19th and beginning of the 20th century to fully understand the real number line.

While not necessarily beyond our abilities, proving existence of a structure satisfying these axioms is a job for the set theorists and logicians that we will not tackle here.

Beyond providing justification for our usual way of speaking, the uniqueness of the reals is an important result to the history of mathematics. Its statement and proof in 1903 by Huntington marked the end of the era of searching for the fundamental principles behind the real numbers, and the beginning of the modern point of view, completely specifying their structure axiomatically.

Remark 6.2. The completeness axiom is what sets analysis apart from algebra, as it does not tell us how elements behave with respect to a given operation, but rather tells us about the *existence* of new elements. Indeed, this assertive ability of the completeness axiom is more radical than it seems at first, and can even be captured by mathematical logic: the other axioms are all *first order axioms*, whereas the completeness axiom is *second order.*

6.0.1. Dubious Numbers

Now that we have a precise definition of the real number line, we can make precise the philosophical questions that were raised during the origin story of The Calculus.

The first of these was the concept of a *nilpotent number*, something so small that its square was literally equal to zero.

Definition 6.2 (Nilpotent Numbers)**.** A number is *nilpotent* if $\epsilon \neq 0$ but $\epsilon^2 = 0$.

Such numbers were often used in justifying various calculations of the derivative (and continue to be used, as heuristic arguments in introductory calculus and science courses). But it is immediate from even just the *field axioms* that no such numbers exist.

Proposition 6.1 (Fields have no Nilpotent Numbers)**.** *Let $\$\mathbb{F}$ be any field, and ϵ some number where $\epsilon^2 = 0$. By the zero-product-property (Example 3.7), this implies $\epsilon = 0$. Thus there are no nonzero elements that square to zero.*

The other two classes of numbers proposed to help make sense of the calculus were infinite numbers (to represent the number of summands in an infinite sum, for instance) and infinitesimal numbers (as we saw with differentiation).

Definition 6.3 (Infinite Numbers)**.** A number x is finite if its bounded above and below by integers. If a number is not finite, its said to be infinite.

Equivalently, a number x is *infinite* if either it or its negation is greater than all positive integers.

Definition 6.4 (Infinitesimal Numbers)**.** A positive number ϵ is infinitesimal if it is smaller than $1/n$ for all n.

So of course, the next thing we should do is figure out if these numbers exist!

6.1. Infinites and Infinitesimals

Theorem 6.2 (Infinite Numbers Do Not Exist)**.** *There are no infinite elements of* $\mathbb{R}$.

Proof. Assume for the sake of contradiction that there is some infinite number: without loss of generality (perhaps after multiplying by -1) we may assume its positive. Thus, this number is greater than every natural number, and so the natural numbers are bounded above.

Thus, by the completeness axiom, we find that the natural numbers must have a supremum. Denote this by $X = \sup \mathbb{N}$. So far, everything seems fine. But consider the number $X - 1$. This is smaller than X, and since X is the *least upper bound*, $X - 1$ cannot be an upper bound to $\mathbb{N}$. This means there must be some element $n \in \mathbb{N}$ with $n > X - 1$. But this means $X < n + 1$, and as $n + 1$ is a natural number whenever n is, we've run headfirst into a contradiction: X is not an upper bound at all! □

It is an immediate corollary of this that infinitesimals also do not exist (but, because this is such an important result, we call it a theorem on its own.)

Theorem 6.3 (Infinitesimals Do Not Exist)**.** *There are no infinitesimal elements of* $\mathbb{R}$.

Proof. Let x be a positive element of $\mathbb{R}$, and consider its reciprocal $1/x$. By Theorem 6.2 $1/x$ is finite, so there's some $n \in \mathbb{N}$ with $n > 1/x$. Re-arranging the inequality shows $x > 1/n$ as required, so x is not infinitesimal. □

This argument shows that for a field, containing infinite elements and infinitesimal elements are *logically equivalent*: thanks to division, you can't have one without the other.

Exercise 6.1 (An Flawed Argument for the Nonexistence of Infinitesimals)**.** Its possible to show that the infimum of the set of all positive real numbers is zero, just from the definition of infimum.

- Prove this
- Explain why this is not enough to conclude the nonexistence of infinitesimals.

These theorems are fundamental to the foundations of calculus. We know the *rationals* do not contain infinites or infinitesimals, but to go from the rationals to the reals its quite possible that such numbers were added. After all, the field of real numbers is *defined axiomatically* - we don't know what the elements are we just know how to work with them! And many throughout the history of calculus assumed the reals contained infinitesimals....but it turns out they were all wrong.

6.1.1. Archimedean Property

A useful way to repackage the nonexistence of infinite numbers and infinitesimals into a *usable statement* known as the *Archimedean property*, as Archimedes took it as an axiom describing the number system in his paper *The Sphere and the Cylinder.* It also appears (earlier) as a definition in Euclid's elements: Book V Definition 4:

> Magnitudes are said to have a ratio to one another which can, when multiplied, exceed one another.

We rephrase this in precise modern terminology below:

Definition 6.5 (Archimedean Field)**.** A field $\mathbb{F}$ is *archimedean* if for every positive $a, b \in \mathbb{F}$ there is a natural number n with

$$na > b$$

Remark 6.3. While Archimedes himself attributes this to Eudoxus of Cnidus, it was named after Archimedes in the 1880s.

The important applications of this property all come from the case where b is really large, and a is really small. In an archimedean field, no matter how small a is you can always collect enough of them $na = a + a + a + \cdots + a$ to surpass b. A common way to remember this property is to poetically rephrase it as *you can empty the ocean with a teaspoon.*

Its possible to give an elementary proof (directly from the definition of rational numbers as fradtions p/q for $p, q \in \mathbb{Z}$, $q \neq 0$) that $\mathbb{Q}$ is an archimedean field:

Exercise 6.2 (The Rationals are Archimedean)**.** Prove the rationals are an archimedean field. *Hint: write a and b as fractions, can you figure out from the inequality you want, what n can be?*

Such a proof is not possible for $\mathbb{R}$ as we don't have an explicit description of its elements! All we know is its axiomatic properties. However, a proof is immediate using Theorem 6.2:

Theorem 6.4 (The Reals are Archimedean)**.** *Complete ordered fields satisfy the Archimedean property.*

Proof. **(In the book, Lemma 1.26, page 28)** Let a, b be positive real numbers. Since $b/a \in \mathbb{R}$ it is finite (by Theorem 6.2), so there is some $n \in \mathbb{N}$ with $n > \frac{b}{a}$, and thus $na > b$. □

Its also a short proof to show that archimedean fields cannot contain infinite elements (and thus also cannot contain infinitesimals), providing a useful equivalence:

Theorem 6.5. *The following three conditions are equivalent, for an ordered field* $\mathbb{F}$*:*

- $\mathbb{F}$ *is archimedean.*
- $\mathbb{F}$ *contains no infinite elements.*
- $\mathbb{F}$ *contains no infinitesimal elements.*

Proof. We already know the existence of infinite elements and infinitesimal elements are equivalent, so all we need to show is that $\mathbb{F}$ is archimedean if and only if all elements are finite.

But the proof of Theorem 6.4 already provides an argument that a field with only finite elements is necessarily archimedean, so we seek only the converse.

If $\mathbb{F}$ is archimedean, then for any positive $b \in \mathbb{F}$ we may take $a = 1$ and apply the archimedean property to get an $n \in \mathbb{N}$ with $n \cdot 1 > b$. For negative b, applying the ame to $-b$ results in a $n \in \mathbb{N}$ where $-n < b$, and together these imply all elements of $\mathbb{F}$ are finite. □

Remark 6.4. In fact one can be more precise than this: it turns out that the real numbers are the *largest possible archimedean field* - and every archimedean field fits somewhere between the rationals and the reals.

Originating in the foundations of analysis, the archimedean property has proven a useful guide in the general study of ordered fields. When encountering a new ordered field, one of the first questions one usually asks is *is it archimedean*? If so, we know immediately that it does not contain any infinites or infinitesimals, and one one can use intuition from the rationals or real numbers. *Non-archimedean fields* on the other hand are a totally different beast, and lead to theories rather qualitatively different from real analysis.

Exercise 6.3. Prove that the supremum of the set $S = \{\frac{n}{n+1} \mid n \in \mathbb{N}\}$ is 1.

6.2. Irrationals

Definition 6.6 (Irrational Numbers)**.** A number $x \in \mathbb{R}$ is irrational if it is not rational.

6.2.1. Existence of $\sqrt{2}$

Our first goal is to prove that irrational numbers exist, by exhibiting one. We will use the example of the square root of two, and rigorously prove that $\sqrt{2}$ is a real number. (Just so you don't brush this off as trivial, its not immediately obvious: after all, $\sqrt{-2}$ is not a real number!)

Theorem 6.6. *Let $\mathbb{F}$ be archimedean, and consider the set*

$$S = \{r \in \mathbb{F} \mid r^2 < 2\}$$

Then if $\sigma = \sup S$ exists, $\sigma^2 = 2$.

We prove this rather indirectly, showing that both $\sigma^2 > 2$ and $\sigma^2 < 2$ are impossible, so the only remaining option is $\sigma^2 = 2$.

Example 6.1 ($\sigma^2 > 2$ is impossible.)**.** To show this is impossible, we will show if you have *any upper bound $b \in \mathbb{F}$* with $b^2 > 2$, it's not the *least upper bound*, as we can make a smaller one.

Let b be any upper bound with $b^2 > 2$. To find a smaller upper bound, one idea is to try and find a natural number n where $\beta = b - 1/n$ works. That is,

$$\left(b - \frac{1}{n}\right)^2 > 2$$

Expanding this out, we see $b^2 - 2b/n + 1/n^2 > 2$, or after moving terms around, $b^2 - 2 > 2b/n - 1/n^2$. Now we need a little ingenuity: notice that $2b/n - 1/n^2$ is less than $2b/n$ (because we're subtracting something) so in fact, if we can find an n where $2b/n < b^2 - 2$ we're already good. Re-arranging this equation, we need to find n with

$$(b^2 - 2)n > 2b$$

But this is possible using the Archimedean property! Since $A = b^2 - 2$ and $B = 2b$ are both positive numbers, we can always find an $n \in \mathbb{N}$ where $nA > B$. Thus, we may choose this value of n, and note that $\beta = b - \frac{1}{n}$ is an upper bound for S that is smaller than b. Thus b was not the least upper bound!

Exercise 6.4 ($\sigma^2 < 2$ is impossible.)**.** Can you preform an argument similar to Example 6.1, to prove that $\sigma^2 < 2$ also leads to contradiction?

Since both the real numbers and the rationals are archimedean, the above applies to a consideration of either field

However applying the same knowledge to the reals yields the opposite conclusion, by virtue of the completeness axiom.

Theorem 6.7 ($\sqrt{2}$ is a Real Number). *There exists a positive real number which squares to* 2.

Proof. Let $S = \{r \in \mathbb{R} \mid r^2 < 2\}$. Then, S is nonempty, as $0 \in S$ since $O^2 = 0$ and $0 < 2$. Next, we show that S is bounded above by 10:

Let $r \in S$ is arbitrary. Without loss of generality we may assume $r > 0$ as if $r < 0$ then certainly $r < 10$. By the definition of S, we know $r^2 < 2$ and thus clearly $r^2 < 100$. But recall Proposition 4.5: for positive a, b if $a^2 < b^2$ then $a < b$, so from $r^2 < 100$ we may conclude $r < 10$.

Knowing that S is both nonempty and bounded above, the completeness axiom applies to furnish us with a least upper bound $\sigma = \sup S$. And knowing its existence, Theorem 6.6 immediately implies that $\sigma^2 = 2$, so σ is by definition a square root of 2. □

Theorem 6.8 (The Rationals are Incomplete). *Within the field of rational numbers, the set* $S = \{r \in \mathbb{Q} \mid r^2 < 2\}$ *is bounded above and nonempty, but does not have a supremum.*

Proof. The argument that S is nonempty and bounded above is identical to that in Theorem 6.7. And, Theorem 6.6 implies that *if* the supremum exists it must square to 2. But we know by Theorem 1.1 that there is no such rational number. Thus, the supremum must not exist, and so $\mathbb{Q}$ fails the completeness axiom. □

There is nothing special about 2 in the above argument, other than it is easy for us to work with. We could stop right now to prove the more general statement that all square roots exist:

Theorem 6.9 (Square Roots Exist). *If* $x \in \mathbb{R}$ *is positive, then* $\sqrt{x}$ *is a real number.*

Though to not be *too* repetitive, we will hold off and prove this a different way, to illustrate more powerful tools in CITE.

Exercise 6.5. Prove that the product of a nonzero rational and an irrational number is irrational.

Exercise 6.6. The sum of two irrational numbers need not be irrational, as the example $\sqrt{2} - \sqrt{2} = 0$ shows. Prove or disprove: the sum of two *positive* irrational numbers is irrational.

6.2.2. Density

Definition 6.7 (Density). Let S be a subset of an ordered field $\mathbb{F}$. Then S is *dense* in $\mathbb{F}$ if between any two elements $a, b \in \mathbb{F}$ with $a < b$ there is some $s \in S$ with

$$a < s < b$$

Theorem 6.10 (Density of the Rationals). *The rational numbers are dense in the real numbers.*

Proof. We need to start with two arbitrary real numbers $a < b$, and find a rational number r between them. Let's do some scratch work: if $r = m/n$ and we want $a < m/n < b$ then it suffices to find an integer m between na and nb. This sounds doable!

Precisely, since $b - a > 0$, we can use the archimedean property to find some $n \in \mathbb{N}$ with $n(b - a) > 1$. Now since $nb - na > 1$, we just need to prove there's an integer m between them, and this'll be the number we want!

To rigorously prove this m exists, we can reason as follows: we know there are integers greater than na (since $\mathbb{R}$ has no infinite elements), so let m be the smallest such. Then by definition $m > na$, so all we need to show is $m < nb$. Since m is the smallest integer greater than na, we know $m - 1 < na$, or $m < na + 1$. But $na + 1 < nb$ so $m < nb$ as required.

Now we have a natural number n and an integer m with $na < m < nb$. Dividing through by n gives

$$a < \frac{m}{n} < b$$

□

As we have gotten used to being very careful in our arguments, you may think while working out the above argument to fill in a little lemma showing that every set of integers bounded below has a *minimum.* And, you could indeed do so by induction (try it - but fair warning, the argument is a little tricky! It's easiest with "strong induction" - what are we inducting over?). However this fact is actually *logically equivalent* to the principle of induction, and in foundations of arithmetic things are often reversed: we take this as an axiom, and prove induction from it! The statement is called the *well ordering principle.*

Definition 6.8 (The Well Ordering Principle). Every nonempty subset of $\mathbb{N}$ has a least element.

Exercise 6.7 (Density of the Irrationals). Use Theorem 6.10 above to prove that the irrationals are also dense in the reals.

Exercise 6.8. The *dyadic* rationals are the subset of $\mathbb{Q}$ which have denominators that are a power of 2 when written in lowest terms.

Prove the dyadic rationals are dense in $\mathbb{R}$.

6.3. Uncountability

We can use this to prove the uncountability of the reals using Cantor's original argument. (We will give the better known Cantor diagonalization argument later, once we've introduced decimals)

Theorem 6.11 ($\mathbb{R}$ is Uncountable)**.** *There is no bijection between* $\mathbb{N}$ *and* $\mathbb{R}$

Proof. Let $f : \mathbb{N} \to [0,1]$ be any function whatsoever. We can use this function to produce a sequence of points as follows:

$$f(1) = x_1, f(2) = x_2, f(3) = x_3 \dots$$

From this we can construct a set of nested intervals.
Let $I_1 \subset [0,1]$ be any closed interval that doesn't contain x_1. Then let $I_2 \subset I_1$ be a closed interval which does not contain x_2 (if x_2 was outside I_1, you could just take I_1 again, otherwise if its inside I_1 just take an interval on one side or the other of it). Continuing, we can easily choose an interval $I_{n+1} \subset I_n$ which doesn't contain x_{n+1}.

This gives us an infinite sequence of closed nested intervals inside a complete ordered field, so Theorem 5.2 tells us that their intersection must be nonempty. That is, there is some point $y \in [0,1]$ where $y \in I_n$ for all n.

What does this mean? Well, since $y \in I_1$ we know $y \neq x_1$ since I_1 was purpose-built to exclude x_1. Similarly $y \in I_2$ guarantees $y \neq x_2$, and so on...$y \in I_n$ means $y \neq x_n$. Thus, y is some point in $[0,1]$ which is not in our list!

Since $y \neq f(n)$ for any n, we see that our original (arbitrary) function cannot have been surjective. And, since bijections are both injective and surjective, this proves there *is no bijection* from $\mathbb{N}$ to $[0,1]$, so $[0,1]$ is uncountable! Then, as $[0,1] \subset \mathbb{R}$ we see $\mathbb{R}$ is uncountable as well. □

This has some pretty wild corollaries if you have studied countable sets before. Here's a couple examples

Corollary 6.1 (Transcendental Numbers)**.** *There exist real numbers which are not the solution of any algebraic equation with rational coefficients.*

Corollary 6.2 (Uncomputable Numbers)**.** *There exist real numbers which cannot be computed by any computer program.*

These are additional motivation for why we really need a precise theory of the real numbers: with very little work we've already *proven* that there is no way to study this number system with algebra alone - or even with the most powerful computer you could imagine.

6.4. ⋆ Topology

One final basic property of $\mathbb{R}$ that we will show follows from completeness is that its "connected" - it really does form a continuous line.

Definition 6.9 (Connected)**.** Let S be a subset of a topological space. Then a *separation* of S is a pair of disjoint open sets U, V whose union is S.

A subset is called *disconnected* if there is a separation, and *connected* if there is no way to make a separation.

Example 6.2 (A disconnected set)**.** Let $S = \{x \in \mathbb{R} \mid x > 0, x < 2, \text{ and } x \neq 1\}$. Then S is disconnected as we can write

$$S = (0,1) \cup (1,2)$$

And note these two intervals are both open, and dont share any points in common (so they are disjoint).

It's harder to imagine doing this for the interval $(0,2)$ however: if you try to imagine cutting it into two disjoint intervals at some point x, you're going to end up with $(0,x) \cup [x,2)$ or $(0,x] \cup (x,2)$. In either case, these intervals are not both open! To make them both open you could try $(0,x) \cup (x,2)$ but now they miss the point x (so their union isnt the whole space) or $(0, x+0.01) \cup (x-0.01, 2)$ but now they overlap and aren't disjoint. Intuitively there's no way to do it - the interval $(0,2)$ is connected!

Theorem 6.12 (The Real Line is Connected)**.**

Proof. Assume for the sake of contradiction that $U \cup V$ is a separation of $\mathbb{R}$ (so, U, V are nonempty open sets and every point of $\mathbb{R}$ is in exactly one of them).

Choose some $x \in U$ and $y \in V$ - we can do this because they're nonempty - and without loss of generality assume that $x < y$. Considering the interval $[x,y]$ we know the left side is in U and the right in V, so we can define the

$$Z = \{z \in [x,y] \mid [x,z] \subset U\}$$

This set is nonempty (as $x \in Z$) and its bounded above (by y), so by completeness it has some supremum $\zeta = \sup Z$. Now the question is, which set is ζ in, U or V?

If $\zeta \in V$ then we know that since V is open ther's some small interval $(\zeta - \epsilon, \zeta + \epsilon)$ fully contained in V. But this means there's a number *smaller* than ζ contained in V, which means the interval $[0, \zeta]$ isnt fully contained in U, a contradiction!

If $\zeta \in U$ then we know since U is open, that there must be some tiny open interval $(\zeta - \epsilon, \zeta + \epsilon)$ around ζ contained U. This means there's a number *larger than ζ (for example, $\zeta + \epsilon/2$) where $[x, \zeta + \epsilon/2]$ is contained in U. So, ζ can't even be an upper bound to the set of all such numbers, a contradiction!

Both cases lead to contradiction, so there must be no such ζ, and hence no such separation. ☐

Exercise 6.9 (Open Intervals of ℝ are connected)**.** Prove that every open interval $(a, b) \subset \mathbb{R}$ is connected, mimicking the proof style above.

This *fails* for the rational numbers - they are not connected!

Theorem 6.13 (The Rationals are Not Connected)**.** *Consider the following two subsets of the rational numbers:*

$$A = \{x > 0 \mid x^2 > 2\}$$

$$B = \{x \in \mathbb{Q} \mid x \notin A\}$$

Then A and B form a separation of $\mathbb{Q}$.

Proof. A and B are open intervals in $\mathbb{Q}$ (they're the rational points of the open intervals $(\sqrt{2}, \infty)$ and $(-\infty, \sqrt{2})$). By definition every point of $\mathbb{Q}$ is in either A or B and they're disjoint. Since we just showed they are open, they form a separation, so $\mathbb{Q}$ is disconnected. ☐

In fact, $\mathbb{Q}$ is *extremely disconnected* - this same argument applies at every irrational number of ℝ.

7. ⋆ Infinity

Definition 7.1. The symbol ∞ is a *formal symbol*: that is, a symbol that we agree to write, but do not attach any specific *value* to.

By default, any expression involving the symbol ∞ is considered *undefined.* We will use define certain contexts where the symbol ∞ is meaningful below.

7.1. Order

Our first use of the symbol ∞ is to expand *interval notation* of the real numbers. Right now, using the order $<$ we have rigorously defined intervals such as (a, b), $[a, b)$ and $[a, b]$ for $a, b \in \mathbb{R}$.

Definition 7.2. For any real number a, we define the following intervals with $\pm\infty$ as an endpoint:

$$(-\infty, a) = \{x \in \mathbb{R} \mid x < a\}$$
$$(-\infty, a] = \{x \in \mathbb{R} \mid x \leq a\}$$
$$(a, \infty) = \{x \in \mathbb{R} \mid x > a\}$$
$$[a, \infty) = \{x \in \mathbb{R} \mid x \geq a\}$$

But we can take this farther, by actually adding the formal symbols $\pm\infty$ to our number system, to create a set called the *extended reals.*

Definition 7.3 (The Extended Reals)**.** The *extended real number line* is the set

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}.$$

Definition 7.4 (Ordering on $\overline{\mathbb{R}}$)**.** The order $<$ on $\mathbb{R}$ can be extended to $\overline{\mathbb{R}}$ by the following two rules:

$$\forall x \in \mathbb{R},\ x < \infty \qquad \forall x \in \mathbb{R},\ -\infty < x$$

This allows for interval notation on $\overline{\mathbb{R}}$ where, we may may write intervals such as $[-\infty 1]$ to mean the points $\{x \mid \overline{R} \mid x \leq 1\}$ etc.

In $\overline{\mathbb{R}}$ then, ∞ is an upper bound for every set, and $-\infty$ is a lower bound for every set. On the real numbers alone, the completeness axiom tells us that the supremum of *bounded* nonempty sets exist, but unbounded sets do not have a supremum. In the extended reals, we see that $\pm\infty$ naturally satisfy the definitions of

Proposition 7.1 (Unbounded Above means $\sup = \infty$). *Let A be a nonempty subset of $\mathbb{R}$ which is not bounded above. Then as a subset of of the extended reals,* $\sup A = \infty$.

Proof. By the definition of ∞, we see that ∞ is an upper bound for A always, so we need only show it is the supremum. Let $x \in \overline{\mathbb{R}}$ be any element less than ∞. Then x must be an element of $\mathbb{R}$, and since A is not bounded above in $\mathbb{R}$, there is some $a \in A$ with $a > x$. Thus x is not an upper bound, and so every element less than ∞ fails to be an upper bound: that is, ∞ is the *least upper bound* as claimed. □

Exercise 7.1 (Unbounded Below means $\inf = -\infty$).

Corollary 7.1 (Sup and Inf in the Extended Reals). *Every nonempty subset of the extended real line has both an infimum and a supremum.*

Proof. Let A be a nonempty subset of $\overline{\mathbb{R}}$. First, if A contains ∞, then $\sup A = \infty$ as it is the maximum. So, we can consider the case that $\infty \notin A$. If A is bounded above by a real number, then $\sup A$ is also a real number by completeness, and if A is not bounded above, then $\sup A = \infty$ by Proposition 7.1.

The same logic applies to lower bounds: after taking care of the case where $\inf A = \min A = -\infty$, if A is bounded below completeness furnishes a real infimum, and if it is not, Exercise 7.1 shows the infimum to be $-\infty$. □

In the extended reals, it is still common to take the infimum and supremum of the empty set to be undefined. But there is also another option: one can assign $\inf \emptyset = \infty$ and $\sup \emptyset = -\infty$: if we do this then every set in the extended reals has an infimum and supremum!

7.2. Arithmetic

We know from the previous chapter that complete ordered fields cannot contain infinite numbers, yet in the section above we added $\pm\infty$ to $\mathbb{R}$ in a way that did not mess up the order, or completeness properties. So, the addition of this new symbol must cause trouble with the field axioms. And, indeed it does!

Its extremely important to remember that $\overline{\mathbb{R}}$ is **not** a field. We have not extended any of the operations to the formal symbol ∞, so things like $\infty + 1$ or $\infty - \infty$ or $3\infty + \infty/2$ are currently undefined.

Part III.

Sequences

8. Definitions

Highlights of this Chapter: we define the notion of converges, and discuss examples of how to prove a sequence converges directly from this definition.

Definition 8.1 (Sequence)**.** A sequence is an infinite ordered list of numbers

$$(s_1, s_2, s_3, \dots, s_n, \dots)$$

Each individual element is a *term* of the sequence, with an subscript (the *index*) denoting its position in the list.

Most often, we take the set of indices to be $1, 2, 3, \dots$, but any infinite subset of the integers will do. For example, the sequence p_n of perimeters of inscribed n-gons starts with index 3 (the triangle), as this is the smallest polygon. And, the subsequence Archimedes used to calculate π started with the hexagon and then iterated doubling: $P_6, P_{12}, P_{24}, \dots$ so has index set

$$\{6, 12, 24, 48, 96, 192, 384, \dots\}$$

Formally, we note all of this is captured using functions, though we will not need this perspective during our day-to-day usage of sequences.

Remark 8.1. Let $I \subset \mathbb{Z}$ be any infinite set of indices. Then a sequence is a *function* $s\colon I \to \mathbb{R}$.

While sequence itself is just an infinite ordered list of numbers, to work with such an object we often require a way to compute its terms. Sometimes this is hard! For example, the sequence

$$\pi_n = \text{ the number of prime numbers} \leq n$$

Is called the *prime counting function*, and being able to compute its exact values efficiently would be monumental progress in number theory. In practice, sequences that we *can* compute with efficiently are often presented to us in one of two ways:

- **Closed Formula** For each n, we are given some formula of the type familiar from high school mathematics, and plugging n into this formula yields the n^{th} term of the sequence. Some examples are

$$a_n = \frac{n^2+1}{3n-2}, \qquad b_n = \sin\left(\frac{1}{n}\right) \qquad c_n = \sqrt{1+\frac{\sqrt{n}}{n+1}}$$

- **Recursive Definition** For each n, we are not given a formula to compute s_n directly, but rather we are given a formula to compute it from the previous value s_{n-1}.

Here's some example sequences that are important both to us, and the history of analysis:

Example 8.1 (Babylonians and $\sqrt{2}$)**.** Starting from rectangle of width and height w, h, the Babylonians created a new rectangle whose width was the average of these, and whose height was whatever is required to keep the area 2:

$$w_{\text{new}} = \frac{w+h}{2} \qquad h_{\text{new}} = \frac{2}{w_{\text{new}}}$$

This because we can solve for h in terms of w, this induces a recursive sequence for the widths. Starting from some (w_n, h_n) we have

$$w_{n+1} = \frac{w_n+h_n}{2} = \frac{w_n+\frac{2}{w_n}}{2} = \frac{w_n}{2}+\frac{1}{w_n}$$

Thus, in modern terminology the babylonian procedure defines a recursive sequence, given any starting rectangle. If we begin with the rectangle of wdith 2 and height 1, we get

$$w_0 = 2, \qquad w_{n+1} = \frac{w_n}{2}+\frac{1}{w_n}$$

Exercise 8.1 (Babylonians and $\sqrt{2}$)**.** Following the same type of reasoning as for width, use the babylonian procedure to produce a recursive formula for the sequence of heights h_n, for a rectangle starting with $h = 1$.

Example 8.2. An *infinite sum* is a type of recursively defined sequence, built from another sequence called its *terms*. Assume that a_n is any sequence. Then we build a sequence s_n by

$$s_0 = a_0 \qquad s_{n+1} = s_{n-1} + a_n$$

Unpacking this, we see that $s_1 = s_0 + a_1 = a_0 + a_1$, and thus $s_2 = s_1 + a_2 = a_0 + a_1 + a_2$ etc.

Exercise 8.2. Come up with a recursive sequence that could be used to formally understand the infinite expression below:

$$\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}}}}$$

8.1. Convergence

The reason to define a sequence precisely is that we are interested in making rigorous the idea of *infinitely many steps*, the way the Babylonians may have pictured running their procedure an infinite number of times to produce a perfect square, or Archimedes who ran his side-doubling procedure infinitely many times to produce a circle.

In both cases, there was some number L *out there at infinity* that they were probing with a sequence. We call such a number L the *limit of the sequence*.

Definition 8.2 (Convergent Sequence)**.** A sequence s_n *converges* to a limit L if for all $\epsilon > 0$ there is some threshold N past which every further term of the sequence is within ϵ of L. Formally, this is the logic expression

$$\forall \epsilon > 0\, \exists N\, \forall n > N\;\; |s_n - L| < \epsilon$$

When a sequene converges to L we write

$$\lim s_n = L \qquad \text{or} \qquad s_n \to L$$

A sequence is *divergent* if its not convergent. The definition of convergence formalizes the idea the ancients sought **if you keep calculating terms, you'll get as close as you like to the number you seek**

That is, the definition sets up a challenge between you (the computer of the sequence) and the error tolerance. Once you set a certain amount of acceptable error ϵ, the definition furnishes an N and guarantees that if you compute the sequence out until N you'll be within the tolerated error - and if you keep computing more terms, the approximation will never get worse. Its good to look at some specific examples, while getting comfortable with this:

Exercise 8.3 (Understanding Convergence)**.** Consider the sequence $a_n = \frac{1}{n^2+13}$. Feel free to use a calculator (even just the google search bar) to experiment and answer these questions.

- What value L do you think this sequence converges to?
- If $\epsilon = 1/10$, what value of N ensures that a_n is always within ϵ of L for, $n > N$?
- If $\epsilon = 1/100$, what value of N ensures that a_n is always within ϵ of L for, $n > N$?

Exercise 8.4 (Convergence and $\sqrt{2}$)**.** This problem concerns the babylonian sequence for $\sqrt{2}$ in Example 8.1. Again, use a calculator to play around and answer the following

- For which value of N are we guaranteed that w_n calculates the first **two** decimal places $\sqrt{2}$ correctly, when $n > N$?
- For which value of N are we guaranteed that w_n calculates the first **eight** decimal places $\sqrt{2}$ correctly, when $n > N$?

To prove a sequence converges, we need to work through the string of quantifiers $\forall\epsilon\exists N\forall n...$ This sets up a sort of imagined *battle* between an imagined foe setting a value of ϵ, and you needing to come up with an N such that you can get the sequene within ϵ of the limit.

Here's one incredibly useful example, that will serve as the basis of many future calculations.

Proposition 8.1 ($1/n$ converges to 0.)**.** *Prove that the sequence $s_n = 1/n$ of reciprocals of the natural numbers converges to* 0.

Proof. Let $\epsilon > 0$. Then set $N = 1/\epsilon$, and choose arbitrary $n > N$. Since $n > 1/\epsilon$ it follows that $1/n < \epsilon$, and hence that

$$\left|\frac{1}{n} - 0\right| < \epsilon$$

Since $n > N$ was arbitrary, this holds for all such n, and we have proved for this ϵ, theres an N with $n > N$ implying the sequence $1/n$ is within ϵ of the proposed limit 0. Since ϵ was also arbitrary, we have in fact proved this for *all positive epsilon*, and thus we conclude

$$\frac{1}{n} \to 0$$

□

Exercise 8.5 ($\frac{n}{n+1}$ converges to 1)**.** Prove this directly from the limit definition.

Exercise 8.6 ($\frac{n}{n^2+1}$ converges to 0.)**.** Prove this directly from the limit definition.

Example 8.3 ($\frac{1}{2^n} \to 0$)**.** Here's a sketch of an argument: you should fill in the details. Let $\epsilon > 0$. Then we want to find an N where $n > N$ implies $1/2^n < \epsilon$. First, we prove by induction that $2^n \geq n$ for all n. Thus, $1/2^n < 1/n$, and so it suffices to find N where $1/n < \epsilon$. But this is exactly what we did above in the proof that $1/n \to 0$. So this is possible, and hence $1/2^n \to 0$.

Exercise 8.7. Give an example of the following, or explain why no such example can exist.

- A sequence with infinitely many terms equal to zero, but does not converge to zero.
- A sequence with infinitely many terms equal to zero, which converges to a nonzero number.
- A sequence of irrational numbers that converges to a rational number.
- A convergent sequence where every term is an integer.

Exercise 8.8. Prove, directly from the definition of convergence, that

$$\frac{2n-2}{5n+1} \to \frac{2}{5}$$

8.1.1. Some Useful Limits

We will soon develop several theorems that let us calculate many limits without tediously chasing down an N for every ϵ. But there are still several 'basic limits' that we will need to know, that will prove useful as building blocks of more complicated limits, as well as foundations to further theory in analysis. We compute several of them here: you should not worry too hard about committing these to memory; but rather read the proofs as examples of how to play the $\epsilon - N$ game in tricky situations.

This first is useful in developing the theory of geometric series: :::{#exm-an-to-o} Let $|a| < 1$, then the sequence a^n of repeated powers of a converges to 0. ::: :::{.proof} :::

This next is an essential building block of the theory of exponential functions: :::{#exm-nth-root-n} The sequence $n^{1/n}$ converges to 1. ::: :::{.proof} :::

8.2. Divergence

The definition of convergence picks out a very nice class of sequences: those that *get arbitrarily close to a fixed value, as their index grows.* The rest of sequences - anything that does not have this nice property, are all lumped into the category of *divergent.*

Definition 8.3 (Divergence)**.** A sequence diverges if its *not true* that for any ϵ you can find an N where beyond that, all terms of the sequence differ from some constant (the limit) less than ϵ.

Phrasing this positively: a sequence a_n diverges if for *every value of* a, there exists some $\epsilon > 0$ where no matter which N you pick, there's always some $n > N$ where $|a_n - a| > \epsilon$. There's a lot of quantifiers here! Written out in first order logic:

$$\forall a \in \mathbb{R}\ \exists \epsilon > 0\ \forall N\ \exists n > N \quad |a_n - a| > \epsilon$$

Again, its easiest to illustrate with an example:

Example 8.4 (($(-1)^n$ Diverges). **(Example 3.18, on page 78 of text)**

8.2.1. ⋆ Divergence to ∞

Definition 8.4 (Diverging to $\pm\infty$). A sequence s_n diverges to ∞ if for all $M > 0$ there exists an threshold past which the sequence is always larger than M. As a logic statement,

$$\forall M > 0\, \exists N\, \forall n > N\; s_n > M$$

Exercise 8.9 (n^2 diverges to ∞.).

Proposition 8.2. *If $a > 1$ then a^n diverges to infinity.*

While we are not often interested in sequences going to infinity themselves, they are a useful tool to prove that other sequences diverge. But they also provide a tool to help prove certain sequences *converge*

Proposition 8.3. *A sequence s_n of positive terms converges to 0 if and only if its reciprocals $1/s_n$ diverge to ∞.*

Proof.

□

Exercise 8.10.

- Give an example of two divergent sequences a_n, b_n where $a_n + b_n$ is convergent.
- Give an example of two divergent sequences a_n, b_n where a_nb_n is convergent.

8.3. Uniqueness

Theorem 8.1 (Limits are unique). *Let a_n be a convergent sequence. Then there exists a unique $a \in \mathbb{R}$ with $a_n \to a$.*

Proof. **Proposition 3.19 in the book** □

The proof of this theorem shows a couple important points that will occur time and again throughout analysis:

- We proved uniqueness by showing that if x and y were both limits, then $x = y$.
- We proved $x = y$ by showing that for every $\epsilon > 0$ the difference $|x - y| < \epsilon$.
- We proved $|x - y| < \epsilon$ by an $\epsilon/2$ argument:

- We added zero in a clever way: $|x-y| = |x-a_n+a_n-y|$
- We used the triangle inequality $|x-a_n+a_n-y| \leq |x-a_n| + |a_n-y|$
- We used the fact that $a_n \to x$ and $a_n \to y$ to make each of $|a_n - x|$ and $|a_n - y|$ less than $\epsilon/2$.

There's one more *uniqueness-type* theorem about limits that's useful to get a handle on. We just saw that the limit is uniquely determined by the sequence, but we can say something slightly stronger. Its uniquely determined by the *end of the sequence*: if you throw away the first finitely many terms, it won't change the limit.

Definition 8.5. A *shifted sequence* the result of shifting the indices by a constant k, deleting the first k terms. Precisely, given a sequence a_n and some $k \in \mathbb{N}$, the sequence $s_n = a_{n+k}$ is the result of shifting a by k.

$$s_0 = a_k, s_1 = a_{k+1}, s_2 = a_{k+2}, \ldots$$

Proposition 8.4. *Shifting a convergent sequence does not change its limit.*

Scratch Work. Assume that a_n converges to a, and define the sequence s_n by deleting the first k terms of a_n, that is, $s_n = a_{n+k}$. We claim that $s_n \to a$.

Let $\epsilon > 0$ and choose an N such that if $n > N$ we know that $|a_n - a| < \epsilon$ (we know such an N exists by the assumption $a_n \to a$). Now consider $|s_n - a|$. Since $s_n = a_{n+k}$, we know $|s_n - a| < \epsilon$ because we already knew $|a_{n+k} - a| < \epsilon$: we knew this for *every single index bigger than* N.

Thus, for all $n > N$ we have $|s_n - a| < \epsilon$, which is the definition of $s_n \to a$. □

This can be generalized, to show that any two sequences which are eventually the same have the same limit. Since the first finite part of any sequence is irrelevant to its limiting behavior, its nice to have a word for "the rest of the sequence, after throwing away an unspecified amount at the beginning". This is called the *tail*.

Definition 8.6 (Tail of a Sequence)**.** The *tail of a sequence* is what remains after chopping off an arbitrary (finite) number of terms from the beginning of the sequence. Two sequences *have the same tail* if they agree after some point: more precisely, a_n and b_n have the same tail if there is an N_a and N_b such that for all $k \in \mathbb{N}$

$$a_{N_a+k} = b_{N_b+k}$$

Example 8.5 (Tail of a Sequence)**.** The following two sequences have the same tail:

$$a_n = 1, 1, 4, 3, 1, 5, 1, 3, 1, 4, 7, 8, 9, 10, 11, 12, 13, 14, \ldots$$

$$b_n = -4, 3, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 \ldots$$

We can see this because $a_{13} = b_3 = 9$, and $a_{14} = b_4 = 10$, and $a_{15} = b_5 = 11$...for every k we have that $a_{13+k} = b_{3+k}$ so they agree after chopping the first 12 terms off of a_n and the first two terms off of b_n.

Exercise 8.11 (Convergence only depends on the tail)**.** If two sequences have the same tail, then they either both converge or both diverge, and if they converge, they have the same limit.

8.4. ⋆ Topology

With an eye to topology, everything about sequences and convergence can be rephrased in terms of open sets, instead of with talk about ϵ and inequalities.

Definition 8.7 (Neighborhoods)**.** A *neighborhood* of a point x is any open set U containing x. The ϵ-neighborhood of x is the neighborhood $U_\epsilon = (x - \epsilon, x + \epsilon)$

Definition 8.8 (Convergence and ϵ-Neighborhoods)**.** A sequence a_n converges to a if every ϵ neighborhood contains all but finitely many terms of the sequence.

That this is equivalent to Definition 8.2, because the definition of epsilon neighborhood exactly captures the interval discussed in the original definition of convergence.

Exercise 8.12 (Convergence and ϵ-Neighborhoods)**.** The definition of convergence in terms of epsilon neighborhoods is equivalent to the usual definition in terms of absolute values and inequalities.

The definition of an epsilon neighborhood makes sense only somewhere like the real line, where we can talk about intervals. So, the general topological definition must dispense with this notion and talk just about open sets:

Definition 8.9. A sequence a_n converges to a if every neighborhood contains all but finitely many terms of the sequence.

Exercise 8.13 (Convergence and Neighborhoods)**.** Prove this is equivalent to convergence using ϵ neighborhoods. *Hint: show that every neighborhood contains some epsilon neighborhood. Can you show that is enough?*

9. Calculation

Highlights of this Chapter: We develop techniques for bounding limits by inequalities, and computing limits using the field axioms. We use these techniques to prove two interesting results:

- The Babylonian sequence approximating $\sqrt{2}$ truly does converge to this value.
- Given any real number, there exists a sequence of rational numbers converging to it.

Now that we have a handle on the definition of convergence and divergence, our goal is to develop techniques to avoid using the definition directly, wherever possible (finding values of N for an arbitrary ϵ is difficult, and not very enlightening!)

The natural first set of questions to investigate then are how our new definition interacts with the ordered field axioms: can we learn anything about limits and inequalities, or limits and field operations? We tackle both of these in turn below.

9.1. Limits and Inequalities

Proposition 9.1 (Limits of nonnegative sequences)**.** *Let a_n be a convergent sequence of nonnegative numbers. Then* $\lim a_n$ *is nonnegative.*

Proof. Assume for the sake of contradiction that $a_n \to L$ but $L < 0$. Since L is negative, we can find a small enough epsilon (say, $\epsilon = |L|/2$) such that the entire interval $(L-\epsilon, L+\epsilon)$ consists of negative numbers.

The definition of convergence says for this ϵ, there must be an N where for all $n > N$ we know a_n lies in this interval. Thus, we've concluded that for large enough n, that a_n *must be negative*! This is a contradiction, as a_n is a nonnegative sequence. □

Exercise 9.1. If a_n is a convergent $a_n \geq L$ for all n, then $\lim a_n \geq L$. Similarly prove if a_n is a convergent $a_n \leq U$ for all n, then $\lim a_n \leq U$.

This exercise provides the following useful corollary, telling you that if you can bound a sequence, you can bound its limit.

Corollary 9.1 (Inequalities and Convergence). *If a_n is a convergent sequence with $L \leq a_n \leq U$ for all n, then*

$$L \leq \lim a_n \leq U$$

In fact, a kind of converse of this is true as well: if a sequence converges, then we know the limit 'is bounded' (as it exists, as a real number, and those can't be infinite). But this is enough to conclude that the entire sequence is bounded!

Proposition 9.2 (Convergent Sequences are Bounded). *Let s_n be a convergent sequence. Then there exists a B such that $|s_n| < B$ for all $n \in \mathbb{N}$.*

Proof. Let $s_n \to L$ be a convergent sequence. Then we know for any $\epsilon > 0$ eventually the sequence stays within ϵ of L. So for example, choosing $\epsilon = 1$, this means there is some N where for $n > N$ we are assured $|s_n - L| < 1$, or equivalently $-1 < s_n - L < 1$. Adding L,

$$L - 1 < s_n < L + 1$$

Thus, we have both upper and lower bounds for the sequence after N and all we are left to worry about is the finitely many terms before this. For an upper bound on these we can just take the max of $s_1, \ldots, s_N$ and for a lower bound we can take the min.

Thus, to get an overall upper bound, we can take

$$M = \max\{s_1, s_2, \ldots, s_N, L + 1\}$$

and for an overall lower bound we can take

$$m = \min\{s_1, s_2, \ldots, s_N, L - 1\}$$

Then for all n we have $m \leq s_n \leq M$ so the sequence s_n is bounded. □

Theorem 9.1 (The Squeeze Theorem). *Let a_n, b_n and c_n be sequences with $a_n \leq b_n \leq c_n$ for all n. Then if a_n and c_n are convergent, with $\lim a_n = \lim c_n = L$, then b_n is also convergent, and*

$$\lim b_n = L$$

Proof. **Theorem 3.23 on page 87 of the textbook** □

9.1.1. Example Computations

The squeeze theorem is *incredibly useful* in practice as it allows us to prove the convergence of complicated looking sequences by replacing them with two (hopefully simpler) sequences, an upper and lower bound. To illustrate, let's look back at Exercise 8.6, and re-prove its convergence.

Example 9.1 ($\frac{n}{n^2+1}$ converges to 0.)**.** Since we are trying to converge to zero, we want to bound this sequence above and below by sequences that converge to zero. Since n is always positive, a natural lower bound is the constant sequence $0, 0, 0, \ldots$.

One first thought for an upper bound may be $\frac{n}{n+1}$: its easy to prove that $\frac{n}{n^2+1} < \frac{n}{n+1}$ (as we've made the denominator smaller), and so we have bounded our sequence $0 < a_n < \frac{n}{n+1}$. Unfortunately this does not help us, as $\lim \frac{n}{n+1} = 1$ (Exercise 8.5) so the two bounds do not squeeze a_n to zero!

Another attempt at an upper bound may be $1/n$: we know this goes to zero (Proposition 8.1) and it is also an upper bound:

$$\frac{n}{n^2+1} < \frac{n}{n^2} = \frac{1}{n}$$

Thus since $\lim 0 = 0$ and $\lim \frac{1}{n} = 0$, we can conclude via squeezing that $\lim \frac{n}{n^2+1} = 0$ as well.

This theorem is particularly useful for calculating limits involving functions whose values are difficult to compute. While we haven't formally introduced the sine function yet in this class, we know (and will later confirm) that $-1 \leq \sin(x) \leq 1$ for all $x \in \mathbb{R}$. We can use this to compute many otherwise difficult limits:

Example 9.2 ($s_n = \frac{\sin n}{n}$ converges to 0.)**.** Since $-1 \leq \sin(x) \leq 1$ we know $0 \leq |\sin x| \leq 1$ for all x, and thus

$$0 \leq \frac{\sin n}{n} \leq \frac{1}{n}$$

Since both of these bounding sequences converge to zero, we know the original does as well, by the squeeze theorem.

This sort of estimation can be applied to even quite complicated looking limits:

Example 9.3. Compute the following limit:

$$\lim \left(\frac{n^2 \sin(n^3 - 2n + 1)}{n^3 + n^2 + n + 1} \right)^n$$

Lets begin by estimating as much as we can: we know $|\sin(x)| \le 1$, so we can see that

$$\left|\frac{n^2\sin(n^3-2n+1)}{n^3+n^2+n+1}\right| < \frac{n^2}{n^3+n^2+1}$$

Next, we see that by shrinking the denominator we can produce yet another over estimate:

$$\frac{n^2}{n^3+n^2+1} < \frac{n^2}{n^3} = \frac{1}{n}$$

Bringing back the n^{th} power

$$\left|\frac{n^2\sin(n^3-2n+1)}{n^3+n^2+n+1}\right|^n < \frac{1}{n^n}$$

And, unpacking the definition of absolute value:

$$-\frac{1}{n^n} < \left(\frac{n^2\sin(n^3-2n+1)}{n^3+n^2+n+1}\right)^n < \frac{1}{n^n}$$

It now suffices to prove that $1/n^n$ converges to zero, as we ve squeezed our sequence with it. But this is easiest to do with *another squeeze*: namely, since $n^n > 2^n$ we see $0 < 1/n^n < 1/2^n$, and we already proved that $1/2^n \to 0$, so we're done!

$$\lim\left(\frac{n^2\sin(n^3-2n+1)}{n^3+n^2+n+1}\right)^n = 0$$

Exercise 9.2. Use the squeeze theorem to prove that

$$\lim\left(\frac{n^3-2-\frac{1}{n^3}}{3n^3+5}\right)^{2n+7} = 0$$

A nice corollary of the squeeze theorem tells us when a sequence converges by estimating its difference from the proposed limit:

Exercise 9.3. Let a_n be a sequence, and L be a real number. If there exists a sequence α_n where $|a_n - L| \le \alpha_n$ for all n, and $\alpha_n \to 0$, then $\lim a_n = L$.

This is useful as unpacking the definition of absolute value (Definition 4.5), a sequence α_n with

$$-\alpha_n \leq a_n - L \leq \alpha_n$$

can be thought of as giving "error bounds" on the difference of a_n from L. In this language, the proposition says if we can bound the error between a_n and L by a sequence going to zero, then a_n must actually go to L.

9.2. Limits and Field Operations

Just like inequalities, the field operations themselves play nicely with limits.

Theorem 9.2 (Constant Multiples)**.** *Let s_n be a convergent sequence, and k a real number. Then the sequence ks_n is convergent, and*

$$\lim ks_n = k \lim s_n$$

Proof. We distinguish two cases, depending on k. If $k = 0$, then ks_n is just the constant sequence $0, 0, 0 \ldots$ and $k \lim s_n = 0$ as well, so the theorem is true.

If $k \neq 0$, we proceed as follows. Denote the limit of s_n by L, and let $\epsilon > 0$. Choose N such that $n > N$ implies $|s_n - L| < \frac{\epsilon}{|k|}$ (we can do so, as $s_n \to L$). Now, for this same value of N, choose arbitrary $n > N$ and consider the difference $|ks_n - kL|$:

$$|ks_n - kL| = |k(s_n - L)| = |k||s_n - L| < |k|\frac{\epsilon}{|k|} = \epsilon$$

Thus, $ks_n \to kL$ as claimed! □

To do a similar calculation for the sum of sequences requires an $\epsilon/2$ type argument:

Theorem 9.3 (Limit of a Sum)**.** *Let s_n, t_n be convergent sequences. Then the sequence of term-wise sums $s_n + t_n$ is convergent, with*

$$\lim(s_n + t_n) = \lim s_n + \lim t_n$$

Exercise 9.4 (Limit of Sums and Differences)**.** Prove Theorem 9.3, that if s_n and t_n converge so does $s_n + t_n$ and

$$\lim(s_n + t_n) = \lim s_n + \lim t_n$$

Use this together with other limit theorems to prove the same holds for differences: $s_n - t_n$ also converges, and

$$\lim(s_n - t_n) = \lim s_n - \lim t_n$$

The case of products is a little more annoying to prove, but the end result is the same - the limit of a product is the product of the limits.

Theorem 9.4 (Limit of a Product)**.** *Let s_n, t_n be convergent sequences. Then the sequence of term-wise products $s_n t_n$ is convergent, with*

$$\lim(s_n t_n) = (\lim s_n)(\lim t_n)$$

Sketch. Let $s_n \to S$ and $t_n \to T$ be two convergent sequences and choose $\epsilon > 0$. We wish to find an N beyond which we know $s_n t_n$ lies within ϵ of $ST.

To start, we consider the difference $|s_n t_n - ST|$ and we add zero in a clever way:

$$|s_n t_n - ST| = |s_n t_n - s_n T + s_n T - ST| = |(s_n t_n - s_n T) + (s_n T - ST)|$$

applying the triangle inequality we can break this apart

$$|s_n t_n - ST| \leq |s_n t_n - s_n T| + |s_n T - ST| = |s_n||t_n - T| + |s_n - S||T|$$

The second term here is easy to bound: if $T = 0$ then its just literally zero, and if $T \neq 0$ then we can make it as small as we want: we know $s_n \to S$ so we can make $|s_n - S|$ smaller than anything we need (like ϵ/T, or even $\epsilon/2T$ if necessary).

For the first term we see it includes a term of the form $|t_n - T|$ which we know we can make as small as we need to by choosing sufficiently large N. But its being multiplied by $|s_n|$ and we need to make sure the whole thing can be made small, so we should worry about what if $|s_n|$ is getting really big? But this isn't actually a worry - we know s_n is convergent, so its bounded, so there is some B where $|s_n| < B$ for all n. Now we can make $|t_n - T|$ as small as we like, (say, smaller than ϵ/B or $\epsilon/2B$ or whatever we need).

Since each of these terms can be made small as we need individually, choosing large enough n's we can make them both simultaneously small, so the whole difference $|s_n t_n - ST|$ is small (less than ϵ) which proves convergence. □

Exercise 9.5. Write the sketch of an argument above in the right order, as a formal proof.

Corollary 9.2. *If p is a positive integer then*

$$\lim \frac{1}{n^p} = 0$$

Hint: Induction on the power p

The next natural case to consider after sums and differences and products is quotients. We begin by considering the limit of a reciprocal:

Proposition 9.3 (Limit of a Reciprocal)**.** *Let s_n be a convergent nonzero sequence wtih a nonzero limit. Then the sequence $1/s_n$ of reciprocals is convergent, with*

$$\lim \frac{1}{s_n} = \frac{1}{\lim s_n}$$

Sketch. For any $\epsilon > 0$, want to show when n is very large, we can make

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| < \epsilon$$

We can get a common denominator and rewrite this as

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| = \frac{|s - s_n|}{|ss_n|}$$

Since s_n is not converging to zero, we should be able to bound it away from zero: that is, find some m such that $|s_n| > m$ for all $n \in \mathbb{N}$ (we'll have to prove we can actually do this). Given such an m we see the denominator $|ss_n| > m|s|$, and so

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| < \frac{|s_n - s|}{m|s|}$$

We want this less than ϵ so all we need to do is choose N big enough that $|s_n - s|$ is less than $\epsilon m|s|$ and we're good. □

Exercise 9.6. Turn the sketch argument for $\lim \frac{1}{s_n} = \frac{1}{s_n}$ in Proposition 9.3 into a formal proof.

From here, its quick work to understand the limit of a general quotient.

Theorem 9.5 (Limit of a Quotient)**.** *Let s_n, t_n be convergent sequences, with $t_n \neq 0$ and $\lim t_n \neq 0$. Then the sequence s_n/t_n of quotients is convergent, with*

$$\lim \frac{s_n}{t_n} = \frac{\lim s_n}{\lim t_n}$$

Proof. Since t_n converges to a nonzero limit, by Proposition 9.3 we know that $1/t_n$ converges, with limit $1/\lim t_n$. Now, we can use Theorem 9.4 for the product $s_n \cdot \frac{1}{t_n}$:

$$\lim \frac{s_n}{t_n} = \lim s_n \cdot \frac{1}{t_n} = (\lim s_n)\left(\lim \frac{1}{t_n}\right)$$

$$= \lim s_n \frac{1}{\lim t_n} = \frac{\lim s_n}{\lim t_n}$$

□

Finally we look at square roots. We have already proven in Theorem 6.9 that nonnegative numbers have square roots, and so given a nonnegative sequence s_n we can consider the sequence $\sqrt{s_n}$ of its roots. Below we see that the limit concept respects roots just as it does the other field operations:

Theorem 9.6 (Root of Convergent Sequence)**.** *Let $s_n > 0$ be a convergent sequence, and $\sqrt{s_n}$ its sequence of square roots. Then $\sqrt{s_n}$ is convergent, with*

$$\lim \sqrt{s_n} = \sqrt{\lim s_n}$$

Sketch. Assume $s_n \to s$, and fix $\epsilon > 0$. We seek an N where $n > N$ implies $|\sqrt{s_n} - \sqrt{s}| < \epsilon$. This looks hard: because the fact we know is about $s_n - s$ and the fact we need is about $\sqrt{s_n} - \sqrt{s}$.
But what if we multiply and divide by $\sqrt{s_n} + \sqrt{s}$ so we can simplify using the difference of squares?

$$|\sqrt{s_n} - \sqrt{s}| \frac{\sqrt{s_n} + \sqrt{s}}{\sqrt{s_n} + \sqrt{s}} = \frac{|s_n - s|}{\sqrt{s_n} + \sqrt{s}}$$

This has the quantity $|s_n - s|$ that we know about in it! We know we can make this as small as we like by the assumption $s_n \to s$, so as long as the denominator does not go to zero, we can make this happen!

□

Formal. Let s_n be a positive sequence with $s_n \to s$ and assume $s \neq 0$ (we leave that case for the exercise below). Let $\epsilon > 0$, and choose N such that if $n > N$ we have $|s_n - s| < \epsilon\sqrt{s}$.

Now for any n, rationalizing the numerator we see

$$|\sqrt{s_n} - \sqrt{s}| = \frac{|s_n - s|}{\sqrt{s_n} + \sqrt{s}} < \frac{|s_n - s|}{\sqrt{s}}$$

Where the last inequality comes from the fact that $\sqrt{s_n} > 0$ by definition, so $\sqrt{s} + \sqrt{s_n} > \sqrt{s}$. When $n > N$ we can use the hypothesis that $s_n \to s$ to see

$$|\sqrt{s_n} - \sqrt{s}| < \frac{|s_n - s|}{\sqrt{s}} = \frac{\epsilon\sqrt{s}}{\sqrt{s}} = \epsilon$$

And so, $\sqrt{s_n}$ is convergent, with limit $\sqrt{s}$. □

Exercise 9.7. Prove that if $s_n \to 0$ is a sequence of nonnegative numbers, that the sequence of roots also converges to zero $\sqrt{s_n} \to 0$.

Hint: you don't need to rationalize the numerator or do fancy algebra like above

Together this suite of results provides an effective means of calculating limits from simpler pieces. They are often referred to together as *the limit theorems*

Theorem 9.7 (The Limit Theorems)**.** *Let a_n and b_n be any two convergent sequences, and $k \in \mathbb{R}$ a constant. Then*

$$\lim ka_n = k \lim a_n$$

$$\lim(a_n \pm b_n) = (\lim a_n) \pm (\lim b_n)$$

$$\lim a_n b_n = (\lim a_n)(\lim b_n)$$

If $b_n \neq 0$ and $\lim b_n \neq 0$,

$$\lim \frac{a_n}{b_n} = \frac{\lim a_n}{\lim b_n}$$

And, if $a_n \geq 0$, then $\sqrt{a_n}$ is convergent, with

$$\lim \sqrt{a_n} = \sqrt{\lim a_n}$$

9.2.1. ⋆ Infinity

Given the formal defintion of *divergence to infinity* as meaning *eventually gets larger than any fixed number*, we can formulate analogs of the limit theorems for such divergent sequences. We will not need any of these in the main text but it is good practice to attempt their proofs:

Exercise 9.8. If $s_n \to \infty$ and $k > 0$ then $ks_n \to \infty$.

Exercise 9.9. If t_n diverges to infinity, and s_n either converges, or also diverges to infinity, then $s_n + t_n \to \infty$.

Exercise 9.10. If t_n diverges to infinity, and s_n either converges, or also diverges to infinity, then $s_n t_n \to \infty$.

Note that there is *not* an analog of the division theorem: if $s_n \to \infty$ and $t_n \to \infty$, with only this knowledge we can learn nothing about the quotient s_n/t_n.

Exercise 9.11. Give examples of sequences $s_n, t_n \to \infty$ where

$$\lim \frac{s_n}{t_n} = 0$$

$$\lim \frac{s_n}{t_n} = 2$$

$$\lim \frac{s_n}{t_n} = \infty$$

These limit laws are the precise statement behind the "rules" often seen in a calculus course, where students may write $2+\infty = \infty$, $\infty+\infty = \infty$, or $\infty\cdot\infty = \infty$, but they may not write ∞/∞. (If you are looking at this last case and thinking *l'Hospital*, we'll get there in **?@thm-Lhospital**!)

9.2.2. Example Computations

Example 9.4. Compute the limit of the following sequence s_n:

$$s_n = \frac{3n^3 + \frac{n^6-2}{n^2+5}}{n^3 - n^2 + 1}$$

Example 9.5. Compute the limit of the sequence s_n

$$s_n = \sqrt{\frac{1}{2^n} + \sqrt{\frac{n^2-1}{n^2-n+1}}}$$

9.3. Applications

9.3.1. Babylon and $\sqrt{2}$

We know that $\sqrt{2}$ exists as a real number (Theorem 6.7), and we know that the babylonian procedure produces excellent rational approximations to this value (Exercise 1.5), in the precise sense that the numerator squares to just one more than twice the square of the denominator.

Now we finally have enough tools to combine these facts, and prove that the babylonian procedure really does limit to $\sqrt{2}$.

Theorem 9.8. *Let $s_n = \frac{p_n}{q_n}$ be a sequence of rational numbers where both $p_n, q_n \to \infty$ and for each $p_n^2 = 2q_n^2 - 1$. Then $s_n \to \sqrt{2}$.*

Proof. We compute the limit of the sequence s_n^2. Using that $p_n^2 = 2q_n^2 + 1$ we can replace the numerator and do algebra to see

$$s_n^2 = \frac{p_n^2}{q_n^2} = \frac{2q_n^2 + 1}{q_n^2} = 2 + \frac{1}{q_n^2}.$$

Now, as by assumption $q_n \to \infty$ we have that $q_n^2 = q_n q_n$ also diverges to infinity (Exercise 9.10), and so its reciprocal converges to 0 (Proposition 8.3). Thus, using the limit theorems for sums,

$$\lim \frac{p_n^2}{q_n^2} = \lim\left(2 - \frac{1}{q_n^2}\right) = 2 - \lim \frac{1}{q_n^2} = 2$$

That is, the limit of the squares approaches 2. Now we apply Theorem 9.6 to this sequence s_n^2, and conclude that

- $s_n = \sqrt{s_n^2}$ converges.
- $\lim s_n = \lim \sqrt{s_n^2} = \sqrt{\lim s_n^2} = \sqrt{2}$

□

This provides a *rigorous justification* of the babylonian's assumption that if you are patient, and compute more and more terms of this sequence, you will *always* get better and better approximations of the square root of 2.

Exercise 9.12. Build a sequence that converges to $\sqrt{n}$ by following the babylonian procedure, starting with a rectangle of area n.

9.3.2. Rational and Irrational Sequences

Combining the squeeze theorem and limit theorems with the density of the (ir)rationals allows us to prove the existence of certain sequences that will prove quite useful:

Theorem 9.9. *For every $x \in \mathbb{R}$ there exists a sequence r_n of rational numbers with $r_n \to x$.*

Proof. Let $x \in \mathbb{R}$ be arbirary, and consider the sequence $x + \frac{1}{n}$. Because the constant sequence $x, x, x \ldots$ and the sequence $1/n$ are convergent, by the limit theorem for sums we know $x + \frac{1}{n}$ is convergent and

$$\lim\left(x + \frac{1}{n}\right) = x + \lim \frac{1}{n} = x$$

Now for each $n \in \mathbb{N}$, by the density of the rationals we can find a rational number r_n with $x < r_n < x + \frac{1}{n}$. This defines a sequence of rational numbers squeezed between x and $x + \frac{1}{n}$: thus, by the squeeze theorem we hav

$$x < r_n < x + \frac{1}{n} \implies \lim r_n = x$$

□

Through a similar argument using Exercise 6.7 we find the existence of a sequence of irrational numbers converging to any real number.

Exercise 9.13. For every $x \in \mathbb{R}$ there exists a sequence y_n of irrationals with $y_n \to x$.

10. Monotone Convergence

Highlights of this Chapter: We prove the monotone convergence theorem, which is our first theorem that tells us a sequence converges, without having to first know its limiting value. We show how to use this theorem to find the limit of various recursively defined sequences, including two important examples.

- We prove the infinite sequence of roots $\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}$ converges to the golden ratio.
- We prove the sequence $\left(1+\frac{1}{n}\right)^n$ converges to the number $e = 2.71828\ldots$
- We begin a treatment of irrational exponents, by looking at the limit of sequences with rational exponents.

The motivation for inventing sequences is to work with infinite processes, where we have a precise description of each finite stage, but cannot directly grasp the "completed" state "at infinity". In the first section of this chapter we computed a few specific limits, and then in the second we showed *how to find new, more complicated limits* given that you know the value of some simpler ones via algebra.

But what we haven't done, since our original motivating discussion with the nested intervals theorem, is actually return to the part of the theory we are most interested in: rigorously assuring that certain sequences converge, *without* knowing the value of their limit ahead of time! The most useful theorem in this direction is the *monotone convergence theorem*, which deals with monotone sequences.

Definition 10.1 (Monotone Sequences)**.** A sequence s_n is monotone increasing (or more precisely, monotone *non-decreasing*) if

$$m \leq n \implies s_m \leq s_n$$

A sequence is monotone decreasing (non-increasing) if

$$m \leq n \implies s_m \geq s_n$$

Note: constant sequences are monotone: both monotone increasing and monotone decreasing.

The original inspiration for a monotone sequence is the sequence of upper bounds or lower bounds from a collection of nested intervals: as the intervals get smaller, the lower bounds monotonically increase, and the upper bounds monotonically decrease. The *Monotone convergence theorem* guanatees that such sequences always converge. Its proof is below, but could actually be extracted directly from Theorem 5.2.

Theorem 10.1 (Monotone Convergence)**.** *Let s_n be a monotone bounded sequence. Then s_n is a convergent sequence.*

Proof. Here we consider the case that s_n is monotone increasing, and leave the decreasing case as an exercise. Let $S = \{s_n \mid n \in \mathbb{N}\}$. Then S is nonempty, and is bounded above (by any upper bound for the sequence s_n, which we assumed is bounded). Thus by completeness, it has a supremum $s = \sup S$.

We claim that s_n is actually a convergent sequence, which limits to s_n. To prove this, choose $\epsilon > 0$, and note that as s is the *least* upper bound, $s - \epsilon$ is not an upper bound for S, so there must be some N where $s_N > s - \epsilon$. But s_n is monotone increasing, so if $n > N$ it follows that $s_n > s_N$. Recalling that for all n we know $s_n \leq s$ (since s is an upper bound), we have found some N where for all $n > N$ we know $s - \epsilon < s_n < s$. This further implies $|s_n - s| < \epsilon$, which is exactly the definition of convergence! Thus

$$s_n \to s$$

So it is a convergent sequence, as claimed. □

Though straightforward to prove, this theorem has tons of applications, as it assures us that many of the difficult to describe recursively defined sequences that show up in practice actually do converge, and thus we may rigorously reason about their limits. We will give several interesting ones below.

10.1. Infinite Recursion

Remember a *recursively defined sequence* is one given by iterating some function f starting from an initial value a_0, as $a_{n+1} = f(a_n)$. We saw one such sequence previously, defined by $s_{n+1} = \sqrt{1 + s_n}$ starting from $s_0 = 1$:

$$s_0 = 1$$

$$s_1 = \sqrt{1 + \sqrt{1}}$$

$$s_2 = \sqrt{1 + \sqrt{1 + \sqrt{1}}}$$

$$\ldots$$

Because such sequences follow a regular pattern, we can use a shorthand notation with ellipsis for their terms. For example, in the original sequence above, writing the first couple steps of the pattern followed by an ellipsis

$$\sqrt{1+\sqrt{1+\sqrt{1+\ldots}}}$$

we take to mean the sequence of terms s_n where $s_{n+1} = \sqrt{1+s_n}$ itself. Thus, writing $\lim\sqrt{1+\sqrt{1+\cdots}}$ means the limit of this sequence, implicitly defined by this infinite expression.

Exercise 10.1. Here are some other infinite expressions defined by recursive sequences: can you give the recursion relation they satisfy?

$$\sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdots}}}$$

$$\frac{1}{1+\frac{1}{1+\cdots}}$$

$$\cos(\cos(\cos(\cdots\cos(5)\cdots)))$$

In all of these sequences it is not clear *at all* how to find their limit value from scratch, or how we could possibly apply any of the limit theorems about field axioms and inequalities. But, recursive sequences are set up for using induction, and monotone convergence! We can build a sort of recipe for dealing with them:

Recursive Sequence Operation Manual:

- Prove its bounded, by induction.
- Prove its monotone, by induction.
- Use Monotone convergence to conclude its convergent.
- Use the recursive definition, and the limit theorems, to find an equation satisfied by the limit.
- Solve that equation, to find the limit.

A beautiful and interesting example of this operations manual is carried out below:

Proposition 10.1. *The sequence* $\sqrt{1+\sqrt{1+\cdots}}$ *converges to the golden ratio.s*

Proof. The infinite expression $\sqrt{1+\sqrt{1+\cdots}}$ defines the recursive sequence $s_{n+1} = \sqrt{1+s_n}$ with $s_1 = 1$.

Step 1: s_n is monotone increasing, by induction First we show that $s_2 > s_1$. Using the formula, $s_2 = \sqrt{1+\sqrt{1}} = \sqrt{2}$, which is larger than $s_1 = 1$. Next, we assume for induction $s_n > s_{n-1}$ and we use this to prove that $s_{n+1} > s_n$. Starting from our induction hypothesis, we add one to both sides yielding $1+s_n > 1+s_{n-1}$ and then we take the square root (which preserves the inequality, by Proposition 4.5) to get

$$\sqrt{1+s_n} > \sqrt{1+s_{n-1}}$$

But now, we simply note that the term on the left is the definition of s_{n+1} and the term on the right is the definition of s_n. Thus we have $s_{n+1} > s_n$ as claimed, and our induction proof works for all n.

Step 2: s_n is bounded, by induction It is hard to guess an upper bound for s_n without doing a little calculation, but plugging the first few terms into a calculator shows them to be less than 2, so we might try to prove $\forall n\, s_n < 2$. The base case is immediate as $s_1 = 1 < 2$, so assume for induction $s_n < 2$. Then $1+s_n < 3$ and so $\sqrt{1+s_n} < \sqrt{1+2} = \sqrt{3}$, and $\sqrt{3} < 2$ (as $3 < 2^2 = 4$) so our induction has worked, and the entire sequence is bounded above by 2.

Conclusion: s_n converges! We have proven the sequence s_n is both monotone increasing and bounded above by 2. Thus the monotone convergence theorem assures us that there exists some L with $s_n \to L$. It only remains to figure out what number this is!

Step 3: The Limit Theorems Because truncating the beginning of a sequence does not change its limit, we see that $\lim s_n = \lim s_{n+1} = L$. But applying the limit theorems to $s_{n+1} = \sqrt{1+s_n}$, we see that as $s_n \to L$, it follows that $1+s_n \to 1+L$ and thus that $\sqrt{1+s_n} \to \sqrt{1+L}$. This gives us an equation that L must satisfy!

$$\sqrt{1+L} = L$$

Simplifying this becomes $1+L = L^2$, which has solutions $(1 \pm \sqrt{5})/2$. This argument only tells us so far that one of these numbers must be our limit L: to figure out which we need to bring in more information. Noticing that only one of the two is positive, and all the terms of our sequence are positive singles it out:

$$\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}} = \frac{1+\sqrt{5}}{2} \approx 1.618\ldots$$

This number is known as the *golden ratio.* □

Example 10.1. The final step of the proof above suggests a way one might find a recursive sequence to use as a calculating tool: if we started with the golden ratio

$$\phi = \frac{1+\sqrt{5}}{2}$$

we could observe that ϕ solves the quadratic equation $1+L=L^2$, and hence $L=\sqrt{1+L}$. This sets up a recursive sequence, as we can plug this relation into itself over and over:

$$L=\sqrt{1+L}=\sqrt{1+\sqrt{1+L}}=\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}$$

Which immediately suggests the recursion $s_{n+1}=\sqrt{1+s_n}$ as a candidate for generating a sequence that would solve the original equation.

Exercise 10.2. Find a recursive sequence whose limit is the positive real root of x^2-2x-5. Then prove that your proposed sequence actually converges to this value.

Exercise 10.3. What number is this?

$$\sqrt{1-2\sqrt{1-2\sqrt{1-2\sqrt{\cdots}}}}$$

10.2. ⋆ The Number *e*

In this section we aim to study, and prove the convergence of the following sequence of numbers

$$\left(1+\frac{1}{n}\right)^n$$

We will later see that the limit of this sequence is the number e (indeed, many authors take this sequence itself as *the definition of e* as it is perhaps the first natural looking sequence limiting to this special value. We will instead define e in terms of *exponential functions* to come, and then later show its value coincides with this limit).

We begin by proving a_n is monotone as a prelude to applying monotone convergence.

Example 10.2. The sequence $a_n=\left(\frac{n+1}{n}\right)^n$ is monotone increasing.

Proof. To show a_n is increasing we will show that the ratio $\frac{a_n}{a_{n-1}}$ is greater than 1. Simplifying,

$$\frac{a_n}{a_{n-1}}=\frac{\left(\frac{n+1}{n}\right)^n}{\left(\frac{n}{n-1}\right)^{n-1}}=\left(\frac{n+1}{n}\right)^n\left(\frac{n-1}{n}\right)^{n-1}$$

Multiplying by $\frac{n-1}{n}$ and its inverse we can make the powers on each of these terms the same, and combine them:

$$= \left(\frac{n+1}{n}\right)^n \left(\frac{n-1}{n}\right)^n \frac{n}{n-1} = \left(\frac{n^2-1}{n^2}\right)^n \frac{n}{n-1}$$

Simplifying what is in parentheses, we notice that we are actually in a perfect situation to apply Bernoulli's Inequality (Exercise 4.6) to help us estimate this term. Recall this says that if r is any number such that $1+r$ is positive, $(1+r)^n \geq 1+nr$. When $n \geq 2$ we can apply this to $r = -\frac{1}{n^2}$, yielding

$$\frac{a_n}{a_{n-1}} = \left(1 - \frac{1}{n^2}\right)^n \frac{n}{n-1} \geq \left(1 - \frac{n}{n^2}\right) \frac{n}{n-1}$$
$$= \frac{n-1}{n} \frac{n}{n-1} = 1$$

Thus $\frac{a_n}{a_{n-1}} \geq 1$, so $a_n \geq a_{n-1}$ and the sequence is monotone increasing for all n, as claimed. □

Next we need to show that a_n is bounded above. Computing terms numerically, it *seems* that a_n is bounded above by 3, but of course no amount of computation can substitute for a proof. And after a bit of trying, it seems hard to prove *directly* that it actually is bounded above.

So instead, we will employ a bit of an ingenious trick. We will study a second sequence, which appears very similar to the first:

$$b_n = \left(\frac{n+1}{n}\right)^{n+1}$$

Indeed, this is just our sequence a_n multiplied by one extra factor of $\frac{n+1}{n}$! But this extra factor changes its behavior a bit: computing the first few terms, we see that it appears to be decreasing:

$$b_1 = (1+1)^2 = 4,\ b_2 = \left(1 + \frac{1}{2}\right)^3 = \frac{27}{8} = 3.375,\ b_3 = \left(1 + \frac{1}{3}\right)^4 \approx 3.1604$$

Indeed, a proof that its decreasing can be constructed following an identical strategy to a_n in Example 10.2.

Exercise 10.4. The sequence $b_n = \left(\frac{n+1}{n}\right)^{n+1}$ is monotone decreasing.

Now that we understand the behavior of b_n we can use it to prove that a_n is bounded above:

Corollary 10.1. *The sequence* $a_n = \left(1+\frac{1}{n}\right)^n$ *is convergent*

Proof. Note that the sequence b_n and a_n are related by

$$b_n = \left(\frac{n+1}{n}\right)^{n+1} = a_n\left(\frac{n+1}{n}\right)$$

Since $\frac{n+1}{n} > 1$ we see that $b_n > a_n$ for all n. But b_n is decreasing, so $b_n \leq b_1 = 2^2 = 4$, and so a_n is bounded above by 4. □

Note that we can also readily see that b_n is itself convergent (though we did not actually need that fact for our analysis of a_n): we proved its monotone decreasing, and its a sequence of positive terms - so its trivially bounded below by zero!

We can also see that a_n and b_n have the same limit, using the limit theorems. Since $\frac{1}{n} \to 0$, we know that $1+\frac{1}{n} \to 1$, and hence that

$$\begin{aligned}\lim b_n &= \lim\left[a_n\left(\frac{n+1}{n}\right)\right]\\ &= (\lim a_n)\cdot\left(\lim\frac{n+1}{n}\right)\\ &= \lim a_n\end{aligned}$$

As mentioned previously, we will later see that this limit is the number called e. But believing for a moment that we *should* be interested in this particular limit, having the two sequences a_n and b_n lying around actually proves quite practically useful for estimating its value.

Since $\lim a_n = e = \lim b_n$ and $a_n < b_n$ for all n, we see that the number e is contained in the interval $I_n = [a_n, b_n]$, and hence is the limit of the nested intervals:

Corollary 10.2.

$$\{e\} = \bigcap_{n\geq 1}\left[\left(1+\frac{1}{n}\right)^n, \left(1+\frac{1}{n}\right)^{n+1}\right]$$

Taking any finite n, this interval gives us both an upper and lower bound for e: for example

$$\begin{aligned}n = 10 &\implies 2.59374 \leq e \leq 2.85311\\ n = 100 &\implies 2.7048 \leq e \leq 2.73186\\ n = 1,000 &\implies 2.71692 \leq e \leq 2.71964\\ n = 1,000,000 &\implies 2.71826 \leq e \leq 2.71829\end{aligned}$$

Thus, correct to four decimal places we know $e \approx 2.7182$

10.3. Application: Defining Irrational Powers

We have already defined rational powers of a number in terms of iterated multiplication/division, and the extraction of roots: but how does one define a real numbered power? We can use sequences to do this! To motivate this, let's consider the example of defining 2^π. We can write π as the limit of a sequence of rational numbers, for instance

$$3,\ 3.1,\ 3.14,\ 3.141,\ 3.1415\ \dots$$

And since rational exponents make sense, from this we can produce a sequence of exponentials

$$2^3,\ 2^{\frac{31}{10}},\ 2^{\frac{314}{100}},\ 2^{\frac{3141}{1000}},\ 2^{\frac{31415}{10000}},\dots$$

Then we may ask if this sequence has a limit: if it does, it's natural to try and define this value two to the power of pi. To make sure this makes sense, we need to check several potential worries:

- Does this sequence converge?
- Does the limit depend on the particular sequence chosen?

For example if you tried to define $3^{\sqrt{2}}$ using the babylonian sequence for $\sqrt{2}$, and your friend tried to use the sequence coming from the partial fraction, you'd better get the same number if this is a reasonable thing to define! Because we are in the section on monotone convergence, we will restrict ourselves at the moment to *monotone sequences* though we will see later we can dispense with this if desired.

Proposition 10.2. *If $r_n \to x$ is a monotone sequence of rational numbers converging to x, and $a > 0$ then the sequence a^{r_n} converges.*

Proof. Recall for a fixed positive base a, exponentiation by rational numbers is monotone increasing, so $r < s$ implies $a^r < a^s$.
Thus, given a monotone sequence r_n, the exponentiated sequence a^{r_n} remains monotone (for monotone increasing we see $r_n \le r_{n+1} \implies a^{r_n} \le a^{r_{n+1}}$ and the equalities are reversed if r_n is monotone decreasing).

Now that we know a^{r_n} is monotone, we only need to see its bounded to apply Monotone Convergence. Again we have two cases, and will deal here with the monotone increasing case. As $r_n \to x$ and x is a real number, there must be some natural number $N > x$. Thus, N is greater than r_n for all n, and so a^N is greater than a^{r_n}: our sequence is bounded above by a^N. Thus all the hypotheses of monotone convergence are satisfied, and $\lim a^{r_n}$ exists. ☐

Now that we know such sequences make sense, we wish to clear up any potential ambiguity, and show that if two different sequences both converge to x, the value we attempt to assign to a^x as a limit is the same for each. As a lemma in this direction, we look at sequences converging to zero.

Exercise 10.5. Let r_n be any sequence of rationals converging to zero. Then for any $a > 0$ we have

$$\lim a^{r_n} = 1$$

Corollary 10.3. *If r_n, s_n are two monotone sequences of rationals each converging to x, then*

$$\lim a^{r_n} = \lim a^{s_n}$$

for any $a > 0$.

Proof. Let $z_n = r_n - s_n$, so that $z_n \to 0$. Because r_n and s_n are monotone, we know $\lim a^{r_n}$ and $\lim a^{s_n}$ exist. And by the exercise above, we have $a^{z_n} \to 1$. Noting that $r_n = s_n + z_n$ and that the laws of exponents apply for rational exponents, we have

$$a^{r_n} = a^{s_n+z_n} = a^{s_n}a^{z_n}$$

But as all quantities in question converge we can use the limit theorems to compute:

$$\begin{aligned}\lim a^{r_n} &= \lim a^{s_n+z_n}\\ &= \lim a^{s_n}a^{z_n}\\ &= (\lim a^{s_n})(\lim a^{z_n})\\ &= \lim a^{s_n}\end{aligned}$$

□

Thus, we can unambiguously define the value of a^x as the limit of *any monotone sequence* a^{r_n} without specifying the sequence itself.

Definition 10.2. ## Irrational Powers Let $a > 0$ and $x \in \mathbb{R}$. Then we define a^x as a limit

$$a^x = \lim a^{r_n}$$

For r_n any monotone sequence of rational numbers converging to x.

Perhaps upon reading this definition to yourself you wonder, is the restriction to monotone sequences important, or just an artifact of our currently limited toolset? Once we build more tools we will see the latter is the case; you will show on homework that arbitrary convergent sequences $r_n \to x$ can be used to unambiguously define a^x.

11. Subsequences

Highlights of this Chapter: We define the concept of subsequence, and investigate examples where subsequences behave much simpler than the overall sequence with the example of continued fractions. We then investigate the relationship between the convergence of subsequences and the convergence of a sequence as a whole. This leads to several nice theorems:

- A continued fraction description of the golden ratio and $\sqrt{2}$
- Theorem: a sequence converges if it is a union of subsequences converging to the same limit.
- Theorem: every bounded sequence contains a convergent subsequence.

Definition 11.1. A *subsequence* is a *subset* of a sequence which is itself a sequence. As sequences are infinite ordered lists of real numbers, an equivalent definition is that a subsequence is *any* infinite subset of a sequence.

We often denote an abstract subsequence like s_{n_k}, meaning that we have kept only the n_k terms of the original, and discarded the rest.

Example 11.1 (Example Subsequences)**.** In the sequence of all n-gons inscribed in a circle, the collection studied by archimedes (CITE EALRIER CHAP) by doubling is the subsequence

$$\begin{aligned} P_{3\cdot 2^k} &= (P_{3\cdot 2^1}, P_{3\cdot 2^2}, P_{3\cdot 2^3}, P_{3\cdot 2^4}, \dots) \\ &= (P_6, P_{12}, P_{24}, P_{48}, \dots) \end{aligned}$$

Archimedes began his estimation of π using a simple idea: create a sequence of nested intervals (upper and lower bounds) from inscribing and circumscribing n-gons. But then he realized calculations would be much simpler if he focused only on a *subsequence*, namely that generated by side-doubling. We too will often run into situations like Archimedes, where the overall behavior of a sequence is difficult to understand, but we can pull out subsequences that are much easier to work with.

We will begin our exploration with an extended example, that illuminates the main idea.

11.1. Continued Fractions

In the previous section, we uncovered a beautiful formula for the golden ratio as the limit of an infinite process of square roots. However, practically speaking (if you were interested in calculating the value of the golden ratio, as the ancient mathematicians were) this series is *useless*. The golden ratio itself involves a square root, so if you are seeking a method of approximations its fair to assume that you cannot evaluate the square root function exactly. But what does our sequence of approximations look like? To calculate the n^{th} term, you need to take n square roots! The very terms of our convergent sequence are actually much *much* more algebraically complicated than their limiting value.

To be practical, we would like a sequence that (1) contains easy to compute terms, and (2) converges to the number we seek to understand. By **?@thm-rational-sequence**, we know for any real number there exists a sequence of rationals that converges to it, and so it's natural to seek a method of producing such a thing.

One method is the *continued fraction*, which is best illustrated by example. We know that the golden ratio L satisfies the equation $L^2 = L + 1$, and dividing by L this gives us an equation satisfied by L and $1/L$:

$$L = 1 + \frac{1}{L}$$

Just like we did above, we can use this self-referential equation to produce a series, by plugging it into itself over and over. After one such substitution we get

$$L = 1 + \frac{1}{1 + \frac{1}{L}}$$

And then after another such we get

$$L = 1 + \frac{1}{1 + \frac{1}{1+\frac{1}{L}}}$$

Continuing this way over and over, we push the L "off to infinity" on the right hand side, and are left with an infinite expression for L, as a limit of a sequence of fractions.

$$L = 1 + \frac{1}{1 + \frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\cdots}}}}}}}$$

Of course, this 'infinite manipulation' is not itself rigorous, but we can interpret this as a recursive sequence exactly as above. Setting $s_1 = 1$, we have the rule $s_{n+1} = 1 + \frac{1}{s_n}$, and we wish to understand $\lim s_n$.

Example 11.2 (Continued Fraction of the Golden Ratio)**.** The continued fration

$$1 + \cfrac{1}{1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}}}}}}}$$

defined by the recursive sequence $s_1 = 1$, $s_{n+1} = 1 + \frac{1}{s_n}$ limits to the golden ratio.

A continued fraction is a recursive sequence, so we can compute everything with the starting value and a single simple rule. To get a feel for the sequence at hand, let's compute the first few terms:

$$s_1 = 1, s_2 = 2, s_3 = \frac{3}{2}, s_4 = \frac{5}{3}, s_5 = \frac{8}{5}, s_6 = \frac{13}{8}, s_6 = \frac{21}{13}, \ldots$$

What's one thing we notice about this sequence from its first few terms? Well - it looks like the fractions are all ratios of Fibonacci numbers! This won't actually be relevant but it's a good practice of induction with the sequence definition, so we might as well confirm it:

Example 11.3 (Fibonacci Numbers and the Golden Ratio)**.** Recall that the Fibonacci numbers are defined by the recurrence relation $F_1 = F_1 = 2$ and $F_{n+2} = F_{n+1} + F_n$. Show that the n^{th} convergent s_n of the continued fraction for the golden ratio is the ratio of the Fibonacci numbers F_{n+1}/F_n.

Proof. This is true for the first convergent which is 1, and $F_2/F_1 = 1/1 = 1$. Assume the n^{th} convergent is $s_n = F_{n+1}/F_n$, and consider the $n + 1^{st}$: this is

$$s_{n+1} = 1 + \frac{1}{s_n} = 1 + \frac{1}{\frac{F_{n+1}}{F_n}}$$

$$= 1 + \frac{F_n}{F_{n+1}} = \frac{F_{n+1} + F_n}{F_{n+1}} = \frac{F_{n+2}}{F_{n+1}}$$

□

The more important thing we notice is that looking at the magnitude of the terms, it is neither increasing or decreasing, but it appears the sequence is zig-zagging up and down. Its straightforward to prove this is actually the case:

Example 11.4. If n is odd, then $s_n < s_{n+1}$. If n is even, $s_n > s_{n+1}$.

Proof. Again, we proceed by induction: we prove only the first case, and leave the second as an exercise. Note first $s_1 = 1$, $s_2 = 2$ and $s_3 = \frac{3}{2}$ so $s_1 < s_2$ and $s_2 > s_3$: the base case of each is true.

Now, assume that n is odd, and $s_n < s_{n+1}$. Computing from here

$$s_n < s_{n+1} \implies \frac{1}{s_n} > \frac{1}{s_{n+1}} \implies 1 + \frac{1}{s_n} > 1 + \frac{1}{s_{n+1}}$$

The last line of this computation is the definition of $s_{n+1} > s_{n+2}$,so we see the next one is decreasing as claimed. And applying the recurrence once more:

$$s_{n+1} > s_{n+2} \implies \frac{1}{s_{n+1}} < \frac{1}{s_{n+2}} \implies 1 + \frac{1}{s_{n+1}} < 1 + \frac{1}{s_{n+2}}$$

Where now the last line of the calculation is the definition of $s_{n+2} < s_{n+3}$, fininshing our induction step! □

While the overall sequence isn't monotone, it seems to be built of two different monotone sequences, interleaved with one another! In particular the *odd subsequence* $s_1, s_3, s_5, \ldots$ is monotone increasing, and the *even subsequence* $s_2, s_4, s_6, \ldots$ is monotone decreasing.

To study these subsequences separately, we first need to find a recurrence relation that gives us s_{n+2} in terms of s_n: applying this to either s_1 or s_2 will then produce the entire even or odd subsequence.

$$s_{n+2} = 1 + \frac{1}{s_{n+1}} = 1 + \frac{1}{1 + \frac{1}{s_n}}$$

Example 11.5. The subsequence $s_1, s_3, s_5, s_7, \ldots$ is monotone increasing.

Proof. We prove this by induction. Starting from $s_1 = 1$, we compute

$$s_3 = 1 + \frac{1}{1 + \frac{1}{1}} = 1 + \frac{1}{2} = \frac{3}{2}$$

So $s_1 < s_3$, completing the base case. Next, assume for induction that $s_{n+2} > s_n$. We wish to show that $s_{n+4} > s_{n+2}$. Calculating from our assumption:

$$\begin{aligned}
s_{n+2} > s_n &\implies \frac{1}{s_{n+2}} < \frac{1}{s_n} \\
&\implies 1 + \frac{1}{s_{n+2}} < 1 + \frac{1}{s_n} \\
&\implies \frac{1}{1 + \frac{1}{s_{n+2}}} > \frac{1}{1 + \frac{1}{s_n}} \\
&\implies 1 + \frac{1}{1 + \frac{1}{s_{n+2}}} > 1 + \frac{1}{1 + \frac{1}{s_n}} \\
&\implies s_{n+4} > s_{n+2}
\end{aligned}$$

This completes the induction step, so the subsequence of odd terms is monotone increasing as claimed! □

A nearly identical argument applies to the even subsequence:

Exercise 11.1. The subsequence $s_2, s_4, s_6, s_8, \ldots$ is monotone decreasing.

Exercise 11.2. Let $f(x) = 1 + \frac{1}{x}$. Show that if $x < y$ then $f(x) > f(y)$; that is, f reverses the ordering of numbers. Use this to give a more streamlined proof that the even and odd subsequences are both monotone, but the overall sequence zigzags.

Now that we know each sequene is monotone, we are in a position similar to the previous chapter where we played two sequences off one another to learn about e. The same trick works to show they are bounded.

Example 11.6. The odd subsequence of s_n is bounded above, and the even subsequence is bounded below.

Proof. The even subsequece is monotone decreasing, but consists completely of positive terms. Thus, its bounded below by zero. Now we turn our attention to the odd subsequence: if n is odd, we know that s_n is bounded above by s_{n+1}, but s_{n+1} is a member of the monotone decreasing even subsequence, so $s_{n+1} < s_2 = 2$. Thus, for all odd n, s_n is bounded above by 2. □

Now we know by monotone convergence that both the even and odd subsequences converge! Next, we show they converge to the same value:

Example 11.7. Both the even and odd subsequences converge to the same value.

Proof. Let $e_n = s_{2n}$ be the even subsequence and $o_n = s_{2n-1}$ the odd subsequence, and write $\lim e_n = E$ and $\lim o_n = \Theta$. We wish to show $E = \Theta$.

Using the recurrence relation we see

$$o_{n+1} = 1 + \frac{1}{e_n} \qquad e_n = 1 + \frac{1}{o_n}$$

and so, using the limit laws and the convergence of e_n, o_n

$$\Theta = 1 + \frac{1}{E} \qquad E = 1 + \frac{1}{\Theta}$$

Therefore we see $\Theta - E = \frac{1}{E} - \frac{1}{\Theta}$, which after getting a common denominator implies

$$\Theta - E = \frac{\Theta - E}{\Theta E}$$

So whatever number $\Theta - E$ is, it has the property that it is unchanged when divided by the number ΘE. But the only number unchanged by multiplication and division is zero! Thus

$$\Theta - E = 0$$

□

Now we know that not only the even and odd subsequences converge but that they converge to the same limit! Its not too much more work to show that the entire sequence converges.

Example 11.8. The sequence s_n converges.

Proof. Call the common limit of the even and odd subsequences L. Let $\epsilon > 0$ Since $s_{2n-1} \to L$ we know there is an N_1 with $n > N_1$ implying $|s_{2n-1} - L| < \epsilon$. Similarly since $s_{2n} \to L$ we can find an N_2 where $n > N_2$ implies $|s_{2n} - L| < \epsilon$.

Set $N = \max\{N_1, N_2\}$. Then if $n > N$ we see both the even and odd subsequences are within ϵ of L by construction, and thus all terms of the sequence are within ϵ of L. But this is the definition of convergence! Thus s_n is convergent and $\lim s_n = L$. □

Finally! Starting with a zigzag sequence where monotone convergene did not apply, we broke it into two subsequences, each of which were monotone, and each of which we could prove converge. Then we showed these subsequences have the same limit and hence the overall sequence converges. We made it! Now its quick work to confirm the limit is what we expected from our construction: the golden ratio.

Example 11.9. The sequence s_n converges to the golden ratio.

Proof. Since throwing away the first term of the sequence does not change the limit, we see $\lim s_{n+1} = \lim s_n = L$. Using the recurrence relation and the limit laws, this implies

$$\lim s_{n+1} = \lim 1 + \frac{1}{s_n} = 1 + \frac{1}{L}$$

THus, the limit L satisfies the equation $L = 1 + 1/L$ or $L^2 = L + 1$. This has two solutions

$$\frac{1 \pm \sqrt{5}}{2}$$

Only one of which is positive. Thus this must be the limit

$$1 + \cfrac{1}{1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}}}}}}} = \frac{1+\sqrt{5}}{2}$$

s □

We can apply this same process to discover another sequence of rational approximations to $\sqrt{2}$, by algebraic means (in contrast wtih the geomeric approach of the babylonians). To start, we need to find a recursive formula that is satisfied by $\sqrt{2}$, and involves a reciprocal: something like

$$\sqrt{2} = \text{Rational stuff} + \frac{1}{\text{Rational stuff and } \sqrt{2}}$$

We can get such a formula through some trickery: first, using the difference of squares $a^2-b^2 = (a+b)(a-b)$ we see that $1 = 2-1 = (\sqrt{2}+1)(\sqrt{2}-1)$, which can be re-written

$$\sqrt{2} - 1 = \frac{1}{1+\sqrt{2}}$$

Now, substitute this into the obvious $\sqrt{2} = 1 + \sqrt{2} - 1$ to get

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

This is a *self-referential equation*, meaning $\sqrt{2}$ appears on both sides.

Example 11.10 (Continued Fraction of $\sqrt{2}$)**.** The continued fraction

$$1 + \cfrac{1}{2 + \cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cdots}}}}}$$

converges to the square root of 2.

Exercise 11.3. For any fixed n, prove that the following continued fraction exists, and find its vallue.

$$\cfrac{1}{n+\cfrac{1}{n+\cfrac{1}{n+\cfrac{1}{n+\cfrac{1}{n+\cfrac{1}{n+\cfrac{1}{n+\cdots}}}}}}}$$

Exercise 11.4 (Continued Fractions for Roots)**.** Let p be any prime number, find the continued fraction for $\sqrt{p}$.

Knowing such sequences is extremely useful for computation, in the age before computers: if n is a composite number we can find $\sqrt{n}$ by multiplying together the square roots of its prime factorization

Exercise 11.5. Find a rational approximation to $\sqrt{6}$ by calculating the first three terms in the continued fraction expansions for $\sqrt{2}$ and $\sqrt{3}$.

We could also find a contined fraction directly for cases like this, with a little more care:

Exercise 11.6. Find the continued fraction expansion for $\sqrt{pq}$ if p and q are primes. What happens to your procedure when $p = q$?

11.2. Subsequences and Convergence

Hopefully this exploration into continued fractions has shown the usefulness of looking for easy-to-work-with subsequences, when theorems such as monotone convergence don't automatically apply. It is then our gaol to try and piece this information back together: if we know the limits of various subsequences, what can we say about the entire sequence?

First of all, a simple example shows its not enough to say "if the even and odd subsequences converge, then the sequence converges".

Example 11.11. The sequence $s_n = (-1)^n$ diverges, but its even and odd subsequences form constant (thus convergent) subsequences:

$$s_{2n} = (-1)^{2n} = 1, 1, 1, \ldots$$
$$s_{2n+1} = (-1)^{2n-1} = -1, -1, -1, \ldots$$

Indeed, if you can find any two subsequences which limit to different values, then the sequence itself must diverge. This is a useful thing to try yourself when developing intuition.

Exercise 11.7. If a sequence s_n has two subsequences which converge to different values, then the overall sequence diverges.

The converse of this: that a sequence converges if all subsequences converge to the same limit - is trivial to prove because the entire sequence is a subsequence of itself, so we've actually assumed convergence! Thus, we get a nice characterization:

Theorem 11.1 (Convergence in terms of Subsequences)**.** *A sequence s_n converges if and only if all of its subsequences converge, and have the same limit.*

Remark 11.1. Its just as true if we even wish to *weaken the hypothesis to only include proper subsequences.* In this case, we could consider the proper subsequences of even and odd terms (for example): these converge to the same limit by hypothesis, and so via the argument in Example 11.8 we see the entire sequence converges.

Dealing with the collection of *all subsequences* is often technically difficult to do - so while this theorem statement is pretty clean sounding, its difficult to put into practice.

This showed us how to modify our original claim, about even and odd sequences however: so long as we assume they converge to the same limit, everything works.

Proposition 11.1. *If the even and odd subsequences of s_n converge to the same limit L, the s_n converges to L.*

Proof. This is actually done already in Example 11.8, when studying the golden ratio! □

Here we generalize this slightly, to any finite collection of subsequences.

Theorem 11.2. *Let s_n be a sequence, and assume that s_n is the union of N subsequences, all of which converge to the same limit L. Then s_n is convergent, with limit L.*

Sketch. One can prove this directly, but choosing useful notation is tedious. The idea is as follows: for each of the N sequences, let $M_1, M_2, \dots M_N$ be the threshold beyond which the subsequence is within ϵ of L for some fixed $\epsilon > 0$. Then set $M = \max\{M_1, \dots, M_N\}$ and note that for all $n > M$ each of the subsequences is within ϵ of L. Because the entire sequence is just the union of these N subsequences, this means that every term of the sequence is within ϵ of L. But this is precisely the definition of $s_n \to L$. So we are done. □

11.3. The Bolzano Weierstrass Theorem

What about sequences that *don't* converge? The theorem above says that it cannot be true that all their subsequences converge, but Example 11.11 does show that a divergent sequence can still contain convergent subsequences. A natural question then is - do they always? Alas, a simple counterexample shows us that is too much to ask:

Example 11.12. The sequence $s_n = n^2$ diverges, and all subsequences of it diverge.

But the problem here is not serious, its simply that the original sequence is *unbounded* and cannot possibly contain anything that converges. The perhaps surprising fact that this is the *only constraint preventing the existence of a convergent subsequence* is known as the Bolzano Weierstrass theorem.

Theorem 11.3 (Bolzano-Weierstrass)**.** *Every bounded sequence has a convergent subsequence*

There are many ways to prove this, but a particularly elegant one uses (of course!) the monotone convergence theorem.

At first this sounds suspicious: we must confront head on the issue we ran into above, that not every sequence is monotone! However, the weaker property we actually need *is* true: while not every sequence is monotone, every sequence contains a monotone subsequence. There is a very clever argument for this, which needs one new definition.

Definition 11.2 (Peak of a Sequence)**.** Let s_n be a sequence and $N \in \mathbb{N}$. Then s_N is a *peak* if it is larger than all following terms of the sequence:

$$s_N \geq s_m \;\; \forall m > N$$

Theorem 11.4 (Monotone Subsequences)**.** *Every sequence contains a monotone subsequence*

Proof. Let s_n be an arbitrary sequence. Then there are two options: either s_n contains infinitely many peaks or it does not.

If s_n contains infinitely many peaks, we can build the subsequence of all peaks. This is monotone decreasing: if p_1 is the first peak, then its greater than or equal to all subsequent terms s_n, and so its greater than or equal to the second peak p_2. (But, nothing here is special about 1 and 2, this holds for the n^{th} and $n+1^{st}$ peak without change).

Otherwise, if s_n contains only finitely many peaks, we will construct a monotone increasing subsequence as follows. Since there are finitely many peaks there must be a *last peak*, say this occurs at position N. Then s_{N+1} is not a peak, and we will take this as the first term of our new sequence (let's call it q_1). Because its not a peak, by definition there is some term farther down the sequence which is *larger* than s_{N+1} - say this happens at index N_2 and call it q_2. But q_2 is *also* not a peak (as there were only finitely many, and we are past all of them), so there's a term even farther down - say at index N_3 which is larger: call it q_3. Now we have $q_1 < q_2 < q_3$, and we can continue this procedure inductively to build a monotone increasing subsequence for all n. □

Now, given that every sequence has a monotone subsequence, we know that every *bounded* sequence has a *monotone and bounded subsequence.* Such things converge by MCT, so we know every sequence has a convergent subsequence!

We will not have much immediate use for this theorem in this or the following chapter, but in time will come to appreciate it as one of the most elegant tools available to us. There will come many times (soon, when dealing with functions) where we can easily produce a sequence of points satisfying some property, but to make progress we need a *convergent* sequence of such points. The BW theorem assures us that we don't have to worry - we can always make one by just throwing out some terms, so long as the sequence we have is bounded.

11.3.1. ⋆ An Alternative Proof of Bolzano-Weierstrass

An alternative argument for the BW theorem proceeds via the nested interval property. Here's an outline of how this works

- If s_n is bounded then there is some a, b with $a \leq s_n \leq b$ for all n. Call this interval I_0, and inductively build a sequence of nested closed intervals as follows
- At each stage $I_k = [a_k, b_k]$, bisect the interval with the midpoint $m_k = \frac{a_k+b_k}{2}$. This divides I_k into two sub-intervals, and since I_k contains infinitely many points of the sequence, one of these two halves must still contain infinitely many points. Choose this as the interval I_{k+1}.
- Now, this sequence of nested intervals has nonempty intersection by the Nested Interval Property. So, let $L \in \mathbb{R}$ be a point in the intersection.
- Now, we just need to build a subsequence of s_n which converges to L. We build it inductively as follows: let the first term be s_1, and then for each k choose some point $s_{n_k} \in I_k$ that is distinct from all previously chosen points (we can do this because there are infinitely many points available in I_k and we have only used finitely many so far in our subsequence).
- This new sequence is trapped between a_k and b_k, which both converge to L. Thus it converges by the squeeze theorem!

Exercise 11.8. In this problem, you are to check the main steps of this proof to ensure it works. Namely, given the above situation prove that

- If $I_k = [a_k, b_k]$, the sequences a_k and b_k of endpoints converge. *Hint: Monotone Convergence*
- $\lim a_k = \lim b_k$, so the Squeeze theorem really does apply *Hint: use that at each stage we are bisecting the intervals: can you find a formula for the sequence $b_k - a_k$, and prove this converges to zero?

12. Cauchy Sequences

> Highlights of this Chapter: We define the notion of a Cauchy sequence, and we prove that being Cauchy is equivalent to being convergent.

One reasonably ambitious sounding goal in the study of sequences is to find a nice criterion to determine exactly when a sequence converges or not. We made partial progress towards this in the previous two chapters, and our goal in this chapter is to provide an alternative complete characterization, by a single simple property. But what could such a property be? One (good!) thought is the following

> When a sequence converges, terms eventually get close to some limit L. Thus the terms of the sequence eventually get close to one another.

This condition is certainly necessary: if the terms of a sequence do *not* get close together, then they cannot get close to any limit! But is it sufficiently precise to actually work? For that we need to turn it into a mathematical definition: perhaps

> For all $\epsilon > 0$ there is an N where if $n > N$ then $|a_n - a_{n+1}| < \epsilon$

Unfortunately this doesn't quite seem to work: perhaps surprisingly, it *is possible* for consecutive terms of a sequence to all get within ϵ of one another, but for the overall sequence to diverge.

Example 12.1 ($|a_n - a_{n+1}|$ small but a_n diverges!)**.** Consider the sequence $a_n = \sqrt{n}$. Then for all $\epsilon > 0$ there is an N where $n > N$ implies $|\sqrt{n} - \sqrt{n+1}| < \epsilon$, but nonetheless a_n diverges (to infinity).

Proof. We can estimate the difference between consecutive terms with some algebra:

$$\begin{aligned}\sqrt{n+1} - \sqrt{n} &= \left(\sqrt{n+1} - \sqrt{n}\right)\frac{\sqrt{n+1}+\sqrt{n}}{\sqrt{n+1}+\sqrt{n}}\\ &= \frac{(n+1)-n}{\sqrt{n+1}+\sqrt{n}}\\ &= \frac{1}{\sqrt{n+1}+\sqrt{n}}\\ &< \frac{1}{\sqrt{n}}\end{aligned}$$

Thus for any $\epsilon > 0$ we can just take $N = \frac{1}{\epsilon^2}$ and see that if $n > N$ we have

$$|a_{n+1} - a_n| < \frac{1}{\sqrt{n}} < \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{\frac{1}{\epsilon^2}}} = \epsilon$$

Nowever, a_n is not converging to any finite number, as for any $M > 0$, if $n > M^2$ then $a_n = \sqrt{n} > M$, so $a_n \to \infty$ by Definition 8.4 □

Example 12.2 ($|a_n - a_{n+1}|$ small but a_n diverges, again!)**.** Perhaps the most famous example with this property is the *harmonic series*

$$a_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$$

Here it is clear that $a_n - a_{n+1} = \frac{1}{n+1}$ and we know this can be made smaller than any $\epsilon > 0$. However, as we will prove in CITE, this sequence nonetheless diverges to infinity.

So, we need to ask for a *stronger condition.* What went wrong? Well, even though we forced a_n to be close to a_{n+1} for all n, the small differences between consecutive terms could still manage to add up to *big differences* between terms: even if a_n was within 0.01 of a_{n+1} for all n, its totally possible that $a_{n+100,000}$ could differ from a_n by $(0.01)(10,000) = 100$! So, to strengthen our definition we might try to impose that *all terms of the sequence eventually stay close together*:

Definition 12.1 (Cauchy Sequence)**.** A sequence s_n is Cauchy if for all $\epsilon > 0$ there is a threshold past which any two terms of the sequence differ from one another by at most ϵ. As a logic sentence,

$$\forall \epsilon > 0 \ \exists N \ \forall m, n > N \ |s_n - s_m| < \epsilon$$

Example 12.3 (Cauchy Sequences: An Example)**.** The sequence $s_n = \frac{1}{n}$ is cauchy: we can see this because for any n, m

$$\left|\frac{1}{n} - \frac{1}{m}\right| < \left|\frac{1}{n} - 0\right| = \frac{1}{n}$$

And we already know that for any ϵ we can choose N with $n > N$ implying $1/n < \epsilon$.

Example 12.4 (Cauchy Sequences: A Nonexample)**.** The sequence $s_n = 1, 0, 1, 0, 1, 0 \ldots$ is not Cauchy, as the difference between any two consecutive terms is 1. Thus for $\epsilon = 1/2$ there is no N where past that N, every s_n is within $1/2$ of each other.

Exercise 12.1. Is the sequence $s_n = 1 - \frac{(-1)^n}{n}$ cauchy nor not? Prove your claim.

Exercise 12.2. Let s_n be a periodic sequence (meaning after some period P we have $s_n = s_{P+n}$ for all n). Prove that if s_n is Cauchy then it is constant. *Hint: what's the contrapositive?*

12.1. ⋆ Properties of Cauchy Sequences

A good way to get used to a new definition is to *use it.* This definition looks very similar to the limit definition, which means we can often formulate analogous theorems and proofs to things we've seen before:

Note the proofs in this section are not *logically required* as the next section will render them superfluous: once we know Cauchy and convergent are equivalent, these all follow as immediate corollaries of the limit laws! Nonetheless it is instructive to see their direct proofs:

Proposition 12.1 (Cauchy Implies Bounded)**.** *If s_n is Cauchy then its bounded: there exists a B such that $|s_n| < B$ for all $n \in \mathbb{N}$.*

Very similar to proof for convergent seqs Proposition 9.2 in style, where we show after some N all the terms are bounded by some particular number, and then take the max of this and the (finitely many!) previous terms to get a bound on the entire sequence. :::{.proof} Set $\epsilon = 1$. Since a_n is Cauchy we know there is some N beyond which $|a_n - a_m| < 1$ for all $n, m > N$. In particular, this means every |a_n-a_{N+1}|<1$ so

$$a_{N+1} - 1 < a_n < a_{N+1} + 1$$

Thus for the (infinitely many terms!) after a_N, we can bound all of them above by $a_{N+1} + 1$ and below by $a_{N+1} - 1$. To extend these to bounds for the whole sequence, we just take the max or min with the (finitely many!) previous terms:

$$L = \min\{a_1, a_2, \dots, a_N, a_{N+1} - 1\}$$

$$U = \max\{a_1, a_2, \dots, a_N, a_{N+1} + 1\}$$

Now we have for all n, $L \leq a_n \leq U$ so $\{a_n\}$ is bounded. :::

Proposition 12.2 (Sums of Cauchy Sequences)**.** *If a_n and b_n are Cauchy sequences, so is $a_n + b_n$.*

Proof. Let $\epsilon > 0$. Then choose N_a and N_b such that for all n, m greater than N_a, N_b respectively, we have $|a_n - a_m| < \epsilon/2$ and $|b_n - b_m| < \epsilon/2$. Set $N = \max\{N_a, N_b\}$ and let $n, m > N$. Then each of the above two inequalities hold, and so by the triangle inequality

$$|(a_n + b_n) - (a_m - b_m)| = |(a_n - a_m) + (b_n - b_m)|$$
$$\leq |a_n - a_m| + |b_n - b_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Thus, $a_n + b_n$ is Cauchy as well. □

Exercise 12.3 (Constant Multiples of Cauchy Sequences)**.** Let a_n be Cauchy, and $k \in \mathbb{R}$ be constant. Then ka_n is Cauchy.

Proposition 12.3 (Products of Cauchy Sequences)**.** *Let a_n, b_n be Cauchy. Then $s_n = a_nb_n$ is a cauchy sequence.*

First, some scratch work: we are going to want to work with the condition $|s_n - s_m| = |a_nb_n - a_mb_m|$. But we only know things about the quantities $|a_n - a_m|$ and $|b_n - b_m|$. So, we need to do some algebra, involving *adding zero in a clever way* and *applying the triangle inequality*:

$$\begin{aligned}
|a_nb_n - a_mb_m| &= |a_nb_n + (a_nb_m - a_nb_m) - a_mb_m| \\
&= |(a_nb_n - a_nb_m) + (a_nb_m - a_mb_m)| \\
&= |a_n(b_n - b_m) + b_m(a_n - a_m)| \\
&\leq |a_n(b_n - b_m)| + |b_m(a_n - a_m)| \\
&= |a_n||b_n - b_m| + |b_m||a_n - a_m|
\end{aligned}$$

Because we know Cauchy sequences are bounded, we can get upper estimates for both $|a_n|$ and $|b_n|$. And then as we know that the sequences are Cauchy, we can make $|a_n - a_m|$ and $|b_n - b_m|$ as small as we need, so that this overall term is small. We carry this idea out precisely in the proof below.

Proof. Let a_n and b_n be Cauchy, and choose an $\epsilon > 0$. Then each are bounded, so we can choose some M_a with $|a_n| < M_a$ and M_b where $|b_n| < M_b$ for all n. To make notation easier, set $M = \max\{M_a, M_b\}$ so that we know both a_n and b_n are bounded by the same constant M.

Using that each is Cauchy, we can also get an N_a and N_b where if n, m are greater than these respectively, we know that

$$|a_n - a_m| < \frac{\epsilon}{2M} \qquad |b_n - b_m| < \frac{\epsilon}{2M}$$

Then set $N = \max\{N_a, N_b\}$, and choose arbitrary $n, m > N$. Since in this case both of the above hypotheses are satisfied, we know that

$$|a_n||b_n - b_m| \leq M\frac{\epsilon}{2M} = \frac{\epsilon}{2} \qquad |b_m||a_n - a_m| \leq M\frac{\epsilon}{2M} = \frac{\epsilon}{2}$$

Together, this means their sum is less than ϵ, and from our scratch work we see their sum is already an upper bound for the quantity we are actually interested in:

$$|a_nb_n - a_mb_m| \leq |a_n||b_n - b_m| + |b_m||a_n - a_m| \leq \epsilon$$

□

Exercise 12.4 (Reciprocals of Cauchy Sequences)**.** Let a_n be a Cauchy sequence with $a_n \neq 0$ for all n, which does not converge to zero. Then the sequence of reciprocals $s_n = \frac{1}{a_n}$ is Cauchy.

Just like for convergence, once we know the results products and reciprocals, quotients follow as an immediate corollary:

Corollary 12.1 (Quotients of Cauchy Sequences)**.** *If a_n and b_n are Cauchy with $b_n \neq 0$ and* $\lim b_n \neq 0$ *then the quotients a_n/b_n form a Cauchy sequence.*

Exercise 12.5. Show the hypothesis $b_n \not\to 0$ is necessary in Corollary 12.1 by giving an example of two Cauchy sequences a_n, b_n where $b_n \neq 0$ for all n, yet $\frac{a_n}{b_n}$ is *not* a Cauchy sequence.

12.2. Cauchy $\iff$ Convergent

Now we move on to the main act, where we prove convergence is equivalent to Cauchy by showing an implication in both directions.

Exercise 12.6 (Convergent Implies Cauchy)**.** If s_n is a convergent sequence, then s_n is Cauchy. *Hint: The triangle inequality and $|a_n - a_m|$ for a sequence converging to L can tell you….what?*

More difficult, and more interesting, is the converse:

Proposition 12.4 (Cauchy Implies Convergent)**.** *If s_n is a Cauchy sequence, then s_n is convergent.*

Proof. Let s_n be a Cauchy sequence. Then it is bounded, by Proposition 12.1, so by the Bolzano Weierstrass theorem (**?@thm-thm-bolzano-weierstrass**) we can extract a subsequence s_{n_k} which converges to some real number L.

Now we have something to work with, and all we need to show is that the rest of the sequence also converges to L. So, let's fix an $\epsilon > 0$. Since $s_{n_k} \to L$ there exists an N_1 where if $n_k > N_1$ we know $|s_{n_k} - L| < \epsilon/2$. And, since s_n is Cauchy, we know there is an N_2 where for any $n, m > N_2$ we know $|s_n - s_m| < \epsilon/2$.

Let $N = \max\{N_1, N_2\}$, and choose any $n > N$. If s_n is in the subsequence, we are good because $n > N_1$ and we know for such elements of the subsequence $|s_n - L| < \epsilon/2 < \epsilon$. But if s_n is not in the subsequence, choose any m such that $m > N$ and s_m is in the subsequence, and apply the triangle inequality:

$$|s_n - L| = |s_n - s_m + s + m - L| \leq |s_n - s_m| + |s_m - L| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Where the first inequality is because of the Cauchy condition, and the second is the convergence of the subsequence. □

Together these imply the main theorem we advertised.

Theorem 12.1 (Cauchy $\iff$ Convergent)**.** *The conditions of being a Cauchy sequence and a convergent sequence are logically equivalent.*

If a sequence converges, then every subsequence converges to the same limit (Theorem 11.1). This has a nice application: if you can *find* any subsequence where it's easier to compute the values, you can use that subsequence to compute the limit.

Exercise 12.7. Prove directly from the definition of Cauchy: if s_n is Cauchy and s_{n_k} is a subsequence whose limit is L then $s_n \to L$.

Part IV.

Functions

13. Definition

> Highlights of this Chapter: we briefly explore the evolution of the modern conception of a function, and give foundational definitions for reference.

While sequences may be the main *tool* of real analysis, *functions* are its main object of study. The term *function* was first introduced to mathematics by Leibniz during his development of the Calculus in the 1670s (he also introduced the idea of *parameters* and *constants* familiar in calculus courses to this day). In the first centuries of its mathematical life, the term function usually denoted what we would think of today as a *formula* or *algebraic expression*. For example, Euler's definition of function from his 1748 book *Introductio in analysin infinitorum* embodies the sentiment:

> A function of a variable quantity is *an analytic expression* composed in any way whatsoever of the variable quantity and numbers or constant quantities.

As a first step to adding functions to our theory of real analysis, we would somehow like to make this definition rigorous. But upon closer inspection, this concept, of "something expressible by a (single) analytic expression" is actually logically incoherent! For example, say that we decide, after looking at the definition of $|x|$, that it cannot be a function as it is not expressed as a single formula:

$$|x| = \begin{cases} -x & x \leq 0 \\ x & x > 0 \end{cases}$$

But we also agree that x^2 and $\sqrt{x}$ are both (obviously!) functions as they are given by nice algebraic expressions. What are we then to make of the fact that for all real numbers x,

$$\sqrt{x^2} = |x|$$

It seems we have found a perfectly good "single algebraic expression" for the absolute value after all! This even happens for functions with infinitely many pieces (which surely would have been horrible back then)

$$f(x) = \begin{cases} \vdots & \vdots \\ 3 + \sin(x) & x \in (0, \pi] \\ 1 + \sin(x) & x \in (\pi, 2\pi] \\ 3 + \sin(x) & x \in (2\pi, 3\pi] \\ \vdots & \vdots \end{cases}$$

This can be written as a composition involving just one piecewise function

$$f(x) = |1 + \sin x| + 2$$

Which can, by the earlier trick, be reduced to a function with no "pieces" at all:

$$f(x) = 2 + \sqrt{1 + 2\sin(x) + \sin^2(x)}$$

So the idea of "different pieces" or different rules, seemingly so clear to us, is not a good mathematical notion at all! We are *forced* by logic to include such things, whether we aimed to or not. This became clear rather quickly, as even Euler had altered a bit his notion of functions by 1755:

> When certain quantities depend on others in such a way that they undergo a change when the latter change, then the first are called functions of the second. This name has an extremely broad character; it encompasses all the ways in which one quantity can be determined in terms of others.

The modern approach is to be much more open minded about functions, and define a function as *any rule whatsoever* which uniquely specifies an output given an input. This seems to have first been clearly articulated by Lobachevsky (of hyperbolic geometry fame) in 1834, and independently by Dirichlet in 1837

> The general concept of a function requires that a function of x be defined as a number given for each x and varying gradually with x. The value of the function can be given either by an analytic expression, or by a condition that provides a means of examining all numbers and choosing one of them; or finally the dependence may exist but remain unknown. (Lobachevsky)

> If now a unique finite y corresponding to each x, and moreover in such a way that when x ranges continuously over the interval from *a*$ to b, $y = f(x)$ also varies continuously, then y is called a *continuous function of* x for this interval. It is not at all necessary here that y be given in terms of x by one and the same law throughout the entire interval, and it is not necessary that it be regarded as a dependence expressed using mathematical operations. (Dirichlet)

13.1. Definition and Examples

Through this definitions added generality comes *simplicity*: we are not trying to poliece what sort of rules can be used to define a function, and so the notion can be efficiently captured in the language of sets and logic.

Definition 13.1. A function from a set X to a set Y is an assignment to each element of X a *unique* element of Y. If we call the function f, we write the unique element of Y assigned to $x \in X$ as $y = f(x)$, and the entire function as

$$f : X \to Y$$

The definition of a function comes with three parts, so its good to have precise names for all of these.

Definition 13.2. If f is a function, its input set X is called the *domain*, and the set of possible outputs Y is called the *codomain*. The set of *actual outputs*, that is $R = \{f(x) \mid x \in X\}$ is called the *range*.

If the codomain of a function f is the real numbers, we call f a **real-valued function**. We will be most interested in real valued function throughout this course.

Now, because the definition itself is conspicuously quiet on *what a function looks like*, it's good to start ourselves off (as usual) with a collection of examples and non-examples. First, an example

Example 13.1 ($y = x^2$ is a function)**.** The assignment taking every real number x to the real number x^2 is a function: for every input, there is exactly one output. We write this $f(x) = x^2$, and note the domain and codomain are $\mathbb{R}$, and the range is just the nonnegative reals.

This example comes with a *formula*, much like Euler had hoped for: it tells us exactly what to do with our input to get the output - multiply it by itself! This example immediately generalizes to a whole host of functions-defined-by-formulas, by using the field operations of $\pm$ and $\times$:

Example 13.2 (Polynomial Functions)**.** A polynomial function is an assignment $p : \mathbb{R} \to \mathbb{R}$ which takes each x to a linear combination of powers of x:

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots a_1 x + a_0$$

The highest power of x appearing in p is called the *degree* of the polynomial.

The idea of a *function defined by a formula* can be extended even farther by allowing the field operation of division; though this time we must be careful about the inputs.

Example 13.3 (Rational Functions)**.** A rational function is a an assignment

$$f(x) = \frac{p(x)}{q(x)}$$

where p and q are polynomials. Rational functions are real-valued, but their domain is not all of $\mathbb{R}$: at any zero of q the formula above is undefined, a rational function is only defined on the set of points where q is nonzero.

We already saw that piecewise formulas count in our modern definition, but perhaps didn't fully think through the implications: they can be *very, very piecewise*

Example 13.4 (The Characteristic Function of $\mathbb{Q}$)**.** The function $f : \mathbb{R} \to \mathbb{R}$ defined as follows

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Here's another monstrous piecewise function we will encounter again soon:

Example 13.5 (Thomae's Function)**.** This is the function $\tau : \mathbb{R} \to \mathbb{R}$ defined by

$$\tau(x) = \begin{cases} \frac{1}{q} & x \in \mathbb{Q} \text{ and } \frac{p}{q} \text{ is lowest terms.} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

We've stressed that functions don't need to be given by explicit formulas, so we should give an example of that: here's a function that is defined at each point as a different limit (using the completeness axiom)

Example 13.6. The exponential function may be defined for each $x \in \mathbb{R}$ by the following limit

$$\exp(x) = \lim_{n\to\infty} a_n$$

Where a_n is the recursive sequence $a_0 = 1$, $a_n = a_{n-1} + \frac{x^n}{n!}$.

A function can also be defined by an *existence proof* telling us that a certain relationship determines a function, without giving us any hint on how to compute its value:

Example 13.7 ($\sqrt{\cdot}$ defined by an existence theorem)**.** We proved that for every $x \geq 0$ that there exists some number $y > 0$ with $y^2 = x$, back in our original study of completeness (Theorem 6.9).

We can easily see that such a number is *unique*: if $y_1 \neq y_2$ then by the order axioms one is greater: without loss of generality $0 < y_1 < y_2$. Thus $y_1^2 < y_2^2$, so we can't have both $y_1^2 = x$ and $y_2^2 = x$, and $x \to y = \sqrt{x}$ is a function.

Alright - that's plenty of examples to get ourselves in the right mindset. Let's give a non-example, to remind us that while there need not be formulas, the modern notion of function is not '*anything goes*'!

Example 13.8. The assignment taking an integer to one of its prime factors does *not* define a function. This would take the integer 6 to both 2 and 3, and part of the definition of a function is that the output is *unique* for a given input.

13.2. Composition and Inverses

Likely familiar from previous math classes, but it is good to get rigorous definitions down on paper when we are starting anew.

Definition 13.3 (Composition)**.** If $f: X \to Y$ and $g: Y \to Z$ then we may use f to send an element of X into Y, and follow it by g to get an element of Z. The result is a function from X to Z, known as the *composition*

$$g \circ f: X \to Z \qquad g \circ f(x) := g(f(x))$$

Every set has a particularly simple function defined on it known as the *identity function*: $\mathrm{id}_X : X \to X$ is the function that takes each element $x \in X$ and *does nothing*: $\mathrm{id}_X(x) = x$. These play a role in concisely defining inverse functions below:

Definition 13.4 (Inverse Functions)**.** If $f: X \to Y$ is a function, and $g: Y \to X$ is another function such that

$$g \circ f = \mathrm{id}_X \qquad f \circ g = \mathrm{id}_Y$$

Then f and g are called *inverse functions* of one another, and we write $g = f^{-1}$ if we wish to think of g as inverting f, or $f = g^{-1}$ rather we started with g, and think of f as undoing it.

Example 13.9. The function $f(x) = 2x$ and $g(x) = x/2$ are inverses of one another as functions $\mathbb{R} \to \mathbb{R}$.

The squaring function $s: \mathbb{R} \to \mathbb{R}$ defined by $s(x) = x^2$ has the square root as an inverse, only if the domain and codomain are restricted to the nonnegative reals. Otherwise, we see that $s(-2) = 4$ and $\sqrt{4} = 2$ so $\sqrt{\ } \circ s$ is not the identity: it takes -2 to 2!

13.3. Increasing, Decreasing & Convexity

Finally we end our introductory march through definitions with several that make sense for functions on ordered fields, but not necessarily for general functions.

Definition 13.5 (Increasing Functions)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is *increasing* if whenever $x < y$, it follows that $f(x) \leq f(y)$. A function is *strictly increasing* if this inequality is strict ($<$).

Definition 13.6 (Decreasing Functions)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is *decreasing* if whenever $x < y$, it follows that $f(x) \geq f(y)$. A function is *strictly increasing* if this inequality is strict ($>$).

Definition 13.7 (Monotone)**.** A function is monotone if it is either increasing or decreasing.

Definition 13.8 (Convexity)**.** Let f be a function defined on some interval (possibly all of $\mathbb{R}$). Then f is convex if for any interval $[x, y] \subset \mathrm{dom} f$, the value of f at the midpoint exceeds the average value of f at the endpoints:

$$\forall x, y \qquad f\left(\frac{x+y}{2}\right) \geq \frac{f(x)+f(y)}{2}$$

Exercise 13.1. Prove that if f is convex then for any x, y in the domain, the the secant line connecting $f(x)$ to $f(y)$ lies above the graph of f.

Hint: the equation of secant line is $L(t) = tf(x) + (1-t)f(y)$: so need to show $L(t) \geq f(t)$.

Proposition 13.1. *If f is a convex function and $a \in \mathbb{R}$ Then for $x < a$ the function*

$$\ell(x) = \frac{f(x) - f(a)}{x - a}$$

is monotone increasing.

Proof.

□

The same is true for $x > a$: on this domain the difference quotient also defines a monotone increasing function (so, its monotone *decreasing* when going "backwards" towards a).

13.4. ⋆ Functions of a Real Variable

Our main concern in this class will be *real valued* functions of a *real variable*, meaning the input and output are both real numbers.

However the theory of real analysis allows one easy generalization of this: we can consider functions of more complicated outputs so long as we can understand what *convergence* means in the range.

For example: the complex numbers $\mathbb{C}$ are pairs of real numbers, functions $f : \mathbb{R} \to \mathbb{C}$ are just pairs of real valued functions. We can define *convergence* for complex numbers to mean that each component converges (as a real sequence), and then discuss *continuity* of a complex-valued function

Even more generally, we could look at *vector valued functions* or even *matrix valued functions* of a real variable, if we define convergence component-wise.

So, functions of a more complicated range can be easily incorporated into our theory. Its only when the *domain* gets more complicated that analysis needs a real extension: to complex analysis, or multivariate calculus. We will not discuss these topics in this course.

14. Continuity

Highlights of this Chapter: we formalize the concept of continuity, one of the foundational definitions in the analysis of functions. We provide an equivalent definition built out of sequences, and use it to prove 'continuity analogs' of the limit theorems. Finally, we prove that continuous functions are determined by their values on a dense set, an oft-useful result allowing one to reduce various arguments to considerations about rational numbers.

What does continuity mean? In pre-calculus classes, we often first hear something like "you can draw the graph without picking up your pencil". This is a good guide to start with for a formal definition: its clearly capturing some property that is easy to check by visual inspection! But it's not precise: terms like "you" and "pencil", as well as modal phrases like "can draw" are nowhere to be found in the axioms of ordered fields! How can we say the same thing, using words we have access to?

First, a function is an input-output machine, so we should rephrase things in terms of inputs and outputs. When a graph makes a jump (where you'd have to pick up your pencil), the output changes a lot even when the input barely does. Thus, not having to pick up your pencil means you change the input by a little bit, the output changes by a little bit.

This is totally something we can make precise! A good start is by giving names to things: we want to say for any change in the input smaller than some δ, we know the output cant change that much: maybe its maximum is some other small change ϵ:

Definition 14.1 (ϵ-δ continuity). A function f is continuous at a point a in its domain if for every $\epsilon > 0$ there is some threshold δ where if x is within δ of a, then $f(x)$ is within ϵ of $f(a)$. As a logic sentence:

$$\forall \epsilon > 0\, \exists \delta > 0\, \forall x\, |x - a| < \delta \implies |f(x) - f(a)| < \epsilon$$

A function is continuous on a set $X \subset \mathbb{R}$ if it is continuous at a for each $a \in \mathbb{R}$. A function is *continuous* if it is continuous on its domain.

14.1. Using the $\epsilon - \delta$ Definition

This definition looks a lot like the sequence definition, at least in terms of the order of the quantifiers. And this is a good thing for us, who are now experts at the sequence definition!

Example 14.1. Any constant function $f(x) = c$ is continuous at every real number a.

Example 14.2. The function $y = x$ is continuous at every real number a.

This generalizes directly to functions like $f(x) = 2x + 1$, where now for a fixed ϵ we may wish to take $\delta = \epsilon/2$ after some scratch work:

Exercise 14.1. Show that linear functions $y = mx + b$ are continuous at every $a \in \mathbb{R}$.

Exercise 14.2. Prove that the function $f(x) = |x|$ is continuous at $x = 0$. Then use the fact that $f(x) = x$ for $x > 0$ and $f(x) = -x$ for $x < 0$ (which are linear functions) to conclude that $|x|$ is continuous at every real number.

Like any definition, its good after seeing a few examples to also turn and look at non-examples:

Example 14.3. The step function

$$h(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

is discontinuous at 0, but is continuous at all other real numbers.

Thus, a function with a jump in it is discontinuous right at the jump, as we expect. This shows its possible for a function to be *discontinuous* at a single point, but things can get much stranger!

Example 14.4. The characteristic function of the rational numbers is discontinuous *everywhere.*

$$b(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

We saw above a function that is discontinuous at a single point, and then one that is discontinuous everywhere. What's harder to imagine, is a function that is *continuous at a single point.* Try thinking about what this might mean!

Exercise 14.3. Show that the following function is continuous at 0 and discontinuous everywhere else:

$$g(x) = \begin{cases} x & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

While the $\epsilon - \delta$ definition is nice in that it *looks* like the sequence definition, we still end up having to play the ϵ game with every argument. Indeed, while some functions are well-suited these, for other relatively simple looking arguments, picking the right δ actually turns out to be a bit of work!

Exercise 14.4. Prove that $f(x) = x^2$ is a continuous function using the $\epsilon - \delta$ definition.

To avoid having to do such hard work on a regular basis, we will seek to broaden our theoretical toolkit.

14.2. Continuity With Sequences

We spent a lot of time working with sequences so far, so it would be nice if we could leverage some of that knowledge as more than just analogy. And indeed we can! In this section, we introduce an alternative definition of continuity, and prove that it is equivalent to our original.

Definition 14.2 (Continuity)**.** Let f be a real valued function with domain $D \subset \mathbb{R}$ and $a \in D$ a point. Then f is continuous at a if for every convergent sequence $\{x_n\} \subset D$ with $x_n \to a$, the limit can be taken either before or after applying f:

$$\lim f(x_n) = f(\lim x_n) = f(a)$$

A function is continuous on a set $S \subset D$ if it is continuous at each point of S.

Thus, we can think of continuity as the condition that allows us to "pull the limit inside of f". It is immediate from the definition that constant functions are continuous at every point of their domain, as is the function $f(x) = x$.

Example 14.5. The function $f(x) = x^2$ is continuous on the entire real line.

Proof. Let $a \in \mathbb{R}$ be arbitrary, and let x_n be an arbitrary sequence converging to a. Then by the limit theorem for products, we see that since $x_n \to a$, it follows that $x_n^2 \to a^2$. Thus, if $f(x) = x^2$ we have

$$\lim f(x_n) = \lim x_n^2 = a^2 = f(a) = f(\lim x_n)$$

So, f is continuous at $x = a$. Since a was an arbitrary real number, f is continuous on the entire real line. □

Theorem 14.1 (Equivalence of Continuity Definitions)**.** *Let f be a real function, and a a point of its domain. Then f is continuous by the sequence definition if and only if it is continuous by the ϵ-δ definition.*

This theorem is an *equivalence of definitions* or an *if-and-only-if* result, so the proof requires two parts: first we show that continuity implies sequence continuity, and then we show the converse.

Continuity Implies Sequence Continuity. Let f be continuous at a, and x_n an arbitrary sequence converging to a. We wish to show the sequence $f(x_n)$ converges to $f(a)$. Choosing an $\epsilon > 0$, we use the assumed continuity to get a $\delta > 0$ where $|x - a| < \delta$ implies that $|f(x) - f(a)| < \epsilon$.

But since $x_n \to a$, we know there must be some N such that for $n > N$ we have $|x_n - a| < \delta$: thus for this same N we have $|f(x_n) - f(a)| < \epsilon$.

Putting this all together, this is just the definition of convergence for the sequence $f(x_n)$ to $f(a)$: starting with $\epsilon > 0$ we got an N which for $n > N$ we can guarantee $|f(x_n) - f(a)| < \epsilon$. So we are done. □

Sequence Continuity Implies Continuity. Here we prove the contrapositive: that if f is *not continuous* at a then it is also *not sequence continuous* there.

If f is not continuous at a then there is some ϵ where for every $\delta > 0$ we can find points within δ of a where $f(x)$ is more than ϵ away from $f(a)$. From this we need to somehow produce a *sequence*, so we will take a sequence of such δ's and for each pick some such bad point x.

For example, if we let $\delta = 1/n$ then call x_n the point with $|x_n - a| < 1/n$ but $|f(x_n) - f(a)| > \epsilon$. Doing this for all n produces a sequence where

$$a - \frac{1}{n} < x_n < a + \frac{1}{n}$$

And so by the squeeze theorem we see that x_n converges, and its limit is a. But we also know (by our choices of x_n) that for *every element of this sequence* |f(x_n)-f(a)|>⊠$, so there's no way that $f(x_n)$ converges to $f(a)$.

Thus, we've shown by example that our function is not sequence continuous at a, as required. □

When working with this definition of continuity, its important to remember that we need to check $f(\lim x) = \lim f(x_n)$ for *all sequences* $x_n \to a$. If it fails for any individual sequence, that is enough to show the function is not continuous at that point. Thus when proving continuity we will always start with *let* x_n *be an arbitrary sequence converging to* a, and make use of convergence theorems to help us (since we cannot know the particular sequence), whereas for proving discontinuity all we need to do is produce a specific example sequence that fails.

Exercise 14.5. The function

$$\operatorname{sgn}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

is discontinuous at $x = 0$, but continuous at every other real number.

Its useful to have two definitions, as often one will be easier to use than the other. Below we will see *many* examples where sequence continuity is easier to apply, but here's an example where $\epsilon - \delta$ continuity makes things clearer.

Proposition 14.1. *Let f be a continuous function, and assume that $f(c) \neq 0$ for some point $c \in \mathbb{R}$. Then there exists a small interval $(c - \delta, c + \delta)$ on which $f(x) \neq 0$ for all x in the interval.*

Proof. Let $\epsilon = |f(c)|/2$. Then by continuity, there is some δ such that if $|x - c| < \delta$ we know $|f(x) - f(c)| < \epsilon$. Unpacking this, for all $x \in (c - \delta, c + \delta)$ we know

$$-\epsilon = \frac{-|f(c)|}{2} < f(x) - f(c) < \frac{|f(c)|}{2} = \epsilon$$

And thus

$$f(c) - \frac{|f(c)|}{2} < f(x) < f(c) + \frac{|f(c)|}{2}$$

If $f(c)$ is positive, then the lower bound here is $f(c)/2$ which is still positive, so $f(x)$ is always positive in the interval. And, if $f(c)$ is negative, the upper bound here is $f(c)/2$ which is still negative: thus $f(x)$ is always negative in the interval. □

14.3. Analogs of the Limit Theorems

Beause we have an equivalent characterization of continuity in terms of sequence convergence, and we have many theorems about this, we can use our characterization to rephrase these as results about *continuity.*

Proposition 14.2 (Continuity of Multiples). *If f is continuous at $a \in \mathbb{R}$ and $k \in \mathbb{R}$ is a constant, then the function $kf : x \mapsto kf(x)$ is continuous at a.*

Proof. Let $a \in \mathbb{R}$ be arbitrary, and x_n a sequence converging to a. Then by the limit theorem for multiples, $kx_n \to ka$. Rephrasing this in terms of the function $f(x) = kx$, this just says that $\lim f(x_n) = f(\lim x_n)$ so f is continuous at a. □

Theorem 14.2 (Continuity of Field Operations). *Let f, g be functions which are continuous at a point a. Then the functions $f(x) + g(x)$, $f(x) - g(x)$ and $f(x)g(x)$ are all continuous at a. Furthermore if $g(a) \neq 0$ then $f(x)/g(x)$ is also continuous at a.*

Proof. Let f, g be any two continuous functions and let $a \in \mathbb{R}$ be a point in their domains. Let x_n be any sequence converging to a. Since f is continuous we know that $\lim f(x_n) = f(\lim x_n) = f(a)$ and similarly by the continuity of g, $\lim g(x_n) = f(\lim x_n) = g(a)$. Thus by the limit theorem for sums, the sequence $f(x_n) + g(x_n)$ is convergent, with

$$\lim (f(x_n) + g(x_n)) = \lim f(x_n) + \lim g(x_n) = f(a) + g(a)$$

So, $f + g$ is continuous at a. Since a was arbitrary, we see that $f + g$ is continuous at every point of its domain. The same argument applies for subtraction, multiplication, and division using the respective limit theorems for sequences. □

One of the most important operations for functions is that of *composition*: if $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ then the function $g \circ f : \mathbb{R} \to \mathbb{R}$ is defined as $g \circ f(x) := g(f(x))$. More generally, so long as the domain of g is a subset of the range of f, the composition $g \circ f$ is well defined.

Theorem 14.3 (Continuity of Compositions). *Let f, g be functions such that f is continuous at a, and g is continuous at $f(a)$. Then the composition $g \circ f(x) := g(f(x))$ is continuous at a.*

Proof. Let x_n be an arbitrary sequence converging to $a \in \mathbb{R}$: we wish to show that $\lim g(f(x_n)) = g(f(\lim x_n)) = g(f(a))$. Since f is continuous at $x = a$ we see immediately that $f(x_n)$ is a convergent sequence with $f(x_n) \to f(a)$. And now, since g is assumed to be continuous at $x = f(a)$ and $f(x_n)$ is a sequence converging to this point, we know $g(f(x_n)) = g(f(a))$ as required. □

Exercise 14.6. Let $f(x)$ be a continuous function, and assume that $f(x)^2$ is a constant function. Prove that $f(x)$ is constant.

Give an example of an $f(x)$ where $f(x)^2$ is constant, but f is not.

Theorem 14.4 (Continuity of Roots)**.** *The function* $R(x) = \sqrt{x}$ *is continuous on* $[0, \infty)$.

Proof. Actually we already proved this, before we had the terminology! Re-read **?@exr-limit-of-root**: it shows that if $x_n \to a$ is a convergent sequence with $x_n \geq 0$ and $a \geq 0$, then $\sqrt{x_n} \to \sqrt{a}$. So $\lim \sqrt{x_n} = \sqrt{\lim x_n}$, and $\sqrt{x}$ is continuous at the arbitrary nonnegative real a. □

The same is true for n^{th} roots, though we do not stop to prove it here, you may wish to for practice!

14.4. Useful Examples

Because continuity is going to be a big part of our course, its good to have a couple examples of functions we already know to be continuous. The ones below are particularly useful:

Exercise 14.7 (Continuity of Polynomials)**.** Prove that every polynomial is a continuous function on the entire real line. *Hint: induction on the degree of the polynomial!*

Exercise 14.8 (Continuity of Rational Functions)**.** A rational function is a quotient of polynomials $r(x) = p(x)/q(x)$. Prove that every rational function is continuous, on every point of its domain.

Exercise 14.9. If f is continuous at a point a, then $|f|$ is continuous there.

*Hint: either use the reverse triangle inequality (**?@exr-reverse-triangle-inequality**) or use that its a composition*

Exercise 14.10 (Continuity of Max and Min)**.** Prove that for any two numbers x, y we can express the max and min by the following formulas:

$$FORMULAS$$

Use this, together with the limit theorems on field operations and continuity to prove that for any two continuous functions $f(x), g(x)$ that the functions $\max\{f(x), g(x)\}$ and $\min\{f(x), g(x)\}$ are continuous.

Putting all this together, we already can build many examples of continuous functions! For example,

$$\frac{|3x^2 + 2x - 1|^7 \max\{7x, 2x^2 + 3\}}{4(3x^2 + 11)}$$

Exercise 14.11. Prove carefully that the above function is continuous at every $a \in \mathbb{R}$, using the theorems and examples developed above.

14.5. Continuity and the Rationals

Before closing the introductory chapter on continuity, we turn to one important theoretical tool: the density of the rationals. Because every real number is the limit of a sequence of rationals, and continuous functions are determined by limits, it seems that continuous functions are rather constrained by their value on the rationals. This is indeed true, and will prove quite useful: we prove it in two steps below.

Proposition 14.3. *If f is a continuous function such that $f(r) = 0$ for every rational number r, then $f = 0$ is the zero function.*

Proof. Let f be such a function, and $a \in \mathbb{R}$ any real number. Then there is a sequence r_n of rational numbers converging to a. Given that f is zero on all rationals, we see that $f(r_n) = 0$ for all n. Thus $f(r_n)$ is the *constant* zero sequence, and so its limit is zero:

$$\lim f(r_n) = \lim 0 = 0$$

But, since f is assumed to be continuous, we know that we can move the limit inside of f:

$$0 = \lim f(r_n) = f(\lim r_n) = f(a)$$

Thus $f(a) = 0$, and since a was arbitrary, we see f is the constant function equal to zero at all real numbers. □

Corollary 14.1. *Let f, g be continuous functions such that for all $r \in \mathbb{Q}$ they are equal: $f(r) = g(r)$. Then in fact, $f = g$: for all $x \in \mathbb{R}$, $f(x) = g(x)$*

Proof. Since f and g are continuous, the function $h = f - g$ is continuous using the theorems for field operations. And, since $f(x) = g(x)$ for all rational x, we see $h(x) = 0$ on the rationals. Thus, by Proposition 14.3, h itself must be the zero function on all of $\mathbb{R}$. Thus for every x, $h(x) = f(x) - g(x) = 0$, or rearranging,

$$\forall x,\ f(x) = g(x)$$

□

This has a the pretty significant consequence that if we have a function and we know it is continuous, then being able to calculate its values at the rational numbers is good enough to completely determine the function on the real line. In particular, this can be used to prove various *uniqueness results*: you can show a certain function is uniquely defined if you can prove that its definition implies (1) continuity and (2) determines the rational points (or more generally, the values on a dense set).

Exercise 14.12. Let $X \subset \mathbb{R}$ be any dense subset. Prove that if f is a continuous function then it is completely determined by its values on X by showing

- Every real number is the limit of some sequence x_n of points in X.
- If f is continuous $f(x) = 0$ on all points of x, then f is the zero function.
- If f, g are two continuous functions with $f(x) = g(x)$ for $x \in X$, then they are the same function.

We will use this property in understanding exponential functions (where their value at rational numbers are determined by powers and roots) and trigonometric functions (whose values on certain dyadic multiples of π are determined by the half-angle identities.)

15. Transcendental Functions

> Highlights of this Chapter: we introduce the idea of defining functions by a *Functional Equation* specifying how a function should behave instead of specifying how to compute it. Following this approach, we give rigorous definitions for exponentials logarithms and trigonometric functions, and investigate some of their consequences. With these definitions in hand, we are able to define the field of *Elementary Functions*, familiar from calculus and the sciences.

At the heart of real analysis is the study of functions: not only the study of their properties (continuity being a prime example) but also their very definition. Exponentials, trigonometric functions and logarithms are all examples of *transcendental functions* or things that *transcend algebra*: they are not built from a finite composition of the field operations and instead are calculated as the result of infinite processes.

In this chapter we will not focus on *how to compute* such functions, but rather on the more pressing question of *how to even define them*: if all we have available to us are the axioms of a complete ordered field how do we rigorously capture aspects of circles in the plane (trigonometry) or continuous growth (exponentials)? The key is the idea of a *functional equation*: something that will let us define a function by how it behaves, instead of by directly specifying a formula to compute it.

15.1. Functional Equations

Recall the great shift in our collective conception of a function that occurred around the time of Euler, where mathematicians stopped insisting that functions were given by formulas and rather began to welcome rather arbitrary rules, so long as they assigned a unique output to each input. This is accompanied by a conceptual leap, removing the focus from *how to compute* a function and turning towards *what is the function doing*?

This is perhaps easiest to illustrate by example, so we give two below for functions that we already know of from algebra: roots and linear functions.

15.1.1. Roots

How should one define the square root, to someone who has never seen it before? Perhaps as "*the square root is a number that when multiplied by itself, gives the number you started with*". Such a description does a good job of telling us exactly what the square root *does*, and is worth trying to translate into formal mathematics.

In symbols, this means if $r(x)$ is the square root, we need for each allowable value of x, that $r^2(x) = x$. In Example 16.2 we will show that exactly two such functions exist, and there is a unique one with $R > 0$. Thus, this approach is fully rigorous and we call this function *the square root* and write $R(x) = \sqrt{x}$, consistent with Definition 4.6.

In general, we make the same definition, justified by the uniqueness result in Theorem 16.3.

Definition 15.1. Let $r\colon [0,\infty) \to [0,\infty)$ satisfy the functional equation

$$r(x)^n = x$$

Then r is called the n^{th} *root function* and denoted $r(x) = \sqrt[n]{x}$.

The utility of functional equations is that if we can take them as the definition of a particular function we are interested in, we *know for sure* that this function has the property we want: that's all the definition specifies! The hard work them comes in figuring out how to actually compute the values of functions which are defined functionally.

15.1.2. Linear Functions

We know how to express linear functions already using the field axioms, as maps $f(x) = kx$ for some real number k. To speak of linear functions *functionally* however, we should not give a definition telling us how to compute their values (take the input, and multiply by a fixed constant k) but rather by *what they're for*: by the defining property of linearity.

This more abstract functional approach was first taken by Cauchy during the development of analysis, and so the resulting equation is called the *Cauchy Functional Equation*

Definition 15.2 (Cauchy's Functional Equation for Linearity)**.** A function $f\colon \mathbb{R} \to \mathbb{R}$ satisfies Cauchy's functional equation if for all $x, y \in \mathbb{R}$,

$$f(x+y) = f(x) + f(y)$$

Such an abstract characterization has had a tremendous influence in mathematics: for example, think of the definition of a *linear map* in linear algebra.

15.1.3. Difficulties

Moving away from defining a function computationally, there are several potential issues that need to be confronted. First, how do we know that there even is a function satisfying our functional equation?

Example 15.1 (An impossible functional equation)**.** Consider the functional equation

$$f(x)^2 = -1$$

There is no real valued function f satisfying this equation, as squaring to a negative requires complex solutions.

The second worry is to make sure the functional equation really is strict enough to capture what you want it to capture. One example is already presented by linearity: its easy to see that any linear function must be zero at $x = 0$, Thus we could propose the functional equation $L(0) = 0$ enforcing this property. But, this is far from actually capturing the notion of a linear function we had in mind: this functional equation has all sorts of solutions like $f(x) = x^3$ which do exactly what was asked (are zero at zero) but are not what we had in mind.

But, its even worse than this: while it seems that Cauchy's equation captures exactly what we want from the idea of linearity (the ability to distribute over addition) it also has pathological solutions beyond $x \mapsto kx$ that we did not intend:

Example 15.2 (Pathological Solutions to Cauchy's Functional Equation)**.**

To avoid such pathological solutions one needs to impose extra conditions - and a hint at which conditions may help comes from the example above, which turns out to be continuous only at the point $x = 0$. What happens if we ask for a *continuous solution* to Cauchy's equation?

Theorem 15.1. *Any continuous solution to Cauchy's functional equation is a function of the form $f(x) = kx$ for some $k \in \mathbb{R}$.*

Exercise 15.1. Prove Theorem 15.1 by following the outline below:

- Define $k = f(1)$, and prove that $f(x) = kx$ for all $x \in \mathbb{Z}$, using the functional equation.
- Extend this to show that $f(1/n) = k/n$ using the functional equation, and then that for any $r \in \mathbb{Q}$ $f(r) = kr$.
- Use continuity to show that for any $a \in \mathbb{R}$ this implies that $f(a) = ka$.

This is one critical way that analysis enters into the very definition of functions - if we specify what we want a function to do that often leaves room for pathological, discontinuous behavior. And, to get what we really want, we need to ask for our function to behave *continuously*. We see this time and again below, where we define exponentials logarithms and trigonometric functions all as the *continuous solutions* to various functional equations.

15.2. Exponentials

Definition 15.3 (The Law of Exponents)**.** A function $E: \mathbb{R} \to \mathbb{R}$ satisfies the *law of exponents* if for every $x, y \in \mathbb{R}$

$$E(x+y) = E(x)E(y)$$

We use this to give a functional definition of exponential functions.

Definition 15.4. An *exponential function* is a continuous nonzero function E that satisfies the law of exponents.

Now that we have a formal definition, we can start seeing what properties exponential functions must have.

Example 15.3. If E satisfies the law of exponents and evaluates to zero at any point, then E is the zero function.

Proof. Let E be an exponential function and assume there is some $z \in \mathbb{R}$ such that $E(z) = 0$. Then for any $x \in \mathbb{R}$ we may write $x = x - z + z = (x - z) + z = y + z$ for $y = x - z \in \mathbb{R}$. Evaluating $E(x)$ using the law of exponents,

$$E(x) = E(y+z) = E(y)E(z) = E(y) \cdot 0 = 0$$

□

Exercise 15.2. Prove that if E is any exponential function, then $E(0) = 1$, and that $E(-x) = 1/E(x)$.

Exercise 15.3 (Convexity of exponentials)**.** Prove that exponential functions are convex (Definition 13.8): their secant lines lie above their graphs.

Proposition 15.1. *Prove that if E is an exponential function, $x \in \mathbb{R}$ and $r \in \mathbb{Q}$ then*

$$E(xr) = E(x)^r$$

Proof. We deal separately with two cases, for nonzero integers p, q. First we see that $E(px) = E(x)^p$ by inductively applying the law of exponents to $px = x + x + \cdots + x$:

$$E(px) = E(x + x + \cdots + x) = E(x)E(x)\cdots E(x) = E(x)^p$$

Next, we see that $E(x/q) = \sqrt[q]{E(x)}$, again by the law of exponents: Since $x = q(x/q) = (x/q) + (x/q) + \cdots + (x/q)$, we can use the above to see

$$E(x) = E\left(q\frac{x}{q}\right) = E\left(\frac{x}{q}\right)^q$$

Thus, $E(x/q)$ is a number such that when raised to the q^{th} power gives $E(x)$. This is the definition of the q^{th} root, so

$$E\left(\frac{x}{q}\right) = \sqrt[q]{E(x)} = E(x)^{\frac{1}{q}}$$

Putting these two cases together completes the argument, as for $r = p/q$ an arbitrary rational number

$$E(rx) = E\left(\frac{p}{q}x\right) = E\left(\frac{x}{q}\right)^p = (E(x)^p)^{\frac{1}{q}} = E(x)^r$$

□

This has a rather strong consequence for the values of an exponential function at the rational numbers, in terms of its value at a single point:

Definition 15.5 (The Base of an Exponential)**.** If E is any exponential function, its value at 1 is called its *base*.

Corollary 15.1. *Let $E(x)$ be an exponential function with base a. Then for every $r \in \mathbb{Q}$ we have*

$$E(r) = a^r$$

Proof. Let $r \in \mathbb{Q}$ and Proposition 15.1 to $r = r \cdot 1$:

$$E(r) = E(r \cdot 1) = E(1)^r = a^r$$

□

This is a pretty strong property: any two exponential functions that agree at 1 actually agree on the entire real line, since they agree at a dense set. In fact, this is true not just of 1, but of *any point*.

Exercise 15.4 (Exponentials that agree at a point)**.** Prove that if E, F are two exponential functions which take the same value at any nonzero $x \in \mathbb{R}$, then they are equal.

Hint: prove that $x\mathbb{Q} = \{xr \mid r \in \mathbb{Q}\}$ *is dense in* $\mathbb{R}$*, and use Proposition 15.1, and Exercise 14.12*

This work all tells us that *if an exponential function exists at all* then it is fully determined by its value at any point: phrased in terms of the value at 1, *the base uniquely determines the exponential function (if it exists).* We will have to do some more work before we can prove that these functions actually exist however!

Exercise 15.5. Prove that if E is an exponential function with base a, then $a > 0$. *Hint: if* $a < 0$ *what is* $E(1/2)$*?*

15.3. Logarithms

Just like we defined an exponential function by what we want it to do, we will define a logarithm based on its desired properties, giving a functional equation. Logarithms were originally invented to speed up computation, by turning multiplication into addition.

do this, think vhen working ny individual ultiply two 5 adding them

Definition 15.6 (The Laws of Logarithms)**.** We say a function $L : \mathbb{R}_+ \to \mathbb{R}$ satisfies the *laws of logarithms* if for every $x, y \in \mathbb{R}_+$

$$L(xy) = L(x) + L(y)$$

Like in the case of exponentials, we are right to worry that there may be many pathological, everywhere discontinuous solutions to this functional equation. To avoid these, we define *logarithms* to be the continuous solutions

Definition 15.7 (Logarithm)**.** A function L is a logarithm if it is a continuous solution to the law of logarithms (Definition 15.6).

Because of the similarity of the logarithm law to that of exponentials, its perhaps no surprise that with some induction we can fully understand the behavior of these functions on rational inputs:

Proposition 15.2. *Let* $r \in \mathbb{Q}$*. Then if* L *is any logarithm, for every* $x \in \mathbb{R}$ *we have*

$$L(x^r) = rL(x)$$

Exercise 15.6. Prove Proposition 15.2 via the following steps:

- Prove that for any $n \in \mathbb{N}$ we have $L(x^n) = nL(x)$ inductively.
- Prove that $L(x^{-1}) = -L(x)$. Use this to conclude that $L(x^{-n}) = -L(x^n)$ for all $n \in \mathbb{N}$. Thus $L(x^p) = pL(x)$ for all $p \in \mathbb{Z}$.
- Prove that $L(x^{1/q}) = \frac{1}{q}L(x)$ for $q \in \mathbb{N}$.
- Put these together to see that for $r = p/q$, $L(x^r) = rL(x)$.

This gives an equality between two functions at *every rational value*. Because the functions are continuous (L is continuous, a^x is continuous, and multiplication is continuous). Thus, these are equal at *every real value*

Corollary 15.2. *Let L be a logarithmic function, $x > 0$, and $y \in \mathbb{R}$. Then*

$$L(x^y) = yL(x)$$

This has a pretty incredible consequence:

Theorem 15.2. *The inverse of an exponential function is a logarithm function!*

Proof. Let E be an exponential and L a logarithm function. Then for any x we have

$$L(E(x)) = L(E(1)^x) = xL(E(1))$$

Thus, for the exponential E of base $a = E(1)$, if we choose a logarithm function L where $L(a) = 1$, we see

$$L(E(x)) = x$$

so they are inverses! □

This makes it natural to try and define the *base* of a logarithm:

Definition 15.8 (Base of a Logarithm)**.** If L is a logarithm, its *base* is the real number a such that $L(a) = 1$.

Unlike for the exponential where the base was a *value* of the function (which then existed by definition), we do not know a priori that every logarithm takes the value 1 at some point, or even that it does so uniquely! So, we will have some work to do to show this actually makes sense.

15.4. Trigonometric

The trigonometric functions are originally defined geometrically, but like the exponentials above, we will specify them by a *functional equation* - specifying how the functions *behave* instead of what they *measure*.

Trigonometric functions satisfy *many* functional equations - these are what we call trigonometric identities! And, as one is perhaps too familiar with from a trigonometry class, there are many *many* trigonometric identities! Here our goal is to pick some small set of identities to impose as the *axioms* for trigonometry, from which all other functional properties can be derived.

The natural candidates are the angle sum or difference identities:

Definition 15.9 (Angle Sum Identities)**.** Two functions S, C satisfy the *angle-sum identities* if for any $x, y \in \mathbb{R}$:

$$S(x+y) = S(x)C(y) + C(x)S(y)$$

$$C(x+y) = C(x)C(y) - S(x)S(y)$$

Definition 15.10 (Angle Difference Identities)**.** Two functions S, C satisfy the *angle-sum identities* if for any $x, y \in \mathbb{R}$:

$$S(x-y) = S(x)C(y) - C(x)S(y)$$

$$C(x-y) = C(x)C(y) + S(x)S(y)$$

In fact, either of these serves just fine, but for technical reasons (shortening some proofs a little bit) it's easier to take the angle difference identities as our functional equations.

Definition 15.11. A pair of functions S, C are called *trigonometric* if they are continuous solutions to the angle difference identities.

This seems perhaps surprisingly non-restrictive: nowhere have we built in tha these functions are periodic, or differentiable, or anything else! Can all of trigonometry really be reduced to this simple rule and the imposition of continuity? Indeed it can! And this development will be the subject of the final project in this course. –>

15.5. ⋆ Elementary Functions

The functions you are used to seeing in a calculus course, and in the sciences are called *elementary functions*, and include all the functions we have discussed so far in this course, as well as messy combinations like

While in the sciences people often non-rigorously think of the elementary functions as simply "those functions which have a formula" we should be more precise as mathematicians. After all, what is to stop us from giving a fancy name like $\mathrm{char}(x)$ to the Characteristic function of the rationals - then we could say something like $\sin(e^{\mathrm{char}}(x))$ 'has a formula we can write down'?

Definition 15.12. The elementary functions $\mathcal{E}$ are the set of real valued functions produced using the field operations and composition from the following basic building blocks:

- Constants
- Powers x^n and their inverses $\sqrt[m]{x}$
- Exponentials $E(x)$ and their inverses $L(x)$
- Trigonometric functions $S(x), C(x)$ and their inverses $S^{-1}(x), C^{-1}(x)$

This list includes all the familiar functions; from the tangent $\tan x = \frac{\sin x}{\cos x}$ to the hyperbolic cosine $\cosh(x) = \frac{e^x+e^{-x}}{2}$.

But this list, as written is not fully 'minimal': we can remove some functions from it without changing the class $\mathcal{E}$!

Exercise 15.7. The function x^n is generated from x by repeated multiplication, and so is not required to be part of the list of basic building blocks for elementary functions so long as x is included.

Show the same is true for the roots $\sqrt[n]{x}$: given an exponential E and its inverse L, can you find a formula for the n^{th} root? (Recall the definition: r is the n^{th} root of x if $r^n = x$)

Thus, we can reduce without loss of generality the line "powers x^n and their inverses $\sqrt[n]{x}$" to just requiring the identity function $f(x) = x$ is elementary!

But even further simplification is possible. As we continue to study the transcendental functions, we will learn a lot more about them from their functional definitions. Indeed, we will see that each of these picks out an essentially unique function:

- There is a unique exponential function up to scaling: if E_1 and E_2 are any exponentials, then there exists a constant $k \in \mathbb{R}$ such that $E_2(x) = E_1(kx)$.
- There is a unique logarithm function up to scaling: if L_1 and L_2 are any logarithms, then there exists a constant $k \in \mathbb{R}$ such that $L_2(x) = L_1(kx)$.
- There exists a unique pair of trigonometric functions up to scaling: if (S_1, C_1) and (S_2, C_2) are two pairs of trigonometric functions, then there exists a constant $k \in \mathbb{R}$ such that $(S_2(x), C_2(x)) = (S_1(kx), C_1(kx))$.

Thus, since constant scaling of the argument is part of the 'construction kit' for elementary functions (k is an elementary function, x is an elementary function kx is a field operation and $f(x) \mapsto f(kx)$ is composition), we will be able to use these results to further simplify our definition by choosing one particular function of each type. We will see that analysis provides a natural choice: the exponential e^x and its inverse $\log(x)$, and the trigonometric functions $\sin(x)$, $\cos(x)$ (from which we will derive the notion of 'radians'). This abbreviated definition is then

Definition 15.13. The elementary functions $\mathscr{E}$ are the set of real valued functions produced using the field operations and composition from the following basic building blocks:

- Constants
- The identity x
- The exponential e^x and its inverse $\log(x)$,
- The trigonometric functions $\sin(x)$, $\cos(x)$ and their inverses $\arcsin(x)$, $\arccos(x)$.

If we further allow ourselves to work with *complex valued* functions of a real argument (which is both mathematically convenient, and relevant to the sciences) even further simplification becomes possible: Euler's formula relates the exponential to the sine and cosine

$$e^x = \cos(x) + i\sin(x)$$

So we may derive formulas for sin and cos in terms of the exponential itself (and formulas for their inverses in terms of the logarithm)

$$\sin(x) = \frac{e^{ix} - e^{-ix}}{2i} \qquad \cos(x) = \frac{e^{ix} + e^{-ix}}{2}$$

Thus, if complex constants are allowed instead of just real constants, the definition of elementary functions reduces even further:

Definition 15.14. The elementary functions $\mathscr{E}$ are the set of real valued functions produced using the field operations and composition from the following basic building blocks:

- Constants $\in \mathbb{C}$
- The identity x
- The exponential e^x and its inverse $\log(x)$,

This shows, more than anything else (in my opinion) how the exponential function e^x truly is fundamental: the large class of elementary functions we've known since childhood is simply what you get from using field operations and composition starting from two functions: the identity and the exponential.

16. Theory

> Highlights of this Chapter: we prove two foundational results about continuous functions whose proofs have several steps in common:
>
> - The Extreme Value Theorem: a continuous function achieves a max and min on any closed interval.
> - The Intermediate Value Theorem: a continuous function must take every value between $f(a)$ and $f(b)$ on the interval $[a, b]$.
>
> We will call the proof style introduced with these theorems "proof by continuity". Finally, we investigate one further topic - uniform continuity - where this proof strategy also helps, and prove that a continuous function on a closed interval is uniformly continuous.

Here cover some of the important theorems about continuous functions that will prove useful during the development of calculus. Just like we have seen various 'proof styles' for sequences (recurrent themes in proofs, like 'an $\epsilon/2$ argument') one of the biggest takeaways of this section is a proof technique for working with continuous functions. It has three steps, summarized below:

- Use whatever information you have to start, to construct a sequence of points.
- Use Bolzano Weierstrass to find a convergent subsequence.
- Apply f to that sequence and use continuity to know the result is also convergent.

This is to vague on its own to be useful, but in reading the proofs of the boundedness theorem, the extreme value theorem, and the intermediate value theorem below, look out for these three recurrent steps.

16.1. Extreme Values

Proposition 16.1. *Let f be a continuous function on a closed interval $[a, b]$. Then the image $f([a, b])$ is bounded.*

Proof. Assume for the sake of contradiction that f is not bounded. Then for each $n \in \mathbb{N}$ there must be some $x_n \in [a, b]$ where $|f(x_n)| > n$. This sequence $\{x_n\}$ need not be convergent, but it lies in the interval $[a, b]$ so it is bounded, and thus contains a convergent subsequence x_{n_k} by Bolzano Weierstrass. Say $x_{n_k} \to x$. Then since

$a \leq x_{n_k} \leq b$ for all k, by the inequalities of limits we see $a \leq x \leq b$ so the limit x lies in the interval $[a, b]$ as well.

But what is the value $f(x)$? Since f is continuous and $x_{n_k} \to x$ we know that

$$f(x_{n_k}) \to f(x)$$

But for each k, x_{n_k} has the property that $f(x_{n_k}) > n_k$ *by definition*. Thus, the sequence $f(x_{n_k})$ is not bounded, and cannot be convergent (since all convergent sequences are bounded). This is a contradiction, as it implies that $f(x)$ is not defined, even though we have assumed f is defined on the entire interval $[a, b]$.

Thus, no such sequence x_n is possible, and so there must be some n where $|f(x)| < n$ for all $x \in [a, b]$. That is, f must be bounded on $[a, b]$. □

Building off this result, one can prove that a continuous function actually *achieves* its upper and lower bounds on any closed interval. This result will play a role several times across the theory of functions and derivatives, so we give it a memorable name: the *extreme value theorem* (as maxima and minima taken collectively are called *extrema*).

Theorem 16.1 (The Extreme Value Theorem)**.** *Let f be a continuous function on a closed interval $[a, b]$. Then f achieves a maximum and minimum value: that is, there exists a point p where $f(p) \geq f(x)$ for all $x \in [a, b]$, and a q where $f(q) \leq f(x)$ for all $x \in [a, b]$.*

Proof. Let f be continuous on $[a, b]$ and let $R = \{f(x) \mid x \in [a, b]\}$ be the set of outputs, or the *range* of f. Since f is bounded we see that R is a bounded subset of $\mathbb{R}$, and so by completeness

$$m = \inf R \qquad M = \sup R$$

must exist. Our goal is to find values $x_m, x_M \in [a, b]$ for which the infimum and supremum are realized:

$$f(x_m) = m \qquad f(x_M) = M$$

Here we show this holds for the supremum, the infimum is left as an exercise below. Since M is the supremum, for any $\epsilon > 0$ we know that $M - \epsilon$ is *not* an upper bound for $R = \{f(x) \mid x \in [a, b]\}$: thus there must be some x where $f(x) > M - \epsilon$. So letting $\epsilon = 1/n$ each n, let x_n be a point where $M - \frac{1}{n} < f(x_n) \leq M$. As $n \to \infty$ we know $M - \frac{1}{n} \to M$ and so by the squeeze theorem we see that $f(x_n) \to M$ as well.

We don't know that the points x_n themselves converge, but we do know that this entire sequence lies inside the closed interval $[a, b]$ so its bounded and Bolzano Weierstrass lets us extract a convergent subsequence $x_{n_k} \to x$. And as $a \leq x_{n_k} \leq b$ it follows that the limit $x \in [a, b]$ as well. Because subsequences of a convgent sequence converge to the same limit, we know that $f(x_{n_k})$ is convergent, and still has limit M. But now we can finally use continuity!

Since f is continuous, we know $\lim f(x_n) = f(\lim x_n)$, and so $M = f(x)$. Thus we managed to find a point $x \in [a, b]$ where $f(x)$ is the supremum: $f(x)$ is an upper bound for all possible values of f on $[a, b]$, which by definition means its the max value! So f achieves a maximum on $[a, b]$. □

Exercise 16.1. Complete the proof of the extreme value theorem by showing that the infimum of a function on a closed interval is also realized as its value at a point.

16.2. Intermediate Values

The intermediate value theorem is the rigorous version of "you can draw the graph of a continuous function without picking up your pencil".

One note: in the statement below we use the phrase y is between $f(a)$ and $f(b)$ as a shorthand to mean that either $f(a) < y < f(b)$ or $f(b) < y < f(a)$ (as we don't know if $f(a)$ or $f(b)$ is larger).

Theorem 16.2 (The Intermediate Value Theorem)**.** *Let f be a continuous function on the interval $[a, b]$, and let y be any number between $f(a)$ and $f(b)$. Then there exists an x between a and b such that $y = f(x)$.*

Proof. Without loss of generality we will assume that $f(a) < f(b)$ so that y lies in the interval $[f(a), f(b)]$ (the other case is analogous, we just instead must write the interval $[f(b), f(a)]$). We wish to find a point $x \in [a, b]$ where $f(x) = y$, so we start by defining the set of points where $f(x)$ is less than *or equal to* y:

$$S = \{x \in [a, b] \mid f(x) \leq y\}$$

This set is nonempty: $a \in S$ as $f(a) < y$ by assumption. And its bounded above by b: if $x \in S$ then $x \in [a, b]$ so $x \leq b$ by definition. Thus, the supremum $\sigma = \sup S$ exists, and $\sigma \in [a, b]$. We will show that $f(\sigma) = y$, by showing both inequalities $f(\sigma) \leq y$ and $f(\sigma) \geq y$.

First, we show $\leq$. Since σ is the supremeum, for each n we know that $\sigma - \frac{1}{n}$ is not an upper bound, and so there must be an point $x_n \in (\sigma - 1/n, \sigma)$ where $f(x_n) \leq y$. The squeeze theorem assures that $x_n \to \sigma$, and the continuity of f assures that $f(x_n)$ converges (since x_n does). But for all n we know $f(x_n) \leq y$, so by the inequalities of limits we also know $\lim f(x_n) = f(\sigma) \leq y$.

Next, we show $\geq$. First note that $\sigma \neq b$ as $f(\sigma) \leq y$ but $f(b) > y$. So, $\sigma < b$ and so after truncating finitely many terms, the sequence $x_n = \sigma + 1/n$ lies strictly between σ and b. Since this sequence is *greater* than the upper bound σ, we know that none of the x_n are in S and so $f(x_n) > y$ by definition, for all n. But as $n \to \infty$ the sequence of x_n's is squeezed to converge to σ, and so by continuity we know

$$f(\sigma) = f(\lim x_n) = \lim f(x_n)$$

Applying the inequalities of limits this time yields the reverse: since for all n we know $f(x_n) > y$, it follows that $\lim f(x_n) \geq y$ so $f(\sigma) \geq y$.

Putting these together we know that $f(\sigma)$ is some number which must simultaneously by $\geq y$ and $\leq y$. The only number satisfying both of these inequalities is y itself, so

$$f(\sigma) = y$$

□

Historically, the intermediate value theorem was one of the reasons for developing much of analysis: mathematicians knew that *whatever the correct formal definition of continuity was, it should certainly imply this!* So, our proof of the intermediate value theorem (which embodies the intuitive notion of continuity) may be seen as evidence that we have chosen good definitions of continuity and convergence: they work as we expect!

Remark 16.1. It may seem at first that this is EQUIVALENT to continuity: if a function satisfies the intermediate value property, then its continuous. Try to prove it! Where do you get stuck?

Example 16.1. Consider the following function

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

Then f satisfies the conclusion of the intermediate value theorem on every closed interval, but f is not continuous at 0.

16.2.1. Applications of the IVT

The intermediate value theorem has many applications, as it is often the case that we know information about a function at several points, and want to infer information about its value at others. One immediate application is a way of finding roots:

Corollary 16.1. *If f is a continuous function on an interval and it is positive one end-point and negative on the other, then f has a zero in-between.*

This suggests a means of finding the zeros of a function, which narrows in on them exponentially fast! Called "bisection": find any two points where function changes sign. Divide region in half, evaluate at midpoint. Keep interval with different sign endpoints, repeat.

Second, this lets us understand something about the range of continuous functions:

Corollary 16.2. *If f is a continuous function and $I \subset \mathbb{R}$ is a closed interval, then $f(I)$ is an interval.*

Here we allow the degenerate case $[a, a] = \{a\}$ to count as an interval, if f is constant.

Another application is to prove the existence of certain *inverse functions* - we will look here at the example of roots. Of course, we already have a rigorous argument for the existence of $\sqrt{x}$ for any nonnegative x, but this argument was quite low-level: working directly with the definition of supremum and the Archimedean property! Now that we have built up more machinery, we can re-prove the same result in a much cleaner way:

Example 16.2. For every $v > 0$ there exists a positive u with $u^2 = v$: we call this the *square root* $u = \sqrt{v}$.

Proof. Let $v > 0$ and consider the function $f(x) = x^2 - v$. This function is continuous, and at $x = 0$ this function is negative, so all we need to do is find a point where the function is positive to be able to apply the IVT. Note $f(v+1) = (v+1)^2 - v = v^2 + v + 1$ is positive: thus there must be some point $u \in [0, v+1]$ such that $u^2 = v$, as required. □

Exercise 16.2. For any $x \in \mathbb{R}$ there exists a unique a with $a^3 = x$: we call this a the *cube root.*

Theorem 16.3. *For every $x \geq 0$ there exists a unique positive number y such that $y^n = x$.*

16.2.2. Fixed Points

Another application of the intermediate value theorem beyond finding roots is to prove various *fixed point theorems* which guarantee that, under certain conditions on a function f there is a point x with $f(x) = x$.

Example 16.3 (A Fixed Point Theorem)**.** Let $f : [0, 1] \to [0, 1]$ be any continuous function. Then there exists a fixed point: an $x \in [0, 1]$ where $f(x) = x$.

Example 16.4. There is a solution to the equation $\cos(x) = x$.

Exercise 16.3. Prove that every cubic polynomial has a real root. *Hint: show its enough to consider monic cubics $p(x) = x^3 + ax^2 + bx + c$. Can you prove there is some number M where $p(M)$ is positive but $p(-M)$ is negative?*

16.3. Uniform Continuity

There is one final tool that often proves useful when working with continuous functions, related to the ϵ-δ definition. That definition specifies that given an ϵ, at each x where f is continuous you can find a δ, but does not give any information about *how* to do so, meaning we cannot say anything about if the δ's at nearby points are related.

Thus, just looking at the definition, one may be tempted to also write down a *stronger* alternative, which says that you can use the same δ at every point:

Definition 16.1 (Uniform Continuity)**.** A function f is uniformly continuous on a domain D if for every $\epsilon > 0$ there exists a *single $\delta * thatcanbeusedateverypointinthecontinuitydefinition: D, |x-a| < \delta \implies |f(x)-f(a)| < \epsilon$$

Sometimes when proving continuity using $\epsilon - \delta$, its easy to directly see that a function is in fact uniform continuous as when playing the $\epsilon - \delta$ game its easy to pick an expression for δ that doesn't depend on a:

Example 16.5. $f(x) = x^2$ is uniformly continuous on the interval $[1, 3]$.

Here's some scratch work: let $\epsilon > 0$. Then at any a we see that $|f(x) - f(a)| = |x^2 - a^2| = |x + a||x - a|$. If $|x - a| < \delta$ and we want $|f(x) - f(a)| < \epsilon$, this tells us that we want

$$|x + a|\delta < \epsilon$$

We don't know what x and a are, but we do know they are points in the interval $[1, 3]$! So, the smallest $x + a$ could be is $1 + 1 = 2$, and the biggest is $3 + 3 = 6$. This means that

$$|x + a|\delta \leq 6\delta$$

So, if we can make $6\delta < \epsilon$, we are good! This is totally possible: just set $\delta = \epsilon/6$. Below is the rigorous proof.

Proof. Let $\epsilon > 0$, and set $\delta = \epsilon/6$. Note that for any $a \in [1, 3]$ and any x within δ of a, we know $a \leq 3$ and $x \leq 3$ so $x + a \leq 6$. But this implies that

$$|x^2 - a^2| = |x + a||x - a| \leq 6|x - a| < 6\delta < 6\frac{\epsilon}{6} = \epsilon$$

And so f is uniformly continuous, as this single choice of δ works for every point $a \in [1, 3]$. □

Intuitively, what this means is that at our function f cannot vary too much over any fixed interval: we can use the same error bar at any point to control the total change in y values! But of course, this is not true for every continuous function - functions that change arbitrarily quickly (say, by having a vertical asymptote, or an accumulation of oscillations) require smaller and smaller choices of δ as one approaches the area where the function is 'behaving badly'.

Example 16.6. $f(x) = \frac{1}{x}$ is *not* uniformly continuous on the interval $(0, \infty)$

Again, lets start with some scratch work. First, notice that we can easily see (via the sequence definition of continuity) that f is continuous on $(0, \infty)$. But, let's actually do it to see what happens:

Looking at $|f(x) - f(a)|$ we can do some algebra to see

$$\left|\frac{1}{x} - \frac{1}{a}\right| = \left|\frac{a-x}{xa}\right| = \frac{\delta}{xa}$$

We want to make this less than ϵ, and we know that x is within δ of a (so the smallest it could be is $a - \delta$). Thus, $\delta/a(a-\delta) < \epsilon$ and we can solve this for δ:

$$\delta = \frac{a^2\epsilon}{1 + a\epsilon}$$

This gives us for each a, a different δ. What we might like to do is to pick the *minimum* δ - that would work for all a! But here we have a problem - there is no minimum: as $a \to 0$, the δ we need to choose goes to zero as well.

Example 16.7. The function $f(x) = \sin\left(\frac{1}{x}\right)$ is continuous, but is not uniformly continuous on $(0, \infty)$, or on any interval of the form $(0, L)$. As a gets closer to 0, one must choose smaller and smaller δs to keep the oscillation of $\sin(1/x)$ less than ϵ.

Both of these functions have problems stemming from a function misbehaving at the boundary of an open interval, as when approaching this endpoint our allowed choices of δ go to zero. A natural question is - is this the only problem that can occur? That is, if we have a function defined on a *closed interval*, can we always pick a uniform δ?

Indeed we can! This tells us that on closed intervals, continuous functions are even nicer behaved than we originally knew: they must be *uniformly continuous.*

Theorem 16.4. *If f is continuous on a closed interval I, f is uniformly continuous.*

Proof. Assume for the sake of contradiction that f is not uniformly continuous, and fix $\epsilon > 0$. Then there is no fixed δ that works, so for any proposed δ, there must be some a where it fails.

We can use this to produce a sequence: for $\delta = 1/n$ let $a_n \in I$ be a point where this δ fails: there is some x_n within $1/n$ of a_n but $|f(x_n) - f(a_n)| > \epsilon$.

Thus, in fact we have two sequences x_n and a_n! We know very little about either except that they are in a closed interval I, so we can apply Bolzano Weierstrass to get convergent subsequences (we have to be a bit careful here, see the exercise below).

We will call the subsequences X_n and A_n (with capital letters). Now that we know they both converge, we can see that they also have the same limit: (as, by construction $|X_n - A_n| < \frac{1}{n}$). Call that limit L.

Then since f is continuous at L, we know that

$$\lim f(X_n) = f(\lim X_n) = f(L) = f(\lim A_n) = \lim f(A_n)$$

Thus, $\lim f(X_n) - f(A_n) = 0$. However this is impossible, since for all values of n we know $|f(X_n) - f(A_n)| > \epsilon$! This is a contradiction, and thus there must have been some uniform δ that worked all along. □

Exercise 16.4. Let x_n and y_n be two bounded sequences. Show that it is possible to choose some subsequence of the indices n_k such that the subsequences x_{n_k} and y_{n_k} both converge.

(Note we can't apply Bolzano Weierstrass individually to x_n and y_n: what if that gives you that the even subsequence of x_n is convergent, and the odd subsequence of y_n is convergent!)

Exercise 16.5 (Uniform Continuity and $\mathbb{R}$:)**.** We know that if f is continuous on any closed interval it is uniformly continuous, but it is also possible that functions on open or infinite intervals are uniformly continuous. Show this by example, confirming that

$$f(x) = \frac{1}{1+x^2}$$

is uniformly continuous on the entire real line.
Hint: try to simplify and overestimate the quantity $f(x) - f(a)$: remember that $1+x^2$ and $1+a^2$ are always ≥ 1!

17. Limits

> Highlights of this Chapter: we introduce the notion of a limit of a function, as well as the limit from above and the limit from below. We prove that a limit exists if and only if these right and left hand limits both exist, and are equal - a fact which will prove useful in various calculations with derivatives to come.

Sometimes we need to understand the behavior of a function *near* a point, without actually being able to compute the function's value *at* that point (perhaps, that point is outside the functions' domain). To do so, we use sequences once again to help us out!

First, one quick definition to make terminology easier: :::{#def-limit-point} Let $D \subset \mathbb{R}$ be a set. Then a point $p \in \mathbb{R}$ is a *limit point of* D if there is at least one sequence of points in D converging to a. :::

For example, 0 is a limit point of $(0,1)$ even though it is not a point of $(0,1)$. Any point in a set is trivially a limit point of that set (just take the constant sequence equal to that point over and over).

Definition 17.1 (Limits of Functions)**.** Let f be a function defined defined on a domain $D \subset \mathbb{R}$ and let a be a limit point of D. Then we write

$$\lim_{x\to a} f(x) = L$$

to mean for every sequence $\{x_n\} \subset D$ with $x_n \to a$ and $x_n \neq a$, we have

$$\lim f(x_n) = L$$

Note, this definition looks a lot like the definition of continuity, except that we are not able to say "approaches $f(a)$" as we are not interested in what f is doing at a (or even if f is defined at a), but rather only on what is happening nearby.

Let's do some quick examples to get a feel for this definition:

Example 17.1.

$$\lim_{x\to 2} 3x^2 + 4$$

Let x_n be any sequence converging to 2, for which $x_n \neq 2$ for all n. Then by the limit theorems we see $x_n^2 \to 4$, so $3x_n^2 \to 12$, and $3x_n^2 + 4 \to 16$. Since x_n was arbitrary, this holds for all such sequences, thus

$$\lim_{x \to 2} 3x^2 + 4 = 16$$

Example 17.2.

$$\lim_{x \to 2} \begin{cases} x^2 & x \neq 2 \\ 3 & x = 2 \end{cases}$$

Let x_n be any sequence converging to 2, for which $x_n \neq 2$ for all n, and f be the piecewise function above. Then $f(x_n) = x_n^2$ for all n as we avoid the case $x = 2$, and using the limit theorems $x_n^2 \to 4$. Since x_n was arbitrary,

$$\lim_{x \to 2} f(x) = 4$$

Note this is true even though $f(2) = 3$.

This example shows why we do not consider sequences that contain the point a: if we looked at sequences converging to 2 that contained infinitely many 2's above, they either diverge, or converge to 3, whereas all other sequences converge to 4 as we showed. The next example shows another utility of this.

Example 17.3.

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2}$$

Let x_n be any sequence converging to 2, for which $x_n \neq 2$ for all n. Then since $x_n \neq 2$ the denominator of $(x^2 - 4)/(x - 2)$ is never zero, and we can simplify with algebra:

$$\frac{x_n^2 - 4}{x_n - 2} = \frac{(x_n + 2)(x_n - 2)}{x_n - 2} = x_n + 2$$

Thus, for all n we have

$$\lim \frac{x_n^2 - 4}{x_n - 2} = \lim x_n + 2 = \lim(x_n) + 2 = 4$$

Since x_n was arbitrary, this holds for all sequences and

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2} = 4$$

Example 17.4.

$$f(x) = \begin{cases} 0 & x < 0 \\ 17 & x = 0 \\ x & x > 0 \end{cases}$$

Then $\lim_{x\to 0} f(x) = 0$

Example 17.5.

$$f(x) = \begin{cases} 0 & x < 0 \\ 17 & x = 0 \\ x^2 + 1 & x > 0 \end{cases}$$

Then $\lim_{x\to 0} f(x)$ does not exist.

Here's the familiar theorem from calculus that you can 'plug in' when taking limits of continuous functions.

Theorem 17.1 (Limit of Continuous Functions)**.** *If f is continuous at a then* $\$\lim_{x\to a} f(x) = f(a)$

Proof. Let f be continuous at a and $x_n \to a$ be an arbitrary sequence with $x_n \neq a$ for all n. By the assumption of continuity, we know that for *all* sequences converging to a,

$$\lim f(x_n) = f(\lim x_n) = f(a)$$

Since our x_n is such a sequence (just with the extra condition that $x_n \neq a$) the same holds. And, as x_n was an arbitrary such sequence

$$\lim_{x\to a} f(x) = f(a)$$

□

17.0.1. Epsilons and Deltas

Just like for continuity - whatever we can do with arbitrary sequences we can also do with ϵs and δs. Here's an alternative conception of functional limits:

Definition 17.2 (Function Limits and $\epsilon - \delta$)**.** Let f be a function with domain D and a a limit point of D. Then we write $\lim_{x\to a} f(x) = L$ if for all $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in D, x \neq a$ with $|x - a| < \delta$, we have $|f(x) - f(a)| < \epsilon$.

Exercise 17.1. Reprove the above theorem using the $\epsilon - \delta$ definition of limit, and the $\epsilon - \delta$ definition of continuity.

Exercise 17.2 (Equivalence of Limit Definitions)**.** Prove that the $\epsilon - \delta$ definition of functional limits is equivalent to the sequence definition (using the same ideas we used to prove the analogous definitions of continuity are equivalent).

17.1. Limits from Above and Below

In some cases, we want to consider a more restricted notion of limit: not one that considers *all sequences* but rather one that only notices sequences larger than, or smaller than the target value. It is easy to modify the sequence definition for

Definition 17.3 (Limit From Above)**.** Let a be a limit point of the domain of a function f. Then we write

$$\lim_{x\to a^+} f(x) = L$$

and say "The limit from above is L" if for all sequences x_n in the domain with $x_n \to a$ and $x_n > a$, we have $\lim f(x_n) = L$.

Definition 17.4 (Limit From Below)**.** Let a be a limit point of the domain of a function f. Then we write

$$\lim_{x\to a^-} f(x) = L$$

and say "The limit from below is L" if for all sequences x_n in the domain with $x_n \to a$ and $x_n < a$, we have $\lim f(x_n) = L$.

Exercise 17.3. Give an $\epsilon - \delta$ definition of limits from above, and limits from below.

One sided limits are very useful as they can be easier to compute, but their values completely determine the value of the overall limit. In particular:

Exercise 17.4. Let f be a function and a a limit point of its domain. Then if

$$\lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x)$$

The overall limit $\lim_{x\to a} f(x)$ exists, and has the same value as the two one sided limits.

*Hint: take an arbitrary sequence $x_n \to a$ with $x_n \neq a$, and decompose it into a union of two subsequences, the "right subsequence" r_n of terms $> a$ and the "left subsequence" ℓ_n of terms $< a$. Use what you know about limits of sequences and subsequences!

The converse of this is immediate, as the sequences ranged over in the definitions of $\lim_{x\to a^+}$ and $\lim_{x\to a^-}$ are just particular cases of the sequences ranged over in the definition of $\lim_{x\to a}$. Together, these yield the following theorem:

Theorem 17.2. *Let f be a function and a a point of its domain. Then $\lim_{x\to a} f(x)$ exists if and only if both one sided limits exist and are equal. And, in this case*

$$\lim_{x\to a} f(x) = \lim_{x\to a^+} f(x) = \lim_{x\to a^-} f(x)$$

We can generalize this to formulate a condition on when gluing two continuous functions together remains continuous. While elementary, this result proves *extremely* useful in analysis and topology, and is called the *pasting lemma* as it allows you to paste continuous functions together.

Exercise 17.5 (The Pasting Lemma)**.** Let f, g be two continuous functions and $a \in \mathbb{R}$ is a point such that $f(a) = g(a)$. Prove that the piecewise function below is continuous at a.

$$h(x) = \begin{cases} f(x) & x \leq a \\ g(x) & x > a \end{cases}$$

When do one-sided limits exist at all? One useful assurance of existence is monotonicity:

Exercise 17.6. Let f be a bounded monotone function on the interval (a, b). Then both of the one sided limits exist

$$\lim_{x \to a^+} f(x) \qquad \lim_{x \to b^-} f(x)$$

Hint: show they are the inf and sup of $\{f(x) \mid x \in (a, b)\}$

This proves useful in many cases where we know only that our function is monotone, but cannot compute its values. For us, the most important application is Proposition 23.2 where we show exponential functions are differentiable, when we have only assumed they are continuous.

One result that will be useful later on is describing continuity with function limits. These definitions are very similar, except that $\lim_{x \to a} f(x)$ only considers sequences with $x_n \neq a$, whereas the definition of continuity at a requires we consider *all sequences* that converge to a. This does not pose any serious issue, as the only difference is terms literally equal to a!

Theorem 17.3. *f is continuous at a if and only if* $\lim_{x \to a} f(x) = f(a)$.

Part V.

Series & Products

18. Definition

Highlights of this Chapter: We define infinite series and infinite products, and relate them through via exponential functions and logarithms, reducing the theory to the study of one or the other. We then introduce two classes of series that we can essentially compute by hand: telescoping sums, and the geometric series.

We return from our excursion into the study of functions back to sequences for a short bit, and discuss two particular types of recursive sequences which prove to be extremely useful across mathematics: infinite series, and infinite products. Most of the material in this section and the following could easily have been covered much earlier - the reason we have postponed them is that the most striking applications of sequences and series involve not numbers but whole *functions*, and now that we have that technology available we will be able to present the theory in its fullest.

Definition 18.1 (Series)**.** A series s_n is a recursive sequence defined in terms of another sequence a_n by the recurrence relation $s_{n+1} = s_n + a_n$. Thus, the first terms of a series are

$$s_0 = a_0, \qquad s_1 + a_0 + a_1 \qquad s_2 = a_0 + a_1 + a_2 \ldots$$

We use *summation notation* to denote the terms of a series:

$$s_n =_0 +a_1 + \cdots + a_n = \sum_{k=0}^{n} a_k$$

Remark 18.1. It is important to carefully distinguish between the sequence a_n of *terms* being added up, and the sequence s_n of partial sums.

When a series converges, we often denote its limit using summation notation as well. The traditional 'calculus notation' sets n to infinity as the upper index; and another common notation is to list just the subset of integers over which we sum in as the lower bound: all of the following are acceptable

$$\lim s_n = \lim \sum_{k=0}^{n} a_k = \sum_{k=0}^{\infty} a_k = \sum_{k\geq 0} a_k$$

There are many important infinite series in mathematics: one that we encountered earlier is the Basel series first summed by Euler.

$$\sum_{n\geq 1}\frac{1}{n^2}=\frac{\pi^2}{6}$$

When the sequences a_n consists of functions of x, we may define an infinite series *function* for each x at which it converges. These describe some of the most important functions in mathematics, such as the Riemann zeta function

$$\zeta(s)=\sum_{n\geq 1}\frac{1}{n^s}$$

One of our big accomplishments to come in this class is to prove that exponential functions can be computed via infinite series, and in particular, the *standard exponential* of base e has a very simple expression

$$\exp(x)=\sum_{n\geq 0}\frac{x^n}{n!}$$

Remark 18.2. Because the sum of any finitely many terms of a series is a finite number, we can remove any finite collection without changing whether or not the series converges. In particular, when proving convergence we are free to ignore the first finitely many terms when convenient.

Because of this, we often will just write $\sum a_n$ when discussing a series, without giving any lower summation bound.

The other infinite algebraic expression we can conjure up is infinite products:

Definition 18.2 (Infinite Products)**.** An infinite product p_n is a recursive sequence defined in terms of another sequence a_n by the recurrence relation $p_{n+1}=p_na_n$. Thus, the first terms of a series are

$$s_0=a_0,\qquad s_1+a_0a_1\qquad s_2=a_0a_1a_2\ldots$$

We use *product notation* to denote the terms of a series:_

$$s_n=_0 a_1\cdots a_n=\prod_{k=0}^{n}a_k$$

Again, like for series, when such a sequence converges there are multiple common ways to write its limit:

$$\lim p_n = \lim \prod_{k=0}^{n} p_k = \prod_{k=0}^{\infty} p_k = \prod_{k\geq 0} p_k$$

The first infinite product to occur in the mathematics literature is Viete's Product for π

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2}\frac{\sqrt{2+\sqrt{2}}}{2}\cdots$$

This product is derived from Archimedes' side-doubling procedure for the areas of circumscribed n-gons; hence the collections of nested roots!

Another early and famous example being Wallis' infinite product for $2/\pi$, which instead is derived from Euler's infinite product for the sine function.

$$\frac{\pi}{2} = \prod_{n\geq 1} \frac{4n^2}{4n^2-1}$$
$$= \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\frac{8}{7}\frac{8}{9}\frac{10}{9}\frac{10}{11}\frac{12}{11}\frac{12}{13}\frac{14}{13}\frac{14}{15}\cdots$$

In 1976, the computer scientist N. Pippenger discovered a modification of Wallis' product which converges to e:

$$\frac{e}{2} = \left(\frac{2}{1}\right)^{\frac{1}{2}}\left(\frac{2}{3}\frac{4}{3}\right)^{\frac{1}{4}}\left(\frac{4}{5}\frac{6}{5}\frac{6}{7}\frac{8}{7}\right)^{\frac{1}{8}}\left(\frac{8}{9}\frac{10}{9}\frac{10}{11}\frac{12}{11}\frac{12}{13}\frac{14}{13}\frac{14}{15}\frac{16}{15}\right)^{\frac{1}{16}}\cdots$$

Pippenger wrote up his result as a paper...but due to the relatively ancient tradition of mathematics he was adding to - he decided to write it in Latin! The paper appears as *"Formula nova pro numero cujus logarithmus hyperbolicus unitas est"*. in IBM Research Report RC 6217. I am still trying to track down a copy of this! So if any of you are better at the internet than me, I would be very grateful if you could locate it.

Alluded to above, one of the most famous functions described by an infinite product is the sine function, which Euler expanded in his proof of the Basel sum

$$\frac{\sin \pi x}{\pi x} = \prod_{n\geq 1}\left(1-\frac{x^2}{n^2}\right)$$

AAs well as our friend the Riemann zeta function from above, which can be written as a product over all the primes! (Alluding to its deep connection to number theory)

$$\zeta(s) = \prod_{p\,\text{prime}} \frac{1}{1-p^{-s}}$$

Perhaps in a calculus class you remember seeing many formulas for the convergence of series (we will prove them here in short order), but did not see many infinite products. The reason for this is that it is enough to study one class of these recursive sequences, once we really understand exponential functions and logarithms: we can use these to convert between the two.

Proposition 18.1 (Relating Products to Series)**.** *Let $p_n = \prod_{k=0}^{n} a_k$ be a product, and L a logarithm function. Then p_n converges if and only if the series $s_n = \sum_{k=0}^{n} L(a_k)$ converges.*

Furthermore, if we can explicitly sum the series $s_n \to s$ then $p_n \to E(s)$, where E is an exponential with the same base as L.

Proof. Let p_n be a convergent product, so $p_n \to p$ and L a logarithm. Then since L is continuous, we know $L(p_n) \to L(p)$. But each term p_n is a finite product so inductively using the law of logarithms yields

$$\begin{aligned} L(p_n) &= L(p_0 p_1 p_2 \cdots p_n) \\ &= L(p_0) + L(p_1) + L(p_2) + \cdots + L(p_n) \\ &= \sum_{k=0}^{n} L(p_k) \end{aligned}$$

Thus, $\sum L(p_k)$ converges, as claimed. Now for the reverse, let E be the exponential function whose base is the same as L and assume that $s_n \sum L(p_k) \to s$ converges. Since E is continuous, we see that $E(s_n)$ is convergent. Using the law of exponents on each finite term shows

$$\begin{aligned} E(s_n) &= E\left(L(p_0) + L(p_1) + \ldots L(p_n)\right) \\ &= \prod_{k=0}^{n} E(L(p_k)) \\ &= \prod_{k=0}^{n} p_k \end{aligned}$$

Thus p_k converges, to $E(s)$. □

This is yet another reason that we should desire a rigorous theory of the transcendental functions, as the ability to turn multiplication into addition is useful in theory, just as it was in practice centuries ago.

In this chapter we begin our study of series by looking at some common types of series, and how to check if these converge or diverge. In the following chapters we will develop more powerful convergence tests, and use them to study series of functions, as well as series of numbers.

18.1. Telescoping

Definition 18.3 (Telescoping Series)**.** A telescoping series is a series $\sum a_n$ where the terms themselves can be written as *differences* of consecutive terms of another sequence, for example if $a_n = t_n - t_{n-1}$.

Telescoping series are the epitome of a math problem that looks difficult, but is secretly easy. Once you can express the terms as differences, everything but the first and last cancels out! For example:

Once a series has been identified as telescoping, often proving its convergence is straightforward: you get a direct formula for the partial sums, and then all that remains is to calculate the limit of a *sequence*.

Example 18.1. The sum $\sum_{k\geq 1} \frac{1}{k} - \frac{1}{k+1}$ telescopes. Writing out a partial sum s_n, everything collapses so $s_n = 1 - \frac{1}{n+1}$.

Now we no longer have a series to deal with, as we've found the partial sums! All that remains is the sequence $s_n = 1 - \frac{1}{n+1}$. And this limit can be computed immediately from the limit laws:

$$s = \lim s_n = 1 - \lim \frac{1}{n+1} = 1$$

Of course, sometimes a bit of algebra needs to be done to reveal that a series is telescoping:

Exercise 18.1. Show that the following series is telescoping, and then find its sum

$$\sum_{n\geq 1} \frac{4}{n^2+n}$$

Hint: factor the denominator, and do a partial fractions decomposition!

A telescoping product is defined analogously

Definition 18.4 (Telescoping Product)**.** A telescoping product is a product $\prod a_n$ where the terms themselves can be written as *ratios* of consecutive terms of another sequence, for example $a_n = \frac{t_n}{t_{n-1}}$.

Exercise 18.2. Prove that if p_n is a telescoping product, then $s_n = L(p_n)$ is a telescoping series.

18.2. Geometric Series

Definition 18.5. A series $\sum a_n$ is *geometric* if all consecutive terms share a common ratio: that is, there is some $r \in \mathbb{R}$ with $a_n/a_{n-1} = r$ for all n.

In this case we can see inductively that the terms of the series are all of the form ar^n. Thus, often we factor out the a and consider just series like $\sum r^n$.

Exercise 18.3 (Geometric Partial Sums)**.** For any real r, the partial sum of the geometric series is:

$$1 + r + r^2 + \cdots + r^n = \sum_{k=0}^{n} r^n = \frac{1 - r^{n+1}}{1 - r}$$

Like telescoping series, now that we have explicitly computed the partial sums, we can find the exact value by just taking a limit.

Theorem 18.1. *If* $|r| < 1$ *then* $\sum r^n$ *converges, and*

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1 - r}$$

Proof. By the partial sum formula, we have

$$\sum_{n \geq 0} r^n = \lim \sum_{k=0}^{n} r^n = \lim \frac{1 - r^{n+1}}{1 - r}$$

Since $|r| < 1$, we know that $r^n \to 0$, and so $r^{n+1} = rr^n \to 0$ by the limit theorems (or by truncating the first term of the sequence). Again by the limit theorems, we may then calculate

$$\lim \frac{1 - r^{n+1} 1 - r}{=} \frac{1 - \lim r^{n+1}}{1 - r} = \frac{1 - 0}{1 - r} = \frac{1}{1 - r}$$

□

Remark 18.3. Its often useful to commit to memory the formula also for when the sum starts at 1:

$$\sum_{k=1}^{\infty} r^k = \frac{r}{1 - r}$$

Exercise 18.4. Prove that for all $|r| \geq 1$, the geometric series $\sum r^n$ diverges. *Hint: use the formula for the partial sums and the limit theorems, which reduces this to the study of the sequence* r^{n+1}.

Because this holds for *all values of* r between -1 and 1, this gives us our first taste of a function defined as an infinite series. For any $x \in (-1, 1)$ we may define the function

$$f(x) = 1 + x + x^2 + x^3 + \cdots + x^n + \cdots$$

and the argument above shows that $f(x) = 1/(1-x)$. Thus, we have two expressions of the same function: one in terms of an infinite sum, and one in terms of familiar algebraic operations. This sort of thing will prove extremely useful in the future, where switching between these two viewpoints can often help us overcome difficult problems.

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + x^4 + x^5 + \cdots$$

18.2.1. Quadrature of the Parabola

We may now revisit Archimedes' other famous argument - discovering the area of the parabola.

Theorem 18.2. *The area of the segment bounded by a parabola and a chord is* $4/3^{rd}$*s the area of the largest inscribed triangle.*

After first describing how to find the largest inscribed triangle (using a calculation of the *tangent lines* to a parabola), Archimedes notes that this triangle divides the remaining region into two more parabolic regions. And, he could fill these with their largest triangles as well!

These two triangles then divide the remaining region of the parabola into *four new parabolic regions*, each of which has their own largest triangle, and so on.

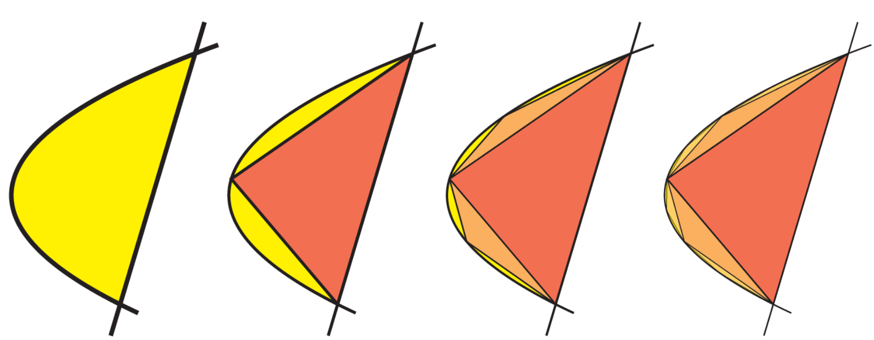


Figure 18.1.: Archimedes' infinite construction of the parabolic segment from triangles

The key geometric step to Archimedes argument is to realize that the total area of triangles added at each stage is proportional to the area of triangles added at the previous stage:

Proposition 18.2 (Area of the n^{th} stage). *The total area of the triangles in each stage is* $1/4$ *the total area of triangles in the previous stage.*

That is, if a_n is the area in the n^{th} stage, Archimedes is saying that $a_{n+1} = \frac{1}{4}a_n$. From here, the calculation step of the argument can be made rigorous with the real analysis of infinite series.

Exercise 18.5. Archimedes has defined a_n as a recursive sequence above. Use this to get an explicit formula for a_n in terms of T, the original area of the first triangle. Now, let A_n be the total area of the triangles up to the n^{th} stage. Show that this gives a geometric series, whose sum is $4/3T$.

There is a final part to this argument, that takes some more real-analysis work: since each of these triangles is a *subset* of the original parabola, the overall shape constructed from their union is also a *subset*, and so the area of the limit is less than or equal to the area of the parabola. But why is it *equal*? This requires us to show that area *missed* by each finite stage converges to zero.

Archimedes carefully works out the geometry to prove that this sequence of errors E_n must go to zero. Thus, as the area A_P of the parabola at each stage is $A_P = A_n + E_n$, and since both A_n and E_n converge we can use the limit theorems:

$$A_P = \lim(A_n + E_n) = \lim A_n + \lim E_n = \frac{4}{3}T + 0 = \frac{4}{3}T$$

Exercise 18.6 (Parabola Error (Challenge)). Try to sketch an argument for why E_n goes to zero. It's hard to write down a formula directly, as this describes the area of a shape with a curved side, and if we knew how to do that we would have sovled the entire probelm directly!

Instead, can you prove that if p is any point inside the original parabola segment, that p must be contained in the n^{th} stage of triangles for some n? Then, since at some finite stage every point can be removed from E_n, the limit of E_n is empty, which has area zero.

18.2.2. The Koch Fractal

The Koch Snowflake is a *fractal*, defined as the limit of an infinite process starting from a single equilateral triangle. To go from one level to the next, every line segment of the previous level is divided into thirds, and the middle third replaced with the other two sides of an equilateral triangle built on that side.

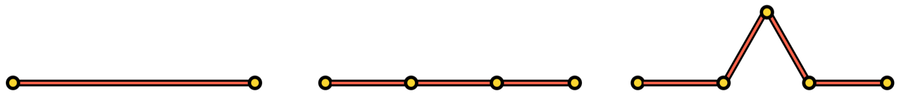


Figure 18.2.: The Koch subdivision rule: replace the middle third of every line segment with the other two sides of an equilateral triangle.

Doing this to *every line segment* quickly turns the triangle into a spiky snowflake like shape, hence the name. Denote by K_n the result of the n^{th} level of this procedure.

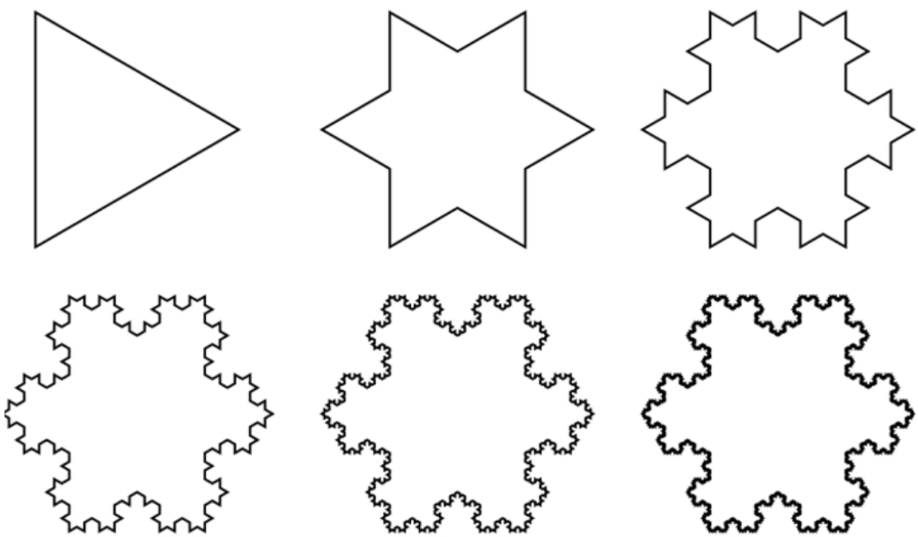


Figure 18.3.: The first six stages K_0, K_1, K_2, K_3, K_4 and K_5 of the Koch snowflake procedure. K_∞ is the fractal itself.

Say the initial triangle at level 0 has perimeter P, and area A. Then we can define the numbers P_n to be the perimeter of the n^{th} level, and A_n to be the area of the n^{th} level..

Exercise 18.7 (The Koch Snowflake Length)**.** What are the perimeters P_1, P_2 and P_3 of the first iterations? From this conjecture (and then prove by induction) a formula for the perimeter P_n and prove that P_n diverges. Thus, the limit cannot be assigned a length!

Next we turn to the area: recall that the area of an equilateral triangle can be given in terms of its side length as $A = \sqrt{3}2s^2$

Exercise 18.8 (The Koch Snowflake Area)**.** What are the areas A_1, A_2 and A_3 in terms of the original area A? Find an infinite series that represents the area of the n^{th} stage A_n, and prove that your formula is correct by induction.

Now, use what we know about geometric series to prove that this converges: in the limit, the Koch snowflake has a finite area even though its perimeter diverges!

19. Convergence

Highlights of this Chapter: Finding the value of a series explicitly is difficult, so we develop some theory to determine convergence without explicitly finding the limit. Our main tool is comparison, which is built using the Monotone convergence theorem; and in particular comparison with a geometric series - the Ratio Test. Along the way to developing this theory we study a few important special series:

- We prove the harmonic series $\sum \frac{1}{n}$ diverges.
- In contrast, we prove that the sum of reciprocal squares $\sum \frac{1}{n^2}$ *converges*. In the final project we will show its value is $\pi^2/6$.

In this section, we build up some technology to prove the convergence (and divergence) of series, without explicitly being able to compute the limit of partial sums. Such results will prove incredibly useful, as in the future we will encounter many theorems of the form *if* $\sum a_n$ *converges, then...* and we will need to a method of proving convergence to continue.

19.1. The Cauchy Criterion

For sequences, after some work we were able to find a definition equivalent to the original notion of convergence, which did not mention the precise value of the limit. This is exactly the sort of thing we seek for our investigation into series, so we carry it over directly here:

Definition 19.1 (Cauchy Criterion)**.** A series $s_n = \sum a_n$ satisfies the Cauchy criterion if for every $\epsilon > 0$ there is an N such that for any $n, m > N$ we have

$$\left|\sum_m^n a_k\right| < \epsilon$$

Exercise 19.1. Prove a series satisfies the Cauchy criterion if and only if its sequence of partial sums is a Cauchy sequence.

Because we know that being convergent and cauchy are equivalent, this means that all series that satisfy the Cauchy criterion are convergent, and conversely if a series does not, then it must diverge. We use this second observation to construct an easy-to-apply test for divergence:

Corollary 19.1 (Divergence Test)**.** *If a series* $\sum a_n$ *converges, then* $\lim a_n = 0$*. Equivalently, if* $a_n \not\to 0$ *then* $\sum a_n$ *diverges.*

Proof. Let's apply the cauchy condition to the single value m. This says for all $\epsilon > 0$ there is some N where for $m > N$ we have

$$\left|\sum_{k=m}^{m} a_k\right| = |a_m| < \epsilon$$

But making $|a_m| < \epsilon$ for all $m > N$ is exactly the definition of $a_m \to 0$. □

This is useful mostly to immediately rule out the possibility that certain series converge. For instance it tells us that $\sum(1 + \frac{1}{n})$ must diverge as the terms approach 1, not zero. But, when the terms approach zero its not very helpful: there are many series with $a_n \to 0$ which do converge, and many which diverge. To distinguish between these, we need to build up some more powerful tools.

19.1.1. Absolute Convergence

Below we will develop several theorems that apply exclusively to *series of positive terms*. That may seem at first to be a significant obstacle, as many series involve both addition and subtraction! So, we take some time here to assuage such worries, and provide a means of probing a general series using information about its nonnegative counterpart.

Definition 19.2 (Absolute Convergence)**.** A series $\sum a_n$ converges absolutely if the associated series of absolute values $\sum |a_n|$ is convergent.

Of course, such a definition is only *useful* if facts about the nonnegative series imply facts about the original. Happily, that is the case.

Theorem 19.1 (Absolute Convergence Implies Convergence)**.** *Every absolutely convergent series is a convergent series.*

Proof. Let $\sum a_n$ be absolutely convergent. Then $\sum |a_n|$ converges, and its partial sums satisfy the Cauchy criterion. This means for any ϵ we can find an N where

$$|a_n| + |a_{n+1}| + \cdots + |a_m| < \epsilon$$

But, by the triangle inequality we know that

$$|a_n + a_{n+1} + \cdots + a_n| \leq |a_n| + |a_{n+1}| + \cdots + |a_m|$$

Thus, our original series $\sum a_k$ satisfies the Cauchy Criterion, as

$$\left|\sum_{k=m}^{n} a_k\right| < \epsilon$$

And, since Cauchy is equivalent to convergence, this implies $\sum a_k$ is a convergent series. □

19.2. Comparison

One of the very most useful convergence tests for a series is comparison. This lets us show that a series we care about (that may be hard to compute with) converges or diverges by comparing it to a simpler series - much like the squeeze theorem did for us with sequences. This theorem gives less information than the squeeze theorem (it doesn't give us the exact value of the series we are interested in) but it is also easier to use (it only requires a bound, not an upper and lower bound with the same limit).

Theorem 19.2 (Comparison For Series)**.** *Let $\sum a_n$ and $\sum b_n$ be two series of nonnegative terms, with $0 \leq a_n \leq b_n$.*

- *If $\sum b_n$ converges, then $\sum a_n$ converges.*
- *If $\sum a_n$ diverges, then $\sum b_n$ diverges.*

The proof is just a rehashing of our old friend, Monotone Convergence.

Proof. We prove the first of the two claims, and leave the second as an exercise. If $x_n \geq 0$ then the series $s_n = \sum_{k=0}^{n} x_k$ is monotone increasing (as by definition $s_n = s_{n-1} + x_n$ and $x_n \geq 0$ we see $s_n \geq s_{n-1}$ for all n).

Thus. $\sum a_n$ and $\sum b_n$ are monotone sequences. If $\sum b_n$ converges, we know by the Monotone Convergence Theorem that it its limit β is the supremum of the partial sums, so for all n

$$\sum_{k=0}^{n} b_k \leq \beta$$

But, since $a_k \leq b_k$ for all k, we see the same is true of the partial sums

$$\sum_{k=0}^{n} a_k \leq \sum_{k=0}^{n} b_k$$

Stringing these inequalities together, we see that $\sum a_k$ is bounded above by β. Since it is monotone (as the sum of nonnegative terms) as well, Monotone convergence assures us that it converges, as claimed. □

Exercise 19.2. Let $\sum a_n$ and $\sum b_n$ be two series of nonnegative terms, with $0 \leq a_n \leq b_n$. Prove that if $\sum a_n$ diverges, then $\sum b_n$ diverges.

The comparison test is incredibly useful: two of the most famous series it lets us understand are left as exercises below.

Exercise 19.3. Prove that $\sum \frac{1}{n^2}$ converges. *Hint: compare with $1/((n-1)n)$, which telescopes.

Exercise 19.4. Show the harmonic series $\sum \frac{1}{n}$ diverges, by comparing it with the partial sums of

$$1, 1/2, 1/4, 1/4, 1/8, 1/8, 1/8, 1/8, 1/16, \ldots$$

19.3. The Ratio Test

We saw in the last chapter that geometric series - where the consecutive ratios of every pair of terms is constant - are particularly easy to sum. Now that we have comparison, we can leverage this to provide a powerful convergence test for a much larger collection of series: those whose consecutive rations are *constant in the limit.*

Theorem 19.3 (The Ratio Test). *Let $\sum a_n$ be a series, and assume that the sequence of consecutive ratios converges,*

$$\lim \left|\frac{a_{n+1}}{a_n}\right| = \alpha$$

Then $\sum a_n$ converges if $\alpha < 1$, and diverges if $\alpha > 1$.

Proof. We prove the convergence claim for $\alpha < 1$ here, and leave the divergence for $\alpha > 1$ as an exercise.

Assume that $\lim \left|\frac{a_{n+1}}{a_n}\right| < 1$, and let N be such that for all $n > N$ we have

$$\left|\frac{a_{n+1}}{a_n}\right| < r$$

For some fixed $0 < r < 1$ (perhaps do this by choosing your favorite $\epsilon > 0$, defining $r = 1 - \epsilon$ and using the convergence hypothesis). Thus, for all $n > N$ we know $|a_{n+1}| < r|a_n|$, and so inductively $|a_{N+n}| < r^n|a_N|$. Summing the series, we see that

$$\sum_{k=0}^{n} |a_{N+k}| < \sum_{k=0}^{n} r^k |a_N|$$

Thus, starting from the N^{th} term, our series is bounded above by a multiple of a geometric series! And, since we know geometric series converge, we can use comparison to see that $\sum_{k\geq N} |a_k|$ is convergent.

But the first finitely many terms of a series cannot affect whether or not it converges, so we see that

$$\sum_{k\geq 0} |a_k| \text{ is convergent}$$

This is the definition of $\sum a_k$ being absolutely convergent, and thus $\sum a_k$ is itself convergent. □

Exercise 19.5. Prove that if $\lim \left|\frac{a_{n+1}}{a_n}\right| > 1$, the series $\sum a_n$ diverges.

Note that this test does not tell us anything when $\alpha = 1$: it only says that our series is growing faster than a geometric series with $r < 1$ but slower than such a series with $r > 1$. There is plenty of room for both behaviors in this gap:

Example 19.1. The sequence $\sum \frac{1}{n}$ diverges, but its limiting ratio is

$$\lim \left| \frac{\frac{1}{n+1}}{\frac{1}{n}} \right| = \lim \left| \frac{n}{n+1} \right| = 1$$

But, the sequence $\sum \frac{1}{n^2}$ converges, with the same limiting ratio:

$$\lim \left| \frac{\frac{1}{(n+1)^2}}{\frac{1}{n^2}} \right| = \lim \left| \frac{n^2}{(n+1)^2} \right| = 1$$

Remark 19.1. There is an even more general version of the ratio test were we don't assume that $|a_{n+1}/a_n|$ converges, but only that it is *eventually strictly bounded above by* 1. Precisely, all that's actually required is $\lim_N \sup \left\{ \left|\frac{a_{n+1}}{a_n}\right| \mid n \geq N \right\} < 1$

Exercise 19.6. Prove that the following series converges:

$$\sum_{n\geq 0} \frac{1}{n!}$$

19.4. ⋆ Other Convergence Tests

Because series are ubiquitous throughout mathematics, there are many more convergence theorems that have been developed than we have the time to cover here. Though we will not need them in our course, I list two of the most popular (following the ratio test) below for reference.

Theorem 19.4 (The Root Test). *Let $\sum a_n$ be a series, and assume that the sequence of n^{th} roots converges,*

$$\lim \sqrt[n]{|a_n|} = \alpha$$

Then $\sum a_n$ converges if $\alpha < 1$, and diverges if $\alpha > 1$.

The following test shows up in a Calculus II course; though we are not ready to rigorously discuss it yet as it requires integration. Once we gain some ability with integrals, this will allow us to leverage our abilities with the Fundamental Theorem to prove new facts about series.

Theorem 19.5 (The Integral Test). *If f is a continuous function such that the sequence $a_n = f(n)$ is a defined by evaluating f at integer values, then the sum $\sum a_n$ converges if and only if the integral $\int_0^\infty f(x)dx$ converges.*

19.5. ⋆ Conditionally Convergent Series

Definition 19.3. A series converges *conditionally* if it converges, but is not absolutely convergent.

Such series caused much trouble in the foundations of analysis, as they can exhibit rather strange behavior. We met one such series in the introduction, the alternating sum of $1/n$ which seemed to converge to different values depending on the order we added its terms. Here we begin an investigation into such phenomena.

19.5.1. Alternating Series

Definition 19.4 (Alternating Series). An alternating series is a series of the form $\sum(-1)^n b_n$ for a_n a nonnegative series. That is, every term switches from positive to negative.

Theorem 19.6 (Alternating Series Test). *If $\sum(-1)^n a_n$ is alternating, then it converges if a_n decreases monotonically with limit zero.*

Before jumping in, its helpful to take a look at a few partial sums to start. For example, s_4:

$$s_4 = a_0 - a_1 + a_2 - a_3 + a_4 = (a_0 - a_1) + (a_2 - a_3) + a_4$$

Grouping the terms of this finite sum like so shows that s_4 is a sum of positive numbers (since a_n is decreasing, so $a_n - a_{n-1} \geq 0$): thus $s_4 \geq 0$.

$$s_4 = a_0 - a_1 + a_2 - a_3 + a_4 = a_0 - (a_1 - a_2) - (a_3 - a_4)$$

This grouping shows s_4 is equal to a_0 minus a bunch of nonnegative terms: thus $s_4 \leq a_0$. This extends directly

Exercise 19.7. Let $s_n = \sum_{k=0}^{n}(-1)^k a_k$ be an alternating series with $a_n \to 0$ monotonically. Prove by induction that

- All the partial sums s_n are nonnegative.
- All partial sums are bounded above by the first term a_0.

Corollary 19.2. *Starting the sum at N instead of 0, the same argument shows that $\left|\sum_{k=N}^{n}(-1)^k a_k\right| \leq |a_N|$ for all $n \geq N$.*

What other patterns can we notice? Increasing from s_4 to s_6 we see

$$s_6 = a_0 - a_1 + a_2 - a_3 + a_4 - a_5 + a_6$$

$$= s_4 - a_5 + a_6 = s_4 - (a_5 - a_6)$$

Thus $s_6 \leq s_4$. A similar look at s_3 and s_5 shows

$$s_5 = a_0 - a_1 + a_2 - a_3 + a_4 - a_5 = s_3 + (a_4 - a_5)$$

So $s_5 \geq s_3$! This is a sort of pattern we've seen before, where it's helpful to look at the even versus odd subsequences individually:

Exercise 19.8. Let $s_n = \sum_{k=0}^{n}(-1)^k a_k$ be an alternating series, and prove by induction that

- The even subsequence is monotone decreasing
- The odd subsequence is monotone increasing

Because each of these subsequences is monotone and bounded (by the previous exercise) they converge via monotone convergence. Now, all we need to see is they converge to the same limit to assure convergence of the entire series, by Theorem 11.2.

Proposition 19.1. *Let $s_n = \sum_{k=0}^{n}(-1)^k a_k$ be an alternating series with $a_n \to 0$ monotonically. Then s_n converges.*

Proof. Let $e_n = s_{2n}$ and $o_n = s_{2n+1}$ be the even and odd subsequences respectively, and note that $o_n = e_n - a_{2n+1}$. Then, since we know the subsequence a_{2n+1} converges to zero (as $a_n \to 0$, so all subsequences have the same limit) we can apply the limit theorems and see

$$\lim o_n = \lim e_n - a_{2n+1} = \lim e_n - \lim a_{2n+1} = \lim e_n$$

So, the odd and even subsequences do have the same limit, as required. □

19.5.2. Properties of Conditionally Convergent Series

First we look at the main example of a conditionally convergent series.

Example 19.2. $\sum \frac{(-1)^n}{n}$ is conditionally convergent:

- It converges, by the alternating series test.
- But it is not absolutely convergent, as $\sum \frac{1}{n}$ diverges by EXR

This series is famous from the introduction to our course, where we saw that its value when summed is the natural logarithm of 2, but that this value changes when we reorder the terms! This is a general behavior of conditionally convergent series; and one hint of this is that the sum of their positive and negative terms separately each *diverges* to $\pm\infty$.

Theorem 19.7. *If $\sum a_k$ is conditionally convergent, let p_k be the subsequence of all positive terms of a_k and n_k be the subsequence of all negative terms. Prove that*

$$\sum p_k \to \infty \qquad \sum n_k \to -\infty$$

For an absolutely convergent series, this cannot happen, and the sums of all the positive terms converges, as does the sum of all the negative terms.

Exercise 19.9. Prove that if $\sum a_n$ is absolutely convergent, then its subseries of positive terms and its subseries of negative terms both converge.

20. Power Series

Highlights of this Chapter: we introduce the definition of a power series, and testing for convergence via ratios.

Definition 20.1 (Power Series)**.** A power series is a function defined as the limit of a sequence of polynomials

$$f(x) = \sum_{n\geq 0} a_n x^n$$

For each x, this defines an infinite series. The domain of a power series is the subset $D \subset \mathbb{R}$ of x values where the series converges.

The simplest power series are *polynomials themselves*, which have $a_n = 0$ after some finite N. Perhaps the second simplest power series is the one with $a_n = 1$ for all n:

$$f(x) = 1 + x + x^2 + x^3 + x^4 + \cdots + x^n + \cdots$$

This is none other than the *geometric series in* x! So, it converges whenever the common ratio x satisfies $|x| < 1$: its domain is the interval $(-1, 1)$.

The domain of a general power series will always look like an interval, motivating the definition of a *radius of convergence*.

Definition 20.2. The radius of convergence of a power series is the largest $r > 0$ such that the series converges at each point of $(-r, r)$.

General power series are essentially just slight modifications the geometric series, multiplying each power of x by some coefficient. Thus, its natural to seek their radius of convergence by comparison with geometric series (the Ratio test):

Theorem 20.1 (Power Series and Radius of Convergence)**.** *If $\sum_k a_k x^k$ is a power series, let $\alpha = \lim |a_{k+1}/a_k|$. If $\alpha = 0$ then the power series converges on the entire real line. And, if $\alpha \neq 0$ then its radius of convergence is $R = 1/\alpha$: the series converges in $(-R, R)$ and diverges for $|x| > R$.*

Proof. Choose $x \in \mathbb{R}$ and compute the ratio test for the corresponding series

$$\lim \left|\frac{a_{k+1}x^{k+1}}{a_k x^k}\right| = |x| \lim \left|\frac{a_{k+1}}{a_k}\right| = |x|\alpha$$

If $\alpha = 0$ then this entire quantity is zero, independent of x. And, as $0 < 1$ the ratio test ensures the series converges. For $\alpha \neq 0$, the ratio test gives convergence if $|x|\alpha < 1$, or $|x| < 1/\alpha = R$, and divergence for $|x| > R$, as claimed. □

Inside their radius of convergence power series are always very well behaved:

Corollary 20.1. *Power series converge absolutely within their radius of convergence.*

Proof. If x is within the radius of convergence, then the ratio test on $\sum_k a_k x^k$ yields a number strictly less than 1 by definition, signaling absolute convergence. □

Since it can often be difficult to determine exactly what happens at the endpoints of the interval of convergence, where the series may converge either absolutely, conditionally, or not at all. Thus speaking of the radius (and avoiding the issue of convergence at endpoints) is often useful.

20.1. Example Power Series

Power series provide us a means of describing functions via explicit formulas that we have not been able to thus far, by allowing a limiting process in their definition. For instance, we will soon see that the power series below is an *exponential function.*

Exercise 20.1. Show the power series $\sum \frac{x^n}{n!}$ converges for all $x \in \mathbb{R}$.

When a power series converges on a finite interval, its behavior at each endpoint may require a different argument than the ratio test (as that will give 1, and tell you nothing)

Example 20.1. Show the power series $\sum \frac{x^n}{n}$ has domain $[-1, 1)$.

Exercise 20.2. Show the power series $\sum \frac{x^n}{n^2}$ has domain $[-1, 1]$.

When the radius of convergence is 0, the power series converges at a single point:

Exercise 20.3. Show the power series $\sum n!x^n$ diverges for all $x \neq 0$.

Exercise 20.4. Series $\sum 2^n x^n$ converges on $[-1/2, 1/2)$. *Hint: substitution $y = 2x$*

Example 20.2. Where does $\sum 2^n x^{3n}$ converge? *Trickier! Need to worry about the exponents not being just n*

20.2. Power Series for Functions

In the above section we discovered many new functions: its easy to define functions that no one has ever heard of by writing down new power series! And indeed, this is often how new functions are first described.

But another big use of power series will be to provide formulas for functions we already know about. At the moment we have essentially one function that we know a power series for: the geometric series

$$\sum_{k\geq 0} x^k = \frac{1}{1-x}$$

From this we can build many new functions, via substitution:

Example 20.3. A series for $\frac{1}{1+x}$ can be constructed, by substituting $-x$ for x on both sides of the equality above:

$$\begin{aligned}\frac{1}{1+x} &= \frac{1}{1-(-x)} \\ &= \sum_{k\geq 0}(-x)^k \\ &= \sum_{k\geq 0}(-1)^k x^k \\ &= 1 - x + x^2 - x^3 + x^4 - x^5 + \cdots\end{aligned}$$

Its instructive to try this with a couple examples yourself.

Exercise 20.5. A series for $\frac{1}{1+x^2}$

Exercise 20.6. A series for $\frac{x}{2+3x^2}$

In the study of calculus, we will produce much more powerful tools to find power series of functions.

21. Switching Limits

Highlights of this Chapter: we consider the delicate problem of switching the order a limit and an infinite sum. We prove a theorem - the Dominated Convergence Theorem for Sums - that provides a condition under which this interchange is allowed, and explore a couple consequences for double summations. This Dominated Convergence Theorem is the first of several analogous theorems that will play an important role in what follows.

The fact that an infinite series is *defined* as a limit - precisely the limit of partial sums - has been of great utility so far, as all of our techniques for dealing with series fundamentally rest on limit theorems for sequences!

$$\sum_{k\geq 0} a_k \;:=\; \lim_{N\to\infty}\sum_{k=0}^{N} a_k$$

But once we start to deal with multiple series at a time, this can present newfound difficulties. Indeed, it's rather common in practice to end up with an *infinite sequence of infinite series.*

For example, imagine that a function $f(x)$ is defined by a power series $f(x) = \sum_{k\geq 0} a_k x^k$. If $a \in \mathbb{R}$ is some point in its domain, how could we hope to test continuity of f at a? Using the sequence definition of continuity, all we need to do is choose a sequence $x_n \to a$ and attempt to evaluate $\lim f(x_n)$. But for each n we know $f(x_n)$ is defined as an infinite series! Thus, we are forced to deal with taking a limit of series - a limit of limits.

$$\lim_n f(x_n) = \lim_n \sum_{k\geq 0} a_k x_n^k = \lim_n \lim_N \sum_{k=0}^{N} a_k x_n^k$$

There's an intuitive urge to just switch the order of the limits - equivalently, to "pull the limit inside the sum". But such an operation is not always justified. Its easy to come up with examples of limits that cannot be switched:

$$\lim_n \lim_m \frac{n}{n+m} = \lim_n \left(\lim_m \frac{n}{n+m}\right) = \lim_n 0 = 0$$

$$\lim_m \lim_n \frac{n}{n+m} = \lim_m \left(\lim_n \frac{n}{n+m} \right) = \lim_m 1 = 1$$

Even worse (for us) this behavior can manifest even when dealing with series

Example 21.1.

$$\begin{aligned} 1 &= \frac{1}{2} + \frac{1}{2} \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \end{aligned}$$

Taking the termwise limit and adding them up gives

$$1 = 0 + 0 + 0 + \cdots + 0 = 0$$

This is nonsense! And the nonsense arises from implicitly *exchanging two limits.* To make this precise, one may define for each n the series

$$a_n(k) = \begin{cases} 1/2^n & 0 \leq k < 2^n \\ 0 & \text{else} \end{cases}$$

Then each of the rows above is the sum $1 = \sum_{k\geq 0} a_n(k)$ for $n = 2, 3, 4$. Since this is constant it is true that the limit is 1, but it is *not true* that the limit of the sums is the sum of the limits, which is zero.

$$1 = \lim_n \sum_{k\geq 0} a_n(k) \neq \sum_{k\geq 0} \lim a_n(k) = 0$$

So, its hopefully clear that to be able to use series in realistic contexts, we are in desparate need of a theorem which tells us when we can interchange limits and summantions.

21.1. Dominated Convergence (Tannery's Theorem)

Because limit interchange is so fundamental to analysis, there are many theorems of this sort, of varying strengths and complexities. The one we will visit here is usually called Tannery's theorem (named for Jules Tannery, an analyst at the end of the 1800s). With the luxury of hindsight, we now realize Tannery's theorem is a particularly special case of a much more general result called Dominated Convergence, of which we will meet other special cases in the chapters to come. As such, I will call it by its more descriptive and general name throughout.

First, let's set the stage precisely. For each n, we have an infinite series s_n, and we are interested in the limit $\lim_n s_n$ (here, we will always write subscripts on the limit as multiple variables are involved!) For each fixed n, the series s_n is an infinite sum, over some summation index k:

$$s_n = \sum_{k\geq 0} a_k(n)$$

Where for each k we write the term as $a_k(n)$ to remember that it also depends on n (the notation $a_{k,n}$ is also perfectly acceptable). We seek a theorem that gives us the conditions on which we can take the term-wise limit, that is when

$$\lim_n \sum_{k\geq 0} a_k(n) = \sum_{k\geq 0} \lim_n a_k(n)$$

Dominated convergence assures us that such a switch is justified so long as the entire process - *all of the* $a_k(n)$*s* are bounded by a convergent series.

Theorem 21.1 (Dominated Convergence for Series)**.** *For each k let $a_k(n)$ be a function of n, and assume the following:*

- *For each k, $a_k(n)$ is convergent.*
- *For each n, $\sum_k a_k(n)$ is convergent.*
- *There is an M_k with $|a_k(n)| \leq M_k$ for all n.*
- *$\sum M_k$ is convergent.*

Then $\sum_k \lim_n a_k(n)$ is convergent, and

$$\lim_n \sum_k a_k(n) = \sum_k \lim_n a_k(n)$$

Proof. First, we show that $\sum_k a_k$ converges. Since for all n, $|a_k(n)| \leq M_k$ we know this remains true in the limit, so $\lim_n |a_k(n)| = |a_k| < M_k$. Thus, by comparison we see $\sum_k |a_k|$ converges, and hence so does $\sum_k a_k$.

Now, the main event. Let $\epsilon > 0$. To show that $\lim_n \sum_k a_k(n) = \sum_k a_k$ we will show that there there is some N beyond which these two sums always differ by less than ϵ.

Since $\sum_k M_k$ converges, by the Cauchy criterion there is some L where

$$\sum_{k\geq L} M_k < \frac{\epsilon}{3}$$

For arbitrary n, we compute

$$\begin{aligned}
\left|\sum_{k\geq 0} a_k(n) - \sum_{k\geq 0} a_k\right| &= \left|\sum_{k<L}(a_k(n)-a_k) + \sum_{k\geq L} a_k(n) + \sum_{k\geq L} a_k\right| \\
&\leq \left|\sum_{k<L}(a_k(n)-a_k)\right| + \left|\sum_{k\geq L} a_k(n)\right| + \left|\sum_{k\geq L} a_k\right| \\
&\leq \sum_{k<L} |a_k(n)-a_k| + \sum_{k<L} |a_k(n)| + \sum_{k\geq L} |a_k| \\
&\leq \sum_{k<L} |a_k(n)-a_k| + 2\sum_{k>L} M_k \\
&< \sum_{k<L} |a_k(n)-a_k| + \frac{2\epsilon}{3}
\end{aligned}$$

That is, for an arbitrary n we can bound the difference essentially in terms of the first L terms: the rest are uniformly less than $2\epsilon/3$. But for each of these L terms, we know that $a_k(n) \to a_k$ so we can find an N making that difference as small as we like. Let's choose N_k such that $|a_k(n) - a_k| < \epsilon/3L$ for each $k < L$ and then take

$$N = \max\{N_0, N_1, \dots N_{L-1}\}$$

Now, for any $n > N$ we are guaranteed that |a_k(n)-a_k|<⍰/3L$ and thus that

$$\sum_{k<L} |a_k(n) - a_k| < L\frac{\epsilon}{3L} = \frac{\epsilon}{3}$$

Combining with the above, we now have for all $n > N$,

$$\left|\sum_{k\geq 0} a_k(n) - \sum_{k\geq 0} a_k\right| < \epsilon$$

as required. □

And, a direct generalization to limits of functions (which are after all defined in terms of sequences!)

Theorem 21.2 (Dominated Convergence for Function Limits). *For each k, let $f_k(x)$ be a function of x on a domain D. For a fixed $a \in \mathbb{R}$, assume there is some interval $I \subset D$ containing a such that:*

- *For each k, $\lim_{x\to a} f_k(x)$ exists.*
- *$\sum_k f_k(x)$ is convergent for each $x \in I$.*
- *There is an M_k with $|f_k(x)| \leq M_k$ for all $x \in I$.*

- *$\sum M_k$ is convergent.*

Then, the sum $\sum_k \lim_{x\to a} f_k(x)$ is convergent and

$$\lim_{x\to a}\sum_k f_k(x) = \sum_k \lim_{x\to a} f_k(x)$$

Proof. Let $x_n \subset I$ be an arbitrary sequence with $x_n \to a$ and $x_n \neq a$. We assumed $\lim_{x\to a} f_k(x)$ exists. As $x_n \to a$, it follows by definition that $\lim_n f_k(x_n) = \lim_{x\to a} f_k(x)$, so this limit also exists, and so (1) holds. Additionally for each fixed n, $\sum_k f_k(x_n)$ is convergent, as $x_n \in I$ and we assumed convergence for each $x \in I$.

As we assumed M_k bounds $|f_k(x)|$ for all $x \in I$ it also does so for all x_n in our sequence, so (3) and (4) are satisfied for the original dominated convergence, Theorem 21.1. Thus, we may conclude that the series $\sum_k \lim_n f_k(x_n)$ is convergent, and that

$$\lim_n \sum_k f_k(x_n) = \sum_k \lim_n f_k(x_n) = \sum_k \lim_{x\to a} f_k(x)$$

Because x_n was arbitrary, this applies for *all* sequences $x_n \to a$ with $x_n \neq a$. Thus, the overall limit $\lim_{x\to a}\sum_k f_k(x)$ exists, and is equal to this common value

$$\lim_{x\to a}\sum_k f_k(x) = \sum_k \lim_{x\to a} f_k(x)$$

□

There is a natural version of this theorem for products as well (though we will not need it in this course, I will state it here anyway)

Theorem 21.3 (★ Dominated Convergence for Products). *For each k let $a_k(n)$ be a function of n, and assume the following:*

- *For each k, $a_k(n)$ is convergent.*
- *For each n, $\prod_{k\geq 0} a_k(n)$ is convergent.*
- *There is an M_k with $|a_k(n)| \leq M_k$ for all n.*
- *$\sum M_k$ is convergent.*

Then $\prod_{k\geq 0}\lim_n(1+a_k(n))$ is convergent, and

$$\lim_n \prod_{k\geq 0}(1+a_k(n)) = \prod_{k\geq 0}(1+\lim_n a_k(n))$$

Exercise 21.1. Use Dominated Convergence to prove that

$$\frac{1}{2} = \lim_n\left[\frac{1+2^n}{2^n\cdot 3+4} + \frac{1+2^n}{2^n\cdot 3^2+4^2} + \frac{1+2^n}{2^n\cdot 3^3+4^3} + \cdots\right]$$

- Write in summation notation, and give a formula for the terms $a_k(n)$
- Show that $\lim_n a_k(n) = \frac{1}{3^k}$
- Show that for all n, $|a_k(n)| \leq \frac{2}{3^k}$

Use these facts to show that the hypotheses of dominated convergence hold true, and then use the theorem to help you take the limit.

21.2. Application: Continuity of Power Series

We will find several applications for dominated convergence during our study of calculus, proving analogs for both derivatives () and integrals (). But our most immediate application is to the problem of continuity of power series originally posed at the beginning of this section: we can now easily prove that every power series is continuous on the interior of its interval of convergence.

Theorem 21.4 (Continuity within Radius of Convergence). *Let $f(x) = \sum_k a_k x^k$ be a power series with radius of convergence r. Then if $|x| < r$, f is continuous at x.*

Proof. Without loss of generality take $x > 0$, and let x_n be an arbitrary sequence in $(-r, r)$ converging to x. We aim to show that $f(x_n) \to f(x)$.

As $x < r$ choose some y with $x < y < r$ (perhaps, $y = (x + r)/2$). Since $x_n \to x$ there is some N past which x_n is always less than y (take $\epsilon = y - x$ and apply the definition of $x_n \to x$). As truncating the terms of the sequence before this does not change its limit, we may without loss of generality assume that $x_n < y$ for all n. Thus, we may define $M_k = a_k y^k$, and we are in a situation to verify the hypotheses of Dominated Convergence:

- Since $x_n \to x$, we have $a_k x_n^k \to a_k x^k$ by the limit theorems.
- For each n, $f(x_n) = \sum_k a_k x_n^k$ is convergent as x_n is within the radius of convergence.
- $M_k = a_k y^k$ bounds $a_k x_n^k$ for all n, as $0 < x_n < y$.
- $\sum_k M_k$ converges as this is just $f(y)$ and y is within the radius of convergence.

Applying the theorem, we see

$$\lim_n f(x_n) = \lim_n \sum_k a_k x_n^k = \sum_k \lim a_k x_n^k = \sum_k a_k x^k = f(x)$$

Thus for arbitrary $x_n \to x$ we have $f(x_n) \to f(x)$, so f is continuous at x. □

21.3. Application: Double Sums

Another useful application of dominated convergence is to switching the order of a double sum. A *double sequence* is a map $\mathbb{N}\times\mathbb{N}\to\mathbb{R}$, where we write $a_{m,n}$ for the value $a(m,n)$. Such sequences like $n/(n+m)$ occured in our original example about switching limits above.

Given a double sequence, one may want to define an *double sum*

$$\sum_{m,n\geq 0} a_{m,n}$$

But, how should one do this? Because we have two indices, there are two possible orders we could attempt to compute this sum:

$$\sum_{n\geq 0}\sum_{m\geq 0} a_{m,n} \qquad \text{or} \qquad \sum_{m\geq 0}\sum_{n\geq 0} a_{m,n}$$

Definition 21.1 (Double Sum)**.** Given a double sequence $a_{m,n}$ its double sum $\sum_{m,n\geq 0} a_{m,n}$ is defined if both orders of iterated summation converge, and are equal. In this case, the value of the double sum is defined to be their common value:

$$\sum_{m,n\geq 0} a_{m,n} := \sum_{n\geq 0}\sum_{m\geq 0} a_{m,n} = \sum_{m\geq 0}\sum_{n\geq 0} a_{m,n}$$

We should be worried from previous experience that in general these two things need not be equal, so the double sum may not exist! Indeed, we can make this worry precise, by seeing that to relate one to the other is really an *exchange of order of limits*:

$$\sum_{m\geq 0} = \lim_M \sum_{0\leq m\leq M} \qquad \sum_{n\geq 0} = \lim_N \sum_{0\leq n\leq N}$$

And so, expanding the above with these definitions (and using the limit laws to pull a limit out of a finite sum) we see

$$\sum_{n\geq 0}\sum_{m\geq 0} a_{m,n} = \lim_N \sum_{0\leq n\leq N}\left(\lim_M \sum_{0\leq m\leq M} a_{m,n}\right)$$

$$= \lim_N \lim_M \left(\sum_{0\leq n\leq N}\sum_{0\leq m\leq M} a_{m,n}\right) = \lim_N \lim_M \sum_{\substack{0\leq m\leq M\\ 0\leq n\leq N}} a_{m,n}$$

Where in the final line we have put both indices under a single sum to indicate that it is a finite sum, and the order does not matter. Doing the same with the other order yields the exact same finite sum, but with the order of limits reversed:

$$\sum_{m\geq 0}\sum_{n\geq 0} a_{m,n} = \lim_M \lim_N \sum_{\substack{0\leq m\leq M\\ 0\leq n\leq N}} a_{m,n}$$

Because this is an exchange-of-limits-problem, we can hope to provide conditions under which it is allowed using Tannery's theorem.

Theorem 21.5. *Let $a_{m,n}$ be a double sequence, and assume that either*

$$\sum_{m\geq 0}\sum_{n\geq 0} |a_{m,n}| \qquad \textit{or} \qquad \sum_{n\geq 0}\sum_{m\geq 0} |a_{m,n}|$$

converges. Then the double sum also converges

$$\sum_{m,n\geq 0} a_{m,n}$$

(meaning either both orders of iterated sum converge, and are equal)

Exercise 21.2 (Cauchy's Double Summation Formula)**.** Use Dominated Convergence to prove the double summation formula (Theorem 21.5): without loss of generality, assume that $\sum_{m\geq 0}\sum_{n\geq 0}|a_{m,n}|$ converges, and use this to show that both orders of iterated sum converge and are equal

$$\sum_{m\geq 0}\sum_{n\geq 0} a_{m,n} = \sum_{n\geq 0}\sum_{m\geq 0} a_{m,n}$$

Hint: Assuming $\sum_{m\geq 0}\sum_{n\geq 0}|a_{m,n}|$ converges, set $M_m = \sum_{n\geq 0}|a_{m,n}|$ and show the various hypotheses of Dominated convergence apply

Exercise 21.3 (Applying the Double Sum)**.** Since switching the order of limits involves commuting terms that are arbitrarily far apart, techniques like double summation allow one to prove many identities that are rather difficult to show directly. We will make a crucial use of this soon, in understanding exponential functions. But here is a first example:

For any $k \in \mathbb{N}$, prove the following equality of infinite sums:

$$\frac{z^{1+k}}{1-z} + \frac{(z^2)^{1+k}}{1-z^2} + \frac{(z^3)^{1+k}}{1-z^3} + \cdots = \frac{z^{1+k}}{1-z^{1+k}} + \frac{z^{2+k}}{z^{2+k}} + \frac{z^{3+k}}{1-z^{3+k}} + \cdots$$

Hint: first write each side as a summation:

$$\sum_{n\geq 1} \frac{z^{n(k+1)}}{1-z^n} = \sum_{m\geq 1} \frac{z^{m+k}}{1-z^{m+k}}$$

*Then setting $a_{m,n} = z^{n(m+k)}$, show that Cauchy summation applies to the double sum $\sum_{m,n} \geq 0 a_{m,n}$ and compute the sum in each order, arriving that the claimed equality.

Part VI.

Derivatives

22. Definition

Highlights of this Chapter: we define the derivative and compute a few examples directly from the definition.

Finally - on to some calculus! Here we will define the derivative, and study its properties. This may sound daunting at first, remembering back to the days of calculus when it all seemed so new and advanced. But hopefully, after so much exposure to sequences and series during this course, the rigorous notion of a derivative will feel more just like a nice application of what we've learned, than a whole new theory.

Definition 22.1 (The Derivative)**.** Let f be a function defined on an open interval containing a. Then f is differentiable at a if the following limit of difference quotients exists. In this case, we define the limiting value to be the *derivative of f at a*.

$$f'(a) = Df(a) = \lim_{t \to a} \frac{f(t) - f(a)}{t - a}$$

Exercise 22.1 (Equivalent Formulation)**.** Prove that we may alternatively use the following limit definition to calculate the derivative:

$$f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$

Example 22.1. The function $f(x) = x^2$ is differentiable at $x = 2$.

This is a classic problem from calculus 1, whose argument is already pretty much rigorous! We wish to compute the limit

$$\lim_{x \to 2} \frac{x^2 - 4}{x - 2}$$

So, we choose an arbitrary sequence x_n with $x_n \neq 2$ but $x_n \to 2$ and compute

$$\lim \frac{x_n^2 - 4}{x_n - 2} = \lim \frac{(x_n + 2)(x_n - 2)}{x_n - 2} = \lim x_n + 2$$

Where the arithmetic is justified since $x_n \neq 2$ for all n by definition, so everything is defined. But now, as $x_n \to 2$ we can just use the limit laws to see

$$\lim x_n + 2 = 2 + 2 = 4$$

Since x_n was arbitrary, this holds for all such sequences, so the limit exists and equals 4. Because this limit defines the derivative, we have that f is differentiable at 2 and

$$f'(2) = 4$$

Exercise 22.2. Compute the derivative of $f(x) = x^3$ at an arbitrary point $a \in \mathbb{R}$, directly from the definition and show $f'(a) = 3a^2$.

22.0.1. Multiple Derivatives

Once we've defined the derivative, as a limit, it's easy to iterate to define multiple derivatives:

Definition 22.2. If f is differentiable on a domain $D \subset \mathbb{R}$, and the function f' itself is differentiable at a point $a \in D$, we call the resulting derivative the *second derivative* of f:

$$f''(a) := (f')'(a) = \lim_{t \to a} \frac{f'(t) - f'(t)}{t - a}$$

Proposition 22.1. *Assume that f is twice differentiable at a point a of its domain. Then $f''(a)$ can be calculated via the following limit:*

$$f''(a) = \lim_{h \to 0} \frac{f(x + 2h) - f(x + h) + f(x)}{h^2}$$

22.1. Using the Definition

Derivative is defined as a limit: one of our most useful tools is Theorem 17.2, which tells us we can detect differentiability by comparing the one-sided limits.

Example 22.2. The function $f(x) = |x|$ is not differentiable at $x = 0$.

The difference quotient defining the derivative is

$$\lim_{x \to 0} \frac{|x| - 0|}{x} = \frac{|x|}{x}$$

When $x > 0$ we have $|x| = x$ and so $\frac{|x|}{x} = 1$. Thus, the rihgt hand limit for any sequence $x_n > 0$ with $x_n \to 0$ is just the limit of the constant sequence 1, and

$$\lim_{x\to 0^+} \frac{|x|}{x} = 1$$

However, for $x < 0$ we have $|x| = -x$, and analogous reasoning shows

$$\lim_{x\to 0^-} \frac{|x|}{x} = -1$$

Since these are unequal, the overall limit cannot exist, and thus the function $|x|$ is not differentiable at zero.

It is sometimes useful to define one-sided derivatives, much as we defined one-sided limits (particularly, to discuss differentiability at the endpoint of an interval, for example)

Definition 22.3 (One Sided Derivatives)**.**

$$D^- f(a) = \lim_{x\to a^-} \frac{f(x) - f(a)}{x - a} \qquad D^+ f(a) \lim_{x\to a^+} \frac{f(x) - f(a)}{x - a}$$

Here's the differentiable analog of the pasting lemma you recently proved on homework:

Exercise 22.3. Let f, g be two continuous and differentiable functions with $a \in \mathbb{R}$ a point such that $f(a) = g(a)$. Prove that the piecewise function

$$h(x) = \begin{cases} f(x) & x \leq a \\ g(x) & x > a \end{cases}$$

is differentiable at a *if and only if* $f'(a) = g'(a)$. (recall we saw such a function is always continuous at a in Exercise 17.5).

One sided derivatives let us more easily prove that the derivative exists in cases where it is easy to take limits from above and below, but not arbitrary limits. A great example use case is when the difference quotient is *monotone*: then the right and left limits exist Exercise 17.6 (they are the inf and sup for any sequence, respectively). When is the difference quotient monotone? One particularly useful case: this holds whenever the function is convex (Proposition 13.1)

Corollary 22.1 (★ Derivatves and Convexity)**.** *If f is convex then at any point $a \in \mathbb{R}$ the one sided difference quotients $D^- f(a)$ and $D^+ f(a)$ both exist.*

Exercise 22.4. These one sided difference quotients need not be equal, however. Prove the convex function $f(x)$ below is not differentiable at $x = 1$:

$$f(x) = \begin{cases} x & x \leq 1 \\ x^2 & x > 1 \end{cases}$$

22.2. The Derivative as a Function

Definition 22.4. Let f be a function, and suppose that the derivative of f exists at each point of a set $D \subset \mathbb{R}$. Then we may define a function $f' : D \to \mathbb{R}$ by

$$f' : x \mapsto f'(x) = \lim_{t \to x} \frac{f(t) - f(x)}{t - x}$$

If f' is continuous, f is called *continuously differentiable* on D.

For example, $f(x) = x^3$ is continuously differentiable on $\mathbb{R}$ since by Exercise 22.2 we see its derivative is the function $x \mapsto 3x^2$, and this is a polynomial: we proved all polynomials are continuous in Exercise 14.7.

Example 22.3. While its hard to imagine a function that is differentiable at every point but *not continuously differentiable* such things exist. For example

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x^2}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

Its possible to find a formula for $f'(x)$ when $x \neq 0$, and show that $\lim_{x \to 0} f'(x)$ does not exist (similar to the previous exercise on $\sin \frac{1}{x}$). However one can also calculate directly the derivative *at zero*: and find $f'(0) = 0$. This means $\lim_{x \to 0} f'(x) \neq f'(\lim_{x \to 0} x)$ as one side does not exist and the other is zero: thus f' is not continuous at 0.

Exercise 22.5. For $f(x)$ as above in Example 22.3, calculate $f'(0)$ directly using the limit definition. (Perhaps surprisingly, all you need to know about the sine function here is that it is bounded between -1 and 1!)

23. Properties

> Highlights of this Chapter: we prove many foundational theorems about the derivative that one sees in an early calculus course. We see how to take the derivative of scalar multiples, sums, products, quotients and compositions. We also compute - directly from the definition - the derivative of exponential functions. This leads to an important discovery: there is a unique *simplest*, or *natural exponential*, whose derivative is itself. This is the origin of e in Analysis.

From the definition, we move on to confirm the basic properties of the derivative well known and loved in introductory calculus courses. Most of these are straightforward, the only exception whose proof requires more thought than usually let on in Calculus I is the chain rule.

However before jumping in we prove one small oft-useful result often not mentioned in a calculus class, relating differentiability to continuity.

The converse
the absolute
differentiable

Proposition 23.1. *Let f be differentiable at $a \in \mathbb{R}$. Then f is continuous at a.*

Proof. Since f is differentiable at a, we know the limit of the difference quotient is finite

$$\lim_{x\to a}\frac{f(x)-f(a)}{x-a} = f'(a)$$

We also know that $\lim_{x\to a}(x-a) = 0\$$ So, using the limit theorems we may multiply these together and get what we want. Precisely, let $x_n \to a$ be any sequence with $x_n \neq a$ for all n. Then we have

$$\begin{aligned}
0 &= (0)(f'(a)) \\
&= (\lim x_n - a)\left(\lim \frac{f(x_n)-f(a)}{x_n-a}\right) \\
&= \lim\left((x_n-a)\frac{f(x_n)-f(a)}{x_n-a}\right) \\
&= \lim\left(f(x_n)-f(a)\right)
\end{aligned}$$

Thus $\lim(f(x_n) - a) = 0$ so by the limit theorems we see $\lim f(x_n) = a$. Since x_n was arbitrary with $x_n \neq a$ this holds for any such sequence, we see that f is continuous at a using the sequence definition. □

There is a little gap not explicitly spelled out at the end of the proof above, that we should fill in now (to assure ourselves this style of reasoning always works). We just proved that for sequences $x_n \neq a$ the property we want holds, but continuity requires this fact for *all arbitrary sequences*. How do we bridge this gap? Let $y_n \to a$ be an arbitrary sequence: then we split into the subsequences $x_n \neq a$ and the subsequence of all terms $= a$. If either of these is finite, we can just truncate the original sequence at a point past which all terms are of one or the other: each of these has $\lim f(x_n) = f(a)$ so we are done. In the case that both are infinite, we just use that we have separated our sequence into a union of two subsequences, each with the same limit! Thus the overall limit exists.

23.1. Differentiation and Field Operations

Here we prove the 'derivative laws' of Calculus I:

Theorem 23.1. *Let f be a function and $c \in \mathbb{R}$. Then if f is differentiable at a point $a \in \mathbb{R}$ so is cf, and*

$$(cf)'(a) = c\,(f'(a))$$

Proof. Let's use the difference quotient with $a + h_n$ to change things up: Let $h_n \to 0$ be arbitrary, and we wish to compute the limit

$$\lim \frac{cf(a+h_n) - cf(a)}{h_n}$$

By the limit laws we can pull out the constant c, and the remainder converges to $f'(a)$, as f is assumed to be differentiable at a.

$$= c \lim \frac{f(a+h_n) - f(a)}{h_n} = cf'(a)$$

Because this is true for all sequences $h_n \to 0$ with $h_n \neq 0$, the limit exists, and equals $cf'(a)$. □

Exercise 23.1. Let f, g be functions which are both differentiable at a point $a \in \mathbb{R}$. Then $f + g$ is also differentiable at a, and

$$(f+g)'(a) = f'(a) + g'(a)$$

Theorem 23.2 (The Product Rule)**.** *Let f, g be functions which are both differentiable at a point $a \in \mathbb{R}$. Then fg is differentiable at a and*

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a)$$

Proof. Let f, g be differentiable at $a \in \mathbb{R}$, and choose an arbitrary sequence $a_n \to a$. Then we wish to compute

$$\lim \frac{f(a_n)g(a_n) - f(a)g(a)}{a_n - a}$$

To the numerator we add $0 = f(a_n)g(a) - f(a_n)g(a)$ and regroup with algebra:

$$= \lim \frac{f(a_n)g(a_n) - f(a_n)g(a) + f(a_n)g(a) - f(a)g(a)}{a_n - a}$$
$$= \lim \frac{f(a_n)g(a_n) - f(a_n)g(a)}{a_n - a} + \frac{f(a_n)g(a) - f(a)g(a)}{a_n - a}$$

Using the limit laws, we can take each of these limits individually so long as they exist (which we will show they do). But even more, note that the first term has a common factor of $f(a_n)$ in the numerator that can be factored out, and the second a common factor of $g(a)$. Thus, by the limit laws, we see

$$= (\lim f(a_n)) \left(\lim \frac{g(a_n) - g(a)}{a_n - a} \right) + g(a) \left(\frac{f(a_n) - f(a)}{a_n - a} \right)$$

Because f is differentiable at a, its continuous at a, and so we know $\lim f(a_n) = f(a)$. The other two limits above converge to the derivatives $f'(a)$ and $g'(a)$ respectively. Thus, alltogether we find the resulting limit to be

$$f(a)g'(a) + f'(a)g(a)$$

As this was the result for an arbitrary sequence $a_n \to a$ with $a_n \neq a$, it must be the same for all sequences, meaning the limit exists, and

$$(f \cdot g)'(a) = f(a)g'(a) + f'(a)g(a)$$

□

Exercise 23.2 (The Reciprocal Rule)**.** Let f be a function and $a \in \mathbb{R}$ be a point such that $f(a) \neq 0$ and f is differentiable at a. Then $1/f$ is also differentiable at a and

$$\left(\frac{1}{f}\right)'(a) = \frac{-f'(a)}{f(a)^2}$$

Theorem 23.3 (The Quotient Rule)**.** *Let f, g be a functions which are differentiable at a point $a \in \mathbb{R}$ and assume $g(a) \neq 0$. Then the function f/g is also differentiable at a and*

$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2}$$

Exercise 23.3. Use the Reciprocal Rule and Product Rule to prove the quotient rule.

23.2. The Chain Rule

Theorem 23.4 (The Chain Rule)**.** *If $g(x)$ is differentiable at $a \in \mathbb{R}$ and $f(x)$ is differentiable at $g(a)$ then the composition $f \circ g$ is differentiable at a, with*

$$(f \circ g)'(a) = f'(g(a))g'(a)$$

Wish this Worked! We are taking the derivative at a, so let $x_n \to a$ wtih $x_n \neq a$ be arbitrary. Then the limit defining $[f(g(a))]'$ is

$$\lim \frac{f(g(x_n)) - f(g(a))}{x_n - a}$$

We multiply the numerator and denominator of this fraction by $\$g(x_n) - g(a)$ and regroup:

$$\begin{aligned}
\frac{f(g(x_n)) - f(g(a))}{x_n - a} &= \frac{f(g(x_n)) - f(g(a))}{x_n - a}\frac{g(x_n) - g(a)}{g(x_n) - g(a)} \\
&= \lim \frac{f(g(x_n)) - f(g(a))}{g(x_n) - g(a)}\frac{g(x_n) - g(a)}{x_n - a}
\end{aligned}$$

Because g is continuous at a, we know $g(x_n) \to a$, and because f is differentiable at $g(a)$ we recognize the first term here as the *limit defining f' at $g(a)$*! Since the second term is the limit defining the derivative of g, both of these exist by our assumptions, and so by the limit theorems we can compute

$$= \left(\lim \frac{f(g(x_n)) - f(g(a))}{g(x_n) - g(a)}\right)\left(\lim \frac{g(x_n) - g(a)}{x_n - a}\right)$$

$$= f'(g(a))g'(a)$$

□

Unfortunately, this proof fails at one crucial step! Wile we do know that $x_n - a \neq 0$ (in the definition of $\lim_{x\to a}$, we only choose sequences $x_n \to a$ with $x_n \neq a$) we do *not* know that the other denominator $g(x_n) - g(a)$ is nonzero.

If this problem could only happen *finitely many times* it would be no trouble - we could just truncate the beginning of our sequence and rest assured we had not affected the value of the limit. But functions - even differentiable functions - can be pretty wild. The function $x^2\sin(1/x)$ (from Example 22.3) ends up equaling zero infinitely often in any neighborhood of zero! So such things are a real concern.

Happily the fix - while tedious - is straightforward. It's given below.

Exercise 23.4. We define the auxiliary function $d(y)$ as follows:

$$d(y) = \begin{cases} \frac{f(y)-f(g(a))}{y-g(a)} & y \neq g(a) \\ f'(g(a)) & y = g(a) \end{cases}$$

This function equals our problematic difference quotient most of the time, but equals the quantity we *want it to be* when the denominator is zero.

Prove that d is continuous at $g(c)$ and we may use d in place of the difference quotient in our computation: that for all $x \neq a$, the following equality holds:

$$\frac{f(g(x)) - f(g(a))}{x-a} = d(g(x))\frac{g(x)-g(a)}{x-a}$$

Given this, the original proof is rescued:

Proof. We are taking the derivative at a, so let $x_n \to a$ with $x_n \neq a$ be arbitrary. Then the limit defining $[f(g(a))]'$ is (by the exercise)

$$\lim \frac{f(g(x_n)) - f(g(a))}{x_n - a} = \lim d(g(x_n))\frac{g(x_n)-g(a)}{x_n - a}$$

Because d is continuous at $g(a)$ and $g(x_n) \to g(a)$ we know $d(g(x_n)) \to d(g(a)) = f'(g(a))$. And, as g is differentiable at a we know the limit of the difference quotient exists. Thus, by the limit laws we can separate them and

$$= (\lim d(g(x_n)))\left(\frac{g(x_n)-g(a)}{x_n-a}\right) = f'(g(a))g'(a)$$

□

23.3. Exponentials and Logs

In this section we look at how to find derivatives of functions which are defined not explicitly, but by *functional equations.* We will take the exponential as our example case; on the final project you will analyze the trigonometric functions this way.

Proposition 23.2. *Let $E(x)$ be an exponential function. Then E is differentiable on the entire real line, and*

$$E'(x) = E'(0)E(x)$$

First we show that this formula holds so long as E is actually differentiable at zero. Thus, differentiability at a single point is enough to ensure differentiability everywhere *and* fully determine the formula!

Proof. Let $x \in \mathbb{R}$, and $h_n \to 0$. Then we compute $E'(x)$ by the following limit:

$$E'(x) = \lim \frac{E(x+h_n) - E(x)}{h_n}$$

Using the property of exponentials and the limit laws, we can factor an $E(x)$ out of the entire numerator:

$$= \lim \frac{E(x)E(h_n) - E(x)}{h_n} = E(x) \lim \frac{E(h_n) - 1}{h_n}$$

But, $E(0) = 1$ so the limit here is actually the *derivative of E at zero$!

$$E'(x) = E(x)E'(0)$$

□

Next, we tackle the slightly more subtle problem of showing that E is in fact differentiable at zero. This is tricky because *all we have assumed* is that E is continuous and satisfies the law of exponents: how are we going to pull differentiability out of this? The trick is two parts (1) show the right and left hand limits defining the derivative exist, and (2) show they're equal.

ts the natural
ove this later)

Proof. STEP 1: Show that the left and right hand limits defining the derivative exist: E is convex (Exercise 15.3) so the difference quotient is monotone increasing (Proposition 13.1), and so the limit $\lim_{x\to 0^-}$ exists (as a sup) and $\lim_{x\to 0^+}$ exists (as an inf), Corollary 22.1.

STEP2: Now that we know each of these limits exist, let's show they are equal using the definition:

To compute the lower limit, we can choose any sequence approaching 0 from below: let h_n be a positive sequence with $h_n \to 0$, then $-h_n$ will do:

$$\lim_{h\to 0^-}\frac{E(h)-1}{h} = \lim\frac{E(-h_n)-1}{-h_n}$$

And by Exercise 15.2 we see $E(-h_n) = 1/E(h_n)$. Thus

$$\begin{aligned}\lim\frac{E(-h_n)-1}{-h_n} &= \lim\frac{\frac{1}{E(h_n)}-1}{-h_n}\\ &= \lim\frac{1-E(h_n)}{-h_n}\frac{1}{Eh_n}\\ &= \lim\frac{E(h_n)-1}{h_n}\frac{1}{E(h_n)}\end{aligned}$$

But, since E is continuous (by definition) and $E(0)=1$ (Exercise 15.2) the limit theorems imply

$$\lim\frac{1}{E(h_n)} = \frac{1}{\lim E(h_n)} = \frac{1}{E(\lim h_n)} = \frac{1}{E(0)} = 1$$

Thus,

$$\begin{aligned}&\lim\left(\frac{E(h_n)-1}{h_n}\frac{1}{E(h_n)}\right)\\ &= \left(\lim\frac{E(h_n)-1}{h_n}\right)\left(\lim\frac{1}{E(h_n)}\right)\\ &= \lim\frac{E(h_n)-1}{h_n}\end{aligned}$$

But this last limit evaluates exactly to the *limit from above* since $h_n > 0$ and $h_n \to 0$. Stringing all of this together, we finally see

$$\lim_{h\to 0^-}\frac{E(h)-1}{h} = \lim_{h\to 0^+}\frac{E(h)-1}{h}$$

Thus, by Theorem 17.2 we see that since both one sided limits exist and are equal the entire limit exists: E is differentiable at 0. □

This theorem tells us that the exponential functions have a remarkable property: they are their own derivatives, up to a constant multiple! While the functional equation alone did not provide us any means of distinguishing between different exponential functions, differentiation selects a single *best*, or *simplest* exponential out of the lot: the one where that constant multiple is just 1!

Definition 23.1. We write $\exp(x)$ for the exponential function which has $\exp'(0) = 1$.

Note that by the chain rule we know such a thing exists so long as *any* exponential exists. If $E(x)$ is any exponential then $E(x/E'(0))$ has derivative 1 at $x = 0$!

Recalling our work with irrational exponents, we saw that if E is an exponential with $E(1) = a$, then we may write $E(x) = a^x$ for any $x \in \mathbb{R}$ (defined as a limit of rational exponents). So, our special exponential exp comes with a special number as its base.

Definition 23.2. We denote by the letter e the base of the exponential $\exp(x)$: that is, $e = \exp(1)$, and

$$\exp(x) = e^x$$

Note that *by definition* we have

$$(e^x)' = e^x$$

This is the origin of the number e from the perspective of analysis! At this point in the story we do not know it's value, but we now have a hint on how to get it: we just need to construct a means of *computing* the exponential function, and then plug in 1.

23.3.1. Logarithms

For every exponential, the inverse function is a logarithm (Theorem 15.2). So, E be any exponential, and L a logarithm. Then $L(E(x)) = x$, and differentiating with the chain rule yields

$$[L(E(x))]' = L'(E(x))E'(x) = L'(E(x))E(x)E'(0)$$

The other side of the equality was x, whose derivative is 1: thus

$$1 = L'(E(x))E(x)E'(0)$$
$$\implies L'(E(x)) = \frac{1}{E'(0)E(x)}$$

Thus, $L'(-)$ is a function that takes the positive number $E(x)$ to $E'(0)/E(x)$: it divides $E'(0)$ by its input!

Proposition 23.3. *If $L(x)$ is a logarithm function, then for some positive $k \in \mathbb{R}$*

$$L'(x) = \frac{1}{kx}$$

(Indeed $k = E'(0)$ where E is the inverse of L)

This tells us that like the exponential function, there is a *natural logarithm* - the one where the arbitrary constant appearing during differentiation is equal to 1.

Definition 23.3. The *natural logarithm* $\log(x)$ is the logarithm function for which

$$\log(x)' = \frac{1}{x}$$

Corollary 23.1 (log and exp are Inverses)**.**

Proof. Since exp is an exponential, we know its derivative is some logarithm L. But, differentiating L yields

$$L'(x) = \frac{1}{\exp'(0)}\frac{1}{x}$$

Since $\exp'(0) = 1$ by definition this says that $L'(x) = 1/x$, which is the defining property of the natural logarithm. Thus $L = \log$. □

23.4. The Power Rule

Perhaps the most memorable fact from Calculus I is the power rule, that $(x^n)' = nx^{n-1}$. In this short section, we prove the power level at various levels of generality, starting with natural number exponents and proceeding to arbitrary real exponents.

Proposition 23.4 (The Power Rule: Natural Number Exponents)**.** *If n is a natural number, x^n is differentiable at all real numbers and*

$$(x^n)' = nx^{n-1}$$

Proof. This is directly proved via induction on n, starting from the base case $x' = 1$, which holds as if $f(x) = x$ and $a \in \mathbb{R}$,

$$\lim_{x\to a}\frac{f(x)-a}{x-a} = \frac{x-a}{x-a} = 1$$

Now, assume $(x^n)' = nx^{n-1}$ and consider x^{n+1}. Using the product rule, we compute the derivative of $x^{n+1} = xx^n$

$$\begin{aligned}(xx^n)' &= (x)'x^n + x(x^n)' \\ &= 1x^n + x(nx^{n-1}) \\ &= x^n + nx^n \\ &= (n+1)x^{n+1}\end{aligned}$$

□

Exercise 23.5 (The Power Rule: Integer Exponents)**.** Let $n \in \mathbb{Z}$ and consider the function x^n (which is defined as $1/x^{|n|}$ when $n < 0$). Then x^n is differentiable at all $x \neq 0$ and

$$(x^n)' = nx^{n-1}$$

Using this, we can extend what we know to rational exponents:

Proposition 23.5 (The Power Rule: Rational Exponents)**.** *Let $r = p/q$ be any rational number and $f(x) = x^r$. Then f is differentiable for all $x > 0$ and*

$$f'(x) = rx^{r-1}$$

Proof. Let $r = p/q$ where without loss of generality $p, q \neq 0$ and $q > 1$ (as if $q = 1$ we are in the integer exponent case). Then let $f(x) = x^{p/q}$, and note that $f(x)^q = x^p$. Then we can differentiate both sides of this inequality:

$$[f(x)^q]' = qf(x)^{q-1}f'(x)$$
$$[x^p]' = px^{p-1}$$

Equating these gives $qf(x)^{q-1}f'(x) = px^{p-1}$, and solving for f':

$$f'(x) = \frac{px^{p-1}}{qf(x)^{q-1}}$$

Using that $f(x) = x^{p/q}$ we can simplify the right hand side further:

$$f'(x) = \frac{px^{p-1}}{q(x^{p/q})^{q-1}} = \frac{px^{p-1}}{qx^{p\frac{q-1}{a}}} = \frac{p}{q}x^{(p-1)-p\frac{q-1}{q}}$$

This exponent simplifies as expected, yielding

$$f'(x) = \frac{p}{q}x^{\frac{p}{q}-1}$$

□

Now that we know the power rule for all rational exponents, it is time to consider arbitrary real exponents, recalling that we define x^a as a limit of rational exponents.

Theorem 23.5 (⋆ The General Power Rule)**.** *If $a \in \mathbb{R}$ and $f(x) = x^a$. Then f is differentiable for all $x > 0$, and*

$$(x^a)' = ax^{a-1}$$

Exercise 23.6. Recall the definition of x^a for irrational a is $\lim x^{a_n}$ for a_n a sequence of rational numbers converging to a. Use this definition to attempt a proof of the general power rule, by computing

$$(x^a)' = (\lim x^{a_n})'$$

In your proof, you will end up getting stuck at a point where you need to interchange two limits: point out where this happens, and then show that *if* you are justified in interchanging the limits, that the generalized power rule holds.

Alternatively, we can give a complete proof of the power rule using exponentials and logarithms.

Proof. Let exp be the natural exponential, and log be the natural log. Then $\exp(\log(x)) = x$, and so $\exp(\log(x^a)) = x^a$. Using the property of logarithms and powers (Corollary 15.2) this simplifies

$$x^n = \exp(\log(x^a)) = \exp(a\log(x))$$

By the chain rule,

$$\begin{aligned}[\exp(a\log(x))]' &= \exp(a\log(x))\,[a\log(x)]' \\ &= \exp(a\log(x))a\log'(x) \\ &= \exp(a\log(x))a\frac{1}{x}\end{aligned}$$

But, recalling that $\exp(a\log(x)) = \exp(\log(x^a)) = x^a$ this simplifies to

$$= x^a a\frac{1}{x} = ax^{a-1}$$

□

24. Theory

Highlights of this Chapter: we study the relationship between the behavior of a function and its derivative, proving several foundational results in the theory of differentiable functions:

- Fermat's Theorem: A differentiable function has derivative zero at an extremum.
- Rolle's Theorem: if a differentiable function is equal at two points, it must have zero derivative at some point in-between.
- The Mean Value Theorem: the average slope of a differentiable function on an interval is realized as the instantaneous slope at some point inside that interval.

The Mean Value theorem is really the star of the show, and we go on to study several of its prominent applications:

- The Second Derivative Test for function extrema
- The ambiguity in antidifferentiation is at most a constant
- L'Hospital's rule

24.1. Derivatives and Extrema

The most memorable procedure from calculus I is likely *to find maxes and mins, set the derivative equal to zero and solve.* This is not precisely correct (nor is it exactly what is taught in Calculus I; just what is remembered!) so here we will give the precise story.

Definition 24.1 (Local Extrema)**.** Let f be a real-valued function with domain $D \subset \mathbb{R}$. Then a point $m \in D$ is a local maximum if $f(m) \geq f(x)$ for x near a, and is a local minimum if $f(m) \leq f(x)$ for x near m. A point that is either a local minimum or local maximum is known as a *local extremum.*

By 'x near m' we mean that there is some interval $(a, b) \subset R$ containing m, where the claimed inequality holds for all $x \in (a, b)$.

That the derivative (rate of change) should be able to detect local extrema is an old idea, even predating the calculus of Newton and Leibniz. Though certainly realized earlier in certain cases, it is Fermat who is credited with the first general theorem (so, the result below is often called *Fermat's theorem*)

Theorem 24.1 (Finding Local Extrema (Fermat's Thm)). *Let f be a function with a local extremum at m. Then if f is differentiable at m, we must have $f'(m) = 0$.*

Proof. Without loss of generality we will assume that m is the location of a local minimum (the same argument applies for local maxima, except the inequalities in the numerators reverse). As f is differentiable at m, we know that both the right and left hand limits of the difference quotient exist, and are equal.

First, some preliminaries that apply to both right and left limits. Since we know the limit exists, it's value can by computed via any appropriate sequence $x_n \to m$. Choosing some such sequence we investigate the difference quotient

$$\frac{f(x_n) - f(m)}{x_n - m}$$

Because m is a local minimum, there is some interval (say, of radius ϵ) about m where $f(x) \geq f(m)$. As $x_n \to m$, we know the sequence eventually enters this interval (by the definition of convergence) thus for all sufficiently large n we know

$$f(x_n) - f(m) \geq 0$$

Now, we separate out the limits from above and below, starting with $\lim_{x \to m^-}$. If $x_n \to m$ but $x_n < m$ then we know $x_n - m$ is negative for all n, and so

$$\frac{f(x_n) - f(m)}{x_n - m} = \frac{\text{pos}}{\text{neg}} = \text{neg}$$

Thus, for all n the difference quotient is ≤ 0, and so the limit must be as well! That is,

$$\lim_{x \to m^-} \frac{f(x) - f(m)}{x - m} \leq 0$$

Performing the analogous investigation for the limit from above, we now have a sequence $x_n \to m$ with $x_n \geq m$. This changes the sign of the denominator, so

$$\frac{f(x_n) - f(m)}{x_n - m} = \frac{\text{pos}}{\text{pos}} = \text{pos}$$

Again, if the difference quotient is ≥ 0 for all n, we know the same is true of the limit.

$$\lim_{x \to m^+} \frac{f(x) - f(m)}{x - m} \geq 0$$

But, by our assumption that f is differentiable at m we know both of these must be equal! And if one is ≥ 0 and the other ≤ 0 the only possibility is that $f'(m) = 0$. □

This provides a clear strategy for tracking down local extrema, especially for functions that are only occasionally not differentiable (piecewise functions, for example): we only need to check the points where f' is either zero, or undefined. This motivates the below definition, giving a uniform term to these disparate categories:

Definition 24.2 (Critical Point)**.** A critical point of a function f is a point where either (1) f is not differentiable, or (2) f is differentiable, and the derivative is zero.

Note that not all critical points are necessarily local extrema - Fermat's theorem only claims that extrema are critical points - not the converse! There are many examples showing this is not an if and only if:

Example 24.1. The function $f(x) = x^3$ has a critical point at $x = 0$ (as the derivative is zero), but does not have a local extremum there. The function $g(x) = 2x + |x|$ has a critical point at 0 (because it is not differentiable there) but also does not have a local extremum.

To classify exactly when a critical point is a local max/min (and crucially, *which* it is) will require a bit more theory, to come. But if one is only interested in the *absolute max and min* of the function over its entire domain, this already provides a reasonable strategy, which is one of the early highlights of Calculus I.

Theorem 24.2 (Finding The Global Max and Min)**.** *Let f be a continuous function defined on a closed interval I with finitely many critical points. Then the absolute maximum and minimum value of f are explicitly findable via the following procedure:*

- *Find the value of f at the endpoints of I*
- *Find the value of f at the points of non-differentiability*
- *Find the value of f at the points where $f'(x) = 0$.*

The absolute max of f is the largest of these values, and the the absolute min is the smallest.

Proof. Because I is a closed interval and f is continuous, we are guaranteed by the extreme value theorem that f achieves both a maximum and minimum value. Let these be max, min respectively, realized at points M, m with

$$f(M) = \max \qquad f(m) = \min$$

Without loss of generality, we will consider M (the same argument applies to m).

First, M could be at one of the endpoints of f. If it is not, then M lies in the interior of I, and there is some small interval (a, b) containing M totally contained in the domain I. Since M is the location of the global max, we know for all $x \in I$, $f(x) \leq f(M)$. Thus, for all $x \in (a, b)$, $f(x) \leq f(M)$ so M is the location of a local max.

But if M is the location of a local maximum, if f is differentiable there by Fermat's theorem we know $f'(M) = 0$. Thus, M must be a critical point of f (whether differentiable or not).

Thus, M occurs in the list of critical points and endpoints, which are the points we checked. □

24.2. The Mean Value Theorem

One of the most important theorems relating f and f' is the *mean value theorem*. This is an excellent example of a theorem that is intuitively obvious (from our experience with reasonable functions) but yet requires careful proof (as we know by know many functions have non-intuitive behavior). Indeed, when I teach calculus I, I often paraphrase the mean value theorem as follows:

> If you drove 60 miles in one hour, then at some point you must have been driving 60 miles per hour

How can we write this mathematically? Say you drove D miles in T hours. If $f(t)$ is your *position* as a function of time*, and you were driving between $t = a$ and $t = b$ (where $b - a = T$), your average speed was

$$\frac{D}{T} = \frac{f(b) - f(a)}{b - a}$$

To then say *at some point you were going D miles per hour implies that there exists some $t^\star$ between a and b where the instantaneous rate of change - the derivative - is equal to this value. This is exactly the Mean Value Theorem:

Theorem 24.3 (The Mean Value Theorem)**.** *If f is a function which is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there exists some $x^\star \in (a, b)$ where*

$$f'(x^\star) = \frac{f(b) - f(a)}{b - a}$$

Note: The reason we require differentiability only ont he *interior* of the interval is that the two sided limit defining the derivative may not exist at the endpoints, (if for example, the domain of f is only $[a, b]$).

In this section we will prove the mean value theorem. It's simplest to break the proof into two steps: first the special case were $f(a) = f(b)$ (and so we are seeking $f'(x^\star =$

0)), and then apply this to the general version. This special case is often useful in its own right and so has a name: *Rolle's Theorem.*

Theorem 24.4 (Rolle's Theorem)**.** *Let f be continuous on the closed interval $[a, b]$ and differentiable on (a, b). Then if $f(b) = f(a)$, there exists some $x^\star \in (a, b)$ where $f'(x^\star) = 0$.*

Proof. Without loss of generality we may take $f(b) = f(a) = 0$ (if their common value is k, consider instead the function $f(x) - k$, and use the linearity of differentiation to see this yields the same result).

There are two cases: (1) f is constant, and (2) f is not. In the first case, $f'(x) = 0$ for all $x \in (a, b)$ so we may choose any such point. In the second case, since f is continuous, it achieves both a maximum and minimum value on $[a, b]$ by the extreme value theorem. Because f is nonconstant these values are distinct, and so at least one of them must be nonzero. Let $c \in (a, b)$ denote the location of either a (positive) absolute max or (negative) absolute min.

Then, $c \in (a, b)$ and for all $x \in (a, b)$, $f(x) \leq f(c)$ if c is the absolute min, and $f(x) \geq f(c)$ if its the max. In both cases, c satisfies the definition of a *local extremum.* And, as f is differentiable on (a, b) this implies $f'(c) = 0$, as required. □

Now, we return to the main theorem:

Proof. Let f be a function satisfying the hypotheses of the mean value theorem, and L be the secant line connecting $(a, f(a))$ to $(b, f(b))$. Computing this line,

$$L = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Now define the auxiliary function $g(x) = f(x) - L(x)$. Since $L(a) = f(a)$ and $L(b) = f(b)$, we see that g is zero at both endpoints. Further, since both L and f are continuous on $[a, b]$ and differentiable on (a, b), so is g. Thus, g satisfies the hypotheses of Rolle's theorem, and so there exists some $\star \in (a, b)$ with

$$g(\star) = 0$$

But differentiating g we find

$$\begin{aligned} 0 &= f'(\star) - L'(\star) \\ &= f'(\star) - \frac{f(b) - f(a)}{b - a} \end{aligned}$$

Thus, at $\star$ we have $f'(\star) = \frac{f(b)-f(a)}{b-a}$ as claimed □

Exercise 24.1. Verify the mean value theorem holds for $f(x) = x^2 + x - 1$ on the interval $[4, 7]$.

24.2.1. ⋆ The Simultaneous Mean Value Theorem

One natural extension is to wonder if this can be done for two functions at once: given f and g can we find a single point c where $f'(c)$ *and* $g'(c)$ equal the average slopes of f and g respectively?

Exercise 24.2. Show that this is impossible in general, by considering $f(x) = x^2$ and $g(x) = x^3$ on the interval $[0, 1]$. Show that each has a unique point c_f, c_g satisfying the Mean Value Theorem, and $c_f \neq c_g$: thus there is no point that works for both.

A slight weakening of the question may be as follows: perhaps there is no c giving each of the average slopes *individually*, but could there be a c such that the *ratio* of the instantaneous slopes is equal to the *ratio* of the average slopes? That is, a $c \in (a, b)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{\frac{f(b)-f(a)}{b-a}}{\frac{g(b)-g(a)}{b-a}}$$

Exercise 24.3. Show that this revised notion *does hold* for the functions $f(x) = x^2$ and $g(x) = x^3$ on the interval $[0, 1]$. In fact, show that this holds on *any interval* $[a, b]$: there is some $c \in (a, b)$ where

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}$$

In fact, this holds for all functions f, g, in a result called the *generalized*, or *simultaneous* mean value theorem.

Theorem 24.5 (The Simultaneous Mean Value Theorem)**.** *Prove that if f and g are both continuous on $[a, b]$ and differentiable on (a, b) with g' nonzero, then there exists some $c \in (a, b)$ where*

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}$$

Exercise 24.4. Prove Theorem 24.5.

Hint: Define some function h built from f and g to which you can apply the mean value theorem, and conclude

$$[f(b) - f(a)]\, g'(c) = [g(b) - g(a)]\, f'(c)$$

24.3. MVT Applications

The mean value theorem is a particularly useful piece of technology, as it lets us connect information about the derivative of a function, to the values of the function itself. This sort of relationship is used all the time in calculus I: three prominent examples are below.

24.3.1. Function Behavior

Proposition 24.1. *If f is is continuous and differentiable on $[a, b]$, then $f(x)$ is monotone increasing on $[a, b]$ if and only of $f'(x) \geq 0$ for all $x \in [a, b]$.*

As this is an if and only if statement, we prove the two claims separately. First, we assume that $f' \geq 0$ and show f is increasing:

Proof. Let $x < y$ be any two points in the interval $[a, b]$: we wish to show that $f(x) \leq f(y)$. By the Mean Value Theorem, we know there must be some point $\star \in (x, y)$ such that

$$f'(\star) = \frac{f(y) - f(x)}{y - x}$$

But, we've assumed that $f' \geq 0$ on the entire interval, so $f'(\star) \geq 0$. Thus $\frac{f(y)-f(x)}{y-x} \geq 0$, and since $y - x$ is positive, this implies

$$f(y) - f(x) \geq 0$$

That is, $f(y) \geq f(x)$. Note that we can extract even more information here than claimed: if we know that f' is *strictly greater than 0* then following the argument we learn that $f(y) > f(x)$, so f is *strictly monotone increasing.* □

Next, we assume f is increasing and show $f' \geq 0$:

Proof. Assume f is increasing on $[a, b]$, and let $x \in (a, b)$ be arbitrary. Because we have assumed f is differentiable, we know that the right and left limits both exist and are equal, and that either of them equals the value of the derivative. So, we consider the right limit

$$f'(x) = \lim_{t \to x^+} \frac{f(t) - f(x)}{t - x}$$

For any $t > x$ we know $f(t) \geq f(x)$ by the increasing hypothesis, and we know that $t - x > 0$ by definition. Thus, for all such t this difference quotient is nonnegative, and hence remains so in the limit:

$$f'(x) \geq 0$$

$\square$

Exercise 24.5. Prove the analogous statement for negative derivatives: $f'(x) \leq 0$ on $[a, b]$ if and only if $f(x)$ is monotone decreasing on $[a, b]$.

Corollary 24.1 (Distinguishing Maxes and Mins)**.** *Let f be a continuously differentiable function on $[a, b]$ and $c \in (a, b)$ be a critical point where $f'(x) < 0$ for $x < c$ and $f'(x) > 0$ if $x > 0$, for all x in some small interval about c.*

Then c is a local minimum of f.

Proof. By the above, we know that $f'(x) < 0$ for $x < c$ implies that f is monotone decreasing for $x < c$: that is, $x < c \implies f(x) \geq f(c)$. Similarly, as $f'(x) > 0$ for $x > 0$, we have that f is increasing, and $c < x \implies f(c) \leq f(x)$.

Thus, for x on either side of c we have $f(x) \geq f(c)$, so c is the location of a local minimum. $\square$

This is even more simply phrased in terms of the *second derivative*, as is common in Calculus I.

Corollary 24.2 (The Second Derivative Test)**.** *Let f be a twice continuously differentiable function on $[a, b]$, and c a critical point. Then if $f''(c) > 0$, the point c is the location of a local minimum, and if $f''(x) > 0$ then c is the location of a local maximum.*

Proof. We consider the case that $f''(c) > 0$, the other is analogous. Since f'' is continuous and positive at c, we know that there exists a small interval $(c - \delta, c + \delta)$ about c where f'' is positive (by Proposition 14.1).

Thus, by Proposition 24.1, we know on this interval that f' is an increasing function. Since $f'(c) = 0$, this means that if $x < c$ we have $f'(x) < 0$ and if $x > c$ we have $f'(x) > 0$. That is, f' changes from negative to positive at c, so c is the location of a local minimum by Corollary 24.1. $\square$

24.3.2. Convexity

Recall back from the very introduction to functions we defined the property of *convexity*, saying that a function was convex if the secant line L connecting any two points lies strictly above the graph of f, or $L(x) - f(x) \geq 0$.

It's good to have a quick review: if a, b are two points in the domain, the secant line connecting $(a, f(a))$ to $(b, f(b))$ is familiar from our proof of the Mean Value Theorem:

$$L_{a,b}(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Exercise 24.6. Show that you can equivalently express this secant line as below, via algebraic manipulation:

$$L_{a,b}(x) = f(a)\left(\frac{b - x}{b - a}\right) + f(b)\left(\frac{x - a}{b - a}\right)$$

Working even harder, we can come up with a rather simple looking condition that is equivalent to f lying below its secant line $L_{a,b}$ for all $x \in (a, b)$. This is all still strictly algebraic manipulations, encapsulated into a lemma below.

Lemma 24.1. *If f is a function defined on $[a, b]$ the, f lies below its secant line $L_{a,b}(x)$ everywhere on the interval if and only if*

$$\frac{f(b) - f(x)}{b - x} - \frac{f(a) - f(x)}{x - a} > 0$$

for all $x \in (a, b)$.

Proof. Because $1 = \frac{b-x}{b-a} + \frac{x-a}{b-a}$, multiplying through by $f(x)$ yields the identity

$$f(x) = f(x)\frac{b - x}{b - a} + f(x)\frac{x - a}{b - a}$$

Substituting this into the simplified form of Exercise 24.6, we can collect like terms and see

$$\begin{aligned} L_{a,b}(x) - f(x) &= [f(b) - f(x)]\,\frac{b - x}{b - a} + [f(a) - f(x)]\,\frac{x - a}{b - a} \\ &= \frac{x - a}{b - a}\,[f(b) - f(x)] - \frac{b - x}{b - a}\,[f(x) - f(a)] \end{aligned}$$

We are trying to set ourselves up to use the Mean Value Theorem, so there's one more algebraic trick we can employ: we can multiply and divide the first term by $b-x$, and multiply and divide the second term by $x-a$: This gives

$$\begin{aligned} L_{a,b}(x) - f(x) &= \frac{b-x}{b-x}\frac{x-a}{b-a}[f(b)-f(x)] - \frac{x-a}{x-a}\frac{b-x}{b-a}[f(x)-f(a)] \\ &= \frac{(b-x)(x-a)}{b-a}\frac{f(b)-f(x)}{b-x} - \frac{(b-x)(x-a)}{b-a}\frac{f(x)-f(a)}{x-a} \end{aligned}$$

Note that each of these terms has the factor $\frac{(b-x)(x-a)}{b-a}$ in common, and that this factor is positive (as $x \in (a,b)$ implies $b-x>0$ and $x-a>0$). Thus, we can factor it out and see that $L_{a,b}(x) - f(x)$ is positive if and only if the remaining term is positive: that is, if and only if

$$\frac{f(b)-f(x)}{b-x} - \frac{f(a)-f(x)}{x-a} > 0$$

as claimed □

Now, our goal is to use the Mean Value Theorem to relate this expression (which is a property of f) to a property of one of its derivatives (here f'').

Exercise 24.7. If $f''>0$ on the interval $[a,b]$ prove that f lies below its secant line $L_{a,b}$.

Hint: Here's a sketch of how to proceed

- For $x \in (a,b)$, start with the expression $\frac{f(b)-f(x)}{b-x} - \frac{f(a)-f(x)}{x-a}$, which you eventually want to show is positive.
- Apply the MVT for f to find points $c_1 \in (a,x)$ and $c_2 \in (x,b)$ where $f'(c_i)$ equals the respective average slopes.
- Using this, show that your original expression is equivalent to $(c_2 - c_1)\frac{f'(c_2)-f'(c_1)}{c_2-c_1}$, and argue that it is sufficient to show that $\frac{f'(c_2)-f'(c_1)}{c_2-c_1}$ is positive.
- Can you apply the MVT again (this time to f') and use our assumption on the second derivative to finish the argument?

Using this, we can quickly prove the main claimed result:

Theorem 24.6. *If f is twice differentiable on an interval and $f''>0$ on that interval, then f is convex on the interval.*

Proof. Let I be the interval in question, and let $a < b$ be any two points in I. Restricting our function to the interval $[a, b]$ we have $f''(x) > 0$ for all $x \in [a, b]$ by hypothesis; so Exercise 24.7 implies that the secant line lies strictly above the graph. Since the interval $[a, b]$ was arbitrary, this holds for any two such points, which is the definition of convexity. □

In fact (though we will not need it) the converse of this is true as well. I've stated it below for reference

Theorem 24.7. *If f is convex on an interval, then f'' is positive on that interval.*

24.3.3. ⋆ Antidifferentiation

Proposition 24.2. *If f is a differentiable function where $f'(x) = 0$ on an interval I, then f is constant on that interval.*

Proof. Let a, b be any two points in the interval: we will show that $f(a) = f(b)$, so f takes the same value at all points. If $a < b$ we can apply the mean value theorem to this pair, which furnishes a point $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

But, $f'(c) = 0$ by assumption! Thus $f(b) - f(a) = 0$, so $f(b) = f(a)$. □

Corollary 24.3. *If f, g are two functions which are differentiable on an interval I and $f' = g'$ on I, then there exists a $C \in \mathbb{R}$ with*

$$f(x) = g(x) + C$$

Proof. Consider the function $h(x) = f(x) - g(x)$. Then by the differentiation laws,

$$h'(x) = f'(x) - g'(x) = 0$$

as we have assumed $f' = g'$. But now Proposition 24.2 implies that h is constant, so $h(x) = C$ for some C. Substituting this in yields

$$f(x) = g(x) + C$$

□

Definition 24.3. Let f be a function. If F is a differentiable function with the same domain such that $F' = f$, we say F is an *antiderivative* of f.

Thus, another way of saying Corollary 24.3 is that any two antiderivatives of a function can only differ by a constant. This is the origin of the $+C$ from calculus, that we will see in the Fundamental Theorem.

24.3.4. ⋆ L'Hospital's Rule

L'Hospital's rule is a very convenient trick for computing tricky limits in calculus: it tells us that when we are trying to evaluate the limit of a quotient of continuous functions and 'plugging in' yields the undefined expression $0/0$ we can attempt to find the limit's value by differentiating the numerator and denominator, and trying again. Precisely:

Theorem 24.8. *Let f and g be continuous functions on an interval containing a, and assume that both f and g are differentiable on this interval, with the possible exception of the point a.*

Then if $f(a) = g(a) = 0$ and $g'(x) \neq 0$ for all $x \neq a$,

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = L \qquad \textit{implies} \qquad \lim_{x\to a} \frac{f(x)}{g(x)} = L$$

Proof. Assume that $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$. Now, we wish to compute the limit of $f(x)/g(x)$. Recalling that $f(a) = g(a) = 0$ we have that for any $x \neq a$

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)}$$

Applying the simultaneous mean value theorem gives a $c \in (a, b)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(x) - f(a)}{g(x) - g(a)}$$

Now, our goal is to calculate $\lim_{x\to a} \frac{f(x)}{g(x)}$, so we begin by choosing an arbitrary sequence $x_n \to a$ with $x_n \neq a$. Applying the above result gives us a sequences c_n trapped between x_n and a: so, by the squeeze theorem we know $c_n \to a$. But, our assumption on f'/g' tells us that since $c_n \neq a$

$$\lim \frac{f'(c_n)}{g'(c_n)} = \lim_{x\to a} \frac{f'}{g'} = L$$

But, for each n we know that

$$\frac{f'(c_n)}{g'(c_n)} = \frac{f(x_n)}{g(x_n)}$$

So, in fact we know $\lim \frac{f(x_n)}{g(x_n)} = L$. Since x_n was an arbitrary sequence $x_n \to a$ with $x_n \neq a$, this holds for all such sequences, and so as claimed,

$$\lim_{x\to a} \frac{f(x)}{g(x)} = L$$

□

Exercise 24.8. Give an alternate proof of L'Hospitals rule using just the ordinary mean value theorem (not the generalized version) following the steps below:

- Show that for any x, we have

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}}$$

- For any x, use the MVT to get points c, k such that $f'(c) = \frac{f(x)-f(a)}{x-a}$ and $g'(k) = \frac{g(x)-g(a)}{x-a}$.
- Choose a sequence $x_n \to a$: for each x_n, the above furnishes points c_n, k_n: show these sequences converge to a by squeezing.
- Use this to show that the sequence $s_n = \frac{f'(c_n)}{g'(k_n)}$ converges to L, using our assumption $\lim_{x\to a} \frac{f'}{g'} = L$.
- Conclude that the sequence $\frac{f(x_n)}{g(x_n)} \to L$, and that $\lim_{x\to a} \frac{f(x)}{g(x)} = L$ as claimed.

Hint: Use the $\epsilon - \delta$ definition of a functional limit our assumption $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$ to help: for any ϵ, theres a δ where $|x - a| < \delta$ implies this quotient is within ϵ of L. Since $c_n, k_n \to a$ can you find an N beyond which $f'(c_n)/g'(k_n)$ is always within ϵ of L?

25. Power Series

Highlights of this Chapter: we prove to marvelous results about power series: we show that they are differentiable (and get a formula for their derivative), and we *also* prove a formula about how to approximate functions well with a power series, and in the limit get a *power series representation* of a known function, in terms of its derivatives at a single point.

25.1. Differentiating Power Series

The goal of this section is to prove that power series are differentiable, and that we can differentiate them term by term. That is, we seek to prove

$$\left(\sum_{k\geq 0} a_k x^k\right)' = \sum_{k\geq 0}(a_k x^k)' = \sum_{k\geq 1} k a_k x^{k-1}$$

Because a derivative is defined as a limit, this process of bringing the derivative inside the sum is really an exchange of limits: and we know the tool for that! Dominated Convergence

25.1.1. ⋆ Dominated Convergence

The crux of differentiating a power series is to be able to bring the derivative *inside* the sum. Because derivatives are limits, we can use dominated convergence to understand when we can switch sums and limits. One crucial step here is the Mean Value Theorem.

Theorem 25.1. *Let $f_k(x)$ be a series of functions on a domain D.*

- *For each k, $f_k(x)$ is differentiable at all $x \in D$.*
- *For each $x \in D$, $\sum_k f_k(x)$ is convergent.*
- *There is an M_k with $|f_k'(x)| < M_K$, for all $x \in D$.*
- *The sum $\sum M_k$ is convergent.*

Then, the sum $\sum_k f_k'(x)$ *is convergent, and*

$$\left(\sum_k f_k(x)\right)' = \sum_k f_k'(x)$$

Proof. Recall the limit definition of the derivative (Definition 22.1):

$$\left(\sum_k f_k(x)\right)' = \lim_{y\to x} \frac{\sum_k f_k(y) - \sum_k f_k(x)}{y-x}$$

Writing each sum as the limit of finite sums, we may use the limit theorems (Theorem 9.3,Theorem 9.2) to combine this into a single sum

$$\lim_{y\to x} \frac{\lim_N \sum_{k=0}^N f_k(y) - \lim_N \sum_{k=0}^N f_k(x)}{y-x} = \lim_{y\to x} \lim_N \sum_{k=0}^N \frac{f_k(y)-f_k(x)}{y-x}$$

And now, rewriting the limit of partial sums as an infinite sum, we see

$$\left(\sum_k f_k(x)\right)' = \lim_{y\to x} \sum_k \frac{f_k(y)-f_k(x)}{y-x}$$

If we are justified in switching the limit and the sum via Theorem 21.2, this becomes

$$\sum_k \lim_{y\to x} \frac{f_k(y)-f_k(x)}{y-x} = \sum_k f_k'(x)$$

which is exactly what we want. Thus, all we need to do is justify that the conditions of Theorem 21.2 are satisfied, for the terms

$$g_k(y) = \frac{f_k(y)-f_k(x)}{y-x}$$

with x a fixed constant and y the variable, as we take the limit $y \to x$.

Step 1: Show $\lim_{y\to x} g_k(y)$ **exists** We have assumed that f_k is differentiable at each point of D, which is exactly the assumption that $\lim_{y\to x} g_k(y)$ exists.

Step 2: Show $\sum_k g_k(y)$ **is convergent** We have assumed that $\sum_k f_k(t)$ exists for all $t \in D$. Let $x \neq y$ be two points in D. Then both $\sum_k f_k(x)$ and $\sum_k f_k(y)$ exist, and by the limit theorems, the following limit also exists:

$$\frac{1}{y-x}\left(\sum_k f_k(y) - \sum_k f_k(x)\right) = \sum_k \frac{f_k(y)-f_k(x)}{y-x} = \sum_k g_k(y)$$

Step 3: Find an M_k with $|g_k(y)| < M_k$ for all $y \neq x$. We are given by assumption that there is such an M_k bounding the derivative f_k on D: we need only show this suffices. If $x \neq y$ then $g_k(y)$ measures the slope of the secant line of f_k between x and y, so by the Mean Value Theroem (Theorem 24.3) there is some c between x and y with

$$|g_k(y)| = \left|\frac{f_k(y) - f_k(x)}{y - x}\right| = \left|f_k'(c)\right|$$

Since $|f_k'(c)| \leq M_k$ by assumption (as $c \in D$), M_k is a bound for g_k as required.

Step 4: Show $\sum M_k$ is convergent This is an assumption, as the M_k's are the same as originally given. Thus there's nothing left to show, and dominated convergence applies! □

25.1.2. Term - By - Term Differentiation

Now, we will attempt to apply dominated convergence for derivatives to a power series. Should this work, we will find the derivative can be calculated via term-by-term differentiation:

$$\sum_{k\geq 0} a_k x^k \mapsto \sum_{k\geq 1} k a_k x^{k-1}$$

So, let's begin by investigating this series: can we figure out when it converges?

Proposition 25.1. *Let $f(x) = \sum_{k\geq 0} a_k x^k$ be a power series with radius of convergence R. Then the series of term-wise derivatives also has radius of convergence R:*

$$g(x) = \sum_{k\geq 1} k a_k x^{k-1}$$

Proof. Say we computed the radius of convergence of f using the ratio test, which implies (Theorem 20.1) $\lim |a_{n+1}/a_n| = 1/R$. Now, we wish to apply the ratio test to our new series $g(x) = \sum_{k\geq 1} k a_k x^{k-1}$. That is, we must compute the limit

$$\lim \left|\frac{(n+1)a_{n+1}x^n}{n a_n x^{n-1}}\right|$$

Simplifying this fraction and breaking into components gives

$$\lim \left(\frac{n+1}{n}\right)\left|\frac{a_{n+1}}{a_n}\right||x|$$

We can compute the limit of the first term here directly, as $\lim(n+1)/n = \lim(1 + 1/n) = 1$ and we know the limit of the second term is $1/R$ by our initial assumption. As x is constant, this is all we need to apply the limit theorems and conclude

$$\lim = \frac{1}{R}|x|$$

and this is < 1 so long as $|x| < R$: that is, our new series converges with radius of convergence R. □

(A small note: while for all series we will see we can easily compute the radius of convergence via the ratio test; if we were not able to we would need a more involved argument above to help us fill in that first line).

Now that we know our proposed derivative actually makes sense (converges), its time to show we are actually justified in exchanging the sum limit and the derivative limit, using Dominated Convergence.

Theorem 25.2 (Differentiation of Power Series). *Let $f = \sum_{k\geq 0} a_k x^k$ be a power series with radius of convergence R. Then for $x \in (-R, R)$:*

$$f'(x) = \sum_{k\geq 1} k a_k x^{k-1}$$

Proof. The terms on the right are the term-by-term derivatives of f. That is, we are trying to show

$$\left(\sum_k a_k x^k\right)' = \sum_k \left(a_k x^k\right)'$$

which is precisely the situation to which dominated convergence for derivatives (Theorem 25.1) is suited. This theorem has several hypotheses we have to verify on the functions $f_k(x) = a_k x^k$.

To start, let $x \in (-R, R)$ be arbitrary. Since x lies strictly within the interval of convergence, we may choose some closed interval $I \subset (-R, R)$ containing x. Without loss of generality we may take $I = [-y, y]$ for some $y < R$, and we do so for concreteness, and use this for the domain of the power series when applying Theorem 25.1.

Requirement 1: f_k is differentiable on I This is immediate, as $f_k(x) = a_k x^k$ is a polynomial and polynomials are differentiable on the entire real line.

Requirement 2: $\sum_k a_k x^k$ converges for each $x \in I$ This is also immediate by definition, since I is a proper subset of the interval of convergence for f.

Requirement 3: There is an M_k bounding $|(a_k x^k)'|$ on I This derivative is $k|a_k||x|^{k-1}$, which is a monotone increasing function of $|x|$. Thus, if $I = [-y, y]$ we may set $M_k = k|a_k|y^{k-1}$ and note

$$\forall x \in I,\ \ k|a_k||x|^{k-1} \leq M_k$$

Requirement 4: $\sum_k M_k$ is convergent. Consider the sum:

$$\sum_k M_k = \sum_k k|a_k|y^{k-1}$$

Because y is within the radius of convergence of the original function f, we know that $\sum_k a_k y^k$ is absolutely convergent, and thus that the power series $\sum_k |a_k|x^k$ converges at y. But now applying Proposition 25.1, we see $\sum_k k|a_k|x^{k-1}$ is also convergent at y. But evaluating at y gives exactly $\sum_k M_k$!

Thus, all the requirements are satisfied, and dominated convergence allows us to switch the order of the sum with differentiation. □

Example 25.1. We know the geometric series converges to $1/(1-x)$ on $(-1, 1)$:

$$\sum_{k\geq 0} x^k = \frac{1}{1-x}$$

Differentiating term by term yields a power series for $1/(1-x)^2$:

$$\begin{aligned}
\frac{1}{(1-x)^2} &= \left(\frac{1}{1-x}\right)' \\
&= \left(\sum_{k\geq 0} x^k\right)' \\
&= \sum_{k\geq 0} x^k \\
&= \sum_{k\geq 1} kx^{k-1} \\
&= 1 + 2x + 3x^2 + 4x^3 + \cdots
\end{aligned}$$

The fact that power series are differentiable on their entire radius of convergence puts a strong constraint on which sort of functions can ever be written as the limit of such a series.

Example 25.2. The absolute value $|x|$ is not expressible as a power series.

25.2. Power Series Representations

Definition 25.1. A *power series representation* of a function f at a point a is a power series p where $p(x) = f(x)$ on some neighborhood of a.

How could one try to track down a power series representation of a given function? Power series - being limits of polynomials - are actually pretty constrained objects: it turns out with a little thought that for a given f there is only *one possible formula for a power series representation*

Theorem 25.3 (Candidate Series Representation). *Let f be a smooth real valued function whose domain contains a neighborhood of* 0*, and let $p(x) = \sum_{k\geq 0} a_k x^k$ be a power series which equals f on some neighborhood of zero. Then, the power series p is uniquely determined:*

$$p(x) = \sum_{k\geq 0} \frac{f^{(k)}(0)}{k!} x^k$$

Proof. Let $f(x)$ be a smooth function and $p(x) = \sum_{k\geq 0} a_k x^k$ be a power series which equals f on some neighborhood of zero. Then in particular, $p(0) = f(0)$, so

$$\begin{aligned} f(0) &= \lim_N (a_0 + a_1 x + a_2 x^2 + \cdots + a_N x^N) \\ &= \lim_N (a_0 + 0 + 0 + \cdots + 0) \\ &= a_0 \end{aligned}$$

Now, we know the first coefficient of p. How can we get the next? Differentiate!

$$p'(x) = \left(\sum_{k\geq 0} a_k x^k\right)' = \sum_{k\geq 0} (a_k x^k)' = \sum_{k\geq 1} k a_k x^{k-1}$$

Since $f(x) = p(x)$ on some small neighborhood of zero and the derivative is a limit, $f'(0) = p'(0)$. Evaluating this at 0 will give the constant term of the power series p'

$$\begin{aligned} f'(0) &= \lim_N (a_1 + 2a_2 x + 3a_3 x^2 \cdots + N a_N x^{N-1}) \\ &= \lim_N (a_1 + 0 + 0 + \cdots + 0) \\ &= a_1 \end{aligned}$$

Continuing in this way, the second derivative will have a multiple of a_2 as its constant term:

$$p''(x) = 2a_2 + 3 \cdot 2 \cdot a_3 x + 4 \cdot 3 \cdot a_4 x^2 + \cdots$$

And evaluating the equality $f''(x) = p''(x)$ at zero yields

$$f''(0) = 2a_2, \qquad \text{so} \qquad a_2 = \frac{f''(0)}{2}$$

This pattern continues indefinitely, as f is infinitely differentiable. The term a_n arrives in the constant term after n differentiations (as it was originally the coefficient of x^n), at which point it becomes

$$a_n x^n \mapsto n a_n x^{n-1} \mapsto n(n-1)a_n x^{n-2} \mapsto \cdots \mapsto n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 a_n$$

As the constant term of $p^{(n)}$ this means $p^{(n)}(0) = n!a_n$, and so using $f^{(n)}(0) = p^{(n)}(0)$,

$$a_n = \frac{f^{(n)}(0)}{n!}$$

In each case there was no choice to be made, so long as $f = p$ in any small neighborhood of zero, the unique formula for p is

$$p(x) = \sum_{k\geq 0} \frac{f^{(k)}(0)}{k!} x^k$$

□

Definition 25.2 (Taylor Series)**.** For any smooth function $f(x)$ we define the *Taylor Polynomial* (centered at 0) of degree N to be

$$p_N(x) = \sum_{0\leq k\leq N} \frac{f^{(k)}(0)}{k!} x^k$$

In the limit as $N \to \infty$, this defines the *Taylor Series* $p(x)$ for f.

We've seen for example, that the geometric series $\sum_{k\geq 0} x^k$ is a power series representation of the function $1/(1-x)$ at zero: it actually converges on the entire interval $(-1, 1)$. There are many reasons one may be interested in finding a power series representation of a function - and the above theorem tells us that if we were to search for one, there is a single natural candidate. If there is *any power series representation*, its this one!

So the next natural step is to study this representation: does it actually converge to $f(x)$?

25.2.1. Taylor's Error Formula

Our next goal is to understand how to create power series that converge to specific functions, and more importantly *prove* that our series actually do what we want! To do so, we are going to need some tools relating a functions derivatives to its values. Rolle's Theorem / the Mean Value Theorem does this for the first derivative, and so we present a generalization here the *polynomial mean value theorem*, which does so for n^{th} derivatives.

Theorem 25.4 (Generalized Rolle's Theorem)**.** *Let f be a function which is $n+1$ times differentiable on the interior of an interval $[a,b]$. Assume that $f(a) = f(b) = 0$, and further that the first n derivatives at a are zero:*

$$f(a) = f'(a) = f''(a) = \cdots = f^{(n)}(a) = 0$$

Then, there exists some $c \in (a,b)$ where $f^{(n+1)}(c) = 0$.

Proof. Because f is continuous and differentiable, and $f(a) = f(b)$, the original Rolle's Theorem implies that there exists some $c \in (a,b)$ where $f'(c_1) = 0$. But now, we know that $f'(a) = f'(c_1) = 0$, so we can apply Rolle's theorem to f' on $[a,c_1] to get a point $c_2 \in (a,c_1)$ with $f''(c_2) = 0$.

Continuing in this way, we get a $c_3 \in (a,c_2)$ with $f^{(3)}(c) = 0$, all the way up to to a $c_n \in (a,c_{n-1})$ where $f^n(c_n) = 0$. This leaves one more application of Rolle's theorem possible, as we assumed $f^{(n)}(a) = 0$, so we get a $c \in (a,c_n)$ with $f^{(n+1)}(c) = 0$ as claimed. ☐

Corollary 25.1 (A polynomial Mean Value Theorem)**.** *Let $f(x)$ be an $n+1$-times differentiable function on $[a,b]$ and $h(x)$ a polynomial which shares the first n derivatives with f at zero:*

$$f(a) = h(a), \quad f'(a) = h'(a), \ldots, \quad f^{(n)}(a) = p^{(n)}(a)$$

Then, if additionally $f(b) = h(b)$, there must exist some point $c \in (a,b)$ where

$$f^{(n+1)}(c) = h^{(n+1)}(c)$$

Proof. Define the function $g(x) = f(x) - h(x)$. Then all the first n derivatives of g at $x = a$ are zero (as f and h had the same derivatives), and furthermore $g(b) = 0$ as well, since $f(b) = h(b)$. This means we can apply the generalized Rolle's theorem and find a $c \in (a,b)$ with

$$g^{(n+1)}(c) = 0$$

That is, $f^{(n+1)}(c) = h^{(n+1)}(c)$. ☐

Theorem 25.5 (Taylor's Error Formula)**.** *Let $f(x)$ be an $n+1$-times differentiable function, and $p_n(x)$ the degree n Taylor polynomial $p(x) = \sum_{0\le k\le n} \frac{f^{(k)}(0)}{k!}x^k$.*

Then for any fixed $b \in \mathbb{R}$, we have

$$f(b) = p_n(b) + \frac{f^{(n+1)}(c)}{(n+1)!}b^{n+1}$$

For some $c \in [0,b]$.

Proof. Fix a point b, and consider the functions $f(x)$ and $p_n(x)$ on the interval $[0,b]$. These share their first n derivatives at a, but $f(b) \neq p_n(b)$: in fact, it is precisely this error we are trying to quantify.

We need to modify p_n in some way without affecting its first n derivatives at zero. One natural way is to add a multiple of x^{n+1}, so define

$$q(x) = p_n(x) + \lambda x^{n+1}$$

for some $\lambda \in \mathbb{R}$, where we choose λ so that $f(b) = q(b)$. Because we ensured $q^{(k)}(0) = f^{(k)}(0)$ for $k \le n$, we can now apply the polynomial mean value theorem to these two functions, and get some $c \in (0,b)$ where

$$f^{(n+1)}(c) = q^{(n+1)}(c)$$

Since p_n is degree n its $n+1^{st}$ derivative is zero, and

$$q^{(n+1)}(x) = 0 + \left(\lambda x^{n+1}\right)^{(n+1)} = (n+1)!\lambda$$

Putting these last two observations together yields

$$f^{(n+1)}(c) = (n+1)!\lambda \implies \lambda = \frac{f^{(n+1)}(c)}{(n+1)!}$$

As $q(b) = f(b)$ by construction, this in turn gives what we were after:

$$f(b) = p_n(b) + \frac{f^{(n+1)}(c)}{(n+1)!}b^{n+1}$$

□

25.2.2. Series Based at $a \neq 0$

All of our discussion (and indeed, everything we will need about power series for our course) dealt with defining a power series based on derivative information at zero. But of course, this was an arbitrary choice: one could do exactly the same thing based at any point $a \in \mathbb{R}$.

Theorem 25.6. *Let f be a smooth function, defined in a neighborhood of $a \in \mathbb{R}$. Then there is a unique power series which has all the same derivatives as f at a:*

$$p(x) = \sum_{k \geq 0} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

And, for any N the error between f and the N^{th} partial sum is quantified as

$$f(x) - p_N(x) = \frac{f^{(N+1)}(\xi)}{(N+1)!}(x-a)^{N+1}$$

For some $c \in [a, x]$.

Exercise 25.1. Prove this.

25.3. ⋆ Smoothness & Analyticity

The above theorem is extremely useful for calculational purposes: it tells us *how* to find the power series of a derivative. But it also provides a window into the special nature of power series themselves. For, not only did we learn that a power series is differentiable, but we learned that its derivative is *another power series* (Theorem 25.2) with the same radius of convergence (Proposition 25.1). Since its a power series, we can apply Theorem 25.2 again to find its derivative, which is another power series, and so on.

Thus a power series isn't only differentiable, but can be differentiated over and over again! Recall such functions are called *smooth* CITE DEF.

Proposition 25.2 (Power Series are Smooth Functions)**.** *Let f be a power series with radius of convergence R. Then f is infinitely differentiable on the interval $(-R, R)$.*

Proof. Let $f(x) = \sum_{k \geq 0} a_k x^k$, which converges absolutely on $(-R, R)$ by assumption. Then

$$f'(x) = \sum_{k \geq 0} (k+1) a_{k+1} x^k$$

is also a power series, which converges absolutely on $(-R, R)$ (here I have re-indexed the sum by powers of k for clarity, whereas the cited theorem has it indexed by powers of $k-1$). □

This is a much stronger requirement, which allows us to much finer recognize when a function cannot be written as a power series:

Example 25.3. Consider the function

$$f(x) = \begin{cases} x^m & x \leq 0 \\ x^n & x \geq 0 \end{cases}$$

This is a function which is continuous and differentiable $\min\{m, n\}$ many times, but not differentiable infinitely many times! If we assume without loss of generality that $m < n$ then after m differentiations we find the left hand derivative to be $m!$ at $x = 0$ whereas the right hand derivative is 0. Thus, this function cannot be represented by a power series.

However, the ability to be represented by a power series is even stricter than being smooth, motivating the definition of *analytic functions* which pervades much of advanced analysis.

25.3.1. Analytic Functions

Definition 25.3 (Analytic Functions)**.** An analytic function is a function $f(x)$ where in a neighborhood of every a in its domain, f can be written as a power series $\sum_k a_k(x-a)^k$.

Corollary 25.2 (The exponential is analytic)**.** *The function* $\exp(x)$ *has a power series which converges for all* $x \in \mathbb{R}$*, and moreover converges to the actual exponential at all points. Thus, its analytic.*

Corollary 25.3 (Sine and Cosine are Analytic)**.** *As you'll prove on the final project, these functions have power series that converge on the entire real line, and further work (in the project, via complex exponentials; or alternatively with the Taylor Error formula) shows the limits equal the sine and cosine at all points. Thus these functions are analytic.*

In both of these examples, we needed only a single power series to verify analyticity as it converged everywhere! But for functions whose power series have limited radii of convergence, one may need to use *many* power series to cover the entire domain of the function

Exercise 25.2 (The function $1/1 + x^2$ is analytic)**.** Derive a power series for $\frac{1}{1+x^2}$ by substitution from the geometric series. Show that this power series has radius of convergence 1.

But, then show at every $a \in \mathbb{R}$, the power series centered at a for $\frac{1}{1+x^2}$ given by Theorem 25.6 has a nonzero radius of convergence, and converges to $1/(1+x^2)$ within it.

Thus, while we need *infinitely many power series* to fully cover the graph of $\frac{1}{1+x^2}$, its still analytic.

In fact, probably every smooth function you have ever heard of is analytic: its hard to imagine what could go wrong - somehow you can take *infinitely many derivatives*, but in the end, the error term does not go to zero?

It's a surprising fact of real analysis with - with very wide implications - that there exist smooth but non-analytic functions.

Exercise 25.3. Consider the function

$$s(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 x \leq 0 \end{cases}$$

Show that s is *infinitely differentiable at zero*, and for all n

$$s^{(n)}(0) = 0$$

This implies that the unique power series centered at $a = 0$ is the zero function. But, $s(x) \neq 0$ on any neighborhood of 0 (show for any $x > 0$, $s(x) > 0$). Thus s is smooth, but *not analytic.*

Hint: compute the derivative via right and left hand limits. We know the left hand limit is always zero, so you just need to show the right hand limit is zero for each derivative...

26. The Exponential

Highlights of this Chapter: we reach a culmination of several topics, drawing in theory from across series and differentiability to come up with a formula for the natural exponential $\exp(x)$, and an explicit formula for its base e.

26.1. Prior Work

It's useful to start by summarizing what we already know. We defined the exponential function as a nonconstant solution to the law of exponents

$$E(x+y) = E(x)E(y)$$

26.1.1. Properties

Such a definition does not guarantee that any such function exists, but using the functional equation one can readily begin to prove many propositions about exponentials, assuming they exist. For example some of the first we proved were

- **If** $E(x)$ is an exponential then $E(x)$ is never zero.
- **If** $E(x)$ is an exponential, then $E(0) = 1$
- **If** $E(x)$ is an exponential, then so is $E(kx)$

Through the introduction to differentiation, we can prove even more about the exponential, such as

- **If** $E(x)$ is an exponential, then $E(x)$ is differentiable, and $E'(x) = cE(x)$ for some $c \neq 0$, and in fact $c = E'(0)$.

Combining this with previous facts and the chain rule, we can see that *if* $E(x)$ is any exponential, then $E(x/c)$ is an exponential whose derivative at zero is 1. We called such a function the *natural exponential*, and so have proven

- **If any exponential exists at all** then there is a natural exponential $\exp(x)$ satisfying $\exp(x)' = \exp(x)$.

From here, we can actually learn quite a lot about this function exp, if it exists. For instance

Example 26.1. If exp exists, then it is a strictly increasing function on the entire real line.

To start the proof of this, note since $\exp(0) = 1$ and $\exp(x)$ is never zero, in fact exp is always positive: were it not then there'd be some y with $\exp(y) < 0$, and since 0 lies between $1 = \exp(0)$ and $\exp(y)$, by the intermediate value theorem there would have to be a z with $\exp(z) = 0$. But we know no such points exist.

Now, because $\exp(x) > 0$ and $\exp(x)' = \exp(x)$, we see that the derivative is strictly positive. And, by an argument using the mean value theorem, we know that on any interval where the derivative is positive, the function is increasing. So exp is increasing on all of $\mathbb{R}$.

26.1.2. Existence

This simplifies things a bit: proving the existence of any exponential at all is enough to get us to the existence of exp. But no arguments starting from the functional equation alone can prove that there are exponential functions at all! For that we need to do some additional work: and you did this, over the course of Assignment 7, where you showed

- We can define 2^x as $\lim 2^{r_n}$ for r_n an arbitrary sequence of rational numbers converging to x. That is
  - For any $x \in \mathbb{R}$, and any sequence $r_n \to x$, the sequence 2^{r_n} converges
  - The value of $\lim 2^{r_n}$ does not depend on the choice of sequence: so long as that sequence converges to x.
- The function 2^x defined this way is continuous
- The function 2^x defined this way satisfies the law of exponents on the rationals (by definition), and so by continuity, satisfies the law of exponents for all real inputs.

Corollary 26.1. *Exponential functions exist.*

Thus, at this point we are *certain* that there is a mysterious real function out there called the natural exponential. We just don't know anything about how to compute it! We are even ignorant of the most basic question: if we were to write $\exp(x) = a^x$ in the form above for some base a, what number is a?

26.2. Finding a Power Series

To work with the natural exponential efficiently, we need to find a *formula* that lets us compute it. And this is exactly what power series are good at! However, the theory of power series is a little tricky, as we saw in the last chapter. Not every function has a power series representation, but *if* a function does, there's only one possibility:

Proposition 26.1. If *the natural exponential has a power series representation, then it is*

$$p(x) = \sum_{k\geq 0} \frac{x^k}{k!}$$

Proof. We know the only candidate series for a function $f(x)$ is $\sum_{k\geq 0} \frac{f^{(k)}(0)}{k!}x^k$, so for exp this is

$$p(x) = \sum_{k\geq 0} \frac{\exp^{(k)}(0)}{k!}x^k$$

However, we know that $\exp' = \exp$ and so inductively $\exp^{(k)} = \exp$, and so

$$\exp^{(k)}(0) = \exp(0) = 1$$

Thus

$$p(x) = \sum_{k\geq 0} \frac{1}{k!}x^k$$

□

So now, while we know exp exists we are *back to talking about hypotheticals* because we don't know if it is representable by a power series! The first step to fixing this is to show that the proposed series at least converges.

Proposition 26.2. *The series* $p(x) = \sum_{k\geq 0} \frac{x^k}{k!}$ *converges for all* $x \in \mathbb{R}$.

Proof. This series converges for all $x \in \mathbb{R}$ by the Ratio test, as

$$\lim \left|\frac{x^{n+1}/(n+1)!}{x^n/n!}\right| = \lim \frac{|x|}{n+1} = 0 < 1$$

□

Now, all that remains is to show that $p(x) = \exp(x)$. Since p is a power series, this really means that *the limit of its partial sums equals* $\exp(x)$, or

$$\forall x \in \mathbb{R} \quad \exp(x) = \lim_N p_N(x)$$

For any finite partial sum p_N, we know that it is not exactly equal to $\exp(x)$ (as this finite sum is just a polynomial!). Thus there must be some error term $R_N = \exp - p_N$, or

$$\exp(x) = p_N(x) + R_N(x)$$

This is helpful, as we know from the previous chapter how to calculate such an error, using the Taylor Error Formula: for each fixed $x \in \mathbb{R}$ and each fixed $N \in \mathbb{N}$, there is some point $c_N \in [0, x]$ such that

$$R_N(x) = \frac{\exp^{(N+1)}(c_N)}{(N+1)!} x^{N+1}$$

And, to show the power series becomes the natural exponential in the limit, we just need to show this error tends to zero!

Proposition 26.3. *As $N \to \infty$, for any $x \in \mathbb{R}$ the Taylor error term for the exponential goes to zero:*

$$R_N(x) \to 0$$

Proof. Fix some $x \in \mathbb{R}$. Then for an arbitrary N, we know

$$R_N(x) = \frac{\exp^{(N+1)}(c_N)}{(N+1)!} x^{N+1}$$

where $c_N \in [0, x]$ is some number that we don't have much control of (as it came from an existence proof: Rolle's theorem in our derivation of the Taylor error). Because we don't know c_N explicitly, its hard to directly compute the limit and so instead we use the squeeze theorem:

We know that exp is an increasing function: thus, the fact that $0 \leq c_N \leq x$ implies that $1 = \exp(0) \leq \exp(c_N) \leq \exp(x)$, and multiplying this inequality through by $x^{N+1}(N+1)!$ yields the inequality

$$\frac{x^{N+1}}{(N+1)!} \leq R_N(x) = \exp(c_N)\frac{x^{N+1}}{(N+1)!} \leq \exp(x)\frac{x^{N+1}}{(N+1)!}$$

(Here I have assumed that $x \geq 0$: if $x < 0$ then the inequalities reverse for even values of N as x^{N+1} is negative and we are multiplying through by a negative number. But this does not affect the fact that the error term $R_N(x)$ is still sandwiched between the two.)

So now our problem reduces to showing that the upper and lower bounds converge to zero. Since $\exp(x)$ is a constant (remember, N is our variable here as we take the limit), a limit of both the upper and lower bounds comes down to just finding the limit

$$\lim_N \frac{x^{N+1}}{(N+1)}$$

But this is just the $N+1$st term of the power series $p(x) = \sum_{n\geq 0} x^n/n!$ we studied above! And since this power series converges, we know that as $n \to \infty$ its terms must go to zero (the divergence test). Thus

$$\lim_N \frac{x^{N+1}}{(N+1)} = 0 \qquad \lim_N \exp(x)\frac{x^{N+1}}{(N+1)} = 0$$

and so by the squeeze theorem, $R_N(x)$ converges and

$$\lim_N R_N(x) = 0$$

□

Now we have all the components together at last: we know that exp exists, we have a candidate power series representation, that candidate converges, and the error between it and the exponential goes to zero!

Theorem 26.1. *The natural exponential is given by the following power series*

$$\exp(x) = \sum_{k\geq 0} \frac{x^k}{k!}$$

Proof. Fix an arbitrary $x \in \mathbb{R}$. Then for any N we can write

$$\exp(x) = p_N(x) + R_N(x)$$

For p_N the partial sum of $p(x) = \sum_{k\geq 0} x^k/k!$ and $R_N(x)$ the error. Since we have proven both p_N and R_N converge, we can take the limit of both sides using the limit theorems (and, as $\exp(x)$ is constant in N, clearly $\lim_N \exp(x) = \exp(x)$):

$$\begin{aligned}
\exp(x) &= \lim_N (p_N(x) + R_N(x)) \\
&= \lim_N p_N(x) + \lim_N R_N(x) \\
&= p(x) + 0 \\
&= \sum_{k\geq 0} \frac{x^k}{k!}
\end{aligned}$$

□

Its incredible in and of itself to have such a simple, explicit formula for the natural exponential. But this is just the beginning: this series actually gives us a means to express *all* exponentials:

Theorem 26.2. *Let $E(x)$ be an arbitrary exponential function. Then E has a power series representation on all of $\mathbb{R}$ which can be expressed for some real nonzero c as*

$$E(x) = \sum_{n\geq 0} \frac{c^n}{n!} x^n$$

Proof. Because E is an exponential we know E is differentiable, and that $E'(x) = E'(0)E(x)$ for all x. Note that $E'(0)$ is nonzero; else we would have $E'(x) = 0$ constantly, and so $E(x)$ would be constant. Set $c = E'(0)$.

Now, inductively take derivatives at zero:

$$E'(0) = c \qquad E''(0) = c^2 \qquad E^{(n)}(0) = c^n$$

Thus, *if* E has a power series representation it must be

$$\sum_{n\geq 0} \frac{c^n}{n!} x^n = \sum_{n\geq 0} \frac{1}{n!} (cx)^n$$

This is just the series for exp evaluated at cx: since exp exists and is an exponential, so is this function (as its defined just by a substitution). So there is such an exponential. □

From this, we can directly get a formula to calculate the base of this exponential, the natural constant e:

Corollary 26.2 (A series for e:). *The base of the natural exponential is given by*

$$e := \exp(1) = \sum_{k\geq 0} \frac{1}{k!}$$

Since we know for a general exponential $E(x) = E(1)^x$ can be written as powers of its base (where the power is defined as the limit of rational exponents…) this finally gives us our standard looking exponential function

$$\exp(x) = \exp(1)^x = e^x$$

26.2.1. Estimating e

We finally found e! And we have a relatively simple, explicit formula to compute it. As some final practice with our new tools, lets use what we know here to do some estimation

Proposition 26.4. *The base of the natural exponential is between* 2 *and* 3.

Proof. The series defining e is all positive terms, so we see that e is greater than any partial sum. Thus

$$2 = 1 + 1 = \frac{1}{0!} + \frac{1}{1!} < \sum_{k\geq 0} \frac{1}{k!} = e$$

so we have the lower bound. To get the upper bound, we need to come up wtih a computable upper bound for our series. This turns out to be not that difficult: as the factorial grows so quickly, we can produce many upper bounds by just fining something that grows slower than the reciprocal and summing up their reciprocals. For instance, when $k \geq 2$

$$k(k-1) \leq k!$$

and so,

$$e = \sum_{k\geq 0} \frac{1}{k!} = 1 + 1 + \sum_{k\geq 2} \frac{1}{k!} \leq 1 + 1 + \sum_{k\geq 2} \frac{1}{k(k-1)}$$

But this upper bound now is our favorite telescoping series! After a rewrite with partial fractions, we directly see that it sums to 1. Plugging this in,

$$e < 1 + 1 + 1 = 3$$

□

How can we get a better estimate? Since we do have a convergent infinite series just sitting here defining e for us, the answer seems obvious - why don't we just sum up more and more terms of the series? And of course - that is *part* of the correct strategy, but it's missing one key piece. If you add up the first 10 terms of the series and you get some number, how can you know how accurate this is?

Just because the first two digits are 2.7, who is to say that after adding a million more terms (all of which are positive) it won't eventually become 2.8? To give us any confidence in the value of e we need a way of measuring how far off any of our partial sums could be.

Our usual approach is to try and produce sequences of upper and lower estimates: nested intervals of error bars to help us out. But here we have only one sequence (and producing even a single upper bound above was a bit of work!) so we need to look elsewhere. It turns out, the correct tool for the job is the Taylor Error formula once more!

Proposition 26.5. *Adding up the first N terms of the series expansion of e results in a an estimate of the true value accurate to within* $3/(N+1)!$.

Proof. The number e is defined as $\exp(1)$, and so using $x = 1$ we are just looking at the old equation

$$\exp(1) = p_N(1) + R_N(1)$$

Where $R_N(1) = \exp(c_N)\frac{1^{N+1}}{(N+1)!}$ for $c_N \in [0,1]$. Since exp is increasing, we can bound $\exp(c_N)$ below by $\exp(0) = 1$ and above by $\exp(1) = e$, and e above by 3: thus

$$\frac{1}{(N+1)!} \leq R_N(x) \leq \frac{3}{(N+1)!}$$

And so, the difference $|e - p_N(1)| = |R_N(1)|$ is bounded above by the upper bound $3/(N+1)!$ □

This gives us a readily computable, explicit estimate. Precisely adding up to the $N =$ 5th term of the series yields

$$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} \approx 2.71666\ldots$$

with the total error between this and e is less than $\frac{3}{6!} = \frac{1}{240} = 0.0041666\ldots$. Thus we can be confident that the first digit after the decimal is a 7, as $2.7176 - 0.0041 = 2.7135 \leq e \leq 2.7176 + 0.0041 = 2.7217$.

Adding up five more terms, to $N = 10$ gives

$$1 + 1 + \frac{1}{2} + \frac{1}{3!} + \cdots + \frac{1}{10!} = 2.71828180114638\ldots$$

now with a maximal error of $3/11! = 0.000000075156\ldots$. This means we are now absolutely confident in the first six digits:

$$e \approx 2.718281$$

Pretty good, for only having to add eleven fractions together! Thats the sort of calculation one could even manage by hand.

27. ⋆ Exponentials PDEs and ODEs

Highlights of this Chapter: to close out our work on derivatives, we take a look at where power series come up in analysis beyond this first course.

- We study the complex exponential, in preparation to prove $e^{i\pi} = -1$ in the final project.
- We study the matrix exponential, and see its utility in solving systems of linear differential equations.
- We look to extend exponentiation to even more abstract settings, and consider the meaning of $e^{\frac{d}{dx}}$
- We use this idea of *e to the power of a differential operator* as a window into functional analysis.

27.1. Generalizing the Exponential

Power series are wonderful functions for many reasons, but one of the most powerful is that they are so *simple*. Like polynomials, to make sense of a power series you just need is addition/subtraction and multiplication, now with one more ingredient: convergence. This makes power series a very natural jumping off point to *generalize* familiar objects to unfamiliar places. As a first step, we will look at complex numbers:

Definition 27.1. A complex number is a pair (a, b) of real numbers, which we will write as $a + bi$, where i is a number such that $i^2 = -1$. Just as the real numbers form a line, pairs of real numbers form a *plane*.

- Addition of complex numbers is defined component-wise:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

- Multiplication of a complex number by a real number can also be computed component-wise:

$$a(b + ci) = (ab) + (ac)i$$

- Multiplication of two complex numbers is defined by the field axioms together with the definition $i^2 = -1$:

$$\begin{aligned}(a+bi)(c+di) &= ac + adi + bci + bdi^2 \\ &= (ac-bd) + (ad+bc)i\end{aligned}$$

Limits of sequences of complex numbers are defined using the fact that they are built from pairs of real numbers. A sequence $z_n = x_n + iy_n$ of complex numbers is said to *converge* if and only if both of the real sequences x_n and y_n converge, and in this case

$$\lim(x_n + iy_n) := (\lim x_n) + i(\lim y_n)$$

We know have a fully rigorous theory of what it means to exponentiate a real number: but what does it mean to raise something to the i power? Or the $3-7i$ power? Because we know the power series for the real exponential, we can attempt to *define* a complex exponential just using the power series directly

Definition 27.2. The complex exponential is defined for any $z \in \mathbb{C}$ by the series

$$\begin{aligned}\exp(z) &= \sum_{k\geq 0} \frac{z^k}{k!} \\ &= 1 + z + \frac{1}{2}z^2 + \frac{1}{3!}z^3 + \cdots\end{aligned}$$

Of course, after making such a bold definition we should ask ourselves, *does this make sense*? That is, does the series of complex numbers converge? This may sound daunting at first, but in fact the theory of complex power series inherits much from the real theory: complex numbers are built from pairs of real numbers after all!

Theorem 27.1. *Let $\sum_n z_n$ be a series of complex numbers, $z_n = x_n + iy_n$, and let $|z_n| = \sqrt{x_n^2 + y_n^2}$ be their magnitudes. Then $\sum_n z_n$ converges if $\sum_n |z_n|$ does.*

Proof. Let $z = x + iy$ be an arbitrary point in $\mathbb{C}$, and define $|z| = \sqrt{x^2+y^2}$ Note that since $x^2, y^2 \geq 0$ we know

$$|z| = \sqrt{x^2+y^2} \geq |x| \qquad |z| = \sqrt{x^2+y^2} \geq |y|$$

Thus, looking at the first inequality we see that

$$\sum_{n\geq 0} |z_n| \geq \sum_{n\geq 0} |x_n|$$

We've assumed that $\sum |z_n|$ is convergent and so by comparison this implies that $\sum |x_n|$ is convergent. But this means $\sum x_n$ is *absolutely convergent*, and hence convergent. Thus, $\sum_n x_n = \chi$ for some $\chi \in \mathbb{R}$.

A similar argument applies to the sequence of y's: since $\sum |z_n| \geq \sum |y_n|$ comparison shows that $\sum y_n$ is absolutely convergent and thus convergent, so $\sum y_n = \eta$ for some $\eta \in \mathbb{R}$.

Now, using the definition of convergence for complex numbers, since both real sequences converge the overall sequence does as well, and we can write

$$\sum_{n\geq 0} x_n + iy_n = \sum_{n\geq 0} x_n + i\sum_{n\geq 0} y_n = \chi + i\eta$$

Thus, $\sum_n z_n$ converges, as claimed. □

Corollary 27.1. *Let $p(z) = \sum a_n z^n$ be a power series of a complex number $z = x + iy$. The $p(z)$ converges if and only if the corresponding real power series $p(|z|)$ converges, with $|z| = \sqrt{x^2 + y^2}$ the complex magnitude.*

Proof. If $p(|z|)$ converges, then we know $\sum_n a_n |z|^n$ converges. But using properties of the absolute value (complex magnitude) we see

$$p(|z|) = \sum_{n\geq 0} a_n |z|^n = \sum_{n\geq 0} a_n |z^n| = \sum_{n\geq 0} |a_n z^n|$$

Thus, we have assumed that the series of magnitudes $|a_n z^n|$ converges, and so by Theorem 27.1 the sequence itself converges,

$$\sum_{n\geq 0} a_n z^n$$

as claimed. □

The complex magnitude $|z|$ defines a kind of *absolute value* on the complex numbers, and so we can extend our notion of absolute convergence, saying a series $\sum a_n z^n$ converges *absolutely* if the series $\sum a_n |z|^n$ converges. The theorem above can be translated into this new language, to reveal a familiar theme:

Corollary 27.2. *A complex power series is convergent, if it is absolutely convergent.*

Applying this to the complex exponential, we can confirm this series makes sense for all complex number inputs:

Corollary 27.3. *For any $z \in \mathbb{C}$ the power series $\exp(z) = \sum_n \frac{z^n}{n!}$ converges.*

Proof. Let $r = |z|$. Since $\exp(x)$ converges for all real inputs, we know $\exp(r) = \sum \frac{r^n}{n!} = \sum \frac{|z|^n}{n!}$ converges, which means $\sum \frac{z^n}{n!}$ converges absolutely. But absolute convergence implies convergence, so

$$\exp(z) = \sum_{n\geq 0} \frac{z^n}{n!}$$

is convergent. □

But we can go much further than this. There are many objects we know how to add/subtract and multiply in mathematics - structures with these operations are called *rings*. So, in any ring where one can make sense of limits, we can attempt to define an exponential function by this power series! A natural example is the ring of $n \times n$ matrices

Definition 27.3. Denote by $M_n(\mathbb{R})$ the set of all $n \times n$ matrices with real number entries. We write the ij^{th} entry of such a matrix $A \in M_n(\mathbb{R})$ as A_{ij}. This space has the following operations that are important for us

- Addition: defined entry-wise $(A + B)_{ij} = A_{ij} + B_{ij}$
- Scalar Multiplication: defined entry-wise $(cA)_{ij} = cA_{ij}$
- Multiplication: defined by usual matrix multiplication $(AB)_{ij} = A_{i\star} \cdot B_{\star j}$

We define limits in $M_n(\mathbb{R})$ using our definition of limits for the real numbers. If M_k is a sequence of matrices, then looking at its entries we can think of M_k as an array of $n \times n$ different real-number sequences. We say that M_n converges if and only if every sequence of entires converges, and in this case define

$$\lim \begin{pmatrix} (a_{11})_n & (a_{12})_n & \cdots \\ (a_{21})_n & (a_{21})_n & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} := \begin{pmatrix} \lim(a_{11})_n & \lim(a_{12})_n & \cdots \\ \lim(a_{21})_n & \lim(a_{21})_n & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Definition 27.4. The matrix exponential is defined for any $n \times n$ matrix A as

$$e^A = \exp(A) = \sum_{k\geq 0} \frac{1}{k!} A^k$$

$$= I + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \cdots$$

Again, we are faced with the problem of convergence: for which matrices A does this power series make sense? A matrix itself is just an $n \times n$ array of real numbers, so working entry-by-entry, this is just an $n \times n$ array of *sequences*, that we need to assure converges. It turns out that like for the complex numbers, the natural thing to do is consider some sory of *absolute value* on the space of matrices, and try to prove an analog of absolute convergence implies convergence.

Definition 27.5. Let $A = [a_{ij}]$ be an $n \times n$ matrix. Then we define the *matrix norm* of A as

$$|A| = \sqrt{\sum_{1 \leq i,j \leq n} a_{ij}^2}$$

So for example

$$\left|\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\right| = \sqrt{1^2 + 2^2 + 3^2 + 4^2} = \sqrt{30}$$

Exactly analogous to the complex case, we can prove that absolute convergence implies convergence

Theorem 27.2. *If $p(x) = \sum a_n x^n$ is a power series and A is an $n \times n$ matrix, then $p(A)$ converges if and only if $p(|A|)$ converges as a real power series. That is, absolute convergence implies convergence.*

Exercise 27.1. Prove a version of Theorem 27.1 for matrices: if $\sum_n A_n$ is a series of matrices, then it converges if the real series $\sum_n |A_n|$ does. Then use this to show that a power series $p(A)$ converges if the real power series converges at $|A|$.

Exercise 27.2. Find the exponential of the matrix

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

Exercise 27.3. Prove that if A is a diagonalizable matrix, then

$$\det(e^A) = e^{\text{trace}(A)}$$

Exercise 27.4. Compute the function

$$R(t) = \exp\left[\begin{pmatrix} 0 & -t \\ t & 0 \end{pmatrix}\right]$$

what matrices do you get? *Hint: think about the power series for functions you learned back in calculus*

27.1.1. How far can we go?

In linear algebra, one key use of matrices is to represent Linear transformations (another is just as arrays to store information, like systems of equations). But linear operators on a vector space V are also things that can be added, and multiplied together (composition, as maps $L : V \to V$) so one can attempt to make sense directly of the exponential of a linear map!

Definition 27.6. If $L : V \to V$ is a linear transformation, we define $\exp(L)$ to be the linear transformation $V \to V$ given by

$$e^L = I + L + \frac{1}{2}L \circ L + \frac{1}{3!}L \circ L \circ L + \cdots$$

For this to make sense, we need V to be a vector space where it makes sense to take limits (for example V could be a real or complex vector space, among some other examples). When V is finite dimensional this is equivalent to the matrix examples already given (as matrix multiplication is composition of linear maps) but things get really interesting if we let ourselves go beyond this.

The derivative after all, is a linear map on the vector space of smooth real valued functions, and these functions are things we know how to take limits of (as their inputs and outputs are real numbers!). This might make us wonder: what is the exponential of the derivative?

Definition 27.7. Let $\mathcal{S}$ be the space of all infinitely differentiable functions, and let

$$\frac{d}{dx} : \mathcal{S} \to \mathcal{S}$$

be the derivative, $f \mapsto \frac{d}{dx}f = f'$. Then the exponential of the derivative is the operator $\mathcal{S} \to \mathcal{S}$ defined by the series

$$e^{\frac{d}{dx}} = I + \frac{d}{dx} + \frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{3!}\frac{d^3}{dx^3} + \cdots$$

This acts on functions as follows, taking a function $f : \mathbb{R} \to \mathbb{R}$ to the function $e^{\frac{d}{dx}}(f) : \mathbb{R} \to \mathbb{R}$ given by

$$e^{\frac{d}{dx}}(f(x)) = f(x) + f'(x) + \frac{1}{2}f''(x) + \frac{1}{3!}f'''(x) + \cdots$$

You might rightly worry about convergence here: when does this expression even make sense?! The general theory of such things is beyond the scope of this course, but for functions which are themselves power series, we can actually come up with a beautifully simple answer:

Exercise 27.5. Prove that for any power series $p(x)$ that the exponentiated derivative actually performs a remarkably simple operation: it *shifts* the functions input

$$e^{\frac{d}{dx}}[p(x)] = p(x+1)$$

Hint: show this happens for x^n, use linearity of the derivative, and take some limits

Corollary 27.4. *Let $t\frac{d}{dx}$ be the operator on functions which takes a function (x) to the function $tf'(x)$. Then the exponential of this operator equal the* shift by t *operator on analytic functions,*

$$e^{t\frac{d}{dx}}[f(x)] = f(x+t)$$

27.2. Solving Differential Equations

One interesting application of exponentiation is to *solving differential equations.* We will not dive deeply into this topic but only take a quick view of some interesting examples, for those who enjoy differential equations.

27.2.1. $y' = cy$

If c is a constant, a solution to the differential equation $y' = cy$ is a function whose derivative is c times itself. The simplest such equation is when $c = 1$, which asks for a function $y' = y$ that is its own derivative. We know the answer to this! The natural exponential, $y = \exp(x)$ has this property - and so does any constant multiple. The same ideas carry over to more general c:

Exercise 27.6. Let $c \in \mathbb{R}$ be any constant. Then the solutions to the differential equation $y' = cy$ are the functions

$$f(t) = Ae^{ct}$$

for $A \in \mathbb{R}$.

Usually, a differential equation is given with an *initial condition*, specifying the functions behavior at a certain point. This picks out one solution from the many: if y_0 is the value at $y = 0$ then the specific solution satisfying both this and $y' = cy$ is

$$f(t) = y_0 e^{ct}$$

27.2.2. Linear Systems

A beautiful generalization of the relatively simple idea above allows one to solve essentially all linear systems of differential equations (with constant coefficients). One learns in a differential equations course how to turn any such system into a system of *first order equations* so we focus on those here. For specificity, assume we have the following three differential equations, for unknown functions $x(t), y(t), z(t)$:

$$x'(t) = 2x(t) + 3y(t) - 4z(t)$$
$$y'(t) = 3x(t) - y(t) + z(t)$$
$$z'(t) = x(t) - z(t)$$

And suppose further that these are constrained by specific initial conditions: $x(0) = 7, y(0) = 3, z(0) = 2$.

Because the right hand side of each is a linear combination of x, y, z we can rewrite this more compactly using *matrix notation*:

$$\begin{pmatrix} x'(t) \\ y'(t) \\ z'(t) \end{pmatrix} = \begin{pmatrix} 2 & 3 & -4 \\ 3 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

In our continued attempt to simplify notation and make this problem more manageable, we define $s(t) : \mathbb{R} \to \mathbb{R}^3$ to be the vector valued function

$$s(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

And let M be the matrix appearing in the system above. Then as $s' = \left(x', y', z'\right)$ we can rewrite this system much more succinctly as

$$s'(t) = Ms(t)$$

So, we are looking for a vector valued function $s(t)$ for which differentiation equals multiplication by the matrix M. This gives a hint of why exponentials may be involved: if everything were one dimensional, s would just be a function and M a number - we are back in the case considered in the previous section, where solutions are multiples of e^{Mt}!

Taking this as a hint, we might attempt to solve this differential equation using the matrix exponential. First, we consider a *matrix valued function*, and then will come back to think about the initial conditions.

Proposition 27.1. *Let $M \in M_n(\mathbb{R})$ be an $n \times n$ matrix, and define the function $F : \mathbb{R} \to M_n(\mathbb{R})$ as follows:*

$$F(t) = e^{tM}$$

Then F satisfies the differential equation $F'(t) = MF(t)$ in the space $M_n(\mathbb{R})$ of $n \times n$ matrices.

Proof. If $M \in M_n(\mathbb{R})$ is a matrix, its norm $|M| = \sqrt{\sum_{1\le i,j\le n} m_{ij}^2}$ is a real number, and so the power series for $e^{t|M|}$ converges for all t (since it converges on the entire real line).

Thus, by Exercise 27.1 the power series $F(t) = e^{tM}$ converges in the space of matrices

$$F(t) = e^{tM} \ := \lim_k \left[I + (tM) + \frac{1}{2}(tM)^2 + \cdots + \frac{1}{k!}(tM)^k\right]$$

Now we wish to take the derivative. Recalling M is a fixed matrix, this power series defines an $n \times n$ array of power series (one for each entry), as a function of t:

$$F(t) = \sum_{k\ge 0} \frac{1}{k!} M^k t^k \qquad [F(t)]_{ij} = \sum_{k\ge 0} \frac{1}{k!}[M^k]_{ij} t^k$$

Where we know from the above that each of these $n\times n$ many power series converges absolutely for all values of t. Thus, using our result on differentiating power series within their radius of convergence, we know for each of these we have

$$\begin{aligned}
[F(t)]'_{ij} &= \left(\sum_{k\ge 0} \frac{1}{k!}[M^k]_{ij} t^k\right)' \\
&= \sum_{k\ge 0}\left(\frac{1}{k!}[M^k]_{ij} t^k\right)' \\
&= \sum_{k\ge 0} \frac{1}{k!}[M^k]_{ij} k t^{k-1} \\
&= \sum_{k\ge 1} \frac{1}{(k-1)!}[M^k]_{[}ij]t^{k-1} \\
&= \sum_{k\ge 0} \frac{1}{k!}[M^{k+1}]_{ij} t^k
\end{aligned}$$

Because we know this equation holds for each entry, we have an equation for the matrices themselves:

$$F(t)' = \sum_{k\ge 0} \frac{1}{k!} M^{k+1} t^k$$

Each term on the right side shares a common factor of M. For any finite sum we may factor out such a term:

$$M + tM^2 + \frac{1}{2}t^2M^3 + \cdots + \frac{1}{k!}t^k M^{k+1}$$

$$= M\left[I + tM + \frac{1}{2}t^2M^2 + \cdots + \frac{1}{k!}t^kM^k\right]$$

As $k \to \infty$ the series on the right converges exactly to the original series $F(t)$ - now multiplied by this common factor of M. And, the series on the left converges to $F'(t)$ by our calculation above. So, both sides of this equality converge and taking the limit yields

$$F(t)' = MF(t)$$

□

Now we utilize this to solve our *particular* differential equation. We've constructed a function whose derivative is M - the thing we wanted - but its not a solution to our differential equation because its a matrix valued function, and we are looking for a vector $(x(t), y(t), z(t))$.

Proposition 27.2. *Given a vector $v_o \in \mathbb{R}^n$, define the vector-valued function $s : \mathbb{R} \to \mathbb{R}^n$*

$$s(t) = e^{tM}v_o$$

for a matrix $M \in M_n(\mathbb{R})$. Then s satisfies the vector-valued differential equation

$$s' = Ms \qquad s(0) = v_0$$

Proof. Defining $s(t)$ by this formula, since $F(t) = e^{tM}$ converges to an $n \times n$ matrix for each t we are assured that $s(t)$ is a well defined vector for all time. So we need only check it has the required properties.

First, at $t = 0$ the series for e^{tM} collapses to its first term, the identity matrix and so

$$s(0) = e^{0M}v_0 = Iv_0 = v_0$$

Next, we wish to take the derivative of the vector equation $s(t)$. Writing this out using the limit definition of the derivative (now applied to a vector):

$$s'(t) = \lim_{h\to 0}\frac{e^{t+h}v_0 - e^t v_0}{h}$$

For each value of h we can factor out the constant vector v_0, and thus are left with a *limit of matrices* applied to this vector:

$$s'(t) = \left(\lim \frac{e^{t+h} - e^t}{h}\right)v_0 = \left(\lim_{h\to 0}\frac{F(t+h) - F(t)}{h}\right)v_0$$

But we already know the derivative of F! So,

$$s'(t) = F'(t)v_0 = MF(t)v_0$$

And, $F(t)v_0$ is the definition of $s(t)$: thus as claimed,

$$s'(t) = Ms(t)$$

□

This gives us an explicit solution to our example system:

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \exp\left[\begin{pmatrix} 2t & 3t & -4t \\ 3t & -t & t \\ t & 0 & -t \end{pmatrix}\right]\begin{pmatrix} 7 \\ 3 \\ 2 \end{pmatrix}$$

And in fact, provides a glimpse at just how powerful of a tool we've created. The matrix valued function $F(t) = \exp(tM)$ is a *solution generator*: it produces *every* solution to the differential equation $s' = Ms$ simply by multiplication by the initial condition. We can think of e^{tM} itself as a map from initial conditions to solutions.

$$e^{tM} : \text{Initial Condition} \mapsto \text{Solution}$$
$$v_0 \mapsto e^{tM} v_0$$

Such a perspective becomes even more important when we turn an eye towards *partial differential equations* below.

27.2.3. Exponential Operators

A partial differential equation is a differential equation for multivariate functions which involves derivatives with respect to multiple variables. Partial differential equations are a cornerstone of applied mathematics, and the applications of Analysis to the natural sciences. Some common examples are the heat equation from thermodynamics

$$\partial_t H(x,t) = \partial_x^2 H(x,y)$$

The wave equation from fluids, material science, and electromagnetism

$$\partial_t^2 W(x,t) = \partial_x^2 W(x,t)$$

the Schrodinger equation of Quantum mechanics

$$i\partial_t \Psi(x,t) = \frac{1}{2}\partial_x^2 \Psi(x,t) + V(x)\Psi(x,t)$$

and the Black-Scholes equation from economic theory:

$$\partial_t V(s,t) = \frac{\sigma^2}{2} s^2 \partial_s^2 V(s,t) = rV(s,t) - rs\partial_s V(s,t)$$

Solving partial differential equations in general is a much more difficult process than the ordinary differential equations discussed above, and so we do not attempt a comprehensive or rigorous treatment here. Instead, we content ourselves to simply explore a few simple cases where exponentiation can play an important role.

Example 27.1 (The Equation $\partial_t f = \partial_x f$)**.** Consider perhaps the simplest partial differential equation for a function $f(x,t)$, with the initial condition given as a function of x at $t = 0$

$$\partial_t f(x,t) = \partial_x f(x,t) \qquad f(x,0) = h(x)$$

One way to think about a function $f(x,t)$ is as a *one parameter family of functions of* x. That is, for each fixed value of t, the function $f(x,t)$ is just a function of x (which is often written $f_t(x)$ instead, to emphasize that t is a parameter and at each fixed t we get a *function* just of x). From this perspective, the differential equation is telling us something about how this collection of functions changes over time: that the rate of change in time is the same as applying the x-partial derivative at the given moment.

We can reason about this in analogy with the system of equations we discussed above. Indeed, just as the matrix M was a *linear transformation on vectors*, the x- derivative is a *linear transformation on functions* $h(x) \mapsto \frac{dh}{dx}$. If we were trying to build a *solution generator*, it would be a family of linear operators $F(t)$ where given a function $h(x)$, we could produce a solution via $s(t) = F(t)[h(x)]$. One might also try to determine the nature of $F(t)$ by analogy with the matrix case:

$$\partial_t s(x,t) = \partial_t \left[F(t)[h(x)]\right] = F'(t)[h(x)]$$

and notice that if $F'(t) = \partial_x F(t)$, then we would have

$$\partial_t s(x,t) = F'(t)[h(x)] = \partial_x F(t)[h(x)] = \partial_x s(x,t)$$

How could we attempt to build an family of operators with this property, that differentiating would "bring down an x-deriavtive?" Why not propose $F(t) = e^{t\frac{d}{dx}}$? We can attempt to define this as a power series

$$e^{t\frac{d}{dx}} = I + t\frac{d}{dx} + \frac{t^2}{2}\frac{d^2}{dx^2} + \cdots + \frac{t^n}{n!}\frac{d^n}{dx^n} + \cdots$$

And use this power series to give an explicit definition for how $e^{t\frac{d}{dx}}$ should act on functions:

$$\begin{aligned} e^{t\frac{d}{dx}}[h(x)] &= h(x) + th'(x) + \frac{t^2}{2}h''(x) + \cdots + \frac{t^n}{n!}h^{(n)}(x) + \cdots \\ &= \sum_{n\geq 0} \frac{t^n}{n!}h^{(n)}(x) \end{aligned}$$

Upon seeing this formula, one should certainly be thinking about convergence: is this infinite sum going to make sense, for all values of x and t? Or could it be that this diverges? An exercise below looks at the case that h is a power series, where we can prove explicitly using our real-analysis techniques that the series converges, so for now let's assume convergence, and investigate the behavior of our proposed function.

Proposition 27.3. *Let $h\colon \mathbb{R} \to \mathbb{R}$ be a function and assume that $s(x,t) = \sum_{n\geq 0} \frac{t^n}{n!}h^{(n)}(x)$ converges absolutely for all x, t.*
Furthermore assume the following technical condition on h, essentially saying the derivatives of h can't grow too fast: for each n the derivative $|h^{(n)}(x)|$ is bounded by some constant K_n, and $\lim \frac{K_n}{nK_{n-1}} \to 0$.
Then s solves the differential equation $\partial_t s = \partial_x s$ with $s(x,0) = h$.

Proof. For any fixed x, the proposed function $s(x,t)$ is a power series in t, and we have proven that inside the radius of convergence such a power series can be differentiated term by term. Thus,

$$\begin{aligned} \partial_t s(x,t) &= \partial_t \left(\sum_{n\geq 0} \frac{t^n}{n!}h^{(n)}(x) \right) \\ &= \sum_{n\geq 0} \partial_t \left(\frac{t^n}{n!}h^{(n)}(x) \right) \\ &= \sum_{n\geq 0} \frac{nt^{n-1}}{n!}h^{(n)}(x) \\ &= \sum_{n\geq 1} \frac{t^{n-1}}{(n-1)!}h^{(n)(x)} \\ &= \sum_{n\geq 0} \frac{t^n}{n!}h^{(n+1)}(x) \end{aligned}$$

Because our original power series converged absolutely for all t, x, this remains true after differentiation (as a corollary of the power series differentiation theorem, proved using dominated convergence.) For each n, we can rewrite $h^{(n+1)}(x)$ as $\frac{d}{dx}h^{(n)}(x)$ and thus

$$\frac{t^n}{n!}h^{(n+1)}(x) = \frac{t^n}{n!}\frac{d}{dx}h^{(n)}(x) = \partial_x \left(\frac{t^n}{n!}h^{(n)}(x) \right)$$

Plugging this back into our series, we see

$$\partial_t s(x,t) = \sum_{n\geq 0} \partial_x \left(\frac{t^n}{n!} h^{(n)}(x)\right)$$

Now we investigate the right-hand-side further. For any *finite sum* we know that

$$\sum_{n\leq N} \partial_x \left(\frac{t^n}{n!} h^{(n)}(x)\right) = \partial_x \left(\sum_{n\leq N} \frac{t^n}{n!} h^{(n)}(x)\right)$$

so all that needs to be justified is that this property remains true in the limit. But this is exactly what dominated convergence is built for, exchanging the limit and sum! Let's check the conditions of dominated convergence apply:

- For each n the term $\frac{t^n}{n!} h^{(n)}(x)$ is differentiable.
- For each x, the sum $\sum_n \frac{t^n}{n!} h^{(n)}(x)$ is convergent.

These follow immediately from our assumptions on h and s. Next we need to define a dominating series M_n of constants. Our technical assumption assures us that for each n there is some uniform constant K_n bounding the derivative $|h^{(n)}(x)|$, so we propose $M_n = \frac{t^n}{n!} K_n$. By definition this is greater than or equal to the n^{th} term in the series, so all we need to see is that it converges:

$$\sum_n M_n = \sum_n \frac{t^n}{n!} K_n$$

Performing the ratio test, we find

$$\begin{aligned}
\lim \left| \frac{\frac{t^n}{n!} K_n}{\frac{t^{n-1}}{(n-1)!} K_{n-1}} \right| &= \lim \left| \frac{t K_n}{n K_{n-1}} \right| \\
&= |t| \lim \frac{K_n}{n K_{n-1}} \\
&= 0
\end{aligned}$$

Where the last equality follows directly from our technical assumption. Thus, our proposed dominating series converges absolutely, and dominated convergence allows us to switch the order of the differentiation and summation to see

$$\partial_x s(x,t) = \partial_x \left(\sum_{n\geq 0} \frac{t^n}{n!} h^{(n)}(x)\right) = \sum_{n\geq 0} \partial_x \left(\frac{t^n}{n!} h^{(n)}(x)\right)$$

But the right hand side here is exactly what we found earlier must equal the partial t derivative! Stringing these together,

$$\partial_t s(x,t) = \sum_{n\geq 0} \partial_x \left(\frac{t^n}{n!} h^{(n)}(x)\right) = \partial_x s(x,t)$$

So, s satisfies the proposed differential equation. Last but not least, we check the intial condition by evaluating at $t = 0$:

$$s(x,0) = h(x) + 0 \cdot h'(x) + \frac{0^2}{2} h''(x) + \cdots = h(x)$$

□

We can rephrase the result above in more abstract language:

Corollary 27.5. *Let $\mathcal{F}$ be the space of functions, and $F(t) = e^{t\frac{d}{dx}}$ the operator $\mathcal{F} \to \mathcal{F}$ defined by*

$$e^{t\frac{d}{dx}}[h] = \sum_{n\geq 0} \frac{t^n}{n!} h^{(n)}$$

Then $e^{t\frac{d}{dx}}$ is a solution generator *for the differential equation $\partial_t f(x,t) = \partial_x f(x,t)$. Given any initial condition $h(x)$ with slow enough growing derivatives, $s = e^{t\frac{d}{dx}}[h(x)]$ is a solution to this differential equation.*

This is pretty incredible: just by analogy with the matrix case we were able to *propose* a solution using the power series for the exponential, and then with some real analysis prove this solution works! But we can go even farther, and understand the solution *geometrically* using what we know about the exponentiated derivative operator. Indeed, in Corollary 27.4 we show (following an exercise for you to complete) that at least if h is a *power series* we can readily understand the action of $e^{t\frac{d}{dx}}$:

$$e^{t\frac{d}{dx}} h(x) = h(x+t)$$

Thus, after *all of this hard work* we end up with a ridiculously simple solution:

Corollary 27.6. *If $h(x)$ is a differentiable function of x, then*

$$s(x,t) = h(x+t)$$

is the solution to $\partial_t s = \partial_x s$ with initial condition h.

This is trivial to confirm by hand, using the chain rule!

$$\partial_t h(x+t) = h'(x+t) = \partial_x h(x+t)$$

And in retrospect, we could have come up with this solution if we just thought hard enough, instead of diving into calculations! But our ability to write this solution in a geometrically - obvious manner is special to this case, and to the differential equation in question being particularly simple. The power of the technique above was that it *did not require us to be clever* the exponential may to the rescue even when - and especially when - our intuition and foresight fail us.

Part VII.

Integrals

28. Definition

Highlights of this Chapter: we give an *axiomatic* definition of the integral, and use these axioms to prove the fundamental theorem of calculus, as well as several corollaries such as u-substitution and integration by parts.

The *integral* is meant to measure the (net) area. When f is positive, for instance, we learn in calculus that $\int_a^b f dx$ should be the area under f between a and b. That is, it should be the area of the region $\mathscr{R} \subset \mathbb{R}^2$ below:

$$\int_a^b f\,dx = \text{Area}\,(\mathscr{R})$$

$$\mathscr{R} = \{(x, y) \in \mathbb{R}^2 \mid a \le x \le b \;\; 0 \le y \le f(x)\}$$

Thus a good theory of area would immediately lead to a good theory of integration. But how does one measure area? Perhaps surprisingly, this turns out to be much more difficult than it sounds - and all the difficulties weren't worked out until the beginning of the 20th century with the advent of *measure theory*.

We will not need the full power of this theory here - areas under the graphs of functions are a special enough case that we can develop a theory of integration independently. But our beginnings will be the same: area is a concept we struggle to define *explicitly* even though we know many rules it should behave. Thus, area is a prime target to try and characterize *axiomatically*, and then seek an explicit definition that realizes our axioms.

What are some natural axioms for area? Perhaps the most fundamental is that area is *additive*: if U, V are two disjoint subsets of the plane, then

$$\text{Area}(U \cup V) = \text{Area}(U) + \text{Area}(V)$$

It turns out this simple rule alone is enough to provide some axioms, which completely determine the theory of integration (for continuous functions, at least).

Remark 28.1. In fact, the definition of a measure is just a slight generalization of this: a measure μ is a function from a collection M of subsets (called measurable subsets) of a space X to $\mathbb{R}$ such that

- For all $A \in M$, $\mu(A) \ge 0$

- If A_n is a countable set of disjoint sets in M then

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$$

A measure on $\mathbb{R}^n$ is typically also asked to be *translation invariant* or more generally *isometry invariant*: if A, B are congruent subsets then $\mu(A) = \mu(B)$.

28.1. Properties of Area

We will try to produce some axioms for integration as a net area. Consider a function f on an interval $[a, b]$ (picture a positive function, if you want the analogy with area to be exact). We will write

$$\int_{[a,b]} f\,dx$$

For the *integral of* f. A first axiom: given a constant function $f(x) = k$, the region under its graph is a rectangle, and the area of a rectangle is base times height. Thus,

$$\int_{[a,b]} k\,dx = k(b-a)$$

Next, if f and g are two functions on $[a, b]$ where $f(x) \leq g(x)$ for all x, then the graph of f lies underneath the graph of g so since area is additive, we should require our integral to satisfy

$$\int_{[a,b]} f\,dx \leq \int_{[a,b]} g\,dx$$

And, finally if $c \in [a, b]$ we can divide the interval into two intervals $[a, c]$ and $[c, b]$ disjoint except for their boundary point c. Since area is additive, we should impose

$$\int_{[a,b]} f\,dx = \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx$$

These rules do not seem like much, but we will see that they are quite powerful: they completely determine the behavior of the integral for continuous functions.

28.2. Axioms

To take an axiomatic treatment seriously, we need to first make these rules more precise. The integral we have described is a function: it takes an input (a function on an interval) and gives a unique output (a real number). But what is the domain? A first hopeful thought might be “all functions” - but we might want to be wary of imposing this from the outset. After all, we have seen that real analysis allows many monstrous functions (like the function which is 0 on the irrationals and 1 on the rationals) that we might not want to - or even be able to - assign an area to!

In fact, this worry is quite real: the Mathematician Guiseppe Vitali showed in 1905 that there are subsets of $\mathbb{R}$ which cannot coherently be given a length with the example below.

Example 28.1 (⋆ Vitali Sets)**.** Two nubmers x and y are said to be *rationally related* if $x-y \in \mathbb{Q}$. This defines an equivalence relation on the interval $[0,1]$, and we choose V to be a set of equivalence class representatives (that is, V contains exactly one element from each equivalence class).

We now show that V has the following property: a countable number of disjoint copies of V cover all of $[0,1]$. Let R be the set of all rational numbers in $[-1,1]$, and note that R is countable. We can then define for each r the set $V_r = \{v+r \mid v \in V\}$, and note that if $r \neq s$ then V_r and V_s are disjoint. But, the union of all the V_r contains the interval $[0,1]$ since *every* point $x \in [0,1]$ is equivalent to some $v \in V$ by definition, meaning there is some $r \in \mathbb{Q}$ with $v + r = x$, and for both v and x to lie in $[0,1]$ requires $|v-x| = |r| \leq 1$ so $r \in [-1,1]$. Indeed, the union itself must be a subset of $[-1,2]$.

Now, if V can be assigned a length by our measure μ, there are two options either $\mu(V) = 0$ or $\mu(V) \neq 0$. We will show both lead to contradiction, so in fact V cannot be assigned such a length.

First, if $\mu(V) = 0$ then $\mu(V_r) = 0$ for all r as these are just translated copies of V. And, as the V_r are all disjoint, we can compute the total measure by adding the measure of each individually:

$$\mu(\cup_r V_r) = 0 + 0 + 0 + \cdots = 0$$

But this union *contains the unit interval $[0,1]$, which has length 1! And $1 > 0$ so this is a contradiction.

A similar argument prevents V from having any finite measure. If $\mu(V) = \epsilon$ for *any* positive ϵ, then the area of the union diverges

$$\mu(\cup_r V_r) = \epsilon + \epsilon + \cdots = \infty$$

But the union s contained in $[-1,2]$ which has length 3, and as $3 < \infty$ this is a contradiction as well.

It follows from this that certain subsets of the plane cannot be given an area: in particular, if V is one of Vitali's non-measurable sets, then the set of points under the graph of

$$\chi_V(x) = \begin{cases} 1 & x \in V \\ 0 & x \notin V \end{cases}$$

cannot be coherently assigned an area, and thus we cannot assign a value to the integral of χ_V on the interval $[0, 1]$.

Remark 28.2. This construction of non-measurable sets is a fancier version of the following argument that there can be no uniform probability distribution on that natural numbers: (for instance, this is the kind of thing you implicitly assume exists when you ask someone to *pick a random number*)

Say you want to assign each integer the same probability ϵ. Recall the total probability needs to be 100%: this leads to a problem because we need to solve

$$1 = \epsilon + \epsilon + \epsilon + \epsilon + \cdots$$

And if $\epsilon \neq 0$ this sum diverges, but if $\epsilon = 0$ this sum is zero, and in neither case is it 1.

Thus, because we've realized that trying to assign an area to *all* subsets of the plane (or even *all* regions under the graph of a function) is too much to ask, we need to specify *as part of our theory* a class of 'integrable functions', and impose our axioms only on those.

Definition 28.1. For any closed interval $J = [a, b]$ we denote by $\mathscr{I}(J)$ the set of integrable functions on J. Then a collection of functions $\mathscr{I}(J) \to \mathbb{R}$ is an *integral*, and denoted

$$f \mapsto \int_J f\,dx$$

if it satisfies the following axioms:

- If $k \in \mathbb{R}$ then $f(x) = k$ is an element of $\mathscr{I}([a, b])$ for any interval $[a, b]$ and

$$\int_{[a,b]} k\,dx = k(b - a).$$

- If $f, g \in \mathscr{I}([a, b])$ and $f(x) \leq g(x)$ for all $x \in [a, b]$ then

$$\int_{[a,b]} f\,dx \leq \int_{[a,b]} g\,dx$$

- If $[a, b]$ is an interval and $c \in (a, b)$, then $f \in \mathcal{I}([a, b])$ if and only if $f \in \mathcal{I}([a, c])$ and $f \in \mathcal{I}([c, b])$. Furthermore, in this case their values are related by

$$\int_{[a,b]} f\,dx = \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx$$

Note these axioms do not aim to *uniquely* specify an integral, but rather to delineate properties that *anything worthy of being called an integral* must have.

Example 28.2 (The "Constant Integral")**.** The first axiom tells us that *if* a constant is integrable, *then* we must have $\int_{[a,b]} k\,dx = k(b-a)$.

So, let $\mathcal{C}$ be the set of constant functions, and define an integral on $\mathcal{C}$ exactly by this formula. Then our integral satsifeies

- Axiom I, by definition
- Axiom II: if $k < K$ and $b - a > 0$ then $k(b-a) < K(b-a)$ so

$$k \leq K \implies \int [a,b] k\,dx \leq \int_{[a,b]} K\,dx$$

- Axiom III: Since for any c whatsoever $b - a = b - c + c - a$ we have $k(b-a) = k(b-c) + k(c-a)$. Using this for an arbitrary $c \in (a, b)$ yields

$$\int_{[a,b]} k\,dx = \int_{[a,c]} k\,dx + \int_{[c,b]} k\,dx$$

Thus, the assignment $\{k, [a,b]\} \mapsto k(b-a)$ defines an integral on the space of constant functions.

This integral is not particularly useful, as it is undefined for any non-constant function. One can make it *slightly* better by extending to an integral for linear functions $f(x) = mx + b$.

Exercise 28.1 (The "Linear Integral")**.** Let $\mathcal{L}(J)$ denote the set of linear functions $y = mx + b$ on the interval J. Show that the following rule defines an integral on $\mathcal{L}(J)$, satisfying the axioms.

$$\int_{[u,v]}^{\text{Lin}} mx + b\,dx = m\frac{v^2 - u^2}{2} + b(v-u)$$

Here we have spelled out the domain (of functions) clearly for the proposed integral, and given a formula by fiat. This is not the usual means of constructing an integral of course, as it requires we sort of *already know the answer*! The usual way we will let $\mathcal{I}(J)$ be determined is to write down a particular definition for the integral (as a limiting process of some kind) and then take $\mathcal{I}(J)$ to be the set of all functions on J for which that process converges.

28.2.1. Properties from the Axioms

In all of the following we assume that $\int$ is some integral satisfying the axioms above axioms.

Proposition 28.1. *If* $\{c\}$ *is the degenerate closed interval containing a single point, and* f *is a function which is integrable on any interval containing* a*, then*

$$\int_{\{a\}} f\,dx = 0$$

Proof. Let f be integrable on the interval $[u, v]$ and $a \in [u, v]$ be a point. Without loss of generality we can in fact take a to be one of the endpoints of the interval, by subdivision: if $a \in (u, v)$ then Axiom III implies that f is integrable on $[u, a]$ and on $[a, v]$ as well.

Thus, we assume f is integrable on $[a, v]$, and further subdivide this interval as

$$[a, v] = [a, a] \cup [a, v] = \{a\} \cup [a, v]$$

By subdivision, we see that f is integrable on $\{a\}$ and that

$$\int_{[a,v]} f\,dx = \int_{\{a\}} f\,dx + \int_{[a,v]} f\,dx$$

Subtracting the common integral over $[a, v]$ from both sides yields the result,

$$\int_{\{a\}} f\,dx = 0$$

□

Proposition 28.2. *If* $f \in \mathcal{I}([a, b])$ *is an integrable function, then there exists a function* $F : [a, b] \to \mathbb{R}$ *defined by*

$$F(x) = \int_{[a,x]} f\,dx$$

Proof. Again this is just subdivision at work: for any $x \in [a, b]$ we may write

$$[a, b] = [a, x] \cup [x, b]$$

. Then Axiom III implies that f is integrable on $[a, x]$, and so the number $\int_{[a,x]} f\,dx$ is defined. This assignment describes a real valued function

$$x \mapsto \int_{[a,x]} f\,dx$$

□

The above proposition has a short proof because it did not claim much: we learned nothing about the *nature* of the area function F. Drawing Calculus I style pictures of F makes one readily believe more should be possible - and indeed it is. With one additional assumption (that f is bounded on the interval in question) we can prove a pretty strong claim - its area function is *continuous*!

Theorem 28.1. *If $f \in \mathcal{I}([a,b])$ is a bounded integrable function, then its integral $F(x) = \int_{[a,x]} f\,dx$ is* continuous.

Proof. Let f be integrable and bounded by M on $[a,b]$, and set $F(x) = \int_{[a.x]} f\,dx$. Begin by choosing an $\epsilon > 0$. We will prove something even strong than asked - that f is *uniformly continuous* by finding a $\delta > 0$ where if $|y-x| < \delta$ we have $|F(y)-F(x)| < \epsilon$. Let's unpack this a bit: if $x < y$ are two points of $[a,b]$,

$$F(y) - F(x) = \int_{[a,y]} f\,dx - \int_{[a,x]} f\,dx$$

But subdivision (Axiom III) implies

$$\begin{aligned} F(y) &= \int_{[a,y]} f\,dx \\ &= \int_{[a,x]} f\,dx + \int_{[x,y]} f\,dx \\ &= F(x) + \int_{[x,y]} f\,dx \end{aligned}$$

Thus $F(y) - F(x)$ is just the integral of f on the subinterval $[x,y] \subset [a,b]$. Because f is bounded by M we know $-M \leq f(x) \leq M$. By subdivsion, f is then integrable on every sub-interval $I \subset [a,b]$, and by comparison (Axiom II) this implies

$$-M|I| \leq \int_I f\,dx \leq M|I|$$

So, we choose $\delta = \epsilon/M$. This immediately yields what we want, as if $|y-x| < \delta$,

$$-\epsilon = -M\delta < -M|y-x| \leq \int_{[x,y]} f\,dx \leq M|y-x| < M\delta = \epsilon$$

Thus $|F(y) - F(x)| = \left|\int_{[x,y]} f, dx\right| < \epsilon$. □

Remark 28.3. Of course, the proven result is not *really* stronger than what was asked, since we began on a closed interval, and we know that continuous on a closed interval implies uniformly continuous.

However, if you look carefully at the proof you see we nowhere used that the original domain was a closed interval! So what we have really proven is that the area function $F(x) = \int_{[a,b]} f\,dx$ is *uniformly continuous* anytime f is bounded!

28.3. The Fundamental Theorem

We've already seen that these meager axioms hide great power: we could prove that the integral of a bounded function was continuous *directly* without anything else! But this is only the start of an incredible story. Here, we jump straight to the main event - and prove that these axioms characterize the fundamental theorem of calculus!

Theorem 28.2 (The Fundamental Theorem of Calculus)**.** *Let f be a continuous function and assume that f is integrable on $[a,b]$. Denote its area function by*

$$F(x) = \int_{[a,x]} f\,dx$$

Then F is differentiable, and for all points $x \in (a,b)$,

$$F' = f$$

Proof. Because f is continuous on a closed interval, it is bounded (by the Extreme Value theorem), and so the area function F is continuous (Theorem 28.1).

Choose an arbitrary $c \in (a,b)$. We wish to show that $F'(c) = f(c)$: that is, we need

$$\lim_{x\to c} \frac{F(x) - F(c)}{x - c} = f(c)$$

In terms of ϵs and δs, this means for arbitrary ϵ we need to find a δ such that if x is within δ of c, this difference quotient is within ϵ of $f(c)$.

It will be convenient to separate this argument into two cases, depending on if $x < c$ or $c < x$ (both arguments are analogous, all that changes is whether the interval in question is $[c,x]$ or $[x,c]$). Below we proceed under the assumption that $c < x$. In this case, looking at the numerator, we see by subidvision (Axiom III) that

$$\begin{aligned}
F(x) &= \int_{[a,x]} f\,dx \\
&= \int_{[a,c]} f\,dx + \int_{[c,x]} f\,dx \\
&= F(c) + \int_{[c,x]} f\,dx
\end{aligned}$$

$$\implies F(x) - F(c) = \int_{[c,x]} f\,dx$$

Thus the real quantity of interest is this integral over $[c, x]$. Choose $\epsilon > 0$. Since f is continuous, there is some $\delta > 0$ where $|x - c| < \delta$ implies $|f(x) - f(c)| < \epsilon$. Equivalently, for all $x \in [c - \delta, c + \delta]$ we have

$$f(c) - \epsilon < f(x) < f(c) + \epsilon$$

By subdivison (Axiom III), we know that f is integrable on $[c, x]$, and so by comparison (Axiom II) and the area of rectangles (Axiom I) we have

$$(f(c) - \epsilon)(x - c) \leq \int_{[c,x]} f\,dx \leq (f(c) + \epsilon)(x - c)$$

Dividing through by $x - c$

$$f(c) - \epsilon \leq \frac{\int_{[c,x]} f\,dx}{x - c} \leq f(c) + \epsilon$$

and subtracting $f(c)$

$$-\epsilon \leq \frac{\int_{[c,x]} f\,dx}{x - c} - f(c) \leq \epsilon$$

We arrive at the inequality

$$\left|\frac{\int_{[c,x]} f\,dx}{x - c}\right| < \epsilon$$

But the numerator here is none other than $F(x) - F(c)$! So, we've done it: for all $x > c$ with $|x - c| < \delta$, we have the difference quotient within ϵ of $f(c)$. This implies the limit exists, and that

$$F'(c) = f(c)$$

□

Exercise 28.2. Write out the case for $x < c$ following the same logic as above.

This tells us that the area function of f is one of its antiderivatives! The theory of area is the *inverse* of the theory of rates of change. But which antiderivative? The mean value theorem assures us that the collection of all possible antiderivatives are easy to understand - any two differ by a constant (Corollary 24.3). So to uniquely specify an antiderivative its enough to give its value at one point. And we can do this!

Corollary 28.1. *Let f be a continuous function which is integrable on $[a, b]$. Then the function $F(x) = \int_{[a,x]} f\,dx$ is uniquely determined as the antiderivative of F such that $F(a) = 0$.*

This connection of integration with antidifferentiation and the classification of antiderivatives has a useful corollary for computation, which is often called the *second fundamental theorem*

Theorem 28.3 (FTC Part II)**.** *Let f be continuous and integrable on $[a, b]$ and let F be* any *antiderivative of f. Then*

$$\int_{[a,b]} f\,dx = F(b) - F(a)$$

Proof. Denote the area function for f as $A(x) = \int_{[a,x]} f\,dx$. Then the quantity we want to compute is $A(b)$.

Now, let F be any antiderivative of f. The first part of the fundamental theorem assures us that A is an antiderivative of f, and so Corollary 24.3 implies there is some constant C such that $A(x) - F(x) = C$, or $F(x) = A(x) + C$. Now computing,

$$\begin{aligned} F(b) - F(a) &= (A(b) + C) - (A(a) + C) \\ &= A(b) - A(a) + (C - C) \\ &= A(b) - A(a) \\ &= A(b) \end{aligned}$$

Where the last equality comes from the fact that $A(a) = \int_{\{a\}} f\,dx = 0$ (Proposition 28.1). □

We are going to have a lot of endpoint-subtraction going on, so its nice to have a notation for it.

Definition 28.2. Let $[a, b]$ be an interval and f a function. We write

$$f\Big|_{[a,b]} = f(b) - f(a)$$

as a shorthand for evaluation at the endpoints.

Remark 28.4. It is often convenient when doing calculations to introduce a slight generalization of the integral, which depends on an *oriented interval.* A natural notation for this is already in use in calculus, using the top and bottom of the integral sign for the locations of the 'ending' and 'starting' bound respectively:

$$\int_a^b f\,dx = \begin{cases} \int_{[a,b]} f\,dx & a \leq b \\ -\int_{[b,a]} f\,dx & a \geq b \end{cases}$$

Show that using this notation, we have a clean *generalized subdivision rule*: for **all points a, b, c irrespective of their orderings,

$$\int_a^b f\,dx = \int_a^c f\,dx + \int_c^b f\,dx$$

This notation helps shorten the computations in the proof of the fundamental theorem (at the expense of adding one new thing to remember).

The fundamental theorem of calculus is a beautiful result for many different reasons. One of course, is that it forges a deep connection between the theory of areas and the theory of derivatives - something missed by the ancients and left undiscovered until the modern advent of the calculus. But second, it shows how incredibly constraining our simple axioms are: we did not prove the fundamental theorem of calculus for any particular definition of the integral (Riemann's, Lebesgue's, Darboux's, etc) but rather showed that *if continuous functions are integrable* then your theory of integration has no choice whatsoever on how to integrate them!

We've seen above that it is possible to construct explicit models of the integration axioms by artificially limiting the domain of integrable functions (to constants, or linear functions for instance). But even these are constrained by the Fundamental theorem: since our example functions were continuous, there *was really no choice at all*!

The remaining question is of course, is there a theory of integration where *all* continuous functions are integrable? We will call any definition of integration *interesting* if it is general enough to include all continuous maps.

Definition 28.3. An integral $\int$ is *interesting* if $\mathscr{I}(J)$ contains the continuous functions on J, for each closed interval $J \subset \mathbb{R}$.

28.3.1. Application: Integration Techniques

Given the fundamental theorem holds for continuous functions, its immediate to build up a strong theory of integration

Theorem 28.4. *Let* $\int$ *be an interesting integral, and* f *be continuous,* g *be continuously differentiable on* $[a, b]$*. Then*

$$\int_{[a,b]} (f \circ g)g' \, dx = \int_{[g(a),g(b)]} f \, dx$$

Proof. Because g is differentiable it is continuous, so $f \circ g$ is the composition of continuous functions, which is continuous. And as the product of continuous functions is continuous, $f(g(x))g'(x)$ is also continuous. Thus this function is integrable on $[a, b]$.

By the Fundamental Theorem, we can evaluate this by antidifferentiation: let $F(x)$ be any antiderivative of f, then the chain rule gives

$$(F(g(x)))' = F'(g(x))g'(x) = f(g(x))g'(x)$$

Using this antiderivative yields

$$\int_{[a,b]} (f \circ g)g' \, dx = F(g(x))\Big|_{[a,b]} = F(g(b)) - F(g(a))$$

A crucial but seemingly simple observation is to note this is the same value one would get by evaluating the function F on the endpoints of the interval $[g(a), g(b)]$:

$$F(g(x))\Big|_{[a,b]} = F(x)\Big|_{[g(a),g(b)]}$$

And as $F' = f$, this second expression is exactly what one would get from integrating f on the interval $[g(a), g(b)]$ using the Fundamental Theorem. □

Similarly without any further theory we can construct the other main integration technique of the calculus: integration by parts!

Theorem 28.5. *Let* $\int$ *be an interesting integral, and* f, g *be two continuously differentiable functions on* $[a, b]$*. Then*

$$\int_{[a,b]} fg' \, dx = fg\Big|_{[a,b]} - \int_{[a,b]} f'g \, dx$$

Exercise 28.3. Prove this using a similar strategy as to what we did above, but using the product rule instead of the chain rule as a starting point.

28.4. The Work to Come

Its pretty incredible that even though we did not set out to *uniquely* define the integral via our axioms, they manage to completely determine the integral for any function f which is (1) continuous and (2) integrable.

A natural and important question then is which continuous functions *are* integrable? (Or, in our terminology above, is there an *interesting integral* at all?). As soon as we know f is integrable, we get the existence of the area function F via subdivision and the proof of the Fundamental Theorem goes through without issue. But how does one construct an integral where one can actually prove all continuous functions are integrable?

28.4.1. Failure of the 'Calculus Integral'

The example below shows this is actually a difficult problem to answer: one might try to define the integral using a right endpoint Riemann sum (as one would in a calculus course): from this definition one can prove that all continuous functions *are* integrable, but then when one goes to try and verify the axioms, one finds this is actually not an integral at all!

Definition 28.4 (The "Calculus Integral")**.** Let f be a function defined on the interval $[a, b]$, and N a natural number. With $\Delta = (b - a)/N$ we define the (right endpoint) Riemann sum for f with N subintervals is

$$\sum_{i=1}^{n} f(a + i\Delta)\Delta$$

Such a function f is *Calculus - integrable* if the limit of its Riemann sums exists as the number of subintervals goes to infinity. In this case, the *Calculus Integral* is defined as the limiting value:

$$\int_{[a,b]}^{\text{Calc}} f(x)dx = \lim_{N\to\infty} \sum_{i=1}^{N} f(a + i\Delta x)\Delta x$$

It turns out that while this definition seems unproblematic when applied to elementary functions seen in a calculus course, it has some rather surprising behavior in general: and taking it as our definition would destroy some of the familiar pillars of integration theory!

To find the trouble, we need to look away from the well behaved functions, and investigate the integrability of some monsters. Here we'll look at the characteristic function of the rationals.

$$\chi(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Example 28.3. Let χ be the above function, equal to 1 on the rationals and 0 on the irrationals. Then f is Calculus - Integrable on every interval of the form $[0, a]$ but

$$a \in \mathbb{Q} \implies \int_{[0,a]}^{\text{Calc}} \chi dx = a$$

$$a \notin \mathbb{Q} \implies \int_{[0,a]}^{\text{Calc}} \chi dx = 0$$

In fact, its worse than this! As a natural extension of the above, one can show the following:

Exercise 28.4. The function χ is Calculus-Integrable on any closed interval in $\mathbb{R}$, and the resulting value is:

- The length of the interval, when both endpoints are rational.
- Zero, when one endpoint is rational and the other irrational

This has a very important consequence to our theory: our proposed definition of the integral *violates* the subdivision rule.

Exercise 28.5. The subdivison rule

$$\int_{[a,b]}^{\text{Calc}} f\, dx = \int_{[a,c]}^{\text{Calc}} f\, dx + \int_{[c,b]}^{\text{Calc}} f\, dx$$

is **false** for the integral as defined in Definition 28.4.
Hint: look at the interval $[0, 2]$, *and note* $0 < \sqrt{2} < 2$.

29. Construction

29.1. Definition

Here we give some preliminary definitions leading to the construction of an integral that we will then try to prove satisfies the axioms. The main difficulty is to put down a strict criterion that determines when a function is integrable, and when it is not.

The idea here - due to Darboux - is to try to overestimate and underestimate the true value of an integral by increasingly precise estimates. If the two estimates coincide as they get better and better we say the function is integrable. If they do not, we say it is not.

We build the estimates as *bar-graph-like* functions; similar to the familiar Riemann Sums from calculus.

Definition 29.1 (Partition)**.** A partition of the interval $I = [a, b]$ is a finite ordered set $P = \{t_0, t_1, \dots, t_n\}$ with $a = t_0 < t_1 < \dots < t_{n-1} < t_N = b$.

- N is called the *length* of the partition
- We write $P_i = [t_i, t_{i+1}]$ for the i^{th} interval of P, and $|P_i| = (t_{i+1} - t_i)$ for its width.

- The *maxwidth* of P is the maximal width of the P's intervals, $\text{maxwidth}(P) = \max_{0\leq i<N}\{|P_i|\}$.
- The set of all partitions on a fixed interval I is denoted $\mathscr{P}_I$.

$$\mathscr{P}_I = \{P \,:\, P \text{ is a partition of } I\}$$

Definition 29.2. Let f be a function, and P a partition of the closed interval I. For each segment $P_i = [t_i, t_{i+1}]$, we define

$$m_i = \inf_{x\in P_i}\{f(x)\} \qquad M_i = \sup_{x\in P_i}\{f(x)\}$$

We then define the *upper sum* $U_I(f, P)$ and the *lower sum* $L_I(f, P)$ as

$$L_I(f, P) = \sum_{0\leq i<N} m_i|P_i|$$

$$U_I(f, P) = \sum_{0\leq i<N} M_i|P_i|$$

Definition 29.3. Let f be a function on the closed interval I. Then we define the *upper integral* $U_I(f)$ and the *lower integral* $L_I(f)$ as

$$U(f) = \inf_{P\in\mathscr{P}_I}\{U_I(f,P)\}$$

$$L(f) = \sup_{P\in\mathscr{P}_I}\{L_I(f,P)\}$$

Definition 29.4 (The Darboux Integral). Let f be a function on the closed interval I. Then f is *Darboux-Integrable* on I if $U(f) = L(f)$, and we define the integral to be this common value:

$$\int_{[a,b]} f\,dx = U(f) = L(f)$$

29.2. Partitions

The goal of this section is to prove the seemingly obvious fact $L_I(f) \leq U_I(f)$. This takes more work than it seems at first because of the definitions of $L_I(f)$ as a supremum and $U_I(f)$ as an infimum, but proves an invaluable tool in analyzing integrability.

Definition 29.5. A partition Q is a refinement of a partition P if Q contains all the points of P (that is, $P \subset Q$).

Proposition 29.1 (Refinement Lemma). *If Q is a refinement of the partition P on a closed interval I, then for any bounded function f the following inequalities hold*

$$L_I(f,P) \leq L_I(f,Q) \leq U_I(f,Q) \leq U_I(f,P)$$

Proof. Here we give the argument for lower sums, the analogous case for upper sums is asked in Exercise 29.1. Since $P \subset Q$ and both P, Q are finite sets we know Q contains finitely many more points than P. Here we will show that if Q contains exactly one more point than P, that the claim holds; the general case follows by induction.

In this case we may write $Q = P \cup \{z\}$, where z lies within the partition $P_k = [t_k, t_{k+1}]$. Thus, $Q_k = [t_k, c]$ for the left half after subdivision, and $Q_{k+1} = [c, t_{k+1}]$ for the right half. Outside of P_k, the two partitions are identical, so their difference is given only by the difference of their values on P_k:

$$L_I(f,Q) - L_I(f,P) =$$

$$\left(\inf_{x\in Q_k}\{f(x)\}\,|Q_k| + \inf_{x\in Q_{k+1}}\{f(x)\}\,|Q_{k+1}|\right) - \left(\inf_{x\in P_k}\{f(x)\}\,|P_k|\right)$$

Since both Q_k and Q_{k+1} are *subsets* of P_k, the infimum over each of them is at its smallest the infimum over the whole set. This implies

$$
\begin{aligned}
&\inf_{x\in Q_k}\{f(x)\}\,|Q_k| + \inf_{x\in Q_{k+1}}\{f(x)\}\,|Q_{k+1}| \\
&\geq \inf_{x\in P_k}\{f(x)\}|Q_k| + \inf_{x\in P_k}\{f(x)\}|Q_{k+1} \\
&= \inf_{x\in P_k}\{f(x)\}\,(|Q_k| + |Q_{k+1}) \\
&= \inf_{x\in P_k}\{f(x)\}|P_k|
\end{aligned}
$$

Thus, the first term in the difference above is bigger than the second, so the overall difference is positive. Thus $L_I(f,Q) - L_I(f,P) \geq 0$ and so as claimed,

$$L_I(f,Q) \geq L_I(f,P)$$

□

Exercise 29.1. Following the structure above, prove that if Q refines P, that

$$U_I(f,Q) \leq U_I(f,P)$$

Proposition 29.2. *Lower sums are always smaller than upper sums, independent of partition. That is, if P,Q be two arbitrary partitions of a closed interval I, for any bounded function f,*

$$L_I(f,P) \leq U_I(f,Q)$$

Proof. Let P and Q be two arbitrary partitions of the interval I, and consider the partition $P\cup Q$. This contains both P and Q as subsets, so is a *common refinement* of both.

Using our previous work, this implies

$$L(f,P) \leq L(f,P\cup Q) \qquad U(f,P\cup Q) \leq U(f,Q)$$

We also know that for the partition $P\cup Q$ itself,

$$L(f,P\cup Q) \leq U(f,P\cup Q)$$

Taken together these produce the the string of inequalities

$$L(f,P) \leq L(f,P\cup Q) \leq U(f,P\cup Q) \leq U(f,Q)$$

From which immediately follows that $L(f,P) \leq U(f,Q)$, as desired. □

Proposition 29.3. *Let I be any closed interval and f a bounded function on I. Then the lower integral is less than or equal to the upper integral,*

$$L_I f \leq U_I f.$$

Proof. Recall that $U(f)$ is the *infimum* over all partitions of the upper sums.
Let P be an arbitrary partition. By Proposition 29.2 we know the upper sum with respect to *any partition whatsoever* is greater than or equal to $L(f,P)$, so $L(f,P)$ is a *lower bound* for the set of all upper sums.

Thus, the infimum of the upper sums - the *greatest* of all lower bounds - must be at greater or equal to this specific lower bound,

$$L(f,P) \leq \inf_{Q\in\mathscr{P}}\{U(f,Q)\} = U(f)$$

But this holds for *every partition P*. That means this number $U(f)$ is actually an *upper bound* for the set of all $L(f,P)$. And so, it must be greater than or equal to the *least upper bound* $L(f)$:

$$L(f) \leq U(f)$$

□

Corollary 29.1. *To show integrability it is enough to prove $U_I f \leq L_I f$.*

Proof. We know in general that $L_I(f) \leq U_I(f)$ from Proposition 29.3. So, if $U_I f \leq L_I f$ then in fact they are equal, which is the definition of f being integrable. □

29.3. Integrability Criteria

Here we prove a very useful condition to test if a function is integrable, by finding sufficient partitions.

Theorem 29.1 (Darboux Integrability Criterion)**.** *Let f be a bounded function on a closed interval I. Then f is integrable if and only if for all $\epsilon > 0$ there exists a partition P such that*

$$U_I(f,P) - L_I(f,P) < \epsilon$$

Here we prove one direction of this theorem, namely that if such partitions exist for all $\epsilon > 0$ then f is integrable. We prove the converse below.

Proof. Let $\epsilon > 0$, and assume there is a partition P with

$$U_I(f,P) - L_I(f,P) < \epsilon$$

Then, recalling $L_I(f,P) \le L_I(f)$ and $U_I(f) \le U_I(f,P)$ by definition, we chain these together with $L_I(f) \le U_I(f)$ to get

$$L_I(f,P) \le L_I(f) \le U_I(f) \le U_I(f,P)$$

Thus, the interval $[L_I(f), U_I(f)]$ is contained within the interval $[L_I(f,P), U_I(f,P)]$ which has length $< \epsilon$. Thus its length must also be less than ϵ:

$$0 \le U_I(f) - L_I(f) \le \epsilon$$

But ϵ was arbitrary! Thus the only possibility is that $U_I(f) - L_I(f) = 0$, and so the two are equal, meaning f is integrable as claimed. □

Now we prove the second direction of Theorem 29.1: the proof is reminiscent of the triangle inequality, though without absolute values (as we know terms of the form $U - L$ are always nonnegative already)

Proof. Assume that f is integrable, so $L_I(f) = U_I(f)$. Since $U_I(f)$ is the greatest lower bound of all the upper sums, for any $\epsilon > 0$, $U_I(f) + \frac{\epsilon}{2}$ is not a lower bound: that is, there must be some partition P_1 where

$$U_I(f,P_1) < U_I(f) + \frac{\epsilon}{2}$$

Similarly, since $L_I(f)$ is the least upper bound of the lower sums, there must be some partition P_2 with

$$L_I(f,P_2) > L_I(f) - \frac{\epsilon}{2}$$

Now, define $P = P_1 \cup P_2$ to be the common refinement of these two partitions, and observe that

$$\begin{aligned} U_I(f,P) - L_I(f,P) &\le U_I(f,P_1) - L_I(f,P_2) \\ &< \left(U_I(f) + \frac{\epsilon}{2}\right) - \left(L_I - \frac{\epsilon}{2}\right) \\ &= U_I(f) - L_I(f) + \epsilon \\ &= \epsilon \end{aligned}$$

Where the last inequality uses $L_I(f) = U_I(f)$. Thus, for our arbitrary ϵ we found a partition on which the upper and lower sums differ by less than that, as claimed. □

And finally, we provide an even stronger theorem than ϵ-integrability, that lets us prove a function is integrable *and* calculate the resulting value, by taking the limit of carefully chosen sequences of partitions. More precisely, we want to consider any sequence of partitions that's *getting finer and finer*:

Definition 29.6 (Shrinking Paritions)**.** A sequence $P_n \in \mathcal{P}_I$ of partitions is said to be *shrinking* if the corresponding sequence of max-widths converges to 0.

We often abbreviate the phrase P_n *is a shrinking sequence of partitions* by $P_n \to 0$.

Theorem 29.2. *Let f be a function on the interval I, and assume that P_n, P_n' are two sequences of shrinking partitions such that*

$$\lim L_I(f, P_n) = \lim U_I(f, P_n')$$

Then, f is integrable on I and $\int_I f\,dx$ is equal to this common value.

Proof. Call this common limiting value X. As $L_I f$ is defined as a supremum over all lower sums

$$\begin{aligned}\lim L_I(f, P_n) &\leq \sup_{\{n\in\mathbb{N}\}} \{L_I(f, P_n)\} \\ &\leq \sup_{P\in\mathcal{P}_I} \{L_I(f, P)\} \\ &= L_I(f)\end{aligned}$$

Similiarly, as $U_I(f)$ is the *infimum* over all upper sums, we have

$$\lim U_I(f, P_n') \geq U_I(f)$$

By Proposition 29.3 we know $L_I(f) \leq U_I(f)$, which allows us to string these inequalities together:

$$\lim L_I(f, P_n) \leq L_I(f) \leq U_I(f) \leq \lim U_I(f, P_n')$$

Under the assumption that these two limits are equal, all four quantities in this sequence must be equal, and in particular $L_I(f) = U_I(f)$. Thus f is integrable, and its value coincides with the limit of either of these sequences of shrinking partitions, as claimed. □

29.4. Riemann Sums

There is an alternative (and equivalent) construction of this integral which predates the construction above. Originally due to Riemann, this alternative version can be more complicated to work with, but has certain advantages: it makes a clear path from the general theory to *numerical integration*, and occasionally allows alternative, more algebraic proofs of various integral identities, avoiding discussions of suprema and infima.

We can compute an integral via Riemann sums, when the integral exists. To write down the definition we need to talk of *sample points* for a partition.

Definition 29.7 (Sample Set)**.** Let P be a partition of $[a,b]$. Then an *sample set* for P is a set $S=\{s_1,\ldots,s_n\}\subset[a,b]$ of points with $s_i\subset P_i$ for all i.

The set of all sample sets for a fixed partition P is denoted $\mathcal{S}_P$.

$$\mathcal{S}_P=\{S\mid S\text{ is a sample set for }P\}$$

Given a partition and a set of sample points, we can define a Riemann sum

Definition 29.8. If P is a partition of an interval I, S is an evaluation set for P, and f is a function defined on I, the *Riemann sum of f with respect to (P,S) is

$$\sum\nolimits_I(f,P,S):=\sum_{i=1}^{n}f(s_i)|P_i|$$

Definition 29.9 (The Riemann Integral)**.** Let f be a function defined on the closed interval $[a,b]$. Then f is Riemann-integrable on $[a,b]$ if for *every sequence* of P_n shrinking partitions, and for *every choice of sample sets* S_n for these partitions, $\lim\sum_I(f,P_n,S_n)$ exists, and is independent of the choice of sequences P_n,S_n. In this case, we write

$$\int_I^{\text{Riem.}}f=\lim\sum\nolimits_I(f,P_n,S_n)$$

Theorem 29.3. *Let f be a function on a closed interval I. Then f is Darboux integrable on I if and only if f is Riemann-Integrable on I, and the two integrals agree*

$$\int_I f\,dx=\int_I^{Riem.}f\,dx$$

Riemann $\implies$ Darboux.

□

Darboux $\implies$ *Riemann.*

□

Corollary 29.2. *If f is integrable, then $\int_{[a,b]} f dx$ can by computed via* Riemann Sum*: choosing any sequence P_n of shrinking partitions, and any sequence S_n of sample points for each partition,*

$$\int_{[a,b]} f\, dx = \lim_n \sum\nolimits_{[a,b]}(f, P_n, S_n)$$

30. ⋆ Examples

Highlights of this Chapter: we compute several integrals *directly from the definition*. The examples of x and x^2 are perhaps familiar from many sources, but we also compute the integral of an exponential $E(x)$, and prove that $\int_{[1,x]} \frac{1}{t}\, dt$ is a logarithm.

This entire section is *incredibly superfluous*. After all, we already know that *if* our construction really yields an integral then *the fundamental theorem of calculus must hold*, and we can compute any of these integrals below by antidifferentiation.

So, as efficient mathematicians, we should not pause to try and compute any integrals by hand, but rather move immediately to try and prove the integration axioms for our construction (we do this right away, at the beginning of the next chapter). Nonetheless, when learning a new definition it is often instructive to *use it*: and below are several example integrals computed directly from the construction.

30.1. Some Polynomials

Proposition 30.1 (Integrating a Constant)**.** *Let $f(x) = k$ be a constant function. Then f is integrable on any closed interval of $\mathbb{R}$, and*

$$\int_{[a,b]} k = k(b-a)$$

Proof. Let P be any partition of $[a,b]$ then since $f(x) = k$ is constant, on every subinterval P_i we have $m_i = k = M_i$, and so

$$L(f,P) = \sum_i m_i |P_i| = \sum_i k|P_i| = k(b-a)$$

$$U(f,P) = \sum_i M_i |P_i| = \sum_i k|P_i| = k(b-a)$$

Thus for *all partitions* the upper sum and lower sum are constant - and equal the same value! Taking the supremum over all lower sums and infimum over all upper sums then just yields this same constant, so the upper and lower integrals are equal. Thus

$$\int_{[a,b]} f\,dx = k(b-a)$$

□

Proposition 30.2 (Integrating x). *Let $[a,b]$ be any closed interval in $\mathbb{R}$. Then $f(x)=x$ is integrable on $[a,b]$ and*

$$\int_{[a,b]} x = \frac{b^2-a^2}{2}$$

Proof. Start with $[0,b]$, then look at $0<a<b$ using interval subdivision. To show x is integrable, we use Theorem 29.2, which assures us it is enough to find a sequence P_n of shrinking partitions where $\lim L(f,P_n) = \lim U(f,P_n)$.

For each n, let P_n be the evenly spaced partition with n subintervals, of width $\Delta_n = (b-a)/n$. Since $f(x)=x$ is monotone increasing, we know on each subinterval $[t_{i-1},t_i]$ that

$$m_i = t_{i-1} = (i-1)\Delta_n \qquad M_i = t_i = i\Delta_n$$

Thus, the upper and lower sums for these partitions are

$$\begin{aligned} L(x,P_n) &= \sum_{1\le i\le n} m_i\Delta_n = (i-1)\Delta_n\Delta_n \\ &= \Delta_n^2\,(0+1+2+\cdots+(n-1)) \end{aligned}$$

$$\begin{aligned} U(x,P_n) &= \sum_{1\le i\le n} M_i\Delta_n = i\Delta_n\Delta_n \\ &= \Delta_n^2\,(1+2+\cdots+n) \end{aligned}$$

These are nearly identical formulae: the upper sum is just one term longer than the lower sum and so their difference is

$$U(x,P_n) - L(x,P_n) = n\Delta_n^2 = n\frac{b^2}{n^2} = \frac{b^2}{n}$$

As $n\to\infty$ this converges to zero: thus, if either the upper *or* lower sum converges, then both do, and both converge to the same value by the limit theorems. For example, if we prove $U(f,P_n)$ converges then

$$\begin{aligned}\lim L(x, P_n) &= \lim \left(U(x, P_n) - U(f, P_n) + L(x, P_n)\right)\\&= \lim U(x, P_n) - \lim(U(x, P_n) - L(x, P_n))\\&= \lim U(xs, P_n) + 0s\end{aligned}$$

So, we focus on just proving that $U(x, P_n)$ converges and finding its value. Because $U(x, P_n)$ is a multiple of $1 + 2 + \cdots + n$, we start by finding a closed form using the formula for the sum of the first n positive integers: $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.

$$U(x, P_n) = \Delta_n^2 \frac{n(n+1)}{2} = \frac{b^2}{n^2}\frac{n(n+1)}{2} = \frac{b^2}{2}\frac{n(n+1)}{n^2}$$

The factor $b^2/2$ out front is a constant independent of n, and the remainder simplifies directly with some algebra:

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

Thus $\lim U(x, P_n) = \frac{b^2}{2}\lim(1 + 1/n) = \frac{b^2}{2}$. Since this converges our previous work ensures that the lower sum does as well, and to the same value. Thus x is integrable on $[0, b]$ and

$$\int_{[0,b]} x\,dx = \frac{b^2}{2}$$

Knowing this, we complete the case for a general positive interval $[a, b]$ with $0 < a < b$ by subdivision:

$$\int [a, b] x\,dx = \int_{[0,a]} x\,dx + \int_{[a,b]} x\,dx$$

Since we know the value of all integrals over intervals beginning at 0, this simplifies to

$$\frac{b^2}{2} = \frac{a^2}{2} + \int_{[a,b]} x\,dx$$

And, subtracting to the other side gives our answer

$$\int [a, b] x\,dx = \frac{b^2 - a^2}{2}$$

□

Exercise 30.1. Complete the general proof by dealing with the cases where a, b may be negative.

The same style of argument works to integrate any x^n for which we know how to sum the n^{th} powers of the positive integers.

Proposition 30.3 (Integrating x^2). *Let $[a, b]$ be any closed interval in $\mathbb{R}$. Then $f(x) = x^2$ is integrable on $[a, b]$ and*

$$\int_{[a,b]} x^2\,dx = \frac{b^3 - a^3}{3}$$

Exercise 30.2. Following the same technique as above, show that x^2 is integrable on $[a, b]$:

- First, restrict yourself to intervals of the form $[0, b]$ for $b > 0$.
- Use the monotonicity of x^2 on these intervals to explicitly write out upper and lower sums.
- Use the following identity on sums of squares from elementary number theory to compute their value

$$\sum_{1 \leq k \leq N} k^2 = \frac{N(N+1)(2N+1)}{6}$$

- Explain how to generalize this to intervals of the form $[a, 0]$ for $a < 0$, and finally to general intervals $[a, b]$ for any $a < b \in \mathbb{R}$ using subdivision.

30.2. Exponentials

Here's a quite long calculation showing that it's possible to integrate exponential functions directly from first principles. The length of this calculation alone is a good selling point for the fundamental theorem of calculus!

Proposition 30.4. *Let E be an exponential function, and $[a, b]$ an interval. Then E is integrable on $[a, b]$ and*

$$\int_{[a,b]} E = \frac{E(b) - E(a)}{E'(0)}$$

Proof. We will show the argument for E an increasing exponential (its base $E(1) > 1$): an identical argument applies to decreasing exponentials (only switching U and L in the computations below).

To show $E(x)$ is integrable, we use Theorem 29.2, which assures us it is enough to find a sequence P_n of shrinking partitions where $\lim L(f, P_n) = \lim U(f, P_n)$. Indeed - for each n, let P_n denote the evenly spaced partition of $[a, b]$ with widths $\Delta_n = (b - a)/n$

$$P_n = \{a, a + \Delta_n, a + 2\Delta_n, \cdots, a + n\Delta_n = b\}$$

We will begin by computing the lower sum. Because E is continuous, it achieves a maximum and minimum value on each interval $P_i = [t_i, t_{i+1}]$. And, since E is monotone increasing, this value occurs at the leftmost endpoint. Thus,

$$\begin{aligned} L(E, P_n) &= \sum_{0 \le i < n} \inf_{P_i}\{E(x)\}|P_i| \\ &= \sum_{0 \le i < n} E(t_i)\Delta_n \\ &= \sum_{0 \le i < n} E(a + i\Delta_n)\Delta_n \end{aligned}$$

Using the law of exponents for E we can simplify this expression somewhat:

$$\begin{aligned} E(a + i\Delta_n) &= E(a)E(i\Delta_n) \\ &= E(a)E(\Delta_n + \Delta_n + \cdots + \Delta_n) \\ &= E(a)E(\Delta_n)E(\Delta_n)\cdots E(\Delta_n) \\ &= E(a)E(\Delta_n)^i \end{aligned}$$

Plugging this back in and factoring out the constants, we see that the summation is actually a partial sum of a geometric series:

$$\begin{aligned} \sum_{0 \le i < n} E(a + i\Delta_n)\Delta_n &= \sum_{0 \le i < n} E(a)E(\Delta_n)^i\Delta_n \\ &= E(a)\Delta_n \sum_{0 \le i < n} E(\Delta_n)^i \end{aligned}$$

Having previously derived the formula for the partial sums of a geometric series, we can write this in closed form:

$$\sum_{0 \le i < n} E(\Delta_n)^i = \frac{1 - E(\Delta_n)^n}{1 - E(\Delta_n)}$$

But, we can simplify even further! Using again the laws of exponents we see that $E(\Delta_n)^n$ is the same as $E(n\Delta_n)$, and $n\Delta_n$ is nothing other than the width of our entire interval, so $b - a$. Thus the numerator becomes $1 - E(b - a)$, and putting it all together yields a simple expression for $L(E, P_n)$:

$$L(E, P_n) = E(a)\Delta_n \frac{1 - E(b - a)}{1 - E(\Delta_n)}$$

Some algebraic re-arrangement is beneficial: first, note that by the laws of exponents we have

$$\begin{aligned} E(a)(1-E(b-a)) &= E(a) - E(b-a)E(a) \\ &= E(a) - E(b) \end{aligned}$$

Thus for every n we have

$$L(E, P_n) = (E(a) - E(b)) \frac{\Delta_n}{1 - E(\Delta_n)}$$

We are interested in the limit as $n \to \infty$: by the limit laws we can pull the constant $E(a) - E(b)$ out front, and only concern ourselves with the fraction involving Δ_n. There's one final trick: look at the negative reciprocal of this fraction:

$$\frac{-1}{\frac{\Delta_n}{1-E(\Delta_n)}} = \frac{E(\Delta_n) - 1}{\Delta_n}$$

Because we know $E(0) = 1$ for all exponentials, this latter term is none other than the *difference quotient defining the derivative* for E! Since we have proven E to be differentiable, we know that evaluating this along any sequence converging to zero yields the derivative at zero. And as $\Delta_n \to 0$ this implies

$$\lim \frac{E(\Delta_n) - E(0)}{\Delta_n} = E'(0)$$

Thus, our original limit $\Delta_n/(1 - E(\Delta_n))$ is the negative reciprocal of this, and

$$\begin{aligned} \lim L(E, P_n) &= \lim (E(a) - E(b)) \frac{\Delta_n}{1 - E(\Delta_n)} \\ &= (E(a) - E(b)) \lim \frac{\Delta_n}{1 - E(\Delta_n)} \\ &= (E(a) - E(b)) \frac{-1}{E'(0)} \\ &= \frac{E(b) - E(a)}{E'(0)} \end{aligned}$$

Phew! That was a lot of work! Now we have to tackle the upper sum. But luckily this will not be nearly as bad: we can reuse most of what we've done! Since E is monotone increasing, we know that the maximum on any interval occurs at the rightmost endpoint, so

$$\begin{aligned}
U(E, P_n) &= \sum_{0 \le i < n} \sup_{P_i}\{E(x)\}|P_i| \\
&= \sum_{0 \le i < n} E(t_{i+1})\Delta_n \\
&= \sum_{0 \le i < n} E(a + (i+1)\Delta_n)\Delta_n
\end{aligned}$$

Comparing this with our previous expression for $L(E, P_n)$, we see (unsurprisingly) its identical except for a shift of $i \mapsto i + 1$. The law of exponents turns this additive shift into a multiplicative one:

$$\begin{aligned}
U(E, P_n) &= \sum_{0 \le i < n} E(a + (i+1)\Delta_n)\Delta_n \\
&= \sum_{0 \le i < n} E(\Delta_n)E(a + i\Delta_n)\Delta_n \\
&= E(\Delta_n) \sum_{0 \le i < n} E(a + i\Delta_n)\Delta_n \\
&= E(\Delta_n)L(E, P_n)
\end{aligned}$$

Thus, $U(E, P_n) = E(\Delta_n)L(E, P_n)$ for every n. Since E is continuous,

$$\lim E(\Delta_n) = E(\lim \Delta_n) = E(0) = 1$$

And, as $L(E, P_n)$ converges (as we proved above) we can apply the limit theorem for products to get

$$\begin{aligned}
\lim U(E, P_n) &= \lim(E(\Delta_n)L(E, P_n)) \\
&= (\lim E(\Delta_n))(\lim L(E, P_n)) \\
&= \lim L(E, P_n) \\
&= \frac{E(b) - E(a)}{E'(0)}
\end{aligned}$$

Thus, the limits of our sequence of upper and lower bounds are equal! And, by the argument at the beginning of this proof, that squeezes $L(E)$ and $U(E)$ to be equal as well. Thus, E is integrable on $[a, b]$ and its value is what we have squeezed:

$$\int_{[a,b]} E = \frac{E(b) - E(a)}{E'(0)}$$

□

Corollary 30.1. *On any interval $[a, b]$ the natural exponential is integrable, and*

$$\int_{[a,b]} \exp\, dx = \exp(b) - \exp(a)$$

30.3. A Logarithm

Proposition 30.5. *Let $a < b$ be positive numbers. Then the function $f(x) = 1/x$ is integrable on the interval $[a, b]$.*

Proof. Here we attempt to prove integrability *without* necessarily computing the value of the function at the same time. So, its enough to use the ϵ-integrability criterion, where we show that for any $\epsilon > 0$ there exists some partition P where $U(1/x, P) - L(1/x, P) < \epsilon$.

Note that $1/x$ is monotone decreasing on the positive reals, so for any sub-interval $[t_{i-1}, t_i]$ of any partition, we have

$$m_i = \frac{1}{t_i} \qquad M_i = \frac{1}{t_{i-1}}$$

If P is an evenly spaced partition of $[a, b]$ with $|P_i| = \Delta$ for some $\Delta > 0$ this lets us express the difference $U - L$ as a telescoping sum:

$$\begin{aligned} U - L &= \sum_{1\leq i\leq N} M_i\Delta - \sum_{1\leq i\leq N} m_i\Delta \\ &= \Delta \sum_{1\leq i\leq N} (M_i - m_i) \\ &= \Delta \sum_{1\leq i\leq N} \frac{1}{t_{i-1}} - \frac{1}{t_i} \\ &= \Delta\left(\left(\frac{1}{t_0} - \frac{1}{t_1}\right) + \left(\frac{1}{t_1} - \frac{1}{t_2}\right) + \cdots + \left(\frac{1}{t_{N-1}} - \frac{1}{t_N}\right)\right) \\ &= \Delta\left(\frac{1}{t_0} - \frac{1}{t_N}\right) \\ &= \Delta\left(\frac{1}{a} - \frac{1}{b}\right) \end{aligned}$$

Write $L = \frac{1}{a} - \frac{1}{b}$ for this constant value. Then to make the difference between upper and lower sums less than ϵ all we need is to set $\Delta < \epsilon/L$. □

Proposition 30.6. *For any positive $k \in \mathbb{R}$, and any interval $[a,b] \subset (0,\infty)$*

$$\int_{[a,b]} \frac{1}{x} = \int_{[ka,kb]} \frac{1}{x}$$

Proof. For any partition P of $[a,b]$ and number k let kP be the partition of $[ka,kb]$ resulting from multiplying all points by k. This assignment determines a bijection between the sets of partitions of $[a,b]$ and the partitions of $[ka,kb]$.

Because we already know $f(x) = 1/x$ to be integrable on both intervals, we may choose to work with just lower sums without loss of generality. We aim to show that for every $P \in \mathcal{P}_{[a,b]}$

$$L_{[a,b]}\left(\tfrac{1}{x}, P\right) = L_{[ka,kb]}\left(\tfrac{1}{x}, kP\right)$$

Assuming we have this, since $P \mapsto kP$ is a bijection $\mathcal{P}_{[a,b]} \cong \mathcal{P}_{[ka,kb]}$, this implies the *sets* of all possible lower sums are equal:

$$\left\{L_{[a,b]}\left(\tfrac{1}{x}, P\right) \;:\; P \in \mathcal{P}_{[a,b]}\right\} = \left\{L_{[ka,kb]}\left(\tfrac{1}{x}, P\right) \;:\; P \in \mathcal{P}_{[ka,kb]}\right\}$$

Thus as the sets are equal, their suprema are equal, which are by definition the lower integrals $L_{[a,b]}\frac{1}{x} = L_{[ka,kb]}\frac{1}{x}$. But, as we already know this function is integrable on each of these intervals, these values are just the integrals themselves, so we are done. Thus, it only remains to prove equality of the upper sums for partitions in bijective correspondence. □

Exercise 30.3. Let P be an arbitrary partition of $[a,b]$. Prove that

$$U_{[a,b]}\left(\tfrac{1}{x}, P\right) = U_{[ka,kb]}\left(\tfrac{1}{x}, kP\right)$$

Hint:$1/x$ *is monotone decreasing, so we know its infimum on each interval is the right endpoint*

Theorem 30.1. *The function $L(x) = \int_{[1,x]} \frac{1}{t}$ is a logarithm.*

Proof 1. For any $x, y \in (1,\infty)$ we directly compute using the above lemma. The idea of the proof is immediate in the first case, where we consider $x, y > 1$:

$$\begin{aligned} L(xy) &= \int_{[1,xy]} \frac{1}{t} \\ &= \int_{[1,x]} \frac{1}{t} + \int_{[x,xy]} \frac{1}{t} \\ &= \int_{[1,x]} \frac{1}{t} + \int_{[1,y]} \frac{1}{t} \\ &= L(x) + L(y) \end{aligned}$$

□

This function extends to all of $(0, \infty)$, if we use the definition of the integral allowing *oriented intervals* (Remark 28.4), as you can check in the exercise below.

Exercise 30.4. What are the other cases? Prove them by similarly breaking into sub-intervals and rescaling (Proposition 30.6).

30.4. Using the Freedom of Partition

We can use the freedom of choice of partition to our advantage even more, in calculating more difficult integrals, such as below:‘

Exercise 30.5. Follow the argument structure of Example 30.1 in the simpler case below to show if $a, b, c \in \mathbb{R}$ and f is an integrable function on the interval $[a+c, b+c]$, then $f(x+c)$ is integrable on $[a, b]$ and

$$\int_{[a+c,b+c]} f(x) = \int_{[a,b]} f(x+c)$$

Exercise 30.6. Let $[a, b]$ be an interval and $k > 0$. Then if $f(x)$ is integrable on $[ka, kb]$, the function $f(kx)$ is integrable on $[a, b]$ and

$$\int_{[ka,kb]} f(x) = k \int_{[a,b]} f(kx)$$

Example 30.1. Let f be an integrable function. Then

$$\int_{[a^2,b^2]} f(x) = 2 \int_{[a,b]} x f(x^2)$$

Proof. We begin with some preliminary calculations involving partitions. Let $P = \{t_i\}$ be a partition of the interval $I = [a, b]$ and S the set of midpoint samples $s_i = (t_i + t_{i-1})/2$ for P. Now consider the *squares* $P^2 = \{t_i^2\}$ and $S^2 = \{s_i^2\}$ of these.

As squaring is monotone, $t_i < t_{i+1}$ implies $t_i^2 < t_{i+1}^2$, so P^2 is still a partition, but now of the interval I^2 from $t_0^2 = a^2$ to $t_n^2 = b^2$. Again by the monotonicity of squaring, since $s_i \in P_i$ for each i, it follows that $s_i^2 \in P_i^2$ so S^2 is a sample set for P^2. The key to our computation is to work out this Riemann sum for f with the partition P^2:

$$\begin{aligned}
\textstyle\sum_{I^2}(f, P^2, S^2) &= \sum_i f(t_i^2)|P_i^2| \\
&= \sum_i f(s_i^2)\left(t_i^2 - t_{i-1}^2\right) \\
&= \sum_i f(s_i^2)(t_i + t_{i-1})(t_i - t_{i-1}) \\
&= \sum_i f(s_i^2)(t_i + t_{i-1})|P_i|
\end{aligned}$$

As the samples occur at interval midpoints, by definition $2s_i = t_i + t_{i-1}$, so the Riemann sum simplifies

$$\begin{aligned}
\sum_i f(s_i^2)(t_i + t_{i-1})|P_i| &= \sum_i f(s_i^2)2s_i|P_i| \\
&= 2\sum_i s_i f(s_i^2)|P_i|
\end{aligned}$$

But this is precisely the Riemann sum for $xf(x^2)$ on I, using the partition P! Thus we've shown for *any* partition P, and midpoint samples S

$$\textstyle\sum_{I^2}(f(x), P^2, S^2) = 2\sum_I(xf(x^2), P, S)$$

Finally, we can begin the computation. Let P_n be a sequence of shrinking partitions on $[a, b]$ and S_n the corresponding sequences of midpoints. Then the squares P_n^2 form a sequence of shrinking partitions on the interval $[a^2, b^2]$ which we may use to compute the integrals $\int_{[a,b]} xf(x^2)$ and $\int_{[a^2,b^2]} f(x)$ respectively.

$$\begin{aligned}
\int_{[a^2,b^2]} f(x) &= \lim \textstyle\sum_{[a^2,b^2]}(f(x), P_n^2, S_n^2) \\
&= 2\lim \textstyle\sum_{[a,b]}(xf(x^2), P_n, S_n) \\
&= 2\int_{[a,b]} xf(x^2)
\end{aligned}$$

□

Exercise 30.7. Verify the fact used in the proof: if P_n is any sequence of partitions of $[a, b]$ with $\max_W(P_n) \to 0$ then the max-widths of the corresponding sequence P_n^2 on $[a^2, b^2]$ also goes to zero.

31. Theory

Highlights of this Chapter: we prove that the Riemann/Darboux integral satisfies our axioms of integration, and thus is truly an integral. We then use this construction to prove useful facts about the integral: includeing that the integral is a linear map, and power series can be integrated term by term within their radius of convergence.

Remark 31.1. Because the definitions of the Riemann and Darboux integral are equivalent, we are free to use whichever we please in deriving properties of this integral. While each definition has its own strengths, the formulation of Darboux will provide shorter proofs the majority of the time.

31.1. Verifying the Axioms

Theorem 31.1 (The Darboux Integral and Constants)**.** *Let $f(x) = k$ be a constant function, and $[a,b]$ an interval. Then k is Darboux integrable on $[a,b]$ and*

$$\int_{[a,b]} k\,dx = k(b-a)$$

Proof. For any partition P, we have

$$M_i = \sup_{x\in P_i}\{f(x)\} = k = \inf x \in P_i\{f(x)\} = m_i$$

as f is constant. Thus,

$$U(f,P) = \sum_{P_i\in P} M_i|P_i| = k\sum_{P_i\in P}|P_i| = k(b-a)$$

$$L(f,P) = \sum_{P_i\in P} m_i|P_i| = k\sum_{P_i\in P}|P_i| = k(b-a)$$

The upper and lower sums are *constant*, independent of partition, and so their respective infima/suprema are also constant, equal to this same value. Thus k is integrable, and the integral is also this common value

$$\int_{[a,b]} k\,dx = k(b-a)$$

□

Theorem 31.2 (The Darboux Integral and Inequality). *Let f, g be Darboux integrable functions on $[a, b]$ and assume that $f(x) \leq g(x)$ for all $x \in [a, b]$. Then*

$$\int_{[a,b]} f\,dx \leq \int_{[a,b]} g\,dx$$

Proof. The constraint $f \leq g$ implies that on any partition P we have

$$L(f, P) \leq L(g, P)$$

Or, equivalently $L(g, P) - L(f, P) \geq 0$. Taking the supremum over all P of this set of nonnegative numbers yields a nonnegative number, so

$$\sup_{P \in \mathscr{P}_{[a,b]}} \{L(g, P) - L(f, P)\} \geq 0$$

$$L(g) - L(f) \geq 0 \implies L(f) \leq L(g)$$

But since we've assumed f and g are integrable we know that $L(f) = U(f) = \int_{a,b} f\,dx$ and $L(g) = U(g) = \int_{[a,b]} g\,dx$. Thus

$$\int_{[a,b]} f\,dx \leq \int_{[a,b]} g\,dx$$

□

Theorem 31.3 (The Darboux Integral and Subdivision). *Let $[a, b]$ be an interval and $c \in (a, b)$. Then a function f defined on $[a, b]$ is Darboux-integrable on this interval if and only if it is Darboux integrable on both $[a, c]$ and $[c, b]$. Furthermore, when defined these three integrals satisfy the identity*

$$\int_{[a,b]} f\,dx = \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx$$

Proof. First, assume that f is integrable on $[a, b]$. By Theorem 29.1, this means for any $\epsilon > 0$ there exists a partition P where $U(f, P) - L(f, P) < \epsilon$. Now consider the refinement $P_c = P \cup \{c\}$. By the refinement lemma,

$$L(f, P) \leq L(f, P_c) \leq U(f, P_c) \leq U(f, P)$$

Thus $U(f, P_c) - L(f, P_c) < \epsilon$ as well. Next we take this partition and divide it into partitions of each subinterval $P_1 = P_c \cup [a, c]$ and $P_2 = P_c \cup [c, b]$. By simply re-grouping the finite sums, we see

$$L(f, P_c) = L(f, P_1) + L(f, P_2) \qquad U(f, P_c) = U(f, P_1) + U(f, P_2)$$

And, by the definitions of upper and lower sums, for each we know $U(f, P_i) - L(f, P_i) \geq 0$. All that remains to insure the integrability of f on $[a, c]$ and $[c, b]$ is to show that these differences are individually less than ϵ. But this is immediate, as for example,

$$\begin{aligned} U(f, P_1) - L(f, P_1) &\leq U(f, P_1) - L(f, P_1) + (U(f, P_2) - L(f, P_2)) \\ &= (U(f, P_1) + U(f, P_2)) - (L(f, P_1) + L(f, P_2)) \\ &= U(f, P_c) - L(f, P_c) \\ &\leq \epsilon \end{aligned}$$

and the same argument applies to $U(f, P_2) - L(f, P_2)$. □

Next we assume integrability on the two subintervals, and prove integrability on the whole interval.

Proof. Let $\epsilon > 0$ and by our integrability assumptions choose partitions P_1 of $[a, c]$ and P_2 of $[c, b]$ such that

$$U(f, P_i) - L(f, P_i) \leq \frac{\epsilon}{2} \qquad i \in \{1, 2\}$$

Now, their union $P = P_1 \cup P_2$ is a partition of $[a, b]$, and re-grouping the finite sums, we see

$$L(f, P) = L(f, P_1) + L(f, P_2) \qquad U(f, P) = U(f, P_1) + U(f, P_2)$$

Thus,

$$\begin{aligned} U(f, P) - L(f, P) &= (U(f, P_1) + U(f, P_2)) - (L(f, P_1) + L(f, P_2)) \\ &= (U(f, P_1) - L(f, P_1)) + (U(f, P_2) - L(f, P_2)) \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon \end{aligned}$$

□

So, we see that integrability on $[a, b]$ is equivalent to integrability on $[a, c]$ and $[c, b]$. Finally, we need to show in the case where all three integrals are defined, the subdivision equality actually holds.

Proof. Let P be any partition of the interval $[a, b]$ and define the usual suspects:

$$P_c = P \cup \{c\} \qquad P_1 = P_c \cup [a, c] \qquad P_2 = P_c \cup [c, b]$$

We need three pieces of data. First, the inequalities relating integrals to upper and lower sums

$$L(f, P_1) \leq \int_{[a,c]} f\,dx \leq U(f, P_1) \qquad L(f, P_2) \leq \int_{[c,b]} f\,dx \leq U(f, P_2)$$

Second, the inequalities of refinements:

$$L(f, P) \leq L(f, P_c) \leq U(f, P_c) \leq U(f, P)$$

and third, the relationships between P_1, P_2 and P_c:

$$L(f, P_c) = L(f, P_1) + L(f, P_2) \qquad U(f, P_c) = U(f, P_1) + U(f, P_2)$$

Putting all of these together, we get both lower and upper estimates for the sum of the integrals over the subdivision:

$$L(f, P) \leq L(f, P_c) = L(f, P_1) + L(f, P_2) \leq \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx$$

$$\int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx \leq U(f, P_1) + U(f, P_2) = U(f, P_c) \leq U(f, P)$$

And concatenating these inequalities gives the overall bound, for any *arbitrary* partition P:

$$L(f, P) \leq \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx \leq U(f, P)$$

Thus, the sum of these integrals lies between the upper and lower sum of f on $[a, b]$ for *every partition*. As f is integrable, we know there is a *single number with this property*, and that number is by definition the integral. Thus

$$\int_{[a,b]} f\,dx = \int_{[a,c]} f\,dx + \int_{[c,b]} f\,dx$$

□

Phew! We've successfully verified all three axioms for the Darboux integral. Taken together, these prove that our construction really is an integral!

Corollary 31.1. *The equality of upper and lower sums satisfies the axioms of integration, and thus the Darboux Integral really does define an integral.*

31.2. Integrability

Now, we show that our constructed integral is actually *interesting* - that all continuous functions are integrable!

Theorem 31.4 (Continuous $\implies$ Integrable)**.** *Every continuous function on a closed interval is Darboux integrable.*

Proof. Let f be continuous on the interval $[a, b]$ and choose $\epsilon > 0$. We will prove integrability by finding a partition P such that $U(f, P) - L(f, P) < \epsilon$.

As f is continuous it is bounded (by the extreme value theorem), so the upper and lower sums are defined for all partitions. It is also *uniformly continuous* (as $[a, b]$ is a closed interval), so we can find a δ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \frac{\epsilon}{b - a}$$

Now, choose a partition P of $[a, b]$ where the width of each interval is less than δ. Comparing upper and lower sums on this interval,

$$U(f, P) - L(f, P) = \sum_{P_i \in P} M_i |P_i| - \sum_{P_i \in P} m_i |P_i| = \sum_{P_i \in P} [M_i - m_i] |P_i|$$

Since $|P_i| < \delta$, we know that for any $x, y \in P_i$ the values $f(x), f(y)$ differ by less than $\epsilon/(b - a)$. Thus the difference of between the infimum and supremum over this interval must be less than or equal to this bound:

$$M_i - m_i \leq \frac{\epsilon}{b - a}$$

Using this to bound our sum, we see

$$U(f, P) - L(f, P) = \sum_{P_i \in P} [M_i - m_i] |P_i| \leq \frac{\epsilon}{b - a} \sum_{P_i \in P} |P_i|$$

$$= \frac{\epsilon}{b - a}(b - a) = \epsilon$$

Thus, f is integrable! □

But the Darboux integral allows us to integrate even more things than the continuous functions. For example, it is quite straightforward to prove that *all monotone functions are integrable* (even those with many discontinuities!)

Theorem 31.5 (Monotone $\implies$ Integrable)**.** *Every monotone bounded function on a closed interval is integrable.*

Proof. Without loss of generality let f be monotone *increasing* and bounded on the interval $[a, b]$ and choose $\epsilon > 0$. We will prove integrability by finding a partition P such that $U(f, P) - L(f, P) < \epsilon$.

Let $B = f(b) - f(a)$ be the difference between values of f at the endpoints. If $B = 0$ then f is constant, and we already know constant functions are integrable so we are done.

Otherwise, let P be an arbitrary evenly spaced partition of widths $\Delta = \epsilon/B$, we consider the difference $U(f, P) - L(f, P)$:

$$\begin{aligned} U(f,P) - L(f,P) &= \sum_{P_i \in P} M_i |P_i| - \sum_{P_i \in P} m_i |P_i| \\ &= \sum_{P_i \in P} [M_i - m_i] |P_i| = \Delta \sum_{P_i \in P} [M_i - m_i] \end{aligned}$$

Since f is increasing, its supremum on each interval occurs on the right, and its infimum on the left. That is, if $P_i = [t_{i-1}, t_i]$ we have

$$m_i = f(t_{i-1}) \qquad M_i = f(t_i)$$

Plugging this into the above gives a telescoping sum!

$$U(f,P) - L(f,P) = \Delta \sum_{1 \leq i \leq n} [f(t_i) - f(t_{i-1})] = \Delta[f(t_n) - f(t_0)]$$

But $t_0 = a$ and $t_n = b$ are the endpoints of our partition, and so this equals

$$= \Delta[f(b) - f(a)] = \frac{\epsilon}{f(b) - f(a)} [f(b) - f(a)] = \epsilon$$

□

And, inductively its straightforward to show (via subdivision) that if the domain of a function can be partitioned into finitely many intervals on which it is integrable, than its integrable on the entire thing. Thus, for example piecewise continuous functions are Darboux Integrable. The precise statement and theorem is below.s

Definition 31.1 (Piecewise Integrable Function)**.** A function f defined on a domain I is *piecewise integrable* if I is the disjoint of a finite sequence of intervals $I = I_1 \cup I_2 \cup \ldots \cup I_n$, and f restricted to each interval is integrable.

Proposition 31.1 (Piecewise Integrable $\implies$ Integrable)**.** *If f is piecewise integrable, then it is integrable.*

Proof. We begin with the case that f is piecewise integrable on two subintervals, $[a, c]$ and $[c, b]$ of the interval $[a, b]$. Then the subdivision axiom immediately implies that f is in fact integrable on the entire interval.

Now, assume for induction that all functions that are piecewise integrable on intervals with $\leq n$ subdivisions are actually integrable, and let f be a piecewise integrable function on a union of $n + 1$ intervals

$$[a, b] = I_1 \cup I_2 \cup \cdots \cup I_n \cup I_{n+1}$$

Set J equal to the union of the first n, so that $[a, b] = J \cup I_{n+1}$. Then when restricted to J, the function f is piecewise integrable on n intervals, so its integrable by assumption. And so, f is integrable on both J and I_{n+1}, so its piecewise integrable with *two intervals*, and hence integrable as claimed. □

Because all continuous functions and all monotone functions are integrable, we have the following useful corollary covering most functions usually seen in a calculus course.

Corollary 31.2. *All piecewise continuous and piecewise monotone functions with finitely many pieces are integrable.*

But monotone functions are even more general than this! A monotone function can have *countably many discontinuities.* Pursuing this further, there is a precise characterization of Darboux-Integrable functions (which we state, but do not prove).

Theorem 31.6 (Riemann-Lebesgue Integrability Theorem)**.** *A function f on the interval $[a, b]$ is Riemann/Darboux-integrable if and only if it is bounded and its set of discontinuities is measure zero.*

We do not prove this theorem here (its proof is long, and requires a precise definition of the concept of a *measure zero set* to even state) nor do we use it in what follows. But it is interesting to note that *if* one were to prove this theorem, many of the results both above and below follow as rather trivial consequences:

- Continuous functions are bounded (the extreme value theorem) and have empty discontinuity set. Thus they are integrable.
- One can prove that monotone functions can have at most countably many discontinuities, and any countable set has measure zero. Thus monotone functions are integrable.
- Piecewise integrability implies integrability as the overall function is bounded by the max of the bounds on each interval, and the overall discontinuity set is the union of the discontinuity sets (and, the finite union of measure zero sets has measure zero).

- Constant multiples of integrable functions are integrable: multiplying a bounded function by a constant leaves it still bounded, and does not change the discontinuity set.
- Sums of integrable functions are integrable: a sum is bounded by the sum of the bounds for its terms, and its discontinuity set is contained in the union of the discontinuity sets of each term.

31.3. Linearity

Theorem 31.7 (Integrability of Multiples). *Let f be an integrable function a closed interval I, and $c \in \mathbb{R}$. Then the function cf is also integrable on I, and furthermore*

$$\int_I cf\,dx = c\int_I f\,dx$$

Proof. Let P_n be an arbitrary sequence of shrinking partitions (of length N_n), and S_n an arbitrary sequence of sample points for P_n. We attempt to evaluate the limit

$$\lim \textstyle\sum_I(cf, P_n, S_n)$$

For any partition P and sample set S, $\sum_I(cf, P, S)$ is a finite sum, and so we may factor out the constant c:

$$\begin{aligned}\textstyle\sum_I(cf, P, S) &= \sum_{0\leq k<N} cf(s_k)|P_k| \\ &= c\sum_{0\leq k<N} f(s_k)|P_k| \\ &= c\textstyle\sum_I(f, P, S)\end{aligned}$$

Doing this for each n yields

$$\lim \textstyle\sum_I(cf, P_n, S_n) = \lim c\sum_I(f, P_n, S_n)$$

Since f is assumed integrable on I we know that $\lim \sum_I(f, P_n, S_n) = \int_I f\,dx$ and so we may use the limit theorems to pull the constant c out:

$$\lim c\textstyle\sum_I(f, P_n, S_n) = c\lim \sum_I(f, P_n, S_n) = c\displaystyle\int_I f\,dx$$

Thus, $\lim \sum_I(cf, P_n, S_n)$ exists, and as P_n, S_n were arbitrary, its value is independent of the particular choice of shrinking partitions. By definition this means that cf is integrable on I, and that

$$\int_I cf\,dx = c\int_I f\,dx$$

□

Exercise 31.1. For practice, provide a proof of this using the Darboux integrability criterion, instead of the definition of the Riemann integral.

We proceed with the same strategy to prove the integrability of a sum of integrable functions: using the limit laws and definitions for finite sums to calculate along the way:

Theorem 31.8 (Integrability of Sums)**.** *Let f, g be integrable functions on a closed interval I. Then their sum $f + g$ is also integrable on I. Furthermore, its integral is the sum of the integrals of f and g:*

$$\int_I (f+g) = \int_I f + \int_I g$$

Proof. Let P_n be an arbitrary sequence of shrinking partitions, and for each n, let S_n be a sample set for P_n. We attempt to evaluate the limit

$$\lim \sum_I(f+g, P_n, S_n)$$

For any partition P and sample S, $\sum_I(f+g, P, S)$ is a finite sum, and so we can re-order its terms by the commutativity of addition, decomposing into two Riemann sums

$$\begin{aligned}
\sum_I(f+g, P, S) &= \sum_{0\le k<N} [f(s_k)+g(s_k)]\,|P_k| \\
&= \sum_{0\le k<N} f(s_k)\,|P_k| + \sum_{0\le k<N} g(s_k)\,|P_k| \\
&= \sum_I(f, P, S) + \sum_I(g, P, S)
\end{aligned}$$

Doing this for each n yields

$$\lim \sum_I(f+g, P_n, S_n) = \lim \left[\sum_I(f, P_n, S_n) + \sum_I(g, P_n, S_n)\right]$$

By hypothesis, both f and g are integrable on I, meaning that

$$\lim \sum_I(f, P_n, S_n) = \int_I f\,dx \qquad \lim \sum_I(g, P_n, S_n) = \int_I g\,dx$$

Since both of these limits exist, we can use the limit law for sums to distribute the limit above, and see

$$\begin{aligned}\lim \textstyle\sum_I(f+g, P_n, S_n) &= \lim\left[\textstyle\sum_I(f, P_n, S_n) + \textstyle\sum_I(g, P_n, S_n)\right] \\ &= \lim \textstyle\sum_I(f, P_n, S_n) + \lim \textstyle\sum_I(g, P_n, S_n) \\ &= \int_I f\,dx + \int_I g\,dx\end{aligned}$$

Since P_n and S_n were arbitrary, this same result must hold for all such shrinking partitions - all such limits converge, and have the same value $\int_I f\,dx + \int_I g\,dx$. Thus, by definition $f+g$ is integrable, and

$$\int_I f+g\,dx = \int_I f\,dx + \int_I g\,dx$$

□

Exercise 31.2. For practice, provide a proof of this using the Darboux integrability criterion, instead of the definition of the Riemann integral.

Each of these theorems does two things: it proves something about the *space of integrable functions* and also about *how the integral behaves* on this space. Below we rephrase the conclusion of these theorems in the terminology of linear algebra - a result so important it deserves the moniker of "Theorem" itself.

Theorem 31.9 (Linearity of the Riemann/Darboux Integral)**.** *For each interval* $[a,b] \subset \mathbb{R}$*, the set* $\mathcal{I}([a,b])$ *of Riemann integrable functions forms a* Vector Subspace *of the set of all functions* $[a,b] \to \mathbb{R}$*. On this subspace, the Riemann integral defines a linear map*

$$\int_{[a,b]} : \mathcal{I}([a,b]) \to \mathbb{R}$$

31.4. Power Series

We now turn to the issue of integrating power series. The theoretical results are in close analogy to the differentiation case: in summary,

- A power series is integrable on the interior of its radius of convergence, and the the antiderivative converges on the same interval
- The antiderivative can be found by antidifferentiating term-by-term.

There are two ways we could try to prove a theorem such as this: one, we could try to mimic the style of the differentiation proof, developing a theory of *dominated convergence* for the Riemann integral. This succeeds without issue, and is carried out in the following section. But alternatively we could attempt to *use the fundamental theorem of calculus* to relate this directly to what we already know about differentiation. This turns out to be a shorter and more elementary argument, and while less general (it applies only to power series, not to general series of functions) it is more than sufficient for our purposes, so we give it here.

Theorem 31.10. *Let $f(x) = \sum_{n\geq 0} a_n x^n$ be a power series with radius of convergence R. Then the power series $F(x) = \sum_{n\geq 0} \frac{a_n}{n+1}x^{n+1}$ has the same radius of convergence, and*

$$F(x) = \int_{[a,x]} f\,dx$$

We will prove this theorem as a sequence of propositions. First, we check that the radius of convergence remains unchanged under term-by-term integration of a series.

Proof. Like for the differentiable case, we prove this here under the assumption that the Ratio test succeeds in computing the radius of convergence for the original series, so for any $x \in (-R, R)$

$$\lim \left|\frac{a_{n+1}}{a_n}\right| |x| < 1$$

We now turn to compute the ratio test for our new series $\sum \frac{a_n}{n+1} x^{n+1}$: the ratio in question is

$$\frac{\frac{a_{n+1}}{n+2}x^{n+2}}{\frac{a_n}{n+1}x^{n+1}} = \left(\frac{a_{n+1}}{a_n}\right)\left(\frac{n+1}{n+2}\right)x$$

Since $(n+1)/(n+2) \to 1$ we can compute the overall limit using the limit theorems and see we end up with *the exact same limit as for the original series*! Thus integrating term by term does not change the radius of convergence at all. □

Now, we turn to the main event: we prove that the term-by-term antiderivative is the integral of the original power series.

Proof. Let $f(x) = \sum_{n\geq 0} a_n x^n$ have radius of convergence R, and let $F(x) = \sum_{n\geq 0} \frac{a_n}{n+1}x^{n+1}$. Choose any $x \in (0, R)$; we wish to show that

$$F(x) = \int_{[0,x]} f\,dx$$

Because term-by-term-differentiation holds within the radius of convergence, for any $|x| < R$ we have

$$\begin{aligned} F'(x) &= \left(\sum_{n\geq 0} \frac{a_n}{n+1} x^{n+1}\right)' \\ &= \sum_{n\geq 0} \left(\frac{a_n}{n+1} x^{n+1}\right)' \\ &= \sum_{n\geq 0} \frac{a_n}{n+1}(n+1)x^n \\ &= \sum_{n\geq 0} a_n x^n \\ &= f(x) \end{aligned}$$

Thus, F is an antiderivative of f! So by the fundamental theorem of calculus, we know that we can use F to evaluate the integral:

$$\int_{[0,x]} f\,dx = F(x) - F(0)$$

But $F(0) = 0$ as F has no constant term! Thus we have it,

$$\int_{[0,x]} \sum_{n\geq 0} a_n x^n\,dx = \sum_{n\geq 0} \frac{a_n}{n+1} x^{n+1}$$

□

31.5. ⋆ Dominated Convergence

Now we turn to the development of a more general theory, based on a new Dominated Convergence Theorem for the Riemann integral.

Theorem 31.11 (Dominated Convergence for the Riemann/Darboux Integral)**.** *Let $\{f_n\}$ be a sequence of Riemann integrable functions on a closed interval I, and assume that the functions f_n converge pointwise to a Riemann integrable function f. Then if there exists some M where $|f_n(x)| < M$ for all $x \in I$, the order of integration and limit may be interchanged:*

$$\lim \int_I f_n = \int_I f$$

Proof. For a short elementary proof, see https://arxiv.org/pdf/1408.1439.pdf. □

Looking back at the original Dominated Convergence for sums, there are some slight differences here: we have to *assume* the integrability of the limit, and we must give a *uniform bound* on all the f_n. However, if we are mindful of these differences the application of the theorem is analogous - we use it to switch a sum and integral, and then integrate term by term.

Remark 31.2. One motivating reason to seek an alternative theory of integration in advanced analysis is to find an integral with a dominated convergence theorem closer to the others we've met. Such an integral exists, and was first constructed by Henri Lebesgue in 1905 (but will not concern us here; dominated convergence for the Riemann integral is plenty powerful!)

Again, both series and integrals are defined by limit statements, so this proof is going to require an interchange of limits. Instead of digging all the way down to the foundations and using the Riemann sum definitions to apply Dominated convergence for series (Theorem 21.1), we can save some trouble by directly using Dominated convergence for integrals (Theorem 31.11)

Just as before we first show how the proof goes *assuming* that we can exchange the sum and integral limits, and *then* we'll justify that this is allowed:

Theorem 31.12 (Integration of Power Series). *Let $f = \sum_{k\geq 0} a_k x^k$ be a power series with radius of convergence R. Then for $x \in (-R, R)$:*

$$\int_{[0,x]} f = \sum_{k\geq 0} \frac{a_k}{k+1} x^{k+1}$$

Proof. Let f_N denote the N^{th} partial sum of the series, $f_N = \sum_{k=0}^{N} a_k x^k$, so $f = \lim_N f_N$. Substituting this into the above,

$$\int_{[0,x]} f = \int_{[0,x]} \lim_N f_N$$

Now **assuming** that dominated convergence for integrals applies, we may switch the integral and limit statement, to get

$$\int_{[0,x]} \lim_N f_N = \lim_N \int_{[0,x]} f_N$$

Now, each f_N is a polynomial - meaning its a *finite* sum! This means we can integrate it term by term using the linearity of the integral (Theorem 31.8):

$$\int_{[0,x]} f_N = \int_{[0,x]} \sum_{k=0}^{N} a_k t^k$$
$$= \sum_{k=0}^{N} a_k \int_{[0,x]} t^k$$
$$= \sum_{k=0}^{N} a_k \frac{x^{k+1}}{k+1}$$

Now, taking the limit $N \to \infty$ gives the series of term by term antiderivatives:

$$\int_{[0,x]} f = \lim_N \sum_{k=0}^{N} a_k \frac{x^{k+1}}{k+1} = \sum_{k\geq 0} \frac{a_k}{k+1} x^{k+1}$$

□

Now, we need to justify that dominated convergence applies. Theorem 31.11 requires two things: (1) that the limit $\lim f_N = f$ is integrable on $[0, x]$, and (2) that each of the functions f_N is uniformly bounded by some constant M on the interval $[0, x]$.

Proposition 31.2. *If f is a power series and x is within the radius of convergence, then f is integrable on $[0, x]$.*

Proof. If $x \in (-R, R)$ then the closed interval $[0, x]$ is completely contained within the interval of convergence. Because a power series is continuous at each point on the interior of its interval of convergence (Theorem 21.4), it is continuous on the closed interval $[0, x]$.
And, as continuous functions on a closed interval are integrable (Theorem 31.4), it is integrable on $[0, x]$ as required. □

The second requirement requires us to dig into the definition of a power series a bit.

Proposition 31.3. *Let f be a power series with radius of convergence R, and f_N be its sequence of partial sums. Then if $x \in (0, R)$, there is a fixed constant M such that*

$$|f_N(t)| \leq M \;\forall N \;\forall t \in [0, x]$$

Proof. As $0 < x < R$ the interval $[0, x]$ is contained in the interior of the interval of convergence, so the power series $f = \sum_{k\geq 0} a_k t^k$ is absolutely convergent for each $t \in [0, x]$. Let g denote the series of term-wise absolute values $g(t) = \sum_{k\geq 0} |a_k| t^k$, and

g_N denote its sequence of partial sums. Then, by the triangle inequality for finite sums, for every $t \in [0, x]$,

$$|f_N(t)| = \left|\sum_{k=0}^{N} a_k t^k\right| \leq \sum_{k=0}^{N} |a_k| t^k = g_N(t)$$

And, since all the terms of g are positive, the sequence $g_N(t)$ is monotone increasing in N, with

$$g_N(t) \leq g(t) \ \forall N$$

Stringing these two inequalities together, we see that for each $t \in [0, x]$, the quantity $g(t)$ is an upper bound for $\{f_N(t)\}$.

But g itself is a power series (with coefficients $|a_k|$) and is convergent for all $t \in [0, x]$ (as f is absolutely convergent at all points on the interior of its radius of convergence). Thus by Theorem 21.4, g is continuous on $[0, x]$. That means we can apply the extreme value theorem (Theorem 16.1) to find an absolute maximum of g on $[0, x]$: a value M such that $g(t) \leq M$ for all $t \in [0, x]$.

Now truly stringing it all together, we see that for each $t \in [0, x]$ and each $N \in \mathbb{N}$,

$$|f_N(t)| \leq g_N(t) \leq g(t) \leq M$$

Thus M is the uniform bound we seek. □

Remark 31.3. Note we could get by here without invoking the extreme value theorem,but rather just to see $\{g(t) \mid t \in [0, x]\}$ is a bounded set, and select M to be any upper bound. We chose the (stronger) extreme value theorem only because it is more memorable.

Exercise 31.3. Use the argument above to show that this holds for any $x \in (-R, R)$; the assumption on positivity is not required.

32. π

Highlights of this Chapter: we prove that π - defined in our final project as the first zero of the sine function - is the area of the unit circle. We then look at several means of approximating the value of π; from computing Riemann sums to integrating power series. In the end, we derive a relatively efficient means of calculating π, which gets 15 digits after adding only 22 terms.

32.1. π and the Circle

In the second project, we have defined π as the first zero of the sine function - a definition, and as a final computation in this class, we will show that this is equal to the geometric definition - the area of a circle!

This provides a relationship between the modern, rigorous theory of trigonometric functions and the ancient quest of Archimedes to measure the area of the circle.

Indeed, since we have defined area rigorously with integration, we can now make sense of *the area of the circle* as long as we can express the unit circle as a function. While this is not directly possible, we can take the implicit equation $x^2 + y^2 = 1$ and solve for y giving *two* functions (one for the top half and one for the bottom). Then we can measure the area of the circle as twice the top half, or

$$\text{Area} = 2\int_{[-1,1]} \sqrt{1-x^2}$$

Now we compute this integral with our newfound integration techniques (substitution), and show it equals the half-period of our trigonometric functions in natural units.

Theorem 32.1.

$$2\int_{[-1,1]} \sqrt{1-x^2} = \pi$$

Proof. By subsitution, we see that the following two integrals are equal

$$\int_{[0,1]} \sqrt{1-x^2} = \int_I \sqrt{1-(\sin(t))^2}(\sin(t))'$$

Where $I = [a,b]$ is the interval such that $[\sin(a), \sin(b)] = [0,1]$. Since $\sin(0) = 0$ and $\sin(\pi/2) = 1$ we see $I = [0, \pi/2]$. Now we focus on simplifying the integrand:

By the Pythagorean identity, $1 - \sin^2(t) = \cos^2(t)$, thus by Example 4.3,

$$\sqrt{1-\sin^2(t)} = \sqrt{\cos^2(t)} = |\cos(t)|$$

and by definition we recall $(\sin t)' = \cos t$. Thus

$$\begin{aligned} \int_{[0,\pi/2]} &= \int_{[0,\pi/2]} |\cos(t)|\cos(t) \\ &= \int_{[0,\pi/2]} \cos^2(t) \end{aligned}$$

Where we can drop the absolute value as cos is nonnegative on $[0, \pi/2]$ (its first zero is at half the period, so π). We can simplify this using the "half angle formula" $\cos^2(x) = (1+\cos(2x))/2$

$$\int_{[0,\pi/2]} \cos^2(t) = \int_{[0,\pi/2]} \frac{1+\cos(2t)}{2}$$

Using the linearity of the integral, this reduces to

$$\begin{aligned} \int_{[0,\pi/2]} \cos^2(t) &= \frac{1}{2}\int_{[0,\pi/2]} 1 + \frac{1}{2}\int_{[0,\pi/2]} \cos(2t) \\ &= \frac{\pi}{4} + \frac{1}{2}\int_{[0,\pi/2]} \cos(2t) \end{aligned}$$

The first of these integrals could be immediately evaluated as the integral of a constant, but the second requires us to do another substitution. If $u = 2t$ then

$$\int_{[0,\pi/2]} \cos(2t) = \frac{1}{2}\int_{[0,\pi]} \cos u$$

We recall again that by definition $\cos u = (\sin u)'$, so by the first fundamental theorem

$$\int_{[0,\pi]} \cos u = \int_{[0,\pi]} (\sin u)' = \sin u\Big|_{[0,\pi/]}$$

But, sin is equal to 0 both at 0 and π! So after all this work, this integral evaluates to zero. Thus

$$\begin{aligned}
\int_{[0,1]} \sqrt{1-x^2} &= \int_{[0,\pi/2]} \cos^2 t \\
&= \frac{\pi}{4} + \frac{1}{2}\int_{[0,\pi]} \cos(2t) \\
&= \frac{\pi}{4} + 0
\end{aligned}$$

Now, we are ready to assemble the pieces. Because x^2 is an even function so is $\sqrt{1-x^2}$, and so its integral over $[-1,1]$ is twice its integral over $[0,1]$. Thus

$$\text{Area} = 2\int_{[-1,1]} \sqrt{1-x^2} = 4\int_{[0,1]} \sqrt{1-x^2} = 4\frac{\pi}{4} = \pi$$

□

This single result ties together so many branches of analysis, and proves a worthy capstone calculation for the course. However after all this work we shouldn't let ourselves be satisfied too quickly! Now that we have related the area of a circle to *trigonometry*, we can hope to use other techniques from analysis to accurately calculate its value.

32.2. Integrals and Inverse Trigonometry

Since π is defined as the *half period* of the trigonometric functions, we can not hope to get its value as the *output* of a trigonometric calculation (rather, it lies in the *inputs*). This signals that it may prove useful to investigate the *inverse trigonometric functions*. For example, because $\sin(\pi/2) = 1$ we expect $\arcsin(1) = \pi/2$, and finding a way to numerically approximate $\arcsin(x)$ at $x = 1$ would yield the value we seek. While we can do this, the sine turns out not to be the best trigonometric function for this purpose, and it is much more productive to investigate the tangent. For practice both are computed below, but feel free to skip the first.

32.2.1. ⋆ The ArcSine

Thus, we find ourselves interested in calculating these functions. Inspired by our previous treatment of logarithms (where we were able to find the derivative of $L(x)$ using that it was the inverse of an exponential, without actually knowing a formula for L) we seek to begin our study of inverse trigonometry via differentiation:

Proposition 32.1. *The derivative of the inverse sine function is*

$$(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}$$

Proof. Let $f(x) = \arcsin(x)$. Then where defined, $f(\sin(\theta)) = \theta$ by definition, and we may differentiate via the chain rule: on the left side

$$\frac{d}{d\theta} f(\sin(\theta)) = f'(\sin(\theta))\cos(\theta)$$

and on the right $\frac{d}{d\theta}\theta = 1$. Equating these and solving for f' yields

$$f'(\sin(\theta)) = \frac{1}{\cos(\theta)}$$

The only remaining problem is that we want to know f' as a function of x and we only know its value implicitly, as a function of $\sin(\theta)$. But setting $x = \sin\theta$ we can express $\cos\theta = \sqrt{1-x^2}$ via the pythagorean identity $\sin^2\theta + \cos^2\theta = 1$. Thus

$$f'(x) = \frac{1}{\sqrt{1-x^2}}$$

□

Before integration this would have been a mere curiosity. But, armed wtih the fundamental theorem this is an extremely powerful fact: indeed, it directly gives us a representation as an integral:

Corollary 32.1. *The inverse sine function is defined on the interval* $[0, 1]$ *by the integral*

$$\arcsin(x) = \int_{[0,x]} \frac{1}{\sqrt{1-x^2}}\, dx$$

Proof. Since $(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}$, the inverse sine is an antiderivative of $\frac{1}{\sqrt{1-x^2}}$, and also $\sin(0) = 0$ implies $\arcsin(0) = 0$, so it is zero at $x = 0$. Thus, it is exactly the area function

$$\arcsin(x) = \int_{[0,x]} \frac{1}{\sqrt{1-t^2}}\, dt$$

□

One may use this to get another integral representation of π. Perhaps the most natural thought is to use that $\sin(\pi/2) = 1$, and attempt to claim

$$\frac{\pi}{2} = \arcsin(1) \stackrel{?}{=} \int_{[0,1]} \frac{1}{\sqrt{1-x^2}}\,dx$$

Where there is a question over the equals here to signify we do not actually have the tools to conclude this: the integrand is not defined at $x = 1$, and even though it is continuous on $[0, 1)$ it is *unbounded* on that interval! With more work one can overcome these obstacles (the equality is true) but we are already uninterested

Remark 32.1. If we were not bothered by the square roots for our computation-focused goals, one could easily replace the problematic integral above with something avoiding its problems. For instance, since $\sin(\pi/4) = 1/\sqrt{2}$, we have

$$\frac{\pi}{4} = \int_{[0,1/\sqrt{2}]} \frac{1}{\sqrt{1-x^2}}\,dx$$

But this is much worse in terms of square roots: if you write out a Riemann sum here it'll be a sum of nested roots.

The same trouble plagues the cosine function, so we turn their ratio - the tangent - to seek something better for computational purposes. In fact, things work out much nicer for the arctangent.

32.2.2. The ArcTangent

Proposition 32.2.

$$(\arctan x)' = \frac{1}{1+x^2}$$

Proof. We again proceed by differentiating the identity $\arctan(\tan\theta) = \theta$. This yields $\arctan'(\tan\theta)\frac{1}{\cos^2\theta} = 1$ and multiplying through by $\cos^2$ we can solve for the derivative of arctangent:

$$\arctan'(\tan\theta) = \cos^2\theta$$

The only problem is again we have the derivative as a function implicitly of of $\tan\theta$, and we need it in terms of just an abstract variable x. Setting $x = \tan\theta$ we see that $x^2 = \tan^2\theta$ and (using the pythagorean identity) $x^2 + 1 = \tan^2\theta + 1 = \frac{1}{\cos^2\theta}$. Thus

$$\cos^2\theta = \frac{1}{1+x^2}$$

and putting these two together, we reach what we are after

$$\arctan'(x) = \frac{1}{1+x^2}$$

□

Proposition 32.3. *The inverse function* $\arctan(x)$ *to the tangent* $\tan(x) = \sin(x)/\cos(x)$ *admits an integral representation*

$$\arctan(x) = \int_{[0,x]} \frac{1}{1+t^2}$$

Proof. This follows as $\arctan'(x) = 1/(1+x^2)$, so both arctan and this integral have the same derivative. As antiderivatives of the same function this means that they differ by a constant. Finally, this constant is equal to zero as $\arctan(0) = 0$ and $\int_{[0,0]} \frac{1}{1+x^2}\,dx = 0$ as it is an integral over a degenerate interval. □

This gives us a much better opportunity to get an explicit formula for π. We know that sin and cos are equal when evaluated at $\pi/4$, which means their ratio is $1 = \tan \pi/4$. Inverting this,

Corollary 32.2.

$$\frac{\pi}{4} = \arctan(1) = \int_{[0,1]} \frac{1}{1+x^2}, dx$$

This function is integrable (its continuous), so we can compute its value as the limit of any shrinking sequence of Riemann sums. Below is an explicit example, given for evenly spaced partitions sampled at their right endpoints.

Example 32.1. The following infinite series converges to π:

$$\pi = \lim_n 4 \sum_{i=1}^{n} \frac{1}{1+(i\Delta)^2}\Delta$$

$$= \lim_n 4 \sum_{i=1}^{n} \frac{n}{n^2+i^2}$$

$$4\sum_{i=1}^{10} \frac{10}{100+i^2} \approx 3.0395\ldots$$

$$4\sum_{i=1}^{100} \frac{100}{10000+i^2} \approx 3.13155\ldots$$

$$4\sum_{i=1}^{1000} \frac{1000}{1000000+i^2} \approx 3.140592\ldots$$

$$4\sum_{i=1}^{1000000}\frac{1000000}{(1000000)^2+i^2}\approx 3.14159165359\ldots$$

This is great - these sums are trivial to do on a computer (I did these in a simple python for loop) and get us an accurate value for π. But we shouldn't be satisfied just yet! First of all, these sums take a while to converge - we need a thousand terms to get the first two digits after the decimal, and a million to get the first five!

Luckily, the theory we have developed over the semester allows us to do better.

32.3. Series with ArcTan

Instead of trying to evaluate the arctangent integral representation via a Riemann sum, we could attempt to find a power series representation. Like the exponential, we *could* find such a series via Taylor's formula, and prove convergence with the Taylor Error formula. But here there is an easier way!

Recall the geometric series

$$\frac{1}{1-x}=\sum_{n\geq 0}x^n$$

We can substitute $-x^2$ for the variable here to get a series for $1/(1+x^2)$:

$$\frac{1}{1+x^2}=\sum_{n\geq 0}(-x^2)^n=\sum_{n\geq 0}(-1)^n x^{2n}$$

$$=1-x^2+x^4-x^6+x^8-\cdots$$

This power series has radius of convergence 1 (inherited from the original geometric series) and converges at neither endpoint. We know from the above that this function is the derivative of the arctangent, so we should integrate it!

$$\arctan(x)=\int_{[0,x]}\frac{1}{1+t^2}\,dt=\int_{[0,x]}\sum_{n\geq 0}(-1)^n t^{2n}\,dt$$

Inside its radius of convergence we can exchange the order of the sum and the integral:

$$\begin{aligned}
\int_{[0,x]} \left(\sum_{n\geq 0} (-1)^n t^{2n} \right) dt &= \sum_{n\geq 0} \int_{[0,x]} (-1)^n t^{2n}\, dt \\
&= \sum_{n\geq 0} (-1)^n \int_{[0,x]} t^{2n} dt \\
&= \sum_{n\geq 0} (-1)^n \frac{x^{2n+1}}{2n+1}
\end{aligned}$$

Theorem 32.2. *For $x \in (0,1)$,*

$$\arctan(x) = \sum_{n\geq 0} (-1)^n \frac{x^{2n+1}}{2n+1}$$

$$= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \cdots$$

After integrating the series, the result has the same radius of convergence, but now converges at the endpoint $x = 1$ by the Alternating Series Test. Hence, its tempting to write

$$\frac{\pi}{4} = \arctan(1) \stackrel{?}{=} \sum_{n\geq 0} \frac{(-1)^n}{2n+1}$$

But again there's a pesky question mark over the equals signaling that our theory isn't actually strong enough to conclude this! We only know that we can switch the sum and integral *inside* the radius of convergence. At the boundary this doesn't even make sense, as the original series *diverges there*!

Remark 32.2. It turns out that it *is* true - a theorem of Abel guarantees that if a power series converges at a boundary point to its radius of convergence, then it is continuous there. Together with dominated convergence, this lets one compute as

$$\begin{aligned}
\arctan(1) &= \lim_{x\to 1} \arctan(x) \\
&= \lim_{x\to 1} \sum_{n\geq 0} \frac{(-1)^n x^{2n+1}}{2n+1} \\
&= \sum_{n\geq 0} \lim_{x\to 1} \frac{(-1)^n x^{2n+1}}{2n+1} \\
&= \sum_{n\geq 0} \frac{(-1)^n}{2n+1}
\end{aligned}$$

However, way out here at the endpoint the series converges *very slowly*. Using a computer to do a little experimenting:

$$4\sum_{n=0}^{10}\frac{(-1)^n}{2n+1} = 3.2323\ldots$$

$$4\sum_{n=0}^{100}\frac{(-1)^n}{2n+1} = 3.1549\ldots$$

$$4\sum_{n=0}^{1{,}000}\frac{(-1)^n}{2n+1} = 3.1425\ldots$$

Like the Riemann sum approach, we needed a thousand terms to get the first two decimals right.

Instead, like we did for the sine function, we need to seek a different point to evaluate arctan (here, inside the radius of convergence). This will make sure our argument is rigorous, and have the added benefit of being much more efficient than the series in the remark above (plugging in any $x < 1$ will have the series converging *geometrically* - that is, exponentially fast!)

How do we find such a value? Here's one clever possibility: we actually realize $\pi/4$ as the *sum* of two different arctangent values:

Proposition 32.4.

$$\frac{\pi}{4} = \arctan\left(\frac{1}{2}\right) + \arctan\left(\frac{1}{3}\right)$$

Proof. Let $\theta = \arctan(1/2)$ and $\psi = \arctan(1/3)$. Now use the tangent addition law $\tan(\theta+\psi) = \frac{\tan\theta+\tan\psi}{1-\tan\theta\tan\psi}$ to compute $\theta + \psi$:

$$\tan(\theta+\psi) = \frac{\frac{1}{2}+\frac{1}{3}}{1-\frac{1}{2}\frac{1}{3}} = \frac{\frac{5}{6}}{1-\frac{1}{6}} = 1$$

Thus, $\tan(\theta+\psi) = 1$ so $\theta + \psi = \pi/4$, as claimed. □

Now, both $1/2$ and $1/3$ lie well within the radius of convergence of the arctangent, so we can add the two together to get a formula for π. Since series converge absolutely within their radii of convergence, we can re-arrange terms as we please, even combining the two into a single sum:

Theorem 32.3.

$$\frac{\pi}{4} = \sum_{k\geq 0} \frac{(-1)^k}{(2k+1)2^{2k+1}} + \sum_{n\geq 0} \frac{(-1)^k}{(2k+1)3^{2k+1}}$$

$$= \sum_{k\geq 0} \frac{(-1)^k}{2k+1}\left(\frac{1}{2^{2k+1}} + \frac{1}{3^{2k+1}}\right)$$

This series converges very quickly, as the exponents 2^{2k+1} and 3^{2k+1} in the denominators grow rapidly. Indeed, summing up to $N = TWO$ already gives the first two decimal digits!

$$\left(\frac{1}{2} + \frac{1}{3}\right) - \frac{1}{3}\left(\frac{1}{8} + \frac{1}{27}\right) + \frac{1}{5}\left(\frac{1}{32} + \frac{1}{243}\right) = 3.14558$$

Using up until $N = 10$ terms in this series gives the approximation

$$\pi \approx 3.14159257960635$$

Which is correct to 7 decimal digits. To get 15 significant digits using 22 terms in this series is enough!

This is truly a marvelous machine we have built - conjuring directly from the lowly geometric series an efficient formula for π.

The End

Example 32.2 (Epilog)**.** Want to be even more clever? In 1796 John Machin showed the following identity:

$$\frac{\pi}{4} = 4\arctan(1/5) - \arctan(1/239)$$

Note: If you wish to prove this, probably the easiest way is to notice that $(5+i)^4(239-i) = -114244(1+i)$ *and use the polar form of complex numbers to get the result. See here: https://people.math.sc.edu/howard/Classes/555c/trig.pdf*

This allows you to compute π to five or six decimals without much trouble. Just using the first five terms in the series gives $\pi \approx 3.14159268240440$ so we are already good to seven decimals. Using nine terms in the series gives you 15 significant digits

Project

This project illustrates just how far we have come in a semester of real analysis: we can now prove the identity that made Euler famous:

$$\sum_{n\geq 1}\frac{1}{n^2}=\frac{\pi^2}{6}$$

And, the identity that bears his name today:

$$e^{i\pi}=-1$$

This project is a collaboration between you and the text: the full proof takes place over as sequence of smaller results, some of which are proven here, and some of which are left as exercises.

Trigonometry

Both of these identities involve π, a number intimately connected to trigonometry. So, one would (correctly) assume that the trigonometric functions will play a crucial role in their proofs. But how can we suitably define sine and cosine (and π itself) in a way that is both fully rigorous and helpful for our purposes?

One such method is to define them via a *functional equation* - below we give the same functional equation presented in class as the definition.

Definition 32.1 (Angle Identities)**.** A pair of two functions (c, s) are *trigonometric* if they are a continuous & differentiable nonconstant solution to the *angle identities*

$$s(x-y)=s(x)c(y)-c(x)s(y)$$

$$c(x-y)=c(x)c(y)+s(x)s(y)$$

Remark 32.3. In fact one can drop the hypothesis that they are differentiable, and actually *prove this* from the others. But such a proof takes us a bit astray from the main goal of the project, so we instead opt for what initially appears to be a stronger characterization.

Identities

A good warm-up to functional equations is using them to prove some identities! I'll do the first one for you

Lemma 32.1 (Values at Zero)**.** *If s, c are trigonometric, then we can calculate their values at* 0*:*

$$s(0) = 0 \qquad c(0) = 1$$

Proof. Setting $x = y$ in the first immediately gives the first claim

$$s(0) = s(x - x) = s(x)c(x) - c(x)s(x) = 0$$

Evaluating the second functional equation also at $x = y$

$$c(0) = c(x - x) = c(x)c(x) + s(x)s(x) = c(x)^2 + s(x)^2$$

From this we can see that $c(0) \neq 0$, as if it were, we would have $c(x)^2 + s(x)^2 = 0$: since both $c(x)^2$ and $s(x)^2$ are nonnegative this implies each are zero, and so we would have $c(x) = s(x) = 0$ are constant, contradicting the definition. Now, plug in 0 to what we've derived, and use that we know $s(0) = 0$

$$c(0) = c(0)^2 + s(0)^2 = c(0)^2$$

Finally, since $c(0)$ is nonzero we may divide by it, which gives $c(0) = 1$ as claimed. □

An important corollary showed up during the proof here, when we observed that $c(0) = c(x)^2 + s(x)^2$: now that we know $c(0) = 1$, we see that (c, s) satisfy the Pythagorean identity!

Corollary 32.3 (Pythagorean Identity)**.** *If s, c are trigonometric, then for every $x \in \mathbb{R}$*

$$s(x)^2 + c(x)^2 = 1$$

Continuing this way, we can prove many other trigonometric identities: for instance, the double angle identity (which will be useful to us later)

Lemma 32.2 (Evenness and Oddness)**.** *If s, are trigonometric, then s is odd and c is even:*

$$s(-x) = -s(x) \qquad c(-x) = c(x)$$

Lemma 32.3 (Angle Sums). *If s, c are trigonometric, then for every $x \in \mathbb{R}$*

$$s(x+y) = c(x)s(y) + s(x)c(y)$$

$$c(x+y) = c(x)c(y) - s(x)s(y)$$

Corollary 32.4 (Double Angles). *If s, c satisfy the angle sum identities, then for any $x \in \mathbb{R}$,*

$$s(2x) = 2s(x)c(x)$$

Exercise 32.1. Prove Lemma 32.2, Lemma 32.3 and Corollary 32.4.

Another useful identity that I will need at the very end (but you will not, for any of the exercises) is the 'Half Angle Identity' for $c(x)$:

Lemma 32.4. *If s, c are trigonometric functions, then*

$$c(x)^2 = \frac{1 + c(2x)}{2}$$

Proof. Using the angle sum identity we see

$$c(2x) = c(x)c(x) - s(x)s(x) = c(x)^2 - s(x)^2$$

Then applying the pythagorean identity

$$\begin{aligned} c(2x) &= c(x)^2 - s(x)^2 \\ &= c(x)^2 - (1 - c(x)^2) \\ &= 2c(x)^2 - 1 \end{aligned}$$

Re-arranging yields the claimed identity. □

Differentiability

We have assumed as part of the definition that the trigonometric functions (s, c) are differentiable: here we calculate their derivatives and use them to provide some important limits and estimates.

Lemma 32.5. *Show that the derivatives of the trigonometric functions are completely determined by their derivatives at zero:*

$$s'(x) = s'(0)c(x) + c'(0)s(x)$$

$$c'(x) = c'(0)c(x) - s'(0)s(x)$$

Further, what we already know about s and c (together with the fact that they are nonconstant) implies a bit about their values at zero:

Lemma 32.6. *If (s, c) are trigonometric, then $c'(0) = 0$ and $s'(0) \neq 0$.*

Exercise 32.2. Prove Lemma 32.5 and Lemma 32.6

As a rather direct corollary of this, we have complete formulas for the derivatives of trigonometric functions, at every point along the real line!

Corollary 32.5 (Derivatives of s, c)**.** *If s, c are trigonometric, then for a fixed nonzero $\lambda \in \mathbb{R}$*

$$s'(x) = \lambda c(x) \qquad c'(x) = -\lambda s(x)$$

Definition 32.2. The sine and cosine functions are the trigonometric pair where $\sin'(0) = 1$.

Thus, we see $\sin'(x) = \cos(x)$ and $\cos'(x) = -\sin(x)$. Recalling the limit definition of the derivative, this gives us the ability to compute one particularly useful limit:

Proposition 32.5. *For any $x \in \mathbb{R}$*

$$\lim n \sin\left(\frac{x}{n}\right) = x$$

Exercise 32.3. Prove Proposition 32.5

Periodicity

We've already learned a lot about the trigonometric functions sin and cos. But you might notice the fundamental constant π is nowhere to be found! How do we rigorously define this number in the context of trigonometry? You may recall from previous classes that the trig functions have periods 2π. This provides a possible path: we may hope to *define* π a half the period of the trigonometric functions!

But to do so, we first need to show the functions even *have* a period. Why must solutions to the angle identities be periodic?

Lemma 32.7. *The cosine function has a root.*

Proof. Since the cosine is not constant and $\cos(0) = 1$, there must be some t_0 for which $\cos(t_0) \neq 0$. Since $\cos^2 + \sin^2 = 1$ we see that $-1 \leq \cos \leq 1$ so $\cos(t_0) < 1$.

If $\cos(t_0)$ is negative, we are done by the intermediate value theorem - there is a zero between 0 and t_0. So, we may assume $0 < \cos(t_0) < 1$, and define the sequence

$$c_n = \cos(2^n t_0)$$

To show cos is eventually negative (and thus, has a root by the intermediate value theorem argument) it suffices to see that c_n is eventually negative; and thus that $L = \inf\{c_n\}$ is negative (note the infimum exists as the set $\{c_n\}$ is bounded below by -1).

First, notice that the half angle identity implies $2c_0^2 - 1 = c_1$. For $x \in (0,1)$, we see $2x^2 - x - 1$ is negative: plugging in c_0 yields $2c_0^2 - c_0 - 1 < 0$, or $c_1 = 2c_0^2 - 1 < c_0$. Thus, c_0 is not the smallest term in our sequence, and we can truncate it without changing the infimum:

$$\inf_{n\geq 0}\{c_n\} = \inf_{n\geq 0}\{c_{n+1}\}$$

Using again the half angle identity, $2c_n^2 - 1 = c_{n+1}$, so

$$L = \inf\{c_n\} = \inf\{c_{n+1}\} = \inf\{2c_n^2 - 1\} = 2\inf\{c_n^2\} - 1$$

If our sequence were never negative, then $\inf\{c_n\} = L \geq 0$ and $\inf\{c_n^2\} = L^2$. Combining with the above, this implies $L = 2L^2 - 1$ whose only positive solution is $L = 1$ (which we know is not the infimum, as $c_0 < 1$). Thus, this is impossible, so it must be that $L < 0$, and our sequence eventually reaches a negative term.

Applying the intermediate value theorem to on the interval between $c(t_0) > 0$ and $c(2^n t_0) < 0$ furnishes a zero. □

This shows that cosine has a zero *somewhere*. Because it will be convenient below, we carry this reasoning a little farther and show that cosine actually has a *first positive zero*.

Lemma 32.8. *There is a $z > 0$ such that $\cos z = 0$, but the cosine is positive on the interval $[0, z)$: that is, z is the first zero of the cosine.*

Proof. Let x be a zero of the cosine function. Since the cosine is even we know $-x$ is also a zero: and, since $\cos(0) = 1$ we know neither $x \neq 0$ so at least one of $\pm x$ is positive. Thus, the cosine has at least one positive real root.

Let $R = \{x > 0 \mid \cos(x) = 0\}$ be the set of all positive roots of the cosine function. We prove this set has a minimum element, which is the first zero. Since R is nonempty (our first observation) and bounded below by zero (by definition) completeness implies $r = \inf R$ exists. For every $n \in \mathbb{N}$, since $r + 1/n$ is not an upper bound we may choose some $x_n \in \mathbb{R}$ with $r \leq x_n \leq r + 1/n$. By the squeeze theorem $x_n \to r$, and by continuity of the cosine this implies

$$\lim \cos(x_n) = \cos(\lim x_n) = \cos(r)$$

However each x_n is a zero of cosine by definition! Thus this is the constant sequence $0, 0, \dots$, which converges to 0. All together this means $\cos(r) = 0$, and so $r \in R$. But if the infimum is an element of the set then that set has a minimum element, so r is the smallest positive zero of the cosine! □

Remark 32.4. The argument above works generally for continuous functions: we did not use special properties of the cosine. Indeed, being familiar with the cosine function itself this argument might seem a little strange: the zeroes of cosine are evenly spaced out (at intervals of size $\pi \approx 3$) so what are the zeroes we are finding between r and $r + 1/n$? In fact - these are all just r, and the sequence constructed in the above argument is *constant* (but we don't need to know that, for the argument to go through).

It turns out that simply knowing the *existence* of a single zero of the cosine function is enough to resolve everything.

Proposition 32.6 (Periodicity of s, c)**.** *The functions* $\sin(x)$ *and* $\cos(x)$ *are periodic, with the same period* $P > 0$.

Exercise 32.4. Prove Proposition 32.6. *Hint This period is four times the first zero of cosine.*

Definition 32.3. π is the half-period of sine and cosine. Equivalently, π is the first positive zero of the sine function.

This brings us to the final major result we need about the trigonometric functions:

Proposition 32.7. *For all* $x \in \mathbb{R}$*, the cosine is just a shifted version of the sine function., and that sine is symmetric about* $\pi/2$*:*

$$\cos x = \sin\left(x + \frac{\pi}{2}\right)$$

$$\sin\left(\frac{\pi}{2} + x\right) = \sin\left(\frac{\pi}{2} - x\right)$$

Exercise 32.5. Prove Proposition 32.7.

Finally, we will need one more fact about the sine function (which will be crucial to us extracting π in the Basel problem), which relates its values before its first maximum to a secant line connecting the origin to that point:

Proposition 32.8. *For all* $x \in [0, \pi/2]$,

$$\frac{2x}{\pi} < \sin(x)$$

Exercise 32.6. Prove Proposition 32.8

Euler's Identity

Starting from the functional equations, we have done a lot of work to understand the *properties* that trigonometric functions must have, but we have not yet proven that there are such functions! We are going to use the functional properties of trigonometry to solve the Basel problem, so to be fully rigorous we better make sure the tools we use actually exist!

There are many routes to doing so, but the path we follow here meanders through some particularly beautiful mathematics in its own right. We will – as a corollary of proving that trigonometry exists – discover the famous identity of Euler $e^{i\pi} = -1$.

First, we use what we learned about differentiation to produce candidate functions as infinite series

Definition 32.4 (The Series $C(x)$ and $S(x)$)**.** Define the following two series

$$C(x) = \sum_{n\geq 0} \frac{(-1)^n}{(2n)!} x^{2n} \qquad S(x) = \sum_{n\geq 0} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$$

Exercise 32.7. Prove that $C(x)$ and $S(x)$ are absolutely convergent on the entire real line, and that they satisfy the differential equation required of sine and cosine:

$$S'(x) = C(x) \qquad C'(x) = -S(x)$$

Now, we show these candidate functions actually satisfy the angle sum identites.

There are many possible arguments here. The one we'll take uses some complex numbers, but does not require any complex analysis. The only fact needed is that i is a number where $i^2 = -1$, which allows complex multiplication to be defined by the distributive property:

$$(a + bi)(c + di) = ac + adi + bci + bdi^2 = (ac - bd) + (ad + bc)i$$

We can use the fact that a complex number $a + bi$ has a real component (a) and an imaginary component b to define convergence: a complex series converges if both its real and imaginary parts converge.

Definition 32.5 (The Function $\mathrm{CIS}(x)$)**.** Define the function $\mathrm{CIS}(x)$ as

$$\mathrm{CIS}(x) = C(x) + iS(x)$$

Using the series for $C(x)$ and $S(x)$ we can produce a series for $\mathrm{CIS}(x)$.

Proposition 32.9.

$$\mathrm{CIS}(x) = \sum_{n\geq 0} \frac{1}{n!}(ix)^n$$

Proof. First, we recall the definition as a limit of finite sums, and work with the limit laws to get to a single partial sum:

$$\begin{aligned}
\mathrm{CIS}(x) &= \sum_{n\geq 0} \frac{(-1)^n}{(2n)!}x^{2n} + i\sum_{n\geq 0}\frac{(-1)^n}{(2n+1)!}x^{2n+1} \\
&= \lim_N \sum_{0\leq n\leq N} \frac{(-1)^n}{(2n)!}x^{2n} + i\lim_N \sum_{0\leq n\leq N}\frac{(-1)^n}{(2n+1)!}x^{2n+1} \\
&= \lim_N \left(\sum_{0\leq n\leq N} \frac{(-1)^n}{(2n)!}x^{2n} + i\sum_{0\leq n\leq N}\frac{(-1)^n}{(2n+1)!}x^{2n+1}\right) \\
&= \lim_N \sum_{0\leq n\leq N}\left(\frac{(-1)^n}{(2n)!}x^{2n} + i\frac{(-1)^n}{(2n+1)}x^{2n+1}\right)
\end{aligned}$$

Where in the last line here we re-arranged the terms of this finite sum (between 0 and N) Writing out the $N = 3$ case for concreteness we see

$$(1+ix) + \left(\frac{-1}{2!}x^2 + i\frac{-1}{3!}x^3\right) + \left(\frac{1}{4!}x^4 + i\frac{1}{5!}x^5\right) + \left(\frac{-1}{6!}x^6 + \frac{-1}{7!}x^7\right)$$

Using that $i^2 = -1$ we see that we can re-write this summation as

$$(1+ix) + \left(\frac{1}{2!}(ix^2) + \frac{1}{3!}(ix)^3\right) + \left(\frac{1}{4!}(ix)^4 + \frac{1}{5!}(ix)^5\right) + \left(\frac{1}{6!}(ix)^6 + \frac{1}{7!}(ix)^7\right)$$

This has a nice enough pattern that we can re-package it into summation notation $\sum_{0\leq n\leq 7}\frac{1}{n!}(ix)^n$, or more generally

$$\sum_{0\leq n\leq 2N+1}\frac{1}{n!}(ix)^n$$

Because this is *exactly equal* to the partial sums defining $\mathrm{CIS}(x)$ (all we did was algebra to finite arithmetic!), taking the limit as $N \to \infty$ gives the claim. □

But this looks awfully familiar: we are acquainted with the power series $E(x) = \sum_{n\geq 0} x^n/n!$: this defines the exponential function!

Corollary 32.6.

$$CIS(x) = e^{ix}$$

From here, it is easy work to show that the the functions C and S satisfy the angle identities: this is just the law of exponents, which was the defining property of any exponential $E(x + y) = E(x)E(y)$

Proposition 32.10. *The functions C and S satisfy the angle identities:*

$$S(x - y) = S(x)C(y) - C(x)S(y)$$

$$C(x - y) = C(x)C(y) + S(x)S(y)$$

Thus, trigonometric functions exist! This is all we need to move on, and use trigonometry in our solution of the Basel problem.

But…let's not rush so quickly. From our work with the functional equations, we know that $C(x) = \cos(x)$ and $S(x) = \sin(x)$ are periodic, with half period π, which is also the first zero of the sine function. Using this immediately yields something beautiful:

Corollary 32.7.

$$e^{i\pi} + 1 = 0$$

Exercise 32.8. Prove Proposition 32.10, and Corollary 32.7

The Basel Problem

Having now essentially developed the theory of trigonometry from scratch, we now aim to put what we've learned to use, in a rigorous solution to the Basel problem. This identity is so surprising because on one side there are only the natural numbers (as reciprocals, squared) and on the other side is the circle constant π (squared, over six)!

The outline of our approach is as follows:

- Find a trigonometric identity that involves the sum of squares of reciprocals (of something).
- Use this to get an expression that has both π in it, and sums of reciprocals of stuff.
- Take a limit to get an infinite sum of reciprocals, and do some algebra turn this into the sum we want.

A Trigonometric Identity

Step one is to find a trigonometric identity that allows us to expand something as a sum of reciprocal squares. This partial fractions decomposition of $1/\sin^2$ is a good start:

Proposition 32.11. *The following trigonometric identity holds for* $1/\sin^2$, *whenever* x *is not a multiple of* π.

$$\begin{aligned}\frac{1}{\sin^2 x} &= \frac{1}{4\sin^2\left(\frac{x}{2}\right)\cos^2\left(\frac{x}{2}\right)}\\ &= \frac{1}{4}\left[\frac{1}{\sin^2\left(\frac{x}{2}\right)}+\frac{1}{\cos^2\left(\frac{x}{2}\right)}\right]\\ &= \frac{1}{4}\left[\frac{1}{\sin^2\left(\frac{x}{2}\right)}+\frac{1}{\sin^2\left(\frac{x+\pi}{2}\right)}\right]\end{aligned}$$

Exercise 32.9. Prove the equalities claimed in Proposition 32.11 hold.

Let's see what we can do with this: applying it twice recursively,

$$\begin{aligned}\frac{1}{\sin^2 x} &= \frac{1}{4}\left[\frac{1}{\sin^2\left(\frac{x}{2}\right)}+\frac{1}{\sin^2\left(\frac{x+\pi}{2}\right)}\right]\\ &= \frac{1}{4}\left[\frac{1}{4}\left[\frac{1}{\sin^2\left(\frac{x/2}{2}\right)}+\frac{1}{\sin^2\left(\frac{x/2+\pi}{2}\right)}\right]+\frac{1}{4}\left[\frac{1}{\sin^2\left(\frac{(x+\pi)/2}{2}\right)}+\frac{1}{\sin^2\left(\frac{(x+\pi)/2+\pi}{2}\right)}\right]\right]\\ &= \frac{1}{16}\left[\left(\frac{1}{\sin^2\left(\frac{x}{4}\right)}+\frac{1}{\sin^2\left(\frac{x}{4}+\frac{\pi}{2}\right)}\right)+\left(\frac{1}{\sin^2\left(\frac{x}{4}+\frac{\pi}{4}\right)}+\frac{1}{\sin^2\left(\frac{x}{4}+\frac{3\pi}{4}\right)}\right)\right]\\ &= \frac{1}{16}\left[\frac{1}{\sin^2\left(\frac{x}{4}\right)}+\frac{1}{\sin^2\left(\frac{x}{4}+\frac{\pi}{4}\right)}+\frac{1}{\sin^2\left(\frac{x}{4}+\frac{2\pi}{4}\right)}+\frac{1}{\sin^2\left(\frac{x}{4}+\frac{3\pi}{4}\right)}\right]\end{aligned}$$

https://homepage.univie.ac.at/josef.hofbauer/02amm.pdf

Applying once more to each term of the sum (and skipping the algebraic simplifications) yields

$$\frac{1}{\sin^2 x} = \frac{1}{64}\left[\frac{1}{\sin^2\left(\frac{x}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{2\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{3\pi}{8}\right)} + \right.$$
$$\left. + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{4\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{5\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{6\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{x}{8}+\frac{7\pi}{8}\right)}\right]$$

Putting this all together and looking at the three rounds of expansion, we see that

$$\begin{aligned}\frac{1}{\sin^2 x} &= \frac{1}{4}\sum_{k=0}^{1}\frac{1}{\sin^2\left(\frac{x}{2}+\frac{k\pi}{2}\right)} \\ &= \frac{1}{4^2}\sum_{k=0}^{3}\frac{1}{\sin^2\left(\frac{x}{4}+\frac{k\pi}{4}\right)} \\ &= \frac{1}{4^3}\sum_{k=0}^{7}\frac{1}{\sin^2\left(\frac{x}{8}+\frac{k\pi}{8}\right)}\end{aligned}$$

Carrying this out inductively n times straightforwardly yields

Proposition 32.12. *For any $n \in \mathbb{N}$, the function $1/\sin^2(x)$ can be expressed as the following finite sum:*

$$\frac{1}{\sin^2 x} = \frac{1}{4^n}\sum_{0\le k<2^n}\frac{1}{\sin^2\left(\frac{x+k\pi}{2^n}\right)}$$

While this trigonometric identity is interesting in its own right, we will only require a special case for our application: evaluating at $x = \pi/2$ we find

Corollary 32.8.

$$1 = \frac{1}{4^n}\sum_{0\le k<2^n}\frac{1}{\sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right)}$$

Where $\mathfrak{o}_k := 2k+1 = 1, 3, 5, 7, \ldots$ *is the sequence of odd numbers.*

This gives, for every n, a large finite sum (its 2^n terms long!) of reciprocals of squares. There are just two obstacles in our way:

- The sum is finite!
- The reciprocals are complicated trigonometric functions

We can solve both by (carefully) taking the limit as $n \to \infty$.

Taking the Limit

Because for all n this finite sum *exactly equals 1* we know in the limit as $n \to \infty$ it equals 1 as well:

$$1 = \lim_{n\to\infty} \sum_{0\leq k<2^n} \frac{1}{4^n \sin^2\left(\frac{\mathfrak{o}_k \pi}{2^{n+1}}\right)}$$

Now we have the delicate issue of taking the limit as $n \to \infty$. It is tempting to take the limit of the terms individually but this is not always justified: as the simple example below shows

Example 32.3.

$$\begin{aligned} 1 &= \frac{1}{2} + \frac{1}{2} \\ &= \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \end{aligned}$$

Taking the termwise limit and adding them up gives

$$1 = 0 + 0 + 0 + \cdots + 0 = 0$$

What's going on here is that we are implicitly *exchanging two limits* and we haven't justified that such an exchange is possible: in the toy example above, one may define for each n the series

$$a_n(k) = \begin{cases} 1/2^n & 0 \leq k < 2^n \\ 0 & \text{else} \end{cases}$$

Then each of the rows above is the sum $1 = \sum_{k\geq 0} a_n(k)$ for $n = 2, 3, 4$. Since this is constant it is true that the limit is 1, but it is *not true* that the limit of the sums is the sum of the limits, which is zero.

$$1 = \lim_n \sum_{k\geq 0} a_n(k) \neq \sum_{k\geq 0} \lim a_n(k) = 0$$

The reason this fails is that our sum does not satisfy the hypotheses of dominated convergence. Recall that requires that all of the $a_n(k)$ (for all n) must be less than some convergent series. Here we can see that $a_n(k) < \frac{1}{k}$ (and we can't do better than this) but the harmonic series diverges! So the hypotheses of dominated convergence are violated, and switching limits leads to disaster.

The form of this example is exactly replicated in our question: if we define for each n an infinite series with terms $a_n(k)$ as

$$a_n(k) = \begin{cases} 1/4^n \sin^2\left(\frac{\mathfrak{o}_k \pi}{2^{n+1}}\right) & 0 \leq k < 2^n \\ 0 & k \geq 2^n \end{cases}$$

we see that $1 = \lim_n \sum_{k\geq 0} a_n(k)$ (Where again, $\mathfrak{o}_k = (2k-1)$ is just shorthand for the sequence of odd numbers) but to make progress computing this limit, we need to be able to exchange it with the order of summation, so we need a means of applying *dominated convergence*.

Unfortunately, this is tricker than one might hope: our series **fails** the hypotheses of dominated convergence as written!

Remark 32.5. Try to see this yourself, by drawing a graph of $a_n(k)$ as a function of k, for various n. For any given n, can you find where the maximal value of $a_n(k)$ occurs (at which k), and what it's value is?

Why does this prevent you from building a dominating sequence?

Luckily, there is a way to proceed: for a fixed n, the coefficients we are summing up involve the sine function evaluated at odd multiples of $\pi/2^{n+1}$. This list of numbers is symmetric about $\pi/2$, and since $\sin(x)$ is symmetric about $\pi/2$, so are our coefficients.

Completing the first half of the sum and doubling the result gives the same value: illustrating with the $n = 2$ iteration,

$$\begin{aligned} 1 &= \frac{1}{16}\left[\frac{1}{\sin^2\left(\frac{\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{3\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{5\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{7\pi}{8}\right)}\right] \\ &= \frac{1}{16}\left[\left(\frac{1}{\sin^2\left(\frac{\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{7\pi}{8}\right)}\right) + \left(\frac{1}{\sin^2\left(\frac{3\pi}{8}\right)} + \frac{1}{\sin^2\left(\frac{5\pi}{8}\right)}\right)\right] \\ &= \frac{1}{16}\left[\frac{2}{\sin^2\left(\frac{\pi}{8}\right)} + \frac{2}{\sin^2\left(\frac{3\pi}{8}\right)}\right] \end{aligned}$$

Carrying out this finite re-arrangement for the n^{th} iteration yields

Corollary 32.9.

$$1 = \frac{1}{4^n} \sum_{0\leq k<2^n} \frac{1}{\sin^2\left(\frac{\mathfrak{o}_k \pi}{2^{n+1}}\right)} = \frac{1}{4^n} \sum_{0\leq k<\frac{1}{2}2^n} \frac{2}{\sin^2\left(\frac{\mathfrak{o}_k \pi}{2^{n+1}}\right)}$$

Where $\mathfrak{o}_k = 1, 3, 5, \ldots$ *is the sequence of odd numbers.*

Since our new upper summation limit is below $\frac{1}{2}2^n = 2^{n-1}$, the values we are plugging into the sine function are all less than $\pi/2$, and so Proposition 32.8 applies, allowing us to bound sin below, and hence $1/\sin$ above.

Proposition 32.13. *The series*

$$\sum_{0\leq k<2^{n-1}} \frac{1}{4^n \sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right)}$$

Satisfies the hypotheses of dominated convergence: precisely, define

$$a_n(k) = \begin{cases} 1/4^n \sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right) & 0 \leq k < 2^{n-1} \\ 0 & k \geq 2^n \end{cases}$$

- *Independently of n the k^{th} term $a_n(k)$ is bounded above by $2/\mathfrak{o}_k^2$; that is $2/(2k-1)^2$*
- *The sum of these bounds converges*

Exercise 32.10. Prove Proposition 32.13

This justifies the exchange of limits, which will be crucial to our the main step.

Theorem 32.4.

$$\lim_n \sum_{0\leq k<2^{n-1}} \frac{2}{4^n \sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right)} = \sum_{k\geq 0} \lim_n \frac{2}{4^n \sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right)}$$

The Termwise Limit

Taking the limit term by term allows us to start by finding the limit of the denominator:

Proposition 32.14.

$$\lim_n 4^n \sin^2\left(\frac{\mathfrak{o}_k\pi}{2^{n+1}}\right) = \left(\frac{\mathfrak{o}_k\pi}{2}\right)^2$$

Exercise 32.11. Prove this

Putting this back into our original trigonometric identity gives

$$1 = \sum_{k\geq 0} \frac{2}{\left(\frac{\mathfrak{o}_k\pi}{2}\right)^2}$$

This is the key to a rigorous derivation of Euler's incredible identity.

Exercise 32.12. Re-arrange the sum above to show

$$\frac{\pi^2}{8} = \sum_{k\geq 0} \frac{1}{\mathfrak{o}_k^2} = 1 + \frac{1}{3^2} + \frac{1}{5}^2 + \cdots$$

And show that this implies

$$\sum_{n\geq 1} \frac{1}{n^2} = \frac{\pi^2}{6}$$

Exercise 32.13. Before taking the termwise limit, we had to be careful and replace our sum with different sum (having the same value) so that we could apply Dominated Convergence. Show that this is necessary: take the termwise limit of the original sum and show you get the wrong answer for $\sum_{k\geq 0} \frac{1}{\mathfrak{o}_k^2}$.

Hints

Below are hints to the various exercises in this document: I encourage you to try them without the hints! But feel free to get inspiration here, and to ask for further hints in office hours if a particular problem has you stuck.

Exercise 1: Look at $s(0-x)$ or $c(0-x)$ now that we know the values of s and c at zero...

Exercise 2: For the first part: use the definition of the derivative, and the angle sum identites. For the follow-up: we can know $c'(0) = 0$ because we can prove 0 is the location of a max and apply theorems. If both s' and c' were zero at $x = 0$, can you show $s(x)$ and $c(x)$ must be constant?

Exercise 3: The definition of the derivative is a limit, and limits of functions are defined in terms of sequences...

Exercise 4: If z is the first zero of cosine, show that $c(4z) = 1$ and $s(4z) = 0$, and that this implies $c(x + 4z) = c(x)$, $s(x + 4z) = s(x)$.

Exercise 5: Use the definition of π and the angle sum/difference identities.

Exercise 6: Use a theorem relating the second derivative to convexity. A convex function always has its secant line *above the graph.* What about the negative of a convex function?

Exercise 7: Show convergence with a test for power series, then use a theorem on differentiating power series within their radius of convergence.

Exercise 8: Use the definition of CIS(x), the properties of complex multiplication, and the definition of π.

Exercise 9: Use the trigonometric identities derived from the original functional equation.

Exercise 10: Use Exercise 6 to get the bound, and comparison to prove convergence.

Exercise 11: Exercise 3 tells us about $\lim n \sin(x/n)$. Can you rewrite what you have here into something like $\lim 2^{n+1} \sin(x/2^{n+1})$? Then use facts about subsequences and limits.

Exercise 12: The first part is just algebra. For the second, can you break the sum of $1/n^2$ into the sum of even and odd terms? Can you do something about the even terms, to get an equation relating the sum of all reciprocals to just the sum of the odd ones?

Exercise 13: Just repeat the calculation you worked through above, but for the other sum.