Honors Calculus

Pete L. Clark

Thanks to Ron Freiwald, Eilidh Geddes, Melody Hine, Nita Jain, Andrew Kane, Bryan Oakley, Elliot Outland, Didier Piau, Jim Propp, Betul Tolgay, Troy Woo, and Sharon Yao.

Contents

Foreword

Spivak and Me

The document you are currently reading began its life as the lecture notes for a year long undergraduate course in honors calculus. Specifically, during the 2011-2012 academic year I taught Math 2400(H) and Math 2410(H), *Calculus With Theory*, at the University of Georgia. This is a course for unusually talented and motivated (mostly first year) undergraduate students. It has been offered for many years at the University of Georgia, and so far as I know the course text has always been Michael Spivak's celebrated *Calculus* [**S**]. The Spivak text's take on calculus is sufficiently theoretical that, although it is much beloved by students and practitioners of mathematics, it is seldomed used nowadays as a course text. In fact, the UGA Math 2400/2410 course is traditionally something of an interpolation between standard freshman calculus and the fully theoretical approach of Spivak. My own take on the course was different: I treated it as being a sequel to, rather than an enriched revision of, freshman calculus.

I began the course with a substantial familiarity with Spivak's text. The summer after my junior year of high school I interviewed at the University of Chicago and visited Paul Sally, then (and still now, as I write this in early 2013, though he is 80 years old) Director of Undergraduate Mathematics at the University of Chicago. After hearing that I had taken AP calculus and recently completed a summer course in multivariable calculus at Johns Hopkins, Sally led me to a supply closet, rummaged through it, and came out with a beat up old copy of Spivak's text. "This is how we do calculus around here," he said, presenting it to me. During my senior year at high school I took more advanced math courses at LaSalle University (which turned out, almost magically, to be located directly adjacent to my high school) but read through Spivak's text. And I must have learned something from it, because by the time I went on to college – of course at the University of Chicago – I placed not into their Spivak calculus course, but the following course, "Honors Analysis in $\mathbb{R}^n$". This course has the reputation of sharing the honor with Harvard's Math 55 of being the hardest undergraduate math course that American universities have to offer. I can't speak to that, but it was certainly the hardest math course I ever took. There were three ten week quarters. The first quarter was primarily taught out of Rudin's classic text [**R**], with an emphasis on metric spaces. The second quarter treated Lebesgue integration and some Fourier analysis, and the third quarter treated analysis on manifolds and Stokes's theorem.

However, that was not the end of my exposure to Spivak's text. In my second year of college I was a grader for the first quarter of Spivak calculus, at the (even then) amazingly low rate of $20 per student for the entire 10 week quarter. Though

the material in Spivak's text was at least a level below the trial by fire that I had successfully endured in the previous year, I found that there were many truly difficult problems in Spivak's text, including a non-negligible percentage that I still did not know how to solve. Grading for this course solidified my knowledge of this elementary but important material. It was also my first experience with reading proofs written by bright but very inexperienced authors: often I would stare at an entire page of text, one long paragraph, and eventually circle a single sentence which carried the entire content that the writer was trying to express. I only graded for one quarter after that, but I was a "drop in tutor" for my last three years of college, meaning that I would field questions from any undergraduate math course that a student was having trouble with, and I had many further interactions with Spivak's text. (By the end of my undergraduate career there were a small number of double-starred problems of which I well knew to steer clear.)

Here is what I remembered about Spivak's text in fall of 2011:

(i) It is an amazing trove of problems, some of which are truly difficult.

(ii) The text itself is lively and idiosyncratic.[1]

(iii) The organization is somewhat eccentric. In particular limits are not touched until Chapter 5. The text begins with a chapter "basic properties of numbers", which are essentially the ordered field axioms, although not called such. Chapter 3 is on Functions, and Chapter 4 is on Graphs. These chapters are essentially contentless. The text is broken up into five "parts" of which the third is *Derivatives and Integrals*.

After teaching the 2400 course for a while, I lost no esteem for Spivak's text, but increasingly I realized it was not the ideal accompaniment for my course. For one thing, I realized that my much more vivid memories of the problems than the text itself had some basis in fact: although Spivak writes with a lively and distinct voice and has many humorous turns of phrase, the text itself is rather spare. His words are (very) well chosen, but few. When one takes into account the ample margins (sometimes used for figures, but most often a white expanse) the chapters themselves are very short and core-minded. When given the chance to introduce a subtlety or ancillary topic, Spivak almost inevitably defers it to the problems.

I have had many years to reflect on Spivak's text, and I now think its best use is in fact the way I first encountered it myself: as a source of self study for bright, motivated students with little prior background and exposure to university level mathematics (in our day, this probably means *high school students*, but I could imagine a student underwhelmed by an ordinary freshman calculus class for which Spivak's text would be a similarly mighty gift). Being "good for self study" is a high compliment to pay a text, and any such text can *a fortiori* also be used in a course...but not in a completely straightforward way. For my course, although the

[1]In between the edition of the book that Sally had given me and the following edition, all instances of third person pronouns referring to mathematicians had been changed from "he" to "she". Initially I found this amusing but silly. More recently I have begun doing so myself.

students who stuck with it were very motivated and hard-working, most (or all) of them needed *a lot* of help from me in all aspects of the course. I had been informed in advance by my colleague Ted Shifrin that signing on to teach this course should entail committing to about twice as many office hours as is typical for an undergraduate course, and this did indeed come to pass: I ended up having office hours four days a week, and I spent the majority of them helping the students make their way through the problems in Spivak's text. Even the best students – who, by the way, I thought were awfully good – could not solve some of the problems unassisted. There were many beautiful, multi-part problems introducing extra material that I wanted my students to experience: and eventually I figured out that the best way to do this was to incorporate the problems into my lectures as theorems and proofs.

All in all I ended up viewing Spivak's book as being something of a "deconstruction"[2] of the material, and much of my teaching time was spent "reconstructing" it.

For me, the lightness of touch of Spivak's approach was ultimately something I appreciated aesthetically but could see was causing my students difficulty at various key points. My experience rather convinced me that "more is more": students wanted to see more arguments in complete detail, and also simply more proofs overall. Aside from the (substantial) content conveyed in the course, a major goal is to give the students facility with reading, understanding, constructing and critiquing proofs. In this regard being in a classroom setting with other motivated students is certainly invaluable, and I tried to take advantage of this by insisting that students present their questions and arguments *to each other* as well as to me.

But students also learn about how to reason and express themselves mathematically by being repeatedly exposed to careful, complete proofs. I think experienced mathematicians can forget the extent to which excellent, even brilliant, students benefit from this exposure. Of course they also benefit immensely, probably more so, by working things out for themselves. The value of this is certainly not in question, and there is no competition here: the amount of "things to do" in honors calculus is *infinite* – and the repository of problems in Spivak's text is nearly so – so by working out more for herself, the instructor is not leaving *less* for the students to do, but only different things for them to do.

This brings me to the current text. As explained above, it is heavily indebted to [**S**]. However, it is – or at least, I mean it to be – a new honors calculus text, and not merely a gloss of [**S**]. Indeed:

• The text is indebted to [**S**], but not uniquely so. It is also heavily indebted to Rudin's classic text [**R**], and it borrows at key points from several other sources, e.g. [**A**], [**Go**], [**H**], [**L**].

I do not view this borrowing as being in any way detrimental, and I certainly do not attempt to suppress my reliance on other texts. The mathematics exposed here is hundreds of years old and has been treated excellently in many famous texts. An undergraduate mathematics text which strove to be unlike its predecessors would

[2]My conception of the meaning of **deconstruction** in the sense of academic humanities is painfully vague. I am more thinking of the term in the sense that chefs use it.

almost certainly do so to the detriment of its audience.

Having said this, most undergraduate texts with "calculus" in the title offer very little innovation indeed. It is distressing to me how many calculus books are written which look like nearly exact replicas of some platonic (but far from perfect) calculus text written circa 1920. Apparently the issue of treating transcendental functions earlier rather than later was enough to warrant many new editions, if not wholly new texts.

This text does contain some novelties, up to and including a proof technique, **real induction**, which to my knowledge has not appeared before in texts. Spivak's text is a great innovation, and its distinctiveness – some might say eccentricity – has not gone unnoticed by instructors. This text is also eccentric, and in its own way: if you like Spivak's text, you need not like this one.

What is Honors Calculus?

It was audacious of Spivak to call his text simply "calculus," since it is light years away from any other text of that name currently in print. I have chosen to call this text: **honors calculus**. What is honors calculus?

By "honors calculus" I mean a certain take on undergraduate real analysis.

Some Features of the Text

• Our approach is heavily *theoretical* but not heavily *abstract*.

The jumping off point for an honors calculus course is a theoretical understanding of the results seen in freshman calculus, especially rigorous definitions of the various limiting processes and proofs of the "hard theorems" of calculus, which inevitably turn on the completeness of $\mathbb{R}$. We are not interested in calculation for its own sake, and we do not give any halfway serious applications to any subject outside mathematics. This theoretical perspective is of course a huge change from the practical, problem-oriented approach of freshman calculus (with which we assume a prior familiarity). It is certainly not as appealing to as broad an audience as freshman calculus. At the University of Georgia, it happens that many freshman students (especially, those in the honors program) are signed up for Math 2400 without any real idea of what they are getting into. The attrition at the beginning of the course is therefore significant, and the students that remain are not necessarily the most talented or experienced but rather those that remain interested in a fully theoretical approach to calculus. This text assumes that interest on the part of the reader, in fact probably even more so than Spivak's text.

The idea of presenting a fully theoretical take on real analysis to a young undergraduate audience is hardly new, of course. But most contemporary texts on real analysis take a more sophisticated approach, making use of concepts from set theory and topology. In this text we rarely speak explicitly about the connectedness and compactness of intervals on the real line. (More precisely, we *never* use these concepts in the main text, but only in some digressions which briefly and optionally present some advanced material: e.g. § 10.10.5 introduces metric spaces and discussed compactness therein.) Instead we make ϵ-δ arguments which amount to this. Let us freely admit that this concreteness makes certain arguments *longer*

and harder (or phrased more positively: one merit of abstraction is to make certain arguments *shorter and easier*).

I am a fan of abstraction, but I have come to believe that it is much less useful – and moreover, much less educational – in basic real analysis than in most other branches of pure mathematics. A turning point in my feelings about this was a second semester undergraduate real analysis course I taught at McGill University in 2005. When preparing to teach the course I was quite surprised at how pedestrian the syllabus seemed: an entire year of real analysis was restricted to the one-dimensional case, and there was no topology or set theory whatsoever.[3] I was strictly forbidden from talking about metric spaces, for instance. The topics of the course were: infinite series, the Riemann integral, and sequences and series of functions, including uniform convergence. By the end of the course I had acquired more respect for the deep content inherent in these basic topics and the efficacy of treating them using no more than ϵ-δ arguments and the completeness axiom. This was also the first course I taught in which I typed up lecture notes, and portions of these notes appear here as parts of Chapters 11 through 14.

Exception: We begin our treatment of the Riemann integral with an *axiomatic approach*: that is, we list some reasonable properties (axioms) that an area functional should satisfy, and before we do the hard work of constructing such a functional we explore the consequences of these axioms. In particular, less than two pages into our discussion of integration we state and prove (given the axioms!) the Fundamental Theorem of Calculus. This approach to the Riemann Integral resonates deeply with me, as it addresses concerns that bubbled up over several years of my teaching this material in the context of freshman calculus. I came up with this approach in late 2004 while preparing for the McGill analysis course. Just a few days later I noticed that Lang had done (essentially) the same thing in his text [**L**]. I was encouraged that this idea had been used by others, and I endorse it still.

Given the choice between pounding out an ϵ-δ argument and developing a more abstract or softer technique that leads to an easier proof, in this text we usually opt for the former. Well, that's not strictly true: sometimes we do both.

• We often spend time giving multiple proofs and approaches to basic results.

For instance, there are two distinct approaches to the Riemann integral: Riemann's original approach using tagged partitions and Riemann sums, and Gaston Darboux's later simplification using upper and lower sums and integrals. Most texts at this level cover only one of these in detail, but here we cover both: first Darboux, then Riemann. I got "permission" to do so while teaching the analysis class at McGill, since this was done in the official course text of R. Gordon [**Go**]. Later I realized that Gordon is in real life an integration theorist!

More significantly, we prove the Interval Theorems using ϵ-δ arguments and later come back to give much quicker proofs using sequences and the Bolzano-Weierstrass Theorem. One may well say that this is evidence that sequences should be treated

[3]In fact, I didn't find out until just after the course ended that the students did not know about countable and uncountable sets. Without conscious thought I had assumed otherwise.

at the beginning of the course rather than towards the end, and many texts do take this approach, most notably [**R**]. However I endorse Spivak's ordering of the material, in which honors calculus essentially begins with a thorough grappling with the ϵ-δ definition of limit. Although there are easier – and especially, **softer** – approaches to most of the key theorems, I feel that manipulation of inequalities that characterized **hard** analysis is a vitally important *skill* for students to learn and this is their best chance to learn it.

• We view the completeness axiom as the star of the show. Thus we do not put it on the stage in Act I, Scene I, when people are still settling into their seats.

CHAPTER 1

Introduction and Foundations

1. Introduction

1.1. The Goal: Calculus Made Rigorous.

The goal of this course is to cover the material of single variable calculus *in a mathematically rigorous way*. The latter phrase is important: in most calculus classes the emphasis is on techniques and applications; while theoretical explanations may be given by the instructor – e.g. it is usual to give some discussion of the meaning of a **continuous function** – the student tests her understanding of the theory mostly or entirely through her ability to apply it to solve problems. This course is very different: not only will **theorems** and **proofs** be presented in class by me, but they will also be presented by you, the student, in homework and on exams. This course offers a strong foundation for a student's future study of mathematics, at the undergraduate level and beyond.

As examples, here are three of the fundamental results of calculus; they are called – by me, at least – the three **Interval Theorems**, because they all concern an arbitrary continuous function defined on a closed, bounded interval.

THEOREM 1.1. *(Intermediate Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on a closed, bounded interval. Suppose that $f(a) < 0$ and $f(b) > 0$. Then there exists c with $a < c < b$ such that $f(c) = 0$.*

THEOREM 1.2. *(Extreme Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on a closed, bonuded interval. Then f is bounded and assumes its maximum and minimum values: there are numbers $m \leq M$ such that*
a) For all $x \in [a, b]$, $m \leq f(x) \leq M$.
b) There exists at least one $x \in [a, b]$ such that $f(x) = m$.
c) There exists at least one $x \in [a, b]$ such that $f(x) = M$.

THEOREM 1.3. *(Uniform Continuity and Integrability) Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on a closed, bounded interval. Then:*
a) f is uniformly continuous.[1]
b) f is integrable: $\int_a^b f$ exists and is finite.

Except for the part about uniform continuity, these theorems are familiar results from freshman calculus. **Their proofs, however, are not.** Most freshman calculus texts like to give at least *some* proofs, so it is often the case that these three theorems are used to prove even more famous theorems in the course, e.g. the

[1]The definition of this is somewhat technical and will be given only later on in the course. Please don't worry about it for now.

Mean Value Theorem and the **Fundamental Theorem of Calculus**.

Why then are the three interval theorems not proved in freshman calculus? *Because their proofs depend upon fundamental properties of the real numbers that are not discussed in such courses.* Thus one of the necessary tasks of the present course is to give a more penetrating account of the real numbers than you have seen before.

1.2. Numbers of Various Kinds.

There are various kinds of "numbers". Here is a list of the ones which will be most important to us:

$$\mathbb{Z}^+ \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}. \tag{1}$$

Let me remind you what these various numbers are.

$$\mathbb{Z}^+ = \{1, 2, 3, \ldots, n, \ldots\}$$

is the set of **positive integers** (a.k.a. "counting numbers"). On them we have defined the operations of addition $+$ and multiplication $\cdot$. Moreover there is an identity element for the multiplication, namely 1. There is no additive identity.

$$\mathbb{N} = \{0\} \cup \mathbb{Z}^+ = \{0, 1, 2, 3, \ldots, n, \ldots\}$$

is the set of **natural numbers**. Again we have operations of addition and multiplication, and now we have an additive identity as well as a multiplicative identity.

Clearly $\mathbb{Z}^+$ and $\mathbb{N}$ are very similar: they differ only as to whether 0 is included or not. In analysis – the subject we are beginning the study of here! – the distinction between $\mathbb{Z}^+$ and $\mathbb{N}$ is not very important, and in fact most analysts I know use $\mathbb{N}$ to denote the positive integers. I am probably showing my stripes as an algebraically minded mathematician by making the distinction, but so be it.

Recall that the operation of subtraction is nothing else than the inverse operation of addition: in other words, to say $a - b = c$ is to say that $a = b + c$. However the operation of subtraction is not everywhere defined on $\mathbb{N}$: for instance, there is a natural number $5 - 3$, but no natural number $3 - 5$.

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$$

is the set of **integers**. This is formed out of the natural numbers $\mathbb{N}$ by formally allowing all subtractions: for instance, -17 is $0 - 17$. Thus in $\mathbb{Z}$ every element n has an additive inverse $-n$. However, the same cannot be said of multiplicative inverses. Recall that the operation of division is nothing else than the inverse operation of multiplication: to say $a/b = c$ is to say that $a = b \cdot c$. However the operation of division is not everywhere defined on $\mathbb{Z}$: thre is an integer $6/3$, but no integer $6/4$.

$$\mathbb{Q} = \{\frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0\}.$$

is the set of **rational numbers**. This is formed out of the integers $\mathbb{Z}$ by formally allowing all divisions by nonzero integers. Note that one subtlety here is that the same rational number has many different expressions as the quotient of two integers: for instance $\frac{6}{4} = \frac{3}{2} = \frac{3n}{2n}$ for any $n \in \mathbb{Z}^+$. So we just need to agree that $\frac{a}{b} = \frac{c}{d}$ iff $ad = bc$. Alternately, any nonzero rational number has a unique expression $\frac{a}{b}$ **in**

lowest terms, i.e., with a and b not both divisible by any integer $n > 1$.[2] Thus in $\mathbb{Q}$ we have an additive identity 0, every element has an additive inverse, we have a multiplicative identity 1, and every nonzero element has a multiplicative inverse.

What then are the real numbers $\mathbb{R}$? The geometric answer is that the real numbers correspond to "points on the number line", but this does not make clear why there are such points other than the rational numbers. An answer that one learns in high school is that every real number has an *infinite decimal expansion*, not necessarily terminating or repeating, and conversely any integer followed by an infinite decimal expansion determines a real number. In fact this is perfectly correct: it gives a complete characterization of the real numbers, but it is not a cure-all: in order to pursue the implications of this definition – and even to really understand it – one needs tools that we will develop later in the course.

Finally, the complex numbers $\mathbb{C}$ are expressions of the form $a + bi$ where a and b are real numbers and $i^2 = -1$. They are extremely important in mathematics generally – e.g. one needs them in order to solve polynomial equations – but in this course they will play at most a peripheral role.

Back to $\mathbb{R}$: let us nail down the fact that there are real numbers which are not rational. One way to see this is as follows: show that the decimal expansion of every rational number is eventually periodic, and then exhibit a decimal expansion which is not eventually periodic, e.g.

$$x = 0.16116111611116111116\ldots$$

where the number of 1's after each 6 increases by 1 each time. But this number x reeks of contrivance: it seems to have been constructed only to make trouble. The ancient Pythagoreans discovered a much more "natural" irrational real number.

THEOREM 1.4. *The square root of* 2 *is not a rational number.*

PROOF. The proof is the most famous (and surely one of the first) instances of a certain important kind of argument, namely a **proof by contradiction**. We assume that what we are trying to prove is false, and from that we reason until we reach an absurd conclusion. Therefore what we are trying to prove must be true.

Here goes: seeking a contradiction, we suppose $\sqrt{2}$ *is* rational: then there are integers a, b with $b > 0$ and $\sqrt{2} = \frac{a}{b}$. Since the defining property of $\sqrt{2}$ is that its square is 2, there is really nothing to do but square both sides to get

$$2 = \frac{a^2}{b^2}.$$

Clearing denominators, we get

$$2b^2 = a^2.$$

This shows that the integer a^2 is *even*, i.e., divisible by 2. It happens that for any integer a, if a^2 is even, then so is a: let us assume this for now; we can come back to it later. Thus we may write $a = 2A$ with $A \in \mathbb{Z}$. Substituting, we get

$$2b^2 = a^2 = (2A)^2 = 4A^2,$$

[2]Like most of the statements we have made recently, this requires proof! Let me reiterate that we are not giving proofs here or even careful definitions; rather, we are reminding the reader of some of her mathematical past.

or

$$b^2 = 2A^2.$$

Thus b^2 is divisible by 2, so as above $b = 2B$ for some $B \in \mathbb{Z}$. Substituting, we get

$$4B^2 = (2B)^2 = b^2 = 2A^2,$$

or

$$2B^2 = A^2.$$

Thus we are back where we started: assuming that $2b^2 = a^2$, we found that both a and b were divisible by 2. This is suspect in the extreme, and we now have our choice of killing blow. One ending is to observe that everything we have said above applies to A and B: thus we must also have $A = 2A_1$, $B = 2B_2$, and so forth. We can continue in this way factoring out as many powers of 2 from a and b as we wish. But the only integer which is arbitrarily divisible by 2 is 0, so our conclusion is $a = b = 0$, whereas we assumed $b > 0$: contradiction.

Alternately – and perhaps more simply – each rational number may be written in lowest terms, so we could have assumed this about $\frac{a}{b}$ at the outset and, in particular, that a and b are not both divisible by 2. Either way we get a contradiction, so $\sqrt{2}$ must not be a rational number. □

1.3. Why do we not do calculus on $\mathbb{Q}$?

To paraphrase the title question, why do we *want* to use $\mathbb{R}$ to do calculus? Is there something stopping us from doing calculus over, say, $\mathbb{Q}$?

The answer to the second question is **no**: we can define limits, continuity, derivatives and so forth for functions $f : \mathbb{Q} \to \mathbb{Q}$ *exactly* as is done for real functions. The most routine results carry over with no change: it is still true, for instance, that sums and products of continuous functions are continuous. However most of the big theorems – especially, the **Interval Theorems** – become false over $\mathbb{Q}$.

For $a, b \in \mathbb{Q}$, let $[a, b]_{\mathbb{Q}} = \{x \in \mathbb{Q} \mid a \le x \le b\}$.

EXAMPLE 1.1. *Consider the function $f : [0, 2]_{\mathbb{Q}} \to \mathbb{Q}$ given by $f(x) = -1$ if $x^2 < 2$ and $f(x) = 1$ if $x^2 > 2$. Note that we do not need to define $f(x)$ at $x = \pm\sqrt{2}$, because by the result of the previous section these are not rational numbers. Then f is continuous – in fact it is differentiable and has identically zero derivative. But $f(0) = -1 < 0$, $f(2) = 1 > 0$, and there is no $c \in [0, 2]_{\mathbb{Q}}$ such that $f(c) = 0$. Thus the Intermediate Value Theorem fails over $\mathbb{Q}$.*

EXAMPLE 1.2. *Consider the function: $f : [0, 2]_{\mathbb{Q}} \to \mathbb{Q}$ given by $f(x) = \frac{1}{x^2-2}$. Again, this function is well-defined at all points of $[0, 2]_{\mathbb{Q}}$ because $\sqrt{2}$ is not a rational number. It is also a continuous function. However it is* not *bounded above: by taking rational numbers which are arbitrarily close to $\sqrt{2}$, $x^2 - 2$ becomes arbitrarily small and thus $f(x)$ becomes arbitarily large.*[3] *In particular, f certainly does not attain a maximum value. Thus the Extreme Value Theorem fails over $\mathbb{Q}$.*

Moreover, it can be shown (and will be – later) that any function on a closed, bounded interval which is either uniformly continuous or integrable is bounded, so the above function f is neither uniformly continuous nor integrable. If you have

[3]We will be much more precise about this sort of thing later on. This is just an overview.

had second semester freshman calculus, you should think about why the analogous function $f : [0,2] \setminus \{\sqrt{2}\} \to \mathbb{R}$ is not improperly Riemann integrable: it builds up infinite area as we approach $\sqrt{2}$.

The point of these examples is in order to succeed in getting calculus off the ground, we need to make use of some fundamental property of the real numbers not possessed by (for intance) the rational numbers. This property, which can be expressed in various forms, is called **completeness**, and will play a major role in this course.

2. Some Properties of Numbers

2.1. Axioms for a Field.

In order to do mathematics in a rigorous way, one needs to identify a starting point. Virtually all mathematical theorems are of the form $A \implies B$. That is, assuming A, B must follow. For instance, in Euclidean geometry one lays down a set of axioms and reasons only from them. The axioms needed for calculus are a lot to swallow in one dose, so we will introduce them gradually. What we give here is essentially a codification of high school algebra, including inequalities.

Specifically, we will give axioms that we want a **number system** to satisfy. At this point we will take it for granted that in our number system we have operations of addition, multiplication and an inequality relation $<$, and that there are distinguished numbers called 0 and 1. We require the following properties:

(P0) $0 \neq 1$.
(P1) (Commutativity of $+$): For all numbers x, y, $x + y = y + x$.
(P2) (Associativity of $+$): For all numbers x, y, z, $(x + y) + z = x + (y + z)$.
(P3) (Identity for $+$): For all numbers x, $x + 0 = x$.
(P4) (Inverses for $+$): For all numbers x, there exists y with $x + y = 0$.
(P5) (Commutativity of $\cdot$): For all numbers x, y, $x \cdot y = y \cdot x$.
(P6) (Associativity of $\cdot$): For all numbers x, y, z $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.
(P7) (Identity for $\cdot$): For all numbers x, $x \cdot 1 = x$.
(P8) (Inverses for $\cdot$) For all numbers $x \neq 0$, there exists a number y with $xy = 1$.
(P9) (Distributivity of $\cdot$ over $+$): For all numbers x, y, z, $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$.

Although it is not important for us now, the above axioms (P0) through (P9) are called the **field axioms**, and a structure which satisfies them is called a **field**.

EXAMPLE 1.3. *Both $\mathbb{Q}$ and $\mathbb{R}$ satisfy all of the above field axioms. (We take this as "known" information.)*

EXAMPLE 1.4. *The complex numbers $\mathbb{C}$ satisfy all of the above field axioms. The only one which is not straightforward is the existence of multiplicative inverses. For this: if $z = x + iy$ is a nonzero complex number – i.e., the real numbers x and y are not both zero – then if $w = \frac{x - iy}{x^2 + y^2}$, $zw = 1$.*

EXAMPLE 1.5. *Let $\mathbb{F}_2 = \{0, 1\}$ be a set consisting of two elements, 0 and 1. We define $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, $1 + 1 = 0$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$, $1 \cdot 1 = 1$. Then $\mathbb{F}_2$ satisfies all of the above field axioms. It is sometimes called the **binary field**.*

PROPOSITION 1.5. *In every system satisfying the field axioms, for every number x we have $x \cdot 0 = 0$.*

PROOF. We have $x \cdot 0 = x \cdot (0 + 0) = (x \cdot 0) + (x \cdot 0)$. Subtracting $(x \cdot 0)$ from both sides gives $0 = x \cdot 0$. □

PROPOSITION 1.6. *In every system satisfying the field axioms:*
a) The only additive identity is 0.
b) Every number x has a unique additive inverse. If -1 denotes the additive inverse of 1*, then the additive inverse of x is $(-1) \cdot x$.*
c) The only multiplicative identity is 1.
b) Every nonzero number has a unique multiplicative inverse.

PROOF. a) Note that 0 is an additive identity by (P3). Suppose that z is another additive identity, and consider $0+z$. Since 0 is an additive identity, $0+z = z$. Since z is an additive identity, $0 + z = 0$. Thus $z = 0$.
b) Suppose y and z are both additive inverses to x: $x + y = x + z = 0$. Adding y to both sides gives

$$y = 0 + y = (x + y) + y = (y + x) + y = y + (x + y)$$
$$= y + (x + z) = (y + x) + z = (x + y) + z = 0 + z = z,$$

so $y = z$. Moreover, for any number x,

$$(-1) \cdot x + x = ((-1) \cdot x) + (1 \cdot x) = (-1 + 1) \cdot x = 0 \cdot x = 0.$$

c), d) The proofs of these are the same as the proofs of parts a) and b) but with all instances of + replaced by $\cdot$ and all instances of 0 replaced by 1. □

PROPOSITION 1.7. *In every system satisfying the field axioms, $(-1)^2 = 1$.*

PROOF. By Proposition 1.6, $-1 \cdot -1$ is the additive inverse of -1, namely 1. □

PROPOSITION 1.8. *In every system satisfying the field axioms, if $x \neq 0$ and $y \neq 0$ then $xy \neq 0$.*

PROOF. Seeking a contradiction, suppose that $xy = 0$. Since $x \neq 0$ it has a multiplicative inverse x^{-1} and then by Proposition 1.5,

$$0 = x^{-1} \cdot xy = (x^{-1} \cdot x)y = 1 \cdot y = y,$$

contradicting the assumption that $y \neq 0$. □

Note that a logically equivalent formulation of Proposition 1.8 is: in any system satisfying the field axioms, if $xy = 0$ then $x = 0$ or $y = 0$.

2.2. Axioms for an ordered field.

The remaining properties of numbers concern the inequality relation $<$. Instead of describing the relation $<$ directly, it turns out to be simpler to talk about the properties of positive numbers. If we are given the inequality relation $<$, then we say that x is **positive** if $x > 0$, thus knowing $<$ we know which numbers are positive. Conversely, suppose that we have identified a subset $\mathcal{P}$ of numbers as positive. Then we can define $x < y$ if $y - x \in \mathcal{P}$. Now we want our set of positive numbers to satisfy the following properties.

(P10) (Trichotomy) For all numbers x, exactly one of the following holds: $x = 0$, x

is positive, $-x$ is positive.
(P11) (Closure under $+$) For all positive numbers x, y, $x + y \in \mathcal{P}$.
(P12) (Closure under $\cdot$) For all positive numbers x, y, $x \cdot y \in \mathcal{P}$.

A number system satisfying (P1) through (P12) is called an **ordered field**.

PROPOSITION 1.9. *In every ordered field:*
a) $1 > 0$ *and* $-1 < 0$.
b) For every nonzero x, $x^2 > 0$.
c) It follows that for all x, $x^2 \geq 0$.

PROOF. a) By (P0), $1 \neq 0$. Thus by trichotomy, either 1 is positive and -1 is negative, or -1 is positive and 1 is negative. But by (P12) the product of two positive numbers is positive, so if -1 is positive and 1 is negative then $1 = (-1)^2$ is positive, a contradiction. So it must be that 1 is positive and -1 is negative.
b) Since x is nonzero, either $x > 0$ or $-x > 0$. If $x > 0$, then $x^2 = x \cdot x$ is the product of two positive numbers, hence positive. If $x < 0$, then $-x > 0$ and then $x^2 = (-1)^2 x^2 = (-x) \cdot (-x)$ is the product of two positive numbers, hence positive.
c) Since $0^2 = 0$, part c) follows immediately from part b). □

EXAMPLE 1.6. *The binary numbers* $\mathbb{F}_2$ *satisfy the field axioms (P0) through (P9), but are they an ordered field? Well, not on the face of it because we have not been given an inequality relation* $<$ *satisfying (P10) through (P12). In fact we will now show that there is no such relation. Indeed, in any ordered field, since* $1 > 0$, *also* $1 + 1 > 0$, *but in* $\mathbb{F}_2$ $1 + 1 = 0$. *In fancy language,* $\mathbb{F}_2$ *is a field which cannot be endowed with the structure of an ordered field.*

EXAMPLE 1.7. *The complex numbers* $\mathbb{C}$ *satisfy the field axioms (P0) through (P9), but are they an ordered field? As above, we have not been given an inequality relation. Also as above we can show that there is no such relation. For in the complex numbers we have an element* i *with* $i^2 = -1$. *But the ordered field axioms imply both that* -1 *is negative and that any square is non-negative, contradiction.*

PROPOSITION 1.10. *For any* x, y, z *in an ordered field:*
a) $x < 0 \iff 0 < -x$. *(We say "x is negative".)*
b) The trichotomy property may be restated as: for any number x, *exactly one of the following holds:* x *is positive,* x *is zero,* x *is negative.*
c) If x *is positive,* $\frac{1}{x}$ *is positive. If* x *is negative,* $\frac{1}{x}$ *is negative.*
d) If x *is positive and* y *is negative, then* xy *is negative.*
e) If x *and* y *are both negative, then* xy *is positive.*

PROOF. a) By definition, $x < 0$ means $0 - x = -x$ is positive. Also $0 < -x$ means $-x - 0 = -x$ is positive. So there is nothing to show here.
b) No further argument for this is needed; we just state it for future reference.
c) Suppose x is positive. Certainly $\frac{1}{x}$ is not zero, so we need to rule out the possibility that it's negative. But if it were, then by part a) $\frac{-1}{x}$ would be positive and thus by (P12) $x \cdot \frac{-1}{x} = -1$ would be positive, contradicting Proposition 1.9a). If x is negative then $-x$ is positive so by what we just showed $\frac{1}{-x} = \frac{-1}{x}$ is positive, and thus $\frac{1}{x} = -(\frac{-1}{x})$ is negative.
d) Suppose x is positive and y is negative. In particular x and y are not zero, so $xy \neq 0$. To show that xy is negative, by part b) it is enough to *rule out* the

possibility that xy is positive. Suppose it is. Then, by part c), since x is positive, $\frac{1}{x}$ is positive, and thus $y = xy \cdot \frac{1}{x}$ would be positive: contradiction.
e) Suppose x and y are both negative. Then $xy \neq 0$, and we need to rule out the possibility that xy is negative. Suppose it is. Then $-xy$ is positive, $\frac{1}{x}$ is negative, so by part d) $-y = -xy \cdot \frac{1}{x}$ is negative and thus y is positive: contradiction. □

PROPOSITION 1.11. *For all a, b, c, d in an ordered field:*
a) If $a < b$ and $c < d$, $a + c < b + d$.
b) If $a < b$ and $c > 0$, then $ac < bc$.
c) If $a < b$ and $c < 0$, then $ac > bc$.
d) If $0 < a < b$, then $0 < \frac{1}{b} < \frac{1}{a}$.
e) If $a > 0$ and $b > 0$, then $a < b \iff a^2 < b^2$.

PROOF. a) Since $a < b$ and $c < d$, $b - a$ is positive and $d - c$ is positive, and then by (P11) $(b - a) + (d - c) = (b + d) - (a + c)$ is positive, so $b + d > a + c$.
b) Since $a < b$, $b - a$ is positive. Since c is positive, by (P12) $bc - ac = (b - c)a$ is positive, so $bc > ac$.
c) Left to the reader.
d) We have $\frac{1}{a} - \frac{1}{b} = (b - a) \cdot \frac{1}{ab}$. The hypotheses imply that $b - a$ and $\frac{1}{ab}$ are both positive, so by (P12) so is their product.
e) Note that $b^2 - a^2 = (b + a)(b - a)$. Since a and b are both positive, $b + a$ is positive, and therefore $b - a$ is positive iff $b^2 - a^2$ is positive. □

2.3. Some further properties of $\mathbb{Q}$ and $\mathbb{R}$.

As we have mentioned before, the ordered field axioms (P0) through (P12) are just a list of *some* of the useful properties of $\mathbb{Q}$ and $\mathbb{R}$. They are not a "complete set of axioms" for either $\mathbb{Q}$ or $\mathbb{R}$ – in other words, there are other properties these fields have that cannot be logically deduced from these axioms alone. In fact this is already clear because $\mathbb{Q}$ and $\mathbb{R}$ each satisfy all the ordered field axioms but are essentially different structures: in $\mathbb{Q}$ the element $2 = 1 + 1$ is not the square of another element, but in $\mathbb{R}$ it is. Here we want to give some further "familiar" properties that do not hold for all ordered fields but both of which hold for $\mathbb{R}$ and one of which holds for $\mathbb{Q}$. (We are still far away from the fundamental **completeness axiom** for $\mathbb{R}$ which is necessary to prove the Interval Theorems.)

The first axiom is called the **Archimedean property**: it says that for any positive number x, there is a positive integer n such that $x \leq n$. This clearly holds for $\mathbb{R}$ according to our description of real numbers as integers followed by infinite decimal expansions: a positive real number x is of the form

$$x = n_0.a_1a_2 \dots a_n \dots$$

with $n_0 \in \mathbb{Z}^+$ and $a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and thus x is less than or equal to the integer $n_0 + 1$.[4]

Since every positive real number is less than or equal to some integer, and every positive rational number is, in particular, a positive real number, then also every positive rational number is less than or equal to some integer. That is, $\mathbb{Q}$ also

[4]When I first typed this I wrote that x is *less than* $n_0 + 1$. But actually this need not be true! Can you think of an example? Beware: decimal expansions can be tricky.

satisfies the Archimedean property. (Or, directly: any positive rational number may be written in the form $\frac{a}{b}$ with $a, b \in \mathbb{Z}^+$, and then $\frac{a}{b} \leq a$.)

This Archimedean property is *so* natural and familiar (not to mention useful...) that the curious student may be well wonder: are there in fact systems of numbers satisfying the ordered field axioms but *not* the Archimedean property?!? The answer is **yes**, there are plenty of them, and it is in fact possible to construct a theory of calculus based upon them (in fact, such a theory is in many ways more faithful to the calculus of Newton and Leibniz than the theory which we are presenting here, which is a 19th century innovation). But we will not see such things in this course!

The next property *does* provide a basic difference between $\mathbb{Q}$ and $\mathbb{R}$.

THEOREM 1.12. *Let x be a real number and $n \in \mathbb{Z}^+$.*
a) If n is odd, there is a unique real number y such that $y^n = x$. We write $y = \sqrt[n]{x}$.
b) If n is even and x is positive, there is a unique positive real number y such that $y^n = x$. We write $y = \sqrt[n]{x}$.
c) If n is even and x is negative, then there is no real number y with $y^n = x$.

The first two parts of Theorem 1.12 rely on the Intermediate Value Theorem so are not accessible to us at this time. (Thus we must guard aginst using the existence of nth roots of real numbers in any of the theorems that lead up to the Intermediate Value Theorem. In fact we will not use such things in the proof of any theorem, but only as examples.) As a supplement tto part b), note that if n is even and y is a positive real number such that $y^n = x$, then there is exactly one other real number with nth power equal to x: $-y$. (You might try to prove this as an exercise.)

We *can* prove part c) now, since it holds in any ordered field. Indeed, if n is even then $n = 2k$ for an integer k, so if for some negative x we have $y^n = x$ then $x = y^{2k} = (y^k)^2$, contradicting Proposition 1.9c).

Here is a special case of Theorem 1.12 important enough to be recorded separately.

COROLLARY 1.13. *A real number x is non-negative if and only if it is a square, i.e., if and only if there exists a real number y with $y^2 = x$.*

Note that Corollary 1.13 does not hold in the number system $\mathbb{Q}$, since $2 = 1 + 1$ is positive but is not the square of any rational number.

Corollary 1.13 leads to a basic strategy for proving inequalities in the real numbers: for $x, y \in \mathbb{R}$, $x \leq y \iff (y - x) = z^2$ for some real number z. In the next section we will see some instances of this strategy in action.

2.4. Some Inequalities.

For an element x of an ordered field, we define the **absolute value** of x to be x if $x \geq 0$ and $-x$ if $x < 0$; it is denoted by $|x|$. Thus $|x| \geq 0$ always and $x = \pm|x|$.

PROPOSITION 1.14. *For any number x in an ordered field, $x \leq |x|$.*

PROOF. If $x \geq 0$ then $x = |x|$. If $x < 0$ then $x < 0 < -x = |x|$, so $x < |x|$. □

THEOREM 1.15. *(Triangle Inequality) For all numbers x, y, $|x + y| \leq |x| + |y|$.*

PROOF. Since $|x|$ is defined to be x if $x \geq 0$ and $-x$ if $x < 0$, it is natural to break the proof into cases.
Case 1: $x, y \geq 0$. Then $|x + y| = x + y = |x| + |y|$.
Case 2: $x, y < 0$. Then $x + y < 0$, so $|x + y| = -(x + y) = -x - y = |x| + |y|$.
Case 3: $x \geq 0$, $y < 0$. Now unfortunately we do not know whether $|x + y|$ is non-negative or negative, so we must consider consider further cases.
Case 3a: $x + y \geq 0$. Then $|x + y| = x + y \leq |x| + |y|$.
Case 3b: $x + y < 0$. Then $|x + y| = -x - y \leq | - x| + | - y| = |x| + |y|$.
Case 4: $x < 0$, $y \geq 0$. The argument is exactly the same as that in Case 3. In fact, we can guarantee it is the same: since the desired inequality is **symmetric** in x and y – meaning, if we interchange x and y we do not change what we are trying to show – we may reduce to Case 3 by interchanging x and y.[5] □

The preceding argument is definitely the sort that one should be prepared to make when dealing with expressions involving absolute values. However, it is certainly not very much fun. Spivak gives an alternate proof of the Triangle Inequality which is more interesting and thematic. First, since both quantities $|x + y|$ and $|x| + |y|$ are non-negative, the inequality will hold iff it holds after squaring both sides (Proposition 1.11e). So it is enough to show

$$(|x + y|)^2 \leq (|x| + |y|)^2.$$

now $(|x+y|)^2 = (x+y)^2 = x^2+2xy+y^2$, whereas $(|x|+|y|)^2 = |x|^2+2|x||y|+|y|^2 = x^2 + |2xy| + y^2$, so subtracting the left hand side from the right, it is equivalent to show that

$$0 \leq (x^2 + |2xy| + y^2) - (x^2 + 2xy + y^2).$$

But

$$(x^2 + |2xy| + y^2 - (x^2 + 2xy + y^2) = |2xy| - 2xy \geq 0$$

by Proposition 1.14. So this gives a second proof of the Triangle Inequality.

A similar argument can be used to establish the following variant.

PROPOSITION 1.16. *(Reverse Triangle Inequality)*
For all numbers x, y, $||x| - |y|| \leq |x - y|$.

PROOF. Again, since both quantities are non-negative, it is sufficient to prove the inequality after squaring both sides:

$$(||x| - |y||)^2 = (|x| - |y|)^2 = |x|^2 - 2|x||y| + |y|^2 = x^2 - |2xy| + y^2$$
$$\leq x^2 - 2xy + y^2 = (x - y)^2 = (|x - y|)^2.$$ □

EXERCISE 1.1. *Let x, y be any numbers.*
a) Show that $|x| - |y| \leq |x - y|$ by writing $x = (x - y) + y$ and applying the usual triangle inequality.
b) Deduce from part a) that $||x| - |y|| \leq |x - y|$.

THEOREM 1.17. *(Cauchy-Bunyakovsky-Schwarz Inequality, $n = 2$)*
a) For all numbers x_1, x_2, y_1, y_2,

$$(x_1y_1 + x_2y_2)^2 \leq (x_1^2 + x_2^2)(y_1^2 + y_2^2). \tag{2}$$

[5]Such **symmetry arguments** can often by used to reduce the number of cases considered.

b) Moreover equality holds in (2) iff $y_1 = y_2 = 0$ or there exists a number λ such that $x_1 = \lambda y_1$ and $x_2 = \lambda y_2$.

PROOF. By pure brute force, one can prove the **two squares identity**:

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1y_2 - x_2y_1)^2 + (x_1y_1 + x_2y_2)^2.$$

Now we need only rewrite it in the form

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) - (x_1y_1 + x_2y_2)^2 = (x_1y_2 - x_2y_1)^2 \geq 0,$$

establishing part a). Moreover, equality holds iff $x_1y_2 = x_2y_1$. If in this equality y_1 and y_2 are both nonzero, we may divide by them to get $\frac{x_1}{y_1} = \frac{x_2}{y_2} = \lambda$. If $y_1 = 0$ and $y_2 \neq 0$ then we must have $x_1 = 0$ and then we may take $\lambda = \frac{x_2}{y_2}$. Similarly, if $y_1 \neq 0$ and $y_2 = 0$, then we must have $x_2 = 0$ and then we may take $\lambda = \frac{x_1}{y_1}$. Finally, if $y_1 = y_2 = 0$ then the equality $x_1y_2 = x_2y_1$ also holds. □

THEOREM 1.18. *(Cauchy-Bunyakovsky-Schwarz Inequality) For any $n \in \mathbb{Z}^+$ and numbers $x_1, \ldots, x_n, y_1, \ldots, y_n$ we have*

$$(x_1y_1 + \ldots + x_ny_n)^2 \leq (x_1^2 + \ldots + x_n^2)(y_1^2 + \ldots + y_n^2).$$

PROOF. Expanding out the right and left hand sides, we get

$$\text{RHS} = \sum_{i=1}^{n} x_i^2y_i^2 + \sum_{i \neq j} x_i^2y_j^2.$$

$$\text{LHS} = \sum_{i=1}^{n} x_i^2y_i^2 + 2\sum_{i<j} x_iy_ix_jy_j,$$

so

$$\text{RHS} - \text{LHS} = \sum_{i \neq j} x_i^2y_j^2 - 2\sum_{i<j} x_iy_jx_jy_i = \sum_{i<j}(x_iy_j - x_jy_i)^2 \geq 0.$$

□

THEOREM 1.19. *(Arithmetic-Geometric Mean Inequality, $n = 2$) For all numbers $0 < a < b$, we have*

$$a^2 < ab < \left(\frac{a+b}{2}\right)^2 < b^2.$$

PROOF. First inequality: Since $a > 0$ and $0 < a < b$, $a \cdot a < a \cdot b$.
Second inequality: Expanding out the square and clearing denominators, it is equivalent to $4ab < a^2 + 2ab + ab^2$, or to $a^2 - 2b + b^2 > 0$. But $a^2 - 2ab + b^2 = (a-b)^2$, so since $a \neq b$, $(a-b)^2 > 0$.
Third inequality: Since $\frac{a+b}{2}$ and b are both positive, it is equivalent to $\frac{a+b}{2} < b$ and thus to $a + b < 2b$. But indeed since $a < b$, $a + b < b + b = 2b$. □

Later we will use the theory of convexity to prove a signficant generalization of Theorem 1.19, the **Weighted Arithmetic-Geometric Mean Inequality**.

CHAPTER 2

Mathematical Induction

1. Introduction

Principle of Mathematical Induction for sets
Let S be a subset of the positive integers. Suppose that:
(i) $1 \in S$, and
(ii) $\forall\ n \in \mathbb{Z}^+, n \in S \implies n+1 \in S$.
Then $S = \mathbb{Z}^+$.

The intuitive justification is as follows: by (i), we know that $1 \in S$. Now apply (ii) with $n = 1$: since $1 \in S$, we deduce $1 + 1 = 2 \in S$. Now apply (ii) with $n = 2$: since $2 \in S$, we deduce $2 + 1 = 3 \in S$. Now apply (ii) with $n = 3$: since $3 \in S$, we deduce $3 + 1 = 4 \in S$. And so forth.

This is not a proof. (No good proof uses "and so forth" to gloss over a key point!) But the idea is as follows: we can keep iterating the above argument as many times as we want, deducing at each stage that since S contains the natural number which is one greater than the last natural number we showed that it contained. Now it is a fundamental part of the structure of the positive integers that every positive integer can be reached in this way, i.e., starting from 1 and adding 1 sufficiently many times. In other words, any **rigorous definition** of the natural numbers (for instance in terms of sets, as alluded to earlier in the course) needs to incorporate, either implicitly or (more often) explicitly, the principle of mathematical induction. Alternately, the principle of mathematical induction is a key ingredient in any axiomatic characterization of the natural numbers.

It is not a key point, but it is somewhat interesting, so let us be a bit more specific. In Euclidean geometry one studies points, lines, planes and so forth, but one does not start by saying what sort of object the Euclidean plane "really is". (At least this is how Euclidean geometry has been approached for more than a hundred years. Euclid himself gave such "definitions" as: "A point is that which has position but not dimensions." "A line is breadth without depth." In the 19th century it was recognized that these are descriptions rather than definitions, in the same way that many dictionary definitions are actually descriptions: "cat: A small carnivorous mammal domesticated since early times as a catcher of rats and mice and as a pet and existing in several distinctive breeds and varieties." This helps you if you are already familiar with the animal but not the word, but if you have never seen a cat before this definition would not allow you to determine with certainty whether any particular animal you encountered was a cat, and still less would it allow you to reason abstractly about the cat concept or "prove theorems about cats.") Rather

"point", "line", "plane" and so forth are taken as **undefined terms**. They are related by certain **axioms**: abstract properties they must satisfy.

In 1889, the Italian mathematician and proto-logician Gisueppe Peano came up with a similar (and, in fact, much simpler) system of axioms for the natural numbers. In slightly modernized form, this goes as follows:

The undefined terms are **zero**, **number** and **successor**.

There are five axioms that they must satisfy, the **Peano axioms**. The first four are:
(P1) Zero is a number.
(P2) Every number has a successor, which is also a number.
(P3) No two distinct numbers have the same successor.
(P4) Zero is not the successor of any number.

Using set-theoretic language we can clarify what is going on here as follows: the structures we are considering are triples $(X, 0, S)$, where X is a set, 0 is an element of X, and $S : X \to X$ is a function, subject to the above axioms.

From this we can deduce quite a bit. First, we have a number (i.e., an element of X) called $S(0)$. Is $0 = S(0)$? No, that is prohibited by (P4). We also have a number $S(S(0))$, which is not equal to 0 by (P4) and it is also not equal to $S(0)$, because then $S(0) = S(S(0))$ would be the successor of the distinct numbers 0 and $S(0)$, contradicting (P3). Continuing in this way, we can produce an infinite sequence of distinct elements of X:

$$0, S(0), S(S(0)), S(S(S(0)), \dots. \tag{3}$$

In particular X itself is infinite. The crux of the matter is this: is there any element of X which is *not* a member of the sequence (3), i.e., is not obtained by starting at 0 and applying the successor function finitely many times?

The axioms so far do not allow us to answer this question. For instance, suppose that the "numbers" consisted of the set $[0, \infty)$ of all non-negative real numbers, we define 0 to be the real number of that name, and we define the successor of x to be $x + 1$. This system satisfies (P1) through (P4) but has much more in it than just the natural numbers we want, so we must be missing an axiom! Indeed, the last axiom is:

(P5) If Y is a subset of the set X of numbers such that $0 \in Y$ and such that $x \in Y$ implies $S(x) \in Y$, then $Y = X$.

Notice that the example we cooked up above fails (P5), since in $[0, \infty)$ the subset of natural numbers contains zero and contains the successor of each of its elements but is a proper subset of $[0, \infty)$.

Thus it was Peano's contribution to realize that mathematical induction is an axiom for the natural numbers in much the same way that the parallel postulate is an axiom for Euclidean geometry.

On the other hand, it is telling that this work of Peano is little more than one hundred years old, which in the scope of mathematical history is quite recent. Traces of what we now recognize as induction can be found from the mathematics of antiquity (including Euclid's Elements!) on forward. According to the (highly recommended!) Wikipedia article on mathematical induction, the first mathematician to formulate it explicitly was Blaise Pascal, in 1665. During the next hundred years various equivalent versions were used by different mathematicians – notably the methods of infinite descent and minimal counterexample, which we shall discuss later – and the technique seems to have become commonplace by the end of the 18th century. Not having an formal understanding of the relationship between mathematical induction and the structure of the natural numbers was not much of a hindrance to mathematicians of the time, so still less should it stop us from learning to use induction as a proof technique.

Principle of mathematical induction for predicates
Let $P(x)$ be a sentence whose domain is the positive integers. Suppose that:
(i) $P(1)$ is true, and
(ii) For all $n \in \mathbb{Z}^+$, $P(n)$ is true $\implies$ $P(n+1)$ is true.
Then $P(n)$ is true for all positive integers n.

Variant 1: Suppose instead that $P(x)$ is a sentence whose domain is the natural numbers, i.e., with zero included, and in the above principle we replace (i) by the assumption that $P(0)$ is true and keep the assumption (ii). Then of course the conclusion is that $P(n)$ is true for all natural numbers n. This is more in accordance with the discussion of the Peano axioms above.[1]

EXERCISE 2.1. *Suppose that N_0 is a fixed integer. Let $P(x)$ be a sentence whose domain contains the set of all integers $n \geq N_0$. Suppose that:*
(i) $P(N_0)$ is true, and
(ii) For all $n \geq N_0$, $P(n)$ is true $\implies$ $P(n+1)$ is true.
Show that $P(n)$ is true for all integers $n \geq N_0$. (Hint: define a new predicate $Q(n)$ with domain $\mathbb{Z}^+$ by making a "change of variables" in P.)

2. The First Induction Proofs

2.1. The Pedagogically First Induction Proof.

There are many things that one can prove by induction, but the first thing that everyone proves by induction is invariably the following result.

PROPOSITION 2.1. *For all $n \in \mathbb{Z}^+$, $1 + \ldots + n = \frac{n(n+1)}{2}$.*

PROOF. We go by induction on n.
Base case ($n = 1$): Indeed $1 = \frac{1(1+1)}{2}$.
Induction step: Let $n \in \mathbb{Z}^+$ and suppose that $1 + \ldots + n = \frac{n(n+1)}{2}$. Then

$$1 + \ldots + n + n + 1 = (1 + \ldots + n) + n + 1 \stackrel{\text{IH}}{=} \frac{n(n+1)}{2} + n + 1$$

[1]In fact Peano's original axiomatization did not include zero. What we presented above is a standard modern modification which is slightly cleaner to work with.

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

Here the letters "IH" signify that the induction hypothesis was used. □

Induction is such a powerful tool that once one learns how to use it one can prove many nontrivial facts with essentially no thought or ideas required, as is the case in the above proof. However thought and ideas are good things when you have them! In many cases an inductive proof of a result is a sort of "first assault" which raises the challenge of a more insightful, noninductive proof. This is certainly the case for Proposition 2.1 above, which can be proved in many ways.

Here is one non-inductive proof: replacing n by $n-1$, it is equivalent to show:

$$\forall n \in \mathbb{Z}, n \geq 2 : \ 1 + \ldots + n - 1 = \frac{(n-1)n}{2}. \tag{4}$$

We recognize the quantity on the right-hand side as the **binomial coefficient** $\binom{n}{2}$: it counts the number of 2-element subsets of an n element set. This raises the prospect of a **combinatorial proof**, i.e., to show that the number of 2-element subsets of an n element set is *also* equal to $1 + 2 + \ldots + n - 1$. This comes out immediately if we list the 2-element subsets of $\{1, 2, \ldots, n\}$ in a systematic way: we may write each such subset as $\{i, j\}$ with $1 \leq i \leq n-1$ and $i < j \leq n$. Then:

The subsets with least element 1 are $\{1,2\}, \{1,3\}, \ldots, \{1,n\}$, a total of $n-1$.
The subsets with least element 2 are $\{2,3\}, \{2,4\}, \ldots, \{2,n\}$, a total of $n-2$.
$\vdots$
The subset with least element $n-1$ is $\{n-1, n\}$, a total of 1.

Thus the number of 2-element subsets of $\{1, \ldots, n\}$ is on the one hand $\binom{n}{2}$ and on the other hand $(n-1) + (n-2) + \ldots + 1 = 1 + 2 + \ldots + n - 1$. This gives a combinatorial proof of Proposition 2.1.

2.2. The (Historically) First(?) Induction Proof.

THEOREM 2.2. *(Euclid) There are infinitely many prime numbers.*

PROOF. For $n \in \mathbb{Z}^+$, let $P(n)$ be the assertion that there are at least n prime numbers. Then there are infinitely many primes if and only if $P(n)$ holds for all positive integers n. We will prove the latter by induction on n.
Base Case ($n = 1$): We need to show that there is at least one prime number. For instance, 2 is a prime number.
Induction Step: Let $n \in \mathbb{Z}^+$, and assume that $P(n)$ holds, i.e., that there are at least n prime numbers $p_1 < \ldots < p_n$. We need to show that $P(n+1)$ holds, i.e., there is at least one prime number different from the numbers we have already found. To establish this, consider the quantity

$$N_n = p_1 \cdots p_n + 1.$$

Since $p_1 \cdots p_n \geq p_1 \geq 2$, $N_n \geq 3$. In particular it is divisible by at least one prime number, say q.[2] But I claim that N_n is not divisible by p_i for any $1 \leq i \leq n$. Indeed, if $N_n = ap_i$ for some $a \in \mathbb{Z}$, then let $b = \frac{p_1 \cdots p_n}{p_i} \in \mathbb{Z}$. Then $kp_i = p_1 \cdots p_n + 1 =$

[2]Later in these notes we will prove the stronger fact that any integer greater than one may be expressed as a product of primes. For now we assume this (familiar) fact.

$bp_i + 1$, so $(k - b)p_i = 1$ and thus $p_i = \pm 1$, a contradiction. So if we take q to be, for instance, the smallest prime divisor of N_n, then there are at least $n + 1$ prime numbers: $p_1, \ldots, p_n, q$. □

The proof that there are infinitely many prime numbers first appeared in Euclid's *Elements* (Book IX, Proposition 20). Euclid did not explicitly use induction, but in retrospect his proof is clearly an inductive argument: what he does is to explain, as above, how given any finite list $p_1, \ldots, p_n$ of distinct primes, one can produce a new prime which is not on the list. (Euclid *does not* verify the base case; he must have regarded it as obvious that there is at least one prime number.) What is strange is that in our day Euclid's proof is generally *not* seen as a proof by induction. Rather, it is often construed as a proof by contradiction – which it isn't! Rather, Euclid's argument is perfectly constructive. Starting with any given prime number – say $p_1 = 2$ – and following his procedure, one generates an infinite sequence of primes. For instance, $N_1 = 2 + 1 = 3$ is prime, so we take $p_2 = 3$. Then $N_2 = 2 \cdot 3 + 1 = 7$ is again prime, so we take $p_3 = 7$. Then $N_3 = 2 \cdot 3 \cdot 7 + 1 = 43$ is also prime, so we take $p_4 = 43$. But this time something more interesting happens:

$$N_4 = 2 \cdot 3 \cdot 7 \cdot 43 + 1 = 13 \cdot 139$$

is *not* prime. For definiteness let us take p_5 to be the smallest prime factor of N_4, so $p_5 = 13$. In this way we generate an infinite sequence of prime numbers – so the proof is unassailably constructive.

By the way, this sequence of prime numbers is itself rather interesting. It is often called the **Euclid-Mullin sequence**, after Albert A. Mullin who asked questions about it in 1963 [**Mu63**]. The next few terms are

$$53, 5, 6221671, 38709183810571, 139, 2801, 11, 17, 5471, 52662739, 23003,$$

$$30693651606209, 37, 1741, 1313797957, 887, 71, 7127, 109, 23, \ldots.$$

Thus one can see that it is rather far from just giving us all of the prime numbers in increasing order! In fact, since to find p_{n+1} we need to factor $N_n = p_1 \cdots p_n + 1$, a quantity which rapidly increases with n, it is in fact quite difficult to compute the terms of this sequence, and as of 2010 only the first 47 terms are known. Perhaps Mullin's most interesting question about this sequence is: does every prime number appear in it eventually? This is an absolutely open question. At the moment the smallest prime which is not known to appear in the Euclid-Mullin sequence is 31.

Remark: Some scholars have suggested that what is essentially an argument by mathematical induction appears in the later middle Platonic dialogue *Parmenides*, lines 149a7-c3. But this argument is of mostly historical and philosophical interest. The statement in question is, very roughly, that if n objects are placed adjacent to another in a linear fashion, the number of points of contact between them is $n - 1$. (Maybe. To quote wikipedia: "It is widely considered to be one of the more, if not the most, challenging and enigmatic of Plato's dialogues.") There is not much mathematics here! Nevertheless, for a thorough discussion of induction in the *Parmenides* the reader may consult [**Ac00**] and the references cited therein.

3. Closed Form Identities

The inductive proof of Proposition 2.1 is a prototype for a certain kind of induction proof (the easiest kind!) in which $P(n)$ is some algebraic identity: say

$LHS(n) = RHS(n)$. In this case to make the induction proof work you need only (i) establish the base case and (ii) verify the equality of successive differences

$$LHS(n+1) - LHS(n) = RHS(n+1) - RHS(n).$$

We give two more familiar examples of this.

PROPOSITION 2.3. *For all* $n \in \mathbb{Z}^+$, $1 + 3 + \ldots + (2n-1) = n^2$.

PROOF. Let $P(n)$ be the statement "$1+3+\ldots+(2n-1) = n^2$". We will show $P(n)$ holds for all $n \in \mathbb{Z}^+$ by induction on n. Base case ($n = 1$): indeed $1 = 1^2$. Induction step: Let n be an arbitrary positive integer and assume $P(n)$:

$$1 + 3 + \ldots + (2n-1) = n^2. \tag{5}$$

Adding $2(n+1) - 1 = 2n + 1$ to both sides, we get

$$(1 + 3 + \ldots + (2n-1) + 2(n+1) - 1 = n^2 + 2(n+1) - 1 = n^2 + 2n + 1 = (n+1)^2.$$

This is precisely $P(n+1)$, so the induction step is complete. □

PROPOSITION 2.4. *For all* $n \in \mathbb{Z}^+$, $1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$.

PROOF. By induction on n.
Base case: $n = 1$.
Induction step: Let $n \in \mathbb{Z}^+$ and suppose that $1^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$. Then

$$1 + \ldots + n^2 + (n+1)^2 \overset{\text{IH}}{=} \frac{n(n+1)(2n+1)}{6} + (n+1)^2 =$$

$$\frac{2n^3 + 3n^2 + n + 6 + 6n^2 + 12n + 1}{6} = \frac{2n^3 + 9n^2 + 13n + 7}{6}.$$

On the other hand, expanding out $\frac{(n+1)((n+1)+1)(2(n+1)+1)}{6}$, we also get $\frac{2n^3+9n^2+13n+7}{6}$. □

Often a non-inductive proof, when available, offers more insight. Again returning to our archetypical example: $1 + \ldots + n$, it is time to tell the story of little Gauss. As a child of no more than 10 or so, Gauss and his classmates were asked to add up the numbers from 1 to 100. Most of the students did this by a laborious calculation and got incorrect answers in the end. Gauss reasoned essentially as follows: put

$$S_n = 1 + \ldots + (n-1) + n.$$

Of course the sum is unchanged if we we write the terms in descending order:

$$S_n = n + (n-1) + \ldots + 2 + 1.$$

Adding the two equations gives

$$2S_n = (n+1) + (n+1) + \ldots + (n+1) = n(n+1),$$

so

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

This is no doubt preferable to induction, so long as one is clever enough to see it.

Mathematical induction can be viewed as a particular incarnation of a much more general proof technique: try to solve your problem by reducing it to a previously

solved problem. A more straightforward application of this philosophy allows us to deduce Proposition 2.3 from Proposition 2.1:

$$1+3+\ldots+(2n-1) = \sum_{i=1}^{n}(2i-1) = 2\sum_{i=1}^{n} i - \sum_{i=1}^{n} 1 = 2\left(\frac{n(n+1)}{2}\right) - n = n^2+n-n = n^2.$$

4. Inequalities

PROPOSITION 2.5. *For all $n \in N$, $2^n > n$.*

PROOF ANALYSIS For $n \in \mathbb{N}$, let $P(n)$ be the statement "$2^n > n$". We want to show that $P(n)$ holds for all natural numbers n by induction.
Base case: $n = 0$: $2^0 = 1 > 0$.

Induction step: let n be any natural number and assume $P(n)$: $2^n > n$. Then

$$2^{n+1} = 2 \cdot 2^n > 2 \cdot n.$$

We would now like to say that $2n \geq n + 1$. But in fact this is true if and only if $n \geq 1$. Well, don't panic. We just need to restructure the argument a bit: we verify the statement separately for $n = 0$ and then use $n = 1$ as the base case of our induction argument. Here is a formal writeup:

PROOF. Since $2^0 = 1 > 0$ and $2^1 = 2 > 1$, it suffices to verify the statement for all natural numbers $n \geq 2$. We go by induction on n.
Base case: $n = 2$: $2^2 = 4 > 2$.
Induction step: Assume that for some natural number $n \geq 2$, $2^n > n$. Then

$$2^{n+1} = 2 \cdot 2^n > 2n > n + 1.$$

□

PROPOSITION 2.6. *There exists $N_0 \in \mathbb{Z}^+$ such that for all $n \geq N_0$, $2^n \geq n^3$.*

PROOF ANALYSIS A little experimentation shows that there are several small values of n such that $2^n < n^3$: for instance $2^9 = 512 < 9^3 = 729$. On the other hand, it seems to be the case that we can take $N_0 = 10$: let's try.
Base case: $n = 10$: $2^{10} = 1024 > 1000 = 10^3$.

Induction step: Suppose that for some $n \geq 10$ we have $2^n \geq n^3$. Then

$$2^{n+1} = 2 \cdot 2^n \geq 2n^3.$$

Our task is then to show that $2n^3 \geq (n+1)^3$ for all $n \geq 10$. (By considering limits as $n \to \infty$, it is certainly the case that the left hand side exceeds the right hand side for all sufficiently large n. It's not guaranteed to work for $n \geq 10$; if not, we will replace 10 with some larger number.) Now,

$$2n^3 - (n+1)^3 = 2n^3 - n^3 - 3n^2 - 3n - 1 = n^3 - 3n^2 - 3n - 1 \geq 0$$
$$\iff n^3 - 3n^2 - 3n \geq 1.$$

Since everything in sight is a whole number, this is in turn equivalent to

$$n^3 - 3n^2 - 3n > 0.$$

Now $n^3 - 3n^2 - 3n = n(n^2 - 3n - 3)$, so this is equivalent to $n^2 - 3n - 3 \geq 0$. The roots of the polynomial $x^2 - 3x - 3$ are $x = \frac{3 \pm \sqrt{21}}{2}$, so $n^2 - 3n - 3 > 0$ if

$n > 4 = \frac{3+\sqrt{25}}{2} > \frac{3+\sqrt{21}}{2}$. In particular, the desired inequality holds if $n \geq 10$, so by induction we have shown that $2^n \geq n^3$ for all $n \geq 10$.

We leave it to to the student to convert the above analysis into a formal proof.

Remark: More precisely, $2^n \geq n^3$ for all natural numbers n *except* $n = 2, 3, 4, 6, 7, 8, 9$. It is interesting that the desired inequality is true for a little while (i.e., at $n = 0, 1$) then becomes false for a little while longer, and then becomes true for all $n \geq 10$. Note that it follows from our analysis that if for any $N \geq 4$ we have $2^N \geq N^3$, then this equality remains true for all larger natural numbers n. Thus from the fact that $2^9 < 9^3$, we can in fact deduce that $2^n < n^3$ for all $4 \leq n \leq 8$.

PROPOSITION 2.7. *For all* $n \in \mathbb{Z}^+$, $1 + \frac{1}{4} + \ldots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$.

PROOF ANALYSIS By induction on n.
Base case ($n = 1$): $1 \leq 2 - \frac{1}{1}$.
Induction step: Assume that for some $n \in \mathbb{Z}^+$, $1 + \frac{1}{4} + \ldots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$. Then

$$1 + \frac{1}{4} + \ldots + \frac{1}{n^2} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2}.$$

We want the left hand side to be less than $2 - \frac{1}{n+1}$, so it will suffice to show

$$2 - \frac{1}{n} + \frac{1}{(n+1)^2} < 2 - \frac{1}{n+1}.$$

Equivalently, it suffices to show

$$\frac{1}{n+1} + \frac{1}{(n+1)^2} < \frac{1}{n}.$$

But we have

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

Everything in sight is positive, so by clearing denominators, the desired inequality is equivalent to

$$n^2 + 2n = n(n+2) < (n+1)^2 = n^2 + 2n + 1,$$

which is true! Thus we have all the ingredients of an induction proof, but again we need to put things together in proper order, a task which we leave to the reader.

Remark: Taking limits as $n \to \infty$, it follows that $\sum_{n=1}^{\infty} \frac{1}{n^2} \leq 2$. In particular, this argument shows that the infinite series converges. The exact value of the sum is, in fact, $\frac{\pi^2}{6}$. A proof of this requires techniques from advanced calculus.

5. Extending Binary Properties to n-ary Properties

Example: All horses have the same color.

Proposed proof: There are only finitely many horses in the world, so it will suffice to show that for all $n \in \mathbb{Z}^+$, $P(n)$ holds, where $P(n)$ is the statement that in any set of n horses, all of them have the same color.

Base case: In any set S of one horse, all of the horses in S have the same color!
Induction step: We suppose that for some positive integer n, in any set of n horses,

all have the same color. Consider now a set of $n+1$ horses, which for specificity we label $H_1,\ H_2, \ldots, H_n,\ H_{n+1}$. Now we can split this into two sets of n horses:

$$S = \{H_1, \ldots,\ H_n\}$$

and

$$T = \{H_2, \ldots, H_n,\ H_{n+1}\}.$$

By induction, every horse in S has the same color as H_1: in particular H_n has the same color as H_1. Similarly, every horse in T has the same color as H_n: in particular H_{n+1} has the same color as H_n. But this means that $H_2, \ldots, H_n, H_{n+1}$ all have the same color as H_1. It follows by induction that for all $n \in \mathbb{Z}^+$, in any set of n horses, all have the same color.

Proof analysis: Naturally one suspects that there is a mistake somewhere, and there is. However it is subtle, and occurs in a perhaps unexpected place. In fact the argument is completely correct, except the induction step is not valid when $n = 1$: in this case $S = \{H_1\}$ and $T = \{H_2\}$ and these two sets are disjoint: they have no horses in common. We have been misled by the "dot dot dot" notation which suggests, erroneously, that S and T must have more than one element.

In fact, if only we could establish the argument for $n = 2$, then the proof goes through just fine. For instance, the result can be fixed as follows: if in a finite set of horses, any two have the same color, then they all have the same color.

There is a moral here: one should pay especially close attention to the smallest values of n to make sure that the argument has no gaps. On the other hand, there is a certain type of induction proof for which the $n = 2$ case is the most important (often it is also the base case, but not always), and the induction step is easy to show, but uses once again the $n = 2$ case. Here are some examples of this.

The following is a fundamental fact of number theory, called **Euclid's Lemma**.

Proposition 2.8. *Let p be a prime, and let $a, b \in \mathbb{Z}^+$. If $p \mid ab$, $p \mid a$ or $p \mid b$.*

Later in this chapter we will give a proof (yes, by induction!). Let's assume it for now. Then we can swiftly deduce the following useful generalization.

Proposition 2.9. *Let p be a prime number, $n \in \mathbb{Z}^+$ and $a_1, \ldots, a_n \in \mathbb{Z}^+$. If $p \mid a_1 \cdots a_n$, then $p \mid a_i$ for some $1 \leq i \leq n$.*

Proof. This is trivial for $n = 1$. We show it for all $n \geq 2$ by induction.
Base case: $n = 2$: This is precisely Euclid's Lemma.
Induction Step: We assume that for a given $n \in \mathbb{Z}^+$ and $a_1, \ldots, a_n \in \mathbb{Z}^+$, if a prime p divides the product $a_1 \cdots a_n$, then it divides at least one a_i. Let $a_1, \ldots, a_n, a_{n+1} \in \mathbb{Z}$, and that a prime p divides $a_1 \cdots a_n a_{n+1}$. Then $p \mid (a_1 \cdots a_n)a_{n+1}$, so by Euclid's Lemma, $p \mid a_1 \cdots a_n$ or $p \mid a_{n+1}$. If the latter, we're done. If the former, then by our inductive hypothesis, $p \mid a_i$ for some $1 \leq i \leq n$, so we are also done. □

Corollary 2.10. *Let p be a prime, and let $a, n \in \mathbb{Z}^+$. Then $p \mid a^n \implies p \mid a$.*

Exercise 2.2. *Use Corollary 2.10 to show that for any prime p, $p^{\frac{1}{n}}$ is irrational.*

6. The Principle of Strong/Complete Induction

Problem: A sequence is defined recursively by $a_1 = 1$, $a_2 = 2$ and $a_n = 3a_{n-1} - 2a_{n-2}$. Find a general formula for a_n and prove it by induction.

Proof analysis: Unless we know something better, we may as well examine the first few terms of the sequence and hope that a pattern jumps out at us. We have

$$a_3 = 3a_2 - 2a_1 = 3 \cdot 2 - 2 \cdot 1 = 4.$$

$$a_4 = 3a_3 - 2a_2 = 3 \cdot 4 - 2 \cdot 2 = 8.$$

$$a_5 = 3a_4 - 2a_3 = 3 \cdot 8 - 2 \cdot 4 = 16.$$

$$a_6 = 3a_5 - 2a_4 = 3 \cdot 16 - 2 \cdot 8 = 32.$$

The evident guess is therefore $a_n = 2^{n-1}$. Now a key point: it is not possible to prove this formula using the version of mathematical induction we currently have. Indeed, let's try: assume that $a_n = 2^{n-1}$. Then

$$a_{n+1} = 3a_n - 2a_{n-1}.$$

By the induction hypothesis we can replace a_n with 2^{n-1}, getting

$$a_{n+1} = 3 \cdot 2^{n-1} - 2a_{n-1};$$

now what?? A little bit of thought indicates that we think $a_{n-1} = 2^{n-2}$. If for some reason it were logically permissible to make that substitution, then we'd be in good shape:

$$a_{n+1} = 3 \cdot 2^{n-1} - 2 \cdot 2^{n-2} = 3 \cdot 2^{n-1} - 2^{n-1} = 2 \cdot 2^{n-1} = 2^n = 2^{(n+1)-1},$$

which is what we wanted to show. Evidently this goes a bit beyond the type of induction we have seen so far: in addition to assuming the truth of a statement $P(n)$ and using it to prove $P(n+1)$, we also want to assume the truth of $P(n-1)$.

There is a version of induction that allows this, and more:

Principle of Strong/Complete Induction:
Let $P(n)$ be a sentence with domain the positive integers. Suppose:
(i) $P(1)$ is true, and
(ii) For all $n \in \mathbb{Z}^+$, if $P(1), \ldots, P(n-1), P(n)$ are all true, then $P(n+1)$ is true.
Then $P(n)$ is true for all $n \in \mathbb{Z}^+$.

Thus, in a nutshell, strong/complete induction allows us to assume not only the truth of our statement for a single value of n in order to prove it for the next value $n+1$, but rather allows us to assume the truth of the statement for all positive integer values less than $n+1$ in order to prove it for $n+1$.

It is easy to see that PS/CI implies the usual principle of mathematical induction. The logical form of this is simply[3]

$$(A \implies C) \implies (A \wedge B \implies C).$$

[3]The symbol $\wedge$ denotes logical conjunction: in other words, $A \wedge B$ means "A and B".

In other words, if one can deduce statement C from statement A, then one can also deduce statement C from A together with some additional hypothesis or hypotheses B. Specifically, we can take A to be $P(n)$, C to be $P(n+1)$ and B to be $P(1) \wedge P(2) \wedge \ldots \wedge P(n-1)$.

Less obviously, one can use our previous PMI to prove PS/CI. The proof is not hard but slightly tricky. Suppose we know PMI and wish to prove PS/CI. Let $P(n)$ be a sentence with domain the positive integers and satisfying (i) and (ii) above. We wish to show that $P(n)$ holds for all $n \in \mathbb{Z}^+$, using only ordinary induction. The trick is to introduce a new predicate $Q(n)$, namely

$$Q(n) = P(1) \wedge P(2) \wedge \ldots \wedge P(n).$$

Notice that $Q(1) = P(1)$; (ii) above tells us that $Q(n) \implies P(n+1)$. But if we know $Q(n) = P(1) \wedge \ldots \wedge P(n)$ and also $P(n+1)$, then we know $P(1) \wedge \ldots \wedge P(n) \wedge P(n+1) = Q(n+1)$. So $Q(1)$ holds and for all n, $Q(n) \implies Q(n+1)$. So by PMI, $Q(n)$ holds for all n, hence certainly $P(n)$ holds for all n.

Exercise 6: As for ordinary induction, there is a variant of strong/complete induction where instead of starting at 1 we start at any integer N_0. State this explicitly.

Here is an application which makes full use of the "strength" of PS/CI.

PROPOSITION 2.11. *Let $n > 1$ be an integer. Then there exist prime numbers $p_1, \ldots, p_k$ (for some $k \geq 1$) such that $n = p_1 \cdots p_k$.*

PROOF. We go by strong induction on n.
Base case: $n = 2$. Indeed 2 is prime, so we're good.
Induction step: Let $n > 2$ be any integer and assume that the statement is true for all integers $2 \leq k < n$. We wish to show that it is true for n.
Case 1: n is prime. As above, we're good.
Case 2: n is not prime. By definition, this means that there exist integers a, b, with $1 < a, b < n$, such that $n = ab$. But now our induction hypothesis applies to both a and b: we can write $a = p_1 \cdots p_k$ and $b = q_1 \cdots q_l$, where the p_i's and q_j's are all prime numbers. But then

$$n = ab = p_1 \cdots p_k q_1 \cdots q_l$$

is an expression of n as a product of prime numbers: done! □

This is a good example of the use of induction (of one kind or another) to give a very clean proof of a result whose truth was not really in doubt but for which a more straightforward proof is wordier and messier.

7. Solving Homogeneous Linear Recurrences

Recall our motivating problem for PS/CI: we were given a sequence defined by $a_1 = 1$, $a_2 = 2$, and for all $n \geq 1$, $a_n = 3a_{n-1} - 2a_{n-2}$. By trial and error we guessed that $a_n = 2^{n-1}$, and this was easily confirmed using PS/CI.

But this was very lucky (or worse: the example was constructed so as to be easy to solve). In general, it might not be so obvious what the answer is, and as above, this is induction's Kryptonite: it has no help to offer in guessing the answer.

Example: Suppose a sequence is defined by $x_0 = 2$, $x_n = 5x_{n-1} - 3$ for all $n \geq 1$.

Here the first few terms of the sequence are $x_1 = 7$, $x_2 = 32$, $x_3 = 157$, $x_4 = 782$, $x_5 = 3907$. What's the pattern? At least to me, it's not evident.

This is a case where more generality brings clarity: it is often easier to detect a pattern involving algebraic expressions than a pattern involving integers. So suppose we have $a, b, c \in \mathbb{R}$, and we define a sequence recursively by

$$x_0 = c;\ \forall n \in \mathbb{N},\ x_{n+1} = ax_n + b.$$

Now let's try again:

$$x_1 = ax_0 + b = ac + b.$$
$$x_2 = ax_1 + b = a(ac + b) + b = ca^2 + ba + b.$$
$$x_3 = ax_2 + b = a(ca^2 + ba + b) + b = ca^3 + ba^2 + ba + b.$$
$$x_4 = ax_3 + b = a(ca^3 + ba^2 + ba + b) + b = ca^4 + ba^3 + ba^2 + ba + b.$$

Aha: it seems that we have for all $n \geq 1$.

$$x_n = ca^n + ba^{n-1} + \ldots + ba + b.$$

Now we have something that induction can help us with: it is true for $n = 1$. Assuming it is true for n, we calculate

$$x_{n+1} = ax_n + b \stackrel{IH}{=} a(ca^n + ba^{n-1} + \ldots + ba + b) + b) = ca^{n+1} + ba^n + \cdots + ba^2 + ba + b,$$

so the desired expression is correct for all n. Indeed, we can simplify it:

$$x_n = ca^n + b\sum_{i=1}^{n} a_i = ca^n + b\left(\frac{a^{n+1} - 1}{a - 1}\right) = \frac{(ab + ac - c)a^n - b}{a - 1}.$$

In particular the sequence x_n grows exponentially in n.

Let us try our hand on a sequence defined by a two-term recurrence:

$$F_1 = F_2 = 1;\ \forall n \geq 1, F_{n+2} = F_{n+1} + F_n.$$

The F_n's are the famous **Fibonacci numbers**. Again we list some values:

$$F_3 = 2,\ F_4 = 3,\ F_5 = 5,\ F_6 = 8,\ F_7 = 13,\ F_8 = 21,\ F_9 = 34,\ F_{10} = 55,$$
$$F_{11} = 89,\ F_{12} = 144,\ F_{13} = 233,\ F_{14} = 377,\ F_{15} = 377,$$
$$F_{200} = 280571172992510140037611932413038677189525,$$
$$F_{201} = 453973694165307953197296969697410619233826.$$

These computations suggest F_n grows exponentially. Taking ratios of successive values suggests that the base of the exponential lies between 1 and 2, e.g.

$$\frac{F_{201}}{F_{200}} = 1.618033988749894848204586834\ldots.$$

Cognoscenti may recognize this as the decimal expansion of the **golden ratio**

$$\varphi = \frac{1 + \sqrt{5}}{2}.$$

However, let's consider a more general problem and make a vaguer guess. Namely, for real numbers b, c we consider an recurrence of the form

$$x_1 = A_1, x_2 = A_2, \ \forall n \geq 1, x_{n+2} = bx_{n+1} + cx_n. \tag{6}$$

In all the cases we've looked at the solution was (roughly) exponential. So let's **guess** an exponential solution $x_n = Cr^n$ and plug this into the recurrence; we get

$$Cr^{n+2} = x_{n+2} = b(Cr^{n+1}) + c(Cr^n),$$

which simplifies to

$$r^2 - br - cr = 0.$$

Evidently the solutions to this are

$$r = \frac{b \pm \sqrt{b^2 + 4c}}{2}.$$

Some cases to be concerned about are the case $c = \frac{-b^2}{4}$, in which case we have only a single root $r = \frac{b}{2}$, and the case $c < \frac{-b^2}{4}$ in which case the roots are complex numbers. But for the moment let's look at the Fibonacci case: $b = c = 1$. Then $r = \frac{1 \pm \sqrt{5}}{2}$. So we recover the golden ratio $\varphi = \frac{1+\sqrt{5}}{2}$ – a good sign! – as well as

$$\frac{1 - \sqrt{5}}{2} = 1 - \varphi = -.618033988749894848204586834\ldots.$$

So we have two different bases – what do we do with that? A little thought shows that if r_1^n and r_2^n are both solutions to the recurrence $x_{n+2} = bx_{n+1}cx_n$ (with any initial conditions), then so is $C_1 r_1^n + C_2 r_2^n$ for any constants C_1 and C_2. Therefore we propose $x_n = C_1 r_1^n + C_2 r_2^n$ as the **general solution** to the two-term homogeneous linear recurrence (6) and the two initial conditions $x_1 = A_1$, $x_2 = A_2$ provide just enough information to solve for C_1 and C_2.

Trying this for the Fibonacci sequence, we get

$$1 = F_1 = C_1\varphi + C_2(1 - \varphi).$$

$$1 = F_2 = C_1(\varphi)^2 + C_2(1 - \varphi)^2.$$

Multiplying the first equation by φ and subtracting it from the second equation will give us a linear equation to solve for C_2, and then we plug the solution into either of the equations and solve for C_1. It turns out that

$$C_1 = \frac{1}{\sqrt{5}}, \ C_2 = \frac{-1}{\sqrt{5}}.$$

INTERLUDE: This is easily said and indeed involves only high school algebra. But we can do something slicker. Instead of determining the constants by evaluating F_n at $n = 1$ and $n = 2$, it's easier to evaluate at $n = 1$ and $n = 0$: then we have

$$F_0 = C_1\varphi^0 + C_2(1 - \varphi)^0 = C_1 + C_2.$$

But for this to work we need to know F_0, which we have not defined. Can it be defined in a sensible way? Yes! Writing the basic recurrence in the form $F_{n+1} = F_n + F_{n-1}$ and solving for F_{n-1} gives:

$$F_{n-1} = F_{n+1} - F_n.$$

This allows us to define F_n for all integers n. In particular, we have

$$F_0 = F_2 - F_1 = 1 - 1 = 0.$$

Thus we get

$$0 = C_1 + C_2,$$

whereas plugging in $n = 1$ gives

$$1 = C_1(\varphi) + C_2(1-\varphi) = C_1(\varphi) - C_1(1-\varphi) = (2\varphi - 1)C_1,$$

$$C_1 = \frac{1}{2\varphi - 1} = \frac{1}{2\left(\frac{1+\sqrt{5}}{2}\right) - 1} = \frac{1}{\sqrt{5}},\ C_2 = \frac{-1}{\sqrt{5}}.$$

Now we are ready to prove the following result.

THEOREM 2.12. *(Binet's Formula) For any $n \in \mathbb{Z}$, the nth Fibonacci number is*

$$F_n = \frac{1}{\sqrt{5}}\left(\varphi^n - (1-\varphi)^n\right),$$

where $\varphi = \frac{1+\sqrt{5}}{2}$.

PROOF. We go by strong/complete induction on n. The base cases are $n = 1$ and $n = 2$, but we have already checked these: we used them to determine the constants C_1 and C_2. So now assume that $n \geq 3$ and that the formula is correct for all positive integers smaller than $n + 2$. Then, using the identities

$$\varphi^2 = \varphi + 1,$$

$$(1-\varphi) = -\varphi^{-1},$$

$$1 - \varphi^{-1} = \varphi^{-2} = (-\varphi)^{-2},$$

we compute

$$F_{n+2} = F_{n+1} + F_n = \frac{1}{\sqrt{5}}(\varphi^{n+1} + \varphi^n - (1-\varphi)^{n+1} - (1-\varphi)^n))$$

$$= \frac{1}{\sqrt{5}}(\varphi^n(\varphi+1) - (1-\varphi)^n(1-\varphi+1) =$$

$$\frac{1}{\sqrt{5}}(\varphi^n(\varphi^2) - (-\varphi)^{-n}((-\varphi)^{-1}+1)$$

$$= \frac{1}{\sqrt{5}}(\varphi^{n+2} - (-\varphi)^{-n}(-\varphi)^{-2}) = \frac{1}{\sqrt{5}}(\varphi^{n+2} - (-\varphi)^{-(n+2)}) = \frac{1}{\sqrt{5}}(\varphi^{n+2} - (1-\varphi)^{n+2}).$$

□

Exercise 7: Find all $n \in \mathbb{Z}$ such that $F_n < 0$.

By the way, it is not quite true that any solution to (6) must have exponential growth. For instance, consider the recurrence

$$x_1 = 1,\ x_2 = 2;\ \forall n \geq 1, x_{n+2} = 2x_{n+1} - x_n.$$

Then

$$x_3 = 2x_2 - x_1 = 2 \cdot 2 - 1 = 3,\ x_4 = 2x_3 - x_2 = 2 \cdot 3 - 2 = 4,\ x_5 = 2 \cdot 4 - 3 = 5.$$

It certainly looks as though $x_n = n$ for all n. Indeed, assuming it to be true for all positive integers smaller than $n+2$, we easily check

$$x_{n+2} = 2x_{n+1} - x_n = 2(n+1) - n = 2n + 2 - n = n + 2.$$

The characteristic polynomial is $r^2 - 2r + 1 = (r-1)^2$: it has repeated roots. One solution is $C_1 1^n = C_1$ (i.e., x_n is a constant sequence). This occurs iff $x_2 = x_1$, so clearly there are nonconstant solutions as well. It turns out that in general, if the characteristic polynomial is $(x-r)^2$, then the two basic solutions are $x_n = r^n$ and also $x_n = nr^n$. It is unfortunately harder to guess this in advance, but it is not hard to check that this gives a solution to a recurrence of the form $x_{n+2} = 2r_0 x_{n+1} - r_0^2 x_n$.

These considerations will be eerily familiar to the reader who has studied differential equations. For a more systematic exposition on "discrete analogues" of calculus concepts (with applications to the determination of power sums as in §3), see [**DC**].

8. The Well-Ordering Principle

There is yet another form of mathematical induction that can be used to give what is, arguably, an even more elegant proof of Proposition 2.11.

THEOREM 2.13. *(Well-Ordering Principle) Let S be a nonempty subset of $\mathbb{Z}^+$. Then S has a least element, i.e., there exists $s \in S$ such that for all $t \in S$, $s \leq t$.*

Intutitively, the statement is true by the following reasoning: first we ask: is $1 \in S$? If so, it is certainly the least element of S. If not, we ask: is $2 \in S$? If so, it is certainly the least element of S. And then we continue in this way: if we eventually get a "yes" answer then we have found our least element. But if for every n the answer to the question "Is n an element of S?" is negative, then S is empty!

The well-ordering principle (henceforth **WOP**) is often useful in its contrapositive form: if a subset $S \subset \mathbb{Z}^+$ does *not* have a least element, then $S = \varnothing$.

We CLAIM WOP is *logically equivalent* to the principle of mathematical induction (PMI) and thus also to the principle of strong/complete induction (PS/CI).

First we will assume PS/CI and show that WOP follows. For this, observe that WOP holds iff $P(n)$ holds for all $n \in \mathbb{Z}^+$, where $P(n)$ is the following statement:

$P(n)$: If $S \subset \mathbb{Z}^+$ and $n \in S$, then S has a least element.

Indeed, if $P(n)$ holds for all n and $S \subset \mathbb{Z}$ is nonempty, then it contains some positive integer n, and then we can apply $P(n)$ to see that S has a least element. Now we can prove that $P(n)$ holds for all n by complete induction: first, if $1 \in S$, then indeed 1 is the least element of S, so $P(1)$ is certainly true. Now assume $P(k)$ for all $1 \leq k \leq n$, and suppose that $n+1 \in S$. If $n+1$ is the least element of S, then we're done. If it isn't, then it means that there exists $k \in S$, $1 \leq k \leq S$. Since we have assumed $P(k)$ is true, therefore there exists a least element of S.

Conversely, let us assume WOP and prove PMI. Namely, let $S \subset \mathbb{Z}$ and suppose that $1 \in S$, and that for all n, if $n \in S$ then $n+1 \in S$. We wish to show that $S = \mathbb{Z}^+$. Equivalently, putting $T = \mathbb{Z}^+ \setminus S$, we wish to show that $T = \emptyset$. If not,

then by WOP T has a least element, say n. Reasoning this out gives an immediate contradiction: first, $n \notin S$. By assumption, $1 \in S$, so we must have $n > 1$, so that we can write $n = m + 1$ for some $m \in \mathbb{Z}^+$. Further, since n is the least element of T we must have $n - 1 = m \in S$, but now our inductive assumption implies that $n + 1 = n \in S$, contradiction.

So now we have shown that PMI $\iff$ PS/CI $\implies$ WOP $\implies$ PMI.

Let us give another proof of Proposition 2.11 using WOP. We wish to show that every integer $n > 1$ can be factored into primes. Let S be the set of integers $n > 1$ which *cannot* be factored into primes. Seeking a contradiction, we assume S is nonempty. In that case, by WOP it has a least element, say n. Now n is certainly not prime, since otherwise it can be factored into primes. So we must have $n = ab$ with $1 < a, b < n$. But now, since a and b are integers greater than 1 which are smaller than the least element of S, they must each have prime factorizations, say $a = p_1 \cdots p_k$, $b = q_1 \cdots q_l$. But then (stop me if you've heard this one before)

$$n = ab = p_1 \cdots p_k q_1 \cdots q_l$$

itself can be expressed as a product of primes, contradicting our assumption. therefore S is empty: every integer greater than 1 is a product of primes.

This kind of argument is often called proof by **minimum counterexample**.

These two proofs of Proposition 2.11 are very close: the difference between a proof by PS/CI and a proof by WOP is more a matter of taste than technique.

9. The Fundamental Theorem of Arithmetic

9.1. Euclid's Lemma and the Fundamental Theorem of Arithmetic.

The following are the two most important theorems in beginning number theory.

THEOREM 2.14. *(Euclid's Lemma) Let p be a prime number and a, b be positive Suppose that $p \mid ab$. Then $p \mid a$ or $p \mid b$.*

THEOREM 2.15. *(Fundamental Theorem of Arithmetic) The factorization of any integer $n > 1$ into primes is unique, up to the order of the factors: suppose*

$$n = p_1 \cdots p_k = q_1 \cdots q_l,$$

are two factorizations of n into primes, with $p_1 \leq \ldots \leq p_k$ and $q_1 \leq \ldots \leq q_l$. Then $k = l$ and $p_i = q_i$ for all $1 \leq i \leq k$.

A prime factorization $n = p_1 \cdots p_k$ is in **standard form** if $p_1 \leq \ldots \leq p_k$. Every prime factorization can be put in standard form by ordering the primes from least to greatest. Dealing with standard form factorizations is a convenient bookkeeping device, since otherwise our uniqueness statement would have to include the proviso "up to the order of the factors."

Given Proposition 2.11 – i.e., the existence of prime factorizations – Theorems 2.14 and 15.13 are equivalent: each can be easily deduced from the other.

EL implies FTA: Assume Euclid's Lemma. As seen above, this implies Proposition 2.9: if a prime divides any finite product of integers it must divide one of the factors. Our proof will be by minimal counterexample: suppose that there are some integers greater than one which factor into primes in more than one way, and let n be the least such integer, so

$$n = p_1 \cdots p_k = q_1 \cdots q_l, \tag{7}$$

where each of the primes is written in nonincreasing order. Evidently $p_1 \mid n = q_1 \cdots q_l$, so by Proposition 2.9, we must have that $p_1 \mid q_j$ for some $1 \leq j \leq l$. But since q_j is also prime, this means that $p_1 = q_j$. Therefore we can cancel them from the expression, getting

$$\frac{n}{p_1} = p_2 \cdots p_k = q_1 \cdots q_{j-1} q_{j+1} \cdots q_l. \tag{8}$$

But $\frac{n}{p_1}$ is less than the least integer which has two different factorizations into primes, so it must have a unique factorization into primes, meaning that the primes on the left hand side of (8) are equal, in order, to the primes on the right hand side of (8). This also implies that $p_1 = q_j$ is less than or equal to all the primes appearing on the right hand side, so $j = 1$. Thus $k = l$, $p_1 = q_j = q_1$ and $p_i = q_i$ for $2 \leq i \leq j$. This means that in (7) the two factorizations are the same after all!

FTA implies EL: Assume every integer greater than one factors *uniquely* into a product of primes, let p be a prime, and let a and b be positive integers such that $p \mid ab$. If either a or b is 1, then the other is just p and the conclusion is clear, so we may assume $a, b > 1$ and therefore have unique prime factorizations

$$a = p_1 \cdots p_r, \ b = q_1 \cdots q_s;$$

our assumption that p divides ab means $ab = kp$ for some $k \in \mathbb{Z}^+$ and thus

$$ab = p_1 \cdots p_r q_1 \cdots q_s = kp.$$

The right hand side of this equation shows that p must appear in the prime factorization of ab. Since the prime factorization is unique, we must have at least one p_i or at least one q_j equal to p. In the first case p divides a; in the second case p divides b.

The traditional route to FTA is via Euclid's Lemma, and the traditional route to Euclid's Lemma (Euclid's route in *Elements*) is via a series of intermediate steps including the **Euclidean algorithm** and finding the set of all integer solutions to equations of the form $ax+by = 1$. This takes some time to develop – perhaps a week in an elementary number theory course. But we can bypass all these intermediate steps and give direct inductive proofs of both EL and FTA. And we will.

9.2. Rogers' Inductive Proof of Euclid's Lemma.

Here is a proof of Euclid's Lemma using WOP, following K. Rogers [**Ro63**].

Seeking a contradiction, we suppose there is at least one prime such that Euclid's Lemma does not hold for that prime. By WOP there is a *least* such prime, say p, so there are $a, b \in \mathbb{Z}^+$ with $p \mid ab$ but $p \nmid a$ and $p \nmid b$. By WOP there is a least $a \in \mathbb{Z}^+$ such that there is at least one $b \in \mathbb{Z}^+$ with $p \mid ab$ and $p \nmid a$, $p \nmid b$.

Now consider the following equation:

$$ab = (a - p)b + pb,$$

which shows that $p \mid ab \iff p \mid (a-p)b$. There are three cases:
Case 1: $a-p$ is a positive integer. Then, since $0 < a-p < a$ and a was by assumption the *least* positive integer such that Euclid's Lemma fails for the prime p, we must have that $p \mid a-p$ or $p \mid b$. By assumption $p \nmid b$, so we must have $p \mid a-p$, but then $p \mid (a-p)+p = a$, contradiction!
Case 2: We have $a = p$. But then $p \mid a$, contradiction.
Case 3: We have $a < p$. On the other hand, $a > 1$ – if $p \mid 1 \cdot b$, then $p \mid b$ – so by Proposition 2.11 a is divisible by some prime q, and $q \mid a < p$, so $q < p$. Therefore q is a prime which is smaller than the least prime for which Euclid's Lemma fails, so Euclid's Lemma holds for q. Since $p \mid ab$, we may write $pk = ab$ for some $k \in \mathbb{Z}^+$, and now $q \mid a \implies q \mid ab = pk$, so by Euclid's Lemma for q, $q \mid p$ or $q \mid k$. The first case is impossible since p is prime and $1 < q < p$, so we must have $q \mid k$. Therefore $p\left(\frac{k}{q}\right) = \left(\frac{a}{q}\right)b$, so $p \mid \frac{a}{q}b$. But $1 < \frac{a}{q} < a$ and a is the *least* positive integer for which Euclid's Lemma fails for p and a, so it must be that $p \mid \frac{a}{q}$ (so in particular $p \mid a$) or $p \mid b$. Contradiction. So Euclid's Lemma holds for all primes p.

9.3. The Lindemann-Zermelo Inductive Proof of FTA.

Here is a proof of FTA using WOP, following Lindemann [**Li33**] and Zermelo [**Ze34**].

We claim that the standard form factorization of a positive integer is unique. Assume not; then the set of positive integers which have at least two different standard form factorizations is nonempty, so has a least element, say n, where:

$$n = p_1 \cdots p_r = q_1 \cdots q_s. \tag{9}$$

Here the p_i's and q_j's are prime numbers, not necessarily distinct from each other. However, $p_1 \neq q_j$ for any j. Indeed, if we had such an equality, then after relabelling the q_j's we could assume $p_1 = q_1$ and then divide through by $p_1 = q_1$ to get a smaller positive integer $\frac{n}{p_1}$. By the assumed minimality of n, the prime factorization of $\frac{n}{p_1}$ must be unique: i.e., $r-1 = s-1$ and $p_i = q_i$ for all $2 \leq i \leq r$. Then multiplying by $p_1 = q_1$ we see that we didn't have two different factorizations after all.

In particular $p_1 \neq q_1$. Without loss of generality, assume $p_1 < q_1$. Then, if we subtract $p_1 q_2 \cdots q_s$ from both sides of (9), we get

$$m := n - p_1 q_2 \cdots q_s = p_1(p_2 \cdots p_r - q_2 \cdots q_s) = (q_1 - p_1)(q_2 \cdots q_s). \tag{10}$$

Evidently $0 < m < n$, so by minimality of n, the prime factorization of m must be unique. However, (10) gives two different factorizations of m, and we can use these to get a contradiction. Specifically, $m = p_1(p_2 \cdots p_r - q_2 \cdots q_s)$ shows that $p_1 \mid m$. Therefore, when we factor $m = (q_1 - p_1)(q_2 \cdots q_s)$ into primes, at least one of the prime factors must be p_1. But $q_2, \ldots, q_j$ are already primes which are different from p_1, so the only way we could get a p_1 factor is if $p_1 \mid (q_1 - p_1)$. But this implies $p_1 \mid q_1$, and since q_1 is also prime this implies $p_1 = q_1$. Contradiction!

9.4. A Generalized Euclid's Lemma.

THEOREM 2.16. *(Generalized Euclid's Lemma) Let $a, b, c \in \mathbb{Z}^+$. Suppose that $x \mid yz$ and x and y are relatively prime. Then $x \mid z$.*

Exercise: a) Prove Theorem 2.16 using the Fundamental Theorem of Arithmetic.
b) Explain why Theorem 2.16 is indeed a generalization of Euclid's Lemma.

CHAPTER 3

Polynomial and Rational Functions

1. Polynomial Functions

Using the basic operations of addition, subtraction, multiplication, division and composition of functions, we can combine very simple functions to build large and interesting (and useful!) classes of functions. For us, the two simplest kinds of functions are the following:

Constant functions: for each $a \in \mathbb{R}$ there is a function $C_a : \mathbb{R} \to \mathbb{R}$ such that for all $x \in \mathbb{R}$, $C_a(x) = a$. In other words, the output of the function does not depend on the input: whatever we put in, the same value a will come out. The graph of such a function is the horizontal line $y = a$. Such functions are called **constant**.

The identity function $I : \mathbb{R} \to \mathbb{R}$ by $I(x) = x$. The graph of the identity function is the straight line $y = x$.

Recall that the identity function is so-called because it is an identity element for the operation of function composition: that is, for any function $f : \mathbb{R} \to \mathbb{R}$ we have $I \circ f = f \circ I = f$.

Example: Let $m, b \in \mathbb{R}$, and consider the function $L : \mathbb{R} \to \mathbb{R}$ by $x \mapsto mx + b$. Then L is built up out of constant functions and the identity function by addition and multiplication: $L = C_m \cdot I + C_b$.

Example: Let $n \in \mathbb{Z}^+$. The function $m_n : x \mapsto x^n$ is built up out of the identity function by repreated multiplication: $m_n = I \cdot I \cdots I$ (n I's altogether).

The general name for a function $f : \mathbb{R} \to \mathbb{R}$ which is built up out of the identity function and the constant functions by finitely many additions and multiplications is a **polynomial**. In other words, every polynomial function is of the form

$$f : x \mapsto a_n x^n + \ldots + a_1 x + a_0 \tag{11}$$

for some constants $a_0, \ldots, a_n \in \mathbb{R}$.

However, we also want to take – at least until we prove it doesn't make a difference – a more algebraic approach to polynomials. Let us define a **polynomial expression** as an expression of the form $\sum_{i=0}^{n} a_i x^i$. Thus, to give a polynomial expression we need to give for each natural number i a constant a_i, while requiring that all but finitely many of these constants are equal to zero: i.e., there exists some $n \in \mathbb{N}$ such that $a_i = 0$ for all $i > n$.

Then every polynomial expression $f = \sum_{i=0}^{n} a_i x^i$ determines a **polynomial function** $x \mapsto f(x)$. But it is at least conceivable that two different-looking polynomial expressions give rise to the *same function*. To give some rough idea of what I mean here, consider the two expressions $f = 2\arcsin x + 2\arccos x$ and $g = \pi$. Now it turns out for all $x \in [-1, 1]$ (the common domain of the arcsin and arccos functions) we have $f(x) = \pi$. (The angle θ whose sine is x is complementary to the angle φ whose cosine is x, so $\arcsin x + \arccos x = \theta + \varphi = \frac{\pi}{2}$.) But still f and g are given by different "expressions": if I ask you what the coefficient of $\arcsin x$ is in the expression f, you will immediately tell me it is 2. If I ask you what the coefficient of $\arcsin x$ is in the expression π, you will have no idea what I'm talking about.

One special polynomial expression is the **zero polynomial**. This is the polynomial whose ith coefficient a_i is equal to zero for all $i \geq 0$.

Every nonzero polynomial expression has a **degree**, which is a natural number, the largest natural number i such that the coefficient a_i of x^i is nonzero. Thus in (11) the degree of f is n *if and only if* $a_n \neq 0$: otherwise the degree is smaller than n.

Although the zero polynomial expression does not have any natural number as a degree, it is extremely convenient to regard $\deg 0$ as *negative*, i.e., such that $\deg 0$ is smaller than the degree of any nonzero polynomial. This means that for any $d \in \mathbb{N}$ "the set of polynomials of degree at most d" includes the zero polynomial. We will follow this convention here but try not to make too big a deal of it.

Let us give some examples to solidify this important concept:

The polynomials of degree at most 0 are the expressions $f = a_0$. The corresponding functions are all constant functions: their graphs are horizontal lines. (The graph of the zero polynomial is still a horizontal line, $y = 0$, so it is useful to include the zero polynomial as having "degree at most zero".)

The polynomials of degree *at most one* are the linear expressions $L = mx + b$. The corresponding functions are linear functions: their graphs are straight lines. The degree of $L(x)$ is one if $m \neq 0$ – i.e., if the line is not horizontal – and 0 if $m = 0$ and $b \neq 0$.

Similarly the polynomials of degree *at most two* are the quadratic expressions $q(x) = ax^2 + bc + c$. The degree of q is 2 unless $a = 0$.

We often denote the degree of the polynomial expression f by $\deg f$.

THEOREM 3.1. *Let f, g be nonzero polynomial expressions.*
a) If $f + g \neq 0$, then $\deg(f + g) \leq \max(\deg f, \deg g)$.
b) We have $\deg(fg) = \deg f + \deg g$.

PROOF. a) Suppose that

$$f(x) = a_m x^m + \ldots + a_1 x + a_0, \ a_m \neq 0$$

and

$$g(x) = b_n x^n + \ldots + b_1 x + b_0, \ b_n \neq 0$$

so that $\deg g = m,\ \deg g = n$.
Case 1: $m > n$. Then when we add f and g, the highest order term will be $a_m x^m$, since the polynomial g only smaller powers of x. In particular the degree of $f + g$ is $m = \max(m, n)$.
Case 2: $m < n$. Similarly, when we add f and g, the highest order term will be $a_n x^n$, so the degree of $f + g$ is $n = \max(m, n)$.
Case 3: Suppose $m = n$. Then

$$(f + g)(x) = (a_m + b_m)x^m + \ldots + (a_1 + b_1)x + (a_0 + b_0).$$

Thus the degree of $f + g$ is *at most* m. It will be exactly m unless $a_m + b_m = 0$, i.e., unless $b_m = -a_m$; in this case it will be strictly smaller than m.
b) If f and g are as above, then the leading term of $f \cdot g$ will be $a_m b_n x^{m+n}$, and since $a_m, b_n \neq 0$, $a_m b_n \neq 0$. Thus $\deg fg = m + n$. □

Exercise: For polynomial expressions $f = \sum_{i=0}^m a_i x^i$ and $g = \sum_{j=0}^n b_j x^j$, define

$$(f \circ g)(x) = \sum_{i=0}^m a_i (\sum_{j=0}^n b_j x^j)^n.$$

Show that $\deg(f \circ g) = (\deg f)(\deg g)$.

The following is the most important algebraic property of polynomials.

THEOREM 3.2. *(Polynomial Division With Remainder) Let $a(x)$ be a polynomial expression and $b(x)$ be a nonzero polynomial expression. There are* unique *polynomial expressions $q(x)$ and $r(x)$ such that*
(i) $a(x) = q(x)b(x) + r(x)$ and
(ii) $\deg r(x) < \deg b(x)$.

PROOF. First note that we really get to choose only $q(x)$, for then $r(x) = a(x) - q(x)b(x)$. Also let us denote the leading coefficient of $a(x)$ by α and the leading coefficient of $b(x)$ by β. Step 1: We prove the *existence* of $q(x)$ and $b(x)$.
Case 1: $\deg a(x) < \deg b(x)$. Take $q(x) = 0$: $\deg r(x) = \deg a(x) < \deg b(x)$.
Case 2: $\deg a(x) \geq deg b(x)$. Take $q_1(x) = \frac{\alpha}{\beta} x^{\deg a(x) - \deg b(x)}$. The point of this is that $q_1(x)b(x)$ has degee $\deg a(x) - \deg b(x) + \deg b(x) = \deg a(x)$ and leading coefficient $\frac{\alpha}{\beta} \cdot \beta = \alpha$, so in $r_1(x) = a(x) - q_1(x)b(x)$ the leading terms cancel and $\deg r_1(x) \leq \deg(a(x)) - 1$. If $\deg r_1(x) < \deg b(x)$ then we're done, take $q(x) = q_1(x)$ and $r(x) = r_1(x)$. On the other hand, if $\deg r_1(x) \geq \deg b(x)$ we apply the process again with $r_1(x)$ in place of $a(x)$: let α_1 be the leading coefficient of $r_1(x)$ and take $q_2(x) = \frac{\alpha_1}{\beta} x^{\deg r_1(x) - \deg b(x)}$, so that in $r_2(x) = r_1(x) - q_2(x)b(x)$ the leading terms cancel to give $\deg r_2(x) \leq \deg(r_1(x)) - 1 \leq \deg(a(x)) - 2$. Continue in this way generating polynomial expressions $q_i(x)$ and $r_i(x)$ until we reach a k with $\deg r_k(x) < \deg b(x)$. Then:

$$\begin{aligned} r_k(x) &= r_{k-1}(x) - q_k(x)b(x) \\ &= r_{k-2}(x) - q_{k-1}(x)b(x) - q_k b(x) \\ &= r_{k-3}(x) - q_{k-2}(x)b(x) - q_{k-1}(x)b(x) - q_k b(x) \\ &\vdots \\ &= a(x) - (q_1(x) + q_2(x) + \ldots q_k(x))b(x). \end{aligned}$$

Thus we may take $q(x) = q_1(x) + \ldots + q_k(x)$, so that $r(x) = a(x) - q(x)b(x) = r_k(x)$ and $\deg r(x) = \deg r_k(x) < \deg b(x)$.
Step 2: We prove the *uniqueness* of $q(x)$ (and thus of $r(x) = a(x) - q(x)b(x)$). Suppose $Q(x)$ is another polynomial such that $R(x) = a(x) - Q(x)b(x)$ has degree less than the degree of $b(x)$. Then

$$a(x) = q(x)b(x) + r(x) = Q(x)b(x) + R(x),$$

so

$$(q(x) - Q(x))b(x) = R(x) - r(x).$$

Since $r(x)$ and $R(x)$ both have degree less than $\deg b(x)$, so does $r(x) - R(x)$, so

$$\deg b(x) > \deg(R(x) - r(x)) = \deg((q(x) - Q(x))b(x)) = \deg(q(x) - Q(x)) + \deg b(x).$$

Thus $\deg(q(x) - Q(x)) < 0$, and the only polynomial with negative degree is the zero polynomial, i.e., $q(x) = Q(x)$ and thus $r(x) = R(x)$.[1] □

Exercise: Convince yourself that the proof of Step 1 is really a careful, abstract description of the standard high school procedure for long division of polynomials.

Theorem 3.2 has many important and useful consequences; here are some of them.

THEOREM 3.3. *(Root-Factor Theorem) Let $f(x)$ be a polynomial expression and c a real number. The following are equivalent:*
*(i) $f(c) = 0$. ("c is a **root** of f.")*
*(ii) There is some polynomial expression q such that as polynomial expressions, $f(x) = (x - c)q(x)$. ("$x - c$ is a **factor** of f.")*

PROOF. We apply the Division Theorem with $a(x) = f(x)$ and $b(x) = x - c$, getting polynomials $q(x)$ and $r(x)$ such that

$$f(x) = (x - c)q(x) + r(x)$$

and $r(x)$ is either the zero polynomial or has $\deg r < \deg x - c = 1$. In other words, $r(x)$ is in all cases a constant polynomial (perhaps constantly zero), and its constant value can be determined by plugging in $x = c$:

$$f(c) = (c - c)q(c) + r(c) = r(c).$$

The converse is easier: if $f(x) = (x - c)q(x)$, then $f(c) = (c - c)q(c) = 0$. □

COROLLARY 3.4. *Let f be a nonzero polynomial of degree n. Then the corresponding polynomial function f has at most n real roots: i.e., there are at most n real numbers a such that $f(a) = 0$.*

PROOF. By induction on n.
Base case ($n = 0$): If $\deg f(x) = 0$, then f is a nonzero constant, so has no roots.
Induction Step: Let $n \in \mathbb{N}$, suppose that every polynomial of degree n has at most n real roots, and let $f(x)$ be a polynomial of degree $n + 1$. If $f(x)$ has no real root, great. Otherwise, there exists $a \in \mathbb{R}$ such that $f(a) = 0$, and by the Root-Factor Theorem we may write $f(x) = (x - a)g(x)$. Moreover by Theorem 3.1, we have $n + 1 = \deg f = \deg(x - a)g(x) = \deg(x - a) + \deg g = 1 + \deg g$, so $\deg g = n$. Therefore our induction hypothesis applies and $g(x)$ has m distinct real

[1] If you don't like the convention that the zero polynomial has negative degree, then you can phrase the argument as follows: *if* $q(x) - Q(x)$ were a nonzero polynomial, these degree considerations would give the absurd conclusion $\deg(q(x) - Q(x)) < 0$, so $q(x) - Q(x) = 0$.

roots $a_1, \ldots, a_m$ for some $0 \leq m \leq n$. Then f has either $m+1$ real roots – if a is distinct from all the roots a_i of g – or m real roots – if $a = a_i$ for some i, so it has at most $m+1 \leq n+1$ real roots. □

LEMMA 3.5. *Let $f = \sum_{i=0}^n a_i x^i$ be a polynomial expression. Suppose that the function $f(x) = \sum_{i=0}^n a_i x^i$ is the zero function: $f(x) = 0$ for all $x \in \mathbb{R}$. Then $a_i = 0$ for all i, i.e., f is the zero polynomial expression.*

PROOF. Suppose that f is not the zero polynomial, i.e., $a_i \neq 0$ for some i. Then it has a degree $n \in \mathbb{N}$, so by Corollary 3.4 there are at most n real numbers c such that $f(c) = 0$. But this is absurd: $f(x)$ is the zero function, so for all (infinitely many!) real numbers c we have $f(c) = 0$. □

THEOREM 3.6. *(Uniqueness Theorem For Polynomials) Let*

$$f = a_n x^n + \ldots + a_1 x + a_0,$$

$$g = b_n x^n + \ldots + b_1 x + b_0$$

be two polynomial expressions. The following are equivalent:
(i) f and g are equal as polynomial expressions: for all $0 \leq i \leq n$, $a_i = b_i$.
(ii) f and g are equal as polynomial functions: for all $c \in \mathbb{R}$, $f(c) = g(c)$.
(iii) There are $c_1 < c_2 < \ldots < c_{n+1}$ such that $f(c_i) = g(c_i)$ for $1 \leq i \leq n+1$.

PROOF. (i) $\implies$ (ii): This is clear, since if $a_i = b_i$ for all i then f and g are being given by the same expression, so they must give the same function.
(ii) $\implies$ (iii): This is also immediate: since $f(c) = g(c)$ for all real numbers c, we may take for instance $c_1 = 1$, $c_2 = 2, \ldots, c_{n+1} = n+1$.
(iii) $\implies$ (i): Consider the polynomial expression

$$h = f - g = (a_n - b_n)x^n + \ldots + (a_1 - b_1)x + (a_0 - b_0).$$

Then $h(c_1) = f(c_1) - g(c_1) = 0, \ldots, h(c_{n+1}) = f(c_{n+1}) - g(c_{n+1}) = 0$. So h is a polynomial of degree at most n which has (at least) $n+1$ distinct real roots. By Corollary 3.4, h must be the zero polynomial expression: that is, for all $0 \leq i \leq n$, $a_i - b_i = 0$. Equivalently, $a_i = b_i$ for all $0 \leq i \leq n$, so f and g are equal as polynomial expressions. □

In particular, Theorem 3.6 says that if two polynomials $f(x)$ and $g(x)$ *look different* – i.e., they are not coefficient-by-coefficient the same expression – then they are actually different functions, i.e., there is some $c \in \mathbb{R}$ such that $f(c) \neq g(c)$.

Finally, we want to prove an *arithmetic* result about polynomials, the **Rational Roots Theorem**. For this we need another number-theoretic preliminary. Two positive integers a and b are **coprime** (or **relatively prime**) if they are not both divisible by any integer $d > 1$ (equivalently, they have no common prime factor).

THEOREM 3.7. *(Rational Roots Theorem) Let $a_0, \ldots, a_n$ be integers, with $a_0, a_n \neq 0$. Consider the polynomial*

$$P(x) = a_n x^n + \ldots + a_1 x + a_0.$$

Suppose that $\frac{b}{c}$ is a rational number, written in lowest terms, which is a root of P: $P(\frac{b}{c}) = 0$. Then a_0 is divisible by b and a_n is divisible by c.

PROOF. We know

$$0 = P\left(\frac{b}{c}\right) = a_n\frac{b^n}{c^n} + \ldots + a_1\frac{b}{c} + a_0.$$

Multiplying through by c^n clears denominators, giving

$$a_n b^n + a_{n-1}b^{n-1}c + \ldots + a_1 bc^{n-1} + a_0 c^n = 0.$$

Rewriting this equation as

$$a_n b^n = c(-a_{n-1}b^{n-1} - \ldots - a_0 c^{n-1})$$

shows that $a_n b^n$ is divisible by c. But since b and c have no prime factors in common and b^n has the same distinct prime factors as does b, b^n and c have no prime factors in common and are thus coprime. So Theorem 2.16 applies to show that a_n is divisible by c. Similarly, rewriting the equation as

$$a_0 c^n = b(-a_n b^{n-1} - a_{n-1}b^{n-2}c - \ldots - a_1 c^{n-1})$$

shows that $a_0 c^n$ is divisible by b. As above, since b and c are coprime, so are b and c^n, so by Theorem 2.16 a_0 is divisible by b. □

In high school algebra the Rational Roots Theorem is often used to generate a finite list of possible rational roots of a polynomial with integer coefficients. This is nice, but there are more impressive applications. For instance, taking $a_n = 1$ and noting that 1 is divisible by c iff $c = \pm 1$ we get the following result.

COROLLARY 3.8. *Let $a_0, \ldots, a_{n-1} \in \mathbb{Z}$, and consider the polynomial*

$$P(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0.$$

Suppose c is a rational number such that $P(c) = 0$. Then c is an integer.

So what? Let p be any prime number, let $n \geq 2$, and consider the polynomial

$$P(x) = x^n - p.$$

By Corollary 3.8, if $c \in \mathbb{Q}$ is such that $P(c) = 0$, then $c \in \mathbb{Z}$. But if $c \in \mathbb{Z}$ is such that $P(c) = 0$, then $c^n = p$. But this means that the *prime* number p is divisible by the integer c, so $c = \pm 1$ or $c = \pm p$. But $(\pm 1)^n = \pm 1$ and $(\pm p)^n = \pm p^n$, so $c^n = p$ is impossible. So there is no $c \in \mathbb{Q}$ such that $c^n = p$: that is, $\sqrt[n]{p}$ is irrational.

This is a doubly infinite generalization of our first irrationality proof, that $\sqrt{2}$ is irrational, but the argument is, if anything, shorter and easier to understand. (However, we did use – and prove, by induction – the Fundamental Theorem of Arithmetic, a tool which was not available to us at the very beginning of the course.) Moral: polynomials can be useful in surprising ways!

PROPOSITION 3.9. *For every polynomial P of positive degree there are irreducible polynomials $p_1, \ldots, p_k$ such that*

$$P = p_1 \cdots p_k.$$

Exercise: Prove it. (Suggestion: adapt the proof that every integer $n > 1$ can be factored into a product of primes.)

2. Rational Functions

A rational function is a function which is a quotient of two polynomial functions: $f(x) = \frac{P(x)}{Q(x)}$, with $Q(x)$ not the zero polynomial. To be sure, Q is allowed to have roots; it is just not allowed to be zero at all points of $\mathbb{R}$.

The "natural domain" of $\frac{P(x)}{Q(x)}$ is the set of real numbers for which $Q(x) \neq 0$, i.e., all but a finite (possibly empty) set of points.

2.1. The Partial Fractions Decomposition.

The goal of this section is to derive the Partial Fractions Decomposition (PFD) for a proper rational function. The PFD is a purely algebraic fact which is useful in certain analytic contexts: especially, it is the key technique used to find explicit antiderivatives of rational functions.

THEOREM 3.10. *Let P_1, P_2, P be polynomials, with P_1 monic irreducible and P_2 not divisible by P_1. Then there are polynomials A, B such that*

$$AP_1 + BP_2 = P.$$

PROOF. Let $\mathcal{S}$ be the set of nonzero polynomials of the form $aP_1 + bP_2$ for some polynomials a, b. Since $P_1 = 1 \cdot P_1 + 0 \cdot P_2, P_2 = 0 \cdot P_1 + 1 \cdot P_2$ are in $\mathcal{S}$, $\mathcal{S} \neq \varnothing$. By the Well Ordering Principle there is $m \in \mathcal{S}$ of minimal degree: write

$$aP_1 + bP_2 = m.$$

If a polynomial lies in $\mathcal{S}$ then so does every nonzero constant multiple of it, so we may assume m is monic. We claim that for any $M = AP_1 + BP_2 \in \mathcal{S}$, we must have $m \mid M$. To see this, apply Polynomial Division to M and m, getting

$$M = Qm + R, \ \deg R < \deg m.$$

Since also

$$R = M - Qm = (A - Qa)P_1 + (B - Qb)P_2,$$

so if $R \neq 0$ we'd have $R \in \mathcal{S}$ of smaller degree than m, contradiction. Thus $R = 0$ and $m \mid M$. Since $P_1 \in \mathcal{S}$, $m \mid P_1$. Because P_1 is monic irreducible, $m = P_1$ or $m = 1$. Since $P_2 \in \mathcal{S}$, also $m \mid P_2$, and so by hypothesis $m \neq P_1$. Thus $m = 1$ and

$$aP_1 + bP_2 = 1.$$

Multiplying through by P we get

$$(aP)P_1 + (bP)P_2 = P,$$

so we may take $A = aP$, $B = bP$. □

COROLLARY 3.11. *(Euclid's Lemma For Polynomials) Let p be an irreducible polynomia, and suppose $p \mid P_1P_2$. Then $p \mid P_1$ or $p \mid P_2$.*

PROOF. As is usual in trying to prove statements of the form "A implies (B or C)", we assume that B is false and show that C is true. Here this means assuming $p \nmid P_1$. By Theorem 3.10, there are polynomials A and B such that

$$Ap + BP_1 = 1.$$

Multiplying through by P_2 gives

$$ApP_2 + BP_1P_2 = P_2.$$

Since $p \mid ApP_2$ and $p \mid BP_1P_2$, $p \mid (Ap_2 + BP_1P_2) = P_2$. □

Exercise: Show that every monic polynomial of positive degree factors *uniquely* into a product of monic irreducible polynomials.

COROLLARY 3.12. *Let $n \in \mathbb{Z}^+$, and let p, P, Q_0 be polynomials, such that p is irreducible and that p does not divide Q_0.*
a) There are polynomials B, A such that we have a rational function identity

$$\frac{P(x)}{p(x)^n Q_0(x)} = \frac{B(x)}{p(x)^n} + \frac{A(x)}{p(x)^{n-1}Q_0(x)}. \tag{12}$$

b) If $\deg P < \deg(p^n Q_0)$, we may choose A and B such that

$$\deg A < \deg(p^{n-1}Q_0),\ \deg B < \deg p^n.$$

PROOF. a) Apply Theorem 3.10a) with $P_1 = p$, $P_2 = Q_0$: we get

$$P = A_1 p + BQ_0. \tag{13}$$

for some polynomials A_1, B. Thus

$$\frac{P}{p^n Q_0} = \frac{A_1 p + BQ_0}{p^n Q_0} = \frac{B}{p^n} + \frac{A_1}{p^{n-1}Q_0}.$$

b) By Polynomial Division (Theorem 3.2), there are C, r such that $B = Cp + r$ and $\deg r < \deg p$. Substituting in (13) and putting $A = A_1 + CQ_0$ we get

$$P = A_1 p + BQ_0 = A_1 p + (Cp + r)Q_0 = (A_1 + CQ_0)p + rq_0 = Ap + rQ_0.$$

We have

$$\deg(rQ_0) = \deg r + \deg Q_0 < \deg p + \deg Q_0 \leq \deg p^n + \deg Q_0 = \deg(p^n Q_0).$$

Since also $\deg P < \deg(p^n Q_0)$ and $Ap = P - rQ_0$, it follows that

$$\deg A + \deg p = \deg(Ap) < \deg(p^n Q_0) = \deg Q_0 + n \deg p,$$

so

$$\deg A \leq \deg Q_0 + (n-1)\deg p = \deg(p^{n-1}Q_0).$$ □

The identity of Corollary 3.12 can be applied repeatedly: suppose we start with a proper rational function $\frac{P}{Q}$, with Q a monic polynomial (as is no loss of generality; we can absorb the leading coefficient into P). Then Q is a polynomial of positive degree, so we may factor it as

$$Q = p_1^{a_1} \cdots p_r^{a_r},$$

where $p_1, \ldots, p_r$ are distinct monic irreducible polynomials. Let us put $Q_0 = p_2^{a_2} \cdots p_r^{a_r}$, so $Q = p_1^{a_1} Q_0$. Applying Corollary 3.12, we may write

$$\frac{P}{Q} = \frac{B_{a_1}}{p_1^{a_1}} + \frac{A_{a-1}}{p_1^{n-1}Q_0}$$

with $\deg B_{a_1} < \deg p_1$ and $\deg A_{a_1} < \deg(p_1^{n-1}Q_0)$. Thus Corollary 3.12 applies to $\frac{A_{a_1}}{p_1^{a_1-1}Q_0}$, as well, giving us overall

$$\frac{P}{Q} = \frac{B_{a_1}}{p_1^{a_1}} + \frac{B_{a_1-1}}{p_1^{a_1-1}} + \frac{A_{a_1-1}}{p_1^{n-2}Q_0}.$$

And we may continue in this way until we get

$$\frac{P}{Q} = \frac{B_{a_1}}{p_1^{a_1}} + \ldots + \frac{B_1}{p_1} + \frac{A_0}{Q_0},$$

with each $\deg B_\cdot < \deg P_1$ and $\deg A_0 < \deg Q_0$. Recalling that $Q_0 = p_2^{a_2} \cdots p_r^{a_r}$, we may put $Q_0 = p_2^{a_2} Q_1$ and continue on. Finally we get an identity of the form

$$\frac{P}{p_1^{a_1} \cdots p_r^{a_r}} = \frac{B_{1,1}}{p_1^{a_1}} + \ldots + \frac{B_{1,a_1}}{p_1} + \ldots + \frac{B_{r,1}}{p_r^{a_r}} + \ldots + \frac{B_{r,a_r}}{p_r}, \tag{14}$$

with $\deg B_{i,j} < \deg p_i$ for all i and j. From an algebraic perspective, (14) is the **Partial Fractions Decomposition**. In fact, though it is not the point for us, everything that we have done works for polynomials with coefficients in *any field* K – e.g. $K = \mathbb{Q}$, $K = \mathbb{C}$, $K = \mathbb{Z}/p\mathbb{Z}$. Conversely, when we work with polynomials over the real numbers, the expression simplifies, because the classification of irreducible polynomials over the real numbers is rather simple. Indeed:

LEMMA 3.13. *The irreducible (real!) polynomials are precisely:*
(i) The linear polynomials $\ell(x) = a(x - c)$ for $a, c \in \mathbb{R}$, $a \neq 0$; and
(ii) The quadratic polynomisls $Q(x) = ax^2 + bx + c$ for $a, b, c \in \mathbb{R}$, $b^2 - 4ac < 0$.

Every linear polynomial is irreducible. Further, from the Rational Roots Theorem, for a polynomial of degree greater than 1 to be irreducible, it cannot have any real roots. For quadratic (and, for the record, cubic) polynomials this is also sufficient, since a nontrivial factorization of a degree 2 or 3 polynomial necessarily has a linear factor and thus a real root. It is a a consequence of the Quadratic Formula that a real quadratic polynomial $Q(x) = ax^2 + bx + c$ has no real roots iff its discriminant $b^2 - 4ac$ is negative. Thus half of Lemma 3.13 – the half that says that a linear polynomials and quadratic polynomial without real roots are irreducible – is easy.

The other half of Lemma 3.13 – the half that says that there are no other irreducible polynomials – is far from easy: in fact it lies well beyond our current means. In several hundred pages' time we will return to prove this result as a consequence of the **Fundamental Theorem of Algebra**: see Theorem 15.17.

Combining (14) and Lemma 3.13 we get the desired result.

THEOREM 3.14. *(Real Partial Fractions Decomposition) Let $\frac{P(x)}{Q(x)}$ be a proper rational function, with*

$$Q(x) = (x - c_1)^{m_1} \cdots (x - c_k)^{m_k} q_1(x)^{n_1} \cdots q_l(x)^{n_l};$$

here $c_1, \ldots, c_k \in \mathbb{R}$, with $a \neq 0$, $c_1, \ldots, c_k$ distinct; and $q_1, \ldots, q_l$ are distinct monic irreducible quadratics. There are real numbers $A_{\cdot,\cdot}, B_{\cdot,\cdot}, C_{\cdot,\cdot}$ such that we have an identity of rational functions

$$\frac{P(x)}{Q(x)} = \sum_{i=1}^{m_1} \frac{A_{1,i}}{(x - c_1)^i} + \ldots + \sum_{i=1}^{m_k} \frac{A_{k,i}}{(x - c_k)^i} + \sum_{j=1}^{n_1} \frac{B_{1,j}x + C_{1,j}}{q_1(x)^j} + \ldots + \sum_{j=1}^{n_l} \frac{B_{l,j}x + C_{l,j}}{q_l(x)^j}. \tag{15}$$

Exercise: Show that the constants in (15) are unique.

CHAPTER 4

Continuity and Limits

1. Remarks on the Early History of the Calculus

We have seen that in order to define the derivative f' of a function $f : \mathbb{R} \to \mathbb{R}$ we need to understand the notion of a *limit* of a function at a point. It turns out that giving a mathematically rigorous and workable definition of a limit is hard – really hard. Let us begin with a quick historical survey.

It is generally agreed that calculus was invented (discovered?) independently by Isaac Newton and Gottfried Wilhelm von Leibniz, roughly in the 1670's. Leibniz was the first to publish on calculus, in 1685. However Newton probably could have published his work on calculus before Leibniz, but held it back for various reasons.[1]

To say that "calculus was discovered by Newton and Leibniz" is an oversimplification. Computations of areas and volumes which we can now recognize as using calculus concepts go back to ancient Egypt, if not earlier. The Greek mathematicians **Eudoxus** (408-355 BCE) and **Archimedes** (287-212 BCE) developed the **method of exhaustion**, a limiting process which anticipates integral calculus. Also Chinese and Indian mathematicians made significant achievements. Even in "modern" Europe, Newton and Leibniz were not functioning in a complete intellectual vacuum. They were responding to and continuing earlier work by **Pierre de Fermat** (on tangent lines) and **John Wallis**, **Isaac Barrow** and **James Gregory**. This should not be surprising: all scientific work builds on work of others. But the accomplishments of Newton and Leibniz were so significant that after their efforts calculus existed as a systematic body of work, whereas before them it did not.

How did Newton and Leibniz construe the fundamental concept, namely that of a limit? Both of their efforts were far from satisfactory, indeed far from making sense. Newton's limiting concept was based on a notion of **fluxions**, which is so obscure as not to be worth our time to describe it. Leibniz, a philosopher and writer as well as a mathematician, addressed the matter more thoroughly and came up with the notion of an **infinitesimal quantity**, a number which is not zero but "vanishingly small", i.e., smaller than any "ordinary" positive number.

The concept of infinitesimals *has* been taken up by mathematicians since Leibniz, and eventually with complete mathematical success...but not until the 1960s![2] For instance one has a definition of an infinitesimal element x of an ordered field K, namely an element x which is positive but smaller than $\frac{1}{n}$ for all $n \in \mathbb{Z}^+$. It is

[1] I highly recommend James Gleick's biography of Newton. If I wanted to distill hundreds of pages of information about his personality into one word, the word I would choose is...**weirdo**.

[2] See http://en.wikipedia.org/wiki/Nonstandard_analysis for an overview of this story.

easy to see that an ordered field admits infinitesimal elements iff it does *not* satisfy the Archimedean axiom, whereas the real numbers $\mathbb{R}$ do satisfy the Archimedean axiom. So at best Lebiniz was advocating a limiting process based on a different mathematical model of the real numbers than the "standard" modern one. And at worst, Lebiniz's writing on infinitesimals seems like equivocation: at different stages of a calculation the same quantity is at one point "vanishingly small" and at another point not. The calculus of both fluxions and infnitesimals required, among other things, some goodwill: if you used them as Newton and Leibniz did in their calculations then at the end you would get a sensible (in fact, correct!) answer. But if you wanted to make trouble and ask why infinitesimals could not be manipulated in other ways which swiftly led to contradictions, it was all too easy to do so.

The calculus of Newton and Leibniz had a famous early critic, **Bishop George Berkeley**. In 1734 he published *The Analyst*, subtitled "A DISCOURSE Addressed to an Infidel MATHEMATICIAN. WHEREIN It is examined whether the Object, Principles, and Inferences of the modern Analysis are more distinctly conceived, or more evidently deduced, than Religious Mysteries and Points of Faith." Famously, Berkeley described fluxions as *the ghosts of departed quantities*. I haven't read Berkeley's text, but from what I am told it displays a remarkable amount of mathematical sophistication and most of its criticisms are essentially valid!

So if the mid 17th century is the birth of systematic calculus it is *not* the birth of a satisfactory treatment of the limiting concept. When did this come? More than 150 years later! The modern definition of limits via inequalities was given by Bolzano in 1817 (but not widely read), in a somewhat imprecise form by Cauchy in his influential 1821 text, and then finally by Weierstrass around 1850.

2. Derivatives Without a Careful Definition of Limits

Example 2.1: Let $f(x) = mx + b$ be a linear function. Then f has the following property: for any $x_1 \neq x_2$, secant line between the two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ *is* the line $y = f(x)$. Indeed, the slope of the secant line is

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{mx_2 + b - (mx_1 + b)}{x_2 - x_1} = \frac{m(x_2 - x_1)}{x_2 - x_1} = m.$$

Thus the secant line has slope m and passes through the point $(x_1, f(x_1))$, as does the linear function f. But there is a unique line passing through a given point with a given slope, so that the secant line must be $y = mx + b$.

Using this, it is now not at all difficult to compute the derivative of a linear function...assuming an innocuous fact about limits.

Example 2.2: Let $f(x) = mx + b$. Then

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} m(x+h) + b - (mx + b)h = \lim_{h \to 0} \frac{mh}{h} = \lim_{h \to 0} m.$$

The above computation is no surprise, since we already saw that the slope of any secant line to a linear function $y = mx + b$ is just m. So now we need to evaluate the limiting slope of the secant lines. But surely if the slope of every secant line

is m, the desired limiting slope is also m, and thus $f'(x) = m$ (constant function). Let us record the fact about limits we used.

FACT 4.1. *The limit of a constant function $f(x) = C$ as x approaches a is C.*

Example 2.3: Let $f(x) = x^2$. Then

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}$$

$$= \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h} = \lim_{h \to 0} \frac{h(2x+h)}{h} = \lim_{h \to 0} 2x + h.$$

Now Leibniz would argue as follows: in computing the limit, we want to take h infinitesimally small. Therefore $2x + h$ is infinitesimally close to $2x$, and so in the limit the value is $2x$. Thus

$$f'(x) = 2x.$$

But these are just words. A simpler and equally accurate description of what we have done is as follows: we simplified the difference quotient $\frac{f(x+h)-f(x)}{h}$ until we got an expression in which it made good sense to plug in $h = 0$, and then we plugged in $h = 0$. If you wanted to give a freshman calculus student practical instructions on how to compute derivatives of reasonably simple functions directly from the definition, I think you couldn't do much better than this!

Example 2.4: $f(x) = x^3$. Then

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{(x+h)^3 - x^3}{h} = \lim_{h \to 0} \frac{x^3 + 3x^2h + 3xh^2 + h^3 - x^3}{h}$$

$$= \lim_{h \to 0} \frac{h(3x^2 + 3xh + h^2)}{h} = \lim_{h \to 0} 3x^2 + 3xh + h^2.$$

Again we have simplified to the point where we may meaningfully set $h = 0$, getting

$$f'(x) = 3x^2.$$

Example 2.5: For $n \in \mathbb{Z}^+$, let $f(x) = x^n$. Then

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{(x+h)^n - x^n}{h} = \lim_{h \to 0} \frac{\sum_{i=0}^n \binom{n}{i} x^{n-i} h^i - x^n}{h}$$

$$= \lim_{h \to 0} \frac{h \sum_{i=1}^n \binom{n}{i} x^{n-i} h^{i-i}}{h} = \lim_{h \to 0} \binom{n}{1} x^{n-1} + h \sum_{i=2}^n \binom{n}{i} x^{n-i} h^{i-2}$$

$$= \binom{n}{1} x^{n-1} = nx^{n-1}.$$

At this point we have seen many examples of a very pleasant algebraic phenomenon. Namely, for $y = f(x)$ a polynomial function, when we compute the difference quotient $\frac{f(x+h)-f(x)}{h}$ we find that the numerator, say $G(h) = f(x+h) - f(x)$, always has h as a factor: thus we can write it as $G(h) = hg(h)$, where $g(h)$ is another polynomial in h. This is exactly what we need in order to compute the derivative, because when this happens we get

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{hg(h)}{h} = \lim_{h \to 0} g(h) = g(0).$$

Can we guarantee that this factorization $f(x+h) - f(x) = G(h) = hg(h)$ always takes place? Yes, we can! By the Root-Factor Theorem, we may factor out $h = h-0$ from the polynomial $G(h)$ iff $G(0) = 0$. But $G(0) = f(x+0) - f(x) = f(x) - f(x) = 0$. Thus some simple polynomial algebra ensures that we will be able to compute the derivative of any polynomial function, provided we assume the following fact.

FACT 4.2. *For any polynomial function g,*

$$\lim_{h \to a} g(h) = g(a).$$

(This generalizes our first fact above, since constant functions are polynomials.)

Differentiating polynomial functions directly from the definition is, evidently, somewhat tedious. Perhaps we can establish some techniques to streamline the process? For instance, suppose we know the derivative of some function f: what can we say about the derivative of $cf(x)$, where c is some real number? Let's see:

$$\lim_{h \to 0} \frac{cf(x+h) - cf(x)}{h} = \lim_{h \to 0} c \left(\frac{f(x+h) - f(x)}{h} \right).$$

If we assume $\lim_{x \to a} cf(x) = c \lim_{x \to a} f(x)$, then we can complete this computation: the derivative of $cf(x)$ is $cf'(x)$. Let us again record our assumption about limits.

FACT 4.3. *If $\lim_{x \to a} f(x) = L$, then $\lim_{x \to a} cf(x) = cL$.*

This tells for instance that the derivative of $17x^{10}$ is $17(10x^9) = 170x^9$. More generally, this tells us that the derivative of the general **monomial** cx^n is cnx^{n-1}. Now what about sums?

Let f and g be two differentiable functions. Then the derivative of $f + g$ is

$$\lim_{h \to 0} \frac{f(x+h) + g(x+h) - f(x) - g(x)}{h} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} + \frac{g(x+h) - g(x)}{h}.$$

If we assume the limit of a sum is the sum of limits, we get

$$(f+g)' = f' + g'.$$

Again, let's record what we've used.

FACT 4.4. *If $\lim_{x \to a} f(x) = L$ and $\lim_{x \to a} g(x) = M$, then $\lim_{x \to}(f+g)(x) = L + M$.*

Exercise 2.6: Show by mathematical induction that if $f_1, \ldots, f_n$ are functions with derivatives $f_1', \ldots, f_n'$, then $(f_1 + \ldots + f_n)' = f_1' + \ldots + f_n'$.

Putting these facts together, we get an expression for the derivative of any polynomial function: if $f(x) = a_n x^n + \ldots + a_1 x + a_0$, then $f'(x) = na_n x^{n-1} + \ldots + a_1$. In particular the derivative of a degree n polynomial is a polynomial of degree $n-1$ (and the derivative of a constant polynomial is the zero polynomial).

3. Limits in Terms of Continuity

We have been dancing around two fundmaental issues in our provisional treatment of derivatives. The first is, of course, the notion of the limit of a function at a point. The second, just as important, is that of **continuity** at a point.

In freshman calculus it is traditional to define continuity in terms of limits. A

true fact which is not often mentioned is that this works just as well the other way around: treating the concept of a continuous function as known, one can define limits in terms of it. Since I think most people have at least some vague notion of what a continuous function is – *very roughly* it is that the graph $y = f(x)$ is a nice, unbroken curve – and I know all too well that many students have zero intuition for limits, it seems to be of some value to define limits in terms of continuity.

Let $f : \mathbb{R} \to \mathbb{R}$ be a function. For any $x \in \mathbb{R}$, f may or may not be continuous *at* x. We say f is simply **continuous** if it is continuous at x for every $x \in \mathbb{R}$.[3]u

Here are some basic and useful properties of continuous functions. (Of course we cannot prove them until we give a formal definition of continuous functionsu.)

FACT 4.5. *a) Every constant function is continuous at every $c \in \mathbb{R}$.*
b) The identity function $I(x) = x$ is continuous at every $c \in \mathbb{R}$.
c) If f and g are continuous at $x = c$, then $f + g$ and $f \cdot g$ are continuous at $x = c$.
d) If f is continuous at $x = c$ and $f(c) \neq 0$, then $\frac{1}{f}$ is continuous at $x = c$.
e) If f is continuous at $x = c$ and g is continuous at $x = f(c)$, then $g \circ f$ is continuous at $x = c$.

From this relatively small number of facts many other facts follow. For instance, since polynomials are built up out of the identity function and the constant functions by repeated addition and multiplication, it follows that all polynomials are continuous at every $c \in \mathbb{R}$. Similarly, every rational function $\frac{f}{g}$ is continuous at every c in its domain, i.e., at all points c such that $g(c) \neq 0$.

We now wish to define the limit of a function f at a point $c \in \mathbb{R}$. Here it is crucial to remark that c need not be in the domain of f. Rather what we need is that f is defined on some **deleted interval** $I_{c,\delta}$ about c: that is, there is some $\delta > 0$ such that all points in $(c - \delta, c + \delta)$ *except possibly at* c, f is defined. To see that this a necessary business, consider the basic limit defining the derivative:

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

Here x is fixed and we are thinking of the difference quotient as a function of h. Note though that this function is not defined at $h = 0$: the denominator is equal to zero. In fact what we are trying to when differentiating is to find the most reasonable extension of the right hand side to a function which is defined at 0. This brings us to the following definition.

Let f be a real-valued function defined on some subset D of $\mathbb{R}$ such that $(c - \delta, c) \cup (c, c + \delta)$ is contained in D. Let L be a real number. Then $\lim_{x \to c} f(x) = L$ if the function $\overline{f}$ with domain $(c - \delta, c + \delta)$ and defined as $\overline{f}(x) = f(x)$, $x \neq c$, $\overline{f}(c) = L$, is continuous at c.

Thus the limit L of a function as $x \to c$ is a value that if we "plug the hole" in the graph of the function $y = f(x)$ by setting $f(c) = L$, we get a graph which is

[3]The concept of continuity also makes sense for functions with domain a proper subset of $\mathbb{R}$, but let's save this for later.

continuous – just think nicely behaved, for now – at $x = c$.

Note that an immediate consequence of the definition is that if $f(x)$ is itself continuous at $x = c$, then it is already defined at c and $\lim_{x\to c} f(x) = f(c)$. Thus the limit of a continuous function at a point is simply the value of the function at that point. This is very important!

In fact, we can now give a better account of what we have been doing when we compute $f'(x)$. We start with the difference quotient $\frac{f(x+h)-f(x)}{h}$ which is defined for all sufficiently small h but *not* for $h = 0$. Then we manipulate / simplify the difference quotient until we recognize it as being equal, for all $h \neq 0$, to some new function $g(h)$ which is continuous at zero. (For instance, when $f(x) = x^2$, that new function was $g(h) = 2x + h$.) Then the limiting value is obtained simply by plugging in $h = 0$, i.e., it is $g(0)$.

4. Continuity Done Right

4.1. The formal defininition of continuity.

Let $D \subset \mathbb{R}$ and $f : D \to \mathbb{R}$ be a function. For $c \in \mathbb{R}$ we say **f is continuous at c** if for all $\epsilon > 0$ there exists $\delta > 0$ such that $(c-\delta, c+\delta) \subset D$ and for all x with $|x - c| < \delta$, $|f(x) - f(c)| < \epsilon$.

Morover we say that f is **continuous** if it is continuous at c for all c in its domain D.

The bit about the domain D is traditionally left a bit more implicit, but since we are trying to be precise we may as well be completely precise. The condition is equivalent to requiring that there be some $\Delta > 0$ such that f is defined on $(c - \Delta, c + \Delta)$ and that for all $\epsilon > 0$, whenever we speak of $\delta > 0$ we always take it as implicit that $\delta \leq \Delta$.

In general, mathematical statements become more complex and harder to parse the more alternating quantifiers they have: i.e., statements of the form "There exists x such that $P(x)$" or "For all x, $P(x)$" have a simple logical structure (if $P(x)$ is itself something reasonably simple, of course). Statements of the form "For all x, there exists y such that $P(x, y)$" and "There exists x such that for all y, $P(x, y)$" are a bit more complex; the untrained mind must stop to remind itself that they are not logically equivalent: e.g. if x and y are real numbers and $P(x, y)$ is $x > y$ then the first statement is true – for every real number, there exists a greatest real number – and the second statement is false – there is no real number which is greater than every eal number. The ϵ-δ definition of continuity has *three alternating quantifiers*: for all, then there exists, then for all. In general, to fully comphrehend the meaning of statements this logically complex takes serious mental energy.

Let us first talk about the geometric meaning of the statement: the inequality $|f(x) - f(c)| < \epsilon$ means $f(c) - \epsilon < f(x) < f(c) + \epsilon$: that is, it determines a horizontal strip centered at the horizontal line $y = f(c)$ of width 2ϵ. Similarly, the inequality $|x-c| < \delta$ means $c-\delta < x < c+\delta$, so determines a vertical strip centered at the vertical line $x = c$ of width 2δ. Thus the statement is saying something about

approximating both the y-values and the x-values of the function.

Now let us talk about the logical meaning of the statement and the sequence of quantifiers. We may think of it as a game: the first player chooses any $\epsilon > 0$ and thereby lays down a horizontal strip bounded above and below by the lines $f(c)+\epsilon$ and $f(c)-\epsilon$. The second player chooses a $\delta > 0$. Moreover, the second player wins if by restricting to x values lying between the vertical lines $c-\delta$ and $c+\delta$, the graph of the function is trapped between the two vertical lines $f(c) \pm \epsilon$; otherwise the first player wins. Now the assertion that f is continuous at $x = c$ is equivalent to the fact that the second player has a winning strategy: in other words, it is possible for her to win no matter which $\epsilon > 0$ the first player names.

Example 4.1: Constant functins $f(x) = C$ are continuous.
...

Example 4.2: The identity function $I(x) = x$ is continuous.
...

Example 4.3: Linear functions $f(x) = mx + b$ are continuous.
...

Example 4.4: $f(x) = x^2$ is continuous.
...

4.2. Basic properties of continuous functions.

LEMMA 4.6. *(Upper and Lower Bounds for Continuous Functions)*
Let f be continuous at $x = c$.
a) For any $\epsilon > 0$, there exists $\delta > 0$ such that $|x-c| < \delta$ implies $|f(x)| \leq |f(c)|+\epsilon$.
b) Suppose $f(c) \neq 0$. Then for any $\alpha \in (0,1)$, there exists $\delta > 0$ such that $|x-c| < \delta$ implies $|f(x)| \geq \alpha|f(c)|$.

PROOF. a) For any $\epsilon > 0$, there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \epsilon$. By the Reverse Triangle Inequality,

$$|f(x)| - |f(c)| \leq |f(x) - f(c)| < \epsilon,$$

so

$$|f(x)| \leq |f(c)| + \epsilon.$$

b) We will prove the result for $\alpha = \frac{1}{2}$, leaving the general case as an extra credit problem. There exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \frac{|f(c)|}{2}$. Again the Reverse Triangle Inequality implies

$$|f(c)| - |f(x)| \leq |f(x) - f(c)| < \frac{|f(c)|}{2},$$

or

$$|f(x)| > |f(c)| - \frac{|f(c)|}{2} = \frac{|f(c)|}{2}.$$

□

THEOREM 4.7. *Let f and g be functions and $c \in \mathbb{R}$.*
a) If f is continuous at c and $A \in \mathbb{R}$, then Af is continuous at c.
b) If f and g are both continuous at c then $f + g$ is continuous at c.

c) If f and g are both continuous at c then fg is continuous at c.
d) If f and g are both continuous at c and $g(c) \neq 0$, then $\frac{f}{g}$ is continuous at c.
e) If f is continuous at c and g is continuous at $f(c)$ then $g \circ f$ is continuous at c.

PROOF. For each part we work out the idea of the proof first and then translate it into a formal ϵ-δ argument.
a) Fix $\epsilon > 0$. We must show that there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|Af(x) - Af(c)| < \epsilon$. but $|Af(x) - Af(c)| = |A||f(x) - f(c)|$. Moreover, precisely because f is continuous at c we may make the quantity $|f(x) - f(c)$— as small as we like by taking x sufficiently close to c. A quantity which we can make as small as we like times a constant can still be made as small as we like!

Now formally: we may assume $A \neq 0$ for othewise Af is the constantly zero function, which we have already proved is continuous. For any $\epsilon > 0$, since f is continuous at $x = c$ there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \frac{\epsilon}{|A|}$. (Note what is being done here: by continuity, we can make $|f(x) - f(c)|$ less than *any* positive number we choose. It is convenient for us to make it smaller than $\frac{\epsilon}{|A|}$, where ϵ is a previously given positive number.) Then $|x - c| < \delta$ implies

$$|Af(x) - Af(c)| = |A||f(x) - f(c)| < |A| \cdot \frac{\epsilon}{|A|} = \epsilon.$$

b) Fix $\epsilon > 0$. We must show that there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) + g(x) - (f(c) + g(c))| < \epsilon$. Now

$$|f(x)+g(x)-(f(c)+g(c))| = |(f(x)-f(c))+(g(x)-g(c))| \leq |f(x)-f(c)|+|g(x)-g(c)|.$$

This is good: since f and g are both continuous at c, we can make each of $|f(x) - f(c)|$ and $|g(x) - g(c)|$ as small as we like by taking x sufficiently close to c. The sum of two quantities which can each be made as small as we like can be made as small as we like!

Now formally: choose $\delta_1 > 0$ such that $|x - c| < \delta_1$ implies $|f(x) - f(c)| < \frac{\epsilon}{2}$. Choose $\delta_2 > 0$ such that $|x - c| < \delta_2$ implies $|g(x) - g(c)| < \frac{\epsilon}{2}$. Let $\delta = \min(\delta_1, \delta_2)$. Then $|x - c| < \delta$ implies $|x - c| < \delta_1$ and $|x - c| < \delta_2$, so

$$|f(x) + g(x) - (f(c) + g(c))| \leq |f(x) - f(c)| + |g(x) - g(c)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

c) Fix $\epsilon > 0$. We must show that there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x)g(x) - f(c)g(c)| < \epsilon$. The situation here is somewhat perplexing: clearly we need to use the continuity of f and g at c, and to do this it stands to reason that we should be estimating $|f(x)g(x) - f(c)g(c)|$ in terms of $|f(x) - f(c)|$ and $|g(x) - g(c)|$, but unfortunately we do not yet see these latter two expressions. So *we force them to appear* by adding and subtracting $f(x)g(c)$:

$$\begin{aligned} |f(x)g(x) - f(c)g(c)| &= |f(x)g(x) - f(x)g(c) + f(x)g(c) - f(c)g(c)| \\ &\leq |f(x)||g(x) - g(c)| + |g(c)||f(x) - f(c)|. \end{aligned}$$

This is much better: $|g(c)||f(x) - f(c)|$ is a constant times something which can be made arbitrarily small, so can be made arbitarily small. Moreover in the term $|f(x)||g(x) - g(c)|$ we can make $|g(x) - g(c)|$ arbitarily small by taking x sufficiently close to c, and then, by continuity of f, $|f(x)|$ gets arbitrarily close to $|f(c)|$. So $|f(x)|$ is nonconstant but *bounded*, and something which is bounded times something which can be made arbitarily small can be made arbitrarily small!
Now formally:

Using Lemma 4.6a) and taking $\epsilon = 1$, there exists $\delta_1 > 0$ such that $|x - c| < \delta_1$ implies $|f(x)| \leq |f(c)| + 1$. There exists $\delta_2 > 0$ such that $|x - c| < \delta_2$ implies $|g(x) - g(c)| < \frac{\epsilon}{2(|f(c)|+1)}$. Finally, there exists $\delta_3 > 0$ such that $|x - c| < \delta_3$ implies $|f(x) - f(c)| < \frac{\epsilon}{2|g(c)|}$. (Here we are assuming that $g(c) \neq 0$. If $g(c) = 0$ then we simply don't have the second term in our expression and the argument is similar but easier.) Taking $\delta = \min \delta_1, \delta_2, \delta_3$, for $|x - c| < \delta$ then $|x - c|$ is less than δ_1, δ_2 and δ_3 so

$$|f(x)g(x) - f(c)g(c)| \leq |f(x)||g(x) - g(c)| + |g(c)||f(x) - f(c)|$$
$$< (|f(c)| + 1) \cdot \frac{\epsilon}{2(|f(c)| + 1)} + |g(c)|\frac{\epsilon}{2|g(c)|} = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

d) Since $\frac{f}{g} = f \cdot \frac{1}{g}$, in light of part c) it will suffice to show that if g is continuous at c and $g(c) \neq 0$ then $\frac{1}{g}$ is continuous at c. Fix $\epsilon > 0$. We must show that there exists $\delta > 0$ such that $|x - c| < \delta$ implies

$$|\frac{1}{g(x)} - \frac{1}{g(c)}| < \epsilon.$$

Now

$$|\frac{1}{g(x)} - \frac{1}{g(c)}| = \frac{|g(c) - g(x)|}{|g(x)||g(c)|} = \frac{|g(x) - g(c)|}{|g(x)||g(c)|}.$$

Since g is continuous at $x = c$, we can make the numerator $|g(x) - g(c)|$ as small as we like by taking x sufficiently close to c. This will make the entire fraction as small as we like provided the denominator is not also getting arbitrarily small as x approaches c. But indeed, since g is continuous at c and $g(c) \neq 0$, the denominator is approaching $|g(c)|^2 \neq 0$. Thus again we have a quantity which we can make arbitarily small times a bounded quantity, so it can be made arbitrarily small!
Now formally:
We apply Lemma 4.6b) with $\alpha = \frac{1}{2}$: there exists $\delta_1 > 0$ such that $|x - c| < \delta_1$ implies $|g(x)| \geq \frac{|g(c)|}{2}$ and thus also

$$\frac{1}{|g(x)||g(c)|} \leq \frac{2}{|g(c)|^2}.$$

Also there exists $\delta_2 > 0$ such that $|x - c| < \delta_2$ implies $|g(x) - g(c)| < \left(\frac{|g(c)|^2}{2}\right)\epsilon$. Taking $\delta = \min(\delta_1, \delta_2$, $|x - c| < \delta$ implies

$$|\frac{1}{g(x)} - \frac{1}{g(c)}| = \left(\frac{1}{|g(x)||g(c)|}\right)|g(x) - g(c)| < \frac{2}{|g(c)|^2}\left(\frac{|g(c)|^2}{2}\right)\epsilon = \epsilon.$$

e) Fix $\epsilon > 0$. Since $g(y)$ is continuous at $y = f(c)$, there exists $\gamma > 0$ such that $|y - f(c)| < \gamma$ implies $|g(y) - g(f(c))| < \epsilon$. Moreover, since f is continuous at c, there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \gamma$. Thus, if $|x - c| < \delta$, $|f(x) - f(c)| = |y - f(c)| < \gamma$ and hence

$$|g(f(x)) - g(f(c))| = |g(y) - g(f(c))| < \epsilon.$$

□

COROLLARY 4.8. *All rational functions are continuous.*

PROOF. Since rational functions are built out of constant functions and the identity by repeated addition, multiplication and division, this follows immediately from Theorem 4.7. □

Other elementary functions: unfortunately if we try to go beyond rational functions to other elementary functions familiar from precalculus, we run into the issue that we have not yet given complete, satisfactory definitions of these functions! For instance, take even the relatively innocuous $f(x) = \sqrt{x}$. We want this function to have domain $[0, \infty)$, but this uses the special property of $\mathbb{R}$ that every non-negative number has a square root: we haven't proved this yet! If $\alpha > 0$is irrational we have not given any definition of the power function x^α. Similarly we do not yet have rigorous definitions of a^x for $a > 1$, $\log x$, $\sin x$ and $\cos x$, so we are poorly placed to rigorously prove their continuity. However (following Spivak) in order so as not to drastically limit the supply of functions to appear in our examples and exercises, we will **proceed for now on the assumption** that all the above elementary functions are continuous. We hasten to make two remarks:

Remark 4.5: This assumption can be justified! That is, all the elementary functions above are indeed continuous at every point of their domain (with the small proviso that for power functions like $\sqrt{x}$ we will need to give a separate definition of continuity at an endpoint of an interval, coming up soon). And in fact we will prove this later in the course...much later.

Remark 4.6: We will not use the continuity of the elementary functions as an assumption in any of our main results (but only in results and examples explicitly involving elementary functions; e.g. we will use the assumed continuity of the sine function to differentiate it). Thus it will be clear that we are not arguing circularly when we finally prove the continuity of these functions.

5. Limits Done Right

5.1. The Formal Definition of a Limit.

In order to formally define limits, it is convenient to have the notion of a **deleted interval** $I_{c,\delta}$ about a point c, namely a set of real numbers of the form

$$0 < |x - c| < \delta$$

for some $\delta > 0$. Thus I_c consists of $(c - \delta, c)$ together with the points $(c, c + \delta)$, or more colloquially it contains all points "sufficiently close to c but not equal to c".

Now comes the definition. For real numbers c and L and a function $f : D \subset \mathbb{R} \to \mathbb{R}$, we say $\lim_{x\to c} f(x) = L$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all x in the deleted interval $I_{c,\delta}$ – i.e., for all x with $0 < |x - c| < \delta$ – f is defined at x and $|f(x) - L| < \epsilon$.

Among all the many problems of limits, perhaps the following is the most basic and important.

THEOREM 4.9. *The limit at a point is unique (if it exists at all): that is, if L and M are two numbers such that $\lim_{x\to c} f(x) = L$ and $\lim_{x\to c} f(x) = M$, then $L = M$.*

PROOF. Seeking a contradiction, we suppose $L \neq M$; it is no loss of generality to suppose that $L < M$ (otherwise switch L and M) and we do so. Now we take $\epsilon = \frac{M-L}{2}$ in the definition of limit: since $\lim_{x\to c} f(x) = L$, there exists $\delta_1 > 0$ such

that $0 < |x - c| < \delta_1$ implies $|f(x) - L| < \frac{M-L}{2}$; and similarly, since $\lim_{x\to c} f(x) = M$, there exists $\delta_2 > 0$ such that $0 < |x - c| < \delta_2$ implies $|f(x) - M| < \frac{M-L}{2}$. Taking $\delta = \min(\delta_1, \delta_2$, then, as usual, for $0 < |x - c| < \delta$ we get both inequalities:

$$|f(x) - L| < \frac{M - L}{2}$$

$$|f(x) - M| < \frac{M - L}{2}.$$

However these inequalities are contradictory! Before we go further we urge the reader to **draw a picture** to see that the vertical strips defined by the two inequalities above are *disjoint*: they have no points in common. Let us now check this formally: since $|f(x) - L| < \frac{M-L}{2}$, $f(x) < L + \frac{M-L}{2} = \frac{M+L}{2}$. On the other hand, since $|f(x) - M| < \frac{M-L}{2}$, $f(x) > M - \frac{M-L}{2} = \frac{M+L}{2}$. Clearly there is not a single value of x such that $f(x)$ is at the same time greater than and less than $\frac{M+L}{2}$, let alone a deleted interval around c of such values of x, so we have reached a contradiction. Therefore $L = M$. □

We have now given a formal definition of continuity at a point and also a formal definition of limits at a point. Previously though we argued that each of limits and continuity can be defined in terms of the other, so we are now in an "overdetermined" situation. We should therefore check the compatibility of our definitions.

THEOREM 4.10. *Let $f : D \subset \mathbb{R} \to \mathbb{R}$, and let $c \in \mathbb{R}$.*
a) f is continuous at $x = c$ if and only if f is defined at c and $\lim_{x\to c} f(x) = f(c)$.
b) $\lim_{x\to c} f(x) = L$ iff f is defined in some deleted interval $I_{c,\delta}$ around x and, defining $\overline{f}$ on $(c - \delta, c + \delta)$ by $\overline{f}(x) = f(x)$, $x \neq c$, $\overline{f}(c) = L$ makes $\overline{f}$ continuous at $x = c$.

PROOF. All the pedagogical advantage here comes from working through this yourself rather than reading my proof, so I leave it to you. □

5.2. Basic Properties of Limits.

Most of the basic properties of continuity discussed above have analogues for limits. We state the facts in the following result.

THEOREM 4.11. *Let f and g be two functions defined in a deleted interval $I_{c,\delta}$ of a point c. We suppose that $\lim_{x\to c} f(x) = L$ and $\lim_{x\to c} g(x) = M$.*
a) For any constant A, $\lim_{x\to c} Af(x) = AL$.
b) We have $\lim_{x\to c} f(x) + g(x) = L + M$.
c) We have $\lim_{x\to c} f(x)g(x) = LM$.
d) If $M \neq 0$, then $\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{L}{M}$.

We leave the proof to the reader. It is possible to prove any/all of these facts in one of two ways: (i) by rephrasing the definition of limit in terms of continuity and appealing to Theorem 4.7 above, or (ii) adapting the proofs of Theorem 4.7 to the current context.

Now what about limits of composite functions? The natural analogue of Theorem 4.7e) above would be the following:

If $\lim_{x\to c} f(x) = L$ and $\lim_{x\to L} g(x) = M$, then $\lim_{x\to c} g(f(x)) = M$.

Unfortunately the above statement is not always true!

Example 5.1: Let $f(x) = 0$ (constant function). Let $g(x)$ be equal to 1 for $x \neq 0$ and $g(0) = 0$. Take $c = 0$. Then $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} g(x) = 1$, but for all $x \in \mathbb{R}$, $g(f(x)) = g(0) = 0$, so $\lim_{x\to 0} g(f(x)) = 0$.

This result can be repaired by requiring the continuity of the "outside" function $g(x)$. Indeed the proof of the following result is almost identical to that of Theorem 4.7e) above in which f is also continuous at c: in fact, it shows that the continuity of the "inside function" f was not really used.

THEOREM 4.12. *Let f, g be functions with $\lim_{x\to c} f(x) = L$ and g continuous at L. Then $\lim_{x\to c} g(f(x)) = g(L)$.*

PROOF. Fix $\epsilon > 0$. Since $\lim_{y\to f(c)} g(y) = g(L)$, there exists $\gamma > 0$ such that $|y - f(c)| < \gamma$ implies $|g(y) - g(L)| < \epsilon$. Since $\lim_{x\to c} f(x) = L$, there exists $\delta > 0$ such that $0 < |x - c| < \delta$ implies $|y - L| = |f(x) - L| < \gamma$ and thus $|g(f(x)) - g(L)| = |g(y) - g(L)| < \epsilon$. □

One may rephrase Theorem 4.12 as: if g is continuous and $\lim_{x\to c} f(x)$ exists, then

$$\lim_{x\to c} g(f(x)) = g(\lim_{x\to c} f(x)).$$

In other words, one can "pull a limit through a continuous function". In this form the result is actually a standard one in freshman calculus.

It happens that one can say *exactly* when the above statement about limits of composite funtions holds. I don't plan on mentioning this in class and you needn't keep it in mind or even read it, but I recently learned that this question has a rather simple answer so I might as well record it here so I don't forget it myself.

THEOREM 4.13. *(Marjanović* [**MK09**]*) Suppose $\lim_{x\to c} f(x) = L$ and $\lim_{x\to L} g(x) = M$. The following are equivalent:*
(i) $\lim_{x\to c} g(f(x)) = M$.
(ii) At least one of the following holds:
a) g is continuous at L.
b) There exists $\Delta > 0$ such that for all $0 < |x - c| < \Delta$, $f(x) \neq L$.

PROOF. (i) $\implies$ (ii): We will argue by contradiction: suppose that neither a) nor b) holds; we will show that $\lim_{x\to c} g(f(x)) \neq M$. Indeed, since b) dooes not hold, for every $\delta > 0$ there exists x with $0 < |x - c| < \delta$ such that $f(x) = L$. For such x we have $g(f(x)) = g(L)$. But since a) does not hold, g is *not* continuous at L, i.e., $M \neq g(L)$. Thus $g(f(x)) = g(L) \neq M$. Taking $\epsilon = |g(L) - M|$ this shows that there is no $\delta > 0$ such that $0 < |x - c| < \delta$ implies $|g(f(x)) - M| < \epsilon$, so $\lim_{x\to c} g(f(x)) \neq M$.
(ii) $\implies$ (i). The case in which a) holds – i.e., g is continuous at L – is precisely Theorem 4.12. So it suffices to assume that b) holds: there exists some $\Delta > 0$ such that $0 < |x - c| < \Delta$ implies $f(x) \neq L$. Now fix $\epsilon > 0$; since $\lim_{x\to L} g(x) = M$, there exists $\gamma > 0$ such that $0 < |y - L| < \gamma$ implies $|g(y) - M| < \epsilon$. Similarly (and familiarly), since $\lim_{x\to c} f(x) = L$, there exists $\delta_1 > 0$ such that $0 < |x - c| < \delta_1$ implies $|f(x) - L| < \gamma$. Here is the point: for $0 < |x - c| < \delta_1$, we have $|f(x) - L| < \gamma$.

If in addition $0 < |f(x) - L| < \gamma$, then we may conclude that $|g(f(x)) - M| < \epsilon$. So our only concern is that pehaps $f(x) = L$ for some c with $0 < |x - c| < \delta_1$, and this is exactly what the additional hypothesis b) allows us to rule out: if we take $\delta = \min(\delta_1, \Delta)$ then $0 < |x - c| < \delta$ implies $0 < |f(x) - L| < \gamma$ and thus $|g(f(x)) - M| < \epsilon$. □

Remark 5.2: The nice expository article [**MK09**] gives some applications of the implication (ii)b) $\implies$ (i) of Theorem 4.13 involving making an *inverse change of variables* to evaluate a limit. Perhaps we may revisit this point towards the end of the course when we talk about inverse functions.

5.3. The Squeeze Theorem and the Switching Theorem.

THEOREM 4.14. *(Squeeze Theorem) Let $m(x)$, $f(x)$ and $M(x)$ be defined on some deleted interval $I^\circ = (c - \Delta, c + \Delta) - \{c\}$ about $x = c$. We suppose that:*
(i) For all $x \in I^\circ$, $m(x) \leq f(x) \leq M(x)$, and
(ii) $\lim_{x \to c} m(x) = \lim_{x \to c} M(x) = L$.
Then $\lim_{x \to c} f(x) = L$.

PROOF. Fix $\epsilon > 0$. There exists $\delta_1 > 0$ such that $0 < |x - c| < \delta_1$ implies $|m(x) - L| < \epsilon$ and $\delta_2 > 0$ such that $0 < |x - c| < \delta_2$ implies $|M(x) - L| < \epsilon$. Let $\delta = \min(\delta_1, \delta_2)$. Then $0 < |x - c| < \delta$ implies

$$f(x) \leq M(x) < L + \epsilon$$

and

$$f(x) \geq m(x) > L - \epsilon,$$

so $L - \epsilon < f(x) < L + \epsilon$, or equivalently $|f(x) - L| < \epsilon$. □

Example 5.3: For $\alpha \geq 0$, define $f_\alpha : \mathbb{R} \to \mathbb{R}$ by $f_\alpha(x) = x^\alpha \sin(\frac{1}{x})$, $x \neq 0$ and $f_\alpha(0) = 0$. By our assumption about the continuity of sign and our results on continuity of rational functions and compositions of continuous functions, f_α is continuous at all $x \neq 0$. We claim that f_α is continuous at $x = 0$ iff $\alpha > 0$. ...

Example 5.4: The function f_α defined above is differentiable at $x = 0$ iff $\alpha > 1$. ...

Example 5.5: $\lim_{x \to 0} \frac{\sin x}{x} = 1$.
Solution: This will be a "17th century solution" to the problem: i.e., we will assume that the trigonometric functions are continuous and use geometric reasoning, including properties of angles and arclength. (Nevertheless, this solution seems a lot better than not giving any proof at all until much later in the course...)

Consider the unit circle and the point on it $P_x = (\cos x, \sin x)$. There is a right triangle T_1 with vertices $(0, 0)$, $(\cos x, 0)$, P_x. This right triangle is contained in the circular sector determined by all points on or inside the unit circle with angle between 0 and x. In turn this circular sector is contained in a second right triangle T_2, with vertices $(0, 0)$, $(1, 0)$, $(1, \tan x)$. Now let us write down the inequalities expressing the fact that since T_1 is contained in the circular sector which is contained in T_2, the area of T_1 is less than or equal to the area of the circular sector, which is less than or equal to the area of T_2.

The area of T_1 is (one half the base times the height) $\frac{1}{2} \cos x \sin x$. The area of

the circular sector is $\frac{x}{2\pi}$ times the area of the unit circle, or $\frac{x}{2\pi} \cdot \pi = \frac{x}{2}$. The area of T_2 is $\frac{1}{2}\tan x = \frac{1}{2}\frac{\sin x}{\cos x}$. This gives us the inequalities

$$\frac{1}{2}\cos x \sin x \leq \frac{1}{2}x \leq \frac{1}{2}\frac{\sin x}{\cos x},$$

or equivalently, for $x \neq 0$,

$$\cos x \leq \frac{x}{\sin x} \leq \frac{1}{\cos x}.$$

Taking reciprocals this inequality is equivalent to

$$\frac{1}{\cos x} \leq \frac{\sin x}{x} \leq \cos x.$$

Now we may apply the Squeeze Theorem: since $\cos x$ is continuous at 0 and takes the value $1 \neq 0$ there, we have

$$\lim_{x\to 0} \frac{1}{\cos x} = \lim_{x\to 0} \cos x = 1.$$

Therefore the Squeeze Theorem implies

$$\lim_{x\to 0} \frac{\sin x}{x} = 1. \tag{16}$$

Example 5.6: We will evaluate $\lim_{x\to 0} \frac{1-\cos x}{x} = 0$. The idea is to use trigonometric identities to reduce this limit to an expression involving the limit (16). Here goes:

$$\lim_{x\to 0} \frac{1-\cos x}{x} = \lim_{x\to 0} \frac{1-\cos x}{x}\left(\frac{1+\cos x}{1+\cos x}\right) = \lim_{x\to 0} \frac{\cos^2 x - 1}{x(1+\cos x)} = \lim_{x\to 0} \frac{-\sin^2 x}{x(1+\cos x)}$$

$$= \left(\lim_{x\to 0} \frac{\sin x}{x}\right)\left(\lim_{x\to 0} \frac{-\sin x}{1+\cos x}\right) = 1 \cdot (\frac{-0}{2}) = 0.$$

Of course we also have $\lim_{x\to 0} \frac{\cos x - 1}{x} = -\lim_{x\to 0} \frac{1-\cos x}{x} = -0 = 0$. In summary:

$$\lim_{x\to 0} \frac{1-\cos x}{x} = \lim_{x\to 0} \frac{\cos x - 1}{x} = 0. \tag{17}$$

Before doing the next two examples we remind the reader of the composite angle formulas from trigonometry: for any real numbers x, y,

$$\sin(x+y) = \sin x \cos y + \cos x \sin y,$$
$$\cos(x+y) = \cos x \cos y - \sin x \sin y.$$

Example 5.7: If $f(x) = \sin x$, then we claim $f'(x) = \cos x$. Indeed

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{\sin(x+h)-\sin x}{h} = \lim_{h\to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h}$$

$$= -\sin x \left(\lim_{h\to 0} \frac{1-\cos h}{h}\right) + \cos x \left(\lim_{h\to 0} \frac{\sin h}{h}\right) = (-\sin x)\cdot 0 + (\cos x)\cdot 1 = \cos x.$$

Example 5.8: If $f(x) = \cos x$, then we claim $f'(x) = -\sin x$. Indeed

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{\cos(x+h)-\cos x}{h} =$$

$$\lim_{h\to 0} \frac{\cos x \cos h - \sin x \sin h - \cos x}{h}$$

$$= -\cos x \left(\lim_{h\to 0} \frac{1-\cos h}{h}\right) - \sin x \left(\lim_{h\to 0} \frac{\sin h}{h}\right) = (-\cos x)\cdot 0 + (-\sin x)\cdot 1 = -\sin x.$$

THEOREM 4.15. *(Switching Theorem) Consider three functions f, g_1, g_2 defined on some deleted interval $I_{c,\Delta}$ of $x = c$. We suppose that:*
(i) $\lim_{x\to c} g_1(x) = \lim_{x\to c} g_2(x) = L$.
(ii) For all x with $0 < |x - c| < \Delta$, either $f(x) = g_1(x)$ or $f(x) = g_2(x)$.
Then $\lim_{x\to c} f(x) = L$.

PROOF. Fix $\epsilon > 0$. Let $\delta_1 > 0$ be such that $0 < |x-c| < \delta_1$ implies $|g_1(x)-L| < \epsilon$, let $\delta_2 > 0$ be such that $0 < |x - c| < \delta_2$ implies $|g_2(x) - L| < \epsilon$, and let $\delta = \min(\delta_1, \delta_2)$. Let x be such that $0 < |x - c| < \delta$. Then either $f(x) = g_1(x)$, in which case $|f(x)-L| = |g_1(x)-L| < \epsilon$, or $f(x) = g_2(x)$, in which case $|f(x)-L| = |g_2(x) - L| < \epsilon$. Either way, $|f(x) - L| < \epsilon$! □

Example 5.9: Let $f(x)$ be defined as x for rational x and $-x$ for irrational x. We may apply the Switching Theorem to show that f is continuous at 0. Indeed, put $g_1(x) = x$ and $g_2(x) = -x$. Then $\lim_{x\to 0} g_1(x) = \lim_{x\to 0} g_2(x) = 0$ and for all x, $f(x) = g_1(x)$ or $f(x) = g_2(x)$. So by the Switching Theorem, $f(0) = 0 = \lim_{x\to 0} f(x)$.

Remark 5.10: The previous example shows that a function may be very strangely behaved and still be continuous at a point. It is thus worth emphasizing that we are not really interested in functions which are continuous at certain points, but rather at functions which are continuous at every point of their domain. Such functions have many pleasant properties that we will prove later on in the course.

Remark 5.11: The "Switching Theorem" is not a standard result. That is to say, I came up with it myself, inspired by the homework problem asking for a function which is continuous at a single point. Although I do not claim it is in the same league as the venerated Squeeze Theorem, I *do plan* to use it later on to give a proof of the Chain Rule which is (I think) simpler than the one Spivak gives.

5.4. Variations on the Limit Concept.

Finally we consider three variations on the limit concept: one-sided limits, infinite limits, and limits at infinity. There is nothing really novel going on here, but we need to be sure that we can adapt our ϵ-δ formalism to the variant notions of limit that apply in calculus.

By way of introducing the first variant, consider the function $f(x) = \sqrt{x}$. We have said earlier that we are assuming for now that this function is continuous on its entire domain. But that statement glossed over a technicality which we now address. Namely, the domain of f is $[0, \infty)$. So f cannot be continuous at 0 according to the definition that we gave because it is not defined on any open interval containing zero. Instead it is defined, for instance, on an interval of the form $[0, \delta)$ for $\delta > 0$: i.e., this contains all points sufficiently close to 0 but greater than or equal to zero.

A similar phenomenon arises when we consider the function $f(x) = \sqrt{x(1-x)}$, which has natural domain $[0, 1]$. The function is defined at 1 but not in any open interval containing 1: only intervals of the form $(1 - \delta, 1]$.

This brings us to our definition. We say that a function $f : D \subset \mathbb{R}$ is **right continuous** at $x = c$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that f is defined on $[c, c+\delta)$ and $x \in [c, c+\delta) \implies |f(x) - f(c)| < \epsilon$. Similarly, we say that a function $f : D \subset \mathbb{R}$ is **left continuous** at $x = c$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that f is defined on $(c-\delta, c]$ and $x \in (c-\delta, c] \implies |f(x) - f(c)| < \epsilon$.

Finally we make the following definition: let $I \subset \mathbb{R}$ be any interval and let $f : I \to \mathbb{R}$ be a function. We say that a is a **left endpoint** of I if $a \in I$ and there is no $x \in I$ with $x < a$; similarly we say b is a **right endpoint** of I if $b \in I$ and there is no $x \in I$ with $x > b$. An interval I has at most one endpoint and at most one endpoint; all four possibilities of having / not having left / right endpoints are of course possible. Let us say $c \in I$ is an **interior point** if it is *not* a left endpoint or a right endpoint. Now we say that $f : I \to\to \mathbb{R}$ is continuous if:

- f is continuous at c for each interior point $c \in I$,
- If I has a left endpoint a, then f is right continuous at a, and
- If I had a right endpoint b, then f is left continuous at b.

Example: $f(x) = \sqrt{x}$ is right continuous at $x = 0$.

As above, it is necessary to require left/right continuity when discussing behavior at right/left endpoints of an interval. On the other hand one may still discuss left/right continuity at interior points of an interval, and it is sometimes helpful to do so.

Example: Let $f(x) = \lfloor x \rfloor$ be the greatest integer function. Then f is continuous at c for all $c \in \mathbb{R} \setminus \mathbb{Z}$, whereas for any $c \in \mathbb{Z}$, f is right continuous but not left continuous at c.

This example suggests the following simple result.

PROPOSITION 4.16. *For a function $f : D \subset \mathbb{R} \to \mathbb{R}$ and $c \in D$, the following are equivalent:*
(i) f is left continuous at c and right continuous at c.
(ii) f is continuous at c.

We leave the proof to the reader.

In a similar way we can define **one-sided limits** at a point c.

We say $\lim_{x \to c^-} f(x) = L$ – and read this as *the limit as x approaches c from the left of $f(x)$ is L* – if for all $\epsilon > 0$ there exists $\delta > 0$ such that for all x with $c < x < c + \delta$, $|f(x) - L| < \epsilon$.

We say $\lim_{x \to c^+} f(x) = L$ – and read this as *the limit as x approaches c from the right of $f(x)$ is L* – if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all x with $c - \delta < x < c$, $|f(x) - L| < \epsilon$.

PROPOSITION 4.17. *For a function $f : D \subset \mathbb{R} \to \mathbb{R}$ and $c \in D$, the following are equivalent:*

(i) The left hand and right hand limits at c exist and are equal.
(ii) $\lim_{x \to c} f(x)$ *exists.*

Again we leave the proof to the reader.

Example: Let $f(x) = \lfloor x \rfloor$ be the greatest integer function, and let $n \in \mathbb{Z}$. Then $\lim_{x \to n^-} f(x) = n - 1$ and $\lim_{x \to n^+} f(x) = n$, so f is not continuous at n.

There is some terminology here – not essential, but sometimes useful. If for a function f and $c \in \mathbb{R}$, the left and right hand limits at c both exist but are unequal, we say that f has a **jump discontinuity** at c. If the left and right hand limts at c both exist and are equal – i.e., if $\lim_{x \to c} f(x) = L$ exists – but still $f(x)$ is not continuous c (this can happen if either $f(c) \neq L$ or, more plausibly, if c is not in the domain of f) we say that f has a **removable discontinuity** at c. This terminology comes from our earlier observation that if we (re)define f at c to be the limiting value L then f becomes continuous at c. One sometimes calls a discontinuity which is either removable or a jump discontinuity a **simple discontinuity**: i.e., this is the case whenever both one-sided limits exist at c but f is not continuous at c.

Infinite limits: Consider $\lim_{x \to 0} \frac{1}{x^2}$. This limit does not exist: indeed, if it did, then there would be some deleted interval $I_{0,\delta}$ on which f is bounded, whereas just the opposite is happening: the closer x is to 0, the larger $f(x)$ becomes. In freshman calculus we would say $\lim_{x \to 0} f(x) = \infty$. And we still want to say that, but in order to know what we mean when we say this we want to give an ϵ-δ style definition of this. Here it is:

We say $\lim_{x \to c} f(x) = \infty$ if for all $M \in \mathbb{R}$, there exists $\delta > 0$ such that $0 < |x - c| < \delta \implies f(x) > M$.

Geometrically, this is similar to the ϵ-δ definition of limit, but instead of picking two horizontal lines arbitrarily close to $y = L$, we pick one horizontal line which is arbitrarily large and require that on some small deleted interval the graph of $y = f(x)$ always lie above that line. Similarly:

We say $\lim_{x \to c} f(x) = -\infty$ if for all $m \in \mathbb{R}$, there exists $\delta > 0$ such that $0 < |x - c| < \delta$ implies $f(x) < m$.

Example: Let us indeed prove that $\lim_{x \to 0} \frac{1}{x^2} = \infty$. Fix $M \in \mathbb{R}$. We need to find δ such that $0 < |x| < \delta$ implies $\frac{1}{x^2} > M$. It is no loss of generality to assume $M > 0$ (why?). Then $\frac{1}{x^2} > M \iff |x| < \frac{1}{\sqrt{M}}$, so we may take $\delta = \frac{1}{\sqrt{M}}$.

CHAPTER 5

Differentiation

1. Differentiability Versus Continuity

Recall that a function $f : D \subset \mathbb{R} \to \mathbb{R}$ is differentiable at $a \in D$ if

$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists, and when this limit exists it is called the **derivative** $f'(a)$ of f at a. Moreover, the **tangent line** to $y = f(x)$ at $f(a)$ exists if f is differentiable at a and is the unique line passing through the point $(a, f(a))$ with slope $f'(a)$.

Note that an equivalent definition of the derivative at a is

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

One can see this by going to the ϵ-δ definition of a limit and making the "substitution" $h = x - a$: then $0 < |h| < \delta \iff 0 < |x - a| < \delta$.

THEOREM 5.1. *Let $f : D \subset \mathbb{R} \to \mathbb{R}$ be a function, and let $a \in D$. If f is differentiable at a, then f is continuous at a.*

PROOF. We have

$$\lim_{x \to a} f(x) - f(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \cdot (x - a)$$

$$= \left(\lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right) \left(\lim_{x \to a} x - a \right) = f'(a) \cdot 0 = 0.$$

Thus

$$0 = \lim_{x \to a} (f(x) - f(a)) = \left(\lim_{x \to a} f(x) \right) - f(a),$$

so

$$\lim_{x \to a} f(x) = f(a).$$

□

The converse of Theorem 5.1 is far from being true: a function f which is continuous at a need not be differentiable at a. An easy example is $f(x) = |x|$ at $a = 0$.

In fact the situation is worse: a function $f : \mathbb{R} \to \mathbb{R}$ can be continuous *everywhere* yet still fail to be differentiable at many points. One way of introducing points of non-differentiability while preserving continuity is to take the absolute value of a differentiable function.

THEOREM 5.2. *Let $f : D \subset \mathbb{R} \to \mathbb{R}$ be continuous at $a \in D$.*
a) Then $|f|$ is continuous at a.
b) The following are equivalent:
(i) f is differentiable at a, and either $f(a) \neq 0$ or $f(a) = f'(a) = 0$.
(ii) $|f|$ is differentiable at a.

Exercise: Prove Theorem 8.31.

2. Differentiation Rules

2.1. Linearity of the Derivative.

THEOREM 5.3. *(Constant Rule) Let f be differentiable at $a \in \mathbb{R}$ and $C \in \mathbb{R}$. Then the function Cf is also differentiable at a and*

$$(Cf)'(a) = Cf'(a).$$

PROOF. There is nothing to it:

$$(Cf)'(a) = \lim_{h\to 0} \frac{(Cf)(a+h) - (Cf)(a)}{h} = C\left(\lim_{h\to 0} \frac{f(a+h) - f(a)}{h}\right) = Cf'(a).$$

□

THEOREM 5.4. *(Sum Rule) Let f and g be functions which are both differentiable at $a \in \mathbb{R}$. Then the sum $f + g$ is also differentiable at a and*

$$(f+g)'(a) = f'(a) + g'(a).$$

PROOF. Again, no biggie:

$$(f+g)'(a) = \lim_{h\to 0} \frac{(f+g)(a+h) - (f+g)(a)}{h} = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} + \frac{g(a+h) - g(a)}{h}$$
$$= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} + \lim_{h\to 0} \frac{g(a+h) - g(a)}{h} = f'(a) + g'(a).$$

□

These results, simple as they are, have the following important consequence.

COROLLARY 5.5. *(Linearity of the Derivative) For any differentiable functions f and g and any constants C_1, C_2, we have*

$$(C_1 f + C_2 g)' = C_1 f' + C_2 g'.$$

2.2. Product Rule(s).

THEOREM 5.6. *(Product Rule) Let f and g be functions which are both differentiable at $a \in \mathbb{R}$. Then the product fg is also differentiable at a and*

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a).$$

PROOF.

$$(fg)'(a) = \lim_{h\to 0} \frac{f(a+h)g(a+h) - f(a)g(a)}{h}$$
$$= \lim_{h\to 0} \frac{f(a+h)g(a+h) - f(a)g(a+h) + (f(a)g(a+h) - f(a)g(a))}{h}$$
$$= \left(\lim_{h\to 0} \frac{f(a+h) - f(a)}{h}\right)\left(\lim_{h\to 0} g(a+h)\right) + f(a)\left(\lim_{h\to 0} \frac{g(a+h) - g(a)}{h}\right).$$

Since g is differentiable at a, g is continuous at a and thus $\lim_{h\to 0} g(a+h) = \lim_{x\to} ag(x) = g(a)$. The last expression above is therefore equal to

$$f'(a)g(a) + f(a)g'(a).$$

□

Dimensional analysis and the product rule: Leibniz kept a diary in which he recorded his scientific discoveries, including his development of differential calculus. One day he got to the product rule and wrote a formula equivalent to

$$(fg)' = f'g'.$$

But this entire entry of the diary is crossed out. Three days later there is a new entry with a correct statement of the product rule and the derivation, which Leibniz says he has known "for some time".

I confess that I have taken this story from a third-party account. Every once in a while I think about trying to verify it, but I am always stopped by the fact that if it turns out to be apocryphal, I don't really want to know: it's too good a story! It is fundamentally honest about the way mathematics is done and, especially, the way new mathematics is created. That is, to go forward we make guesses. Often we later realize that our guesses were not only wrong but silly, to the extent that we try to hide them from others and perhaps even from ourselves. But the process of "silly guesswork" is vital to mathematics, and if the great genius Gottfried von Leibniz was not above them, what chance do the rest of us have?

It is natural, like Leibniz, to want to hide our silly guesses. However, they play an important role in teaching and learning. Many veteran instructors of higher mathematics come to lament the fact that for the sake of efficiency and "coverage of the syllabus" we often give the student the correct answer within minutes or seconds of asking the question (or worse, do not present the question at all!) and thus deprive them of the opportunity to make a silly guess and learn from their mistake. What can we learn from "Leibniz's" guess that $(fg)' = f'g'$? Simply to write down the correct product rule is not enough: I want to try to persuade you that "Leibniz's guess" was not only incorrect but truly silly: that in some sense the product rule could have turned out to be any number of things, but certainly not $(fg)' = f'g'$.

For this I want to follow an approach I learned in high school chemistry: **dimensional analysis**. We can give physically meaningful dimensions (the more common, but less precise, term is "units") to all our quantities, and then our formulas must manipulate these quantities in certain ways or they are meaningless.

For a simple example, suppose you walk into an empty room and find on the blackboard a drawing of a cylinder with its radius labelled r and its height labelled h. Below it are some formulas, and this one jumps out at you: $2\pi(rh + r^3)$. You can see right away that something has gone wrong: since r and h are both lengths – say in meters – rh denotes a product of lengths – say, square meters. But r^3 is a product of three lengths – i.e., cubic meters. How on earth will we get anything physically meaningful by adding square meters to cubic meters? It is much more likely that the writer meant $2\pi(rh + r^2)$, since that is a meaningful quantity of dimension length squared (i.e., area). In fact, it happens to be the correct formula for

the surface area of the cylinder with the top and bottom faces included, although dimensional considerations won't tell you that.

Something similar can be applied to the "formula" $(fg)' = f'g'$. Let $f = f(t)$ and view t as being time, say measured in seconds. Since $f'(t)$ is a limit of quotients $\frac{f(t+h)-f(t)}{h}$, its dimension is the dimension of f divided by time. Now suppose both $f(t)$ and $g(t)$ are lengths, say in meters. Then $f'(t)$ and $g'(t)$ both have units meters per second, so $f'(t) \cdot g'(t)$ has units meters squared per second squared. On the other hand, the units of $(fg)'$ are meters squared per second. Thus the "formula" $(fg)' = f'g'$ is asserting that some number of meters squared per second is equal to some number of meters squared per second squared. That's not only wrong, it's *a priori* meaningless.

By contrast the correct formula $(fg)' = f'g + fg'$ makes good dimensional sense: as above, $(fg)'$ is meters squared per second; so is $f'g$ and so is fg', so we can add them to get a meaningful number of meters squared per second. Thus the formula at least *makes sense*, which is good because above we proved it to be correct.

Taking these ideas more seriously suggests that we should look for a *proof* of the product rule which explicitly takes into account that both sides are rates of change of areas. This is indeed possible, but we omit it for now.

Suppose we want to find the derivative of a function which is a product of not two but three functions whose derivatives we already know, e.g. $f(x) = x \sin x e^x$. We can – of course? – still use the product rule, in two steps:

$$f'(x) = (x \sin x e^x)' = ((x \sin x)e^x)' = (x \sin x)' e^x + (x \sin x)(e^x)'$$
$$= (x' \sin x + x(\sin x)')e^x + x \sin x e^x = \sin x + x \cos x e^x + x \sin x e^x.$$

Note that we didn't use the fact that our three differentiable functions were x, $\sin x$ and e^x until the last step, so the same method shows that for any three functions f_1, f_2, f_3 which are all differentiable at x, the product $f = f_1 f_2 f_3$ is also differentiable at a and

$$f'(a) = f_1'(a) f_2(a) f_3(a) + f_1(a) f_2'(a) f_3(a) + f_1(a) f_2(a) f_3'(a).$$

The following result rides this train of thought to its final destination.

THEOREM 5.7. *(Generalized Product Rule) Let $n \geq 2$ be an integer, and let $f_1, \ldots, f_n$ be n functions which are all differentiable at a. Then $f = f_1 \cdots f_n$ is also differentiable at a, and*

$$(f_1 \cdots f_n)'(a) = f_1'(a) f_2(a) \cdots f_n(a) + \ldots + f_1(a) \cdots f_{n-1}(a) f_n'(a). \tag{18}$$

PROOF. By induction on n.
Base Case ($n = 2$): This is precisely the "ordinary" Product Rule (Theorem 5.6).
Induction Step: Let $n \geq 2$ be an integer, and suppose that the product of any n functions which are each differentiable at $a \in \mathbb{R}$ is differentiable at a and that the derivative is given by (18). Now let $f_1, \ldots, f_n, f_{n+1}$ be functions, each differentiable at a. Then by the usual product rule

$$(f_1 \cdots f_n f_{n+1})'(a) = ((f_1 \cdots f_n) f_{n+1})'(a) = (f_1 \cdots f_n)'(a) f_{n+1}(a) + f_1(a) \cdots f_n(a) f_{n+1}'(a).$$

Using the induction hypothesis this last expression becomes

$$\left(f_1'(a) f_2(a) \cdots f_n(a) + \ldots + f_1(a) \cdots f_{n-1}(a) f_n'(a)\right) f_{n+1}(a) + f_1(a) \cdots f_n(a) f_{n+1}'(a)$$

$$= f_1'(a)f_2(a)\cdots f_n(a)f_{n+1}(a) + \ldots + f_1(a)\cdots f_n(a)f_{n+1}'(a). \qquad \square$$

Example: We may use the Generalized Product Rule to give a less computationally intensive derivation of the power rule

$$(x^n)' = nx^{n-1}$$

for $n \in \mathbb{Z}^+$. Indeed, taking $f_1 = \cdots = f_n = x$, we have $f(x) = x^n = f_1 \cdots f_n$, so applying the Generalized Power rule we get

$$(x^n)' = (x)'x\cdots x + \ldots + x\cdots x(x)'.$$

Here in each term we have $x' = 1$ multiplied by $n-1$ factors of x, so each term evaluates to x^{n-1}. Moreover we have n terms in all, so

$$(x^n)' = nx^{n-1}.$$

No need to dirty our hands with binomial coefficients!

THEOREM 5.8. *(Generalized Leibniz Rule) Let $n \in \mathbb{Z}^+$, and let f, g be functions which are each n times differentiable at $a \in \mathbb{R}$. Then fg is is n times differentiable at a and*

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}.$$

Exercise: Prove Theorem 5.8.

THEOREM 5.9. *(Quotient Rule) Let f and g be functions which are both differentiable at $a \in \mathbb{R}$, with $g(a) \neq 0$. Then $\frac{f}{g}$ is differentiable at a and*

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{g(a)^2}.$$

PROOF. Step 0: First observe that since g is continuous and $g(a) \neq 0$, there is some interval $I = (a-\delta, a+\delta)$ about a on which g is nonzero, and on this interval $\frac{f}{g}$ is defined. Thus it makes sense to consider the difference quotient

$$\frac{f(a+h)/g(a+h) - f(a)/g(a)}{h}$$

for h sufficiently close to zero.
Step 1: We first establish the **Reciprocal Rule**, i.e., the special case of the Quotient Rule in which $f(x) = 1$ (constant function). Then

$$\left(\frac{1}{g}\right)'(a) = \lim_{h\to 0} \frac{\frac{1}{g(a+h)} - \frac{1}{g(a)}}{h}$$

$$= \lim_{h\to 0} \frac{g(a) - g(a+h)}{hg(a)g(a+h)} = -\left(\lim_{h\to 0} \frac{g(a+h) - g(a)}{h}\right)\left(\lim_{h\to 0} \frac{1}{g(a)g(a+h)}\right) = \frac{-g'(a)}{g(a)^2}.$$

We have again used the fact that if g is differentiable at a, g is continuous at a.
Step 2: We now derive the full Quotient Rule by combining the Product Rule and the Reciprocal Rule. Indeed, we have

$$\left(\frac{f}{g}\right)'(a) = \left(f \cdot \frac{1}{g}\right)'(a) = f'(a)\frac{1}{g(a)} + f(a)\left(\frac{1}{g}\right)'(a)$$

$$= \frac{f'(a)}{g(a)} - f(a)\frac{g'(a)}{g(a)^2} = \frac{g(a)f'(a) - g'(a)f(a)}{g(a)^2}.$$

$\square$

LEMMA 5.10. *Let $f : D \subset \mathbb{R} \to \mathbb{R}$. Suppose:*
(i) $\lim_{x\to a} f(x)$ exists, and
(ii) There exists a number $L \in \mathbb{R}$ such that for all $\delta > 0$, there exists at least one x with $0 < |x - a| < \delta$ such that $f(x) = L$.
Then $\lim_{x\to a} f(x) = L$.

PROOF. Left as an exercise. (Suggestion: suppose $\lim_{x\to a} f(x) = M \neq L$, and derive a contradiction by taking ϵ to be small enough compared to $|M - L|$.) □

THEOREM 5.11. *(Chain Rule) Let f and g be functions, and let $a \in \mathbb{R}$ be such that f is differentiable at a and g is differentiable at $f(a)$. Then the composite function $g \circ f$ is differentiable at a and*

$$(g \circ f)'(a) = g'(f(a))f'(a).$$

PROOF. Motivated by Leibniz notation, it is tempting to argue as follows:

$$(g \circ f)'(a) = \lim_{x\to a} \frac{g(f(x)) - g(f(a))}{x - a} = \lim_{x\to a} \left(\frac{g(f(x)) - g(f(a))}{f(x) - f(a)}\right) \cdot \left(\frac{f(x) - f(a)}{x - a}\right)$$

$$= \left(\lim_{x\to a} \frac{g(f(x)) - g(f(a))}{f(x) - f(a)}\right) \left(\lim_{x\to a} \frac{f(x) - f(a)}{x - a}\right)$$

$$= \left(\lim_{f(x)\to f(a)} \frac{g(f(x)) - g(f(a))}{f(x) - f(a)}\right) \left(\lim_{x\to a} \frac{f(x) - f(a)}{x - a}\right) = g'(f(a))f'(a).$$

The replacement of "$\lim_{x\to a} \ldots$ by $\lim_{f(x)\to f(a)} \ldots$" in the first factor above is justified by the fact that f is continuous at a. However, this argument has a **gap**: when we multiply and divide by $f(x) - f(a)$, how do we know that we are not dividing by zero?? The answer is that we cannot rule this out: it is possible for $f(x)$ to take the value $f(a)$ on arbitarily small deleted intervals around a: again, this is exactly what happens for the function $f_\alpha(x)$ of the above example near $a = 0$.

I maintain that the above gap can be mended so as to give a complete proof. The above argument is valid *unless* for all $\delta > 0$, there is x with $0 < |x - a| < \delta$ such that $f(x) - f(a) = 0$. In this case, it follows from Lemma 5.10 that if

$$\lim_{x\to a} \frac{f(x) - f(a)}{x - a}$$

exists at all, it must be equal to 0. But we are *assuming* the above limit exists, since we are assuming f is differentiable at a. So $f'(a) = 0$, and therefore, since the Chain Rule reads $(g \circ f)'(a) = g'(f(a))f'(a)$, here we are trying to show $(g \circ f)'(a) = 0$. For $x \in \mathbb{R}$ we have two possibilities: the first is $f(x) - f(a) = 0$, in which case also $g(f(x)) - g(f(a)) = g(f(a)) - g(f(a)) = 0$, so the difference quotient is zero at these points. The second is $f(x) - f(a) \neq 0$, in which case

$$g(f(x)) - g(f(a)) = \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \cdot \frac{f(x) - f(a)}{x - a}$$

holds, and the above argument shows this expression tends to $g'(f(a))f'(a) = 0$ as $x \to a$. So *whichever holds*, the difference quotient $\frac{g(f(x))-g(f(a))}{x-a}$ is close to (or equal to!) zero. Thus the limit tends to zero no matter which alternative obtains. Somewhat more formally, if we fix $\epsilon > 0$, then the first step of the argument shows that there is $\delta > 0$ such that for all x with $0 < |x - a| < \delta$ such that $f(x) - f(a) \neq 0$, $|\frac{g(f(x))-g(f(a))}{x-a}| < \epsilon$. On the other hand, when $f(x) - f(a) = 0$,

then $|\frac{g(f(x))-g(f(a))}{x-a}| = 0$, so it is certainly less than ϵ! Therefore, all in all we have $0 < |x-a| < \delta \implies |\frac{g(f(x))-g(f(a))}{x-a}| < \epsilon$, so that

$$\lim_{x\to a} \frac{g(f(x)) - g(f(a))}{x-a} = 0 = g'(f(a))f'(a).$$

□

3. Optimization

3.1. Intervals and interior points.

At this point I wish to digress to formally define the notion of an **interval** on the real line and and **interior point** of the interval. ...

3.2. Functions increasing or decreasing at a point.

Let $f : D \to \mathbb{R}$ be a function, and let a be an interior point of D. We say that f is **increasing at a** if for all x sufficiently close to a and to the left of a, $f(x) < f(a)$ and for all x sufficiently close to a and to the right of a, $f(x) > f(a)$. More formally phrased, we require the existence of a $\delta > 0$ such that:
• for all x with $a - \delta < x < a$, $f(x) < f(a)$, and
• for all x with $a < x < a + \delta$, $f(x) > f(a)$.

We say f is **decreasing at a** if there exists $\delta > 0$ such that:
• for all x with $a - \delta < x < a$, $f(x) > f(a)$, and
• for all x with $a < x < a + \delta$, $f(x) < f(a)$.

We say f is **weakly increasing at a** if there exists $\delta > 0$ such that:
• for all x with $a - \delta < x < a$, $f(x) \leq f(a)$, and
• for all x with $a < x < a + \delta$, $f(x) \geq f(a)$.

Exercise: Give the definition of "f is decreasing at a".

Exercise: Let $f : I \to \mathbb{R}$, and let a be an interior point of I.
a) Show that f is increasing at a iff $-f$ is decreasing at a.
b) Show that f is weakly increasing at a iff $-f$ is weakly decreasing at a.

Example: Let $f(x) = mx + b$ be the general linear function. Then for any $a \in \mathbb{R}$: f is increasing at a iff $m > 0$, f is weakly increasing at a iff $m \geq 0$, f is decreasing at a iff $m < 0$, and f is weakly decreasing at a iff $m \leq 0$.

Example: Let n be a positive integer, let $f(x) = x^n$. Then:
If x is odd, then for all $a \in \mathbb{R}$, $f(x)$ is increasing at a.
If x is even, then if $a < 0$, $f(x)$ is decreasing at a, if $a > 0$ then $f(x)$ is increasing at a. Note that when n is even f is *neither* increasing at 0 nor decreasing at 0 because for every nonzero x, $f(x) > 0 = f(0)$.[1]

[1]We do not stop to prove these assertions as it would be inefficient to do so: soon enough we will develop the right tools to prove stronger assertions. But when given a new definition, it is always good to find one's feet by considering some examples and nonexamples of that definition.

If one looks back at the previous examples and keeps in mind that we are supposed to be studying derivatives (!), one is swiftly led to the following fact.

THEOREM 5.12. *Let $f : I \to \mathbb{R}$. Suppose f is differentiable at $a \in I^\circ$.*
a) If $f'(a) > 0$, then f is increasing at a.
b) If $f'(a) < 0$, then f is decreasing at a.
c) If $f'(a) = 0$, then no conclusion can be drawn: f may be increasing at a, decreasing at a, or neither.

PROOF. a) We use the ϵ-δ interpretation of differentiability at a to our advantage. Namely, take $\epsilon = f'(a)$: there exists $\delta > 0$ such that for all x with $0 < |x - a| < \delta$, $|\frac{f(x)-f(a)}{x-a} - f'(a)| < f'(a)$, or equivalently

$$0 < \frac{f(x) - f(a)}{x - a} < 2f'(a).$$

In particular, for all x with $0 < |x-a| < \delta$, $\frac{f(x)-f(a)}{x-a} > 0$, so: if $x > a$, $f(x)-f(a) > 0$, i.e., $f(x) > f(a)$; and if $x < a$, $f(x) - f(a) < 0$, i.e., $f(x) < f(a)$.
b) This is similar enough to part a) to be best left to the reader as an exercise.[2]
c) If $f(x) = x^3$, then $f'(0) = 0$ but f is increasing at 0. If $f(x) = -x^3$, then $f'(0) = 0$ but f is decreasing at 0. If $f(x) = x^2$, then $f'(0) = 0$ but f is neither increasing nor decreasing at 0. □

3.3. Extreme Values.

Let $f : D \to \mathbb{R}$. We say $M \in \mathbb{R}$ is the **maximum value** of f on D if
(MV1) There exists $x \in D$ such that $f(x) = M$, and
(MV2) For all $x \in D$, $f(x) \leq M$.

It is clear that a function can have at most one maximum value: if it had more than one, one of the two would be larger than the other! However a function need not have any maximum value: for instance $f : (0, \infty) \to \mathbb{R}$ by $f(x) = \frac{1}{x}$ has no maximum value: $\lim_{x \to 0^+} f(x) = \infty$.

Similarly, we say $m \in \mathbb{R}$ is the **minimum value** of f on D if
(mV1) There exists $x \in D$ such that $f(x) = m$, and
(mV2) For all $x \in D$, $f(x) \geq m$.

Also, a function can have at most one minimum value but need not have any. The function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ by $f(x) = \frac{1}{x}$ has no minimum value: $\lim_{x \to 0^-} f(x) = -\infty$.

Exercise: For a function $f : D \to \mathbb{R}$, the following are equivalent:
(i) f assumes a maximum value M, a minimum value m, and $M = m$.
(ii) f is a constant function.

Recall that $f : D \to \mathbb{R}$ is **bounded above** if there is $B \in \mathbb{R}$ such that for all $x \in D$, $f(x) \leq B$. A function is **bounded below** if there is $b \in \mathbb{R}$ such that for all $x \in D$, $f(x) \geq b$. A function is **bounded** if it is both bounded above and bounded

[2]Two ways to go: (i) Revisit the above proof flipping inequalities as appropriate. (ii) Use the fact that f is decreasing at a iff $-f$ is increasing at a and $f'(a) < 0 \iff (-f)'(a) > 0$ to apply the *result* of part a).

below: equivalently, there exists $B \geq 0$ such that for all $x \in D$, $|f(x)| \leq B$: i.e., the graph of f is "trapped between" the horizontal lines $y = B$ and $y = -B$.

Exercise: Let $f : D \to \mathbb{R}$ be a function.
a) Show: if f has a maximum value, it is bounded above.
b) Show: if f has a minimum value, it is bounded below.

Exercise: a) If a function has both a maximum and minimum value on D, then it is bounded on D: indeed, if M is the maximum value of f and m is the minimum value, then for all $x \in B$, $|f(x)| \leq \max(|m|, |M|)$.
b) Give an example of a bounded function $f : \mathbb{R} \to \mathbb{R}$ which has neither a maximum nor a minimum value.

We say f **assumes its maximum value at a** if $f(a)$ is the maximum value of f on D, or in other words, for all $x \in D$, $f(x) \leq f(a)$. Simlarly, we say f **assumes its minimum value at a** if $f(a)$ is the minimum value of f on D, or in other words, for all $x \in D$, $f(x) \geq f(a)$.

Example: The function $f(x) = \sin x$ assumes its maximum value at $x = \frac{\pi}{2}$, because $\sin \frac{\pi}{2} = 1$, and 1 is the maximum value of the sine function. Note however that $\frac{\pi}{2}$ is not the only x-value at which f assumes its maximum value: indeed, the sine function is periodic and takes value 1 precisely at $x = \frac{\pi}{2} + 2\pi n$ for $n \in \mathbb{Z}$. Thus there may be more than one x-value at which a function attains its maximum value. Similarly f attains its minimum value at $x = \frac{3\pi}{2}$ – $f(\frac{3\pi}{2}) = -1$ and f takes no smaller values – and also at $x = \frac{3\pi}{2} + 2\pi n$ for $n \in \mathbb{Z}$.

Example: Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3 + 5$. Then f does not assume a maximum or minimum value. Indeed, $\lim_{x \to \infty} f(x) = \infty$ and $\lim_{x \to -\infty} f(x) = -\infty$.

Example: Let $f : [0, 2] \to \mathbb{R}$ be defined as follows: $f(x) = x + 1$, $0 \leq x < 1$.
$f(x) = 1$, $x = 1$.
$f(x) = x - 1$, $1 < x \leq 2$.
Then f is defined on a closed, bounded interval and is bounded above (by 2) and bounded below (by 0) but does not have a maximum or minimum value. Of course this example of a function defined on a closed bounded interval without a maximum or minimum value feels rather contrived: in particular it is *not continuous* at $x = 1$.

This brings us to one of the most important theorems in the course.

THEOREM 5.13. *(Extreme Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f has a maximum and minimum value, hence is bounded above and below.*

The proof will be given in the next chapter, following a discussion of **completeness**.

Ubiquitously in (pure and applied) mathematics we wish to **optimize** functions: that is, find their maximum and or minimum values on a certain domain. Unfortunately, as we have seen above, a general function $f : D \to \mathbb{R}$ need not have a maximum or minimum value! But the Extreme Value Theorem gives rather mild hypotheses on which these values are guaranteed to exist, and in fact is a useful tool for establishing the existence of maximia / minima in other situations as well.

Example: Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2(x-1)(x-2)$. Note that f does not have a maximum value: indeed $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} = \infty$. However, we claim that f *does have* a minimum value. We argue for this as follows: given that f tends to ∞ with $|x|$, there must exist $\Delta > 0$ such that for all x with $|x| > \Delta$, $f(x) \geq 1$. On the other hand, if we restrict f to $[-\Delta, \Delta]$ we have a continuous function on a closed bounded interval, so by the Extreme Value Theorem it must have a minimum value, say m. In fact since $f(0) = 0$, we see that $m < 0$, so in particular $m < 1$. This means that the minimum value m for f on $[-\Delta, \Delta]$ must in fact be the minimum value for f on all of $\mathbb{R}$, since at the other values – namely, on $(-\infty, -\Delta)$ and (Δ, ∞), $f(x) > 1 > 0 \geq m$.

We can be at least a little more explicit: a **sign analysis** of f shows that f is positive on $(-\infty, 1)$ and $(2, \infty)$ and negative on $(1, 2)$, so the minimum value of f will be its minimum value on $[1, 2]$, which will be strictly negative. But exactly what is this minimum value m, and for which x value(s) does it occur? Stay tuned:we are about to develop tools to answer this question!

3.4. Local Extrema and a Procedure for Optimization.

We now describe a type of "local behavior near a" of a very different sort from being increasing or decreasing at a.

Let $f : D \to \mathbb{R}$ be a function, and let $a \in D$. We say that f has a **local maximum** at a if the value of f at a is greater than or equal to its values at all sufficiently close points x. More formally: there exists $\delta > 0$ such that for all $x \in D$, $|x - a| < \delta \implies f(x) \leq f(a)$. Similarly, we say that f has a **local minimum** at a if the vaalue of f at a is greater than or equal to its values at all sufficiently close points x. More formally: there exists $\delta > 0$ such that for all $x \in D$, $|x - a| < \delta \implies f(x) \geq f(a)$.

THEOREM 5.14. *Let $f : D \subset \mathbb{R}$, and let a be an interior point of a. If f is differentiable at a and has a local extremum – i.e., either a local minimum or a local maximum – at $x = a$, then $f'(a) = 0$.*

PROOF. Indeed, if $f'(a) \neq 0$ then either $f'(a) > 0$ or $f'(a) < 0$.
If $f'(a) > 0$, then by Theorem 5.12, f is increasing at a. Thus for x slightly smaller than a, $f(x) < f(a)$, and for x slightly larger than a, $f(x) > f(a)$. So f does not have a local extremum at a.
Similarly, if $f'(a) < 0$, then by Theorem 5.12, f is decreasing at a. Thus for x slightly smaller than a, $f(x) > f(a)$, and for x slightly larger than a, $f(x) < f(a)$. So f does not have a local extremum at a. □

THEOREM 5.15. *(Optimization Theorem) Let $f : I \to \mathbb{R}$ be continuous. Suppose that f attains a minimum or maximum value at $x = a$. Then a is either:*
(i) an endpoint of I,
*(ii) a **stationary point**: $f'(a) = 0$, or*
*(iii) a **point of nondifferentiability**.*

PROOF. This is an immediate consequence of Theorem 5.14. □

Exercise: State and prove a version of Theorem 5.15 for the maximum value.

Often one lumps cases (ii) and (iii) of Theorem 5.15 together under the term **critical point** (but there is nothing very deep going on here: it's just terminology). Clearly there are always exactly two endpoints. In favorable circustances there will be only finitely many critical points, and in very favorable circumstances they can be found exactly: suppose they are $c_1, \ldots, c_n$. (There may not be *any* critical points, but that only makes things easier...) Suppose further that we can explicitly compute all the values $f(a), f(b), f(c_1), \ldots, f(c_n)$. Then **we win**: the largest of these values is the maximum value, and the smallest is the minimum value.

Example: Let $f(x) = x^2(x-1)(x-2) = x^4 - 3x^3 + 2x^2$. Above we argued that there is a Δ such that $|x| > \Delta \implies |f(x)| \geq 1$: let's find such a Δ explicitly. We intend nothing fancy here:

$$f(x) = x^4 - 3x^2 + 2x^2 \geq x^4 - 3x^3 = x^3(x-3).$$

So if $x \geq 4$, then

$$x^3(x-3) \geq 4^3 \cdot 1 = 64 \geq 1.$$

On the other hand, if $x < -1$, then $x < 0$, so $-3x^3 > 0$ and thus

$$f(x) \geq x^4 + 2x^2 = x^2(x^2+2) \geq 1 \cdot 3 = 3.$$

Thus we may take $\Delta = 4$.

Now let us try the procedure of Theorem 5.15 out by finding the maximum and minimum values of $f(x) = x^4 - 3x^3 + 2x^2$ on $[-4, 4]$.

Since f is differentiable everywhere on $(-4, 4)$, the only critical points will be the stationary points, where $f'(x) = 0$. So we compute the derivative:

$$f'(x) = 4x^3 - 9x^2 + 4x = x(4x^2 - 9x + 4).$$

The roots are $x = \frac{9 \pm \sqrt{17}}{8}$, or, approximately,

$$x_1 \approx 0.6094\ldots, \ x_2 = 1.604\ldots.$$

$$f(x_1) = 0.2017\ldots, \ f(x_2) = -0.619\ldots.$$

Also we always test the endpoints:

$$f(-4) = 480, \ f(4) = 96.$$

So the maximum value is 480 and the minimum value is $-.619\ldots$.

4. The Mean Value Theorem

4.1. Statement of the Mean Value Theorem.

Our goal in this section is to prove the following important result.

THEOREM 5.16. *(Mean Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there is at least one c with $a < c < b$ and*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Remark:[3] I still remember the calculus test I took in high school in which I was asked to state the Mean Value Theorem. It was a multiple choice question, and I didn't see the choice I wanted, which was as above except with the subtly stronger assumption that $f'_R(a)$ and $f'_L(b)$ exist: i.e., f is one-sided differentiable at both endpoints. So I went up to the teacher's desk to ask about this. He thought for a moment and said, "Okay, you can add that as an answer if you want", and so as not to give special treatment to any one student, he announced to all that he was adding a possible answer to the Mean Value Theorem question. So I marked my added answer, did the rest of the exam, and then had time to come back to this question. After more thought I decided that one-sided differentiability at the endpoints was not required. So in the end I selected this pre-existing choice and submitted my exam. As you can see, my final answer was correct. **But** many other students figured that if I had successfully lobbied for an additional answer then "my" answer was probably correct, so they changed their answer from the correct answer to my added answer. They were not so thrilled with me or the teacher, but in my opinion he behaved admirably: talk about a "teachable moment"!

One should certainly draw a picture to go with the Mean Value Theorem, as it has a very simple geometric interpretation: under the hypotheses of the theorem, there exists at least one interior point c of the interval such that the tangent line at c is parallel to the secant line joining the endpoints of the interval.

And one should also interpret it physically: if $y = f(x)$ gives the position of a particle at a given time x, then the expression $\frac{f(b)-f(a)}{b-a}$ is nothing less than the average velocity between time a and time b, whereas the derivative $f'(c)$ is the instantaneous velocity at time c, so that the Mean Value Theorem says that there is at least one instant at which the instantaneous velocity is equal to the average velocity.

Example: Suppose that cameras are set up at checkpoints along an interstate highway in Georgia. One day you receive timestamped photos of yourself at two checkpoints. The two checkpoints are 90 miles apart and the second photo is taken 73 minutes after the first photo. You are issued a ticket for violating the speeed limit of 70 miles per hour. Your average velocity was (90 miles) / (73 minutes) $\cdot$ (60 minutes) / (hour) $\approx$ 73.94 miles per hour. Thus, although no one saw you violating the speed limit, they may mathematically deduce that at some point your instantaneous velocity was over 70 mph. Guilt by the Mean Value Theorem!

4.2. Proof of the Mean Value Theorem.

We will deduce the Mean Value Theorem from the (as yet unproven) Extreme Value Theorem. However, it is convenient to first establish a special case.

THEOREM 5.17. *(Rolle's Theorem) Let $f : [a, b] \to \mathbb{R}$. We suppose:*
(i) f is continuous on $[a, b]$.
(ii) f is differentiable on (a, b).
(iii) $f(a) = f(b)$.
Then there exists c with $a < c < b$ and $f'(c) = 0$.

PROOF. By Theorem 5.13, f has a maximum M and a minimum m.
Case 1: Suppose $M > f(a) = f(b)$. Then the maximum value does not occur

[3]Please excuse this personal anecdote.

at either endpoint. Since f is differentiable on (a,b), it must therefore occur at a stationary point: i.e., there exists $c \in (a,b)$ with $f'(c) = 0$.
Case 2: Suppose $m < f(a) = f(b)$. Then the minimum value does not occur at either endpoint. Since f is differentiable on (a,b), it must therefore occur at a stationary point: there exists $c \in (a,b)$ with $f'(c) = 0$.
Case 3: The remaining case is $m = f(a) = M$. Then f is constant and $f'(c) = 0$ at *every* point $c \in (a,b)$. □

To deduce the Mean Value Theorem from Rolle's Theorem, it is tempting to tilt our head until the secant line from $(a, f(a))$ to $(b, f(b))$ becomes horizontal and then apply Rolle's Theorem. The possible flaw here is that if we start a subset in the plane which is the graph of a function and rotate it too much, it may no longer be the graph of a function, so Rolle's Theorem does not apply.

The above objection is just a technicality. In fact, it suggests that more is true: there should be some version of the Mean Value Theorem which applies to curves in the plane which are not necessarily graphs of functions. Nevertheless formalizing this argument is more of a digression than we want to make, so the "official" proof that follows is (slightly) different.

Proof of the Mean Value Theorem: Let $f : [a,b] \to \mathbb{R}$ be continuous on $[a,b]$ and differentiable on (a,b). There is a unique linear function $L(x)$ such that $L(a) = f(a)$ and $L(b) = f(b)$: indeed, L is nothing else than the secant line to f between $(a, f(a))$ and $(b, f(b))$. Here's the trick: by subtracting $L(x)$ from $f(x)$ we reduce ourselves to a situation where we may apply Rolle's Theorem, and then the conclusion that we get is easily seen to be the one we want about f. Here goes: define

$$g(x) = f(x) - L(x).$$

Then g is defined and continuous on $[a,b]$, differentiable on (a,b), and $g(a) = f(a) - L(a) = f(a) - f(a) = 0 = f(b) - f(b) = f(b) - L(b) = g(b)$. Applying Rolle's Theorem to g, there exists $c \in (a,b)$ such that $g'(c) = 0$. On the other hand, since L is a linear function with slope $\frac{f(b)-f(a)}{b-a}$, we compute

$$0 = g'(c) = f'(c) - L'(c) = f'(c) - \frac{f(b) - f(a)}{b-a},$$

and thus

$$f'(c) = \frac{f(b) - f(a)}{b-a}.$$

4.3. The Cauchy Mean Value Theorem.

We present here a generalization of the Mean Value Theorem due to A.L. Cauchy. Although perhaps not as fundamental and physically appealing as the Mean Value Theorem, it has its place: later we will use it to prove L'Hôpital's Rule.

THEOREM 5.18. *(Cauchy Mean Value Theorem) Let $f, g : [a,b] \to \mathbb{R}$ be continuous and differentiable on (a,b). Then there exists $c \in (a,b)$ such that*

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c). \tag{19}$$

PROOF. Case 1: Suppose $g(a) = g(b)$. By Rolle's Theorem, there is $c \in (a,b)$ such that $g'(c) = 0$. For this c, both sides of (19) are zero, hence they are equal.

Case 2: Suppose $g(a) \neq g(b)$, and define

$$h(x) = f(x) - \left(\frac{f(b)-f(a)}{g(b)-g(a)}\right) g(x).$$

Then h is continuous on $[a,b]$, differentiable on (a,b), and

$$h(a) = \frac{f(a)(g(b)-g(a)) - g(a)(f(b)-f(a))}{g(b)-g(a)} = \frac{f(a)g(b)-g(a)f(b)}{g(b)-g(a)},$$

$$h(b) = \frac{f(b)(g(b)-g(a)) - g(b)(f(b)-f(a))}{g(b)-g(a)} = \frac{f(a)g(b)-g(a)f(b)}{g(b)-g(a)},$$

so $h(a) = h(b)$.[4] By Rolle's Theorem there exists $c \in (a,b)$ with

$$0 = h'(c) = f'(c) - \left(\frac{f(b)-f(a)}{g(b)-g(a)}\right) g'(c),$$

or equivalently,

$$(f(b)-f(a))g'(c) = (g(b)-g(a))f'(c).$$

□

Exercise: Which choice of g recovers the "ordinary" Mean Value Theorem?

The Cauchy Mean Value Theorem can be used to vindicate the "head-tilting argument" contemplated before the formal proof of the Mean Value Theorem. To do so takes us a bit outside our wheelhouse (i.e., a seriously theoretical approach to single variable calculus), but it is interesting enough to be worth sketching here.

First recall (if possible) the notion of a **parameterized curve**: this is given by a function f with domain an interval I in $\mathbb{R}$ and codomain the plane $\mathbb{R}^2$: we may write $v : t \mapsto (x(t), y(t))$, where $t \mapsto x(t)$ and $y \mapsto y(t)$ are each functions from I to $\mathbb{R}$. We say that v is continuous at t if both x and y are continuous at t. Further, when both x and y are differentiable at t we decree that v is differentiable at t and set $v'(t) = (x'(t), y'(t))$, which we think of as being a *vector* in the plane with tail at the point $(x(t), y(t))$. When $x'(t) = y'(t) = 0$ we get the zero vector, which we don't want: let's call this a **singular point**. For any nonsingular point $(x(t), y(t))$ we can define the **tangent line** to be the unique line passing through the points $(x(t), y(t))$ and $(x(t) + x'(t), y(t) + y'(t))$. (When $x'(t) \neq 0$, we can define the tangent line in a more familiar way, as the unique line passing through $(x(t), y(t))$ with slope $\frac{y'(t)}{x'(t)}$; the above definition is phrased so as to make sense also if $x'(t) = 0$ and $y'(t) \neq 0$, in which case we get a vertical tangent line.)

Example: Let $v : [0, 2\pi] \to \mathbb{R}^2$ by $v(t) = (\cos t, \sin t)$. This point traces out a path which winds once around the unit circle counterclockwise. We have $v'(t) = (-\sin t, \cos t)$: since for all t, $(-\sin t)^2 + (\cos t)^2 = 1$, in particular $v'(t)$ is never zero: there are no singular points. At the points $t = 0, \pi, 2\pi$, $x'(t) = 0$ and the tangent line is vertical at the points $(\pm 1, 0)$ (as you would certainly expect from contemplation of the unit circle). When $t = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi$, one of $(x(t), y(t))$ and $(x'(t), y'(t))$ is horizontal and the other is vertical, so in particular the tangent line at $v(t)$ is perpendicular to $v(t)$. This holds at all other points as well, e.g. by noting that the slope of the radial line from $(0,0)$ to $v(t))$ is $\frac{y(t)}{x(t)} = \frac{\sin t}{\cos t}$ and the slope of

[4] Don't be so impressed: we wanted a constant C such that if $h(x) = f(x) - Cg(x)$, then $h(a) = h(b)$, so we set $f(a) - Cg(a) = f(b) - Cg(b)$ and solved for C.

the tangent line to $v(t)$ is $\frac{y'(t)}{x'(t)} = \frac{-\cos t}{\sin t}$ and observing that these two slopes are opposite reciprocals. Thus we get a derivation using calculus of a familiar (or so I hope!) fact from elementary geometry.

Here is a version of the Mean Value Theorem for nonsingular parameterized curves.

THEOREM 5.19. *(Parameteric Mean Value Theorem) Let $v : I \to \mathbb{R}^2$, $t \mapsto (x(t), y(t))$, be a nonsingular parametrized curve in the plane, i.e., for all $t \in I$, $x'(t)$ and $y'(t)$ both exist and are not both 0. Let $a < b \in I$, and suppose that either $x(a) \neq x(b)$ or $y(a) \neq y(b)$. Then there is $c \in (a, b)$ such that the tangent vector $(x'(c), y'(c))$ at c is parallel to the secant line from $(a, f(a))$ to $(b, f(b))$.*

PROOF. We apply the Cauchy Mean Value Theorem with $f = x$ and $g = y$: there is $c \in (a, b)$ such that

$$(x(b) - x(a))y'(c) = (y(b) - y(a))x'(c).$$

By assumption, $x'(c)$ and $y'(c)$ are not both 0.
Case 1: Suppose $x'(c) \neq 0$. Then we must have $x(b) - x(a) \neq 0$: if not,

$$0 = (x(b) - x(a))y'(c) = (y(b) - y(a))x'(c)$$

and thus, since $x'(c) \neq 0$, $y(b) - y(a) \neq 0$, and then we have $x(a) = x(b)$ and $y(a) = y(b)$, contrary to our hypothesis. So we may divide to get

$$\frac{y'(c)}{x'(c)} = \frac{y(b) - y(a)}{x(b) - x(a)},$$

which says that the tangent line to v at c has the same slope as the secant line between $(x(a), y(a))$ and $(x(b), y(b))$, hence these two lines are parallel.
Case 2: Suppose $x'(c) = 0$, i.e., the tangent line to v at c is vertical. By our nonsingularity assumption $y'(c) \neq 0$, so

$$0 = (y(b) - y(a))x'(c) = (x(b) - x(a))y'(c)$$

implies $x(a) = x(b)$, so the secant line from $(x(a), x(b))$ to $(y(a), y(b))$ is vertical, hence the two lines are parallel. □

Conversely, it is possible (indeed, similar) to deduce the Cauchy Mean Value Theorem from the Parametric Mean Value Theorem: try it. Thus really we have one theorem with two moderately different phrasings, and indeed it is common to also refer to Theorem 5.19 as the Cauchy Mean Value Theorem.

5. Monotone Functions

A function $f : I \to \mathbb{R}$ is **monotone** if it is weakly increasing or weakly decreasing.

5.1. The Monotone Function Theorems.

The Mean Value Theorem has several important consequences. Foremost of all it will be used in the proof of the Fundamental Theorem of Calculus, but that's for later. At the moment we can use it to establish a criterion for a function f to be monotone on an interval in terms of sign condition on f'.

THEOREM 5.20. *(First Monotone Function Theorem) Let I be an open interval, and let $f : I \to \mathbb{R}$ be a function which is differentiable on I.*
a) Suppose $f'(x) > 0$ for all $x \in I$. Then f is increasing on I: for all $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) < f(x_2)$.
b) Suppose $f'(x) \geq 0$ for all $x \in I$. Then f is weakly increasing on I: for all $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \leq f(x_2)$.
c) Suppose $f'(x) < 0$ for all $x \in I$. Then f is decreasing on I: for all x_1, x_2 in$I$ with $x_1 < x_2$, $f(x_1) > f(x_2)$.
d) Suppose $f'(x) \leq 0$ for all $x \in I$. Then f is weakly decreasing on I: for all $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \geq f(x_2)$.

PROOF. a) We go by contraposition: suppose that f is *not* increasing: then there exist $x_1, x_2 \in I$ with $x_1 < x_2$ such that $f(x_1) \geq f(x_2)$. Apply the Mean Value Theorem to f on $[x_1, x_2]$: there exists $x_1 < c < x_2$ such that $f'(c) = \frac{f(x_2)-f(x_1)}{x_2-x_1} \leq 0$.
b) Again, we argue by contraposition: suppose that f is *not* weakly increasing: then there exist $x_1, x_2 \in I$ with $x_1 < x_2$ such that $f(x_1) > f(x_2)$. Apply the Mean Value Theorem to f on $[x_1, x_2]$: there exists $x_1 < c < x_2$ such that $f'(c) = \frac{f(x_2)-f(x_1)}{x_2-x_1} < 0$.
c),d) We leave these proofs to the reader. One may either proceed exactly as in parts a) and b), or reduce to them by multiplying f by -1. □

COROLLARY 5.21. *(Zero Velocity Theorem) Let $f : I \to \mathbb{R}$ be a differentiable function with identically zero derivative. Then f is constant.*

PROOF. Since $f'(x) \geq 0$ for all $x \in I$, f is weakly increasing on I: $x_1 < x_2 \implies f(x_1) \leq f(x_2)$. Since $f'(x) \leq 0$ for all $x \in I$, f is weakly decreasing on I: $x_1 < x_2 \implies f(x_1) \geq f(x_2)$. But a function which is weakly increasing *and* weakly decreasing satsifies: for all $x_1 < x_2$, $f(x_1) \leq f(x_2)$ and $f(x_1) \geq f(x_2)$ and thus $f(x_1) = f(x_2)$: f is constant. □

Remark: Above, we deduced Corollary 5.21 from Theorem 5.20. If we had argued directly from the Mean Value Theorem the proof would have been shorter: try it!

COROLLARY 5.22. *Suppose $f, g : I \to \mathbb{R}$ are both differentiable and such that $f' = g'$ (equality as functions, i.e., $f'(x) = g'(x)$ for all $x \in I$). Then there exists a constant $C \in \mathbb{R}$ such that $f = g + C$, i.e., for all $x \in I$, $f(x) = g(x) + C$.*

PROOF. Let $h = f - g$. Then $h' = (f - g)' = f' - g' \equiv 0$, so by Corollary 5.21, $h \equiv C$ and thus $f = g + h = g + C$. □

Remark: Corollary 5.22 can be viewed as a *uniqueness theorem* for differential equations. Let $f : I \to \mathbb{R}$ be a function, and consider the set of all functions $F : I \to \mathbb{R}$ such that $F' = f$. Then Corollary 5.22 asserts that *if* there is a function F such that $F' = f$, then there is a **one-parameter family** of such functions, and more specifically that the general such function is of the form $F + C$.

On the other hand, the **existence** question lies deeper: namely, given $f : I \to \mathbb{R}$, must there exist $F : I \to \mathbb{R}$ such that $F' = f$? In general the answer is *no*.

Exercise: Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = 0$ for $x \leq 0$ and $f(x) = 1$ for $x > 0$. Show that there is no function $F : \mathbb{R} \to \mathbb{R}$ such that $F' = f$.

In other words, not every function $f : \mathbb{R} \to \mathbb{R}$ has an **antiderivative**, i.e., is

the derivative of some other function. It turns out that every *continuous* function has an antiderivative: this will be proved later. (Much more subtly, there are also *some* discontinuous functions which have antiderivatives...)

COROLLARY 5.23. *Let $f : I \to \mathbb{R}$ be a function whose nth derivative $f^{(n)}$ is identically zero. Then f is a polynomial function of degree at most $n-1$.*

PROOF. Exercise. (Hint: use induction.) □

The setting of the Increasing Function Theorem is that of a differentiable function defined on an *open* interval I. This is just a technical convenience: for continuous functions, the increasing / decreasing / weakly increasing / weakly decreasing behavior on the interior of I implies the same behavior at an endpoint of I.

THEOREM 5.24. *Let $f : [a,b] \to \mathbb{R}$ be continuous at $x = a$ and $x = b$.*
a) If f is weakly increasing on (a,b), it is weakly increasing on $[a,b]$.
b) If f is incereasing on (a,b), it is increasing on $[a,b]$.

PROOF. Step 1: Suppose that f is continuous at a and weakly increasing on (a,b). We will show that f is weakly increasing on $[a,b)$. Indeed, assume not: then there exists $x_0 \in (a,b)$ such that $f(a) > f(x_0)$. Now take $\epsilon = f(a) - f(x_0)$; since f is (right-)continuous at a, there exists $\delta > 0$ such that for all $a \leq x < a + \delta$, $|f(x) - f(a)| < f(a) - f(x_0)$, which implies $f(x) > f(x_0)$. By taking $a < x < x_0$, this contradicts the assumption that f is weakly increasing on (a,b).
Step 2: Suppose that f is continuous at a and increasing on (a,b). We will show that f is increasing on $[a,b)$. Note first that Step 1 applies to show that $f(a) \leq f(x)$ for all $x \in (a,b)$, but we want slightly more than this, namely strict inequality. So, seeking a contradiction, we suppose that $f(a) = f(x_0)$ for some $x_0 \in (a,b)$. But now take $x_1 \in (a,x_0)$: since f is increasing on (a,b) we have $f(x_1) < f(x_0) = f(a)$, contradicting the fact that f is weakly increasing on $[a,b)$.
Step 3: In a similar way one can handle the right endpoint b. Now suppose that f is increasing on $[a,b)$ and also increasing on $(a,b]$. It remains to show that f is increasing on $[a,b]$. The only thing that could go wrong is $f(a) \geq f(b)$. To see that this cannot happen, choose any $c \in (a,b)$: then $f(a) < f(c) < f(b)$. □

Exercise: Show that we may replace each instance of "increasing" in Theorem 5.24 with "decreasing" (and still get a true statement!).

THEOREM 5.25. *(Second Monotone Function Theorem) Let $f : I \to \mathbb{R}$ be a function which is continuous on I and differentiable on the interior I° of I (i.e., at every point of I except possibly at any endpoints I may have).*
a) The following are equivalent:
(i) f is monotone.
(ii) Either we have $f'(x) \geq 0$ for all $x \in I^\circ$ or $f'(x) \leq 0$ for all $x \in I^\circ$.
b) Suppose f is monotone. The following are equivalent:
(i) f is not *increasing or decreasing.*
(ii) There exist $a, b \in I^\circ$ with $a < b$ such that the restriction of f to $[a,b]$ is constant.
(iii) There exist $a, b \in I^\circ$ with $a < b$ such that $f'(x) = 0$ for all $x \in [a,b]$.

PROOF. Throughout the proof we restrict our attention to increasing / weakly increasing functions, leaving the other case to the reader as a routine exercise.
a) (i) $\implies$ (ii): Suppose f is weakly increasing on I. We claim $f'(x) \geq 0$ for all $x \in I^\circ$. If not, there is $a \in I^\circ$ with $f'(a) < 0$. Then f is decreasing at a, so there

exists $b > a$ with $f(b) < f(a)$, contradicting the fact that f is weakly decreasing.
(ii) $\implies$ (i): Immediate from the Increasing Function Theorem and Theorem 5.24.
b) (i) $\implies$ (ii): Suppose f is weakly increasing on I but not increasing on I. By Theorem 5.24 f is still not increasing on I°, so there exist $a, b \in I^\circ$ with $a < b$ such that $f(a) = f(b)$. Then, since f is weakly increasing, for all $c \in [a, b]$ we have $f(a) \leq f(c) \leq f(b) = f(a)$, so f is constant on $[a, b]$.
(ii) $\implies$ (iii): If f is constant on $[a, b]$, f' is identically zero on $[a, b]$.
(iii) $\implies$ (i): If f' is identically zero on some subinterval $[a, b]$, then by the Zero Velocity Theorem f is constant on $[a, b]$, hence is not increasing. □

The next result follows immediately.

COROLLARY 5.26. *Let $f : I \to \mathbb{R}$. Suppose $f'(x) \geq 0$ for all $x \in I$, and $f'(x) > 0$ except at a finite set of points $x_1, \ldots, x_n$. Then f is increasing on I.*

Example: A typical application of Theorem 5.26 is to show that the function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3$ is increasing on all of $\mathbb{R}$. Indeed, $f'(x) = 3x^2$ which is strictly positive at all $x \neq 0$ and 0 at $x = 0$.

5.2. The First Derivative Test.

We can use Theorem 5.24 to quickly derive another staple of freshman calculus.

THEOREM 5.27. *(First Derivative Test) Let I be an interval, a an interior point of I, and $f : I \to \mathbb{R}$ a function. We suppose that f is continuous on I and differentiable on $I \setminus \{a\}$. Then:*
a) If there exists $\delta > 0$ such that $f'(x)$ is negative on $(a - \delta, a)$ and is positive on $(a, a + \delta)$. Then f has a strict local minimum at a.
b) If there exists $\delta > 0$ such that $f'(x)$ is positive on $(a - \delta, a)$ and is negative on $(a, a + \delta)$. Then f has a strict local maximum at a.

PROOF. a) By the First Monotone Function Theorem, since f' is negative on the open interval $(a-\delta, a)$ and positive on the open interval $(a, a+\delta)$ f is decreasing on $(a - \delta, a)$ and increasing on $(a, a + \delta)$. Moreover, since f is differentiable on its entire domain, it is continuous at $a-\delta$, a and $a+\delta$, and thus Theorem 5.24 applies to show that f is decreasing on $[a - \delta, a]$ and increasing on $[a, a + \delta]$. This gives the desired result, since it implies that $f(a)$ is strictly smaller than $f(x)$ for any $x \in [a - \delta, a)$ or in $(a, a + \delta]$.
b) As usual this may be proved either by revisiting the above argument or deduced directly from the result of part a) by multiplying f by -1. □

Remark: This version of the First Derivative Test is a little stronger than the familiar one from freshman calculus in that we have not assumed that $f'(a) = 0$ nor even that f is differentiable at a. Thus for instance our version of the test applies to $f(x) = |x|$ to show that it has a strict local minimum at $x = 0$.

5.3. The Second Derivative Test.

THEOREM 5.28. *(Second Derivative Test) Let a be an interior point of an interval I, and let $f : I \to \mathbb{R}$. We suppose:*
(i) f is twice differentiable at a, and
(ii) $f'(a) = 0$.
Then if $f''(a) > 0$, f has a strict local minimum at a, whereas if $f''(a) < 0$, f has a strict local maximum at a.

PROOF. As usual it suffices to handle the case $f''(a) > a$.
Notice that the hypothesis that f is twice differentiable at a implies that f is differentiable on some interval $(a-\delta, a+\delta)$ (otherwise it would not be meaningful to talk about the derivative of f' at a). Our strategy will be to show that for sufficiently small $\delta > 0$, $f'(x)$ is negative for $x \in (a-\delta, a)$ and positive for $x \in (a, a+\delta)$ and then apply the First Derivative Test. To see this, consider

$$f''(a) = \lim_{x \to a} \frac{f'(x) - f'(a)}{x - a} = \lim_{x \to a} \frac{f'(x)}{x - a}.$$

We are assuming that this limit exists and is positive, so that there exists $\delta > 0$ such that for all $x \in (a-\delta, a) \cup (a, a+\delta)$, $\frac{f'(x)}{x-a}$ is positive. And this gives us exactly what we want: suppose $x \in (a-\delta, a)$. Then $\frac{f'(x)}{x-a} > 0$ and $x - a < 0$, so $f'(x) < 0$. On the other hand, suppose $x \in (a, a+\delta)$. Then $\frac{f'(x)}{x-a} > 0$ and $x - a > 0$, so $f'(x) > 0$. So f has a strict local minimum at a by the First Derivative Test. □

Remark: When $f'(a) = f''(a) = 0$, no conclusion can be drawn about the local behavior of f at a: it may have a local minimum at a, a local maximum at a, be increasing at a, decreasing at a, or none of the above.

5.4. Sign analysis and graphing.

When one is graphing a function f, the features of interest include number and approximate locations of the roots of f, regions on which f is positive or negative, regions on which f is increasing or decreasing, and local extrema, if any. For these considerations one wishes to do a **sign analysis** on both f and its derivative f'.

Let us agree that a **sign analysis** of a function $g : I \to \mathbb{R}$ is the determination of regions on which g is positive, negative and zero.

The basic strategy is to determine first the set of roots of g. As discussed before, finding exact values of roots may be difficult or impossible even for polynomial functions, but often it is feasible to determine at least the number of roots and their approximate location (certainly this is possible for all polynomial functions, although this requires justification that we do not give here). The next step is to test a point in each region between consecutive roots to determine the sign.

This procedure comes with two implicit assumptions. Let us make them explicit.

The first is that the roots of f are sparse enough to separate the domain I into "regions". One precise formulation of of this is that f has only finitely many roots on any bounded subset of its domain. This holds for all the elementary functions we know and love, but certainly not for all functions, even all differentiable functions: we have seen that things like $x^2 \sin(\frac{1}{x})$ are not so well-behaved. But this is a convenient assumption and in a given situation it is usually easy to see whether it holds.

The second assumption is more subtle: it is that if a function f takes a positive value at some point a and a negative value at some other point b then it must take the value zero somewhere in between. Of course this does not hold for all functions: it fails very badly, for instance, for the function f which takes the value

1 at every rational number and -1 at every irrational number.

Let us formalize the desired property and then say which functions satisfy it.

A function $f : I \to \mathbb{R}$ has the **intermediate value property** if for all $a, b \in I$ with $a < b$ and all L *in between* $f(a)$ and $f(b)$ – i.e., with $f(a) < L < f(b)$ or $f(b) < L < f(a)$ – there exists some $c \in (a, b)$ with $f(c) = L$.

Thus a function has the intermediate value property when it does not "skip" values.

Here are two important theorems, each asserting that a broad class of functions has the intermediate value property.

THEOREM 5.29. *(Intermediate Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f has the intermediate value property.*

Example of a continuous function $f : [0, 2]_{\mathbb{Q}} \to \mathbb{Q}$ failing the intermediate value property. Let $f(x)$ be -1 for $0 \leq x < \sqrt{2}$ and $f(x) = 1$ for $\sqrt{2} < x \leq 1$.

The point of this example is to drive home the point that the Intermediate Value Theorem is the second of our three "hard theorems" in the sense that we have no chance to prove it without using special properties of the real numbers beyond the ordered field axioms. And indeed we will not prove IVT right now, but we will use it, just as we used but did not yet prove the Extreme Value Theorem. (However we are now not so far away from the point at which we will "switch back", talk about completeness of the real numbers, and prove the three hard theorems.)

The Intermediate Value Theorem (**IVT**) is ubiquitously useful. Such innocuous statements as every non-negative real number having a square root contain an implicit appeal to IVT. Further, IVT justifies the following observation.

Let $f : I \to \mathbb{R}$ be a continuous function, and suppose that there are only finitely many roots, i.e., there are $x_1, \ldots, x_n \in I$ such that $f(x_i) = 0$ for all i and $f(x) \neq 0$ for all other $x \in I$. Then $I \setminus \{x_1, \ldots, x_n\}$ is a finite union of intervals, and on each of them f has constant sign: it is either always positive or always negative.

So this is how sign analysis works for a function f when f is continuous – a very mild assumption. But as above we also want to do a sign analysis of the derivative f': how may we justify this?

Well, here is one very reasonable justification: if the derivative f' of f is itself continuous, then by IVT it too has the intermediate value property and thus, at least if f' has only finitely many roots on any bounded interval, sign analysis is justified. This brings up the following basic question.

QUESTION 5.30. *Let $f : I \to \mathbb{R}$ be a differentiable function. Must its derivative $f' : I \to \mathbb{R}$ be continuous?*

Let us first pause to appreciate the subtlety of the question: we are not asking whether f differentiable implies f continuous: we well know this is the case. Rather we are asking whether the new function f' can exist at every point of I but fail to itself be a continuous function. In fact the answer is yes.

Example: Let $f(x) = x^2 \sin(\frac{1}{x})$. I claim that f is differentiable on all of $\mathbb{R}$ but that the derivative is discontinuous at $x = 0$, and in fact that $\lim_{x\to 0} f'(x)$ does not exist. ...

THEOREM 5.31. *(Darboux) Let $f : I \to \mathbb{R}$ be a differentiable function. Suppose that we have $a, b \in I$ with $a < b$ and $f'(a) < f'(b)$. Then for every $L \in \mathbb{R}$ with $f'(a) < L < f'(b)$, there exists $c \in (a, b)$ such that $f'(c) = L$.*

PROOF. Step 1: First we handle the special case $L = 0$, which implies $f'(a) < 0$ and $f'(b) > 0$. Now f is a differentiable – hence continuous – function defined on the closed interval $[a, b]$ so assumes its minimum value at some point $c \in [a, b]$. If c is an interior point, then as we have seen, it must be a stationary point: $f'(c) = 0$. But the hypotheses guarantee this: since $f'(a) < 0$, f is decreasing at a, thus takes smaller values slightly to the right of a, so the minimum cannot occur at a. Similarly, since $f'(b) > 0$, f is increasing at b, thus takes smaller values slightly to the left of b, so the minimum cannot occur at b.
Step 2: We now reduce the general case to the special case of Step 1 by defining $g(x) = f(x) - Lx$. Then g is still differentiable, $g'(a) = f'(a) - L < 0$ and $g'(b) = f'(b) - L > 0$, so by Step 1, there exists $c \in (a, b)$ such that $0 = g'(c) = f'(c) - L$. In other words, there exists $c \in (a, b)$ such that $f'(c) = L$. □

Remark: Of course there is a corresponding version of the theorem when $f(b) < L < f(a)$. Darboux's Theorem also often called the **Intermediate Value Theorem For Derivatives**, terminology we will understand better when we discuss the Intermediate Value Theorem (for arbitrary continuous functions).

Exercise: Let a be an interior point of an interval I, and suppose $f : I \to \mathbb{R}$ is a differentiable function. Show that the function f' cannot have a **simple discontinuity** at $x = a$. (Recall that a function g has a simple discontinuity at a if $\lim_{x\to a^-} g(x)$ and $\lim_{x\to a^+} g(x)$ both exist but either they are unequal to each other or they are unequal to $g(a)$.)

5.5. A Theorem of Spivak.

The following theorem is taken directly from Spivak's book (Theorem 7 of Chapter 11): it does not seem to be nearly as well known as Darboux's Theorem (and in fact I think I encountered it for the first time in Spivak's book).

THEOREM 5.32. *Let a be an interior point of I, and let $f : I \to \mathbb{R}$. Suppose:*
(i) f is continuous on I,
(ii) f is differentiable on $I \setminus \{a\}$, i.e., at every point of I except possibly at a, and
(iii) $\lim_{x\to a} f'(x) = L$ exists.
Then f is differentiable at a and $f'(a) = L$.

PROOF. Choose $\delta > 0$ such that $(a - \delta, a + \delta) \subset I$. Let $x \in (a, a + \delta)$. Then f is differentiable at x, and we may apply the Mean Value Theorem to f on $[a, x]$: there exists $c_x \in (a, x)$ such that

$$\frac{f(x) - f(a)}{x - a} = f'(c_x).$$

Now, as $x \to a$ every point in the interval $[a, x]$ gets arbitrarily close to x, so $\lim_{x \to a} c_x = x$ and thus

$$f'_R(a) = \lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a^+} f'(c_x) = \lim_{x \to a^+} f'(x) = L.$$

By a similar argument involving $x \in (a - \delta, a)$ we get

$$f'_L(a) = \lim_{x \to} f'(x) = L,$$

so f is differentiable at a and $f'(a) = L$. □

6. Inverse Functions I: Theory

6.1. Review of inverse functions.

Let X and Y be sets, and let $f : X \to Y$ be a function between them. Recall that an **inverse function** is a function $g : Y \to X$ such that

$$g \circ f = 1_X : X \to X, \ f \circ g = 1_Y : Y \to Y.$$

Let's unpack this notation: it means the following: first, that for all $x \in X$, $(g \circ f)(x) = g(f(x)) = x$; and second, that for all $y \in Y$, $(f \circ g)(y) = f(g(y)) = y$.

PROPOSITION 5.33. *(Uniqueness of Inverse Functions) Let $f : X \to Y$ be a function. Suppose that $g_1, g_2 : Y \to X$ are both inverses of f. Then $g_1 = g_2$.*

PROOF. For all $y \in Y$, we have

$$g_1(y) = (g_2 \circ f)(g_1(y)) = g_2(f(g_1(y))) = g_1(y).$$

□

Since the inverse function to f is always unique provided it exists, we denote it by f^{-1}. (Caution: this has nothing to do with $\frac{1}{f}$. Thus $\sin^{-1}(x) \neq \csc(x) = \frac{1}{\sin x}$.)

We now turn to giving conditions for the existence of the inverse function. Recall that $f : X \to Y$ is **injective** if for all $x_1, x_2 \in X$, $x_1 \neq x_2 \implies f(x_1) \neq f(x_2)$. In other words, distinct x-values get mapped to distinct y-values. (And in yet other words, the graph of f satisfies the horizontal line test.) Also $f : X \to Y$ is **surjective** if for all $y \in Y$, there exists at least one $x \in X$ such that $y = f(x)$.

Putting these two concepts together we get the important notion of a **bijective** function $f : X \to Y$, i.e., a function which is both injective and surjective. Otherwise put, for all $y \in Y$ there exists *exactly one* $x \in X$ such that $y = f(x)$. It may well be intuitively clear that bijectivity is exactly the condition needed to guarantee existence of the inverse function: if f is bijective, we define $f^{-1}(y) = x_y$, the unique element of X such that $f(x_y) = y$. And if f is not bijective, this definition breaks down and thus we are unable to define f^{-1}. Nevertheless we ask the reader to bear with us as we give a slightly tedious formal proof of this.

THEOREM 5.34. *(Existence of Inverse Functions) For $f : X \to Y$, TFAE:*
(i) f is bijective.
(ii) f admits an inverse function.

PROOF. (i) $\implies$ (ii): If f is bijective, then as above, for each $y \in X$ there exists exactly one element of X – say x_y – such that $f(x_y) = y$. We may therefore define a function $g : Y \to X$ by $g(y) = x_y$. Let us verify that g is in fact the inverse function of f. For any $x \in X$, consider $g(f(x))$. Because f is injective, the only element $x' \in X$ such that $f(x') = f(x)$ is $x' = x$, and thus $g(f(x)) = x$. For any $y \in Y$, let x_y be the unique element of X such that $f(x_y) = y$. Then $f(g(y)) = f(x_y) = y$.
(ii) $\implies$ (i): Suppose that f^{-1} exists. To see that f is injective, let $x_1, x_2 \in X$ be such that $f(x_1) = f(x_2)$. Applying f^{-1} on the left gives $x_1 = f^{-1}(f(x_1)) = f^{-1}(f(x_2)) = x_2$. So f is injective. To see that f is surjective, let $y \in Y$. Then $f(f^{-1}(y)) = y$, so there is $x \in X$ with $f(x) = y$, namely $x = f^{-1}(y)$. □

For any function $f : X \to Y$, we define the **image** of f to be $\{y \in Y \mid \exists x \in X \mid y = f(x)\}$. The image of f is often denoted $f(X)$.[5]

We now introduce the dirty trick of **codomain restriction**. Let $f : X \to Y$ be any function. Then if we replace the codomain Y by the image $f(X)$, we still get a well-defined function $f : X \to f(X)$, and this new function is tautologically surjective. (Imagine that you manage the up-and-coming band **Yellow Pigs**. You get them a gig one night in an enormous room filled with folding chairs. After everyone sits down you remove all the empty chairs, and the next morning you write a press release saying that *Yellow Pigs* played to a "packed house". This is essentially the same dirty trick as codomain restriction.)

Example: Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. Then $f(\mathbb{R}) = [0, \infty)$, and although $x^2 : \mathbb{R} \to \mathbb{R}$ is not surjective, $x^2 : \mathbb{R} \to [0, \infty)$ certainly is.

Since a codomain-restricted function is always surjective, it has an inverse iff it is injective iff the original functionb is injective. Thus:

COROLLARY 5.35. *For a function $f : X \to Y$, the following are equivalent:*
(i) The codomain-restricted function $f : X \to f(X)$ has an inverse function.
(ii) The original function f is injective.

6.2. The Interval Image Theorem.

Next we want to return to earth by considering functions $f : I \to \mathbb{R}$ and their inverses, concentrating on the case in which f is continuous.

THEOREM 5.36. *(Interval Image Theorem) Let $I \subset \mathbb{R}$ be an interval, and let $f : I \to \mathbb{R}$ be a continuous function. Then the image $f(I)$ of f is also an interval.*

PROOF. For now we will give the proof when $I = [a, b]$, i.e., is closed and bounded. The general case will be discussed later.

Suppose $f : [a, b] \to \mathbb{R}$ is continuous. Then f has a minimum value m, say at x_m and a maximum value M, say at x_M. Thus the image $f([a, b])$ of f is a subset of $[m, M]$. Moreover, if $L \in (m, M)$, then by the Intermediate Value Theorem there exists c in between x_m and x_M such that $f(c) = L$. So $f([a, b]) = [m, M]$. □

Exercise: Let I be a nonempty interval which is *not* of the form $[a, b]$. Let J be any nonempty interval. Show: there is a continuous function $f : I \to \mathbb{R}$ with $f(I) = J$.

[5]This is sometimes called the **range** of f, but sometimes not. It is safer to call it the image!

6.3. Monotone Functions and Invertibility.

Recall $f : I \to \mathbb{R}$ is **strictly monotone** if it is either increasing or decreasing. Every strictly monotone function is injective. Therefore our dirty trick of codomain restriction works to show that if $f : I \to \mathbb{R}$ is strictly monotone, $f : I \to f(I)$ is bijective, hence invertible. Thus in this sense we may speak of the inverse of any strictly monotone function.

PROPOSITION 5.37. *Let $f : I \to f(I)$ be a strictly monotone function.*
a) If f is increasing, then $f^{-1} : f(I) \to I$ is increasing.
b) If f is decreasing, then $f^{-1} : f(I) \to I$ is decreasing.

PROOF. As usual, we will content ourselves with the increasing case, the decreasing case being so similar as to make a good exercise for the reader.

Seeking a contradiction we suppose that f^{-1} is *not* increasing: that is, there exist $y_1 < y_2 \in f(I)$ such that $f^{-1}(y_1)$ is *not* less than $f^{-1}(y_2)$. Since f^{-1} is an inverse function, it is necessarily injective (if it weren't, f itself would not be a function!), so we cannot have $f^{-1}(y_1) = f^{-1}(y_2)$, and thus the possibility we need to rule out is $f^{-1}(y_2) < f^{-1}(y_1)$. But if this holds we apply the increasing function f to get $y_2 = f(f^{-1}(y_2)) < f(f^{-1}(y_1)) = y_1$, a contradiction. □

LEMMA 5.38. *(Λ-V Lemma) Let $f : I \to \mathbb{R}$. The following are equivalent:*
(i) f is not *monotone: i.e., f is neither increasing nor decreasing.*
(ii) At least one of the following holds:
(a) f is not injective.
*(b) f admits a Λ-**configuration**: there exist $a < b < c \in I$ with $f(a) < f(b) > f(c)$.*
*(c) f admits a V-**configuration**: there exist $a < b < c \in I$ with $f(a) > f(b) < f(c)$.*

Exercise: Prove Lemma 5.38.

THEOREM 5.39. *If $f : I \to \mathbb{R}$ is continuous and injective, it is monotone.*

PROOF. We will suppose that f is injective and not monotone and show that it cannot be continuous, which suffices. We may apply Lemma 5.38 to conclude that f has either a Λ configuration or a V configuration.

Suppose first f has a Λ configuration: there exist $a < b < c \in I$ with $f(a) < f(b) > f(c)$. Then there exists $L \in \mathbb{R}$ such that $f(a) < L < f(b) > L > f(c)$. If f were continuous then by the Intermediate Value Theorem there would be $d \in (a, b)$ and $e \in (b, c)$ such that $f(d) = f(e) = L$, contradicting the injectivity of f.

Next suppose f has a V configuration: there exist $a < b < c \in I$ such that $f(a) > f(b) < f(c)$. Then there exists $L \in \mathbb{R}$ such that $f(a) > L > f(b) < L < f(c)$. If f were continuous then by the Intermediate Value Theorem there would be $d \in (a, b)$ and $e \in (b, c)$ such that $f(d) = f(e) = L$, contradicting injectivity. □

6.4. Inverses of Continuous Functions.

THEOREM 5.40. *(Continuous Inverse Function Theorem) Let $f : I \to \mathbb{R}$ be injective and continuous. Let $J = f(I)$ be the image of f.*
a) $f : I \to J$ is a bijection, and thus there is an inverse function $f^{-1} : J \to I$.
b) J is an interval in $\mathbb{R}$.
c) If $I = [a, b]$, then either f is increasing and $J = [f(a), f(b)]$ or f is decreasing and $J = [f(b), f(a)]$.
d) The function $f^{-1} : J \to I$ is also continuous.

PROOF. [**S**, Thm. 12.3] Parts a) through c) simply recap previous results. The new result is part d), that $f^{-1} : J \to I$ is continuous. By part c) and Proposition 5.37, either f and f^{-1} are both increasing, or f and f^{-1} are both decreasing. As usual, we restrict ourselves to the first case.

Let $b \in J$. We must show that $\lim_{y \to b} f^{-1}(y) = f^{-1}(b)$. We may write $b = f(a)$ for a unique $a \in I$. Fix $\epsilon > 0$. We want to find $\delta > 0$ such that if $f(a) - \delta < y < f(a) + \delta$, then $a - \epsilon < f^{-1}(y) < a + \epsilon$.

Take $\delta = \min(f(a+\epsilon) - f(a), f(a) - f(a-\epsilon))$. Then:

$$f(a-\epsilon) \leq f(a) - \delta, \ f(a) + \delta \leq f(a+\epsilon),$$

and thus if $f(a) - \delta < y < f(a) + \delta$ we have

$$f(a-\epsilon) \leq f(a) - \delta < y < f(a) + \delta \leq f(a+\epsilon).$$

Since f^{-1} is increasing, we get

$$f^{-1}(f(a-\epsilon)) < f^{-1}(y) < f^{-1}(f(a+\epsilon)),$$

or

$$f^{-1}(b) - \epsilon < f^{-1}(y) < f^{-1}(b) + \epsilon.$$

□

Remark: To be honest, I don't find the above proof very enlightening. After reflecting on my dissatisfaction with it, I came up with an alternate proof that I find conceptually simpler, but which depends on the **Monotone Jump Theorem**, a characterization of the possible discontinuities of a monotone function. The proof uses the completeness of the real numbers, so is postponed to the next chapter.

6.5. Inverses of Differentiable Functions.

In this section our goal is to determine conditions under which the inverse f^{-1} of a differentiable funtion is differentiable, and if so to find a formula for $(f^{-1})'$.

Let's first think about the problem geometrically. The graph of the inverse function $y = f^{-1}(x)$ is obtained from the graph of $y = f(x)$ by interchanging x and y, or, put more geometrically, by *reflecting* the graph of $y = f(x)$ across the line $y = x$. Geometrically speaking $y = f(x)$ is differentiable at x iff its graph has a well-defined, nonvertical tangent line at the point $(x, f(x))$, and if a curve has a well-defined tangent line, then reflecting it across a line should not change this. Thus it should be the case that if f is differentiable, so is f^{-1}. Well, almost. Notice the occurrence of "nonvertical" above: if a curve has a vertical tangent line, then since a vertical line has "infinite slope" it does not have a finite-valued derivative. So we need to worry about the possibility that reflection through $y = x$ carries a nonvertical tangent line to a vertical tangent line. When does this happen? Well, the inverse function of the straight line $y = mx + b$ is the straight line $y = \frac{1}{m}(x - b)$ – i.e., reflecting across $y = x$ takes a line of slope m to a line of slope $\frac{1}{m}$. Morever, it takes a horizontal line $y = c$ to a vertical line $x = c$, so that is our answer: at any point $(a, b) = (a, f(a))$ such that $f'(a) = 0$, then the inverse function will fail to be differentiable at the point $(b, a) = (b, f^{-1}(b))$ because it will have a vertical tangent. Otherwise, the slope of the tangent line of the inverse function at (b, a) is precisely the reciprocal of the slope of the tangent line to $y = f(x)$ at (a, b).

Well, so the geometry tells us. It turns out to be quite straightforward to adapt this geometric argument to derive the desired formula for $(f^{-1})'(b)$, *under the assumption* that f is differentiable. We will do this first. Then we need to come back and verify that indeed f^{-1} is differentiable at b if $f'(f^{-1}(b))$ exists and is nonzero: this turns out to be a bit stickier, but we are ready for it and we will do it.

PROPOSITION 5.41. *Let $f : I \to J$ be a bijective differentiable function. **Suppose** that the inverse function $f^{-1} : J \to I$ is differentiable at $b \in J$. Then $(f^{-1})'(b) = \frac{1}{f'(f^{-1}(b))}$. In particular, if f^{-1} is differentiable at b then $f'(f^{-1}(b)) \neq 0$.*

PROOF. We need only implicitly differentiate the equation

$$f^{-1}(f(x)) = x,$$

getting

$$(f^{-1})'(f(x))f'(x) = 1, \tag{20}$$

or

$$(f^{-1})'(f(x)) = \frac{1}{f'(x)}.$$

To apply this to get the derivative at $b \in J$, we just need to think a little about our variables. Let $a = f^{-1}(b)$, so $f(a) = b$. Evaluating the last equation at $x = a$ gives

$$(f^{-1})'(b) = \frac{1}{f'(a)} = \frac{1}{f'(f^{-1}(b))}.$$

Moreover, since by (20) we have $(f^{-1})'(b)f'(f^{-1}(b)) = 1$, $f'(f^{-1}(b)) \neq 0$. □

As mentioned above, unfortunately we need to work a little harder to show the differentiability of f^{-1}, and for this we cannot directly use Proposition 5.41 but end up deriving it again. Well, enough complaining: here goes.

THEOREM 5.42. *(Differentiable Inverse Function Theorem) Let $f : I \to J$ be continuous and bijective. Let b be an interior point of J and put $a = f^{-1}(b)$. Suppose that f is differentiable at a and $f'(a) \neq 0$. Then f^{-1} is differentiable at b, with the familiar formula*

$$(f^{-1})'(b) = \frac{1}{f'(a)} = \frac{1}{f'(f^{-1}(b))}.$$

PROOF. [**S**, Thm. 12.5] We have

$$(f^{-1})'(b) = \lim_{h \to 0} \frac{f^{-1}(b+h) - f^{-1}(b)}{h} = \lim_{h \to 0} \frac{f^{-1}(b+h) - a}{h}.$$

Since $J = f(I)$, every $b + h \in J$ is of the form

$$b + h = f(a + k_h)$$

for a unique $k_h \in I$.[6] Since $b + h = f(a + k_h)$, $f^{-1}(b + h) = a + k_h$; let's make this substitution as well as $h = f(a + k_h) - f(a)$ in the limit we are trying to evaluate:

$$(f^{-1})'(b) = \lim_{h \to 0} \frac{a + k_h - a}{f(a + k_h) - b} = \lim_{h \to 0} \frac{k_h}{f(a + k_h) - f(a)} = \lim_{h \to 0} \frac{1}{\frac{f(a+k_h)-f(a)}{k_h}}.$$

[6]Unlike Spivak, we *will* include the subscript k_h to remind ourselves that this k is defined in terms of h: to my taste this reminder is worth a little notational complication.

We are getting close: the limit now looks like the reciprocal of the derivative of f at a. The only issue is the pesky k_h, but if we can show that $\lim_{h\to 0} k_h = 0$, then we may simply replace the "$\lim_{h\to 0}$" with "$\lim_{k_h\to 0}$" and we'll be done.

But $k_h = f^{-1}(b+h) - a$, so – since f^{-1} is continuous by Theorem 5.40 – we have

$$\lim_{h\to 0} k_h = \lim_{h\to 0} f^{-1}(b+h) - a = f^{-1}(b+0) - a = f^{-1}(b) - a = a - a = 0.$$

So as $h \to 0$, $k_h \to 0$ and thus

$$(f^{-1})'(b) = \frac{1}{\lim_{k_h\to 0} \frac{f(a+k_h)-f(a)}{k_h}} = \frac{1}{f'(a)} = \frac{1}{f'(f^{-1}(b))}.$$

□

7. Inverse Functions II: Examples and Applications

7.1. $x^{\frac{1}{n}}$.

In this section we illustrate the preceding concepts by defining and differentiating the nth root function $x^{\frac{1}{n}}$. The reader should not now be surprised to hear that we give separate consideration to the cases of odd n and even n.

Either way, let $n > 1$ be an integer, and consider

$$f : \mathbb{R} \to \mathbb{R},\ x \mapsto x^n.$$

Case 1: $n = 2k+1$ is odd. Then $f'(x) = (2k+1)x^{2k} = (2k+1)(x^k)^2$ is non-negative for all $x \in \mathbb{R}$ and not identically zero on any subinterval $[a, b]$ with $a < b$, so by Theorem 5.25 $f : \mathbb{R} \to \mathbb{R}$ is increasing. Moreover, we have $\lim_{x\to\pm\infty} f(x) = \pm\infty$. Since f is continuous, by the Intermediate Value Theorem the image of f is all of $\mathbb{R}$. Moreover, f is everywhere differentiable and has a horizontal tangent only at $x = 0$. Therefore there is an inverse function

$$f^{-1} : \mathbb{R} \to \mathbb{R}$$

which is everywhere continuous and differentiable at every $x \in \mathbb{R}$ except $x = 0$. It is typical to call this function $x^{\frac{1}{n}}$.

Case 2: $n = 2k$ is even. Then $f'(x) = (2k)x^{2k-1}$ is positive when $x > 0$ and negative when $x < 0$. Thus f is decreasing on $(-\infty, 0]$ and increasing on $[0, \infty)$. In particular it is *not* injective on its domain. If we want to get an inverse function, we need to engage in **domain restriction**. Unlike codomain restriction, which can be done in exactly one way so as to result in a surjective function, domain restriction brings with it many choices. Luckily for us, this is a relatively simple case: if $D \subset \mathbb{R}$, then the restriction of f to D will be injective if and only if for each $x \in \mathbb{R}$, at most one of $x, -x$ lies in D. If we want the restricted domain to be as large as possible, we should choose the domain to include 0 and exactly one of $x, -x$ for all $x > 0$. There are still lots of ways to do this, so let's try to impose another desirable property of the domain of a function: namely, if possible we would like it to be an interval. A little thought shows that there are two restricted domains which meet all these requirements: we may take $D = [0, \infty)$ or $D = (-\infty, 0]$.

7.2. $L(x)$ and $E(x)$.

Consider the function $l : (0,\infty) \to \mathbb{R}$ given by $l(x) = \frac{1}{x}$. As advertised, we will soon be able to prove that every continuous function has an antiderivative, so borrowing on this result we define $L : (0,\infty) \to \mathbb{R}$ to be such that $L'(x) = l(x)$. More precisely, recall that when they exist antiderivatives are unique up to the addition of a constant, so we may uniquely specify $L(x)$ by requiring $L(1) = 0$.

PROPOSITION 5.43. *For all $x, y \in (0,\infty)$, we have*

$$L(xy) = L(x) + L(y). \tag{21}$$

PROOF. Let $y \in (0,\infty)$ be regarded as fixed, and consider the function

$$f(x) = L(xy) - L(x) - L(y).$$

We have

$$f'(x) = L'(xy)(xy)' - L'(x) = \frac{1}{xy} \cdot y - \frac{1}{x} = \frac{y}{xy} - \frac{1}{x} = 0.$$

By the zero velocity theorem, the function $f(x)$ is a constant (depending, *a priori* on y), say C_y. Thus for all $x \in (0,\infty)$,

$$L(xy) = L(x) + L(y) + C_y.$$

If we plug in $x = 1$ we get

$$L(y) = 0 + L(y) + C_y,$$

and thus $C_y = 0$, so $L(xy) = L(x) + L(y)$. □

COROLLARY 5.44. *a) For all $x \in (0,\infty)$ and $n \in \mathbb{Z}^+$, we have $L(x^n) = nL(x)$.*
b) For $x \in (0,\infty)$, we have $L(\frac{1}{x}) = -L(x)$.
c) We have $\lim_{x\to\infty} L(x) = \infty$, $\lim_{x\to 0^+} L(x) = -\infty$.
d) We have $L((0,\infty)) = \mathbb{R}$.

PROOF. a) An easy induction argument using $L(x^2) = L(x) + L(x) = 2L(x)$.
b) For any $x \in (0,\infty)$ we have $0 = L(1) = L(x \cdot \frac{1}{x}) = L(x) + L(\frac{1}{x})$.
c) Since $L'(x) = \frac{1}{x} > 0$ for all $x \in (0,\infty)$, L is increasing on $(0,\infty)$. Since $L(1) = 0$, for any $x > 0$, $L(x) > 0$. To be specific, take $C = L(2)$, so $C > 0$. Then by part a), $L(2^n) = nL(2) = nC$. By the Archimedean property of $\mathbb{R}$, this shows that L takes arbitaririly large values, and since it is increasing, this implies $\lim_{x\to\infty} L(x) = \infty$. To evaluate $\lim_{x\to 0^+} L(x)$ we may proceed similarly: by part b), $L(\frac{1}{2}) = -L(2) = -C < 0$, so $L(\frac{1}{2^n}) = -nL(2) = -Cn$, so L takes arbitrarily small values. Again, combined with the fact that L is increasing, this implies $\lim_{x\to 0^+} L(x) = -\infty$. (Alternately, we may evaluate $\lim_{x\to 0^+} L(x)$ by making the change of variable $y = \frac{1}{x}$ and noting that as $x \to 0^+$, $y \to \infty+$. This is perhaps more intuitive but is slightly tedious to make completely rigorous.)
d) Since L is differentiable, it is continuous, and the result follows immediately from part c) and the Intermediate Value Theorem. □

Definition: We define e to be the unique positive real number such that $L(e) = 1$. (Such a number exists because $L : (0,\infty) \to \mathbb{R}$ is increasing – hence injective and has image $(-\infty,\infty)$. Thus in fact for *any* real number α there is a unique positive real number β such that $L(\beta) = \alpha$.)

Since $L(x)$ is everywhere differentiable with nonzero derivative $\frac{1}{x}$, the differentiable inverse function theorem applies: L has a differentiable inverse function

$$E : \mathbb{R} \to (0, \infty),\ E(0) = 1.$$

Let's compute E': differentiating $L(E(x)) = x$ gives

$$1 = L'(E(x))E'(x) = \frac{E'(x)}{E(x)}.$$

In other words, we get

$$E'(x) = E(x).$$

COROLLARY 5.45. *For all $x, y \in \mathbb{R}$ we have $E(x+y) = E(x)E(y)$.*

PROOF. To showcase the range of techniques available, we give three different proofs.

First proof: For $y \in \mathbb{R}$, let $E_y(x) = E(x+y)$. Put $f(x) = \frac{E_y(x)}{E(x)}$. Then

$$f'(x) = \frac{E_y(x)E'(x) - E_y'(x)E(x)}{E(x)^2} = \frac{E(x+y)E(x) - E'(x+y)(x+y)'E(x)}{E(x)^2}$$
$$= \frac{E(x+y)E(x) - E(x+y)\cdot 1 \cdot E(x)}{E(x)^2} = 0.$$

By the Zero Velocity Theorem, there is $C_y \in \mathbb{R}$ such that for all $x \in \mathbb{R}$, $f(x) = E(x+y)/E(x) = C_y$, or $E(x+y) = E(x)C_y$. Plugging in $x = 0$ gives

$$E(y) = E(0)C(y) = 1 \cdot C(y) = C(y),$$

so

$$E(x+y) = E(x)E(y).$$

Second proof: We have

$$L\left(\frac{E(x+y)}{E(x)E(y)}\right) = L(E(x+y)) - L(E(x)) - L(E(y)) = x + y - x - y = 0.$$

The unique $x \in (0, \infty)$ such that $L(x) = 0$ is $x = 1$, so we must have

$$\frac{E(x+y)}{E(x)E(y)} = 1,$$

or

$$E(x+y) = E(x)E(y).$$

Third proof: For any $y_1, y_2 > 0$, we have

$$L(y_1y_2) = L(y_1) + L(y_2).$$

Put $y_1 = E(x_1)$ and $y_2 = E(x_2)$, so that $x_1 = L(y_1)$, $x_2 = L(y_2)$ and thus

$$E(x_1)E(x_2) = y_1y_2 = E(L(y_1y_2)) = E(L(y_1) + L(y_2)) = E(x_1 + x_2).$$

□

Note also that since E and L are inverse functions and $L(e) = 1$, we have $E(1) = e$. Now the previous disucssion must suggest to any graduate of freshman calculus that $E(x) = e^x$: both functions defined and positive for all real numbers, are equal to their own derivatives, convert multiplication into addition, and take the value 1 at $x = 0$. How many such functions could there be?

PROPOSITION 5.46. *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f'(x) = f(x)$ for all $x \in \mathbb{R}$. There is a constant C such that $f(x) = CE(x)$ for all $x \in \mathbb{R}$.*

PROOF. Define a function $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = \frac{f(x)}{E(x)}$. Then for all $x \in \mathbb{R}$,

$$g'(x) = \frac{E(x)f'(x) - E'(x)f(x)}{E(x)^2} = \frac{E(x)f(x) - E(x)f(x)}{E(x)^2} = 0.$$

By the Zero Velocity Theorem $g = \frac{f}{E}$ is constant: $f(x) = CE(x)$ for all x. □

In other words, if there really is a function $f(x) = e^x$ out there with $f'(x) = e^x$ and $f(0) = 1$, then we must have $e^x = E(x)$ for all x. The point of this logical maneuver is that although in precalculus mathematics one learns to manipulate and graph exponential functions, the actual *definition* of a^x for irrational x is not given, and indeed I don't see how it can be given without using key concepts and theorems of calculus. But, with the functions $E(x)$ and $L(x)$ in hand, let us develop the theory of exponentials and logarithms to arbitrary bases.

Let $a > 0$ be a real number. How should we define a^x? In the following slightly strange way: for any $x \in \mathbb{R}$,

$$a^x := E(L(a)x).$$

Let us make two comments: first, if $a = e$ this agrees with our previous definition: $e^x = E(xL(e))) = E(x)$. Second, the definition is *motivated* by the following desirable law of exponents: $(a^b)^c = a^{bc}$. Indeed, *assuming* this holds unrestrictedly for $b, c \in \mathbb{R}$ and $a > 1$, we would have

$$a^x = E(x \log a) = e^{x \log a} = (e^{\log a})^x = a^x.$$

But here is the point: we do not wish to *assume* that the laws of exponents work for all real numbers as they do for positive integers...we want to *prove* them!

PROPOSITION 5.47. *Fix $a \in (0, \infty)$. For $x \in \mathbb{R}$, we define*

$$a^x := E(L(a)x).$$

If $a \neq 1$, we define

$$\log_a(x) = \frac{L(x)}{L(a)}.$$

a) The function a^x is differentiable and $(a^x)' = L(a)a^x$.
b) The function $\log_a x$ is differentiable and $(\log_a x)' = \frac{1}{L(a)x}$.
c) Suppose $a > 1$. Then a^x is increasing with image $(0, \infty)$, $\log_a x$ is increasing with image $(-\infty, \infty)$, and a^x and $\log_a x$ are inverse functions.
d) For all $x, y \in \mathbb{R}$, $a^{x+y} = a^x a^y$.
e) For all $x > 0$ and $y \in \mathbb{R}$, $(a^x)^y = a^{xy}$.
f) For all $x, y > 0$, $\log_a(xy) = \log_a x + \log_a y$.
g) For all $x > 0$ and $y \in \mathbb{R}$, $\log_a(x^y) = y \log_a x$.

PROOF. a) We have

$$(a^x)' = E(L(a)x)' = E'(L(a)x)(L(a)x)' = E(L(a)x) \cdot L(a) = L(a)a^x.$$

b) We have

$$(\log_a(x))' = \left(\frac{L(x)}{L(a)}\right)' = \frac{1}{L(a)x}.$$

c) Since their derivatives are always positive, a^x and $\log_a x$ are both increasing functions. Moreover, since $a > 1$, $L(a) > 0$ and thus

$$\lim_{x\to\infty} a^x = \lim_{x\to\infty} E(L(a)x) = E(\infty) = \infty,$$

$$\lim_{x\to\infty} \log_a(x) = \lim_{x\to\infty} \frac{L(x)}{L(a)} = \frac{\infty}{L(a)} = \infty.$$

Thus $a^x : (-\infty, \infty) \to (0, \infty)$ and $\log_a x : (0, \infty) \to (-\infty, \infty)$ are bijective and thus have inverse functions. Thus check that they are inverses of each other, it suffices to show that *either* one of the two compositions is the identity function. Now

$$\log_a(a^x) = \frac{L(a^x)}{L(a)} = \frac{L(E(L(a)x))}{L(a)} = \frac{L(a)x}{L(a)} = x.$$

d) We have

$$a^{x+y} = E(L(a)(x+y)) = E(L(a)x + L(a)y) = E(L(a)x)E(L(a)y) = a^x a^y.$$

e) We have

$$(a^x)^y = E(L(a^x)y) = E(L(E(L(a)x))y) = E(L(a)xy) = a^{xy}.$$

f) We have

$$\log_a(xy) = \frac{L(xy)}{L(a)} = \frac{L(x) + L(y)}{L(a)} = \frac{L(x)}{L(a)} + \frac{L(y)}{L(a)} = \log_a x + \log_a y.$$

g) We have

$$\log_a x^y = \frac{L(x^y)}{L(a)} = \frac{L(E(L(x)y))}{L(a)} = \frac{L(x)y}{L(a)} = y \log_a x.$$

□

Having established all this, we now feel free to write e^x for $E(x)$ and $\log x$ for $L(x)$.

Exercise: Suppose $0 < a < 1$. Show that a^x is decreasing with image $(0, \infty)$, $\log_a x$ is decreasing with image $(0, \infty)$, and a^x and $\log_a x$ are inverse functions.

Exercise: Prove the **change of base formula**: for all $a, b, c > 0$ with $a, c \neq 1$,

$$\log_a b = \frac{\log_c b}{\log_c a}.$$

PROPOSITION 5.48. *Let $f(x) = e^{x^2}$. Then for all $n \in \mathbb{Z}^+$ there exists a polynomial $P_n(x)$, of degree n, such that*

$$\frac{d^n}{dx^n} f(x) = P_n(x) e^{x^2}.$$

PROOF. By induction on n.
Base case ($n = 1$):
$\frac{d}{dx} e^{x^2} = 2x e^{x^2} = P_1(x) e^{x^2}$, where $P_1(x) = 2x$, a degree one polynomial.

Inductive step: Assume that for some positive integer n there exists $P_n(x)$ of degree n such that $\frac{d^n}{dx^n} e^{x^2} = P_n(x) e^{x^2}$. So $\frac{d^{n+1}}{dx^{n+1}} e^{x^2} =$

$$\frac{d}{dx}\frac{d^n}{dx^n} e^{x^2} \overset{\text{IH}}{=} \frac{d}{dx} P_n(x) e^{x^2} = P_n'(x) e^{x^2} + 2x P_n(x) e^{x^2} = \left(P_n'(x) + 2x P_n(x)\right) e^{x^2}.$$

Now, since $P_n(x)$ has degree n, $P_n'(x)$ has degree $n-1$ and $2xP_n(x)$ has degree $n+1$. If f and g are two polynomials such that the degree of f is different from the degree of g, then $\deg(f+g) = \max(\deg(f), \deg(g))$. In particular, $P_{n+1}(x) := P_n'(x) + 2xP_n(x)$ has degree $n+1$, completing the proof of the induction step. □

7.3. Some inverse trigonometric functions.

We now wish to consider inverses of the trigonometric functions: sine, cosine, tangent, and so forth. Right away we encounter a problem similar to the case of x^n for even n: the trigonometric functions are periodic, hence certainly not injective on their entire domain. Once again we are forced into the *art* of **domain restriction** (as opposed to the *science* of **codomain restriction**).

Consider first $f(x) = \sin x$. To get an inverse function, we need to restrict the domain to some subset S on which f is injective. As usual we like intervals, and a little thought shows that the maximal possible length of an interval on which the sine function is injective is π, attained by any interval at which the function either increases from -1 to 1 or decreases from 1 to -1. This still gives us choices to make. The most standard choice – but to be sure, one that is not the only possible one nor is mathematically consecrated in any particular way – is to take $I = [\frac{-\pi}{2}, \frac{\pi}{2}]$. We claim that f is increasing on I. To check this, note that $f'(x) = \cos x$ is indeed positive on $(\frac{-\pi}{2}, \frac{\pi}{2})$. We have $f([\frac{-\pi}{2}, \frac{\pi}{2}]) = [-1, 1]$. The inverse function here is often called $\arcsin x$ ("arcsine of x") in an attempt to distinguish it from $\frac{1}{\sin x} = \csc x$. This is as good a name as any: let's go with it. We have

$$\arcsin : [-1, 1] \to [\frac{-\pi}{2}, \frac{\pi}{2}].$$

As the inverse of an increasing function, $\arcsin x$ is increasing. Moreover since $\sin x$ has a nonzero derivative on $(\frac{-\pi}{2}, \frac{\pi}{2})$, $\arcsin x$ is differentiable there. Differentiating

$$\sin(\arcsin x)) = x,$$

we get

$$\cos(\arcsin x) \arcsin'(x) = 1,$$

or

$$\frac{d}{dx} \arcsin x = \frac{1}{\cos(\arcsin x)}.$$

This looks like a mess, but a little trigonometry will clean it up. The key is to realize that $\cos \arcsin x$ means "the cosine of the angle whose sine is x" and that there must be a simpler description of this. If we draw a right triangle with angle $\theta = \arcsin x$, then to get the ratio of the opposite side to the hypotenuse to be x we may take the length of the opposite side to be x and the length of the hypotenuse to be 1, in which case the length of the adjacent side is, by the Pythagorean Theorem, $\sqrt{1-x^2}$. Thus $\cos\theta = \sqrt{1-x^2}$, so finally

$$\frac{d}{dx} \arcsin x = \frac{1}{\sqrt{1-x^2}}.$$

Now consider $f(x) = \cos x$. Since f is even, it is not injective on any interval containing 0 in its interior. Reflecting a bit on the graph of $f(x) = \cos x$ one sees that a reasonable choice for the restricted domain is $[0, \pi]$: since $f'(x) = -\sin x$ is

negative on $(0,\pi)$ and 0 and 0 and π, $f(x)$ is decreasing on $[0,\pi]$ and hence injective there. Its image is $f([0,\pi])) = [-1,1]$. Therefore we have an inverse function

$$\arccos : [-1,1] \to [0,\pi].$$

Since $\cos x$ is continuous, so is $\arccos x$. Since $\cos x$ is differentiable and has zero derivative only at 0 and π, $\arccos x$ is differentiable on $(-1,1)$ and has vertical tangent lines at $x = -1$ and $x = 1$. Morever, since $\cos x$ is decreasing, so is $\arccos x$.

We find a formula for the derivative of arccos just as for arcsin: differentiating

$$\cos \arccos x = x$$

gives

$$-\sin(\arccos x)\arccos' x = 1,$$

or

$$\arccos' x = \frac{-1}{\sin \arccos x}.$$

Again, this may be simplified. If $\varphi = \arccos x$, then $x = \cos\varphi$, so if we are on the unit circle then the y-coordinate is $\sin\varphi = \sqrt{1-x^2}$, and thus

$$\arccos' x = \frac{-1}{\sqrt{1-x^2}}.$$

Remark: It is hard not to notice that the derivatives of the arcsine and the arccosine are simply negatives of each other, so for all $x \in [0, \frac{\pi}{2}]$,

$$\arccos' x + \arcsin' x = 0.$$

By the Zero Velocity Theorem, we conclude

$$\arccos x + \arcsin x = C$$

for some constant C. To determine C, simply evaluate at $x = 0$:

$$C = \arccos 0 + \arcsin 0 = \frac{\pi}{2} + 0 = \frac{\pi}{2},$$

and thus for all $x \in [0, \frac{\pi}{2}]$ we have

$$\arccos x + \arcsin x = \frac{\pi}{2}.$$

So the angle θ whose sine is x is complementary to the angle φ whose cosine is x.

Finally, consider $f(x) = \tan x = \frac{\sin x}{\cos x}$. The domain is all real numbers for which $\cos x \neq 0$, so all real numbers except $\pm\frac{\pi}{2}, \pm\frac{3\pi}{2}, \ldots$. The tangent function is periodic with period π and also odd, which suggests that, as with the sine function, we should restrict this domain to the largest interval about 0 on which f is defined and injective. Since $f'(x) = \sec^2 x > 0$, f is increasing on $(\frac{-\pi}{2}, \frac{\pi}{2})$ and thus is injective there. Moreover, $\lim_{x\to\pm\frac{\pi}{2}} \tan x = \pm\infty$, so by the Intermediate Value Theorem $f((\frac{-\pi}{2}, \frac{\pi}{2})) = \mathbb{R}$. Therefore we have an inverse function

$$\arctan : \mathbb{R} \to \left(\frac{-\pi}{2}, \frac{\pi}{2}\right).$$

Since the tangent function is differentiable with everywhere positive derivative, the same is true for $\arctan x$. In particular it is increasing but bounded: $\lim_{x\to\pm\infty} \arctan x = \pm\frac{\pi}{2}$. In other words the arctangent has horizontal asymptotes at $y = \pm\frac{\pi}{2}$.

8. Some Complements

The Mean Value Theorem and its role in "freshman calculus" has been a popular topic of research and debate over the years.

A short paper of W.J. Knight improves upon the Zero Velocity Theorem.

THEOREM 5.49. *(Right-Handed Zero Velocity Theorem* [**Kn80**]*) Let* $f : [a, b] \to \mathbb{R}$ *be continuous. If the right-hand derivative* $f'_+(x)$ *exists and is* 0 *for all* $x \in (a, b)$*, then* f *is constant.*

The proof is modelled upon the usual one: one starts with a right-handed Rolle's Theorem, deduces a right-handed Mean Value Inequality, and then Theorem 5.49.

CHAPTER 6

Completeness

1. Dedekind Completeness

1.1. Introducing (LUB) and (GLB).

Gather round, my friends: the time has come to tell what makes calculus work.

Recall that we began the course by considering the real numbers as a set endowed with two binary operations $+$ and $\cdot$ together with a relation $<$, and satisfying a longish list of familiar axioms (P0) through (P12), the **ordered field** axioms. We then showed that using these axioms we could deduce many other familiar properties of numbers and prove many other identities and inequalities.

However we did not claim that (P0) through (P12) was a *complete* list of axioms for $\mathbb{R}$. On the contrary, we saw that this could not be the case: for instance the rational numbers $\mathbb{Q}$ also satisfy the ordered field axioms but – as we have taken great pains to point out – most of the "big theorems" of calculus are meaningful but false when regarded as results applied to the system of rational numbers. So there must be some further axiom, or property, of $\mathbb{R}$ which is needed to prove the three Interval Theorems, among others.

Here it is. Among structures F satisfying the ordered field axioms, consider the following further property:

(P14): **Least Upper Bound Axiom** (LUB): Let S be a nonempty subset of F which is bounded above. Then S admits a **least upper bound**.

This means exactly what it sounds like, but it is so important that we had better make sure. Recall a subset S of F is **bounded above** if there exists $M \in \mathbb{R}$ such that for all $x \in S$, $x \leq M$. (For future reference, a subset S of $\mathbb{R}$ is **bounded below** if there exists $m \in F$ such that for all $x \in S$, $m \leq x$.) By a **least upper bound** for a subset S of F, we mean an upper bound M which is less than any other upper bound: thus, M is a least upper bound for S if M is an upper bound for S and for any upper bound M' for S, $M \leq M'$.

There is a widely used synonym for "the least upper bound of S", namely the **supremum** of S. We also introduce the notation $\operatorname{lub} S = \sup S$ for the supremum of a subset S of an ordered field (when it exists).

The following is a useful alternate characterization of $\sup S$: the supremum of

S is an upper bound M for S with the property that for any $M' < M$, M' is *not* an upper bound for S: explicitly, for all $M' < M$, there exists $x \in S$ with $M' < x$.

The definition of the least upper bound of a subset S makes sense for any set X endowed with an order relation $<$. Notice that the *uniqueness* of the supremum $\sup S$ is clear: we cannot have two different least upper bounds for a subset, because one of them will be larger than the other! Rather what is in question is the *existence* of least upper bounds, and (LUB) is an assertion about this.

Taking the risk of introducing even more terminology, we say that an ordered field $(F, +, \cdot, <)$ is **Dedekind complete**[1] if it satisfies the least upper bound axiom. Now here is the key fact lying at the foundations of calculus and real analysis.

THEOREM 6.1. *a) The ordered field $\mathbb{R}$ is Dedekind complete.*
b) Conversely, any Dedekind complete ordered field is isomorphic to $\mathbb{R}$.

Part b) of Theorem 6.1 really means the following: if F is any Dedekind complete ordered field then there is a bijection $f : F \to \mathbb{R}$ which preserves the addition, multiplication and order structures in the following sense: for all $x, y \in F$,

- $f(x + y) = f(x) + f(y)$,
- $f(xy) = f(x)f(y)$, and
- If $x < y$, then $f(x) < f(y)$.

This concept of "isomorphism of structures" comes from a more advanced course – **abstract algebra** – so it is probably best to let it go for now. One may take part b) to mean that there is *essentially* only one Dedekind complete ordered field: $\mathbb{R}$.

The proof of Theorem 6.1 involves *constructing* the real numbers in a mathematically rigorous way. This is something of a production, and although in some sense every serious student of mathematics should see a construction of $\mathbb{R}$ at some point of her career, this sense is similar to the one in which every serious student of computer science should build at least one working computer from scratch: in practice, one can probably get away with relying on the fact that many other people have performed this task in the past. Spivak does give a construction of $\mathbb{R}$ and a proof of Theorem 6.1 in the "Epilogue" of his text. And indeed, if we treat this material at all it will be at the very end of the course.

After discussing least upper bounds, it is only natural to consider the "dual" concept of greatest lower bounds. Again, this means exactly what it sounds like but it is so important that we spell it out explicitly: if S is a subset of an ordered field F, then a **greatest lower bound** for S, or an **infimum** of S, is an element $m \in F$ which is a lower bound for S – i.e., $m \leq x$ for all $x \in S$ – and is such that if m' is any lower bound for S then $m' \leq m$. Equivalently, $m = \inf S$ iff m is a lower bound for S and for any $m' > m$ there exists $x \in S$ with $x < m'$. Now consider:

(P14′): **Greatest Lower Bound Axiom** (GLB): Let S be a nonempty subset of F which is bounded below. Then S admits a greatest lower bound, or infimum.

[1]It is perhaps more common to say "complete" instead of "Dedekind complete". I have my reasons for preferring the lengthier terminology, but I won't trouble you with them.

Example 1.1: In any ordered field F, we may consider the subset

$$S_F = \{x \in F \mid x^2 < 2\}.$$

Then S_F is nonempty and bounded: indeed $0 \in S_F$ and if $x \in S_F$, then $|x| \leq 2$. Of course in the previous inequality we could do better: for instance, if $|x| > \frac{3}{2}$, then $x^2 > \frac{9}{4} > 2$, so also $\frac{-3}{2}$ is a lower bound for S_F and $\frac{3}{2}$ is an upper bound for S_F. Of course we could do better still...

Indeed the bounded set S_F will have an infimum and a supremum if and only if there are *best possible* inequalities $x \in S \implies m \leq x \leq M$, i.e., for which no improvement on either m or M is possible. Whether such best possible inequalities exist depends on the ordered field F. Indeed, it is clear that *if* $M = \sup S_F$ exists, then it must be a positive element of F with $M^2 = 2$: or in other words, what in precalculus mathematics one cavalierly writes as $M = \sqrt{2}$. similarly, *if* $m = \inf S_F$ exists, then it must be a negative element of F with $m^2 = 2$, or what we usually write as $-\sqrt{2}$. But here's the point: how do we know that our ordered field F contains such an element $\sqrt{2}$?

The answer of course is that depending on F such an element may or may not exist. As we saw at the beginning of the course, there is no *rational number* x with $x^2 = 2$, so if $F = \mathbb{Q}$ then our set $S_{\mathbb{Q}}$ has neither an infimum nor a supremum. Thus $\mathbb{Q}$ does not satisfy (LUB) or (GLB). On the other hand, we certainly believe that there is a real number whose square is 2. But...why do we believe this? As we have seen, the existence of a real square root of every non-negative real number is a consequence of the Intermediate Value Theorem...which is of course a theorem that we have exalted but not yet proved. A more fundamental answer is that we believe that $\sqrt{2}$ exists in $\mathbb{R}$ *because* of the Dedekind completeness of $\mathbb{R}$, i.e., according to Theorem 6.1 *every* nonempty bounded above subset of $\mathbb{R}$ has a supremum, so in particular $S_{\mathbb{R}}$ has a supremum, which must be $\sqrt{2}$.

An interesting feature of this example is that we can see that $\inf S_{\mathbb{R}}$ exists, even though we have not as yet addressed the issue of whether $\mathbb{R}$ satisfies (GLB). In general, $\inf S_F$ exists iff there is an element $y < 0$ in F with $y^2 = 2$. But okay: if in F we have a positive element x with $x^2 = 2$, we necessarily must also have a negative element y with $y^2 = 2$: namely, $y = -x$.

This turns out to be a very general phenomenon.

THEOREM 6.2. *Let F be an ordered field.*
a) Then F satisfies (LUB) iff it satisfies (GLB).
b) In particular $\mathbb{R}$ satisfies both (LUB) and (GLB) and is (up to isomorphism) the only ordered field with this property.

PROOF. a) I know two ways of showing that (LUB) $\iff$ (GLB). Both of these arguments is very nice in its own way, and I don't want to have to choose between them. So I will show you both, in the following way: I will use the first argument to show that (LUB) $\implies$ (GLB) and the second argument to show that (GLB) $\implies$ (LUB). (In Exercise 1.2 below, you are asked to do things the other way around.)
(LUB) $\implies$ (GLB): Let $S \subset F$ be nonempty and bounded below by m. Consider

$$-S = \{-x \mid x \in S\}.$$

Then $-S$ is nonempty and bounded above by $-m$. By (LUB), it has a least upper bound $\sup(-S)$. We claim that in fact $-\sup(-S)$ is a greatest lower bound for S, or more symbolically:

$$\inf S = -\sup -S.$$

You are asked to check this in Exercise 1.2 below.
(GLB) $\implies$ (LUB): Let S be nonempty and bounded above by M. Consider

$$\mathcal{U}(S) = \{x \in F \mid x \text{ is an upper bound for } S.\}.$$

Then $\mathcal{U}(S)$ is nonempty: indeed $M \in \mathcal{U}(S)$. Also $\mathcal{U}(S)$ is bounded below: indeed any $s \in S$ (there is at least one such s, since $S \neq \varnothing$!) is a lower bound for $\mathcal{U}(S)$. By (GLB) $\mathcal{U}(S)$ has a greatest lower bound $\inf \mathcal{U}(S)$. We claim that in fact $\inf \mathcal{U}(S)$ is a least upper bound for S, or more succinctly,

$$\sup S = \inf \mathcal{U}(S).$$

Once again, Exercise 1.2 asks you to check this.
b) By Theorem 6.1a), $\mathbb{R}$ satisfies (LUB), and thus by part a) it satisfies (GLB). By Theorem 6.1b) $\mathbb{R}$ is the only ordered field satisfying (LUB), so certainly it is the only ordered field satifying (LUB) *and* (GLB). □

Exercise 1.2: a) Fill in the details of the proof of Theorem 6.2a).
b) Let F be an ordered field, and let S be a subset of F. Suppose that $\inf S$ exists. Show that $\sup -S$ exists and

$$\sup -S = -\inf S.$$

c) Use part b) to give a second proof that (GLB) $\implies$ (LUB).
d) Let F be an ordered field, and let S be a subset of F. Define

$$\mathcal{L}(S) = \{x \in F \mid x \text{ is a lower bound for } S.\}.$$

Suppose that $\sup \mathcal{L}(S)$ exists. Show that $\inf S$ exists and

$$\inf S = \sup \mathcal{L}(S).$$

e) Use part d) to give a second proof that (LUB) $\implies$ (GLB).

The technique which was used to prove (LUB) $\implies$ (GLB) is very familiar: we multiply everything in sight by -1. It seems likely that by now we have used this type of argument more than any other single trick or technique. When this has come up we have usually used the phrase "and similarly one can show..." Perhaps this stalwart ally deserves better. Henceforth, when we wish to multiply by -1 to convert $\leq$ to $\geq$, max to min, sup to inf and so forth, we will say **by reflection**. This seems more appealing and also more specific than "similarly..."!

In view of Theorem 6.2 it is reasonable to use the term **Dedekind completeness** to refer to either or both of (LUB), (GLB), and we shall do so.

THEOREM 6.3. *A Dedekind complete ordered field is Archimedean.*

PROOF. We will prove the contrapositive: let F be a non-Archimedean ordered field: thus there exists $x \in F$ such that $n \leq x$ for all $n \in \mathbb{Z}^+$. Then the subset $\mathbb{Z}^+$ of F is bounded above by x, so in particular it is nonempty and bounded above. So, if F were Dedekind complete then $\sup \mathbb{Z}^+$ would exist.

But we claim that in no ordered field F does $\mathbb{Z}^+$ have a supremum. Indeed,

suppose that $M = \sup \mathbb{Z}^+$. It follows that for all $n \in \mathbb{Z}^+$, $n \leq M$. But then it is equally true that for all $n \in \mathbb{Z}^+$, $n+1 \leq M$, or equivalently, for all $n \in \mathbb{Z}^+$, $n \leq M-1$, so $M-1$ is a smaller upper bound for $\mathbb{Z}^+$ than $\sup \mathbb{Z}^+$: contradiction! □

1.2. Calisthenics With Sup and Inf.

The material and presentation of this section is partly based on [**A**, §1.3.13].

CONVENTION: Whenever $\sup S$ appears in the conclusion of a result, the statement should be understood as including the assertion that $\sup S$ exists, i.e., that S is nonempty and bounded above. Similarly for $\inf S$: when it appears in the conclusion of a result then an implicit part of the conclusion is the assertion that $\inf S$ exists, i.e., that S is nonempty and bounded below.[2]

PROPOSITION 6.4. *Let S be a nonempty subset of $\mathbb{R}$.*
a) Suppose S is bounded above. Then for every $\epsilon > 0$, there exists $x \in S$ such that $\sup S - \epsilon < x \leq \sup S$.
b) Conversely, suppose $M \in \mathbb{R}$ is an upper bound for S such that for all $\epsilon > 0$, there exists $x \in S$ with $M - \epsilon < x \leq M$. Then $M = \sup S$.
c) Suppose S is bounded below. Then for every $\epsilon > 0$, there exists $x \in S$ such that $\inf S \leq x < \inf S + \epsilon$.
d) Conversely, suppose $m \in \mathbb{R}$ is a lower bound for S such that for all $\epsilon > 0$, there exists $x \in S$ with $m \leq x \leq m + \epsilon$. Then $m = \inf S$.

PROOF. a) Fix $\epsilon > 0$. Since $\sup S$ is the *least* upper bound of S and $\sup S - \epsilon < \sup S$, there exists $y \in S$ with $\sup S - \epsilon < y$. It follows that

$$\sup S - \epsilon < \min(y, \sup S) \leq \sup S,$$

so we may take $x = \min(y, \sup S)$.
b) By hypothesis, M is an upper bound for S and nothing smaller than M is an upper bound for S, so indeed $M = \sup S$.
c),d) These follow from parts a) and b) by reflection. □

Exercise 1.3: Let $a, b \in \mathbb{R}$. Suppose that for all $\epsilon > 0$, $a \leq b + \epsilon$. Show that $a \leq b$.

PROPOSITION 6.5. *Let X, Y be nonempty subsets of $\mathbb{R}$, and define*

$$X + Y = \{x + y \mid x \in X,\ y \in Y\}.$$

a) Suppose X and Y are bounded above. Then

$$\sup(X + Y) = \sup X + \sup Y.$$

b) Suppose X and Y are bounded below. Then

$$\inf(X + Y) = \inf X + \inf Y.$$

PROOF. a) Let $x \in X$, $y \in Y$. Then $x \leq \sup X$ and $x \leq \sup Y$, so $x + y \leq \sup X + \sup Y$, and thus $\sup(X+Y) \leq \sup X + \sup Y$. Now fix $\epsilon > 0$. By Proposition 6.4 there are $x \in X$ and $y \in Y$ with $\sup X - \frac{\epsilon}{2} < x,\ \sup Y - \frac{\epsilon}{2} < y$, so

$$\sup X + \sup Y \leq x + y + \epsilon.$$

Since this folds for all $\epsilon > 0$, by Exercise 1.3 $\sup X + \sup Y \leq \sup(X + Y)$.
b) This follows from part a) by reflection. □

[2]Notice that a similar convention governs the use of $\lim_{x \to c} f(x)$, so this is nothing new.

Let X, Y be subsets of $\mathbb{R}$. We write $X \leq Y$ if for all $x \in X$ and all $y \in Y$, $x \leq y$. (In a similar way we define $X < Y, X \geq Y$, $X > Y$.)

Exercise 1.4: Let X, Y be subsets of $\mathbb{R}$. Give necessary and sufficient conditions for $X \leq Y$ and $Y \leq X$ both to hold. (Hint: in the case in which X and Y are both nonempty, $X = Y$ is necessary but not sufficient!)

PROPOSITION 6.6. *Let X, Y be nonempty subsets of $\mathbb{R}$ with $X \leq Y$. Then*

$$\sup X \leq \inf Y.$$

PROOF. Seeking a contradiction, we suppose that $\inf Y < \sup X$. Put

$$\epsilon = \frac{\sup X - \inf Y}{2}.$$

By Proposition 6.4 there are $x \in X$, $y \in Y$ with $\sup X - \epsilon < x$ and $y < \inf Y + \epsilon$. Since $X \leq Y$ this gives

$$\sup X - \epsilon < x \leq y < \inf Y + \epsilon$$

and thus

$$\sup X - \inf Y < 2\epsilon = \sup X - \inf Y,$$

a contradiction. □

PROPOSITION 6.7. *Let X, Y be nonempty subsets of $\mathbb{R}$ with $X \subseteq Y$. Then:*
a) If Y is bounded above, then $\sup X \leq \sup Y$.
b) If Y is bounded below, then $\inf Y \leq \inf X$.

Exercise 1.5: Prove Proposition 6.7.

1.3. The Extended Real Numbers.

As exciting and useful as this whole business with sup and inf is, there is one slightly annoying point: $\sup S$ and $\inf S$ are not defined for *every* subset of $\mathbb{R}$. Rather, for $\sup S$ to be defined, S must be nonempty and bounded above, and for $\inf S$ to be defined, S must be nonempty and bounded below.

Is there some way around this? There is. It involves bending the rules a bit, but in a very natural and useful way. Consider the subset $\mathbb{N}$ of $\mathbb{R}$. It is not bounded above, so it does not have a least upper bound in $\mathbb{R}$. Because $\mathbb{N}$ contains arbitrarily large elements of $\mathbb{R}$, it is not completely unreasonable to say that its elements approach *infinity* and thus to set $\sup \mathbb{N} = +\infty$. In other words, we are suggesting the following definition:

• If $S \subset \mathbb{R}$ is unbounded above, then we will say $\sup S = +\infty$.

Surely we also want to make the following definition ("by reflection"!):

• If $S \subset \mathbb{R}$ is unbounded below, then we will say $\inf S = -\infty$.

These definitions come with a **warning**: $\pm\infty$ **are not real numbers**! They are just symbols suggestively standing for a certain type of behavior of a subset of $\mathbb{R}$, in a similar (but, in fact, simpler) way as when we write $\lim_{x \to c} f(x) = \pm\infty$ and mean that the function has a certain type of behavior near the point c.

To give a name to what we have done, we define the **extended real numbers** $[-\infty, \infty] = \mathbb{R} \cup \{\pm\infty\}$ to be the real numbers together with these two formal symbols $-\infty$ and ∞. This extension is primarily *order-theoretic*: that is, we may extend the $\leq$ relation to the extended real numbers in the obvious way:

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

Conversely much of the point of the extended real numbers is to give the real numbers, as an ordered set, the pleasant properties of a closed, bounded interval $[a, b]$: namely we have a largest and smallest element.

The extended real numbers $[-\infty, \infty]$ *are not* a field. In fact, we cannot even define the operations of $+$ and $\cdot$ unrestrictedly on them. However, it is useful to define some of these operations:

$$\forall x \in \mathbb{R},\ -\infty + x = -\infty,\ x + \infty = \infty.$$
$$\forall x \in (0, \infty),\ x \cdot \infty = \infty,\ x \cdot (-\infty) = -\infty.$$
$$\forall x \in (-\infty, 0),\ x \cdot \infty = -\infty,\ x \cdot (-\infty) = \infty.$$
$$\infty \cdot \infty = \infty,\ \infty \cdot (-\infty) = -\infty,\ (-\infty) \cdot (-\infty) = \infty.$$
$$\frac{1}{\infty} = \frac{1}{-\infty} = 0.$$

None of these definitions are really surprising, are they? If you think about it, they correspond to facts you have learned about manipulating infinite limits, e.g. if $\lim_{x\to c} f(x) = \infty$ and $\lim_{x\to c} g(x) = 17$, then $\lim_{x\to c} f(x) + g(x) = \infty$. However, certain other operations with the extended real numbers *are not defined*, for similar reasons. In particular we **do not define**

$$\infty - \infty,$$
$$0 \cdot \infty,$$
$$\frac{\pm\infty}{\pm\infty}.$$

Why not? Well, again we might think in terms of associated limits. The above are **indeterminate forms**: if I tell you that $\lim_{x\to c} f(x) = \infty$ and $\lim_{x\to c} g(x) = -\infty$, then what can you tell me about $\lim_{x\to c} f(x) + g(x)$? Answer: nothing, unless you know what specific functions f and g are. As a simple example, suppose

$$f(x) = \frac{1}{(x-c)^2} + 2011,\ g(x) = \frac{-1}{(x-c)^2}.$$

Then $\lim_{x\to c} f(x) = \infty$, $\lim_{x\to c} g(x) = -\infty$, but

$$\lim_{x\to c} f(x) + g(x) = \lim_{x\to c} 2011 = 2011.$$

So $\infty - \infty$ cannot have a universal definition independent of the chosen functions.[3] In a similar way, when evaluating limits $0 \cdot \infty$ is an indeterminate form: if $\lim_{x\to c} f(x) = 0$ and $\lim_{x\to c} g(x) = \infty$, then $\lim_{x\to c} f(x)g(x)$ depends on *how fast* f approaches zero compared to how fast g approaches infinity. Again, consider something like $f(x) = (x-c)^2$, $g(x) = \frac{2011}{(x-c)^2}$. And similarly for $\frac{\infty}{\infty}$.

These are good reasons. However, there are also more purely algebraic reasons: there is no way to define the above expressions in such a way to make the field

[3]In the unlikely event you think that perhaps $\infty - \infty = 2011$ always, try constructing another example...or wait until next semester and ask me again.

axioms work out. For instance, let $a \in \mathbb{R}$. Then $a + \infty = \infty$. If therefore we were allowed to substract ∞ from ∞ we would deduce $a = \infty - \infty$, and thus $\infty - \infty$ could be any real number: that's not a well-defined operation.

Remark: Sometimes above we have alluded to the existence of ordered fields F which do not satisfy the Archimedean axiom, i.e., for which there exist elements x such that $x > n$ for all $n \in \mathbb{Z}^+$. In speaking about elements like x we sometimes call them *infinitely large*. This is a totally different use of "infinity" than the extended real numbers above. Indeed, no ordered field F can have a largest element x, because it follows easily from the field axioms that for any $x \in F$, $x+1 > x$. The moral: although we call $\pm\infty$ "extended real numbers", one should not think of them as being elements of a number system at all, but rather limiting cases of such things.

One of the merits of this extended definition of $\sup S$ and $\inf S$ is that it works nicely with calculations: in particular, all of the "calisthenics" of the previous section have nice analogues for unbounded sets. We leave it to the reader to investigate this phenomenon on her own. In particular though, let's look back at Proposition 6.7: it says that, under conditions ensuring that the sets are nonempty and bounded above / below, that if $X \subset Y \subset \mathbb{R}$, then

$$\sup X \leq \sup Y,$$

$$\inf Y \leq \inf X.$$

This definition could have motivated our definition of sup and inf for unbounded sets, as follows: for $n \in \mathbb{Z}$ and $X \subset \mathbb{R}$, put

$$X^n = \{x \in X \mid x \leq n\},\ X_n = \{x \in X \mid x \geq n\}.$$

The idea here is that in defining X^n we are cutting it off at n in order to force it to be bounded above, but in increasingly generous ways. We have

$$X^0 \subset X^1 \subset \ldots \subset X$$

and also

$$X = \bigcup_{n=0}^{\infty} X^n;$$

in other words, every element of X is a subset of X^n for some n (this is precisely the Archimedean property). Applying Proposition 6.7, we get that for every nonempty subset X of $\mathbb{R}$,

$$\sup X^0 \leq \sup X^1 \leq \sup X^2 \leq \ldots \sup X^n \leq \ldots.$$

Suppose moreover that X is bounded above. Then some $N \in \mathbb{Z}^+$ is an upper bound for X, i.e., $X = X^N = X^{N+1} = \ldots$, so the sequence $\sup X^n$ is eventually constant, and in particular $\lim_{n\to\infty} \sup X^n = \sup X$. On the other hand, if X is bounded above, then the sequence $\sup X^n$ is not eventually constant; in fact it takes increasingly large values, and thus

$$\lim_{n\to\infty} \sup X^n = \infty.$$

Thus if we take as our definition for $\sup X$, $\lim_{n\to\infty} \sup X^n$, then for X which is unbounded above, we get $\sup X = \lim_{n\to\infty} \sup X^n = \infty$. By reflection, a similar discussion holds for $\inf X$.

There is, however, one last piece of business to attend to: we said we wanted $\sup S$ and $\inf S$ to be defined for *all* subsets of $\mathbb{R}$: what if $S = \varnothing$? There is an answer for this as well, but many people find it confusing and counterintuitive at first, so let me approach it again using Proposition 6.7. For each $n \in \mathbb{Z}$, consider the set $P_n = \{n\}$: i.e., P_n has a single element, the integer n. Certainly then $\inf P_n = \sup P_n = n$. So what? Well, I claim we can use these sets P_n along with Proposition 6.7 to see what $\inf \varnothing$ and $\sup \varnothing$ should be. Namely, to define these quantities in such a way as to obey Proposition 6.7, then for all $n \in \mathbb{Z}$, because $\varnothing \subset \{n\}$, we must have

$$\sup \varnothing \leq \sup\{n\} = n$$

and

$$\inf \varnothing \geq \inf\{n\} = n.$$

There is exactly one extended real number which is less than or equal to every integer: $-\infty$. Similarly, there is exactly one extended real number which is greater than or equal to every integer: ∞. Therefore the inexorable conclusion is

$$\sup \varnothing = -\infty, \ \inf \varnothing = \infty.$$

Other reasonable thought leads to this conclusion: for instance, in class I had a lot of success with the "pushing" conception of suprema and infima. Namely, if your set S is bounded above, then you start out to the right of every element of your set – i.e., at some upper bound of S – and keep pushing to the left until you can't push any farther without passing by some element of S. What happens if you try this with $\varnothing$? Well, every real number is an upper bound for $\varnothing$, so start anywhere and push to the left: you can keep pushing as far as you want, because you will never hit an element of the set. Thus you can push all the way to $-\infty$, so to speak. Similarly for infima, by reflection.

2. Intervals and the Intermediate Value Theorem

2.1. Convex subsets of $\mathbb{R}$.

We say that a subset S of $\mathbb{R}$ is **convex** if for all $x < y \in S$, the entire interval $[x, y]$ lies in S. In other words, a convex set is one that whenever two points are in it, all in between points are also in it.

Example 2.1: The empty set $\varnothing$ is convex. For any $x \in \mathbb{R}$, the singleton set $\{x\}$ is convex. In both cases the definition applies *vacuously*: until we have two distinct points of S, there is nothing to check!

Example 2.2: We claim any interval is convex. This is immediate – or it would be, if we didn't have so many different kinds of intervals to write down and check. One needs to see that the definition applies to invervals of all of the following forms:

$$(a, b), \ [a, b), \ (a, b], \ [a, b], \ (-\infty, b), (-\infty, b], \ (a, \infty), [a, \infty), (-\infty, \infty).$$

All these verifications are trivial appeals to things like the transitivity of $\leq$ and $\geq$.

Are there any nonempty convex sets other than intervals? (Just to be sure, we

count $\{x\} = [x, x]$ as an interval.[4]) A little thought suggests that the answer should be *no*. But more thought shows that if so we had better use the Dedekind completeness of $\mathbb{R}$, because if we work over $\mathbb{Q}$ with all of the corresponding definitions then there are nonempty convex sets which are not intervals, e.g.

$$S = \{x \in \mathbb{Q} \mid x^2 < 2\}.$$

This has a familiar theme: replacing $\mathbb{Q}$ by $\mathbb{R}$ we would get an interval, namely $(-\sqrt{2}, \sqrt{2})$, but once again $\pm\sqrt{2} \notin \mathbb{Q}$. When one looks carefully at the definitions it is no trouble to check that *working solely in the rational numbers* S is a convex set but is not an interval.

Remark: Perhaps the above example seems legalistic, or maybe even a little silly. It really isn't: one may surmise that contemplation of such examples led Dedekind to his *construction* of the real numbers via **Dedekind cuts**. This construction may be discussed at the end of this course. Most contemporary analysts prefer a rival construction of $\mathbb{R}$ due to Cauchy using **Cauchy sequences**. I agree that Cauchy's construction is simpler. However, both are important in later mathematics: Cauchy's construction works in the context of a general **metric space** (and, with certain modifications, in a general **uniform space**) to construct an associated **complete space**. Dedekind's construction works in the context of a general **linearly ordered set** to construct an associated Dedekind-complete ordered set.

THEOREM 6.8. *Every nonempty convex subset D of $\mathbb{R}$ is an interval.*

PROOF. We are given a nonempty convex subset D of $\mathbb{R}$ and we want to show it is an interval, but as above an interval can have any one of nine basic shapes. It may be quite tedious to argue that one of nine things must occur!

So we just need to set things up a bit carefully: here goes: let $a \in [-\infty, \infty)$ be the infimum of D, and let $b \in (-\infty, \infty]$ be the supremum of D. Let $I = (a, b)$, and let $\overline{I}$ be the **closure** of I, i.e., if a is finite, we include a; if b is finite, we include b. Step 1: We claim that $I \subset D \subset \overline{I}$. Let $x \in I$.

Case 1: Suppose $I = (a, b)$ with $a, b \in \mathbb{R}$. Let $z \in (a, b)$. Then, since $z > a = \inf D$, there exists $c \in D$ with $c < z$. Similarly, since $z < b = \sup D$, then there exists $d \in D$ with $z < d$. Since D is convex, $z \in D$. Now suppose $z \in D$. We must have $\inf D = a \leq z \leq b = \sup D$.

Case 2: Suppose $I = (-\infty, b)$, and let $z \in I$. Since D is unbounded below, there exists $a \in D$ with $a < z$. Moreover, since $z < \sup D$, there exists $b \in D$ such that $z < b$. Since D is convex, $z \in D$. Next, let $z \in D$. We wish to show that $z \in \overline{I} = (-\infty, b]$; in other words, we want $z \leq b$. But since $z \in D$ and $b = \sup D$, this is immediate. Thus $I \subset D \subset \overline{I}$.

Case 3: Suppose $I = (a, \infty)$. This is similar to Case 2 and is left to the reader.

Case 4: Suppose $I = (-\infty, \infty) = \mathbb{R}$. Let $z \in \mathbb{R}$. Since D is unbounded below, there exists $a \in D$ with $a < z$, and since D is unbounded above there exists $b \in D$ with $z < b$. Since D is convex, $z \in D$. Thus $I = D = \overline{I} = \mathbb{R}$.

Step 2: We claim that any subset D which contains I and is contained in $\overline{I}$ is an interval. Indeed I and $\overline{I}$ are both intervals, and the only case in which there is any subset D strictly in between them is $I = (a, b)$ with $a, b \in \mathbb{R}$ – in this case D could also be $[a, b)$ or $(a, b]$, and both are intervals. □

[4]However, we do not wish to say whether the empty set is an interval. Throughout these notes the reader may notice minor linguistic contortions to ensure that this issue never arises.

Exercise (R. Freiwald): Let F be an ordered field. For a subset S of F, we define the associated **downset**

$$S_\downarrow = \{x \in F \mid x \leq s \text{ for some } s \in S\}.$$

a) Show that $S \neq \varnothing \iff S_\downarrow \neq \varnothing$.
b) Show that S is bounded above $\iff S_\downarrow$ is bounded above.
c) Show that S has a supremum in $F \iff S_\downarrow$ has a supremum in F, and if so then we have $\sup S = \sup S_\downarrow$.
d) In any ordered field F, we say a subset S of F is **convex** if $x, y \in S$ with $x < y$, then for all z with $x < z < y$, $z \in S$. Show that for all nonempty subsets S of F, the downset $S_\downarrow$ is convex.
e) Deduce the converse to Theorem 6.8: if F is an ordered field in which every nonempty convex subset is an interval, then F is Dedekind complete.

Recall that a function $f : D \to \mathbb{R}$ satisfies the **Intermediate Value Property** (IVP) if for all $[a, b] \subset D$, for all L in between $f(a)$ and $f(b)$ is of the form $f(c)$ for some $c \in (a, b)$. As you may well have noticed, the IVP is closely related to the notion of a convex subset. The following result clarifies this connection.

THEOREM 6.9. *For $f : D \subset \mathbb{R} \to \mathbb{R}$, the following are equivalent:*
(i) For all $[a, b] \subset D$, $f([a, b])$ is a convex subset of $\mathbb{R}$.
(ii) f satisfies the Intermediate Value Property.
(iii) For any interval $I \subset D$, $f(I)$ is an interval.

PROOF. (i) $\implies$ (ii): For all $[a, b] \subset D$, $f([a, b])$ is a convex subset containing $f(a)$ and $f(b)$, hence it contains all numbers in between $f(a)$ and $f(b)$.
(ii) $\implies$ (iii): Suppose that f satisfies IVP, and let $I \subset D$ be an interval. We want to show that $f(I)$ is an interval. By Theorem 8.1 it suffices to show that $f(I)$ is convex. Assume not: then there exists $a < b \in I$ and some L in between $f(a)$ and $f(b)$ such that $L \neq f(c)$ for any $c \in I$. In particular $L \neq f(c)$ for any $c \in [a, b]$, contradicting the Intermediate Value Property.
(iii) $\implies$ (i): This is immediate: $[a, b]$ is an interval, so by assumption $f([a, b])$ is an interval, hence a convex subset. □

2.2. The (Strong) Intermediate Value Theorem.

THEOREM 6.10. *(Strong Intermediate Value Theorem) If $f : I \to \mathbb{R}$ is continuous, then f satisfies the Intermediate Value Property and thus $f(I)$ is an interval.*

PROOF. Step 1: We make the following CLAIM: if $f : [a, b] \to \mathbb{R}$ is continuous, $f(a) < 0$ and $f(b) > 0$, then there exists $c \in (a, b)$ such that $f(c) = 0$.
PROOF OF CLAIM: Let $S = \{x \in [a, b] \mid f(x) < 0\}$. Since $a \in S$, S is nonempty. Moreover S is bounded above by b. Therefore S has a least upper bound $c = \sup S$. It is easy to see that we must have $f(c) = 0$. Indeed, if $f(c) < 0$, then – as we have seen several times – there exists $\delta > 0$ such that $f(x) < 0$ for all $x \in (c - \delta, c + \delta)$, and thus there are elements of S larger than c, contradicting $c = \sup S$. Similarly, if $f(c) > 0$, then there exists $\delta > 0$ such that $f(x) > 0$ for all $x \in (c - \delta, c + \delta)$, in which case any element of $(c - \delta, c)$ gives a smaller upper bound for S than c. By the process of elimination we must have $f(c) = 0$!
Step 2: We will show that f satisfies the Intermediate Value Property: for all $[a, b] \subset I$, and any L in between $f(a)$ and $f(b)$, we must find $c \in (a, b)$ such that

$f(c) = L$. If $f(a) = f(b)$ there is nothing to show. If $f(a) > f(b)$, then we may replace f by $-f$ (this is still a continuous function), so it is enough to treat the case $f(a) < L < f(b)$. Now consider the function $g(x) = f(x) - L$. Since f is continuous, so is g; moreover $g(a) = f(a) - L < 0$ and $g(b) = f(b) - L > 0$. Therefore by Step 1 there exists $c \in (a, b)$ such that $0 = g(c) = f(c) - L$, i.e., such that $f(c) = L$.
Step 3: By Step 2 and Theorem 6.9, $f(I)$ is an interval. □

Remark: Theorem 6.10 is in fact a mild improvement of the Intermediate Value Theorem we stated earlier in these notes. This version of IVT applies to continuous functions with domain *any* interval, not just an interval of the form $[a, b]$, and includes a result that we previously called the **Interval Image Theorem**.

2.3. The Intermediate Value Theorem Implies Dedekind Completeness.

THEOREM 6.11. *Let F be an ordered field such that every continuous function $f : F \to F$ satisfies the Intermediate Value Property. Then F is Dedekind complete.*

PROOF. We will prove the contrapositive: suppose F is not Dedekind complete, and let $S \subset F$ be nonempty and bounded above but without a least upper bound in F. Let $\mathcal{U}(S)$ be the set of upper bounds of S. We define $f : F \to F$ by:
• $f(x) = -1$, if $x \notin \mathcal{U}(S)$,
• $f(x) = 1$, $x \in \mathcal{U}(S)$.
Then f is continuous on F – indeed, a point of discontinuity would occur only at the least upper bound of S, which is assumed not to exist. Moreover f takes the value -1 – at any element $s \in S$, which cannot be an upper bound for S because then it would be the *maximum* element of S – and the value 1 at any upper bound for S (we have assumed that S is bounded above so such elements exist), but it never takes the value zero, so f does not satisfy IVP. □

Exercise 2.3: Show in detail that the function $f : F \to F$ constructed in the proof of Theorem 6.11 is continuous at every element of F.

3. The Monotone Jump Theorem

THEOREM 6.12. *(Monotone Jump) Let $f : I \to \mathbb{R}$ be weakly increasing.*
a) Let c be an interior point of I. Then $\lim_{x \to c^-} f(x)$ and $\lim_{x \to c^+} f(x)$ exist, and

$$\lim_{x \to c^-} f(x) \leq f(c) \leq \lim_{x \to c^+} f(x).$$

b) Suppose I has a left endpoint a. Then $\lim_{x \to a^+} f(x)$ exists and is at least $f(a)$.
c) Suppose I has a right endpoint b. Then $\lim_{x \to b^-} f(x)$ exists and is at most $f(c)$.

PROOF. a) Step 0: As usual, we may f is weakly increasing. We define

$$L = \{f(x) \mid x \in I, x < c\}, \ R = \{f(x) \mid x \in I, x > c\}.$$

Since f is weakly increasing, L is bounded above by $f(c)$ and U is bounded below by $f(c)$. Therefore we may define

$$\mathfrak{l} = \sup L, \mathfrak{r} = \inf R.$$

Step 1: For all $x < c$, $f(x) \leq f(c)$, $f(c)$ is an upper bound for L, so $\mathfrak{l} \leq f(c)$. For all $c < x$, $f(c) \leq f(x)$, so $f(c)$ is a lower bound for R, so $f(c) \leq \mathfrak{r}$. Thus

$$\mathfrak{l} \leq f(c) \leq \mathfrak{r}. \tag{22}$$

Step 2: We claim $\lim_{x \to c^-} f(x) = \mathfrak{l}$. To see this, let $\epsilon > 0$. Since $\mathfrak{l}$ is the least upper bound of L and $\mathfrak{l} - \epsilon < \mathfrak{l}$, $\mathfrak{l} - \epsilon$ is not an upper bound for L: there exists $x_0 < c$ such that $f(x_0) > \mathfrak{l} - \epsilon$. Since f is weakly increasing, for all $x_0 < x < c$ we have

$$\mathfrak{l} - \epsilon < f(x_0) \leq f(x) \leq \mathfrak{l} < \mathfrak{l} + \epsilon.$$

Thus we may take $\delta = c - x_0$.
Step 3: We claim $\lim_{x \to c^+} f(x) = \mathfrak{r}$: this is shown as above and is left to the reader.
Step 4: Substituting the results of Steps 2 and 3 into (22) gives the desired result.
b) and c): The arguments at an endpoint are routine modifications of those of part a) above and are left to the reader as an opportunity to check her understanding. □

THEOREM 6.13. *For a monotone function $f : I \to \mathbb{R}$, TFAE:*
(i) $f(I)$ is an interval.
(ii) f is continuous.

PROOF. As usual, it is no loss of generality to assume f is weakly increasing.
(i) $\implies$ (ii): If f is *not* continuous on all of I, then by the Monotone Jump Theorem $f(I)$ fails to be convex. In more detail: suppose f is discontinuous at c. If c is an interior point then either $\lim_{x \to c^-} f(x) < f(c)$ or $f(c) < \lim_{x \to c^+} f(x)$. In the former case, choose any $b \in I$, $b < c$. Then $f(I)$ contains $f(b) < f(c)$ but not the in-between point $\lim_{x \to c^-} f(x)$. In the latter case, choose any $d \in I$, $c < d$. Then $f(I)$ contains $f(c) < f(d)$ but not the in-between point $\lim_{x \to c^+} f(x)$. Similar arguments hold if c is the left or right endpoint of I: these are left to the reader. Thus in all cases $f(I)$ is not convex hence is not an interval.
(ii) $\implies$ (i): This follows immediately from Theorem 6.10. □

With Theorems 6.10 and 6.13 in hand, we get an especially snappy proof of the Continuous Inverse Function Theorem. Let $f : I \to \mathbb{R}$ be continuous and injective. By Theorem 6.10, $f(I) = J$ is an interval. Moreover $f : I \to J$ is a bijection, with inverse function $f^{-1} : J \to I$. Since f is monotone, so is f^{-1}. Moreover $f^{-1}(J) = I$ is an interval, so by Theorem 6.13, f^{-1} is continuous!

4. Real Induction

THEOREM 6.14. *(Principle of Real Induction) Let $a < b$ be real numbers, let $S \subset [a, b]$, and suppose:*
(RI1) $a \in S$,
(RI2) for all $x \in S$, if $x \neq b$ there exists $y > x$ such that $[x, y] \subset S$.
(RI3) For all $x \in \mathbb{R}$, if $[a, x) \in S$, then $x \in S$.
Then $S = [a, b]$.

PROOF. Seeking a contradiction we suppose not: $S' = [a, b] \setminus S$ is nonempty. It is bounded below by a, so has a (finite!) greatest lower bound $\inf S'$. However:
Case 1: $\inf S' = a$. Then by (RI1), $a \in S$, so by (RI2), there exists $y > a$ such that $[a, y] \subset S$, and thus y is a greater lower bound for S' then $a = \inf S'$: contradiction.
Case 2: $a < \inf S' \in S$. If $\inf S' = b$, then $S = [a, b]$. Otherwise, by (RI2) there exists $y > \inf S'$ such that $[\inf S', y'] \subset S$, contradicting the definition of $\inf S'$.
Case 3: $a < \inf S' \in S'$. Then $[a, \inf S') \subset S$, so by (RI3) $\inf S' \in S$: contradiction!
□

Example 4.1: Let us reprove the Intermediate Value Theorem. Recall that the key special case of IVT, from which the full theorem easily follows, is this: if $f : [a, b] \to$

$\mathbb{R}$ is continuous, $f(a) < 0$ and $f(b) > 0$, then there exists $c \in (a, b)$ with $f(c) = 0$. We prove this by real induction, as follows. Let $S = \{x \in [a, b] \mid f(x) \geq 0\}$. We know that S is proper in $[a, b]$, so applying real induction shows that one of (RI1), (RI2) and (RI3) must fail. We have $a \in S$ – so (RI1) holds – and if a continuous function is non-negative on $[a, c)$, then it is also non-negative at c: (RI3). So (RI2) must fail: there exists $y \in (a, b]$ such that $f(y) \geq 0$ but there is no $\epsilon > 0$ such that f is non-negative on $[y, y + \epsilon)$. This implies $f(y) = 0$.

Example 4.1, redux: In class I handled the proof of IVT by Real Induction differently, and in a way which I think gives a better first example of the method (most Real Induction proofs are *not* by contradiction). This strategy follows [**Ka07**]. Namely, IVT is equivalent to: let $f : [a, b] \to \mathbb{R}$ be continuous and nowhere zero. If $f(a) > 0$, then $f(b) > 0$. We prove this by Real Induction. Let

$$S = \{x \in [a, b] \mid f(x) > 0\}.$$

Then $f(b) > 0$ iff $b \in S$. We will show $S = [a, b]$ by real induction, which suffices.
(RI1) By hypothesis, $f(a) > 0$, so $a \in S$.
(RI2) Let $x \in S$, $x < b$, so $f(x) > 0$. Since f is continuous at x, there exists $\delta > 0$ such that f is positive on $[x, x + \delta]$, and thus $[x, x + \delta] \subset S$.
(RI3) Let $x \in (a, b]$ be such that $[a, x) \subset S$, i.e., f is positive on $[a, x)$. We claim that $f(x) > 0$. Indeed, since $f(x) \neq 0$, the only other possibility is $f(x) < 0$, but if so, then by continuity there would exist $\delta > 0$ such that f is negative on $[x-\delta, x]$, i.e., f is both positive and negative at each point of $[x-\delta, x]$: contradiction!

The following result shows that Real Induction does not only uses the Dedekind completeness of $\mathbb{R}$ but actually carries the full force of it.

THEOREM 6.15. *In an ordered field F, the following are equivalent:*
(i) F is Dedekind complete: every nonempty bounded above subset has a supremum.
(ii) F satisfies the Principle of Real Induction: for all $a < b \in F$, a subset $S \subset [a, b]$ satisfying (RI1) through (RI3) above must be all of $[a, b]$.

PROOF. (i) $\implies$ (ii): This is simply a restatement of Theorem 6.14.
(ii) $\implies$ (i): Let $T \subset F$ be nonempty and bounded below by $a \in F$. We will show that T has an infimum. For this, let S be the set of lower bounds m of T with $a \leq m$. Let b be any element of T. Then $S \subset [a, b]$.
Step 1: Observe that $b \in S \iff b = \inf T$. In general the infimum could be smaller, so our strategy is not exactly to use real induction to prove $S = [a, b]$. Nevertheless we claim that S satisfies (RI1) and (RI3).
(RI1): Since a is a lower bound of T with $a \leq a$, we have $a \in S$.
(RI3): Suppose $x \in (a, b]$ and $[a, x) \subset S$, so every $y \in [a, x)$ is a lower bound for T. Then x is a lower bound for T: if not, there exists $t \in T$ such that $t < x$; taking any $y \in (t, x)$, we get that y is not a lower bound for T either, a contradiction.
Step 2: Since F satisfies the Principle of Real Induction, by Step 1 $S = [a, b]$ iff S satisfies (RI2). If $S = [a, b]$, then the element $b \in$ is a lower bound for T, so it must be the infimum of T. Now suppose that $S \neq [a, b]$, so by Step 1 S *does not* satisfy (RI2): there exists $x \in S$, $x < b$ such that for any $y > x$, there exists $z \in (x, y)$ such that $z \notin S$, i.e., z is not a lower bound for T. In other words x is a lower bound for T and no element larger than x is a lower bound for T...so $x = \inf T$. $\square$

Remark: Like Dedekind completeness, "Real Induction" depends only on the ordering relation $<$ and not on the field operations $+$ and $\cdot$. In fact, given any ordered set $(F, <)$ – i.e., we need not have operations $+$ or $\cdot$ at all – it makes sense to speak of Dedekind completeness and also of whether an analogue of Real Induction holds. In [**Cl11**], I proved that Theorem 6.15 holds in this general context: an ordered set F is Dedekind complete iff the only it satisfies a "Principle of Ordered Induction."

5. The Extreme Value Theorem

THEOREM 6.16. *(Extreme Value Theorem)*
Let $f : [a, b] \to \mathbb{R}$ be continuous. Then:
a) f is bounded.
b) f attains a minimum and maximum value.

PROOF. a) Let $S = \{x \in [a, b] \mid f : [a, x] \to \mathbb{R} \text{ is bounded}\}$.
(RI1): Evidently $a \in S$.
(RI2): Suppose $x \in S$, so that f is bounded on $[a, x]$. But then f is continuous at x, so is bounded near x: for instance, there exists $\delta > 0$ such that for all $y \in [x - \delta, x + \delta]$, $|f(y)| \leq |f(x)| + 1$. So f is bounded on $[a, x]$ and also on $[x, x + \delta]$ and thus on $[a, x + \delta]$.
(RI3): Suppose that $x \in (a, b]$ and $[a, x) \subset S$. Now **beware**: this *does not say* that f is bounded on $[a, x)$: rather it says that for all $a \leq y < x$, f is bounded on $[a, y]$. These are really different statements: for instance, $f(x) = \frac{1}{x-2}$ is bounded on $[0, y]$ for all $y < 2$ but it is not bounded on $[0, 2)$. But, as usual, the key feature of this counterexample is a lack of continuity: this f is not continuous at 2. Having said this, it becomes clear that we can proceed almost exactly as we did above: since f is continuous at x, there exists $0 < \delta < x - a$ such that f is bounded on $[x - \delta, x]$. But since $a < x - \delta < x$ we know also that f is bounded on $[a, x - \delta]$, so f is bounded on $[a, x]$.
b) Let $m = \inf f([a, b])$ and $M = \sup f([a, b])$. By part a) we have

$$-\infty < m \leq M < \infty.$$

We want to show that there exist $x_m, x_M \in [a, b]$ such that $f(x_m) = m$, $f(x_M) = M$, i.e., that the infimum and supremum are actually attained as values of f. Suppose that there does not exist $x \in [a, b]$ with $f(x) = m$: then $f(x) > m$ for all $x \in [a, b]$ and the function $g_m : [a, b] \to \mathbb{R}$ by $g_m(x) = \frac{1}{f(x)-m}$ is defined and continuous. By the result of part a), g_m is bounded, but this is absurd: by definition of the infimum, $f(x) - m$ takes values less than $\frac{1}{n}$ for any $n \in \mathbb{Z}+$ and thus g_m takes values greater than n for any $n \in \mathbb{Z}^+$ and is accordingly unbounded. So indeed there must exist $x_m \in [a, b]$ such that $f(x_m) = m$. Similarly, assuming that $f(x) < M$ for all $x \in [a, b]$ gives rise to an unbounded continuous function $g_M : [a, b] \to \mathbb{R}$, $x \mapsto \frac{1}{M-f(x)}$, contradicting part a). So there exists $x_M \in [a, b]$ with $f(x_M) = M$. □

6. The Heine-Borel Theorem

Let $S \subset \mathbb{R}$, and let $\{X_i\}_{i \in I}$ be a family of subsets of $\mathbb{R}$. We say that the family $\{X_i\}$ **covers** S if $S \subset \bigcup_{i \in I} X_i$: in words, this simply means that every element $x \in S$ is also an element of X_i for at least one i.

THEOREM 6.17. *(Heine-Borel) Let $\{U_i\}_{i \in I}$ be any covering of the closed, bounded interval by open intervals U_i. Then the covering $\{U_i\}_{i \in I}$* ***has a finite***

***subcovering**: there is a finite subset $J \subset I$ such that every $x \in [a,b]$ lies in U_j for some $j \in J$.*

PROOF. For an open covering $\mathcal{U} = \{U_i\}_{i \in I}$ of $[a,b]$, let

$$S = \{x \in [a,b] \mid \mathcal{U} \cap [a,x] \text{ has a finite subcovering}\}.$$

We prove $S = [a,b]$ by Real Induction. (RI1) is clear. (RI2): If $U_1, \ldots, U_n$ covers $[a,x]$, then some U_i contains $[x, x+\delta]$ for some $\delta > 0$. (RI3): if $[a,x) \subset S$, let $i_x \in I$ be such that $x \in U_{i_x}$, and let $\delta > 0$ be such that $[x-\delta, x] \in U_{i_x}$. Since $x - \delta \in S$, there is a finite $J \subset I$ with $\bigcup_{i \in J} U_i \supset [a, x-\delta]$, so $\{U_i\}_{i \in J} \cup U_{i_x}$ covers $[a,x]$. □

Exercise: The formulation of the Heine-Borel Theorem given above is superficially weaker than the standard one. A subset of $\mathbb{R}$ is called **open** if it is a union of open intervals. In the usual statement of the Heine-Borel Theorem the U_i's are allowed to be arbitrary open subsets of $\mathbb{R}$. Show that this apparently more general version can be deduced from Theorem 6.17. (Suggestion: for each $x \in [a,b]$, there is an open subset U_x containing x. By definition of open subset, U_x must contain some open interval I_x containing x. Apply Theorem 6.17 to the open covering $\{I_x\}$.)

7. Uniform Continuity

7.1. The Definition; Key Examples.

noindent Let I be an interval and $f : I \to \mathbb{R}$. Then f is **uniformly continuous on I** if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x_1, x_2 \in I$, if $|x_1 - x_2| < \delta$ then $|f(x_1) - f(x_2)| < \epsilon$.

In order to show what the difference is between uniform continuty on I and "mere" continuity on I – i.e., continuity at every point of I – let us rephrase the standard ϵ-δ definition of continuity using the notation above. Namely:

A function $f : I \to \mathbb{R}$ is **continuous on I** if for every $\epsilon > 0$ and every $x_1 \in I$, there exists $\delta > 0$ such that for all $x_2 \in I$, if $|x_1 - x_2| < \delta$ then $|f(x_1) - f(x_2)| < \epsilon$.

These two definitions are eerily (and let's admit it: confusingly, at first) similar: they use all the same words and symbols. The only difference is in the *ordering of the quantifiers*: in the definition of continuity, player two gets to hear the value of ϵ and also the value of x_1 before choosing her value of δ. In the definition of uniform continuity, player two only gets to hear the value of ϵ: thus, her choice of δ must work simultaneously – or, in the lingo of this subject, **uniformly** – across all values of $x_1 \in I$. That's the only difference. Of course, switching the order of quantifiers in general makes a big diffference in the meaning and truth of mathematical statements, and this is no exception. Let's look at some simple examples.

Example 6.1: Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = mx + b$, $m \neq 0$. We claim that f is **uniformly continuous** on $\mathbb{R}$. In fact the argument that we gave for continuity long ago shows this, because for every $\epsilon > 0$ we took $\delta = \frac{\epsilon}{|m|}$. Although we used this δ to show that f is continuous at some arbitrary point $c \in \mathbb{R}$, evidently the choice of δ does not depend on the point c: it works uniformly across all values of c. Thus f is uniformly continuous on $\mathbb{R}$.

Example 6.2: Let $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. This time I claim that our usual proof *did not* show uniform continuity. Let's see it in action. To show that f is continuous at c, we factored $x^2 - c$ into $(x - c)(x + c)$ and saw that to get some control on the other factor $x + c$ we needed to restrict x to some bounded interval around c, say $[c-1, c+1]$. On this interval $|x+c| \leq |x|+|c| \leq |c|+1+|c| \leq 2|c|+1$. So by taking $\delta = \min(1, \frac{\epsilon}{2|c|+1})$ we found that if $|x - c| < \delta$ then

$$|f(x) - f(c)| = |x - c||x + c| \leq \frac{\epsilon}{2|c| + 1} \cdot (2|c| + 1) = \epsilon.$$

But the above choice of δ *depends on* c. So it doesn't show that f is uniformly continuous on $\mathbb{R}$. In fact the function $f(x) = x^2$ is *not* uniformly continuous on $\mathbb{R}$. For instance, take $\epsilon = 1$. If it were uniformly continuous, there would have to be some $\delta > 0$ such that for all $x_1, x_2 \in \mathbb{R}$ with $|x_1 - x_2| < \delta$, $|x_1^2 - x_2^2| < \epsilon$. But this is not possible: take any $\delta > 0$. Then for any $x \in \mathbb{R}$, x and $x + \frac{\delta}{2}$ are less than δ apart, and $|x^2 - (x + \frac{\delta}{2})^2| = |x\delta + \frac{\delta^2}{4}|$. But if I get to choose x after you choose δ, this expression can be made arbitrarily large. In particular, if $x = \frac{1}{\delta}$, then it is strictly greater than 1. So f is not uniformly continuous on $\mathbb{R}$.

Remark: In fact a polynomial function $f : \mathbb{R} \to \mathbb{R}$ is uniformly continuous on $\mathbb{R}$ if and only if it has degree at most one. The reasoning is similar to the above.

So that's sad: uniform continuity is apparently quite rare. But wait! What if the domain is a closed, bounded interval I? For instance, by restricting $f(x) = x^2$ to any such interval, it *is* uniformly continuous. Indeed, we may as well assume $I = [-M, M]$, because any I is contained in such an interval, and uniform continuity on $[-M, M]$ implies uniform continuity on I. Now we need only use the fact that we are assuming $|c| \leq M$ to remove the dependence of δ on c: since $|c| \leq M$ we have $\frac{\epsilon}{2|c|+1}) \geq \frac{1}{2M+1}$, so for $\epsilon > 0$ we may take $\delta = \min(1, \frac{1}{2M+1})$. This shows that $f(x) = x^2$ is uniformly continuous on $[-M, M]$.

It turns out that one can always recover uniform continuity from continuity by restricting to a closed bounded interval: this is the last of our Interval Theorems.

7.2. The Uniform Continuity Theorem.

Let $f : I \to \mathbb{R}$. For $\epsilon, \delta > 0$, let us say that f is (ϵ, δ)-UC on I if for all $x_1, x_2 \in I$, $|x_1 - x_2| < \delta \implies |f(x_1) - f(x_2)| < \epsilon$. This is a sort of halfway unpacking of the definition of uniform continuity. More precisely, $f : I \to \mathbb{R}$ is uniformly continuous iff for all $\epsilon > 0$, there exists $\delta > 0$ such that f is (ϵ, δ)-UC on I.

The following small technical argument will be applied twice in the proof of the Uniform Continuity Theorem, so advance treatment of this argument should make the proof of the Uniform Continuity Theorem more palatable.

LEMMA 6.18. *(Covering Lemma) Let $a < b < c < d$ be real numbers, and let $f : [a, d] \to \mathbb{R}$. Suppose that for real numbers $\epsilon_1, \delta_1, \delta_2 > 0$,*
- *f is (ϵ, δ_1)-UC on $[a, c]$ and*

• *f is (ϵ, δ_2)-UC on $[b, d]$.*
Then f is $(\epsilon, \min(\delta_1, \delta_2, c - b))$-UC on $[a, b]$.

PROOF. Suppose $x_1 < x_2 \in I$ are such that $|x_1 - x_2| < \delta$. Then it cannot be the case that both $x_1 < b$ and $c < x_2$: if so, $x_2 - x_1 > c - b \geq \delta$. Thus we must have either that $b \leq x_1 < x_2$ or $x_1 < x_2 \leq c$. If $b \leq x_1 < x_2$, then $x_1, x_2 \in [b, d]$ and $|x_1 - x_2| < \delta \leq \delta_2$, so $|f(x_1) - f(x_2)| < \epsilon$. Similarly, if $x_1 < x_2 \leq c$, then $x_1, x_2 \in [a, c]$ and $|x_1 - x_2| < \delta \leq \delta_1$, so $|f(x_1) - f(x_2)| < \epsilon$. □

THEOREM 6.19. *(Uniform Continuity Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is uniformly continuous on $[a, b]$.*

PROOF. For $\epsilon > 0$, let $S(\epsilon)$ be the set of $x \in [a, b]$ such that there exists $\delta > 0$ such that f is (ϵ, δ)-UC on $[a, x]$. To show that f is uniformly continuous on $[a, b]$, it suffices to show that $S(\epsilon) = [a, b]$ for all $\epsilon > 0$. We will show this by Real Induction.
(RI1): Trivially $a \in S(\epsilon)$: f is (ϵ, δ)-UC on $[a, a]$ for all $\delta > 0$!
(RI2): Suppose $x \in S(\epsilon)$, so there exists $\delta_1 > 0$ such that f is (ϵ, δ_1)-UC on $[a, x]$. Moreover, since f is continuous at x, there exists $\delta_2 > 0$ such that for all $c \in [x, x+\delta_2]$, $|f(c)-f(x)| < \frac{\epsilon}{2}$. Why $\frac{\epsilon}{2}$? Because then for all $c_1, c_2 \in [x-\delta_2, x+\delta_2]$,

$$|f(c_1) - f(c_2)| = |f(c_1) - f(x) + f(x) - f(c_2)| \leq |f(c_1) - f(x)| + |f(c_2) - f(x)| < \epsilon.$$

In other words, f is (ϵ, δ_2)-UC on $[x-\delta_2, x+\delta_2]$. We apply the Covering Lemma to f with $a < x - \delta_2 < x < x + \delta_2$ to conclude that f is $(\epsilon, \min(\delta, \delta_2, x - (x - \delta_2))) = (\epsilon, \min(\delta_1, \delta_2))$-UC on $[a, x + \delta_2]$. It follows that $[x, x + \delta_2] \subset S(\epsilon)$.
(RI3): Suppose $[a, x) \subset S(\epsilon)$. As above, since f is continuous at x, there exists $\delta_1 > 0$ such that f is (ϵ, δ_1)-UC on $[x-\delta_1, x]$. Since $x - \frac{\delta_1}{2} < x$, by hypothesis there exists δ_2 such that f is (ϵ, δ_2)-UC on $[a, x - \frac{\delta_1}{2}]$. We apply the Covering Lemma to f with $a < x-\delta_1 < x-\frac{\delta_1}{2} < x$ to conclude that f is $(\epsilon, \min(\delta_1, \delta_2, x-\frac{\delta_1}{2}-(x-\delta_1))) = (\epsilon, \min(\frac{\delta_1}{2}, \delta_2))$-UC on $[a, x]$. Thus $x \in S(\epsilon)$. □

8. The Bolzano-Weierstrass Theorem For Subsets

Let $S \subset \mathbb{R}$. We say that $x \in \mathbb{R}$ is a **limit point** of S if for every $\delta > 0$, there exists $s \in S$ with $0 < |s - x| < \delta$. Equivalently, x is a limit point of S if every open interval I containing x also contains an element s of S which is not equal to x.

PROPOSITION 6.20. *For $S \subset \mathbb{R}$ and $x \in \mathbb{R}$, the following are equivalent:*
(i) Every open interval I containing x also contains infinitely many *points of S.*
(ii) x is a limit point of S.

Example: If $S = \mathbb{R}$, then every $x \in \mathbb{R}$ is a limit point. More generally, if $S \subset \mathbb{R}$ is **dense** – i.e., if every nonempty open interval I contains an element of S – then every point of $\mathbb{R}$ is a limit point of S. In particular this holds when $S = \mathbb{Q}$ and when $S = \mathbb{R} \setminus \mathbb{Q}$. Note that these examples show that a limit point x of S may or may not be an element of S: both cases can occur.

Example: If $S \subset T$ and x is a limit point of S, x is a limit point of T.

Example: No finite subset S of $\mathbb{R}$ has a limit point.

Example: The subset $\mathbb{Z}$ has no limit points: indeed, for any $x \in \mathbb{R}$, take $I = (x - 1, x + 1)$. Then I is bounded so contains only finitely many integers.

Example: More generally, let S be a subset such that for all $M > 0$, $S \cap [-M, M]$ is finite. Then S has no limit points.

THEOREM 6.21. *(Bolzano-Weierstrass)*
Every infinite subset $\mathcal{A}$ of $[a, b]$ has a limit point.

PROOF. Let $\mathcal{A} \subset [a, b]$, and let S be the set of x in $[a, b]$ such that *if* $\mathcal{A} \cap [a, x]$ is infinite, it has a limit point. It suffices to show $S = [a, b]$, which we will do by Real Induction. (RI1) is clear. (RI2) Suppose $x \in [a, b) \cap S$. If $\mathcal{A} \cap [a, x]$ is infinite, then it has a limit point and hence so does $\mathcal{A} \cap [a, b]$: thus $S = [a, b]$. If for some $\delta > 0$, $\mathcal{A} \cap [a, x + \delta]$ is finite, then $[x, x + \delta] \subset S$. Otherwise $\mathcal{A} \cap [a, x]$ is finite but $\mathcal{A} \cap [a, x + \delta]$ is infinite for all $\delta > 0$, and then x is a limit point for $\mathcal{A}$ and $S = [a, b]$ as above. (RI3) If $[a, x) \subset S$, then: either $\mathcal{A} \cap [a, y]$ is infinite for some $y < x$, so $x \in S$; or $\mathcal{A} \cap [a, x]$ is finite, so $x \in S$; or $\mathcal{A} \cap [a, y]$ is finite for all $y < x$ and $\mathcal{A} \cap [a, x]$ is infinite, so x is a limit point of $\mathcal{A} \cap [a, x]$ and $x \in S$. □

Remark: Later we will give a "sequential version" of Bolzano-Weierstrass and see that it is equivalent to Theorem 6.21 above.

9. Tarski's Fixed Point Theorem

Mainly as a further showpiece for Real Induction, we give here a special case of a theorem of the great logician A. Tarski [**Ta55**]. It is a fixed point theorem: that is, a theorem which gives conditions on a function $f : X \to X$ for there to be a point $c \in X$ with $f(c) = c$. Fixed point theorems are ubiquitously useful throughout mathematics, and further fixed point theorems for functions defined on subintervals of $\mathbb{R}$ will be given when we study infinite sequences later on. However Tarski's theorem has a quite different flavor from the fixed point theorems to come in that the function f is not assumed to be continuous!

THEOREM 6.22. *(Tarski Fixed Point Theorem) Let $f : [a, b] \to [a, b]$ be a weakly increasing function. Then f has a **fixed point**: there is $c \in [a, b]$ with $f(c) = c$.*

PROOF. Seeking a contradiction we suppose f has no fixed point. Let $S = \{x \in [a, b] \mid f(x) > x\}$. We will show $S = [a, b]$ by Real Induction. But then $b \in S$, so $f(b) > b$, contradicting the fact that b is the largest element of $[a, b]$!
(RI1) Since $a = \min[a, b]$, $f(a) \geq a$; since there is no fixed point, $f(a) > a$.
(RI2) Let $x \in [a, b)$ and suppose $[a, x] \subset S$. Put $y = f(x)$, so $y > x$ by definition of S. We claim $[x, y] \subset S$. If not, there is $x \leq z \leq y$ with $f(z) < z \leq y$; but since $x \leq z$, $y = f(x) \leq f(z)$ and we get $y \leq f(z) < z \leq y$, a contradiction.
(RI3) Let $x \in (a, b]$ and suppose $[a, x) \subset S$. We claim $x \in S$. If not, $y = f(x) < x$, so $y \in S$ and thus $f(y) > y = f(x)$; but since $y < x$, $f(y) \leq f(x)$. □

Exercise: a) Give another (more standard) proof of Theorem 6.22 by showing that $\sup\{x \in [a, b] \mid f(x) \geq x\}$ is a fixed point of f.
b) Which argument do you prefer?

Exercise: Find an application of Theorem 6.22 to calculus.

I am indebted to J. Propp for alerting me to this application of Real Induction.

CHAPTER 7

Differential Miscellany

1. L'Hôpital's Rule

We have come to the calculus topic most hated by calculus instructors: L'Hôpital's Rule. This result gives a criterion for evaluating **indeterminate forms** $\frac{0}{0}$ or $\frac{\infty}{\infty}$. The following succinct formulation is taken from [**R**, Thm. 5.13], as is the proof.

THEOREM 7.1. *Let $-\infty \leq a < b \leq \infty$. Let $f, g : (a, b) \to \mathbb{R}$ be differentiable.*
a) We suppose that

$$\lim_{x \to b^-} \frac{f'(x)}{g'(x)} = A \in [-\infty, \infty]$$

and also that either
(i) $\lim_{x \to b^-} f(x) = \lim_{x \to b^-} g(x) = 0$, *or*
(ii) $\lim_{x \to b^-} g(x) = \pm\infty$.
Then $\lim_{x \to b^-} \frac{f(x)}{g(x)} = A$.
b) The analogous statements with $\lim_{x \to b^-}$ *replaced everywhere by* $\lim_{x \to a^+}$ *hold.*

PROOF. a) Step 1: Suppose that $A < \infty$, let α be any real number which is greater than A, and let β be such that $A < \beta < \alpha$. We will show that there exists $c \in (a, b)$ such that for all $x > c$, $\frac{f(x)}{g(x)} < \alpha$.
First, since $\frac{f'(x)}{g'(x)} \to A < \infty$, there is $c \in (a, b)$ such that for all $x > c$, $\frac{f'(x)}{g'(x)} < \beta$. Let $c < x < y < b$. By Cauchy's Mean Value Theorem, there is $t \in (x, y)$ such that

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(t)}{g'(t)} < \beta \tag{23}$$

Suppose first that (i) holds. Then by letting x approach b in (23) we get $\frac{f(y)}{g(y)} \leq \beta < \alpha$ for all $c < y < b$, which is what we wanted to show.
Next suppose that (ii) holds. Fix y in (23) and choose $c_1 \in (y, b)$ such that $c_1 < x < b$ implies $g(x) > g(y)$ and $g(x) > 0$. Multiplying (23) by $\frac{g(x) - g(y)}{g(x)}$ gives

$$\frac{f(x) - f(y)}{g(x)} < \beta \left(\frac{g(x) - g(y)}{g(x)} \right),$$

and then a little algebra yields

$$\frac{f(x)}{g(x)} < \beta - \beta \frac{g(y)}{g(x)} + \frac{f(y)}{g(x)}. \tag{24}$$

Letting x approach b, we find: there is $c \in (c_1, b)$ such that for all $x > c$, $\frac{f(x)}{g(x)} < \alpha$.
Step 2: Suppose $A > -\infty$. Then arguing in a very similar manner as in Step 1 we may show that for any $\alpha < A$ there exists $c \in (a, b)$ such that for all $x > c$,

$\frac{f(x)}{g(x)} > \alpha$. Putting together these two estimates shows $\lim_{x \to b^-} \frac{f(x)}{g(x)} = A$.
b) This is quite straightforward and left to the reader.[1] □

Remark: Perhaps you were expecting the additional hypothesis $\lim_{x \to b^-} f(x) = \pm\infty$ in condition (ii). As the proof shows, this is not necessary. But it seems to be very risky to present the result to freshman calculus students in this form!

Example 7.1: We claim that for all $n \in \mathbb{Z}^+$, $\lim_{x \to \infty} \frac{x^n}{e^x} = 0$. We show this by induction on n. First we do $n = 1$: $\lim_{x \to \infty} \frac{x}{e^x} = 0$. Since $\lim_{x \to \infty} g(x) = 0$ and $\lim_{x \to \infty} \frac{f'(x)}{g'(x)} = \lim_{x \to \infty} \frac{1}{e^x} = 0$, condition (ii) of L'Hôpital's Rule applies so $\lim_{x \to \infty} \frac{x}{e^x} = 0$. Induction Step: let $n \in \mathbb{Z}^+$ and suppose $\lim_{x \to \infty} \frac{x^n}{e^x} = 0$. Then $\lim_{x \to \infty} \frac{x^{n+1}}{e^x} = \frac{\infty}{\infty} \stackrel{\text{LH}}{=} \lim_{x \to \infty} \frac{(n+1)x^n}{e^x} = (n+1)\left(\lim_{x \to \infty} \frac{x^n}{e^x}\right) = (n+1) \cdot 0 = 0$.

Why do calculus instructors not like L'Hôpital's Rule? Oh, let us count the ways!

1) Every derivative $f'(x) = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$ is of the form $\frac{0}{0}$. Thus many calculus students switch to applying L'Hôpital's Rule instead of evaluating derivatives from the definition. This can lead to painfully circular reasoning. For instance, what is $\lim_{x \to 0} \frac{\sin x}{x}$? Well, both numerator and denominator approach 0 and $\lim_{x \to 0} \frac{(\sin x)'}{x'} = \lim_{x \to 0} \frac{\cos x}{1} = \cos 0 = 1$. What's wrong with this? Well, how do we know that $(\sin x)' = \cos x$? Thinking back, we reduced this to computing the derivative of $\sin x$ at $x = 0$, i.e., to showing that $\lim_{x \to 0} \frac{\sin x}{x} = 1$!

2) Many limits which can be evaluated using L'Hôpital's Rule can also be evaluated in many other ways, and often just by thinking a bit about how functions actually behave. For intance, try to evaluate the limit of Example 7.1 above without using L'Hôpital. There are any number of ways. For instance:

LEMMA 7.2. *(Racetrack Principle) Let $f, g : [a, \infty) \to \mathbb{R}$ be two differentiable functions such that $f'(x) \geq g'(x)$ for all $x \geq a$. Then:*
a) We have $f(x) - f(a) \geq g(x) - g(a)$ for all $x \geq a$.
b) If $f'(x) > g'(x)$ for all $x > a$, then $f(x) - f(a) > g(x) - g(a)$ for all $x > a$.

PROOF. Put $h = f - g : [a, \infty) \to \mathbb{R}$.
a) Then $h'(x) \geq 0$ for all $x \geq a$, so h is weakly increasing on (a, ∞), and thus being continuous, weakly increasing on $[a, \infty)$: for all $x \geq a$, $f(x) - g(x) = h(x) \geq h(a) = f(a) - g(a)$, and thus $f(x) - f(a) \geq g(x) - g(a)$.
b) This is the same as part a) with all instances of $\geq$ replaced by $>$: details may be safely left to the reader. □

PROPOSITION 7.3. *Let $f : [a, \infty) \to \mathbb{R}$ be twice differentiable such that:*
(i) $f'(a) > 0$ and
(ii) $f''(x) \geq 0$ for all $x \geq a$.
Then $\lim_{x \to \infty} f(x) = \infty$.

PROOF. Let g be the tangent line to f at $x = a$, viewed as a function from $[a, \infty)$ to $\mathbb{R}$. Because $f'' = (f')'$ is non-negative on $[a, \infty)$, f' is weakly increasing

[1] In fact, [**R**] states and proves the result with $\lim_{x \to a^+}$ instead of $\lim_{x \to b^-}$. I recast it this way since a natural class of examples concerns $\lim_{x \to \infty}$.

on $[a, \infty)$, and thus for all $x \geq a$,

$$f'(x) \geq f'(a) = g'(x).$$

Applying the Racetrack Principle, we get that for all $x \geq a$,

$$f(x) - f(a) \geq g(x) - g(a),$$

or

$$f(x) \geq g(x) + f(a) - g(a) = g(x).$$

Since g is a linear function with positive slope,

$$\lim_{x \to \infty} f(x) \geq \lim_{x \to \infty} g(x) = \infty.$$

□

(i) Let $f_n(x) = \frac{e^x}{x^n}$. One can therefore establish $\lim_{x \to \infty} \frac{e^x}{x^n} = \infty$ by showing that $f_n'(x), f_n''(x)$ are both positive for sufficiently large x. It is easy to see that $f_n'(x) > 0$ for all $x > n$. The analysis for f_n'' is a bit messier; we leave it to the reader and try something slightly different instead.
(ii) Since $f_n'(x) > 0$ for all $x > n$, f is eventually increasing and thus tends either to a positive limit A or to $+\infty$. But as $x \to \infty$, $x + 1 \to \infty$, so

$$A = \lim_{x \to \infty} \frac{e^{x+1}}{(x+1)^n} = e \lim_{x \to \infty} \frac{e^x}{(x+1)^n} = eA.$$

The only $A \in (0, \infty]$ which satisfies $A = eA$ is $A = \infty$.
(iii) Take logarithms: if $A = \lim_{x \to \infty} \frac{e^x}{x^n}$, then

$$\log A = \lim_{x \to \infty} \log \frac{e^x}{x^n} = \lim_{x \to \infty} x - n \log x.$$

Now if $l(x) = x - n \log x$, then $l'(x) = 1 - \frac{n}{x}$, $l''(x) = \frac{n}{x^2}$; both are positive for large enough x, so by Proposition 7.3, $\log A = \infty$ and thus $A = \infty$.
(iv) When we learn about Taylor series we will have access to a superior expression for e^x, namely $\sum_{n=0}^{\infty} \frac{x^n}{n!}$. From this the desired limit follows almost immediately!

3) The statement of L'Hôpital's Rule is complicated and easy for even relatively proficient practitioners of the mathematical craft to misremember or misapply. A classic rookie mistake is to forget to verify condition (i) or (ii): of course in general

$$\lim_{x \to a} \frac{f(x)}{g(x)} \neq \lim_{x \to a} \frac{f'(x)}{g'(x)};$$

try a random example. But there are subtler pitfalls as well. For instance, even under conditions (i) and (ii), $\lim_{x \to a} \frac{f(x)}{g(x)} = A$ need not imply that $\lim_{x \to a} \frac{f'(x)}{g'(x)}$ exists, so you cannot use L'Hopital's Rule to show that a limit *does not* exist.

Example 7.2: Let $f, g : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2 \sin(\frac{1}{x})$ and $f(0) = 0$ and $g(x) = x$. Then f and g are both differentiable (for f this involves going back to the limit definition of the derivative at $x = 0$ – we have seen this example before), and $\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} x \sin(\frac{1}{x}) = 0$. However,

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right) = -\lim_{x \to 0} \cos\left(\frac{1}{x}\right),$$

and this limit does not exist.

Nevertheless, every once in a while one really does need L'Hôpital's Rule! We will encounter such a situation in our study of Taylor polynomials. And, you know, everyone else is doing it, so you should at least know how to do it...

Exercise: Use L'Hôpital's Rule to give a short proof of Theorem 5.32.

2. Newton's Method

2.1. Introducing Newton's Method.

Newton's Method is an important procedure for approximating roots of differentiable functions. Namely, suppose that $y = f(x)$ is a differentiable function and that we know – perhaps by the methods of calculus! – that there is a real number c such that $f(c) = 0$. Well, in general there are many, but suppose we restrict ourselves to some small interval in which we believe there is a unique root c. Say we do not need to know c exactly, but that we wish to approximate to any prescribed degree of accuracy – e.g. we may need to know its value to 100 decimal places.

The first key idea is that Newton's method is one of **sucessive approximations**. That is, we start with a number x_1 which is an "approximate root" of f, i.e., $f(x_1)$ is rather close to 0 (this is not a precise mathematical statement, but the closer x_1 is to the true root c, the better the method will work. If it's too far away, then it won't work at all.) Then we perform some ameloriation procedure resulting in a second approximate root x_2, which is (in general) closer to the true root c than x_1 is. And then we continue: performing our amelioration procedure again we get x_3, performing it a third time we get x_4, and so forth, resulting in an infinite sequence of approximate roots $\{x_n\}_{n=1}^{\infty}$.

This amelioration procedure is very geometric: let $n \in \mathbb{Z}^+$, and start with the approximate root x_n. What we do is consider the tangent line $l_n(x)$ to $y = f(x)$ at $x = x_n$. The equation of this line is

$$y - f(x_n) = f'(x_n)(x - x_n),$$

so

$$y = l_n(x) = f(x_n) + f'(x_n)(x - x_n).$$

Now we take x_{n+1} to be x-intercept of the line $l_n(x)$, i.e., the unique number such that $l_n(x_{n+1}) = 0$. So let's do it:

$$0 = l_n(x_{n+1}) = f(x_n) + f'(x_n)(x_{n+1} - x_n),$$

so

$$x_{n+1} - x_n = \frac{-f(x_n)}{f'(x_n)}$$

or

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{25}$$

Note that our expression for x_{n+1} is undefined if $f'(x_n) = 0$, as well it should be: if the tangent line at x_n is horizontal, then either it coincides with the x-axis (in which case x_n is already a root of f and no amelioration is needed) or it is parallel

to the x-axis, in which case the method breaks down: in a sense we will soon make precise, this means that x_n is "too far away" from the true root c of f.

2.2. A Babylonian Algorithm.

We can use Newton's method to approximate $\sqrt{2}$. Consider $f(x) = x^2 - 2$; straightforward calculus tells us that there is a unique positive number c such that $f(c) = 2$ and that for instance $c \in [1, 2]$. We compute the amelioration formula in this case:

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{2x_n^2 - (x_n^2 - 2)}{2x_n} = \frac{x_n^2 + 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right). \tag{26}$$

In other words, to get from x_n to x_{n+1} we take the average of x_n and $\frac{2}{x_n}$.

In this case we are allowed to take any $x_1 \neq 0$; if we want to approximate the positive root of $x^2 - 2 = 0$, it is more than plausible that we should start with a positive number. So let's try

$$x_1 = 1.$$

Then

$$x_2 = \frac{1}{2}\left(1 + \frac{2}{1}\right) = \frac{3}{2} = 1.5.$$

$$x_3 = \frac{1}{2}\left(\frac{3}{2} + \frac{4}{3}\right) = \frac{17}{12} = 1.41666\ldots$$

$$x_4 = \frac{1}{2}\left(\frac{17}{12} + \frac{24}{17}\right) = \frac{577}{408} = 1.414215686\ldots.$$

$$x_5 = \frac{1}{2}\left(\frac{577}{408} + \frac{816}{577}\right) = 665857/470832 = 1.41421356237468991\ldots.$$

$$x_6 = \frac{1}{2}\left(\frac{470832}{665857} + \frac{1331714}{470832}\right) = 886731088897/627013566048 = 1.41421356237309504880\ldots.$$

If I now ask my laptop computer to directly compute $\sqrt{2}$, then it tells me[2] that

$$\sqrt{2} = 1.41421356237309504880168872\ldots.$$

Thus x_5 is accurate to 11 decimal places and x_6 is accurate to 23 decimal places. Looking more carefully, it seems that each iteration of the "amelioration process" $x_n \mapsto x_{n+1}$ roughly doubles the number of decimal places of accuracy. If this holds true, it means that the approximations get close to the true root very fast – it we wanted $\sqrt{2}$ to 100 decimal places we would only need to compute x_9.

The formula (26) for successive approximations to $\sqrt{2}$ was known to the ancient Babylonians, thousands of years before tangent lines, calculus of any kind, and Isaac Newton. So in practice Newton's method seems to work very well indeed.

Remark: Similarly, for any $a > 0$, using Newton's method as above with $f(x) = x^2 - a$ leads to the recursion

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right),$$

[2]Of course one should not neglect to wonder how my computer is doing this computation. I don't have access to the source code the software I used, so I don't really know, but it is plausible that it is in fact using some form Newton's method.

application of which with $x_1 = 1$ leads to fantastically good numerical approximations to $\sqrt{a}$. If you ever find yourself on a desert island and needing to compute $\sqrt{a}$ to many decimal placees as part of your engineering research to build a raft that will carry you back to civilization, then this is probably the method you should use. And now if anyone asks you whether "honors calculus" contains any practically useful information, you must tell them that the answer is yes!

2.3. Questioning Newton's Method.

Of course we haven't proven anything yet. Here are two natural questions:

QUESTION 7.4. *Let $f : I \to \mathbb{R}$ be differentiable, and let $c \in I$ be such that $f(c) = 0$.*
a) Is there some subinterval $(c - \delta, c + \delta)$ about the true root c such that starting Newton's method with any $x_1 \in (c - \delta, c + \delta)$ guarantees that the sequence of approximations $\{x_n\}$ gets arbitrarily close to c?
b) Assuming the answer to part a) is yes*, given some $x_1 \in (c-\delta, c+\delta)$ can we give a quantitative estimate on how close x_n is to c as a function of n?*

Questions like these are explored in a branch of mathematics called **numerical analysis**. Most theoretical mathematicians (e.g. me) know little about it, which is a shame because the questions its treats are fundamental and closely related to pure mathematics. (As well as being useful in applications, of course.)

2.4. Introducing Infinite Sequences.

We will give some answers to these questions. First, the business of the x_n's getting arbitrarily close to c should be construed in terms of a limiting process, but one of a kind which is slightly different and in fact simpler than the limit of a real-valued function at a point. Namely, a **real infinite sequence** x_n is simply an ordered list of real numbers $x_1, x_2, \ldots, x_n, \ldots$, or – slightly more formally, is given by a function from the positive integers $\mathbb{Z}^+$ to $\mathbb{R}$, say $f(n) = x_n$. If $L \in \mathbb{R}$, we say the infinite sequence $\{x_n\}$ **converges to L** – and write $x_n \to L$ – if for all $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|x_n - L| < \epsilon$. This is precisely the definition of $\lim_{x\to\infty} f(x) = L$ except that our function f is no longer defined for all (or all sufficiently large) real numbers but only at positive integers. So it is a very close cousin of the types of limit operations we have already studied.

Here is one very convenient property of limits of sequences.

PROPOSITION 7.5. *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers, and let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Suppose that $x_n \to L$. Then $f(x_n) \to f(L)$.*

PROOF. Fix $\epsilon > 0$. Since f is continuous at L, there exists $\delta > 0$ such that $|x - L| < \delta \implies |f(x) - f(L)| < \epsilon$. Moreover, since $x_n \to L$, there exists a positive integer N such that for all $n \geq N$, $|x_n - L| < \delta$. Putting these together: if $n \geq N$ then $|x_n - L| < \delta$, so $|f(x_n) - f(L)| < \epsilon$. This shows that $f(x_n) \to L$. □

Remark: a) Proposition 7.5 is a close cousin of the fact that compositions are continuous functions are continuous, and in particular the proof is almost the same.
b) At the moment we are just getting a taste of infinite sequences. Later in the course we will study them more seriously and show that Proposition 7.5 has a very

important converse: if $f : \mathbb{R} \to \mathbb{R}$ is a function such that whenever $x_n \to L$, we have also $f(x_n) \to f(L)$, then f is continuous. Thus preservation of limits of sequences is actually a *characteristic property* of continuous functions, and this suggests (correctly!) that sequences can be a powerful tool in studying functions $f : \mathbb{R} \to \mathbb{R}$.

Thus Newton's method starts with $x_1 \in \mathbb{R}$ and produces an infinite sequence $\{x_n\}$ of "successive aproximations", and our first question is whether – or more precisely, when, i.e., for which choices of x_1 – this sequence converges to the true root c.

2.5. Contractions and Fixed Points.

Recall that a **fixed point** of $f : I \to \mathbb{R}$ is a point $c \in I$ with $f(c) = c$.

A function $f : I \to \mathbb{R}$ is **contractive** (or is **a contraction**) if there is $\alpha < 1$ such that for all $x, y \in I$, $|f(x) - f(y)| \leq \alpha|x - y|$; a real number α for which such an inequality holds will be called a **contraction constant** for f.

Exercise: a) Which functions $f : I \to \mathbb{R}$ have contraction constant 0?
b) Show that every contraction $f : I \to \mathbb{R}$ is continuous.

PROPOSITION 7.6. *Let $f : I \to \mathbb{R}$ be differentiable, and suppose there is $\alpha < 1$ such that for all $x \in I$, $|f'(x)| \leq \alpha$. Then α is a contraction constant for f.*

PROOF. Let $x < y \in I$. By the Mean Value Theorem, there is $c \in (x, y)$ such that

$$\frac{f(x) - f(y)}{x - y} = f'(c),$$

so

$$|f(x) - f(y)| = |f'(c)||x - y| \leq \alpha|x - y|.$$

□

LEMMA 7.7. *A contractive function $f : I \to \mathbb{R}$ has at most one fixed point.*

PROOF. Suppose there are $x_1 \neq x_2$ in I with $f(x_1) = x_1$ and $f(x_2) = x_2$. Let $\alpha < 1$ be a contraction constant for f. Then $|x_1 - x_2| = |f(x_1) - f(x_2)| \leq \alpha|x_1 - x_2|$; dividing through by $|x_1 - x_2|$ gives $1 \leq \alpha$: contradiction. □

Suppose $f : I \to \mathbb{R}$ is a contraction with constant α, and let $c \in I$ be a fixed point of f. Then for any $\delta > 0$, if $x \in [c - \delta, c + \delta]$, then

$$|f(x) - c| = |f(x) - f(c)| \leq \alpha|x - c| < |x - c| \leq \delta,$$

so $f : [c - \delta, c + \delta] \to [c - \delta, c + \delta]$. This is a key point, because given any $f : I \to I$ and any $x_1 \in [c - \delta, c + \delta] \subset I$, we may define a sequence

$$x_1, x_2 = f(x_1), x_3 = f(x_2) = f(f(x_1)), \ldots, x_{n+1} = f(x_n) = (f \circ \cdots \circ f)(x_1), \ldots.$$

We call this the **sequence of iterates of x_1 under f**.

Example: Let $\alpha \in \mathbb{R}$, and consider the function

$$f : \mathbb{R} \to \mathbb{R}, f(x) = \alpha x.$$

Let us study the **dynamics** of iteration of this very simple function.
Step 1: What are the fixed points? We set $c = f(c) = \alpha c$. Then $c = 0$ is a fixed point no matter what α is. Conversely, if $c \neq 0$ then we divide through to get

$\alpha = 1$, so if $\alpha \neq 1$ then 0 is the only fixed point. Finally, if $\alpha = 1$ then every $c \in \mathbb{R}$ is a fixed point, i.e., f is the identity function.

Step 2: Let us try to figure out the limiting behavior of the sequences of iterates. First observe that for any fixed point c, the sequence of iterates will be constant:

$$c, c, c, c \ldots,$$

so of course it will converge to c. So really we are interested in the case when x_1 is *not* a fixed point of f. In fact we can – and this is an unusual feature arising because of the very simple f we choose – give an explicit formula for the general term x_n in the sequence of iterates. Starting at x_1 we have $x_2 = \alpha x$, $x_3 = \alpha x_2 = \alpha(\alpha x_1) = \alpha^2 x_1$, and so on: in general we have $x_n = \alpha^{n-1} x_1$.
Case 1: If $\alpha = 0$, then no matter what x_1 is, for all $n \geq 2$, $x_n = 0$. Thus the sequence is *eventually constant*, with eventual value the unique fixed point: 0.
Case 2: If $\alpha = 1$, then for all n, $x_n = x_1$, so the sequence is constant, and this constant value is a fixed point of f.
Case 3: If $0 < |\alpha| < 1$, then $x_n = \alpha^{n-1} x_1 \to 0$. Here, no matter what the initial point x_1 is, the sequence of iterates converges to the unique fixed point $c = 0$.
Case 4: If $1 < |\alpha|$, then for all $x_1 \neq 0$, $|x_n| = |\alpha^{n-1} x_1| \to \infty$: the sequence of iterates grows without bound, in fact with exponential speed. In particular it does not converge. One can say more about the signs if desired: e.g. when $x_1 > 0$ and $\alpha > 1$, then every x_n is positive and the sequence of iterates approaches ∞, whereas if $x_1 > 0$ and $\alpha < -1$, then the terms x_n alternate in sign while increasing in absolute value, so they do not approach either ∞ or $-\infty$.

Here are three observations about the above example:

1) If x_1 is a fixed point of f, then the equence of iterates is constant.
2) Whenever the sequence of iterates converges, it converges to a fixed point of f.
3) The sequence of iterates converges for all initial points $x_1 \in \mathbb{R}$ iff $|\alpha| < 1$. But note that $|f(x) - f(y)| = |\alpha||x - y|$, so f is contractive iff $|\alpha| < 1$.

The first observation clearly holds for the iterates of any function $f : I \to I$. What about the last two observations? In fact they also hold quite generally.

LEMMA 7.8. *Suppose $f : I \to I$ is continuous, $x_1 \in I$, and the sequence of iterates $\{x_n\}$ of x_1 under f converges to $L \in I$. Then L is a fixed point of f.*

PROOF. Since f is continuous and $x_n \to L$, by Proposition 7.5 we have $f(x_n) \to f(L)$. But $f(x_n) = x_{n+1}$, and if $x_n \to L$ then certainly $x_{n+1} \to L$ as well. A sequence can have at most one limit, so $f(L) = L$. □

LEMMA 7.9. *(Contraction Lemma) For $c \in \mathbb{R}$ and $\delta > 0$, put $I = [c - \delta, c + \delta]$. Suppose that $f : I \to \mathbb{R}$ is a contraction with constant α and that $f(c) = c$.*
a) For all $x \in I$, $f(x) \in I$, i.e., $f : I \to I$.
b) For any $x_1 \in I$, define a sequence $\{x_n\}_{n=1}^{\infty}$ by $x_{n+1} = f(x_n)$ for all $n \geq 1$. Then for all $n \in \mathbb{Z}^+$, we have $|x_{n+1} - c| \leq \alpha^n |x_1 - c|$. In particular $x_n \to c$.

PROOF. a) This was established above. We repeat the statement here because it is a key point: in order to be able to iterate x_1 under f we need $f : I \to I$.

b) We compute $|x_{n+1} - c| = |f(x_n) - f(c)| \leq \alpha |x_n - c| = \alpha |f(x_{n-1}) - f(c)|$
$$\leq \alpha^2 |x_{n-1} - c| = \alpha^3 |x_{n-2} - c| = \ldots = \alpha^n |x_1 - c|.$$

□

2.6. Convergence of Newton's Method.

Believe it or not, we are now very close to a convergence theorem for the sequence of iterates generated by Newton's method. Let $f : I \to \mathbb{R}$ be a C^2 function – i.e., the second derivative exists and is continuous – and let c be a point of the interior of I such that $f(c) = 0$ and $f'(c) \neq 0$ – a **simple root**. Note that since $f'(c) > 0$, f is increasing through c, so there exists $\delta > 0$ such that f is negative on $[c-\delta, c)$ and positive on $(c, c+\delta]$. In particular c is the *unique* root of f on $[c-\delta, c+\delta]$. What we want to show is that – possibly after shrinking δ – for any choice of initial approximation $x_1 \in [c-\delta, c+\delta]$, the Newton's method sequence converges rapidly to c.

So...what does the material we have just developed have to do with this problem? At first sight it does not seem relevant, because we have been talking about fixed points and are now interested in roots, we have been talking about iterating functions on $[c-\delta, c+\delta]$ and we probably cannot iterate f on this interval: $f(c) = 0$ and 0 need not lie in $[c-\delta, \delta]$, and we have been talking about contractions and f need not be a contraction. What gives?

The answer is that all of our preparations apply not to f but to some auxiliary function defined in terms of f. To figure out what this function is, consider the recursive definition of the Newton's method sequence:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Thus the sequence $\{x_n\}_{n=1}^{\infty}$ *is* generated by repeatedly applying a certain function...it's just not the function f. Rather it is the **amelioration function**

$$T(x) = x - \frac{f(x)}{f'(x)}.$$

Now we have to check that our setup *does* apply to T, quite nicely. First, observe that a point x is a root of f if and only if it is a fixed point of T. Since by our assumption c is the unique root of f in $[c-\delta, c+\delta]$, c is the unique fixed point of T on this interval.

The next order of business is to show that T is contractive, at least in some smaller interval around c. For this we look at the derivative:

$$T'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

This is all kosher, since we have assumed f' is nonzero on $[c-\delta, c+\delta]$ and also that f is C^2. In fact, since f is C^2, T' is continuous. But now a miracle occurs: $T'(c) = 0$. Since T' is continuous at c, this means that by making δ smaller we may assume that $|T'(x)| \leq \alpha$ for all $x \in [c-\delta, c+\delta]$, for any positive α we want! Thus not only can we make T contractive, we can make it contractive with any contractive constant $\alpha \in (0,1)$ we want! Thus we get the following result.

THEOREM 7.10. *(Convergence of Newton's Method) Let $f : I \to \mathbb{R}$ be a C^2 function, i.e., f'' exists and is continuous. Let $c \in \mathbb{R}$ be such that $f(c) = 0$ and $f'(c) \neq 0$. Fix $\alpha \in (0,1)$. Then there exists $\delta > 0$ such that for all $x_1 \in [c-\delta, c+\delta]$, the Newton's method sequence $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ is well-defined and converges rapdily to c in the following sense: for all $n \in \mathbb{Z}^+$, $|x_{n+1} - c| \leq \alpha^n |x_1 - c|$.*

PROOF. Fix $\alpha \in (0,1)$. As above, since $T'(c) = 0$, there exists $\delta > 0$ such that α is a contraction constant for T' on $[c-\delta, c+\delta]$, and therefore, by Lemma 7.9a), T maps $[c-\delta, c+\delta]$ back into $[c-\delta, c+\delta]$, so for any $x_1 \in [c-\delta, c+\delta]$ the sequence of iterates $x_{n+1} = T(x_n)$ is well-defined, and note that it is precisely the Newton's method sequence with initial approximation x_1. By Lemma 7.9b), for all $n \in \mathbb{Z}^+$, $|x_{n+1} - c| \leq \alpha^n |x_1 - c|$. So $x_n \to c$, and indeed it does so exponentially fast. In fact, if we take $\alpha = \frac{1}{10}$ and $\delta \leq 1$, the above estimate ensures that x_n approximates c to within n decimal places of accuracy. □

2.7. Quadratic Convergence of Newton's Method.

Although Theorem 7.10 is a very satisfying result, it is far from the last word on Newton's method. In fact, if we compare the proven convergence rate of Theorem 7.10 with the empirically observed convergence rate of $f(x) = x^2 - 2$, $x_1 = 1$, we see that what we have proved is not as good as what we observed: in our example the number of decimal places of accuracy *doubled* with each iteration, i.e., is an exponential function of n, whereas we proved that the number of decimal places of accuracy must grow at least linearly with n. That's a big difference! And in fact, if you look back at exactly what we proved, it seems to itself suggest that more is true: namely, we get exponential convergence, but the base of the exponential gets better and better as we get closer to the point, i.e., as we get farther out into the sequence of iterates. This is faster than exponential decay with any fixed base.

I believe it should be possible to use the idea expressed at the end of the last paragraph to give a proof of the quadratic convergence of Newton's Method. For this, an essential tool is **Taylor's Theorem With Remainder**, an important result which we will unfortunately not treat until much later on. In fact it is possible to give a short, elementary proof of the quadratic convergence which not only avoids Taylor's theorem but also avoids our analysis of the contractive properties of the amelioration map T. Here it is.

THEOREM 7.11. *(Quadratic Convergence of Newton's Method)*
Let $f : I \to \mathbb{R}$ be twice differentiable, and let $c \in I^\circ$ be a root of f.
a) Suppose there are real numbers $\delta, A, B > 0$ such that for all $x \in [c-\delta, c+\delta]$, $|f'(x)| \geq A$ and $|f''(x)| \leq B$. For any $x_0 \in [c-\delta, c+\delta]$, let $\{x_n\}$ be the Newton's Method sequence with initial value x_0. Then for all $n \in \mathbb{N}$,

$$|x_{n+1} - c| \leq \frac{B}{A} |x_n - c|^2. \tag{27}$$

b) If f'' is continuous on I and $f'(c) \neq 0$, then there are indeed $\delta, A, B > 0$ as in the statement of part a), so that (27) holds for all $x_0 \in [c-\delta, c+\delta]$.

PROOF. a) (W. Miller) First, for $a, b \in \mathbb{R}$, we denote by $|[a,b]|$ the set of all real numbers on the closed line segment from a to b; precisely $|[a,b]| = [a,b]$ if $a \leq b$ and $|[a,b] = [b,a]$ if $a > b$. Similarly, let $|(a,b)| = |[a,b]| \setminus \{a,b\}$ be the open line segment from a to b.

Let $x_0 \in [c-\delta, c+\delta]$, and let $\{x_n\}$ be the Newton's Method sequence of iterates. It will be useful to rewrite the defining recursion as

$$\forall n \in \mathbb{N},\ x_{n+1} - x_n = \frac{-f(x_n)}{f'(x_n)}.$$

Apply the Mean Value Theorem to f on $|[x_n, c]|$: there is $y_n \in |(x_n, c)|$ such that

$$\frac{f(x_n)}{x_n - c} = \frac{f(x_n) - f(c)}{x_n - c} = f'(y_n).$$

Apply the Mean Value Theorem to f' on $|[x_n, y_n]|$: there is $z_n \in |(x_n, y_n))|$ such that

$$\frac{f'(x_n) - f'(y_n)}{x_n - y_n} = f''(z_n).$$

The rest of the proof is a clever calculation: for $n \in \mathbb{N}$, we have

$$|x_{n+1} - c| = |(x_{n+1} - x_n) + (x_n - c)| = \left|\frac{-f(x_n)}{f'(x_n)} + \frac{f(x_n)}{f'(y_n)}\right|$$

$$= \left|\frac{f(x_n)(f'(x_n) - f'(y_n))}{f'(x_n)f'(y_n)}\right| = \left|\frac{f'(x_n)(x_n - c)(f'(x_n) - f'(y_n))}{f'(x_n)f'(y_n)}\right|$$

$$= \left|\frac{f'(x_n) - f'(y_n)}{f'(y_n)}\right| |x_n - c| = \left|\frac{f''(z_n)}{f'(y_n)}\right| |x_n - y_n||x_n - c|$$

$$\leq \left|\frac{f''(z_n)}{f'(y_n)}\right| |x_n - c|^2 \leq \frac{B}{A}|x_n - c|^2.$$

In the second to the last inequality above, we used the fact that since y_n lies between x_n and c, $|x_n - y_n| \leq |x_n - c|$.
b) This is left as an exercise for the reader. □

Exercise: Prove Theorem 7.11b).

Let c be a real number, C a positive real number, and let $\{x_n\}_{n=0}^{\infty}$ be a sequence of real numbers. We say that the sequence $\{x_n\}$ **quadratically converges to c** if
(QC1) $x_n \to c$, and
(QC2) For all $n \in \mathbb{N}$, $|x_{n+1} - c| \leq C|x_n - c|^2$.

Exercise:
a) Show that a sequence to satsify (QC2) but not (QC1).
b) Suppose that a sequence $\{x_n\}_{n=0}^{\infty}$ satisfies $|x_0 - c| < \min(1, \frac{1}{C})$. Show that $\{x_n\}$ quadratically converges to c.
c) Deduce that under the hypotheses of Theorem 7.11, there is $\delta > 0$ such that for all $x_0 \in [c - \delta, c + \delta]$, the Newton's Method sequence quadratically converges to c.

Exercise: Let $\{x_n\}$ be a sequence which is quadrtically convergent to $c \in \mathbb{R}$. Viewing x_n as an approximation to c, one often says that "the number of decimal places of accuracy roughly doubles with each iteration".
a) Formulate this as a statement about the sequence $d_n = \log_{10}(|x_n - c|)$.
b) Prove the precise statement you formulated in part a).

2.8. An example of nonconvergence of Newton's Method.

Consider the cubic function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3 - 2x + 2$, so

$$x_{n+1} = x_n - \frac{x_n^3 - 2x_n + 2}{3x_n^2 - 2}.$$

Take $x_1 = 0$. Then $x_2 = 1$ and $x_3 = 0$, so the sequence of iterates will then alternate between 0 and 1: one calls this type of dynamical behavior a **2-cycle**. (The unique real root is at approximately -1.769, so we are plenty far away from it.)

This example is the tip of a substantial iceberg: **complex dynamics**. If you consider cubic polynomials as functions from the complex numbers $\mathbb{C}$ to $\mathbb{C}$, then you are well on your way to generating those striking, trippy pictures of fractal sets that have been appearing on tee-shirts for the last twenty years. I recommend [**Wa95**] as an especially gentle, but insightful, introduction.

3. Convex Functions

3.1. Convex subsets of Euclidean n-space.

Let n be a positive integer and let $\mathbb{R}^n$ be **Euclidean n-space**, i.e., the set of all ordered n-tuples $(x_1, \ldots, x_n)$ of real numbers. E.g. for $n = 2$ we get the plane $\mathbb{R}^2$, for $n = 3$ one gets "three space" $\mathbb{R}^3$, and for higher n the visual/spatial intuition is less immediate but there is no mathematical obstruction. For the most part $\mathbb{R}^n$ is the subject of the sequel to this course – **multivariable mathematics** – but let's digress a little bit to talk about certain subsets of $\mathbb{R}^n$. In fact for our applications we need only $n = 1$ and $n = 2$.

Given two points $P = (x_1, \ldots, x_n),\ Q = (y_1, \ldots, y_n \in \mathbb{R}^n$ we can add them:

$$P + Q = (x_1 + y_1, \ldots, x_n + y_n),$$

a generalization of "vector addition" that you may be familiar with from physics. Also given any real number λ, we can **scale** any point P by λ, namely

$$\lambda P = \lambda(x_1, \ldots, x_n) := (\lambda x_1, \ldots, \lambda x_n).$$

In other words, we simply multiply every coordinate of P by λ.

Let $P \neq Q$ be points in $\mathbb{R}^n$. There is a unique line passing through P and Q. We can express this line **parameterically** as follows: for every $\lambda \in \mathbb{R}$, let

$$R_\lambda = (1 - \lambda)P + \lambda Q = ((1 - \lambda)x_1 + \lambda y_1, \ldots, (1 - \lambda)x_n + \lambda y_n).$$

In particular $R_0 = P$ and $R_1 = Q$. Thus the line segment $\overline{PQ}$ is given as the set

$$\{R_\lambda \mid 0 \leq \lambda \leq 1\}.$$

Now here is the basic definition: a subset $\Omega \subset \mathbb{R}^n$ is **convex** if for all $P, Q \in \Omega$, the line segment $\overline{PQ}$ is contained in Ω. In other words, if our physical universe is the subset Ω and we stand at any two points of Ω, then we can *see each other*: we have an unobstructed line of view that does not at any point leave Ω. There are many convex subsets in the plane: e.g. interiors of disks, ellipses, regular polygons, the portion of the plane lying on one side of any straight line, and so forth.

Exercise: Let $\Omega_1, \ldots, \Omega_n$ be convex subsets of $\mathbb{R}^n$.
a) Show that $\bigcap_{i=1}^n \Omega_i$ – the set of all points lying in *every* Ω_i – is convex.
b) Show by example that $\bigcup_{i=1}^n \Omega_i$ need not be convex.

When $n = 1$, convex subsets are quite constrained. Recall we have proven:

THEOREM 7.12. *For a subset $\Omega \subset \mathbb{R}$, the following are equivalent:*
(i) Ω is an interval.
(ii) Ω is convex.

3.2. Goals.

In freshman calculus one learns, when graphing a function f, to identify subintervals on which the graph of f is "concave up" and intervals on which it is "concave down". Indeed one learns that the former occurs when $f''(x) > 0$ and the latter occurs when $f''(x) < 0$. But, really, what does this mean?

First, where freshman calculus textbooks say *concave up* the rest of the mathematical world says *convex*; and where freshman calculus textbooks say *concave down* the rest of the mathematical world says *concave*. Moreover, the rest of the mathematical world doesn't speak explicitly of concave functions very much because it knows that f is concave exactly when $-f$ is convex.

Second of all, really, what's going on here? Are we saying that our *definition* of convexity is that $f'' > 0$? If so, exactly why do we care when $f'' > 0$ and when $f'' < 0$: why not look at the third, fourth or seventeenth derivatives? The answer is that we have not a formal definition but an intuitive conception of convexity, which a good calculus text will at least try to nurture: for instance I was taught that a function is convex (or rather "concave up") when its graph *holds water* and that it is concave ("concave down") when its graph *spills water*. This is obviously not a mathematical definition, but it may succeed in conveying some intuition. In less poetic terms, the graph of a convex function has a certain characteristic shape that the eye can see: it looks, in fact, qualitatively like an upward opening parabola or some portion thereof. Similarly, the eye can spot concavity as regions where the graph looks, qualitatively, like a piece of a downward opening parabola. And this explains why one talks about convexity in freshman calculus: it is a qualitative, visual feature of the graph of f that you want to take into account. If you are graphing f and you draw something concave when the graph is actually convex, the graph will "look wrong" and you are liable to draw false conclusions about the behavior of the function.

So, at a minimum, our task at making good mathematical sense of this portion of freshman calculus, comes down to the following:

Step 1: Give a precise *definition* of convexity: no pitchers of water allowed!

Step 2: Use our definition to prove a *theorem* relating convexity of f to the second derivative f'', when f'' exists.

In fact this is an oversimplification of what we will actually do. When we try to nail down a mathematical definition of a convex function, we succeed all too well: there are *five* different definitions, each having some intuitive geometric appeal and each having its technical uses. But we want to be talking about one class

of functions, not four different classes, so we will need to show that all five of our definitions are equivalent, i.e., that any function $f : I \to \mathbb{R}$ which satisfies any one of these definitions in fact satisfies all four. This will take some time.

3.3. Epigraphs.

For a function $f : I \to \mathbb{R}$, we define its **epigraph** to be the set of points $(x, y) \in I \times \mathbb{R}$ which lie on or above the graph of the function. In fewer words,

$$\mathrm{Epi}(f) = \{(x, y) \in I \times \mathbb{R} \mid y \geq f(x)\}.$$

A function $f : I \to \mathbb{R}$ is **convex** if its epigraph $\mathrm{Epi}(f)$ is a convex subset of $\mathbb{R}^2$.

Example: Any linear function $f(x) = mx + b$ is convex.

Example: The function $f(x) = |x|$ is convex.

Example: Suppose $f(x) = ax^2 + bx + c$. Then $\mathrm{Epi}(f)$ is just the set of points of $\mathbb{R}^2$ lying on or above a parabola. From this picture it is certainly *intuitively* clear that $\mathrm{Epi}(f)$ is convex iff $a > 0$, i.e., iff the parabola is "opening upward". But proving from scratch that $\mathrm{Epi}(f)$ is a convex subset is not so much fun.

3.4. Secant-graph, three-secant and two-secant inequalities.

A function $f : I \to \mathbb{R}$ satisfies the **secant-graph inequality** if for all $a < b \in I$ and all $\lambda \in [0, 1]$, we have

$$f((1 - \lambda)a + \lambda b) \leq (1 - \lambda)f(a) + \lambda f(b). \tag{28}$$

As λ ranges from 0 to 1, the expression $(1 - \lambda)a + \lambda b$ is a parameterization of the line segment from a to b. Similarly, $(1 - \lambda)f(a) + \lambda f(b)$ is a parameterization of the line segment from $f(a)$ to $f(b)$, and thus

$$\lambda \mapsto ((1 - \lambda)a + \lambda b, (1 - \lambda)f(a) + \lambda f(b))$$

parameterizes the segment of the **secant line** on the graph of $y = f(x)$ from $(a, f(a))$ to $(b, f(b))$. Thus the secant-graph inequality is asserting that the graph of the function lies on or below the graph of any of its secant line segments.

A function $f : I \to \mathbb{R}$ satisfies the **three-secant inequality** if for all $a < x < b$,

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}. \tag{29}$$

A function $f : I \to \mathbb{R}$ satisfies the **two-secant inequality** if for all $a < x < b$,

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}. \tag{30}$$

Proposition 7.13. *For a function $f : I \to \mathbb{R}$, the following are equivalent:*
(i) f satisfies the three-secant inequality.
(ii) f satisfies the two-secant inequality.
(iii) f satisfies the secant-graph inequality.
(iv) f is convex, i.e., $\mathrm{Epi}(f)$ is a convex subset of $\mathbb{R}^2$.

PROOF. We will show (i) $\implies$ (ii) $\iff$ (iii) $\implies$ (i) and (iii) $\iff$ (iv).
(i) $\implies$ (ii): This is immediate.
(ii) $\iff$ (iii): The two-secant inequality

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a}$$

is equivalent to

$$f(x) \leq f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a) = L_{a,b}(x),$$

say. Now $L_{a,b}(x)$ is a linear function with $L_{a,b}(a) = f(a)$ and $L_{a,b}b() = f(b)$, hence it is the secant line between $(a, f(a))$ and $(b, f(b))$. Thus the two-secant inequality is equivalent to the secant-graph inequality.
(iii) $\implies$ (i): As above, snce the secant line $L_{a,b}(x)$ from $(a, f(a))$ to $(b, f(b))$ has equation $y = f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a)$, the secant graph inequality implies

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a}.$$

To get the other half of the three-secant inequality, note that we also have

$$L_{a,b}(x) = f(b) + \frac{f(a)-f(b)}{b-a}(b-x),$$

and the inequality $f(x) \leq f(b) + \frac{f(a)-f(b)}{b-a}(b-x)$ is easily seen to be equivalent to

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x}.$$

(iii) $\implies$ (iv): Let $P_1 = (x_1, y_1), P_2 = (x_2, y_2) \in \text{Epi}(f)$. We want to show $\text{Epi}(f)$ contains the line segment joining P_1 and P_2. This is clear if $x_1 = x_2 = x$ – in this case the line segment is vertical, since since y_1 and y_2 are both greater than or equal to $f(x)$, so is every point y in between y_1 and y_2. So we may assume $x_1 \neq x_2$ and then that $x_1 < x_2$ (otherwise interchange P_1 and P_2). Seeking a contradiction, we suppose there is $\lambda_1 \in (0,1)$ such that $(1-\lambda_1)P_1 + \lambda_1 P_2 \notin \text{Epi}(f)$: that is,

$$(1-\lambda_1)y_1 + \lambda_1 y_2 < f((1-\lambda_1)x_1 + \lambda_1 x_2).$$

But since $f(x_1) \leq y_1$ and $f(x_2) \leq y_2$, we have

$$(1-\lambda_1)f(x_1) + \lambda_1 f(x_2) \leq (1-\lambda_1)y_1 + \lambda_1 y_2$$

and thus

$$(1-\lambda_1)f(x_1) + \lambda_1 f(x_2) < f((1-\lambda_1)x_1 + \lambda_1 x_2),$$

violating the secant-graph inequality.
(iv) $\implies$ (iii): Let $x < y \in I$. Since $(x, f(x))$ and $(y, f(y))$ lie on the graph of f, they are elements of the epigraph $\text{Epi}(f)$. Since $\text{Epi}(f)$ is convex the line segment joining $(x, f(x))$ and $(y, f(y))$ lies inside $\text{Epi}(f)$. But this line segment is nothing else than the secant line between the two points, and to say that it lies inside the epigraph is to say that the secant line always lies on or above the graph of f. □

COROLLARY 7.14. *(Generalized Two Secant Inequality) Let $f : I \to \mathbb{R}$ be a convex function, and let $a, b, c, d \in I$ with $a < b \leq c < d$. Then*

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(d)-f(c)}{d-c}. \tag{31}$$

PROOF. If $b = c$ then (31) is the Two Secant Inequality, so we may suppose $b < c$. Applying the Two Secant Inequality to the points a, b, c gives

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(b)}{c - b}$$

and applying it to the points b, c, d gives

$$\frac{f(c) - f(b)}{c - b} \leq \frac{f(d) - f(c)}{d - c}.$$

Combining these two inequalities gives the desired result. □

Exercise (Interlaced Secant Inequality): Let $f : I \to \mathbb{R}$ be a convex function, and let $a < b < c < d \in I$. Show that we have

$$\frac{f(c) - f(a)}{c - a} \leq \frac{f(d) - f(b)}{d - b}.$$

3.5. Continuity properties of convex functions.

THEOREM 7.15. *Suppose $f : I \to \mathbb{R}$ is convex, and let $[a, b]$ be any subinterval of I°. Then f is a* Lipschitz function *on $[a, b]$: there exists a constant C such that for all $x, y \in [a, b]$, $|f(x) - f(y)| \leq C|x - y|$.*

PROOF. Choose $u, v, w, z \in I^\circ$ with $u < v < a$ and $b < w < z$. Applying the Generalized Two Secant Inequality twice we get

$$\frac{f(v) - f(u)}{v - u} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(w)}{z - w}.$$

Thus

$$\frac{|f(x) - f(y)|}{|x - y|} \leq \max\left|\frac{f(v) - f(u)}{v - u}\right|, \left|\frac{f(z) - f(w)}{z - w}\right|,$$

so

$$L = \max\left|\frac{f(v) - f(u)}{v - u}\right|, \left|\frac{f(z) - f(w)}{z - w}\right|$$

is a Lipschitz constant for f on $[a, b]$. □

COROLLARY 7.16. *Let I be any* open *interval, and let $f : I \to \mathbb{R}$ be a convex function. Then f is continuous.*

Exercise: Prove Corollary 7.16.

It turns out that – at least according to our definition – a convex function *need not* be continuous on an endpoint of an interval on which it is defined. The following result analyzes endpoint behavior of convex functions.

PROPOSITION 7.17. *Let $a \in \mathbb{R}$, and let I be an interval of the form (a, b), $(a, b]$ or (a, ∞). Let $f : I \to \mathbb{R}$ be convex.*
a) $\lim_{x \to a^+} f(x) \in (-\infty, \infty]$.
b) Suppose that $L = \lim_{x \to a^+} f(x) < \infty$. For $M \in \mathbb{R}$, define a function $\tilde{f} : \{a\} \cup I \to \mathbb{R}$ by $\tilde{f}(a) = M$, $\tilde{f}(x) = f(x)$ for $x \in I$. Then $\tilde{f}$ is convex iff $M \geq L$ and continuous at a iff $M = L$.

Exercise: Prove Proposition 7.17.

3.6. Differentiable convex functions.

THEOREM 7.18. *For a differentiable function $f : I \to \mathbb{R}$, TFAE:*
(i) f is convex.
(ii) f' is weakly increasing.

PROOF. For both directions of the proof it is convenient to consider "fixed" $a < b \in I$ and "variable" $a < x < b$.
(i) $\implies$ (ii): For $x \in (a, b]$, we define

$$s(x) = \frac{f(x) - f(a)}{x - a},$$

and for $x \in [a, b)$ we define

$$S(x) = \frac{f(b) - f(x)}{b - x}.$$

Since f is convex, the three-secant inequality for $a < x < b$ holds:

$$s(x) \leq s(b) = S(a) \leq S(x).$$

Taking limits, we get

$$f'(a) = \lim_{x \to a^+} s(x) \leq s(b) = S(a) \leq \lim_{x \to b^-} S(x) = f'(b).$$

(ii) $\implies$ (i): Let $g(x) = \frac{f(x)-f(a)}{x-a}$. We will show that g is increasing on $(a, b]$: this gibes the two-secant inequality and thus the convexity of f. Since f is differentiable, so is g and

$$g'(x) = \frac{(x - a)f'(x) - (f(x) - f(a))}{(x - a)^2}.$$

By the Mean Value Theorem, $\frac{f(x)-f(a)}{x-a} = f'(y)$ for some $a < y < x$. Since f' is increasing,

$$\frac{f(x) - f(a)}{x - a} = f'(y) \leq f'(x);$$

equivalently

$$(x - a)f'(x) - (f(x) - f(a)) \geq 0.$$

Thus $g'(x) \geq 0$ for all $x \in (a, b]$, so indeed g is increasing on $(a, b]$. □

COROLLARY 7.19. *A differentiable convex function is C^1.*

PROOF. By Theorem 7.18, f' is weakly increasing so has only jump discontinuities. By Darboux's Theorem, a derivative cannot have jump discontinuities. □

COROLLARY 7.20. *(Kane[3] Criterion) For twice differentiable $f : I \to \mathbb{R}$, TFAE:*
(i) f is convex.
(ii) $f'' \geq 0$.

PROOF. By Theorem 7.17, f is convex iff f' is weakly increasing. Moreover, f' is weakly increasing iff $f'' = (f')' \geq 0$. We're done. □

[3]Andrew Kane was a student in the 2011-2012 course who suggested this criterion upon being prompted in class.

3.7. An extremal property of convex functions.

The following result secures the importance of convex functions in applied mathematics: for these functions, a critical point must be a global minimum.

THEOREM 7.21. *Let $f : I \to \mathbb{R}$ be a differentiable convex function. Let $c \in I$ be a stationary point: $f'(c) = 0$. Then:*
a) f attains its global minimum at c.
b) Unless f is constant on some nontrivial subinterval of I, f attains a strict global minimum *at c: for all $x \neq c$, $f(x) > f(c)$.*

PROOF. By Theorem 7.18, f' is weakly increasing.

a) Suppose $d > c$ is such that $f(d) < f(c)$. Then the secant line between $(c, f(c))$ and $(d, f(d))$ has negative slope, so by the Mean Value Theorem there is $z \in (c, d)$ such that $f'(z) < 0$. But then $c \leq z$ and $0 = f'(c) > f'(z)$, contradicting f' being weakly increasing. So $f(c) \leq f(x)$ for all $x \geq c$.

Similarly, suppose $b < c$ is such that $f(b) < f(c)$. Then the secant line between $(b, f(b))$ and $(c, f(c))$ has positive slope, so there is $w \in (b, c)$ such that $f'(w) > 0$. But then $b \leq c$ and $f'(b) > 0 = f'(c)$, contradicting f' being weakly incereasing. So $f(x) \geq f(c)$ for all $x \leq c$. That is, f attains a global minimum at c.

b) If $e \neq c$ is such that $f'(e) = 0$, then since f' is weakly increasing, $f' \equiv 0$ on the subinterval $|[c, e]|$ from c to e, and thus f is constant on $|[c, e]|$. □

Exercise: a) Let $f : (a, b) \to \mathbb{R}$ be a convex function such that $\lim_{x \to a^+} f(x) = \lim_{x \to b^-} f(x) = \infty$. Show that f assumes a global minimum.
b) State and prove an analogue of part a) with (a, b) replaced by $\mathbb{R} = (-\infty, \infty)$.
c) Exhibit a convex function $f : \mathbb{R} \to \mathbb{R}$ which *does not* assume a global minimum.

3.8. Supporting lines and differentiability.

Let $f : I \to \mathbb{R}$ be a function. A **supporting line** for f at $c \in I$ is a linear function $\ell : \mathbb{R} \to \mathbb{R}$ such that $f(c) = \ell(c)$ and $f(x) \geq \ell(x)$ for all $x \in I$.

Example: Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. Observe that the horizontal line $\ell = 0$ is a supporting line at $c = 0$: indeed $f(0) = 0$ and $f(x) \geq 0$ for all x. Notice that $\ell = 0$ is the tangent line to $y = f(x)$ at $c = 0$.

Example: More generally, let $A, B, C \in \mathbb{R}$ with $A \neq 0$. We claim that for all $c \in \mathbb{R}$ the tangent line $\ell_c(x)$ to the parabola $f(x) = Ax^2 + Bx + C$ is the unique line passing through $(c, f(c))$ such that $f(x) \neq \ell_c(x)$ for all $x \neq c$. To see this, consider the function $g(x) = f(x) - \ell_c(x)$. Then g is a quadratic polynomial with leading coefficient A and $g(c) = g'(c) = 0$, so $g(c) = A(x - c)^2$, and thus $f(x) - \ell_c(x) = g(x) \neq 0$ for $x \neq c$. On the other hand, let ℓ be any other line passing through $(c, f(c))$. Then $h(x) = f(x) - \ell(x)$ is a degree two polynomial with $h(c) = 0$. Moreover, since ℓ is not the tangent line at c, $h'(c) \neq 0$, and thus h has a simple root at c, i.e., $h(x) = A(x - c)j(x)$ with $j(x)$ a linear function with $j(c) \neq 0$. Therefore j has a root at some $d \neq c$ and that point d, $\ell(x) = f(x)$.

We claim that the tangent line $\ell_c(x)$ to $f(x) = Ax^2 + Bx + C$ is a supporting line iff $A > 0$. Indeed, on both the intervals $(-\infty, c)$ and (c, ∞) the continuous function $f(x) - \ell_c(x)$ is nonzero, so must have constant sign. But $\lim_{x \to \pm\infty} f(x) - \ell_c(x) = \infty$ if $A > 0$ and $-\infty$ if $A < 0$. It follows that if $A > 0$, $f(x) - \ell_c(x) > 0$ for all $x \neq c$

– so ℓ_c is a supporting line – and also that if $A < 0$, $f(x) - \ell_c(x) < 0$ for all x – so ℓ_c is not a supporting line. Note that since $f''(x) = 2A$, f is convex iff $A > 0$.

Example: Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = |x|$. Since $\mathrm{Epi}(f)$ is a convex subset of $\mathbb{R}^2$, f is convex. For every $c > 0$, the line $y = x$ is a supporting line, and for every $c < 0$, the line $y = -x$ is a supporting line, and in both cases these supporting lines are unique. For $c = 0$, $y = 0$ is a supporting line, but it is not the only one: indeed $y = mx$ is a supporting line at $c = 0$ iff $-1 \leq m \leq 1$. Note that the smallest slope of a supporting line is the left-hand derivative at zero:

$$f'_-(0) = \lim_{h \to 0^-} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0^-} \frac{-h}{h} = -1,$$

and the largest slope of a supporting line is the right-hand derivative at zero:

$$f'_+(0) = \lim_{h \to 0^+} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0^+} \frac{h}{h} = -1.$$

LEMMA 7.22. *Convex functions are closed under suprema. More precisely, if $\{f_i : I \to \mathbb{R}\}_{i \in I}$ is a family of convex functions, $f : I \to \mathbb{R}$ is a function such that for all $x \in I$, $f(x) = \sup_{i \in I} f_i(x)$, then f is convex.*

PROOF. Let $a < b \in I$ and $\lambda \in (0,1)$. Then

$$f((1-\lambda)a + \lambda b) = \sup_{i \in I} f_i((1-\lambda)a + \lambda b)$$

$$\leq (1-\lambda) \sup_{i \in i} f(a) + \lambda \sup_{i \in I} f(b) = (1-\lambda)f(a) + \lambda f(b).$$

□

THEOREM 7.23. *Let I be an open interval. For a function $f : I \to \mathbb{R}$, TFAE:*
(i) f is convex.
(ii) f admits a supporting line at each $c \in I$.

PROOF. (i) $\implies$ (ii): Neither property (i) or (ii) is disturbed by translating the coordinate axes, so we may assume that $c = 0$ and $f(0) = 0$. Let $\alpha \in I \setminus \{0\}$. For all $\lambda_1, \lambda_2 > 0$ such that $\lambda_1 \alpha, -\lambda_2 \alpha \in I$, by the secant-graph inequality we have

$$0 = (\lambda_1 + \lambda_2) f\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}(-\lambda_2 \alpha) + \frac{\lambda_2}{\lambda_1 + \lambda_2}(\lambda_1 \alpha)\right) \leq \lambda_1 f(-\lambda_2 \alpha) + \lambda_2 f(\lambda_1 \alpha),$$

or

$$\frac{-f(-\lambda_2 \alpha)}{\lambda_2} \leq \frac{f(\lambda_1 \alpha)}{\lambda_1}.$$

It follows that $\sup_{\lambda_2} \frac{-f(-\lambda_2\alpha)}{\lambda_2} \leq \inf_{\lambda_1} \frac{f(\lambda_1 \alpha)}{\lambda_1}$, so there is $m \in \mathbb{R}$ with

$$\frac{-f(-\lambda_2 \alpha)}{\lambda_2} \leq m \leq \frac{f(\lambda_1 \alpha)}{\lambda_1}.$$

Equivalently, $f(t\alpha) \geq mt$ for all $t \in \mathbb{R}$ such that $t\alpha \in I$. Thus $\ell(x) = mx$ is a supporting line for f at $c = 0$.
(ii) $\implies$ (i): For each $c \in I$, let $\ell_c : I \to \mathbb{R}$ be a supporting line for f at c. Since for all $x \in I$, $f(x) \geq \ell_c(x)$ for all c and $f(c) = \ell_c(c)$, we have $f(x) = \sup_{c \in I} \ell_c(x)$. Since the linear functions ℓ_c are certainly convex, f is the supremum of a family of convex functions, hence convex by Lemma 7.21 □

Before stating the next result, we recall the notion of *one-sided differentiability*: if f is a function defined (at least) on some interval $[c-\delta, c]$, we say f is **left differentiable at c** if $\lim_{x\to c^-} \frac{f(x)-f(c)}{x-c}$ exists, and if so, we denote this limit by $f'_-(c)$, the **left derivative of f at c**. Similarly, if f is defined (at least) on sme interval $[c, c+\delta)$, we say f is **right differentiable at c** if $\lim_{x\to c^+} \frac{f(x)-f(c)}{x-c}$ exists, and if so, we denote this limit by $f'_+(c)$, the **right derivative of f at c**. (As usual for limits, f is differentiable at c iff $f'_-(c)$ and $f'_+(c)$ both exist and are equal.)

THEOREM 7.24. *Let I be an interval, $c \in I^\circ$, and $f : I \to \mathbb{R}$ be convex.*
a) f is both left-differentiable and right-differentiable at c: $f'_-(c)$ and $f'_+(c)$ exist.
b) For all $c \in (a,b)$, $f'_-(c) \le f'_+(c)$.
c) $f'_-,\ f'_+ : I \to \mathbb{R}$ are both weakly increasing functions.
d) A line ℓ passing through $(c, f(c))$ is a supporting line for f iff its slope m satisfies

$$f'_-(c) \le m \le f'_+(c).$$

PROOF. We follow [**Go**, p. 153] and [**Gr**, pp. 8-10].
a), b) Define $\varphi : (a,b) \setminus \{c\} \to \mathbb{R}$ by

$$\varphi(x) = \frac{f(x)-f(c)}{x-c}.$$

Further, put

$$A = \varphi((a,c)), \quad B = \varphi(c,b).$$

From the three-secant inequality we immediately deduce all of the following: φ is weakly increasing on (a,c), φ is weakly increasing on (c,b), and $A \le B$. Thus

$$f'_-(c) = \lim_{x\to c^-} \varphi(x) = \sup A \le \inf B = \lim_{x\to c^+} \varphi(x) = f'_+(c).$$

c) Let $x_1, x_2 \in (a,b)$ with $x_1 < x_2$, and choose v with $x_1 < v < x_2$. Then by part b) and the three-secant inequality,

$$f'_-(x_1) \le f'_+(x_1) \le \frac{f(v)-f(x_1)}{v-x_1} \le \frac{f(v)-f(x_2)}{v-x_2} \le f'_-(x_2) \le f'_+(x_2).$$

d) The proof of parts a) and b) shows that for $x \in I$,

$$f(x) \ge f(c) + f'_-(c)(x-c), \text{ if } x \le c,$$

$$f(x) \ge f(c) + f'_+(c)(x-c), \text{ if } x \ge c.$$

Thus if $f'_-(c) \le m \le f'_+(c)$, we have

$$f(x) \ge f(c) + f'_-(c)(x-c) \ge f(c) + m(x-c), \text{ if } x \le c,$$

$$f(x) \ge f(c) + f'_+(c)(x-c) \ge m(x_c) \text{ if } x \ge c,$$

so $\ell(x) = f(c) + m(x-c)$ is a supporting line for f at c. That these are the only possible slopes of supporting lines for f at c is left as an exercise for the reader. □

Exercise: We supppose the hypotheses of Theorem 7.24, and for $m \in \mathbb{R}$, put $\ell(x) = f(c) + m(x-c)$.
a) Suppose that $m < f'_-(c)$. Show that there is $\delta > 0$ such that for all $x \in (c-\delta, c)$, $\ell(x) > f(x)$ and thus ℓ is not a supporting line for f at c.
b) Suppose that $m > f'_+(c)$. Show that there is $\delta > 0$ such that for all $x \in (c, c+\delta)$, $\ell(x) > f(x)$ and thus ℓ is not a supporting line for f at c.

Exercise: Assume the hypotheses of Theorem 7.24. Show TFAE:

(i) f'_- is continuous at c.
(ii) f'_+ is continuous at c.
(iii) f is differentiable at c, and the tangent line is a supporting line at c.
(iv) f has a unique supporting line at c.

Remark: Because a weakly increasing function can only have jump discontinuities, it can be shown that such functions are continuous at "most" points of their domain. For those who are familiar with the notions of countable and uncountable sets, we may be more precise: the set of discontinuities of a monotone function must be **countable**. Since the union of two countable sets is countable, it follows that a convex function is differentiable except possibly at a countable set of points.

Remark: One can go further: a convex function is twice differentiable at "most" points of its domain. The sense of "most" here is different (and weaker): the set of points at which f fails to be twice differentiable has **measure zero** in the sense of Lebesgue. This result, which is itself due to Lebesgue, lies considerably deeper than the one of the previous remark.

3.9. Jensen's Inequality.

THEOREM 7.25. *(Jensen's Inequality) Let $f : I \to \mathbb{R}$ be continuous and convex. For any $x_1, \ldots, x_n \in I$ and any $\lambda_1, \ldots, \lambda_n \in [0,1]$ with $\lambda_1 + \ldots + \lambda_n = 1$, we have*

$$f(\lambda_1 x_1 + \ldots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \ldots + \lambda_n f(x_n).$$

PROOF. We go by induction on n, the base case $n = 1$ being trivial. So suppose Jensen's Inequality holds for some $n \in \mathbb{Z}^+$, and consider $x_1, \ldots, x_{n+1} \in I$ and $\lambda_1, \ldots, \lambda_{n+1} \in [0,1]$ with $\lambda_1 + \ldots + \lambda_{n+1} = 1$. If $\lambda_{n+1} = 0$ we are reduced to the case of n variables which holds by induction. Similarly if $\lambda_{n+1} = 1$ then $\lambda_1 = \ldots = \lambda_n = 0$ and we have, trivially, equality. So we may assume $\lambda_{n+1} \in (0,1)$ and thus also that $1 - \lambda_{n+1} \in (0,1)$. Now for the big trick: we write

$$\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = (1 - \lambda_{n+1}) \left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} x_{n+1},$$

so that

$$f\left(\lambda_1 x_1 + \ldots + \lambda_n x_n\right) = f((1 - \lambda_{n+1})(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n) + \lambda_{n+1} x_{n+1})$$

$$\leq (1 - \lambda_{n+1}) f\left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} f(x_{n+1}).$$

Since $\frac{\lambda_1}{1-\lambda_{n+1}}, \ldots, \frac{\lambda_n}{1-\lambda_{n+1}}$ are non-negative numbers that sum to 1, by induction the n variable case of Jensen's Inequality can be applied to give that the above expression is less than or equal to

$$(1 - \lambda_{n+1}) \left(\frac{\lambda_1}{1 - \lambda_{n+1}} f(x_1) + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} f(x_n) \right) + \lambda_{n+1} f(x_{n+1})$$

$$= \lambda_1 f(x_1) + \ldots + \lambda_n f(x_n) + \lambda_{n+1} f(x_{n+1}).$$

□

3.10. Some applications of Jensen's Inequality.

Example: For any $p > 1$, the function $f : [0, \infty) \to \mathbb{R}$ by $x \mapsto x^p$ is twice differentiable on $(0, \infty)$ with $f''(x) > 0$ there, so by the Kane Criterion f is convex on $(0, \infty)$. Also f is continuous at 0, so f is convex on $[0, \infty)$.

Example: The function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto e^x$ has $f''(x) = e^x > 0$ for all x, so by the Kane Criterion f is convex on $\mathbb{R}$.

By plugging these convex functions into Jensen's Inequality and massaging what we get a bit, we will quickly deduce some very important inequalities.[4]

THEOREM 7.26. *(Weighted Arithmetic Geometric Mean Inequality) Let $x_1, \ldots, x_n \in [0, \infty)$ and $\lambda_1, \ldots, \lambda_n \in [0, 1]$ be such that $\lambda_1 + \ldots + \lambda_n = 1$. Then:*

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \ldots + \lambda_n x_n. \tag{32}$$

*Taking $\lambda_1 = \ldots = \lambda_n = \frac{1}{n}$, we get the **arithmetic geometric mean inequality**:*

$$(x_1 \cdots x_n)^{\frac{1}{n}} \leq \frac{x_1 + \ldots + x_n}{n}.$$

PROOF. We may assume $x_1, \ldots, x_n > 0$. For $1 \leq i \leq n$, put $y_i = \log x_i$. Then $x_1^{\lambda_1} \cdots x_n^{\lambda_n} = e^{\log(x_1^{\lambda_1} \cdots x_n^{\lambda_n})} = e^{\lambda_1 y_1 + \ldots + \lambda_n y_n} \leq \lambda_1 e^{y_1} + \ldots + \lambda_n e^{y_n} = \lambda_1 x_1 + \ldots + \lambda_n x_n$. □

THEOREM 7.27. *(Young's Inequality)*
Let $x, y \in [0, \infty)$ and let $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{33}$$

PROOF. When either $x = 0$ or $y = 0$ the left hand side is zero and the right hand side is non-negative, so the inequality holds and we may thus assume $x, y > 0$. Now apply the Weighted Arithmetic-Geometric Mean Inequality with $n = 2$, $x_1 = x^p$, $x_2 = y^q$, $\lambda_1 = \frac{1}{p}$, $\lambda_2 = \frac{1}{q}$. We get

$$xy = (x^p)^{\frac{1}{p}} (y^q)^{\frac{1}{q}} = x_1^{\lambda_1} x_2^{\lambda_2} \leq \lambda_1 x_1 + \lambda_2 x_2 = \frac{x^p}{p} + \frac{y^q}{q}.$$

□

THEOREM 7.28. *(Hölder's Inequality)*
Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$ and let $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$|x_1 y_1| + \ldots + |x_n y_n| \leq (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} (|y_1|^q + \ldots + |y_n|^q)^{\frac{1}{q}}. \tag{34}$$

PROOF. As above, the result is clear if either $x_1 = \ldots = x_n = 0$ or $y_1 = \ldots = y_n = 0$, so we may assume that neither of these is the case. For $1 \leq i \leq n$, apply Young's Inequality with

$$x = \frac{|x_i|}{(|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}}, y = \frac{|y_i|}{(|y_1|^q + \ldots + |y_n|^q)^{\frac{1}{q}}},$$

[4]Unfortunately we will not see here *why* these inequalities are important: you'll have to trust me on that for now. But just as soon as you take graduate level real analysis, you'll see these inequalities again.

and sum the resulting inequalities from $i = 1$ to n, getting

$$\frac{\sum_{i=1}^{n} |x_i y_i|}{(|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} (|y_1|^q + \ldots + |y_n|^q)^{\frac{1}{q}}} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

□

THEOREM 7.29. *(Minkowski's Inequality)*
For $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$ and $p \geq 1$:

$$(|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{p}} \leq (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} + (|y_1|^p + \ldots |y_n|^p)^{\frac{1}{p}} \tag{35}$$

PROOF. When $p = 1$, the inequality reads

$$|x_1 + y_1| + \ldots + |x_n + y_n| \leq |x_1| + |y_1| + \ldots + |x_n| + |y_n|$$

and this holds just by applying the triangle inequality: for all $1 \leq i \leq n$, $|x_i + y_i| \leq |x_i| + |y_i|$. So we may assume $p > 1$. Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$, and note that then $(p-1)q = p$. We have

$$|x_1 + y_1|^p + \ldots + |x_n + y_n|^p$$

$$\leq |x_1||x_1 + y_1|^{p-1} + \ldots + |x_n||x_n + y_n|^{p-1} + |y_1||x_1 + y_1|^{p-1} + \ldots + |y_n||x_n + y_n|^{p-1} \overset{\text{HI}}{\leq}$$

$$(|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{q}} + (|y_1|^p + \ldots + |y_n|^p)^{\frac{1}{p}} (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{q}}$$

$$= \left((|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} + (|y_1|^p + \ldots |y_n|^p)^{\frac{1}{p}} \right) (|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{q}}.$$

Dividing both sides by $(|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{q}}$ and using $1 - \frac{1}{q} = \frac{1}{p}$, we get the desired result. □

COROLLARY 7.30. *(Triangle Inequality in $\mathbb{R}^n$) For all $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$*

$$\sqrt{(x_1 + y_1)^2 + \ldots + (x_n + y_n)^2} \leq \sqrt{x_1^2 + \ldots + x_n^2} + \sqrt{y_1^2 + \ldots + y_n^2}.$$

CHAPTER 8

Integration

1. The Fundamental Theorem of Calculus

Having "finished" with continuity and differentiation, we turn to the third main theme of calculus: integration. The basic idea is this: for a function $f : [a, b] \to \mathbb{R}$, we wish to associate a number $\int_a^b f$, the **definite integral**. When f is non-negative, our intuition is that $\int_a^b f$ should represent the area under the curve $y = f(x)$, or more precisely the area of the region bounded above by $y = f(x)$, bounded below by $y = 0$, bounded on the left by $x = a$ and bounded on the right by $x = b$.

Unfortunately this is not yet a formal definition, because we do not have a formal definition of the area of a subset of the plane! In high school geometry one learns only about areas of very simple figures: polygons, circles and so forth. Dealing head-on with the task of assigning an area to every subset of $\mathbb{R}^2$ is quite difficult: it is one of the important topics of *graduate level* real analysis: **measure theory**.

So we need to back up a bit and give a definition of $\int_a^b f$. As you probably know, the general idea is to construe $\int_a^b f$ as the result of some kind of limiting process, wherein we divide $[a, b]$ into subintervals and take the sum of the areas of certain rectangles which approximate the function f at various points of the interval (**Riemann sums**). As usual in freshman calculus, reasonably careful definitions appear in the textbook somewhere, but with so little context and development that (almost) no actual freshman calculus student can really appreciate them.

But wait! Before plunging into the details of this limiting process, let's take a more **axiomatic approach**: given that we want $\int_a^b f$ to represent the area under $y = f(x)$, what properties should it satisfy? Here are some reasonable ones.

(I1) If $f = C$ is a constant function, then $\int_a^b C = C(b - a)$.
(I2) If $f_1(x) \leq f_2(x)$ for all $x \in [a, b]$, then $\int_a^b f_1 \leq \int_a^b f_2$.
(I3) If $a \leq c \leq b$, then $\int_a^b f = \int_a^c f + \int_b^c f$.

Exercise 1.1: Show (I1) implies: for any $f : [a, b] \to \mathbb{R}$ and any $c \in [a, b]$, $\int_c^c f = 0$.

It turns out that these three axioms already imply many of the other properties we want an integral to have. Even more, there is essentially only one way to define $\int_a^b f$ so as to satisfy (I1) through (I3).

Well, almost. One feature that we haven't explicitly addressed yet is this: for

which functions $f : [a, b] \to \mathbb{R}$ do we expect $\int_a^b f$ to be defined? *For all functions*?? A little thought shows this not to be plausible: there are some functions so pathological that there is no reason to believe that "the area under the curve $y = f(x)$" has any meaning whatsoever, and there are some functions for which this area concept seems meaningful but for which the area is *infinite*.

So it turns out to be useful to think of integration itself as a **real-valued function**, with domain some set of functions $\{f : [a, b] \to \mathbb{R}\}$. That is, for each $a \leq b$ we wish to have a set, say $\mathcal{R}[a, b]$, of **integrable functions** $f : [a, b] \to \mathbb{R}$ and for each $f \in \mathcal{R}[a, b]$, we wish to associate a real number $\int_a^b f$. As to exactly what this set $\mathcal{R}[a, b]$ of integrable functions should be, it turns out that we have some leeway, but to get a theory which is useful and not too complicated, let's assume the following:

(I0) For all real numbers $a < b$:
a) Every continuous $f : [a, b] \to \mathbb{R}$ lies in $\mathcal{R}[a, b]$.
b) Every function $f \in \mathcal{R}[a, b]$ is bounded.

By the Extreme Value Theorem, every continuous function $f : [a, b] \to \mathbb{R}$ is bounded. Thus the class $\mathcal{C}[a, b]$ of all continuous functions $f : [a, b] \to \mathbb{R}$ is contained in the class $\mathcal{B}[a, b]$ of all bounded functions $f : [a, b]$, and axiom (I0) requires that the set of integrable functions lies somewhere in between:

$$\mathcal{C}[a, b] \subseteq \mathcal{R}[a, b] \subseteq \mathcal{B}[a, b].$$

Let's recast the other three axioms in terms of our set $\mathcal{R}[a, b]$ of integrable functions:

(I1) If $f = C$ is constant, then $f \in \mathcal{R}[a, b]$ and $\int_a^b C = C(b - a)$.
(I2) If for $f_1, f_2 \in \mathcal{R}[a, b]$ we have $f_1(x) \leq f_2(x)$ for all $x \in [a, b]$, then $\int_a^b f_1 \leq \int_a^b f_2$.
(I3) Let $f : [a, b] \to \mathbb{R}$, and let $c \in (a, b)$. Then $f \in \mathcal{R}[a, b]$ iff $f \in \mathcal{R}[a, c]$ and $f \in \mathcal{R}[c, b]$. If these equivalent conditions hold, then $\int_a^b f = \int_a^c f + \int_c^b f$.

If this business of "integrable functions" seems abstruse, then on the first pass just imagine that $\mathcal{R}[a, b]$ is precisely the set of all continuous functions $f : [a, b] \to \mathbb{R}$.

Now we have the following extremely important result.

THEOREM 8.1. *(Fundamental Theorem of Calculus) Let $f \in \mathcal{R}[a, b]$ be any integrable function. For $x \in [a, b]$, define $\mathcal{F}(x) = \int_a^x f$. Then:*
a) The function $\mathcal{F} : [a, b] \to \mathbb{R}$ is continuous at every $c \in [a, b]$.
b) If f is continuous at $c \in [a, b]$, then $\mathcal{F}$ is differentiable at c, and $\mathcal{F}'(c) = f(c)$.
c) If f is continuous and F is any antiderivative of f – i.e., a function $F : [a, b] \to \mathbb{R}$ such that $F'(x) = f(x)$ for all $x \in [a, b]$, then $\int_a^b f = F(b) - F(a)$.

PROOF. By (I0), there exists $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in [a, b]$. If $M = 0$ then f is the constant function 0, and then it follows from (I1) that $\mathcal{F}$ is also the constant function zero, and one sees easily that the theorem holds in this case.

So we may assume $M > 0$. For all $\epsilon > 0$, we may take $\delta = \frac{\epsilon}{M}$. Indeed, by (I3)

$$\mathcal{F}(x) - \mathcal{F}(c) = \int_a^x f - \int_a^c f = \int_c^x f. \tag{36}$$

Moreover, let $a \leq A \leq B \leq b$. Then still $-M \leq f(x) \leq M$ for all $x \in [A, B]$, so by (I2) and then (I2) we have

$$-M(B-A) = \int_A^B (-M) \leq \int_A^B f \leq M(B-A),$$

and thus

$$|\int_A^B f| \leq M(B-A). \tag{37}$$

Now suppose $|x-c| < \delta = \frac{\epsilon}{M}$. Using (36) and then (37) with $A = c$, $B = x$, we get

$$|\mathcal{F}(x) - \mathcal{F}(c)| = |\int_c^x f| \leq M|x-c| < M\left(\frac{\epsilon}{M}\right) = \epsilon.$$

b) Suppose f is continuous at c. We wish to compute

$$\mathcal{F}'(x) = \lim_{x \to c} \frac{\mathcal{F}(x) - \mathcal{F}(c)}{x-c}.$$

Since f is continuous at c, for all $\epsilon > 0$, there exists $\delta > 0$ such that $|x-c| < \delta \implies |f(x) - f(c)| < \epsilon$, or equivalently

$$f(c) - \epsilon < f(x) < f(c) + \epsilon.$$

Therefore

$$f(c) - \epsilon = \frac{\int_c^x f(c) - \epsilon}{x-c} \leq \frac{\int_c^x f}{x-c} \leq \frac{\int_c^x f(c) + \epsilon}{x-c} = f(c) + \epsilon,$$

and thus

$$|\frac{\int_c^x f}{x-c} - f(c)| \leq \epsilon.$$

This shows that $\mathcal{F}'(c)$ exists and is equal to $f(c)$.
c) By part b), if f is continuous, $\mathcal{F}(x) = \int_a^x f$ is an antiderivative of f. But we have shown that if antiderivatives exist at all they are unique up to an additive constant. We have just found *an* antiderivative $\mathcal{F}$, so if F is any other antiderivative of f we must have $F(x) = \mathcal{F}(x) + C$ for some constant C, and then

$$F(b) - F(a) = (\mathcal{F}(b) + C) - (\mathcal{F}(a) + C) = \mathcal{F}(b) - \mathcal{F}(a) = \int_a^b f - \int_a^a f = \int_a^b f.$$

□

Remark: Although we introduced the integral "axiomatically", as long as we are only trying to integrate continuous functions we had no choice: the *only* way to assign a value $\int_a^b f$ to each continuous function $f : [a, b] \to \mathbb{R}$ satisfying the (reasonable!) axioms (I1) through (I3) is to take $\int_a^b f$ to be an antidervative F of f with $F(a) = 0$, and again, there is at most one such function.

These same considerations answer the conundrum of why the celebrated Theorem 8.1 has such a short and simple proof.[1] The theorem *assumes* that we already

[1]This is not just florid language. I taught second semester calculus four times as a graduate student and really did become puzzled at how easy it was to prove the Fundamental Theorem of Calculus so soon after integration is discussed. I worked out the answer while teaching an undergraduate real analysis course at McGill University in 2005. The current presentation is an adaptation of my lecture notes from this older course. Soon after I gave my 2005 lectures I found that a very similar "axiomatic" treatment of the integral was given by the eminent mathematician

have an integral, i.e., an assignment $(f : [a, b] \to \mathbb{R}) \mapsto \int_a^b f$ for every continuous function f. We have shown that there is *at most* one such integral on the continuous functions, but we have not yet constructed this integral! In other words, we have settled the problem of *uniqueness* of the definite integral but (thus far) assumed a solution to the much harder problem of *existence* of the definite integral. And again, this existence problem is equivalent to an existence problem that we mentioned before, namely that every continuous function has an antiderivative.

Thus: *if* we could prove by some other means that every continuous function f is the derivative of some other function F, then by the above we may simply *define* $\int_a^b f = F(b) - F(a)$. This is the approach that Newton himself took, although he didn't *prove* that every continuous function was a derivative but rather merely assumed it. It is also what freshman calculus students seem to think is taught in freshman calculus, namely that the definition of $\int_a^b f$ is $F(b) - F(a)$.[2]

But I do not know any way to prove that an arbitrary continuous function has an antiderivative *except* to give a constructive definition of $\int_a^b f$ as a limit of sums and then appeal to Theorem 8.1b) to get that $\int_a^x f$ is an antiderivative of f.

Thus Theorem 8.1 is "easy" because it diverts the hard work elsewhere: we need to give a constructive definition of the definite integral via a (new) kind of limiting process and then show "from scratch" that applied to every continuous $f : [a, b] \to \mathbb{R}$ this limiting process converges and results in a well-defined number $\int_a^b f$.

2. Building the Definite Integral

2.1. Upper and Lower Sums.

Now we begin the proof of the hard fact lurking underneath the Fundamental Theorem of Calculus: that we may define for every continuous function $f : [a, b] \to \mathbb{R}$ a number $\int_a^b f$ so as to satisfy (I1) through (I3) above. For now, we will make a simplifying assumption on our class of integrable functions: namely, let us only consider functions $f : [a, b] \to \mathbb{R}$ such that for every closed subinterval $[c, d] \subset [a, b]$, $f : [c, d] \to \mathbb{R}$ has a maximum and minimum value. Of course this holds for all continuous functions, so it will be a good start.

The basic idea is familiar from freshman calculus: we wish to subdivide our interval $[a, b]$ into a bunch of closed subintervals meeting only at the endpoints, and then we want to consider the **lower sum** and **upper sum** associated to f on each subinterval. Then the lower sum should be less than or equal to the "true area under the curve" which should be less than or equal to the upper sum, and by dividing $[a, b]$ into more and smaller subintervals we should get better and better approximations to the "true area under the curve", so we should define $\int_a^b f$ via some limiting process involving lower sums and upper sums.

Okay, let's do it!

Serge Lang in [**L**]. So the presentation that I give here is not being given by me for the first time and was not originated by me...but nevertheless the material is rarely presented this way.

[2]This is not what the books actually say, but what they actually say they don't say loudly enough in order for the point to really stick.

Step 1: We need the notion of a **partition** of an interval $[a, b]$: we choose finitely many "sample points" in $[a, b]$ and use them to divide $[a, b]$ into subintervals. Formally, a partitition $\mathcal{P}$ is given by a positive integer n and real numbers

$$a = a_0 \leq a_1 \leq \ldots \leq a_{n-1} \leq a_n = b.$$

That is, we require the "first sample point" a_0 to be the left endpoint of the interval, the "last sample point" a_n to be the right endpoint of the interval, and the other (distinct) points are absolutely arbitrary but written in increasing order.

Let $f : [a, b] \to \mathbb{R}$ be any function admitting a minimum and maximum value on every closed subinterval of $[a, b]$ (e.g. any continuous function!). For $0 \leq i \leq n-1$, let $m_i(f)$ denote the minimum value of f on the subinterval $[x_i, x_{i+1}]$ and let $M_i(f)$ denote the maximum value of f on the subinterval $[x_i, x_{i+1}]$. Then we define the **lower sum** associated to $f : [a, b] \to \mathbb{R}$ and the partition $\mathcal{P} = \{x_0, \ldots, x_n\}$ as

$$L(f, \mathcal{P}) = \sum_{i=0}^{n-1} m_i(f)(x_{i+1} - x_i)$$

and also the **upper sum** associated to f and $\mathcal{P}$ as

$$U(f, \mathcal{P}) = \sum_{i=0}^{n-1} M_i(f)(x_{i+1} - x_i).$$

These sums have a simple and important geometric interpretation: for any $0 \leq i \leq n-1$, the quantity $x_{i+1} - x_i$ is simply the length of the subinterval $[x_i, x_{i+1}]$. So consider the constant function $m_i(f)$ on the interval $[x_i, x_{i+1}]$: by definition of m_i, this is the largest constant function whose graph lies on or below the graph of f at every point of $[x_i, x_{i+1}]$. Therefore the quantity $m_i(f)(x_{i+1} - x_i)$ is simply the area of the rectangle with height $m_i(f)$ and width $x_{i+1} - x_i$, or equivalently the area under the constant function $y = m_i(f)$ on $[x_i, x_{i+1}]$.

We say the function $f : [a, b] \to \mathbb{R}$ is **integrable** if there is a *unique* $I \in \mathbb{R}$ such that for every partition $\mathcal{P}$ of $[a, b]$ we have

$$L(f, \mathcal{P}) \leq I \leq U(f, \mathcal{P}).$$

This definition, although correct, is not ideally formulated: it underplays the most important part – the *uniqueness* of I – while making it annoying to show the *existence* of I. (It turn outs that there is always *at least one* I lying between every lower sum and every upper sum, but this is as yet far from clear.) Here are some examples.

Example 2.1: If $f(x) \equiv C$ is a constant function, then for every partition $\mathcal{P}$ on $[a, b]$ we have $L(f, \mathcal{P}) = U(f, \mathcal{P}) = C(b - a)$. Thus the unique I in question is celarly $C(b - a)$: constant functions are integrable.

Example 2.2: Suppose $f(x)$ is constantly equal to 1 on the interval $[a, b]$ except for one interior point c, at which $f(c) = 0$. We claim that despite having a discontinuity at c, f is integrable, with $\int_a^b f = b - a$. To see this, first observe that for any partition $\mathcal{P}$ of $[a, b]$ we have $U(f, \mathcal{P}) = b - a$. Indeed this is because on every subinterval of $[a, b]$ f has 1 as its maximum value. On the other hand, for

any sufficiently small $\epsilon > 0$, we may choose a partition in which c occurs in exactly one subinterval (i.e., c is not one of the points of the partition). Then the lower sum on that subinterval is 0, whereas on every other subinterval the minimum is again 1, so $L(f, \mathcal{P}) = (b-a)(1-\epsilon)$. This shows that the unique number between every $L(f, \mathcal{P})$ and $U(f, \mathcal{P})$ is $b-a$, so $\int_a^b f = (b-a)$.

Exercise 2.3: Show that starting with the constant function C on $[a, b]$ and changing its value at finitely many points yields an integrable function f with $\int_a^b f = C(b-a)$.

The previous examples have the property that the upper sums $U(f, \mathcal{P})$ are constant. When this happens, one can show f is integrable by finding a sequence of partitions for which the lower sums approach this common value $U(f, \mathcal{P})$ which must then be the integral. But constancy of upper sums only occurs in trivial examples. For instance, suppose we want to show that $f(x) = x$ is integrable on $[0, 1]$. If we partition $[0, 1]$ into n equally spaced subintervals – let us call this partition $\mathcal{P}_n$ – then since f is increasing its minimum on each subinterval occurs at the left endpoint and its maximum on each subinterval occurs at the right endpoint. Thus

$$L(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} \left(\frac{i}{n}\right) \cdot \frac{1}{n} = \frac{1}{n^2} \sum_{i=0}^{n-1} i = \frac{1}{n^2} \cdot \left(\frac{(n-1)n}{2}\right) = 1 - \frac{1}{2n}.$$

and

$$U(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} \left(\frac{i+1}{n}\right) \cdot \frac{1}{n} = \frac{1}{n^2} \sum_{i=1}^{n} i = \frac{1}{n^2} \left(\frac{n(n+1)}{2}\right) = 1 + \frac{1}{2n}.$$

Since $\lim_{x\to\infty} \frac{x-1}{2x} = \lim_{x\to\infty} \frac{x+1}{2x} = \frac{1}{2}$, the upper and lower sums can both be made arbitrarily close to $\frac{1}{2}$ by taking n to be sufficiently large. Thus *if* $f(x) = x$ is integrable on $[0, 1]$, its integral must be $\frac{1}{2}$. Unfortunately we have not yet shown that f is integrable according to our definition: to do this we would have to consider an *arbitrary* partition $\mathcal{P}$ of $[0, 1]$ and show that $L(f, \mathcal{P}) \leq \frac{1}{2} \leq U(f, \mathcal{P})$. For this very simple function $f(x) = x$ it *is* possible to grind this out directly, but it's quite a bit of work. And that's just to find the area of a right triangle!

Example 2.4: Consider the function $f : [a, b] \to \mathbb{R}$ which is 1 at every irrational point and 0 at every rational point. Because every subinterval contains both rational and irrational numbers, for every partition $\mathcal{P}$ of $[a, b]$ we have $L(f, \mathcal{P}) = 0$ and $U(f, \mathcal{P}) = 1 \cdot (b-a) = b-a$. Assuming of course that $a < b$, this shows that f is **not** integrable: rather than the lower sums and the upper sums becoming arbitrarily close together, there is an unbridgeable gap between them.

The previous example perhaps suggests a solution to the problem. It did so to the late 19th century mathematician Jean-Gaston Darboux, who came up with an elegant way to see whether there is an unbridgeable gap between the lower and upper sums. Darboux's definition crucially and cleverly relies on the existence of suprema and infima for subsets of $\mathbb{R}$. In the next section we do things Darboux's way, which as we will see is much more pleasant than our current setup.

2.2. Darboux Integrability.

Let $f : [a,b] \to \mathbb{R}$ be *any* function. For a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ and $0 \leq i \leq n-1$, let $m_i(f)$ be the infimum of f on $[x_i, x_{i+1}]$ and $M_i(f)$ be the supreumum of f on $[x_i, x_{i+1}]$. Thus we have $m_i(f) \in [-\infty, \infty)$ and $M_i(f) \in (-\infty, \infty]$. As above, we define the lower and upper sums associated to $\mathcal{P}$:

$$L(f, \mathcal{P}) = \sum_{i=0}^{n-1} m_i(f)(x_{i+1} - x_i) \in [-\infty, \infty),$$

$$U(f, \mathcal{P}) = \sum_{i=0}^{n-1} M_i(f)(x_{i+1} - x_i) \in (-\infty, \infty].$$

For any f and $\mathcal{P}$ we have

$$L(f, \mathcal{P}) \leq U(f, \mathcal{P}). \tag{38}$$

Observe though that the lower sum could take the value $-\infty$ and the upper sum could take the value ∞. The following result clarifies when this is the case.

PROPOSITION 8.2. *Let $f : [a,b] \to \mathbb{R}$ be any function.*
a) The following are equivalent:
(i) For all partitions $\mathcal{P}$ of $[a,b]$, $L(f, \mathcal{P}) = -\infty$.
(ii) There exists a partition $\mathcal{P}$ of $[a,b]$ such that $L(f, \mathcal{P}) = -\infty$.
(iii) f is not *bounded below on $[a,b]$.*
b) The following are equivalent:
(i) For all partitions $\mathcal{P}$ of $[a,b]$, $U(f, \mathcal{P}) = \infty$.
(ii) There exists a partition $\mathcal{P}$ of $[a,b]$ such that $U(f, \mathcal{P}) = \infty$.
(iii) f is not *bounded above on $[a,b]$.*
c) The following are equivalent:
(i) For all partitions $\mathcal{P}$ of $[a,b]$, $L(f, \mathcal{P}) > -\infty$ and $U(f, \mathcal{P}) < \infty$.
(ii) f is bounded on $[a,b]$.

PROOF. a) (i) $\implies$ (ii) is immediate.
(ii) $\implies$ (iii): We prove the contrapositive: suppose that there is $m \in \mathbb{R}$ such that $m \leq f(x)$ for all $x \in [a,b]$. Then for all partitions $\mathcal{P} = \{a = x_0 < \ldots < x_{n-1} < x_n = b\}$ and all $0 \leq i \leq n-1$, we have $m_i(f) \geq m > -\infty$, so $L(f, \mathcal{P}) > -\infty$.
(iii) $\implies$ (i): Suppose f is not bounded below on $[a,b]$, and let $\mathcal{P} = \{a = x_0 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a,b]$. If $m_i(f) > -\infty$ for all $0 \leq i \leq n-1$, then $\min_{i=0}^{n-1} m_i(f)$ is a finite lower bound for f on $[a,b]$, contradicting our assumption. So there is at least one i such that $m_i(f) = -\infty$, which forces $L(f, \mathcal{P}) = -\infty$.
b) This is similar enough to part a) to be left to the reader.
c) If for all partitions $\mathcal{P}$, $L(f, \mathcal{P}) > -\infty$ and $U(f, \mathcal{P}) < \infty$, then by parts a) and b) f is bounded above and below on $[a,b]$, so is bounded on $[a,b]$. Conversely, if f is bounded on $[a,b]$ then it is bounded above and below on $[a,b]$, so by parts a) and b), for all partitions $\mathcal{P}$ we have $L(f, \mathcal{P}) > -\infty$ and $U(f, \mathcal{P}) < \infty$. □

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two partitions of $[a,b]$. We say that $\mathcal{P}_2$ **refines** $\mathcal{P}_1$ if $\mathcal{P}_2$ contains every point of $\mathcal{P}_1$: i.e., if $\mathcal{P}_1 \subset \mathcal{P}_2$.

LEMMA 8.3. *(Refinement Lemma) Let $\mathcal{P}_1 \subset \mathcal{P}_2$ be partitions of $[a,b]$ (i.e., $\mathcal{P}_2$ refines $\mathcal{P}_1$). Then for any bounded function $f : [a,b] \to \mathbb{R}$ we have*

$$L(f, \mathcal{P}_1) \leq L(f, \mathcal{P}_2) \leq U(f, \mathcal{P}_2) \leq U(f, \mathcal{P}_1).$$

PROOF. If $\mathcal{P}_2$ refines $\mathcal{P}_1$, $\mathcal{P}_2$ is obtained from $\mathcal{P}_1$ by a finite number of instances of the following: choose a subinterval $[c,d]$ of $\mathcal{P}_1$ and insert a new point e to subdivide it into two subintervals. So it is enough to address the case of a single interval. Let m and M be the infiumum and supremum of f on $[c,d]$. Futher, let m_1 and M_1 be the infimum and supremum of f on $[c,e]$, and m_2 and M_2 be the infimum and supremum of f on $[e,d]$. Then $m = \inf m_1, m_2$ and $M = \sup M_1, M_2$, so

$$m \leq m_1,\ m \leq m_2,\ M_1 \leq M,\ M_2 \leq M.$$

Then

$$m(d-c) = m(e-c) + m(d-e) \leq m_1(e-c) + m_2(e-c)$$

and

$$M(d-c) = M(e-c) + M(d-e) \geq M_1(e-c) + M_2(e-c).$$

This shows that subdivision causes the lower sum to stay the same or increase and the upper sum to stay the same or decrease. Since this holds each time we add a point to get from $\mathcal{P}_1$ to $\mathcal{P}_2$, we get $L(f,\mathcal{P}_1) \leq L(f,\mathcal{P}_2) \leq U(f,\mathcal{P}_2) \leq L(f,\mathcal{P}_2)$. □

LEMMA 8.4. *Let $f : [a,b] \to \mathbb{R}$ be a function, and let $\mathcal{P}_1, \mathcal{P}_2$ be partitions of $[a,b]$. Then $L(f,\mathcal{P}_1) \leq U(f,\mathcal{P}_2)$. That is,* any *lower sum associated to any partition is less than or equal to the upper sum associated to any other partition.*

PROOF. The idea here is simple but important: we choose a **common refinement** of $\mathcal{P}_1$ and $\mathcal{P}_2$, i.e., a partition which refines (contains) both $\mathcal{P}_1$ and $\mathcal{P}_2$. Any two partitions have infinitely many common refinements, but the most economical choice is simply the union of the two: put $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Then by Lemma 8.3 we have

$$L(f,\mathcal{P}_1) \leq L(f,\mathcal{P}) \leq U(f,\mathcal{P}) \leq U(f,\mathcal{P}_2).$$

□

Now we come to the crux of Darboux's theory of integrability: for *every* function $f : [a,b] \to \mathbb{R}$ we define the **lower integral** $\underline{int}_a^b f$ as the supremum of $L(f,\mathcal{P})$ as $\mathcal{P}$ ranges over all partitions of $[a,b]$ and the **upper integral** $\overline{\int_a^b} f$ as the infimum of $U(f,\mathcal{P})$ as $\mathcal{P}$ ranges over all partitions of $[a,b]$. Finally, we say that f is **Darboux integrable** if $\underline{\int_a^b} f = \overline{\int_a^b} f \in \mathbb{R}$, and we denote this common value by $\int_a^b f$.

LEMMA 8.5. *For any function $f : [a,b] \to \mathbb{R}$, we have*

$$\underline{\int_a^b} f \leq \overline{\int_a^b} f.$$

PROOF. Recall that if $X, Y \subset \mathbb{R}$ are such that $x \leq y$ for all $x \in X$ and all $y \in Y$, then $\sup X \leq \inf Y$. Now, by Lemma 8.4, for any partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ we have $L(f,\mathcal{P}_1) \leq U(f,\mathcal{P}_2)$. Therefore

$$\underline{\int_a^b} f = \sup_{\mathcal{P}_1} L(f,\mathcal{P}_1) \leq \inf_{\mathcal{P}_2} U(f,\mathcal{P}_2) = \overline{\int_a^b} f.$$

□

PROPOSITION 8.6. *Let $f : [a,b] \to \mathbb{R}$ be any function.*
a) We have $\underline{\int_a^b} f = -\infty$ iff f is unbounded below.
b) We have $\overline{\int_a^b} f = \infty$ iff f is unbounded above.
c) Therefore $\underline{\int_a^b} f, \overline{\int_a^b} f \in \mathbb{R}$ iff f is bounded. In particular, if f is Darboux integrable, then it is bounded.

PROOF. a) If f is unbounded below, $\underline{\int_a^b} f = \sup_{\mathcal{P}} L(f,\mathcal{P}) = \sup_{\mathcal{P}} -\infty = -\infty$. If f is bounded below, then for all $\mathcal{P}$, $L(f,\mathcal{P}) \in \mathbb{R}$ and thus $\sup_{\mathcal{P}} L(f,\mathcal{P}) > -\infty$.
b) This is very similar to part a) and is left to the reader.
c) This follows immediately from parts a) and b). □

In particular the class of Darboux integrable functions satisfies axiom (IOb).

Exercise 2.5: Is there a function $f : [a,b] \to \mathbb{R}$ with $\underline{\int_a^b} f = \overline{\int_a^b} f = \infty$?

Finally, here is the result we really want.

THEOREM 8.7. *(Darboux's Integrability Criterion)*
For a bounded function $f : [a,b] \to \mathbb{R}$, the following are equivalent:
(i) f is Darboux integrable.
(ii) For all $\epsilon > 0$ there exists a partition $\mathcal{P}$ of $[a,b]$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$.
(iii) There is exactly one real number I such that for all partitions $\mathcal{P}$ of $[a,b]$, we have $L(f,\mathcal{P}) \leq I \leq U(f,\mathcal{P})$.

PROOF. (i) $\implies$ (ii): Fix $\epsilon > 0$. Since $\int_a^b f = \underline{\int_a^b} f = \sup_{\mathcal{P}} L(f,\mathcal{P})$, there exists $\mathcal{P}_1$ with $L(f,\mathcal{P}_1) > \int_a^b - \frac{\epsilon}{2}$. Similarly, since $\int_a^b f = \overline{\int_a^b} f = \inf_{\mathcal{P}} U(f,\mathcal{P})$, there exists $\mathcal{P}_2$ with $U(f,\mathcal{P}_2) < \int_a^b + \frac{\epsilon}{2}$. Let $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Since $\mathcal{P}$ is a common refinement of $\mathcal{P}_1$ and $\mathcal{P}_2$ we have

$$U(f,\mathcal{P}) \leq U(f,\mathcal{P}_2) < \int_a^b f + \frac{\epsilon}{2} \tag{39}$$

and also

$$L(f,\mathcal{P}) \geq L(f,\mathcal{P}_1) > \int_a^b f - \frac{\epsilon}{2},$$

and thus

$$-L(f,\mathcal{P}) < \frac{\epsilon}{2} - \int_a^b f. \tag{40}$$

Adding (39) and (40) gives

$$U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon.$$

(ii) $\implies$ (i): By assumption, for every $\epsilon > 0$, there exists a partition $\mathcal{P}$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$. But by definition, we have $\overline{\int_a^b} f \leq U(f,\mathcal{P})$ and $\underline{\int_a^b} \geq L(f,\mathcal{P})$ and thus also $-\underline{\int_a^b} f \leq -L(f,\mathcal{P})$. Adding these two inequalities gives

$$\overline{\int_a^b} f - \underline{\int_a^b} f \leq U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon.$$

Since this holds for all $\epsilon > 0$, we have $\overline{\int}_a^b f \leq \underline{\int}_a^b f$. On the other hand, by Lemma 8.5 we have $\underline{\int}_a^b f \leq \overline{\int}_a^b f$, so $\underline{\int}_a^b f = \overline{\int}_a^b f \in \mathbb{R}$ and thus f is Darboux integrable.
(i) $\implies$ (iii): Suppose f is Darboux integrable, so $\int_a^b = \underline{\int}_a^b f = \overline{\int}_a^b f \in \mathbb{R}$. Then for all partitions $\mathcal{P}$ we have

$$L(f, \mathcal{P}) \leq \underline{\int}_a^b f = \int_a^b f \leq \overline{\int}_a^b f \leq U(f, \mathcal{P}).$$

Moreover, suppose $I < \int_a^b f = \underline{\int}_a^b f$. Then I is less than the supremum of the lower sums, so there exists a partition $\mathcal{P}$ with $I < L(f, \mathcal{P})$. Similarly, if $I > \int_a^b f = \overline{\int}_a^b f$, then I is greater than the infimum of the upper sums, so there exists a partition $\mathcal{P}$ with $U(f, \mathcal{P}) < I$. This shows that $\int_a^b f$ is the unique real number which lies in between every lower sum and every upper sum.
(iii) $\implies$ (i): We prove the contrapositive. Suppose that f is *not* Darboux integrable. Then for any partition $\mathcal{P}$ we have

$$L(f, \mathcal{P}) \leq \underline{\int}_a^b f < \overline{\int}_a^b f \leq U(f, \mathcal{P}),$$

and thus every $I \in [\underline{\int}_a^b f, \overline{\int}_a^b f]$ lies between every upper sum and every lower sum. □

2.3. Verification of the Axioms.

Let $\mathcal{R}([a, b])$ denote the set of Darboux integrable functions on $[a, b]$. We now tie together the work of the previous two sections by showing that the assignment $f \in \mathcal{R}([a, b]) \mapsto \int_a^b f$ satisfies the axioms (I0) through (I3) introduced in §1. In particular, this shores up the foundations of the Fundamental Theorem of Calculus and completes the proof that every continuous $f : [a, b] \to \mathbb{R}$ has an antiderivative.

In summary, we wish to prove the following result.

THEOREM 8.8. *(Main Theorem on Integration)*
a) Every continuous function $f : [a, b] \to \mathbb{R}$ is Darboux integrable.
b) The operation which assigns to every Darboux integrable function $f : [a, b] \to \mathbb{R}$ the number $\int_a^b f$ satisfies axioms (I0) through (I3) above.
c) Thus the Fundamental Theorem of Calculus holds for the Darboux integral. In particular, for every continuous function f, $F(x) = \int_a^x f$ is an antiderivative of f.

PROOF. a) Let $f : [a, b] \to \mathbb{R}$ be continuous. The key is that f is **uniformly continuous**, so for all $\epsilon > 0$, there is $\delta > 0$ such that for all $x_1, x_2 \in [a, b]$, $|x_1 - x_2| < \delta \implies |f(x_1) - f(x_2)| < \frac{\epsilon}{b-a}$. Let $n \in \mathbb{Z}^+$ be such that $\frac{b-a}{n} < \delta$, and let $\mathcal{P}_n$ be the partition of $[a, b]$ into n subintervals of equal length $\frac{b-a}{n}$. Then
(41)

$$U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} (M_i(f) - m_i(f)) \left(\frac{b-a}{n}\right) \leq \left(\frac{b-a}{n}\right) \sum_{i=0}^{n-1} M_i(f) - m_i(f).$$

Now for all $0 \leq i < n-1$, $m_i(f) = f(c_i)$ and $M_i(f) = f(d_i)$ for some $c_i, d_i \in [x_i, x_{i+1}]$. Thus $|c_i - d_i| \leq x_{i+1} - x_i = \frac{b-a}{n} < \delta$, so

$$|M_i(f) - m_i(f)| = |f(d_i) - f(c_i)| < \frac{\epsilon}{b-a}. \tag{42}$$

Combining (41) and (42) gives

$$U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n) \leq \left(\frac{b-a}{n}\right) \sum_{i=0}^{n-1} (M_i(f) - m_i(f)) \leq \left(\frac{b-a}{n}\right) \sum_{i=0}^{n-1} \frac{\epsilon}{b-a} = \epsilon.$$

b) (I0): By part a), every continuous function $f : [a, b] \to \mathbb{R}$ is Darboux integrable. By Proposition 8.6, every Darboux integrable function on $[a, b]$ is bounded. (I1): In Example 2.1, we showed that the constant function C is integrable on $[a, b]$ with $\int_a^b C = C(b-a)$. (I2): If $f_1, f_2 : [a, b] \to \mathbb{R}$ are both Darboux integrable and such that $f_1(x) \leq f_2(x)$ for all $x \in [a, b]$, then for every partition $\mathcal{P}$ of $[a, b]$ we have $L(f_1, \mathcal{P}) \leq L(f_2, \mathcal{P})$, and thus

$$\int_a^b f_1 = \sup_{\mathcal{P}} L(f_1, \mathcal{P}) \leq \sup_{\mathcal{P}} L(f_2, \mathcal{P}) = \int_a^b f_2.$$

(I3): Let $f : [a, b] \to \mathbb{R}$, and let $c \in (a, b)$. Suppose first that $f : [a, b] \to \mathbb{R}$ is Darboux integrable: thus, for all $\epsilon > 0$, there exists a partition $\mathcal{P}$ of $[a, b]$ with $U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon$. Let $\mathcal{P}_c = \mathcal{P} \cup \{c\}$. By the Refinement Lemma,

$$L(f, \mathcal{P}) \leq L(f, \mathcal{P}_c) \leq U(f, \mathcal{P}_c) \leq U(f, \mathcal{P}),$$

so $U(f, \mathcal{P}_c) - L(f, \mathcal{P}_c) \leq U(f, \mathcal{P}) - L(f, \mathcal{P}) < \epsilon$. Let $\mathcal{P}_1 = \mathcal{P}_c \cap [a, c]$ and $\mathcal{P}_2 = \mathcal{P}_c \cap [c, b]$. Then

$$L(f, \mathcal{P}_c) = L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2),\ U(f, \mathcal{P}_c) = U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2),$$

and therefore

$$\begin{aligned}(U(f, \mathcal{P}_1) - L(f, \mathcal{P}_1)) + (U(f, \mathcal{P}_2) - L(f, \mathcal{P}_1)) &= (U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2)) - (L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2)) \\ &= U(f, \mathcal{P}_c) - L(f, \mathcal{P}_c) < \epsilon,\end{aligned}$$

so by Darboux's criterion $f : [a, c] \to \mathbb{R}$ and $f : [c, b] \to \mathbb{R}$ are Darboux integrable. Conversely, suppose $f : [a, c] \to \mathbb{R}$ and $f : [c, b] \to \mathbb{R}$ are Darboux integrable; let $\epsilon > 0$. By Darboux's criterion, there is a partition $\mathcal{P}_1$ of $[a, c]$ such that

$$U(f, \mathcal{P}_1) - L(f, \mathcal{P}_2) < \frac{\epsilon}{2}$$

and a partition $\mathcal{P}_2$ of $[c, b]$ such that

$$U(f, \mathcal{P}_2) - L(f, \mathcal{P}_2) < \frac{\epsilon}{2}.$$

Then $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$ is a partition of $[a, b]$, and

$$\begin{aligned}U(f, \mathcal{P}) - L(f, \mathcal{P}) &= U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2) - (L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2)) \\ &= (U(f, \mathcal{P}_1) - L(f, \mathcal{P}_1)) + (U(f, \mathcal{P}_2) - L(f, \mathcal{P}_2)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

As for the value of the integral: fix $\epsilon > 0$. Let $\mathcal{P}$ be any partition of $[a, b]$, $\mathcal{P}_c = \mathcal{P} \cup \{c\}$, $\mathcal{P}_1 = \mathcal{P}_c \cap [a, c]$, $\mathcal{P}_2 = \mathcal{P}_c \cap [c, b]$. Then

$$\begin{aligned}L(f, \mathcal{P}) \leq L(f, \mathcal{P}_c) = L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2) \leq \int_a^c f + \int_c^b f \leq U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2) \\ = U(f, \mathcal{P}_c) \leq U(f, \mathcal{P}).\end{aligned}$$

Thus $\int_a^c f + \int_c^b f$ is a real number lying in between $L(f,\mathcal{P})$ and $U(f,\mathcal{P})$ for every partition $\mathcal{P}$ of $[a,b]$, so by Theorem 8.7 $\int_a^c f + \int_c^b f = \int_a^b f$.
c) This is immediate from Theorem 8.1 (the Fundamental Theorem of Calculus!). □

2.4. An Inductive Proof of the Integrability of Continuous Functions.
In this section we will give a proof of the Darboux integrability of an arbitrary continuous function $f : [a,b] \to \mathbb{R}$ which avoids the rather technical Uniform Continuity Theorem. We should say that we got the idea for doing this from Spivak's text, which first proves the integrability using uniform continuity as we did above and then later goes back to give a direct proof.

THEOREM 8.9. *Let $f : [a,b] \to \mathbb{R}$ be a continuous function on a closed bounded interval. Then f is Darboux integrable.*

PROOF. By Darboux's Criterion, it suffices to show that for all $\epsilon > 0$, there is a partition $\mathcal{P}$ of $[a,b]$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < \epsilon$. It is convenient to prove the following slightly different (but logically equivalent!) statement: for every $\epsilon > 0$, there exists a partion $\mathcal{P}$ of $[a,b]$ such that $U(f,\mathcal{P}) - L(f,\mathcal{P}) < (b-a)\epsilon$.

Fix $\epsilon > 0$, and let $S(\epsilon)$ be the set of $x \in [a,b]$ such that there exists a partition $\mathcal{P}_x$ of $[a,b]$ with $U(f,\mathcal{P}_x) - L(f,\mathcal{P}_x) < \epsilon$. We want to show $b \in S(\epsilon)$; our strategy will be to show $S(\epsilon) = [a,b]$ by Real Induction.
(RI1) The only partition of $[a,a]$ is $\mathcal{P}_a = \{a\}$, and for this partition we have $U(f,\mathcal{P}_a) = L(f,\mathcal{P}_a) = f(a) \cdot 0 = 0$, so $U(f,\mathcal{P}_a) - L(f,\mathcal{P}_a) = 0 < \epsilon$.
(RI2) Suppose that for $x \in [a,b)$ we have $[a,x] \subset S(\epsilon)$. We must show that there is $\delta > 0$ such that $[a,x+\delta] \subset S(\epsilon)$, and by the above observation it is enough to find $\delta > 0$ such that $x + \delta \in S(\epsilon)$: we must find a partition $\mathcal{P}_{x+\delta}$ of $[a,x+\delta]$ such that $U(f,\mathcal{P}_{x+\delta}) - L(f,\mathcal{P}_{x+\delta}) < (x+\delta-a)\epsilon)$. Since $x \in S(\epsilon)$, there is a partition $\mathcal{P}_x$ of $[a,x]$ with $U(f,\mathcal{P}_x) - L(f,\mathcal{P}_x) < (x-a)\epsilon$. Since f is continuous at x, we can make the difference between the maximum value and the minimum value of f as small as we want by taking a sufficiently small interval around x: i.e., there is $\delta > 0$ such that $\max(f,[x,x+\delta]) - \min(f,[x,x+\delta]) < \epsilon$. Now take the smallest partition of $[x,x+\delta]$, namely $\mathcal{P}' = \{x, x+\delta\}$. Then $U(f,\mathcal{P}') - L(f,\mathcal{P}') = (x+\delta-x)(\max(f,[x,x+\delta]) - \min(f,[x,x+\delta])) < \delta\epsilon$. Thus if we put $\mathcal{P}_{x+\delta} = \mathcal{P}_x + \mathcal{P}'$ and use the fact that upper / lower sums add when split into subintervals, we have

$$U(f,\mathcal{P}_{x+\delta}) - L(f,\mathcal{P}_{x+\delta}) = U(f,\mathcal{P}_x) + U(f,\mathcal{P}') - L(f,\mathcal{P}_x) - L(f,\mathcal{P}')$$
$$= U(f,\mathcal{P}_x) - L(f,\mathcal{P}_x) + U(f,\mathcal{P}') - L(f,\mathcal{P}') < (x-a)\epsilon + \delta\epsilon = (x+\delta-a)\epsilon.$$

(RI3) Suppose that for $x \in (a,b]$ we have $[a,x) \subset S(\epsilon)$. We must show that $x \in S(\epsilon)$. The argument for this is the same as for (RI2) except we use the interval $[x-\delta,x]$ instead of $[x,x+\delta]$. Indeed: since f is continuous at x, there exists $\delta > 0$ such that $\max(f,[x-\delta,x]) - \min(f,[x-\delta,x]) < \epsilon$. Since $x - \delta < x$, $x - \delta \in S(\epsilon)$ and thus there exists a partition $\mathcal{P}_{x-\delta}$ of $[a,x-\delta]$ such that $U(f,\mathcal{P}_{x-\delta}) = L(f,\mathcal{P}_{x-\delta}) = (x-\delta-a)\epsilon$. Let $\mathcal{P}' = \{x-\delta, x\}$ and let $\mathcal{P}_x = \mathcal{P}_{x-\delta} \cup \mathcal{P}'$. Then

$$U(f,\mathcal{P}_x) - L(f,\mathcal{P}_x) = U(f,\mathcal{P}_{x-\delta}) + U(f,\mathcal{P}') - (L(f,\mathcal{P}_{x-\delta}) + L(f,\mathcal{P}'))$$
$$= (U(f,\mathcal{P}_{x-\delta}) - L(f,\mathcal{P}_{x-\delta})) + \delta(\max(f,[x-\delta,x]) - \min(f,[x-\delta,x]))$$
$$< (x-\delta-a)\epsilon + \delta\epsilon = (x-a)\epsilon.$$

□

Exercise 2.6: Show that if $x \in S(\epsilon)$ and $a \leq y \leq x$, then also $y \in S(\epsilon)$.

Spivak [**S**, pp. 292-293] gives a different uniform continuity-free proof of Theorem 8.9: he establishes equality of the upper and lower integrals by differentiation. This sort of proof goes back at least to M.J. Norris [**No52**].

3. Further Results on Integration

3.1. The oscillation.

Let $f : D \subset \mathbb{R} \to \mathbb{R}$, and let I be an interval contained in the domain D of f. We define the **oscillation of f on I** as

$$\omega(f, I) = \sup(f, I) - \inf(f, I).$$

Note that $\omega(f, I)$ is in general an extended real number; it is an honest real number iff f is bounded on I (which will almost always be the case for us).

If $J \subset I \subset D$, then $\inf(f, J) \geq \inf(f, I)$ and $\sup(f, J) \leq \sup(f, I)$, and thus

$$\omega(f, J) \leq \omega(f, I). \tag{43}$$

Suppose now that c is a point in the interior of the domain D of f. We define the **oscillation of f at c** to be

$$\omega(f, c) = \lim_{\delta \to 0^+} \omega(f, [c - \delta, c + \delta]).$$

In other words, we are considering the oscillation of f on smaller and smaller intervals centered around c and taking the limit as δ approaches zero. Because of (43) the function $\delta \mapsto \omega(f, [c-\delta, c+\delta])$ is an increasing function of δ, so the limit as δ approaches zero from the right exists as an element of $[0, \infty]$ and is simply equal to the infimum. What's the point? This:

PROPOSITION 8.10. *Let I be an interval, $f : I \to \mathbb{R}$ be a function, and c an interior point of I. The following are equivalent:*
(i) $\omega(f, c) = 0$.
(ii) f is continuous at c.

PROOF. (i) $\implies$ (ii): If $\omega(f, c) = 0$, then for all $\epsilon > 0$, there exists $\delta > 0$ such that $\omega(f, [c - \delta, c + \delta]) = \sup(f, [c - \delta, c + \delta]) - \inf(f, [c - \delta, c + \delta]) < \epsilon$, i.e., there exists $\delta > 0$ such that for all x with $|x - c| \leq \delta$, $|f(x) - f(c)| < \epsilon$. So f is continuous at c.
(ii) $\implies$ (i): This is almost exactly the same. If f is continuous at c, then for all $\epsilon > 0$ there exists $\delta > 0$ such that for all x with $|x - c| \leq \delta$, $|f(x) - f(c)| < \epsilon$, and then $\sup(f, [c-\delta, c+\delta]) \leq f(c)+\epsilon$, $\inf(f, [c-\delta, c+\delta]) \geq f(c)-\epsilon$, so $\omega(f, [c-\delta, c+\delta]) \leq 2\epsilon$, so $\omega(f, c) = \lim_{\delta \to 0^+} \omega(f, [c - \delta, c + \delta]) = 0$. □

Remark: If $f : I \to \mathbb{R}$ and c is an endpoint of I, we can still define the oscillation $\omega(f, c)$, just by taking suitable half-intervals: e.g. if c is the left endpoint we put $\omega(f, c) = \lim_{\delta \to 0^+} \omega(f, [c, c + \delta])$. With this definition and our usual standard conventions about continuity at an endpoint of an interval, Proposition 8.10 remains true even if c is an endpoint of the interval I.

Now let $f : [a,b] \to \mathbb{R}$ and let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a,b]$. We define[3]

$$\Delta(f,\mathcal{P}) = \sum_{i=0}^{n-1} \omega(f,[x_i,x_{i+1}])(x_{i+1}-x_i) = U(f,\mathcal{P}) - L(f,\mathcal{P}).$$

Thus this notation is just a way of abbreviating the quantities "upper sum minus lower sum" which will appear ubiquitously in the near future. We can restate Darboux's Criterion especially cleanly with this new notation: a function $f : [a,b] \to \mathbb{R}$ is integrable iff for all $\epsilon > 0$, there exists a partition $\mathcal{P}$ of $[a,b]$ with $\Delta(f,\mathcal{P}) < \epsilon$.

3.2. Discontinuities of Darboux Integrable Functions.

At this point, I want to discuss the result that a bounded function $f : [a,b] \to \mathbb{R}$ with only finitely many discontinuities is Darboux integrable. So I wrote up a "direct" proof of this and it was long and messy. Afterwards I realized that a better argument is by induction on the number of discontinuities. One then has to prove the result for a function with a single discontinuity (base case), and assuming the result for every function with n discontinuities, prove it for every function with $n+1$ discontinutities (inductive step). Here the inductive step is especially easy: if $f : [a,b] \to \mathbb{R}$ has $n+1$ points of discontinuity, we can choose $c \in (a,b)$ such that $f|_{[a,c]}$ has exactly one discontinuity and $f|_{[c,b]}$ has exactly n discontinuities. The restricted functions are Darboux integrable by the base case and the induction hypothesis, and as we know, this implies that $f : [a,b] \to \mathbb{R}$ is Darboux integrable.

So really it is enough to treat the case of a bounded function with a single discontinuity. It turns out that it is no trouble to prove a stronger version of this.

THEOREM 8.11. *Let $f : [a,b] \to \mathbb{R}$ be bounded. Suppose that for all $c \in (a,b)$, $f|_{[c,b]} : [c,b] \to \mathbb{R}$ is Darboux integrable. Then f is Darboux integrable and*

$$\lim_{c \to a^+} \int_c^b f = \int_a^b f.$$

PROOF. Let $M > 0$ be such that $|f(x)| \leq M$ for all $x \in [a,b]$. Fix $\delta > 0$ and consider partitions $\mathcal{P}$ of $[a,b]$ with $x_1 = a + \delta$. For such partitions,

$$\Delta(f,\mathcal{P}) = \Delta(f,\mathcal{P} \cap [a,a+\delta]) + \Delta(f,\mathcal{P} \cap [a+\delta,b]).$$

Since the infimum of f on any subinterval of $[a,b]$ is at least $-M$ and the supremum is at most M, $\Delta(f,[a,a+\delta]) \leq 2M\delta$, which we can make as small as we wish by taking δ small enough. Similarly, having chosen δ, we may make $\Delta(f,\mathcal{P} \cap [a+\delta,b])$ as small as we like with a suitable choice of $\mathcal{P}$, since f is assumed to be Darboux integrable on $[a+\delta,b]$. Thus we can make the oscillation at most ϵ for any $\epsilon > 0$, so f is Darboux integrable on $[a,b]$. The second statement follows easily:

$$|\int_a^b f - \int_c^b f| = |\int_a^c f| \leq 2M(c-a),$$

and the last quantity goes to zero as $c \to a^+$. □

[3]For once we do not introduce a name but only a piece of notation. In an earlier course on this subject I called this quantity "the oscillation of f on $\mathcal{P}$", but this is not especially apt. Better perhaps would be to call $\Delta(f,\mathcal{P})$ the **discrepancy** of f and $\mathcal{P}$, since it is the difference between the upper and the lower sum. But in fact it is simplest not to call it anything but $\Delta(f,\mathcal{P})$!

Of course there is an analogous result with the roles of a and b reversed.

This result is also telling us that under certain situations we need not bother to consider "improper integrals": the improper integral will exist iff the conventional Darboux integral exists. This will make a lot more sense in the context of a discussion on improper integrals, so we defer the point until then.

THEOREM 8.12. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function which is continuous except at a finite set of points in its domain. Then f is Darboux integrable.*

Exercise 3.1: Use Theorem 8.11 (and its reflected version) to prove Corollary 8.12.

THEOREM 8.13. *A monotone function $f : [a, b] \to \mathbb{R}$ is Darboux integrable.*

PROOF. By reflection it suffices to deal with the case of a weakly increasing case. For such functions a miracle occurs: for every partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$ and for all $0 \leq i \leq n-1$, the infimum of f on $[x_i, x_{i+1}]$ is attained at the left endpoint x_i and the supremum of f on $[x_i, x_{i+1}]$ is attained at the right endpoint x_{i+1}. Therefore

$$L(f, \mathcal{P}) = f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + \ldots + f(x_{n-1})(x_n - x_{n-1}),$$
$$U(f, \mathcal{P}) = f(x_1)(x_1 - x_0) + f(x_2)(x_2 - x_1) + \ldots + f(x_n)(x_n - x_{n-1}).$$

Things simplify further if we simply take $\mathcal{P}_n$ to be the uniform partition of $[a, b]$ into n equal parts. Then

$$L(f, \mathcal{P}_n) = \frac{b-a}{n} \left(f(x_0) + \ldots + f(x_{n-1}) \right),$$
$$U(f, \mathcal{P}_n) = \frac{b-a}{n} \left(f(x_1) + \ldots + f(x_n) \right),$$

so

$$U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n) = \left(\frac{b-a}{n} \right) \left(f(b) - f(a) \right).$$

Thus taking n sufficiently large we may take $U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n)$ arbitrarily small, so f is Darboux integrable by Darboux's criterion. □

Theorem 8.13 goes beyond Theorem 8.12: there are increasing functions which are discontinuous at *infinitely many points*. To construct such things becomes much easier with some knowledge of infinite sequences and series, so we defer this discussion until later, except to make the following advertisement: for any real numbers $a < b$ and any injective function $s : \mathbb{Z}^+ \to [a, b]$, there exists an increasing function $f : [a, b] \to \mathbb{R}$ which is discontinuous at precisely the points $s(n) \in [a, b]$.

Example 3.2 (Thomae's Function): Let $f : [0, 1] \to \mathbb{R}$ be defined by $f(0) = 0$, $f(\frac{p}{q}) = \frac{1}{q}$, and f of any irrational number is zero. Then f is continuous at 0 and at all irrational numbers but discontinuous at every rational number. Thus not only does f have infinitely many points of discontinuity in $[0, 1]$, they are *dense*: any nontrivial subinterval contains at least one point of discontinuity. We claim that nevertheless f is Darboux integrable and $\int_0^1 f = 0$. First observe that since every subinterval $[x_i, x_{i+1}]$ contains an irrational number, the infimum of f on $[x_i, x_{i+1}]$ is zero, so for any partition $\mathcal{P}$ the lower sum is $L(f, \mathcal{P}) = 0$. It follows then that $\underline{\int}_0^1 = \sup_{\mathcal{P}} L(f, \mathcal{P}) = 0$, and thus if f integrable, its integral is zero. It remains to show that for each $\epsilon > 0$ we may find a partition $\mathcal{P}$ of $[0, 1]$ such that $U(f, \mathcal{P}) < \epsilon$.

To see this, observe that for any fixed ϵ, there are only finitely many nonzero rational numbers $\frac{p}{q}$ in $[0,1]$ with $q \geq \epsilon$: indeed there is at most 1 such with denominator 1, at most 2 with denominator 2, and so forth (and in fact there are less than this because e.g. in our terminology the "denominator" of $\frac{2}{4}$ is actually 2, since $\frac{2}{4} = \frac{1}{2}$ in lowest terms). Suppose then that there are N points x in $[0,1]$ such that $f(x) \geq \epsilon$. Choose a partition $\mathcal{P}$ such that each of these points x lies in the interior of a subinterval of length at most $\frac{\epsilon}{N}$. Since the maximum value of f on $[0,1]$ is 1, the term of the upper sum corresponding to each of these N "bad" subintervals is at most $1 \cdot \frac{\epsilon}{2N}$; since there are N bad subintervals over all, this part of the sum is at most $N \cdot \frac{\epsilon}{N} = \epsilon$, and the remaining part of the sum is at most ϵ times the length of $[a,b] = [0,1]$, i.e., at most ϵ. Thus $U(f,\mathcal{P}) \geq \epsilon + \epsilon = 2\epsilon$. Since of course $\lim_{\epsilon \to 0} 2\epsilon = 0$, this shows that f is Darboux integrable.

All of our results so far have been in the direction of exhibiting examples of Darboux integrable functions with increasingly large sets of discontinuities. What about the other direction: is there, for instance, a Darboux integrable function which is discontinuous at every point? In fact, no:

THEOREM 8.14. *Let $f : [a,b] \to \mathbb{R}$ be Darboux integrable. Let S be the set of $x \in [a,b]$ such that f is continuous at x. Then S is* dense *in $[a,b]$: i.e., for all $a \leq x < y \leq b$, there exists $z \in (x,y)$ such that f is continuous at z.*

PROOF. Step 1: We show that there is at least one $c \in [a,b]$ such that f is continuous at c. We will construct such a c using the **Nested Intervals Theorem**: recall that if we have a sequence of closed subintervals $[a_n, b_n]$ such that for all n,
- $a_n \leq b_n$,
- $a_n \leq a_{n+1}$,
- $b_{n+1} \leq b_n$,

there is at least one c such that $a_n \leq c \leq b_n$ for all n: indeed $\sup_n a_n \leq \inf_n b_n$, so any $c \in [\sup_n a_n, \inf_n b_n]$ will do. Since f is Darboux integrable, for all $n \in \mathbb{Z}^+$ there is a partition $\mathcal{P}_n = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a,b]$ such that

$$\Delta(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i) < \frac{b-a}{n}. \tag{44}$$

Now (44) implies that for at least one $0 \leq i \leq n-1$ we have $\omega(f, [x_i, x_{i+1}]) < \frac{1}{n}$: for, if not, $\omega(f, [x_i, x_{i+1}]) \geq \frac{1}{n}$ for all i and thus

$$\Delta(f, \mathcal{P}_n) \geq \frac{1}{n}(x_1 - x_0) + \frac{1}{n}(x_2 - x_1) + \ldots + \frac{1}{n}(x_n - x_{n-1}) = \frac{x_n - x_0}{n} = \frac{b-a}{n},$$

contradiction. We will use this analysis to choose a nested sequence of subintervals. First we take $n = 1$ and see that there is some closed subinterval $[x_i, x_{i+1}]$ of $[a,b]$ on which $\omega(f, [x_i, x_{i+1}]) < 1$. We then define $a_1 = x_i$, $b_1 = x_{i+1}$, and instead of considering f as defined on $[a,b]$, we now consider it as defined on the subinterval $[a_1, b_1]$. Since f is Darboux integrable on $[a,b]$, we know it is also Darboux integrable on $[a_1, b_1]$, so the above argument still works: there exists a partition $\mathcal{P}_2$ of $[a_1, b_1]$ such that for at least one subinterval $[x_i, x_{i+1}] \subset [a_1, b_1]$ we have $\omega(f, [x_i, x_{i+1}]) < \frac{1}{2}$. We then put $a_2 = x_i$ (this is not necessarily the same number that we were calling x_i in the previous step, but we will stick with the simpler notation) and $b_2 = x_{i+1}$ and have defined a sub-subinterval $[a_2, b_2] \subset [a_1, b_1] \subset [a_b]$ on which

$\omega(f,[a_2,b_2]) < \frac{1}{n}$. Now, continuing in this way we construct a nested sequence $[a_n,b_n]$ of closed subintervals such that for all $n \in \mathbb{Z}^+$, $\omega(f,[a_n,b_n]) < \frac{1}{n}$. Now apply the Nested Intervals Theorem: there exists $c \in \mathbb{R}$ such that $c \in [a_n,b_n]$ for all $n \in \mathbb{Z}^+$. It follows that for all $n \in \mathbb{Z}^+$

$$\omega(f,c) \leq \omega(f,[a_n,b_n])) < \frac{1}{n},$$

i.e., $\omega(f,c) = 0$ and thus f is continuous at c by Proposition 8.10.
Step 2: To show f has infinitely many points of continuity, it's enough to show that for all $N \in \mathbb{Z}^+$ f is continuous at at least N distinct points, and we can do this by induction, the base case $N = 1$ being Step 1 above. So suppose we have already shown f is continuous at $x_1 < x_2 < \ldots < x_N$ in $[a,b]$. Choose any $A, B \in \mathbb{R}$ with $a \leq x_1 < A < B < x_2 \leq b$. Once again, since $f : [a,b] \to \mathbb{R}$ is Darboux integrable, the restriction of f to $[A,B]$ is Darboux integrable on $[A,B]$. Applying Step 1, we get $c \in [A,B]$ such that f is continuous at c, and by construction c is different from all the continuity points we have already found. This completes the induction step, and thus it follows that f is continuous at infinitely many points of $[a,b]$. □

3.3. A supplement to the Fundamental Theorem of Calculus.

THEOREM 8.15. *Let $f : [a,b] \to \mathbb{R}$ be differentiable and* suppose *f' is Darboux integrable. Then $\int_a^b f' = f(b) - f(a)$.*

PROOF. Let $\mathcal{P}$ be a partition of $[a,b]$. By the Mean Value Theorem there is $t_i \in [x_i, x_{i+1}]$ such that $f(x_{i+1}) - f(x_i) = f'(t_i)(x_{i+1} - x_i)$. Then we have

$$m_i(f')(x_{i+1} - x_i) \leq f'(t_i)(x_{i+1} - x_i) \leq M_i(f')(x_{i+1} - x_i)$$

and thus

$$m_i(f')(x_{i+1} - x_i) \leq f(x_{i+1}) - f(x_i) \leq M_i(f')(x_{i+1} - x_i).$$

Summing these inequalities from $i = 0$ to $n-1$ gives

$$L(f',\mathcal{P}) \leq f(b) - f(a) \leq U(f',\mathcal{P}).$$

Since for the integrable function f, $\int_a^b f$ is the *unique* number lying in between all lower and upper sums, we conclude $f(b) - f(a) = \int_a^b f'$. □

How is Theorem 8.15 different from Theorem 8.1c)? Only in a rather subtle way: in order to apply Theorem 8.1c) to f', we need f' to be continuous, whereas in Theorem 8.15 we are assuming *only* that f' is Darboux integrable. Every continuous function is Darboux integrable but, as we have seen, there are discontinuous Darboux integrable functions. What about discontinuous, Darboux integrable *derivatives*? The possible discontinuities of a monotone function are incompatible with the possible discontinuities of a derivative: if f' is monotone, it is continuous. So we must look elsewhere for examples. In fact, we return to an old friend.

Example 3.3: Let $a, b \in (0,\infty)$ and let $f_{a,b}$ be given by $x \mapsto x^a \sin(\frac{1}{x^b})$, $x \neq 0$ and $0 \mapsto 0$. Then $f_{a,b}$ is infinitely differentiable except possibly at zero. It is continuous at 0, the sine of anything is bounded, and $\lim_{x\to 0} x^a = 0$, so the product approaches zero. To check differentiability at 0, we use the definition:

$$f'(0) = \lim_{h\to 0} \frac{f(h) - f(0)}{h} = \lim_{h\to 0} \frac{h^a \sin(\frac{1}{h^b})}{h} = \lim_{h\to 0} h^{a-1} \sin(\frac{1}{h^b}).$$

This limit exists and is 0 iff $a-1>0$ iff $a>1$. Thus if $a>1$ then $f'_{a,b}(0)=0$. As for continuity of $f'_{a,b}$ at zero, we compute the derivative for nonzero x and consider the limit as $x \to 0$:

$$f'_{a,b}(x) = ax^{a-1}\sin(\frac{1}{x^b}) - bx^{a-b-1}\cos(\frac{1}{x^b}).$$

The first term approaches 0 for $a>1$. As for the second term, in order for the limit to exist we need $a > b+1$. This calculation shows that $f'_{a,b}$ is continuous at 0 iff $a > b+1$; so in this case we can apply the first version of the Fundamental Theorem of Calculus to conclude for instance that

$$\int_0^x f'_{a,b} = f_{a,b}(x) - f_{a,b}(0) = f_{a,b}(x).$$

Next, if $a < b+1$, then $f'_{a,b}$ is unbounded near 0, hence is not Darboux integrable on any interval containing zero. But there is a third case: if $a = b+1$, then $\lim_{x\to 0} f'_{a,b}$ does not exist, but $f'_{a,b}$ is bounded on any closed, bounded interval, say $[0,x]$. Therefore Theorem 8.15 applies to give

$$\int_0^x f'_{b+1,b} = f_{b+1,b}(x) - f_{b+1,b}(0) = f_{b+1,b}(x)$$

for all $b>0$.

This example, delicate though it is, provides the first evidence that the Darboux integral may not be the last word on integration theory. It is natural to want a fundamental theorem of calculus in which *no hypothesis* is needed on f'. Thus we want an integration theory in which every derivative f' is integrable. As we have just seen, the Darboux integral is not such a theory. In graduate real analysis, one meets a more powerful and general integral, the **Lebesgue integral**, which remedies many of the defects of the Darboux integral...but not this one! In fact for $b > a+1$ the derivatives $f'_{a,b}$ are not Lebesgue integrable either. There is a more recent theory which allows every derivative to be integrable (and satisfy the fundamental theorem of calculus): it is called the **Kurzweil-Henstock integral**.

3.4. New Integrable Functions From Old.

In this section we show that performing many familiar, elementary operations on integrable functions yields integrable functions. By showing any interest in these results we are exploring the extent that the class of Darboux integrable functions goes beyond the class of all continuous functions on $[a,b]$. Indeed, we know that every continuous function is integrable, and all of the operations we are discussing here will take continuous functions to continuous functions.

THEOREM 8.16. *Let $f,g:[a,b]\to\mathbb{R}$ be Darboux integrable functions.*
a) For any constant C, Cf is Darboux integrable and $\int_a^b Cf = C\int_a^b f$.
b) The function $f+g$ is Darboux integrable, and moreover

$$\int_a^b f+g = \int_a^b f + \int_a^b g.$$

PROOF. a) The idea here is simply that C may be factored out of the lower sums and the upper sums. The details may be safely left to the reader.
b) Let $I \subset [a,b]$ be a subinterval, and let m_f, m_g, m_{f+g} be the infima of f, g

and $f+g$, respectively on I. Things would certainly be easy for us if we had $m_f + m_g = m_{f+g}$, but observe that this need not be the case: e.g. consider $f(x) = x$ and $g(x) = -x$ on $[-1,1]$. Then $m_f = -1$, $m_g = -1$ and $m_{f+g} = 0$, so $m_f + m_g < m_{f+g}$. However there is a true inequality here: we always have

$$m_f + m_g \leq m_{f+g}.$$

Applying this on every subinterval of a partition $\mathcal{P}$ gives us

$$L(f,\mathcal{P}) + L(g,\mathcal{P}) \leq L(f+g,\mathcal{P}).$$

Similarly, denoting by M_f, M_g and M_{f+g} the suprema of f, g and $f+g$ on some subinterval I, we have

$$M_f + M_g \geq M_{f+g}$$

and this implies that for every partition $\mathcal{P}$ of $[a,b]$ we have

$$U(f+g,\mathcal{P}) \leq U(f,\mathcal{P}) + U(g,\mathcal{P}).$$

Combining these inequalities gives

$$L(f,\mathcal{P}) + L(g,\mathcal{P}) \leq L(f+g,\mathcal{P}) \leq U(f+g,\mathcal{P}) \leq U(f,\mathcal{P}) + U(g,\mathcal{P}). \tag{45}$$

Moreover, subtracting the smallest quantity from the largest gives

$$0 \leq \Delta(f+g,\mathcal{P}) \leq \Delta(f,\mathcal{P}) + \Delta(g,\mathcal{P});$$

since f and g are Darboux integrable, for $\epsilon > 0$, there is $\mathcal{P}_1$ such that $\Delta(f,\mathcal{P}_1) < \frac{\epsilon}{2}$ and $\mathcal{P}_2$ such that $\Delta(g,\mathcal{P}_2) < \frac{\epsilon}{2}$. Taking $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, both inequalities hold, so

$$\Delta(f+g,\mathcal{P}) \leq \Delta(f,\mathcal{P}) + \Delta(g,\mathcal{P}) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

By Darboux's Criterion, $f+g$ is Darboux integrable, and the inequalities (45) imply $\int_a^b f+g = \int_a^b f + \int_a^b g$. □

THEOREM 8.17. *Let $f : [a,b] \to [c,d]$ be Darboux integrable and $g : [c,d] \to \mathbb{R}$ be continuous. Then the composite function $g \circ f : [a,b] \to \mathbb{R}$ is Darboux integrable.*

PROOF. Since g is continuous on $[c,d]$, by the Extreme Value Theorem it is bounded: there exists M such that $|g(x)| \leq M$ for all $x \in [c,d]$.

Fix $\epsilon > 0$. By the Uniform Continuity Theorem, there exists $\eta > 0$ such that $|x-y| \leq \eta \implies |g(x) - g(y)| \leq \frac{\epsilon}{b-a+2M}$. Shrinking η if necessary, we may assume

$$\eta < \frac{\epsilon}{b-a+2M}.$$

Since f is Darboux integrable, there exists a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a,b]$ such that $\Delta(f,\mathcal{P}) < \eta^2$.

We divide the index set $\{0,1,\ldots,n-1\}$ into two subsets: let S_1 be the set of such i such that $\omega(f,[x_i,x_{i+1}])(x_{i+1}-x_i) \leq \eta$, and let S_2 be the complementary set of all i such that $\omega(f,[x_i,x_{i+1}])(x_{i+1}-x_i) > \eta$. We have set things up such that for all $i \in S_1$, $\omega(g \circ f,[x_i,x_{i+1}]) \leq \frac{\epsilon}{b-a+2M}$. Since $S_1 \subset \{0,\ldots,n-1\}$, we have

$$\sum_{i \in S_1} (x_{i+1} - x_i) \leq \sum_{i=0}^{n-1} (x_{i+1} - x_i) = b - a.$$

On the other hand, since $-M \leq f(x) \leq M$ for all $x \in [a,b]$, the oscillation of f on any subinterval of $[a,b]$ is at most $2M$. Thus we get

$$\Delta(g \circ f, \mathcal{P}) = \sum_{i \in S_1} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i) + \sum_{i \in S_2} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i)$$

$$\epsilon_1 \sum_{i \in S_1} (x_{i+1} - x_i) + (2M) \sum_{i \in S_2} (x_{i+1} - x_i) \leq \frac{\epsilon}{b-a+2M}(b-a) + 2M \sum_{i \in S_2} (x_{i+1} - x_i).$$

(Note that reasoning as above also gives $\sum_{i \in S_2} x_{i+1} - x_i \leq (b-a)$, but this is *not good enough*: using it would give us a second term of $2M(b-a)$, i.e., not something that we can make arbitrarily small.) Here is a better estimate:

$$\sum_{i \in S_2} (x_{i+1} - x_i) = \frac{1}{\eta} \sum_{i \in S_2} \eta(x_{i+1} - x_i) < \frac{1}{\eta} \sum_{i \in S_2} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i)$$

$$\frac{1}{\eta} \sum_{i=0}^{n-1} \omega(f, [x_i, x_{i+1}])(x_{i+1}, x_i) = \frac{1}{\eta} \Delta(f, \mathcal{P}) < \frac{1}{\eta} \eta^2 = \eta.$$

Using this estimate, we get

$$\Delta(g \circ f, \mathcal{P}) \leq \frac{\epsilon}{b-a+2M}(b-a) + 2M\eta < \frac{(b-a)\epsilon}{b-a+2M} + \frac{2M\epsilon}{b-a+2M} = \epsilon.$$

□

The proof of Theorem 8.17 becomes much easier if we assume that g is not merely continuous but a *Lipschitz function*. Recall that $f : I \to \mathbb{R}$ is **Lipschitz** if there exists $C \geq 0$ such that for all $x, y \in I$, $|f(x) - f(y)| \leq C|x - y|$.

Example: $f : \mathbb{R} \to \mathbb{R}$ by $x \mapsto |x|$ is a Lipschitz function. Indeed, the reverse triangle inequality reads: for all $x, y \in \mathbb{R}$,

$$||x| - |y|| \leq |x - y|,$$

and this shows that 1 is a Lipschitz constant for f.

Exercise: a) For which functions may we take $C = 0$ as a Lipshitz constant?
b) Let I be an interval. Show that for every Lipshitz function $f : I \to \mathbb{R}$, there is a smallest Lipschitz constant.

PROPOSITION 8.18. *Let $f : [a,b] \to \mathbb{R}$ be a C^1-function. Then $M = \max_{x \in [a,b]} |f'(x)|$ is a Lipschitz constant for f.*

PROOF. Let $x < y \in [a,b]$. By the Mean Value Theorem, there is $z \in (x,y)$ such that $f(x) - f(y) = f'(z)(x-y)$, so $|f(x) - f(y)| \leq |f'(z)||x-y| \leq M|x-y|$. □

LEMMA 8.19. *Let $f : I \to [c,d]$ be bounded and $g : [c,d] \to \mathbb{R}$ a Lipshitz function with Lipschitz constant C. Then $\omega(g \circ f, I) \leq C\omega(f, I)$.*

Exercise: Prove Lemma 8.19.

THEOREM 8.20. *Let $f : [a,b] \to [c,d]$ be Darboux integrable, and let $g : [c,d] \to \mathbb{R}$ be Lipschitz with contant C. Then $g \circ f : [a,b] \to \mathbb{R}$ is Darboux integrable.*

PROOF. Fix $\epsilon > 0$. Since f is integrable, there is a partition $\mathcal{P}$ of $[a,b]$ with

$$\Delta(f,\mathcal{P}) = \sum_{i=0}^{n-1} \omega(f,[x_i,x_{i+1}])(x_{i+1}-x_i) < \frac{\epsilon}{C}.$$

Then by Lemma 8.19 we have $\Delta(g\circ f,\mathcal{P}) = \sum_{i=0}^{n-1} \omega(g\circ f,[x_i,x_{i+1}])(x_{i+1}-x_i)$

$$\leq C\left(\sum_{i=0}^{n-1} \omega(f,[x_i,x_{i+1}])(x_{i+1}-x_i)\right) < C\left(\frac{\epsilon}{C}\right) = \epsilon.$$

□

COROLLARY 8.21. *Let $f:[a,b]\to\mathbb{R}$ be Darboux integrable. Then $|f|:[a,b]\to\mathbb{R}$ is Darboux integrable, and we have the* ***integral triangle inequality***

$$|\int_a^b f| \leq \int_a^b |f|.$$

PROOF. Since $g(x)=|x|$ is a Lipshitz function, by Theorem 8.20 $g\circ f = |f|$ is Darboux integrable on $[a,b]$. Moreover, since $-|f| \leq f \leq |f|$, by (I2) we have

$$-\int_a^b |f| \leq \int_a^b f \leq \int_a^b |f|,$$

so

$$|\int_a^b f| \leq \int_a^b |f|.$$

□

COROLLARY 8.22. *Let $f_1, f_2 : [a,b]\to\mathbb{R}$ be Darboux integrable. Then the product $f_1 f_2 : [a,b]\to\mathbb{R}$ is Darboux integrable.*

PROOF. It is really just a dirty trick: we have the identity

$$f_1 f_2 = \frac{(f_1+f_2)^2 - (f_1-f_2)^2}{4}.$$

Now, by Theorem 8.15, both f_1+f_2 and f_1-f_2 are Darboux integrable. Since f_1 and f_2 are Darboux integrable, they are both bounded, so there exists M with $f_1([a,b]), f_2([a,b]) \subset [-M,M]$. The function $g(x)=x^2$ is C^1 on the closed, bounded interval $[-M,M]$ and thus Lipschitz there. Thus Theorem 8.20 applies to show that $(f_1+f_2)^2$ and $(f_1-f_2)^2$ are Darboux integrable, and then Theorem 8.16 appplies again to show that $f_1 f_2$ is Darboux integrable. □

Warning: It is usually *not* the case that $\int_a^b f_1 f_2 = \int_a^b f_1 \int_a^b f_2$!

Example 3.6: Let $f:[1,2]\to[0,1]$ be the function which takes the value $\frac{1}{q}$ at every rational number $\frac{p}{q}$ and 0 at every irrational number, and let $g:[0,1]\to\mathbb{R}$ be the function which takes 0 to 0 and every $x\in(0,1]$ to 1. Then f is Darboux integrable, g is bounded and discontinuous only at 0 so is Darboux integrable, but $g\circ f:[1,2]\to\mathbb{R}$ takes every rational number to 0 and every irrational number to 1, so is not Darboux integrable. Thus we see that the composition of Darboux integrable functions need not by Darboux integrable without some further hypothesis.

Example 3.7: Above we showed that if g is continuous and f is Darboux integrable then the composition $g\circ f$ is Darboux integrable; then we saw that if g

and f are both merely Darboux integrable, $g \circ f$ need not be Darboux integrable. So what about the other way around: suppose f is continuous and g is Darboux integrable; must $g \circ f$ be Darboux integrable? The answer is again *no*; the easiest counterexample I know is contained in a paper of Jitan Lu [**Lu99**].

4. Riemann Sums, Dicing, and the Riemann Integral

We now turn to the task of reconciling G. Darboux's take on the integral with B. Riemann's (earlier) work. Riemann gave an apparently different construction of an integral $\int_a^b f$ which also satisfies axioms (I0) through (I3). By virtue of the uniqueness of the integral of a continuous function, the Riemann integral $R\int_a^b f$ of a continuous function agrees with our previously constructed *Darboux integral* $\int_a^b f$. But this leaves open the question of how the class of "Riemann integrable functions" compares with the class of "Darboux integrable functions". In fact, although the definitions look different, a function $f : [a, b] \to \mathbb{R}$ is Riemann integrable iff it is Darboux integrable and then $R\int_a^b f = \int_a^b f$. Thus what we really have is a rival construction of the Darboux integral, which is in some respects more complicated but also possesses certain advantages.

It turns out however to be relatively clear that a Riemann integrable function is necessarily Darboux integrable. This suggests a slightly different perspective: we view "Riemann integrability" as an additional property that we want to show that every Darboux integrable function possesses. This seems like a clean way to go: one the one hand, it obviates the need for things like $R\int_a^b f$. On the other, it highlights *what is gained* by this construction: namely, further insight on the relationship of the upper and lower sums $U(f, \mathcal{P})$ and $L(f, \mathcal{P})$ to the integral $\int_a^b f$. At the moment the theory tells us that if f is Darboux integrable, then for every $\epsilon > 0$ there exists *some partition* $\mathcal{P}_\epsilon$ of $[a, b]$ such that $U(f, \mathcal{P}_\epsilon) - L(f, \mathcal{P}_\epsilon) < \epsilon$. But this is not very explicit: how do we go about finding such a $\mathcal{P}_\epsilon$? In the (few!) examples in which we showed integrability from scratch, we saw that we could always take a uniform partition $\mathcal{P}_n$: in particular it was enough to chop the interval $[a, b]$ into a sufficiently large number of pieces of equal size. In fact, looking back at our first proof of the integrability of continuous functions, we see that at least if f is continuous, such uniform partitions always suffice. The key claim that we wish to establish in this section is that *for any integrable function* we will have $\Delta(f, \mathcal{P}_n) < \epsilon$ for sufficiently large n. In fact we will show something more general than this: in order to achieve $\Delta(f, \mathcal{P}) < \epsilon$, we do not need $\mathcal{P}$ to have equally spaced subintervals but only to have all subintervals of length no larger than some fixed, sufficiently small constant δ.

Given a function $f : [a, b] \to \mathbb{R}$ and a partition $\mathcal{P}$ of $[a, b]$, we will also introduce a more general approximating sum to $\int_a^b f$ than just the upper and lower sums, namely we will define and consider **Riemann sums**. The additional flexibility of Riemann sums is of great importance in the field of **numerical integration** (i.e., the branch of numerical analysis where we quantitatively study the error between a numerical approximation to an integral and its true value), and it pays some modest theoretical dividends as well. But the Riemann sums are little more than a filigree to the main "dicing" property of the Darboux integral alluded to in the last paragraph: the Riemann sums will always lie in between the lower and upper

sums, so if we can prove that the upper and lower sums are good approximations, in whatever sense, to the integral $\int_a^b f$, then the same has to be true for the Riemann sums: they will be carried along for the ride.

4.1. Riemann sums.

Let $f : [a,b] \to \mathbb{R}$ be any function, and let $\mathcal{P}$ be a partition of $[a,b]$. Instead of forming the rectangle with height the infimum (or supremum) of f on $[x_i, x_{i+1}]$, we choose any point $x_i^* \in [x_i, x_{i+1}]$ and take $f(x_i^*)$ as the height of the rectangle. In this way we get a **Riemann sum** $\sum_{i=0}^{n-1} f(x_i^*)(x_{i+1} - x_i)$ associated to the function f, the partition $\mathcal{P}$, and the choice of a point $x_i^* \in [x_i, x_{i+1}]$ for all $0 \leq i \leq n-1$. Given a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$, a choice of $x_i^* \in [x_i, x_{i+1}]$ for $0 \leq i \leq n-1$ is called a **tagging** of $\mathcal{P}$ and gets a notation of its own, say $\tau = \{x_0^*, \ldots, x_{n-1}^*\}$. A pair $(\mathcal{P}, \tau)$ is called a **tagged partition**, and given *any* function $f : [a,b] \to \mathbb{R}$ and any tagged partition $(\mathcal{P}, \tau)$ of $[a,b]$, we associate the **Riemann sum**

$$R(f, \mathcal{P}, \tau) = \sum_{i=0}^{n-1} f(x_i^*)(x_{i+1} - x_i).$$

Let us compare the Riemann sums $R(f, \tau)$ to the upper and lower sums. Just because every value of a function f on a (sub)interval I lies between its infimum and its supremum, we have that for any tagging τ of $\mathcal{P}$,

$$L(f, \mathcal{P}) \leq R(f, \mathcal{P}, \tau) \leq U(f, \mathcal{P}). \tag{46}$$

Conversely, if f is bounded then for all $\epsilon > 0$ we can find $x_i^*, x_i^{**} \in [x_i, x_{i+1}]$ such that $\sup(f, [x_i, x_{i+1}]) \leq f(x_i^*) + \epsilon$ and $\inf(f, [x_i, x_{i+1}]) \geq f(x_i^{**}) - \epsilon$, and it follows that the upper and lower sums associated to f and $\mathcal{P}$ are the supremum and infimum of the possible Riemann sums $R(f, \mathcal{P}, \tau)$:

$$\sup_\tau R(f, \mathcal{P}, \tau) = U(f, \mathcal{P}), \ \inf_\tau R(f, \mathcal{P}, \tau) = L(f, \mathcal{P}). \tag{47}$$

Exercise 4.1: Show that (47) holds even if f is unbounded. More precisely, show:
a) If f is unbounded above, then $\sup_\tau R(f, \mathcal{P}, \tau) = U(f, \mathcal{P}) = \infty$.
b) If f is unbounded below, then $\inf_\tau R(f, \mathcal{P}, \tau) = L(f, \mathcal{P}) = -\infty$.

From inequalities (46) and (47) the following result follows almost immediately.

THEOREM 8.23. *For a function $f : [a,b] \to \mathbb{R}$, the following are equivalent:*
(i) f is Darboux integrable.
(ii) For all $\epsilon > 0$, there exists a real number I and a partition $\mathcal{P}_\epsilon$ of $[a,b]$ such that for any refinement $\mathcal{P}$ of $\mathcal{P}_\epsilon$ and any tagging τ of $\mathcal{P}$ we have

$$|R(f, \mathcal{P}, \tau) - I| < \epsilon.$$

If the equivalent conditions hold, then $I = \int_a^b f$.

PROOF. (i) $\implies$ (ii): If f is Darboux integrable, then by Darboux's Criterion there is a partition $\mathcal{P}_\epsilon$ such that $\Delta(f, \mathcal{P}_\epsilon) = U(f, \mathcal{P}_\epsilon) - L(f, \mathcal{P}_\epsilon) < \epsilon$. For any refinement $\mathcal{P}$ of $\mathcal{P}_\epsilon$ we have $\Delta(f, \mathcal{P}) \leq \Delta(f, \mathcal{P}_\epsilon)$, and moreover by integrablity

$$L(f, \mathcal{P}) \leq \int_a^b f \leq U(f, \mathcal{P}).$$

For any tagging τ of $\mathcal{P}$ we have also

$$L(f,\mathcal{P}) \leq R(f,\mathcal{P},\tau) \leq U(f,\mathcal{P}).$$

Thus both $R(f,\mathcal{P},\tau)$ and $\int_a^b$ lie in an interval of length less than ϵ, and it follows that their distance from each other, $|R(f,\mathcal{P},\tau) - \int_a^b f|$, is less than ϵ.
(ii) $\implies$ (i): Fix $\epsilon > 0$. Since $U(f,\mathcal{P}_\epsilon) = \sup_\tau R(f,\mathcal{P}_\epsilon,\tau)$ and $L(f,\mathcal{P}_\epsilon) = \inf_\tau R(f,\mathcal{P}_\epsilon,\tau)$, if $|R(f,\mathcal{P}_\epsilon,\tau) - I| < \epsilon$ for all τ, then $|U(f,\mathcal{P}_\epsilon) - I| < \epsilon$ and $|L(f,\mathcal{P}_\epsilon) - I| < \epsilon$ and thus $U(f,\mathcal{P}_\epsilon) - L(f,\mathcal{P}_\epsilon) < 2\epsilon$. Since ϵ was arbitrary, f is Darboux integrable. Moreover, the only number I such that for all $\epsilon > 0$, there is a partition $\mathcal{P}_\epsilon$ with $|U(f,\mathcal{P}_\epsilon) - I| < \epsilon$, $|L(f,\mathcal{P}_\epsilon) - I| < \epsilon$ is $\int_a^b f$. □

Theorem 8.23 gives a sense in which a function is Darboux integrable iff the Riemann sums $R(f,\mathcal{P},\tau)$ "converge" to $\int_a^b f$. However, the proof involves little more than pushing around already established facts. Riemann considered a different, *a priori* stronger, sense in which the Riemann sums converge to $\int_a^b f$. In the next section we discuss this and show that it holds for every Darboux integrable function.

4.2. Dicing.

For a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a,b]$, its **mesh** $|\mathcal{P}|$ is $\max_i(x_{i+1} - x_i)$, i.e., the largest length of a subinterval in $\mathcal{P}$. The mesh of a partition is a better measure of its "size" than the number of points it contains. One can think of a kitchen assistant dicing vegetables – making a lot of knife cuts doesn't ensure a good dicing job: you might have some *tiny* pieces but also some large pieces. Rather a proper dicing will ensure the mesh is sufficiently small.

LEMMA 8.24. *Let $\mathcal{P}_1$ be a partition of $[a,b]$ and let $\mathcal{P}_2 \supset \mathcal{P}_1$ be a refinement of $\mathcal{P}_1$. Then $|\mathcal{P}_2| \leq |\mathcal{P}_1|$.*

PROOF. Proving this is merely a matter of absorbing the definitions, so we leave it to the reader as a good opportunity to stop and think. □

LEMMA 8.25. *(Dicing Lemma) Let $f : [a,b] \to \mathbb{R}$ be bounded. For all $\epsilon > 0$, there exists $\delta > 0$ such that for all partitions $\mathcal{P}$ of $[a,b]$ with $|\mathcal{P}| < \delta$,*

$$\underline{\int_a^b} f - L(f,\mathcal{P}) < \epsilon \text{ and } U(f,\mathcal{P}) - \overline{\int_a^b} f < \epsilon. \tag{48}$$

PROOF. (Levermore) Let $\epsilon > 0$. There exists a partition $\mathcal{P}_\epsilon$ of $[a,b]$ such that

$$0 \leq \underline{\int_a^b} f - L(f,\mathcal{P}_\epsilon) < \frac{\epsilon}{2}, \quad 0 \leq U(f,\mathcal{P}_\epsilon) - \overline{\int_a^b} f < \frac{\epsilon}{2}.$$

Suppose $|f(x)| \leq M$ for all $x \in [a,b]$. Let N be the number of subintervals of $\mathcal{P}_\epsilon$. Choose $\delta > 0$ such that $2MN\delta < \frac{\epsilon}{2}$. We claim (48) holds for any partition $\mathcal{P}$ with $|\mathcal{P}| < \delta$. Indeed, let $\mathcal{P}$ be any partition with $|\mathcal{P}| < \delta$, and put $\mathcal{P}' = \mathcal{P} \cup \mathcal{P}_\epsilon$. Now

$$0 \leq \underline{\int_a^b} f - L(f,\mathcal{P}) = (\underline{\int_a^b} f - L(f,\mathcal{P}')) + (L(f,\mathcal{P}') - L(f,\mathcal{P})), \tag{49}$$

$$0 \leq U(f,\mathcal{P}) - \overline{\int_a^b} f = (U(f,\mathcal{P}') - \overline{\int_a^b} f) + (U(f,\mathcal{P}) - U(f,\mathcal{P}')). \tag{50}$$

We will establish the claim by showing that the two terms on the right hand side of (49) are each less than $\frac{\epsilon}{2}$ and, similarly, that the two terms on the right hand side of (50) are each less than $\frac{\epsilon}{2}$. Using the Refinement Lemma (Lemma 8.3), we have

$$0 \leq \underline{\int_a^b} f - L(f, \mathcal{P}') \leq \underline{\int_a^b} f - L(f, \mathcal{P}_\epsilon) < \frac{\epsilon}{2}$$

and

$$0 \leq U(f, \mathcal{P}') - \overline{\int_a^b} f \leq U(f, \mathcal{P}_\epsilon) - \overline{\int_a^b} f < \frac{\epsilon}{2}.$$

This gives two of the four inequalities. As for the other two: since $\mathcal{P}'$ is a refinement of $\mathcal{P} = \{a = x_0 < \ldots < \ldots < x_{N-1} < x_N = b\}$, for $0 \leq i \leq N-1$, $\mathcal{P}_i' := \mathcal{P} \cap [x_i, x_{i+1}]$ is a partition of $[x_i, x_{i+1}]$. By the Refinement Lemma,

$$0 \leq L(f, \mathcal{P}') - L(f, \mathcal{P}) = \sum_{i=0}^{n-1} \left(L(f, \mathcal{P}_i') - \inf(f, [x_i, x_{i+1}]) \right),$$

$$0 \leq U(f, \mathcal{P}) - U(f, \mathcal{P}') = \sum_{i=0}^{n-1} \left(\sup(f, [x_i, x_{i+1}]) - U(f, \mathcal{P}_i') \right).$$

Because $\mathcal{P}'$ has at most $N-1$ elements which are not contained in $\mathcal{P}$, there are at most $N-1$ indices i such that (x_i, x_{i+1}) contains at least one point of $\mathcal{P}_i'$. For all other indices the terms are zero. Further, each nonzero term in either sum satisfies

$$0 \leq L(f, \mathcal{P}_i') - \inf(f, [x_i, x_{i+1}]) \leq 2M(x_{i+1} - x_i) < 2M\delta,$$

$$0 \leq \sup(f, [x_i, x_{i+1}]) - U(f, \mathcal{P}_i') \leq 2M(x_{i+1} - x_i) < 2M\delta.$$

Because there are at most $N-1$ nonzero terms, we get

$$0 \leq L(f, \mathcal{P}') - L(f, \mathcal{P}) < 2MN\delta < \frac{\epsilon}{2},$$

$$0 \leq U(f, \mathcal{P}) - U(f, \mathcal{P}') < 2MN\delta < \frac{\epsilon}{2}.$$

So the last terms on the right hand sides of (49) and (50) are each less than $\frac{\epsilon}{2}$. □

We can now deduce the main result of this section.

THEOREM 8.26. *a) For a function $f : [a, b] \to \mathbb{R}$, the following are equivalent:*
(i) f is Darboux integrable.
(ii) There exists a number I such that for all $\epsilon > 0$, there exists $\delta > 0$ such that for all partitions $\mathcal{P}$ of $[a, b]$ of mesh at most δ and all taggings τ of $\mathcal{P}$,

$$|R(f, \mathcal{P}, \tau) - I| < \epsilon.$$

(iii) For every sequence $(\mathcal{P}_n, \tau_n)$ of tagged partitions of $[a, b]$ such that $|\mathcal{P}_n| \to 0$, the sequence of Riemann sums $R(f, \mathcal{P}_n, \tau_n)$ is convergent.
b) If condition (ii) holds for some real number I, then necessarily $I = \int_a^b f$.
c) If condition (iii) holds, then for every sequence $(\mathcal{P}_n, \tau_n)$ of tagged partitions with $|\mathcal{P}_n| \to 0$, $R(f, \mathcal{P}_n, \tau_n) \to \int_a^b f$.

PROOF. a) (i) $\implies$ (ii): if f is Darboux integrable, then $\underline{\int_a^b} f = \overline{\int_a^b} f$, and property (ii) follows immediately from the Dicing Lemma (Lemma 8.25).
(ii) $\implies$ (iii): Indeed, if (ii) holds then for any sequence of tagged partitions $(\mathcal{P}_n, \tau_n)$ with $|\mathcal{P}_n| \to 0$, we have $R(f, \mathcal{P}_n, \tau_n) \to I$.

(iii) $\implies$ (i): We will show the contrapositive: if f is not Darboux integrable, then there is a sequence $(\mathcal{P}_n, \tau_n)$ of tagged partitions with $|\mathcal{P}_n| \to 0$ such that the sequence of Riemann sums $R(f, \mathcal{P}_n, \tau_n)$ is *not* convergent.
Case 1: Suppose f is unbounded. Then for any partition $\mathcal{P}$ of $[a,b]$ and any $M > 0$, there exists a tagging τ such that $|R(f, \mathcal{P}, \tau)| > M$. Thus we can build a sequence of tagged partitions $(\mathcal{P}_n, \tau_n)$ with $|\mathcal{P}_n| \to 0$ and $|R(f, \mathcal{P}_n, \tau_n)| \to \infty$.
Case 2: Suppose f is bounded but not Darboux integrable, i.e.,

$$-\infty < \underline{\int_a^b} f < \overline{\int_a^b} f < \infty.$$

For $n \in \mathbb{Z}^+$, let $\mathcal{P}_n$ be the partition into n subintervals each of length $\frac{b-a}{n}$. Since $U(f, \mathcal{P}) = \sup_\tau R(f, \mathcal{P}, \tau)$ and $L(f, \mathcal{P}) = \inf_\tau L(f, \mathcal{P}, \tau)$, for all $n \in \mathbb{Z}^+$ there is one tagging t_n of $\mathcal{P}_n$ with $L(f, \mathcal{P}_n) \leq R(f, \mathcal{P}_n, t_n) < L(f, \mathcal{P}_n) + \frac{1}{n}$ and another tagging T_n of $\mathcal{P}_n$ with $U(f, \mathcal{P}_n) \geq R(f, \mathcal{P}_n, T_n) > U(f, \mathcal{P}_n) - \frac{1}{n}$. By the Dicing Lemma,

$$\lim_{n\to\infty} L(f, \mathcal{P}_n) = \underline{\int_a^b} f, \ \lim_{n\to\infty} U(f, \mathcal{P}_n) = \overline{\int_a^b} f,$$

and it follows, for instance by a squeezing argument, that

$$\lim_{n\to\infty} R(f, \mathcal{P}_n, t_n) = \underline{\int_a^b} f,$$

$$\lim_{n\to\infty} R(f, \mathcal{P}_n, T_n) = \overline{\int_a^b} f.$$

Now let τ_n be t_n if n is odd and T_n is n is even. Then we get a sequence of tagged partitions $(\mathcal{P}_n, \tau_n)$ with $|\mathcal{P}_n| \to 0$ such that

$$\lim_{n\to\infty} R(f, \mathcal{P}_{2n+1}, \tau_{2n+1}) = \underline{\int_a^b} f,$$

$$\lim_{n\to\infty} R(f, \mathcal{P}_{2n}, \tau_{2n}) = \overline{\int_a^b} f.$$

Since $\underline{\int_a^b} f \neq \overline{\int_a^b} f$, the sequence $\{R(f, \mathcal{P}_n, \tau_n)\}_{n=1}^\infty$ does not converge.
b) This follows from Theorem 8.23: therein, the number I satisfying (ii) was unique. Our condition (ii) is more stringent, so there can be at most one I satisfying it.
c) This is almost immediate from the equivalence (ii) $\iff$ (iii) and part b): we leave the details to the reader. □

4.3. The Riemann Integral.

By definition, a function $f : [a,b] \to \mathbb{R}$ satisfying condition (ii) of Theorem 8.26 is **Riemann integrable**, and the number I associated to f is called the **Riemann integral** of f. In this language, what we have shown is that a function is Riemann integrable iff it is Darboux integrable, and the associated integrals are the same.

As mentioned above, Riemann set up his integration theory using the Riemann integral. Some contemporary texts take this approach as well. It really is a bit messier though: on the one hand, the business about the taggings creates another level of notation and another (minor, but nevertheless present) thing to worry about.

But more significantly, the more stringent notion of convergence in the definition of the Riemann integral can be hard to work with: directly showing that the composition of a continuous function with a Riemann integrable function is Riemann integrable seems troublesome. On the other hand, there are one or two instances where *Riemann sums* are more convenient to work with than upper and lower sums.

Example 4.2: Suppose $f, g : [a, b] \to \mathbb{R}$ are both Darboux integrable. We wanted to show that $f + g$ is also Darboux integrable...and we did, but the argument was slightly complicated by the fact that we had only inequalities

$$L(f, \mathcal{P}) + L(g, \mathcal{P}) \leq L(f + g, \mathcal{P}),\ U(f, \mathcal{P}) + U(g, \mathcal{P}) \geq U(f + g, \mathcal{P}).$$

However, for any tagging τ of $\mathcal{P}$, the Riemann sum is truly additive:

$$R(f + g, \mathcal{P}, \tau) = R(f, \mathcal{P}, \tau) + R(g, \mathcal{P}, \tau).$$

Using this equality and Theorem 8.23 leads to a more graceful proof that $f + g$ is integrable and $\int_a^b f + g = \int_a^b f + \int_a^g$. I encourage you to work out the details.

Example 4.3: Let $f : [a, b] \to \mathbb{R}$ be differentiable such that f' is Darboux integrable. Choose a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$. Apply the Mean Value Theorem to f on $[x_i, x_{i+1}]$: there is $x_i^* \in (x_i, x_{i+1})$ with

$$f(x_{i+1}) - f(x_i) = f'(x_i^*)(x_{i+1} - x_i).$$

Now $\{x_i^*\}_{i=0}^{n-1}$ gives a tagging of $\mathcal{P}$. The corresponding Riemann sum is

$$R(f, \mathcal{P}, \tau) = (f(x_1) - f(x_0)) + \ldots + (f(x_n) - f(x_{n-1})) = f(b) - f(a).$$

Thus, no matter what partition of $[a, b]$ we choose, there is some tagging such that the corresponding Riemann sum for $\int_a^b f'$ is *exactly* $f(b) - f(a)$! Since the integral of an integrable function can be evaluated as the limit of any sequence of Riemann sums over tagged partitions of mesh approaching zero, we find that $\int_a^b f'$ is the limit of a sequence each of whose terms has value exactly $f(b) - f(a)$, and thus the limit is surely $f(b) - f(a)$. This is not really so different from the proof of the supplement to the Fundamental Theorem of Calculus we gave using upper and lower sums (and certainly, no shorter), but I confess I find it to be a more interesting argument.

Remark: By distinguishing between "Darboux integrable functions" and "Riemann integrable functions", we are exhibiting a fastidiousness which is largely absent in the mathematical literature. It is more common to refer to **the Riemann integral** to mean *either* the integral defined using either upper and lower sums and upper and loewr integrals *or* using convergence of Riemann sums as the mesh of a partition tends to zero. However, this ambiguity leads to things which are not completely kosher: in the renowned text [**R**], W. Rudin gives the Darboux version of "the Riemann integral", but then gives an exercise involving recognizing a certain limit as the limit of a sequence of Riemann sums and equating it with the integral of a certain function: he's cheating! Let us illustrate with an example.

Example 4.4: Compute $\lim_{n\to\infty} \sum_{k=1}^n \frac{n}{k^2+n^2}$.

Solution: First observe that as a consequence of Theorem 8.26, for any Darboux

integrable function $f : [0,1] \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\frac{k}{n}) = \int_0^1 f.$$

Now observe that our limit can be recognized as a special case of this:

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{n}{k^2+n^2} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{n^2}{k^2+n^2} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{1}{(\frac{k}{n})^2+1} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(\frac{k}{n}),$$

where $f(x) = \frac{1}{x^2+1}$. Thus the limit is

$$\int_0^1 \frac{dx}{x^2+1} = \arctan 1 - \arctan 0 = \frac{\pi}{2}.$$

Anyway, we have done our homework: by establishing Theorem 8.26 we have *earned* the right to use "Darboux integral" and "Riemann integral" interchangeably. In fact however we will generally simply drop both names and simply speak of "integrable functions" and "the integral". For us, this is completely safe. However, as mentioned before, you should be aware that in more advanced mathematical analysis one studies other kinds of integrals, especially the **Lebesgue integral**.

5. Lesbesgue's Theorem

5.1. Statement of Lebesgue's Theorem.

In this section we give a characterization of the Riemann-Darboux integrable functions $f : [a,b] \to \mathbb{R}$ due to H. Lebesgue. Lebesgue's Theorem is a powerful, definitive result: many of our previous results on Riemann-Darboux integrable functions are immediate corollaries. Here we give an unusually elementary proof of Lebesgue's Theorem following lecture notes of A.R. Schep.

For an interval I, we denote by $\ell(I)$ its length: for all $-\infty < a \leq b < \infty$,

$$\ell((a,b)) = \ell([a,b)) = \ell((a,b]) = \ell([a,b]) = b-a,$$

$$\ell((a,\infty)) = \ell([a,\infty)) = \ell((-\infty,b)) = \ell((-\infty,b]) = \ell((-\infty,\infty)) = \infty.$$

We define a subset $S \subset \mathbb{R}$ to have **measure zero** if for all $\epsilon > 0$, there is a sequence $\{I_n\}$ of open intervals in $\mathbb{R}$ such that (i) for all $N \geq 1$, $\sum_{n=1}^{N} \ell(I_n) \leq \epsilon$, and (ii) $S \subset \bigcup_{i=1}^{\infty} I_n$, i.e., every point of S lies in at least one of the intervals I.

PROPOSITION 8.27. *Let $\{S_n\}_{n=1}^{\infty}$ be a sequence of subsets of $\mathbb{R}$. If each S_n has measure zero, then so does their union $S = \bigcup_{n=1}^{\infty} S_n$.*

PROOF. Fix $\epsilon > 0$ and $n \in \mathbb{Z}^+$. Since S_n has measure zero, it admits a covering by open intervals $\{I_{n,k}\}_{k=1}^{\infty}$ with $\sum_k \ell(I_{n,k}) < \frac{\epsilon}{2^{k+1}}$. Then if we collect together all the open intervals into one big set $\{I_{n,k}\}_{n,k}$ and use the fact that sums of series with positive terms do not depends on the ordering of the terms, we see that

$$\sum_{n,k} I_{n,k} = \sum_n \sum_k \ell(I_{n,k}) < \sum_n \frac{\epsilon}{2^{k+1}} = \epsilon.$$

Further, $\{I_{n,k}\}$ is a covering of $S = \bigcup_{n=1}^{\infty} S_n$ by open intervals. There is one nagging technical problem: the covering is not given by a sequence – i.e., not indexed by the positive integers – but rather by a *double sequence* – i.e., indexed by ordered

pairs (n,k) of positive integers. But this is easily solved, simply by reindexing the terms of a double sequence $a_{n,k}$ via a single sequence b_n, e.g. as follows:

$$a_{1,1}, a_{1,2}, a_{2,1}, a_{1,3}, a_{2,2}, a_{3,1}, a_{1,4}, a_{2,3}, a_{3,2}, a_{4,1}, a_{1,5} \dots .$$

□

THEOREM 8.28. *(Lebesgue) For a function $f : [a,b] \to \mathbb{R}$, let $D(f)$ be the set of $x \in [a,b]$ at which f is discontinuous. The following are equivalent:*
(i) f is Riemann-Darboux integrable on $[a,b]$.
(ii) f is bounded and $D(f)$ has measure zero.

Thus a function is Riemann-Darboux integrable on $[a,b]$ iff it is bounded (which we already know is necessary) and its set of discontinuities is "small" in the sense of having measure zero.

Exercise: Let $f, g : [a,b] \to \mathbb{R}$ be Riemann-Darboux integrable functions. Use Theorem 8.28 to give a quick proof that $f+g$ and $f \cdot g$ are both Riemann-Darboux integrable. (Suggestion: also use Proposition 8.27.)

5.2. Preliminaries on Content Zero.

The notion of sets of measure zero is a vitally important one in the further study of functions of a real variable. In fact one goes further, by assigning a measure (to be entirely technically precise, an *outer measure*) to any subset $S \subset \mathbb{R}$ as the infimum of all quantities $\sum_{n=1}^{\infty} \ell(I_n)$ as $\{I_n\}_{n=1}^{\infty}$ ranges over all sequences of open intervals with $S \subset \bigoplus_{n=1}^{\infty} I_n$. This is the beginning of a branch of real analysis called **measure theory**. In contrast, the following definition ought to be viewed as merely a technical tool used in the proof of Lebesgue's Theorem.

A subset $S \subset \mathbb{R}$ has **content zero** if for every $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ and open intervals $I_1, \dots, I_N$ such that $\sum_{i=1}^{N} \ell(I_i) \leq \epsilon$ and $S \subset \bigcup_{i=1}^{N} I_i$.

Exercise: Show that if in the definitions of measure zero and content zero we replace "open interval" by "interval", we do not change the classes of sets of measure zero or content zero.

Exercise: a) Show that $\mathbb{Q} \cap [0,1]$ has measure zero.
b) Show that $\mathbb{Q} \cap [0,1]$ does not have content zero. (Suggestion: induct on N, the number of intervals.)

LEMMA 8.29. *Let $f : [a,b] \to \mathbb{R}$ be* non-negative *and Riemann-Darboux integrable, with $\int_a^b f = 0$. Then:*
a) For all $C > 0$, the set

$$S_C = \{x \in [a,b] \mid f(x) \geq C\}$$

has content zero.
b) The set $S = \{x \in [a,b] \mid f(x) > 0\}$ has measure zero.

PROOF. Fix $\epsilon > 0$, and choose a partition $\mathcal{P}$ of $[a,b]$ such that $\mathcal{U}(f,\mathcal{P})) < \epsilon \cdot C$. Let i be the set of indices such that $S_C \cap [x_i, x_{i+1}]$ is nonempty. Thus, $i \in I \iff$

$M_i = \sup(f, [x_i, x_{i+1}]) \geq C$. It follows that

$$\epsilon \cdot C > \mathcal{U}(f, \mathcal{P}) \geq \sum_{i \in I} M_i(x_{i+1} - x_i) \geq C \sum_{i \in I} \ell([x_i, x_{i+1}]).$$

Thus $S_C \subset \bigcup_{i \in I} [x_i, x_{i+1}]$ and $\sum_{i \in I} \ell([x_i, x_{i+1}]) < \epsilon$. Thus S_C has content zero.
b) We have $S = \bigcup_{n=1}^{\infty} S_{\frac{1}{n}}$. Each $S_{\frac{1}{n}}$ has content zero, hence certainly has measure zero. Now apply Proposition 8.27. □

Exercise: Let E be a subset of $\mathbb{R}$, and let $\overline{E}$ be the set of $x \in \mathbb{R}$ such that for all $\delta > 0$, there is $y \in E$ with $|x - y| \leq \delta$. Show that E has content zero iff $\overline{E}$ has measure zero.

5.3. Proof of Lebesgue's Theorem.

The proof of Lebesgue's Theorem uses Heine-Borel (Theorem 6.17) and also the fact that a bounded monotone sequence is convergent (Theorem 10.11), which we will not discuss until Chapter 9. All in all this section should probably be omitted on a first reading, and only the extremely interested student need proceed.

Step 0: Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a, b]$. As usual, for $0 \leq i \leq n-1$, put

$$m_i = \inf(f, [x_i, x_{i+1}]), \ M_i = \sup(f, [x_i, x_{i+1}],$$

so that

$$L(f, \mathcal{P}) = \sum_{i=0}^{n-1} m_i \ell([x_i, x_{i+1}]), \ U(f, \mathcal{P}) = \sum_{i=0}^{n-1} M_i \ell([x_i, x_{i+1}]).$$

Let $\varphi, \Phi : [a, b] \to \mathbb{R}$ be the lower and upper step functions, i.e., φ takes the value m_i on $[x_i, x_{i+1})$ and Φ takes the value M_i on $[x_i, x_{i+1})$. The functions φ, Φ are bounded with only finitely many discontinuities, so are Riemann-Darboux integrable by X.X; moreover

$$\int_a^b \varphi = L(f, \mathcal{P}), \ \int_a^n \Phi = U(f, \mathcal{P}).$$

Step 1: Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann-Darboux integrable. By X.X, f is bounded, so it suffices to show that the set $D(f)$ of discontinuities of f has measure zero. Let $\{P_k\}$ be a sequence of partitions of $[a, b]$ with $\mathcal{P}_k \subset \mathcal{P}_{k+1}$ such that $|\mathcal{P}_k| \to 0$ (e.g. let $\mathcal{P}_k$ be the uniform subdivision of $[a, b]$ into 2^k subintervals). Let φ_k and Φ_k be the lower and upper step functions for $\mathcal{P}_k$, so that for all k and all $x \in [a, b)$ and

$$\int_a^b \varphi_k = \mathcal{L}(f, \mathcal{P}_k) \to \int_a^b f,$$

$$\int_a^b \Phi_k = \mathcal{U}(f, \mathcal{P}_k) \to \int_a^b f.$$

Further, since $\mathcal{P}_k \subset \mathcal{P}_{k+1}$ for all k, for all $x \in [a, b)$, the sequence $\phi_k(x)$ is increasing and bounded above, hence convergent, say to $\varphi(x)$; and similarly the sequence $\Phi_k(x)$ is decreasing and bounded below, hence convergent, say to $\Phi(x)$. For all $x \in [a, b)$, we have

$$\varphi_k(x) \leq \varphi(x) \leq f(x) \leq \Phi(x) \leq \Phi_k(x),$$

and thus

$$\int_a^b \varphi_k \leq \underline{\int_a^b} \varphi \leq \overline{\int_a^b} \varphi \leq \int_a^b f \leq \underline{\int_a^b} \Phi \leq \overline{\int_a^b} \Phi \leq \int_a^b \Phi_k.$$

These inequalities show that φ and Φ are Riemann-Darboux integrable and $\int_a^b \varphi = \int_a^b \Phi = \int_a^b f$. Applying Lemma 8.29 to $\Phi - \varphi$, we get that $\varphi = \Phi$ except on a set of measure zero. Let

$$E = \{x \in [a,b] \mid \varphi(x) \neq \Phi(x)\} \bigcup_k \mathcal{P}_k.$$

Since E is the union of a set of measure zero with a the union of a sequence of finite sets, it has measure zero. We CLAIM that f is continuous at every point of $[a,b] \setminus E$, which will be enough to complete this direction of the proof.

PROOF OF CLAIM Fix $x_0 \in [a,b] \setminus E$. Since $\varphi(x) = \Phi(x)$, there is $k \in \mathbb{Z}^+$ such that $\Phi_k(x_0) - \varphi_k(x_0) < \epsilon$. Further, since $x_0 \notin \mathcal{P}_k$, there is $\delta > 0$ such that $\Phi_k - \varphi_k$ is constant on the interval $(x_0 - \delta, x_0 + \delta)$, and for x in this interval we have

$$-\epsilon < \varphi_k(x_0) - \Phi_k(x_0) \leq f(x) - f(x_0) \leq \Phi_k(x_0) - \Phi_k(x_0) < \epsilon,$$

so f is continuous at x_0.

Step 2: Suppose now that f is bounded on $[a,b]$ and continuous on $[a,b] \setminus E$ for a subset E of measure zero. We must show that f is Riemann-Darboux integrable on $[a,b]$. Let M be such that $|f(x)| \leq M$ for all $x \in [a,b]$. Fix $\epsilon > 0$. Since E has measure zero, there is a sequence $\{I_n\}_{n=1}^\infty$ of open intervals covering E such that $\sum_{n=1}^\infty \ell(I_n) < \frac{\epsilon}{4M}$. For $x \in [a,b] \setminus E$, by continuity of f there is an open interval J_x containing x such that for all $y, z \in J_x$, $|f(y) - f(z)| \leq \frac{\epsilon}{2(b-a)}$. Applying the Heine-Borel Theorem (Theorem 6.17) to the open covering $\{I_k\} \cup \{J_x\}_{x \in [a,b] \setminus E}$ of $[a,b]$, there is a finite subcovering, say $\{I_k\}_{k=1}^N \cup \{J_{x_i}\}$. Let $\mathcal{P} = \{a = t_0 < t_1 < \ldots < t_N = b\}$ be the partition of $[a,b]$ whose points are the endpoints of $I_1, \ldots, I_N$ and the endpoints of the intervals J_{x_i} which lie in $[a,b]$. or $0 \leq j \leq N-1$, the subinterval (t_j, t_{j+1}) is contained in some I_k or some J_{x_i}. Let $\mathcal{I}$ be the set of i such that (t_i, t_{i+1}) is contained in I_k for some k. Then

$$\mathcal{U}(f,\mathcal{P}) - \mathcal{L}(f,\mathcal{P}) = \sum_{i=0}^{N-1} \ell([t_i,t_{i+1}]) \sup(f(x) - f(y) : x,y \in [t_i,t_{i+1}])$$

$$\leq \sum_{i \in \mathcal{I}} \ell([t_i,t_{i+1}]) \cdot 2M + \sum_{i \notin \mathcal{I}} \ell([t_i,t_{i+1}]) \frac{\epsilon}{2(b-a)}$$

$$< \left(\frac{\epsilon}{4M}\right) 2M + (b-a) \frac{\epsilon}{2(b-a)} = \epsilon.$$

Since $\epsilon > 0$ was arbitrary, f is Riemann-Darboux integrable on $[a,b]$.

6. Improper Integrals

6.1. Basic definitions and first examples.

6.2. Non-negative functions.

Things become much simpler if we restrict to functions $f : [a, \infty) \to [0, \infty)$: in words, we assume that f is defined *and non-negative* for all sufficiently large x. As usual we suppose that f is integrable on $[a, b]$ for all $b \geq a$, so we may define $F(x) = \int_a^x f$ for $x \geq a$. Then the improper integral $\int_a^\infty f$ is convergent iff $\lim_{x\to\infty} F(x)$ exists. But here is the point: since f is non-negative, F is *weakly increasing*: indeed, for $x_1 \leq x_2$, $F(x_2) - F(x_1) = \int_a^{x_2} f - \int_a^{x_1} f = \int_{x_1}^{x_2} \geq 0$. Now for any weakly increasing function $F : [a, \infty) \to \mathbb{R}$ we have

$$\lim_{x\to\infty} F(x) = \sup(f, [a, \infty)).$$

In other words, the limit exists as a real number iff F is bounded; otherwise, the limit is ∞: there is no oscillation! We deduce:

PROPOSITION 8.30. *Let $f : [a, \infty) \to [0, \infty)$ be integrable on every finite interval $[a, N]$ with $N \geq a$. Then either $\int_a^\infty f$ is convergent or $\int_a^\infty f = \infty$.*

In view of Proposition 8.30, we may write the two alternatives as
- $\int_a^\infty f < \infty$ (convergent)
- $\int_a^\infty f = \infty$ (divergent).

Example: Suppose we wish to compute $\int_{-\infty}^\infty e^{-x^2}$. Well, we are out of luck: this integral cannot be destroyed – ahem, I mean computed – by any craft that we here possess.[4] The problem is that we do not know any useful expression for the antiderivative of e^{-x^2} (and in fact it can be shown that this antiderivative is not an "elementary function"). But because we are integrating a non-negative function, we know that the integral is either convergent or infinite. Can we at least decide which of these alternatives is the case?

Yes we can. First, since we are integrating an even function, we have

$$\int_{-\infty}^\infty e^{-x^2} = 2\int_0^\infty e^{-x^2}.$$

Now the function $e^{-x^2} = \frac{1}{e^{x^2}}$ is approaching 0 *very rapidly*; in fact a function like $e^{-x} = \frac{1}{e^x}$ exhibits **exponential decay**, and our function is even smaller than that, at least for sufficiently large x. So it seems like a good guess that $\int_0^\infty e^{-x^2} < \infty$. Can we formalize this reasoning?

Yes we can. First, for all $x \geq 1$, $x \leq x^2$, and since $(e^x)' = e^x > 0$, e^x is increasing, so for all $x \geq 1$ $e^x \leq e^{-x^2}$, and finally, for all $x \geq 1$, $e^{-x^2} \leq e^{-x}$. By the familiar (I2) property of integrals, this gives that for all $N \geq 1$,

$$\int_1^N e^{-x^2} \leq \int_1^N e^{-x},$$

and taking limits as $N \to \infty$ we get

$$\int_1^\infty e^{-x^2} \leq \int_1^\infty e^{-x}.$$

[4]Elrond: "The Ring cannot be destroyed, Gimli, son of Glóin, by any craft that we here possess. The Ring was made in the fires of Mount Doom. Only there can it be unmade. It must be taken deep into Mordor and cast back into the fiery chasm from whence it came."

This integral is much less scary, as we know an antiderivative for e^{-x}: $-e^{-x}$. Thus

$$\int_1^\infty e^{-x^2} \leq \int_1^\infty e^{-x} = -e^{-x}|_1^\infty = -(e^{-\infty} - e^{-1}) = \frac{1}{e}.$$

Note that we replaced $\int_0^\infty e^{-x^2}$ with $\int_1^\infty e^{-x^2}$: does that make a difference? Well, yes, the difference between the two quantities is precisely $\int_0^1 e^{-x^2}$, but this is a "proper integral", hence finite, so removing it changes the value of the integral – which we don't know anyway! – but not whether it converges. However we can be slightly more quantitative: for all $x \in \mathbb{R}$, $x^2 \geq 0$ so $e^{-x^2} \leq 1$, and thus

$$\int_0^1 e^{-x^2} \leq \int_0^1 1 = 1,$$

and putting it all together,

$$\int_{-\infty}^\infty e^{-x^2} = 2\int_0^\infty e^{-x^2} = 2\left(\int_0^1 e^{-x^2} + \int_1^\infty e^{-x^2}\right) \leq 2\left(1 + \frac{1}{e}\right) = 2.735\ldots.$$

The exact value of the integral *is* known – we just don't possess the craft to find it:

$$\int_{-\infty}^\infty e^{-x^2} = \sqrt{\pi} = 1.772\ldots.$$

This is indeed smaller than our estimate, but it is a bit disappointingly far off. Later in the course we will develop methods suitable for *approximating* $\int_0^\infty e^{-x^2}$ to any desired finite degree of accuracy, and we will be able to check for instance that this integral agrees with $\sqrt{\pi}$ to, say, 20 decimal places.

The following simple theorem formalizes the argument we used above.

THEOREM 8.31. *(Comparison Test For Improper Integrals) Let $f, g : [a, \infty) \to [0, \infty)$ be integrable on $[a, N]$ for all $N \geq a$. Suppose $f(x) \leq g(x)$ for all $x \geq a$.*
a) If $\int_a^\infty g < \infty$, then $\int_a^\infty f < \infty$.
b) If $\int_a^\infty f = \infty$, then $\int_a^\infty g = \infty$.

PROOF. By property (I2) of integrals, for all $N \geq a$ since $f(x) \leq g(x)$ on $[a, N]$ we have $\int_a^N f \leq \int_a^N g$. Taking limits of both sides as $N \to \infty$ gives

$$\int_a^\infty f \leq \int_a^\infty g; \tag{51}$$

here each side is either a non-negative real number or ∞. From (51) both parts follow: if $\int_a^\infty g < \infty$ then $\int_a^\infty f < \infty$, whereas if $\int_a^\infty f = \infty$ then $\int_a^\infty g = \infty$.[5] □

THEOREM 8.32. *(Limit Comparison Test For Improper Integrals)*
Let $f, g : [a, \infty) \to [0, \infty)$ be integrable on $[a, N]$ for all $N \geq a$. Consider condition (S): there exists $b \geq a$ and $M > 0$ such that $f(x) \leq Mg(x)$ for all $x \geq b$.
a) If (S) holds and $\int_a^\infty g < \infty$, then $\int_a^\infty f < \infty$.
b) If (S) holds and $\int_a^\infty f = \infty$, then $\int_a^\infty g = \infty$.
c) Suppose there exists $b \geq a$ such that $g(x) > 0$ for all $x \geq b$ and that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L < \infty$. Then (S) holds.

[5] In fact parts a) and b) are contrapositives of each other, hence logically equivalent.

d) Suppose there exists $b \geq a$ such that $g(x) > 0$ for all $x \geq b$ and that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L$ with $0 < L < \infty$. Then

$$\int_a^\infty f < \infty \iff \int_a^\infty g < \infty.$$

PROOF. For any $b \geq a$, since f and g are integrable on $[a, b]$, we have

$$\int_a^\infty f < \infty \iff \int_b^\infty f < \infty, \int_a^\infty g < \infty \iff \int_b^\infty g < \infty.$$

a) If $f(x) \leq Mg(x)$ for all $x \geq b$, then $\int_b^\infty f \leq \int_b^\infty Mg = M\int_b^\infty g$. Thus if $\int_a^\infty g < \infty$, $\int_b^\infty g < \infty$, so $\int_b^\infty f < \infty$ and thus finally $\int_a^\infty f < \infty$.
b) Note that this is precisely the contrapositive of part a)! Or to put it in a slightly different way: suppose (S) holds. Seeking a contradiction, we also suppose $\int_a^\infty g < \infty$. Then by part a), $\int_a^f f < \infty$, contradiction.
c) Since $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L < \infty$, there is $b \geq a$ such that for all $x \geq b$, $\frac{f(x)}{g(x)} \leq L + 1$. Thus for all $x \geq b$ we have $f(x) \leq (L + 1)g(x)$, so (S) holds.
d) Suppose $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L \in (0, \infty)$. By part c), (S) holds, so by part a), if $\int_a^\infty g < \infty$, then $\int_a^\infty f < \infty$. Moreover, since $L \neq 0$, $\lim_{x\to\infty} \frac{g(x)}{f(x)} = \frac{1}{L} \in (0, \infty)$. So part c) applies with the roles of f and g reversed: if $\int_a^\infty f < \infty$ then $\int_a^\infty g < \infty$. □

Although from a strictly logical perspective part d) of Theorem 8.32 is the weakest, it is the most useful in practice.

7. Some Complements

Riemann was the first to give a complete, careful treatment of integration as a process that applies to a class of functions containing continuous functions and yields the Fundamental Theorem of Calculus: he was certainly not the last.

In more advanced analysis courses one studies the **Lebesgue integral**. This is a generalization of the Riemann integral that, extremely roughly, is modelled on the idea of approximations in small ranges of y-values rather than small ranges of x-values. Suppose for instance that $f : [0, 1] \to \mathbb{R}$ has a finite range $y_1, \ldots, y_N$. Then the idea is that $\int_a^b f$ is determined by the **size** of each of the sets $f^{-1}(y_i)$ on which f takes the value y_i. For instance this is the perspective of the expected value in probability theory: given a function defined on a probability space, its expected value is equal to the sum of each of the values at each possible outcome multipled by the probability that that outcome occurs. Notice that when f is a step function, the sets $f^{-1}(y_i)$ are just intervals, and the "size" of an interval (a, b), $[a, b)$, $(a, b]$ or $[a, b]$ is just its length $b - a$. However, a general function $f : [0, 1] \to \mathbb{R}$ with finite image need not have the sets $f^{-1}(y_i)$ be intervals, so the first and hardest step in Lebesgue's theory is a rigorous definition of the **measure** of a much more general class of subsets of $[a, b]$. This has grown into a branch of mathematics in its own right, **measure theory**, which is probably at least as important as the integration theory it yields. However, Lebesgue's integral has many technical advantages over the Riemann integral: the supply of Lebesgue integrable functions is strictly larger, and in particular is much more stable under limiting operations. Later we

will study sequences of functions $f_n : [a,b] \to \mathbb{R}$ and see that a crucial issue is the permissibility of the **interchange of limit and integral**: i.e., is

$$\lim_{n\to\infty} \int_a^b f_n = \int_a^n \lim_{n\to\infty} f_n?$$

For the Riemann integral it is difficult to prove general theorems asserting that the left hand side exists, a phenomenon which evidently makes the identity hard to prove! Probably the real showpiece of Lebesgue's theory is that useful versions of such **convergence theorems** become available without too much pain.

In classical probability theory one encounters both "discrete" and "continuous" probability distributions. The idea of a continuous probability distribution on $[a,b]$ can be modelled by a non-negative integrable function $f : [a,b] \to \mathbb{R}$ such that $\int_a^b f = 1$. This is studied briefly in the next chapter. Lebesgue's theory allows for distributions like the **Dirac delta function**, a unit mass concentrated at a single point. It turns out that one can model such discrete distributions using a simpler theory than Lebesgue's theory due to the Dutch mathematician T.J. Stieltjes (born in 1856; Riemann was born in 1826, Lebesgue in 1875). This is a relatively mild souping up of Riemann's theory which views the dx as referring to the function $f(x) = x$ and replaces it by dg for a more general function $g : [a,b] \to \mathbb{R}$. Some undergraduate texts develop this "Riemann-Stieltjes integral", most notably [**R**]. But this integral adds one extra layer of technical complication to what is already the most technical part of the course, so I feel like in a first course like this one should stick to the Riemann integral.

In the twentieth century various generalizations of the Riemann integral were developed which are much more closely related to the Riemann integral than the Lebesgue integral – avoiding any need to develop measure theory – but are yet more powerful than the Lebesgue integral: i.e., with a larger class of integral functions, good convergence theorems, and a simple general form of the Fundamental Theorem of calculus. This integral is now called the **Kurzweil-Henstock integral** although it was also developed by Denjoy, Perron and others. This theory however is not so widely known, whereas every serious student of mathematics studies the Lebesgue integral. As alluded to above, the reason is probably that the measure theory, while a lot to swallow the first time around, is in fact a highly important topic (in particular making a profound connection between analysis and geometry) in its own right. It is also extremely general, making sense in contexts far beyond intervals on the real line, whereas the Kurzweil-Henstock integral is more limited in scope.

Aside from developing his new and more sophisticated theory of integration, Lebesgue did much insightful work in the more classical context, as for instance his remarkable characterization of Riemann integrable functions presented above. I want to mention a further achievement of his which has been almost forgotten until recently: from the perspective of differential calculus, a natural – but very difficult – question is **Which functions are derivatives?** We now know that in particular every continuous function $f : [a,b] \to \mathbb{R}$ is the derivative of some other function F, but in order to construct the antiderivative we had to develop the theory of

the Riemann integral. Is it possible to construct an antiderivative of an arbitrary continuous function f in some other way than as $\int_a^b f$? (I don't mean is there some other function which is antiderivative: as we know, $\int_a^b f + C$ is the most general antiderivative of f. But for instance we certainly don't need an integration theory to know that polynomials have antiderivatives: we can just write them down. The question is whether one can find an antiderivative without expressing it in terms of an integration process.)

Lebesgue gave a considerably more elementary construction of antiderivatives of every continuous function. The basic idea is shockingly simple: on the one hand it is easy to write down an antiderivative of a piecewise linear function: it is simply an appropriate piecewise quadratic function. On the other hand every continuous function can be well approximated by piecewise linear functions, so that one can write down the antiderivative of any continuous function as a suitable limit of piecewise quadratic functions. A nice modern exposition is given in [**Be13**].

CHAPTER 9

Integral Miscellany

1. The Mean Value Therem for Integrals

THEOREM 9.1. *Let $f, g : [a, b] \to \mathbb{R}$. Suppose that f is continuous and g is integrable and non-negative. Then there is $c \in [a, b]$ such that*

$$\int_a^b fg = f(c) \int_a^b g. \tag{52}$$

PROOF. By the Extreme Value Theorem, f assumes a minimum value m and a maximum value M on $[a, b]$. Put $I = \int_a^b g$; since g is non-negative, so is I. We have

$$mI = \int_a^b mg \leq \int_a^b fg \leq \int_a^b Mg = MI. \tag{53}$$

Case 1: Suppose $I = 0$. Then by (53), for all $c \in [a, b]$ we have

$$\int_a^b fg = 0 = f(c) \int_a^b g.$$

Case 2: Suppose $I \neq 0$. Then dividing (53) through by I gives

$$m \leq \frac{\int_a^b fg}{I} \leq M.$$

Thus $\frac{\int_a^b fg}{I}$ lies between the minimum and maximum values of the continuous function $f : [a, b] \to \mathbb{R}$, so by the Intermediate Value Theorem there is $c \in [a, b]$ such that $f(c) = \frac{\int_a^b fg}{\int_a^b g}$. Multiplying through by $\int_a^b g$, we get the desired result. □

Exercise 9.1.1: Show that in the setting of Theorem 9.1, we may take $c \in (a, b)$.

Exercise 9.1.2: Show by example that the conclusion of Theorem 9.1 becomes false – even when $g \equiv 1$ – if the hypothesis on continuity of f is dropped.

Exercise 9.1.3: Suppose that $f : [a, b] \to \mathbb{R}$ is differentiable and f' is continuous on $[a, b]$. Deduce the Mean Value Theorem for f from the Mean Value Theoem for Integrals and the Fundamental Theorem of Calculus.[1]

2. Some Antidifferentiation Techniques

2.1. Change of Variables.

[1]Since the full Mean Value Theorem does not require continuity of f', this is not so exciting. But we want to keep track of the logical relations among these important theorems.

2.2. Integration By Parts.

Recall the product rule: if $f, g : I \to \mathbb{R}$ are differentiable, then so is fg and

$$(fg)' = f'g + fg'.$$

Equivalently,

$$fg' = (fg)' - f'g.$$

We have already (!) essentially derived Integration by Parts.

THEOREM 9.2. *(Integration by Parts) Let $f, g : I \to \mathbb{R}$ be continuously differentiable functions: i.e., f' and g' are defined and continuous on I.*
a) We have $\int fg' = fg - \int f'g$, in the sense that subtracting from fg any antiderivative of $f'g$ gives an antiderivative of fg'.
b) If $[a, b] \subset I$ then

$$\int_a^b fg' = f(b)g(b) - f(a)g(a) - \int_a^b f'g.$$

Exercise 9.2.1: a) Prove Theorem 9.2. (Yes, the point is that it's easy.)
b) Can the hypothesis on continuous differentiability of f and g be weakened?

Exercise 9.2.2: Use Theorem 9.2a) to find antiderivatives for the following functions.
a) $x^n e^x$ for $1 \leq n \leq 6$.
b) $\log x$.
c) $\arctan x$.
d) $x \cos x$.
e) $e^x \sin x$.
f) $\sin^6 x$.
g) $\sec^3 x$, given that $\log(\sec x + \tan x)$ is antiderivative of $\sec x$.

Exercise 9.2.3: a) Show that for each $n \in \mathbb{Z}^+$, there is a unique monic (= leading coefficient 1) polynomial $P_n(x)$ such that $\frac{d}{dx}(P_n(x)e^x) = x^n e^x$.
b) Can you observe/prove anything about the other coefficients of $P_n(x)$? (If you did part a) of the preceding exercise, you should be able to find patterns in at least three of the non-leading coefficients.)

Exercise 9.2.4: Use Theorem 9.2a) to derive the following **reduction formulas**: here $m, n \in \mathbb{Z}^+$ and $a, b, c \in \mathbb{R}$.
a)

$$\int \cos^n x = \frac{1}{n} \cos^{n-1} x \sin x + \frac{n-1}{n} \int \cos^{n-2} x.$$

b)

$$\int \frac{1}{(x^2+a^2)^n} = \frac{x}{2a^2(n-1)(x^2+a^2)^{n-1}} + \frac{2n-3}{2a^2(n-1)} \int \frac{1}{(x^2+a^2)^{n-1}}.$$

c)

$$\int \sin^m ax \cos^n ax = \frac{-1}{a(m+n)} \sin^{m-1} ax \cos^{n+1} ax + \frac{m-1}{m+n} \int \sin^{m-2} ax \cos^n ax$$

$$= \frac{1}{a(m+n)} \sin^{m+1} ax \cos^{n-1} ax + \frac{n-1}{m+n} \int \sin^m ax \cos^{n-2} ax.$$

In this text we are not much interested in antidifferentiation techniques *per se* – we regard this as being somewhat forward-looking, since computer algebra packages are by now better at this than any human ever could be – so most of our interest in integration by parts comes from part b) of Theorem 9.2. In fact this result, simple though it is, is a remarkably powerful tool throughout honors calculus and analysis.

PROPOSITION 9.3. *For all* $n \in \mathbb{N}$, $\int_0^\infty x^n e^{-x} dx = n!$.

PROOF. By induction on n. Base case ($n = 0$):
$\int_0^\infty e^{-x} = -e^{-x}|_0^\infty = -e^{-\infty} - (-e^0) = -0 - (-1) = 1 = 0!$
Induction step: let $n \in \mathbb{N}$ and assume $\int_0^\infty x^n e^{-x} dx = n!$. Now to make progress in evaluating $\int_0^\infty x^{n+1} e^{-x} dx$, we integrate by parts, taking $u = x^n + 1,\ dv = e^{-x} dx$. Then $du = (n+1)x^n dx,\ v = e^{-x}$, and

$$\int_0^\infty x^{n+1} e^{-x} dx = (n+1)x^n e^{-x}|_0^{oo} - \int_0^\infty (-e^{-x}(n+1)x^n) dx$$

$$= (0-0) + (n+1) \int_0^\infty x^n e^{-x} dx \stackrel{\text{IH}}{=} (n+1)n! = (n+1)!$$

□

2.3. Integration of Rational Functions.

3. Approximate Integration

Despite the emphasis on integration (more precisely, antidifferentiation!) techniques in a typical freshman calculus class, it is a dirty secret of the trade that in practice many functions you wish to integrate do not have an "elementary" antiderivative, i.e., one that can be written in (finitely many) terms of the elementary functions one learns about in precalculus mathematics. Thus one wants methods for evaluating definite integrals *other* than the Fundamental Theorem of Calculus. In practice it would often be sufficient to *approximate* $\int_a^b f$ rather than know it exactly.[2]

THEOREM 9.4. *(Endpoint Approximation Theorem) Let* $f : [a,b] \to \mathbb{R}$ *be differentiable with bounded derivative: there is* $M \geq 0$ *such that* $|f'(x)| \leq M$ *for all* $x \in [a,b]$. *For* $n \in \mathbb{Z}^+$, *let* $L_n(f)$ *be the* ***left endpoint Riemann sum*** *obtained by dividing* $[a,b]$ *into* n *equally spaced subintervals: thus for* $0 \leq i \leq n-1$, ${x_i}^* = a + i\left(\frac{b-a}{n}\right)$ *and* $L_n(f) = \sum_{i=0}^{n-1} f({x_i}^*)\frac{b-a}{n}$. *Then*

$$\left| \int_a^b f - L_n(f) \right| \leq \left(\frac{(b-a)^2 M}{2} \right) \frac{1}{n}$$

PROOF. Step 1: We establish the result for $n = 1$. Note that $L_1(f) = (b-a)f(a)$. By the Racetrack Principle, for all $x \in [a,b]$ we have

$$-M(x-a) + f(a) \leq f(x) \leq M(x-a) + f(a)$$

[2]It would be reasonable to argue that if one can approximate the real number $\int_a^b f$ to *any* degree of accuracy, then in some sense one does know it exactly.

and thus

$$\int_a^b (-M(x-a)+f(a))) \leq \int_a^b f \leq \int_a^b (M(x-a)+f(a)).$$

Thus

$$\frac{-M}{2}(b-a)^2+(b-a)f(a) \leq \int_a^b f \leq \frac{M}{2}(b-a)^2+(b-a)f(a),$$

which is equivalent to

$$\left|\int_a^b f - L_1(f)\right| \leq \frac{M}{2}(b-a)^2.$$

Step 2: Let $n \in \mathbb{Z}^+$. Then

$$\left|\int_a^b f - L_n(f)\right| = |\sum_{i=0}^{n-1}\left(\int_{{x_i}^*}^{x_{i+1}^*} f - f({x_i}^*)(\frac{b-a}{n})\right)| \leq \sum_{i=0}^{n-1}\left|\int_{{x_i}^*}^{x_{i+1}^*} f - f({x_i}^*)(\frac{b-a}{n})\right|$$

Step 1 applies to each term in the latter sum to give

$$\left|\int_a^b f - L_n(f)\right| \leq \sum_{n=0}^{n-1}\frac{M}{2}(\frac{b-a}{n})^2 = \left(\frac{(b-a)^2M}{2}\right)\frac{1}{n}.$$

□

Exercise 9.3.1: Show that Theorem 9.4 holds verbatim for with left endpoint sums replaced by right endpoint sums $R_n(f)$.

Exercise 9.3.2: a) Suppose $f : [a,b] \to \mathbb{R}$ is increasing. Show that

$$0 \leq \int_a^b f - L_n(f) \leq (f(b)-f(a))(b-a)\frac{1}{n}.$$

b) Derive a similar result to part a) for the right endpoint sum $R_n(f)$.
c) Derive similar results to parts a) and b) if f is decreasing.

In view of the preceding exercise there is certainly no reason to prefer left endpoint sums over right endpoint sums. In fact, if you stare at pictures of left and right endpoint sums for a while, eventually it will occur to you to take the average of the two of them: with $\Delta = \frac{b-a}{n}$, we get

(54)

$$T_n(f) = L_n(f)+R_n(f) = \frac{1}{2}f(a)+f(a+\Delta)+f(a+2\Delta)\ldots+f(a+(n-1)\Delta)+\frac{1}{2}f(b).$$

On each subinterval $[x_i, x_{i+1}]$ the average of the left and right endpoint sums is $\frac{f(x_i)(x_{i+1}-x_i)+f(x_{i+1})(x_{i+1}-x_i)}{2} = \left(\frac{f(x_i)+f(x_{i+1})}{2}\right)(x_{i+1}-x_i)$. Notice that this is the area of the **trapezoid** whose upper edge is the line segment which joins $(x_i, f(x_i))$ to $(x_{i+1}, f(x_{i+1})$. For this reason the approximation $T_n(f)$ to $\int_a^b f$ is often called the **trapezoidal rule**. That's cute, but a more useful way of thinking about $T_n(f)$ is that it is obtained by **linearly interpolating** f on each subinterval $[x_i, x_{i+1}]$ via the line which matches f at the two endpoints of the interval.

Thinking of the trapezoidal rule this way suggests that $T_n(f)$ should, at least for nice f and large n, a better approximation to $\int_a^b f$ than either the left or right

endpoint sums. For instance, if f is linear then on each subinterval we are approximating f by itself and thus $T_n(f) = \int_a^b f$. This was certainly not the case for the endpoint rules: in fact, our proof of Theorem 9.4 showed that for each fixed M, taking f to be linear with slope M (or $-M$) was the *worst case* scenario: if we approximate a linear function $\ell(x)$ with slope $\pm M$ on the interval $[a,b]$ by the horizontal line given by the left endpoint (say), then we may as well assume that $\ell(a) = 0$ and then we are approximating ℓ by the zero function, whereas $\int_a^b \ell$ is the area of the triangle with base $(b-a)$ and height $M(b-a)$, hence $\frac{M}{2}(b-a)^2$. Another motivation is that if f is increasing, then $L_n(f)$ is a lower estimate and $R_n(f)$ is an upper estimate, so averaging the two of them gives something which is closer to the true value.

The following result confirms our intuitions.

THEOREM 9.5. *(Trapezoidal Approximation Theorem) Let $f : [a,b] \to \mathbb{R}$ be twice differentiable with bounded second derivative: there is $M \geq 0$ such that $|f''(x)| \leq M$ for all $x \in [a,b]$. Then for all $n \in \mathbb{Z}^+$ we have*

$$\left| \int_a^b f - T_n(f) \right| \leq \left(\frac{(b-a)^3 M}{12} \right) \frac{1}{n^2}.$$

PROOF. Put $\Delta = \frac{b-a}{n}$, and for $i = 0, \ldots, n-1$, let $x_i = a + i\Delta$. Also put

$$\varphi_i : [0, \Delta] \to \mathbb{R}, \ \varphi_i(t) = \frac{t}{2}\left(f(x_i) + f(x_i + t)\right) - \int_{x_i}^{x_{i+1}} f.$$

As motivation for this definition we observe that

$$\sum_{i=0}^{n-1} \varphi_i(\Delta) = T_n(f) - \int_a^b f. \tag{55}$$

We have $\varphi_i(0) = 0$ and, using the Fundamental Theorem of Calculus,

$$\varphi_i' = \frac{1}{2}\left(f(x_i) + f(x_i + t)\right) + \frac{t}{2} f'(x_i + t) - f(x_i + t) = \frac{1}{2}(f(x_i) - f(x_i + t)) + \frac{t}{2} f'(x_i + t),$$

so $\varphi_i'(0) = 0$ and

$$\varphi_i'' = \frac{-1}{2} f'(x_i + t) + \frac{1}{2} f'(x_i + t) + \frac{1}{2} t f''(x_i + t) = \frac{1}{2} t f''(x_i + t).$$

Put

$$A = \inf(f'', [a,b]) \quad B = \sup(f'', [a,b]).$$

Then for all $0 \leq i \leq n-1$ and $t \in [0, \Delta]$ we have

$$\frac{A}{2} t \leq \varphi_i''(t) \leq \frac{B}{2} t.$$

Since $\varphi_i'(0) = 0$, integrating and applying the Racetrack Principle gives

$$\frac{A}{4} t^2 \leq \varphi_i'(t) \leq \frac{B}{4} t^2;$$

doing it again (since $\varphi_i(0) = 0$) and plugging in $t = \Delta$ gives

$$\frac{A}{12} \Delta^3 \leq \varphi_i(\Delta) \leq \frac{B}{12} \Delta^3.$$

Summing these inequalities from $i = 0$ to $n-1$ and using (55) we get

$$\frac{A}{12}\Delta^3 n \leq T_n(f) - \int_a^b f \leq \frac{B}{12}\Delta^3 n.$$

Substituting in $\Delta = \frac{b-a}{n}$ gives

$$\left(\frac{A(b-a)^3}{12}\right)\frac{1}{n^2} \leq T_n(f) - \int_a^b f \leq \left(\frac{B(b-a)^3}{12}\right).$$

Since $|A|, |B| \leq M$, we get the desired result:

$$\left|\int_a^b f - T_n(f)\right| \leq \left(\frac{(b-a)^3 M}{12}\right)\frac{1}{n^2}.$$

□

Exercise 9.3.3: In the setting of Theorem 9.5, suppose f'' is continuous on $[a,b]$.
a) Show: there is $c \in [a,b]$ such that $T_n(f) - \int_a^b f = \left(\frac{(b-a)^3 f''(c)}{12}\right)\frac{1}{n^2}$.
b) Show: if f is linear, the trapezoidal rule is exact: $\int_a^b f = T_n(f)$ for all $n \in \mathbb{Z}^+$.
c) Show: if $f = x^2 + bx + c$, the trapezoidal rule is never exact: $\int_a^b f \neq T_n(f)$.

Exercise 9.3.4: Show: if $f : [a,b] \to \mathbb{R}$ is convex, then $T_n(f) \geq \int_a^b f$.

Looking back at the formula (54) for the trapezoidal rule we notice that we have given equal weights to all the interior sample points but only half as much weight to the two endpoints. This is an instance of a heuristic in statistical reasoning: extremal sample points are not as reliable as interior points. This suggests a different kind of approximation scheme: rather than averaging endpoint Riemann sums, let's consider the Riemann sums in which each sample point x_i^* is the **midpoint** of the subinterval $[x_i, x_{i+1}]$. Dividing the interval $[a,b]$ into n subintervals of equal width $\Delta = \frac{b-a}{n}$ as usual, this leads us to the **midpoint rule**

$$M_n(f) = \Delta \sum_{i=0}^{n-1} f(a + (i + \frac{1}{2})\Delta).$$

Exercise 9.3.5: a) Show: if f is linear, the midpoint rule is exact: $\int_a^b f = M_n(f)$.
b) Show: if $f : [a,b] \to \mathbb{R}$ is convex, then $M_n(f) \leq \int_a^b f$.

The following result concerning the midpoint rule is somewhat surprising: it gives a sense in which the midpoint rule is twice as good as the trapezoidal rule.

THEOREM 9.6. *(Midpoint Approximation Theorem) Let $f : [a,b] \to \mathbb{R}$ be twice differentiable with bounded second derivative: there is $M \geq 0$ such that $|f''(x)| \leq M$ for all $x \in [a,b]$. Then for all $n \in \mathbb{Z}^+$ we have*

$$\left|\int_a^b f - M_n(f)\right| \leq \left(\frac{(b-a)^3 M}{24}\right)\frac{1}{n^2}.$$

PROOF. Put $\Delta = \frac{b-a}{n}$; for $i = 0, \ldots, n-1$, let $x_i = a + (i + \frac{1}{2})\Delta$. Also put

$$\Psi_i : [0, \Delta/2] \to \mathbb{R},\ \Psi_i(t) = \int_{x_i - t}^{x_i + t} f - 2tf(x_i).$$

As motivation for this definition we observe that

$$\sum_{i=0}^{n-1} \Psi(\Delta/2) = \int_a^b f - M_n(f). \tag{56}$$

We have $\Psi_i(0) = 0$, and using the Fundamental Theorem of Calculus,

$$\Psi_i'(t) = f(x_i + t) - f(x_i - t)(-1) - 2f(x_i) = (f(x_i + t) + f(x_i - t)) - 2f(x_i).$$

So $\Psi_i'(0) = 0$,

$$\Psi''(t) = f'(x_i + t) - f'(x_i - t).$$

Applying the Mean Value Theorem to f' on $[x_i - t, x_i + t]$ we get a point $x_{i,t} \in (x_i - t, x_i + t)$ such that

$$f''(x_{i,t}) = \frac{f'(x_i + t) - f'(x_i - t)}{2t} = \frac{\Psi_i''(t)}{2t},$$

or

$$\Psi_i''(t) = 2tf''(x_{i,t}).$$

Put

$$A = \inf(f'', [a, b]) \quad B = \sup(f'', [a, b]).$$

Then for all $0 \leq i \leq n - 1$ and $t \in [0, \frac{\Delta}{2}]$ we have

$$2tA \leq \Psi_i''(t) \leq 2tB.$$

Integrate and apply the Racetrack Principle twice as in the proof of Theorem 9.5, and plug in $t = \frac{\Delta}{2}$ to get

$$\frac{A}{24}\Delta^3 \leq \Psi_i(\Delta/2) \leq \frac{B}{24}\Delta^3. \tag{57}$$

Summing (57) from $i = 0$ to $n - 1$, using (56) and substituting $\Delta = \frac{b-a}{n}$ gives

$$\left(\frac{(b-a)^3 A}{24}\right)\frac{1}{n^2} \leq \int_a^b f - M_n(f) \leq \left(\frac{(b-a)^3 B}{24}\right)\frac{1}{n^2}.$$

Since $|A|, |B| \leq M$, we get the desired result:

$$\left|\int_a^b f - M_n(f)\right| \leq \left(\frac{(b-a)^3 M}{24}\right)\frac{1}{n^2}.$$

□

For integrable $f : [a, b] \to \mathbb{R}$ and $n \in \mathbb{Z}^+$, we define **Simpson's Rule**

$$S_{2n}(f) = \frac{2}{3}M_n(f) + \frac{1}{3}T_n(f).$$

Thus $S_{2n}(f)$ is a weighted average of the midpoint rule and the trapezoidal rule. Since our previous results suggest that the midpoint rule is "twice as good" as the trapezoidal rule, it makes some vague sense to weight in this way.

Exercise: a) Show that for any $a, b \in \mathbb{R}$ such that $a+b = 1$, $aM_n(f)+bT_n(f) \to \int_a^b f$. b) Deduce that $S_{2n}(f) \to \int_a^b f$.

Let us better justify Simpson's Rule. As with the other approximation rules, it

is really a *composite rule*, i.e., it is based on iterating one simple approximation on many equally spaced subintervals. So let's first concentrate on $S_2(f)$. We have

$$S_2(f) = \frac{2}{3}M_2(f) + \frac{1}{3}T_2(f) = \frac{2}{3}((b-a)f(\frac{a+b}{2})) + \frac{1}{3}(\frac{f(a)+f(b)}{2}(b-a))$$
$$= \frac{b-a}{6}\left(f(a) + 4f(\frac{a+b}{2}) + f(b)\right).$$

Since we are dividing $[a,b]$ into two subintervals, the width of each is $\Delta = \frac{b-a}{2}$, so with $x_0 = a$, $x_1 = a + \Delta$, $x_2 = a + 2\Delta = b$, we have

$$S_2(f) = \frac{\Delta}{3}(f(x_0) + 4f(x_1) + f(x_2)).$$

For $n \in \mathbb{Z}^+$, putting $\Delta = \frac{b-a}{2n}$, $x_0 = a$, $x_1 = a + \Delta, \ldots, x_{n-1}a + (2n-1)\Delta$, $x_n = b$, we get

$$S_{2n}(f) = \frac{\Delta}{3}(f(x_0) + 4f(x_1) + f(x_2) + f(x_2) + 4f(x_3) + f(x_4) + f(x_4) + \ldots + f(b)$$
$$= \frac{\Delta}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + \ldots + 4f(x_{n-1}) + f(x_n).$$

LEMMA 9.7. *For any quadratic function $f(x) = Ax^2 + Bx + C$, Simpson's Rule is exact: $\int_a^b f = S_{2n}(f)$ for all $n \in \mathbb{Z}^+$.*

PROOF. Let $\Delta = \frac{b-a}{2n}$. By splitting up $[a,b]$ into n pairs of subintervals, it suffices to show $\int_a^b f = S_2(f)$. This is done by a direct, if unenlightening, calculation:

$$\int_a^b (Ax^2 + Bx + C) = \frac{A}{3}(b^3 - a^3) + \frac{B}{2}(b^2 - a^2) + C(b-a)$$
$$= (b-a)\left(\frac{A}{3}(a^2 + ab + b^2) + \frac{B}{2}(a+b) + C)\right),$$

whereas

$$\frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$
$$= \frac{b-a}{6}\left(Aa^2 + Ba + C + 4(A\left(\frac{a+b}{2}\right)^2 + B\left(\frac{a+b}{2}\right) + C) + (Ab^2 + Bb + C)\right)$$
$$= \frac{b-a}{6}\left(A(a^2 + (a^2 + 2ab + b^2) + b^2) + B(a + 2a + 2b + b) + 6C\right)$$
$$= (b-a)\left(\frac{A}{3}(a^2 + ab + b^2) + \frac{B}{2}(a+b) + C\right).$$

□

Thus whereas the endpoint rule is an approximation by constant functions and the trapezoidal rule is an approximation by linear functions, Simpson's rule is an approximation by quadratic functions. We may therefore expect it to be more accurate than either the Trapezoidal or Midpoint Rules. The following exercise (which again, can be solved by a direct, if unenlightening, calculation), shows that it is even a little better than we might have expected.

Exercise 9.3.6: Suppose f is a cubic polynomial. Show that Simpson's Rule is exact: $S_{2n}(f) = \int_a^b f$.

Since Simpson's Rule is exact for polynomials of degree at most 3 and a function f is a polynomial of degree at most three if and only if $f^{(4)} \equiv 0$, it stands to reason that Simpson's Rule will be a better or worse approximation according to the magnitude of $f^{(4)}$ on $[a,b]$. The following result confirms this.

THEOREM 9.8. *(Simpson Approximation Theorem) Let $f : [a,b] \to \mathbb{R}$ be four times differentiable with bounded fourth derivative: there is $M \geq 0$ such that $|f^{(4)}(x)| \leq M$ for all $x \in [a,b]$. Then for all* even *$n \in \mathbb{Z}^+$ we have*

$$\left| \int_a^b f - S_n(f) \right| \leq \left(\frac{(b-a)^4 M}{180} \right) \frac{1}{n^4}.$$

PROOF. For $0 \leq i \leq \frac{n}{2} - 1$, let $x_i = a + (2i+1)\Delta$ and put

$$\Phi_i : [0, \Delta] \to \mathbb{R}, \ \Phi_i(t) = \frac{1}{3} t(f(x_i - t) + 4(x_i) + f(x_i + t)) - \int_{x_i - t}^{x_i + t} f.$$

As motivation for this definition we observe that

$$\sum_{i=0}^{n-1} \Phi_i(t) = S_n(f) - \int_a^b f. \tag{58}$$

We have $\Phi_i(0) = 0$ and using the Fundamental Theorem of Calculus,

$$\Phi_i'(t) = \frac{1}{3} t(-f'(x_i - t) + f'(x_i + t)) + \frac{1}{3}(f(x_i - t) + 4f(x_i) + f(x_i + t)) - f(x_i + t) - f(x_i - t))$$

$$= \frac{1}{3} t(f'(x_i + t) - f'(x_i - t)) - \frac{2}{3}(f(x_i - t) - 2f(x_i) + f(x_i + t)).$$

So $\Phi_i'(0) = 0$ and

$$\Phi_i''(t) = \frac{1}{3} t(f''(x_i - t) + f''(x_i + t)) + \frac{1}{3}(f'(x_i + t) - f'(x_i - t)) + \frac{2}{3} f'(x_i - t) - \frac{2}{3} f'(x_i + t)$$

$$= \frac{1}{3} t(f''(x_i + t) + f''(x_i - t)) - \frac{1}{3}(f'(x_i + t) - f'(x_i - t)).$$

So $\Phi_i''(0) = 0$ and

$$\Phi_i'''(t) = \frac{1}{3} t(f'''(x_i + t) - f'''(x_i - t)) + \frac{1}{3}(f''(x_i + t) + f''(x_i - t)) - \frac{1}{3} f''(x_i + t) - \frac{1}{3} f''(x_i - t)$$

$$= \frac{1}{3} t(f'''(x_i + t) - f'''(x_i - t)).$$

So $\Phi_i'''(0) = 0$. Applying the Mean Value Theorem to f''' on $[x_i - t, x_i + t]$, there is $\xi \in (x_i - t, x_i + t)$ such that

$$\frac{3}{2} \Phi_i''(t)/t^2 = \frac{f'''(x_i + t) - f'''(x_i - t)}{2t} = f^{(4)}(\eta).$$

Let $A = \inf(f^{(4)}, [a,b])$, $B = \sup(f^{(4)}, [a,b])$, so for all i and all $t \in [0, \Delta]$ we have

$$\frac{2A}{3} t^2 \leq \Phi_i'''(t) \leq \frac{2B}{3} t^2.$$

Integrate and apply the Racetrack Principle three times and plug in $t = \Delta$ to get

$$\frac{A}{90} \Delta^3 \leq \Phi_i(\Delta) \leq \frac{B}{90} \Delta^3. \tag{59}$$

Summing (59) from $i = 0$ to $\frac{n}{2} - 1$,[3] using (58) and substituting $\Delta = \frac{b-a}{n}$ gives

$$\left(\frac{A(b-a)^5}{180}\right)\frac{1}{n^4} \leq S_n(f) - \int_a^b f \leq \left(\frac{B(b-a)^5}{180}\right).$$

□

Exercise 9.3.7: a) In the setting of the Trapezoidal Rule, suppose moreover that f'' is continuous on $[a,b]$. Adapt the proof of Theorem 9.5 to show that there is $\eta \in [a,b]$ such that

$$T_n(f) - \int_a^b f = \left(\frac{(b-a)^3 f''(\eta)}{12}\right)\frac{1}{n^2}.$$

b) Derive similar "error equalities" for the Endpoint, Midpoint and Simpson rules.

The proofs of Theorems 9.5, 9.6 and 9.8 are taken from [**BS**, Appendix D]. They are admirably down-to-earth, using only the Mean Value Theorem and the Fundamental Theorem of Calculus. However it must be admitted that they are rather mysterious.

Our motivations for considering the various "rules" were also a little light, but I hope the reader can see that they have something to do with **polynomial interpolation**. This will be taken up in more detail in Chapter 12. For now we just mention the sobering fact that all these rules – and higher-degree analogues of them – were already known to Newton, developed in jointly with his younger collaborator Roger Cotes in the early years of the 18th century.

4. Integral Inequalities

THEOREM 9.9. *(The Hermite-Hadamard Inequality) Let $f : [a,b] \to \mathbb{R}$ be convex and continuous. Then*

$$f\left(\frac{a+b}{2}\right) \leq \frac{\int_a^b f}{b-a} \leq \frac{f(a)+f(b)}{2}.$$

PROOF. Let $s(x) = f(\frac{a+b}{2}) + m(x - \frac{a+b}{2})$ be a supporting line for f at $x = \frac{a+b}{2}$, and let $S(x) = f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a)$ be the secant line, so

$$\forall x \in [a,b],\ s(x) \leq f(x) \leq S(x). \tag{60}$$

Integrating these inequalities and dividing by $b-a$ we get

$$\frac{\int_a^b s}{b-a} \leq \frac{\int_a^b f}{b-a} \leq \frac{\int_a^b S}{b-a}.$$

Now we have

$$\frac{\int_a^b s(x)}{b-a} = \frac{\int_a^b f(\frac{a+b}{2}) + m(x - \frac{a+b}{2})}{b-a} = f(\frac{a+b}{2}) + \frac{m}{b-a}\int_a^b (x - \frac{a+b}{2})$$

$$= f(\frac{a+b}{2}) + \frac{m}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2} - \frac{a+b}{2}(b-a)\right) = f(\frac{a+b}{2}) + \frac{m}{b-a}\cdot 0 = \frac{a+b}{2},$$

[3]Note that there are $\frac{n}{2}$ terms here – half as many as in the previous results. This gives rise to an extra factor of $\frac{1}{2}$.

$$\frac{\int_a^b S(x)dx}{b-a} = \frac{\int_a^b (f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a))dx}{b-a} = \frac{f(a)+f(b)}{2}.$$

Substituting these evaluations of $\frac{\int_a^b s}{b-a}$ and $\frac{\int_a^b S}{b-a}$ into (60) gives

$$f(\frac{a+b}{2}) \leq \frac{\int_a^b f}{b-a} \leq \frac{f(a)+f(b)}{2}.$$

□

Exercise: Show that the hypothesis of continuity in Theorem 9.9 is not necessary: the inequality holds for any convex $f : [a,b] \to \mathbb{R}$.

Let $[a,b]$ be a closed interval, and let $P : [a,b] \to [0,\infty)$ be a **probability density**: i.e., P is integrable on $[a,b]$ and $\int_a^b P = 1$. For an integrable function $f : [a,b] \to \mathbb{R}$, we define the **expected value**

$$E(f) = \int_a^b f(x)P(x)dx.$$

THEOREM 9.10. *(Jensen's Integral Inequality) Let $P : [a,b] \to [0,\infty)$ be a probability density function, $\varphi : [c,d] \to \mathbb{R}$ be convex, and let $f : [a,b] \to [c,d]$ be integrable. Then $E(f) \in [c,d]$ and*

$$\varphi(E(f)) \leq E(\varphi(f)).$$

PROOF. Since $f : [a,b] \to [c,d]$, we have

$$c = \int_a^b cP(x)dx \leq \int_a^b f(x)P(x) \leq \int_a^b dP(x) = d,$$

so $\int_a^b f(x)P(x)dx = E(f) \in [c,d]$ and thus $\varphi(E(f))$ is defined. Now put $x_0 = E(f)$ and let $s(x) = mx + B$ be a supporting line for the convex function φ at x_0, so $s(x) \leq \varphi(x)$ for all $x \in [c,d]$ and $s(x_0) = \varphi(x_0)$. Now

$$E(\varphi(f)) = \int_a^b \varphi(f(x))P(x)dx \geq \int_a^b (mf(x)+B)P(x)dx$$

$$= m\int_a^b f(x)P(x)dx + B = mE(f) + B = mx_0 + b = \varphi(x_0) = \varphi(E(f)).$$

□

THEOREM 9.11 (Hölder's Integral Inequality). *Let $p,q \in \mathbb{R}^{>0}$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Let $f,g : [a,b] \to \mathbb{R}$ be Riemann integrable functions. Then*

$$\left|\int_a^b fg\right| \leq \left(\int_a^b |f|^p\right)^{\frac{1}{p}} \left(\int_a^b |g|^q\right)^{\frac{1}{q}}. \tag{61}$$

PROOF. Step 1: Suppose $\int_a^b |f|^p = 0$. Let M_f and M_g be upper bound for f and g on $[a,b]$. By Lemma 8.29, the set of $x \in [a,b]$ such that $|f|^p > \epsilon$ has content zero: thus, for every $\delta > 0$ there is a finite collection of subintervals of $[a,b]$, of total length at most δ, such that on the complement of those subintervals $|f|^p \leq \epsilon$. On this complement, $|fg| \leq \epsilon^{\frac{1}{p}} M_g$ and thus the sum of the integrals of $|fg|$ over the

complement is at most $(b-a)\epsilon^{\frac{1}{p}}M_g$. The sum of the Riemann integrals over the subintervals of total length δ of $|fg|$ is at most $\delta M_f M_g$, so altogether we get

$$|\int_a^b fg| \leq \int_a^b |fg| \leq (b-a)\epsilon^{\frac{1}{p}}M_g + \delta M_f M_g.$$

Since all the other quantities are fixed and ϵ, δ are arbitrary, we deduce $|\int_a^b fg| = 0$. Thus both sides of (61) are zero in this case and the inequality holds. A similar argument works if $\int_a^b |g|^q = 0$. Henceforth we may suppose

$$I_f = \int_a^b |f|^p > 0, \quad I_g = \int_a^b |g|^q > 0.$$

Step 2: Put $\tilde{f} = f/I_f^{\frac{1}{p}}$ and $\tilde{g} = g/I_g^{\frac{1}{q}}$. Then by Young's Inequality,

$$|\int_a^b \tilde{f}\tilde{g}| \leq \int_a^b |\tilde{f}||\tilde{g}| \leq \int_a^b \left(\frac{|\tilde{f}|^p}{p} + \frac{|\tilde{g}|^q}{q}\right) = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying through by $I_f^{\frac{1}{p}} I_g^{\frac{1}{q}} = \left(\int_a^b |f|^p\right)^{\frac{1}{p}} \left(\int_a^b |g|^q\right)^{\frac{1}{q}}$ gives the desired result. □

Remark: Admittedly Step 1 of the above proof is rather technical. To an extent this points to the finickiness of the class of Riemann integrable functions. The student who continues on will learn the corresponding result for the Lebesgue integral, in which the proof of this step is immediate. Alternately, the student may wish to restrict to continuous f and g, since a non-negative continuous function f with $\int_a^b f = 0$ must be identically zero.

The case $p = q = 2$ of Theorem 9.11 is important enough to be restated as a result in its own right.

COROLLARY 9.12 (Cauchy-Schwarz Integral Inequality). *Let $f, g : [a,b] \to \mathbb{R}$ be Riemann integrable functions. Then*

$$\left(\int_a^b fg\right)^2 \leq \int_a^b f^2 \int_a^b g^2.$$

5. The Riemann-Lebesgue Lemma

THEOREM 9.13. *(Riemann-Lebesgue Lemma) Let $f : [a,b] \to \mathbb{R}$ be Riemann integrable. Then:*
a) $\lim_{\lambda\to\infty} \int_a^b f(x)\cos(\lambda x)dx = 0$.
b) $\lim_{\lambda\to\infty} \int_a^b f(x)\sin(\lambda x)dx = 0$.

PROOF. As the reader surely expects, the arguments for parts a) and b) are virtually identical, so we will establish part a) and leave part b) as an exercise.

Here is the *idea* of the proof: if $f(x) \equiv C$ is a constant function, then $\int_a^b C\cos(\lambda x)$ oscillates increasingly rapidly as λ increases. If the length of the interval $a-b$ were a precise multiple of the period $\frac{2\pi}{\lambda}$ then we get several "complete sine waves" and the positive and negative area cancel out exactly, making the integral exactly zero. As λ increases the remaining "incomplete sine wave" is a bounded function living on a smaller and smaller subinterval, so its contribution to the integral goes to zero.

A similar argument holds for step functions, which become constant functions when split up into appropriate subintervals. And now the key: the integrable functions are *precisely* those which can be well approximated by step functions in the sense that the integral of the difference can be made arbitrarily small.

Okay, let's see the details: fix $\epsilon > 0$. By Darboux's Criterion, there is a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$ such that

$$0 \leq L(f, \mathcal{P}) \leq \int_a^b f < \frac{\epsilon}{2}.$$

Let g be the step function which is constantly equal to $m_i = \inf(f, [x_i, x_{i+1}])$ on the subinterval $[x_i, x_{i+1})$ of $[a, b]$, so $g \leq f$ and $\int_a^b g = L(f, \mathcal{P})$, so

$$0 \leq \int_a^b (f - g) \leq \frac{\epsilon}{2}.$$

Now

$$\left| \int_a^b f(x) \cos(\lambda x) dx \right| \leq \int_a^b |f(x) - g(x)||\cos(\lambda x)| dx + \left| \int_a^b g(x) \cos(\lambda x) dx \right|$$

$$\leq \frac{\epsilon}{2} + \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} m_i \cos(\lambda x) dx \right| \leq \frac{\epsilon}{2} + \left| \sum_{i=0}^{n-1} \frac{m_i}{\lambda} \left(\sin(\lambda x_{i+1}) - \sin(\lambda x_i) \right) \right|. \tag{62}$$

Here we have a lot of expressions of the form $|\sin(A) - \sin(B)|$ for which an obvious upper bound is 2. Using this, the last expression in (62) is at most

$$\frac{\epsilon}{2} + \frac{2 \sum_{i=0}^{n-1} |m_i|}{\lambda}.$$

But this inequality holds for any $\lambda > 0$, so taking λ sufficiently large we can make the last term at most $\frac{\epsilon}{2}$ and thus $|\int_a^b f(x) \cos(\lambda x) dx| < \epsilon$. □

CHAPTER 10

Infinite Sequences

Let X be a set. An **infinite sequence** in X is given by a function $x_\bullet : \mathbb{Z}^+ \to X$. Less formally but entirely equivalently, we are getting an ordered infinite list of elements of X: $x_1, x_2, x_3, \ldots, x_n, \ldots$. Note that the function is not required to be injective: i.e., we may have $x_i = x_j$ for $i \neq j$. In fact, a simple but important example of a sequence is a **constant sequence**, in which we fix some element $x \in X$ and take $x_n = x$ for all n.

The notion of an infinite sequence in a general set X really is natural and important throughout mathematics: for instance, if $X = \{0, 1\}$ then we are considering infinite sequences of binary numbers, a concept which comes up naturally in computer science and probability theory, among other places. But here we will focus on **real infinite sequences** $x_{bullet} : \mathbb{Z}^+ \to \mathbb{R}$. In place of $x_\bullet$, we will write $\{x_n\}_{n=1}^\infty$ or even, by a slight abuse of notation, x_n.

We say that an infinite sequence a_n **converges** to a real number L if for all $\epsilon > 0$ there exists a positive integer N such that for all $n \geq N$, we have $|a_n - L| < \epsilon$. A sequence is said to be **convergent** if it converges to some $L \in \mathbb{R}$ and otherwise **divergent**. Further, we say a sequence a_n **diverges to infinity** – and write $\lim_{n\to\infty} a_n = \infty$ or $a_n \to \infty$ – if for all $M \in \mathbb{R}$ there exists $N \in \mathbb{Z}^+$ such that $n \geq N \implies a_n > M$. Finally, we define divergence to negative infinity: I leave it to you to write out the definition.

This concept is strongly reminiscent of that of the limit of a function $f : [1, \infty) \to \mathbb{R}$ as x approaches infinity. In fact, it is more than reminiscent: there is a direct connection. If $\lim_{x\to\infty} f(x) = L$, then if we form the sequence $x_n = f(n)$, then it follows that $\lim_{n\to\infty} x_n = L$. If $x_n = f(x)$ for a function f which is continuous – or better, differentiable – then the methods of calculus can often be brought to bear to analyze the limiting behavior of x_n.

Given a sequence $\{x_n\}$, we say that a function $f : [1, \infty) \to \mathbb{R}$ **interpolates** f if $f(n) = x_n$ for all $n \in \mathbb{Z}^+$.

Example: Supose $x_n = \frac{\log n}{n}$. Then $f(x) = \frac{\log x}{x}$ interpolates the sequence, and

$$\lim_{x\to\infty} \frac{\log x}{x} = \frac{\infty}{\infty} \stackrel{\text{LH}}{=} \lim_{x\to\infty} \frac{\frac{1}{x}}{1} = \frac{1}{\infty} = 0.$$

It follows that $x_n \to 0$.

Exercise: Let $\{a_n\}_{n=1}^\infty$ be a real sequence. Define $f : [1, \infty) \to \mathbb{R}$ as follows: for $x \in [n, n+1)$, $f(x) = (n + 1 - x)a_n + (x - n)a_{n+1}$.

a) Take, for example, $a_n = n^2$, and sketch the graph of f.
b) Show that the sequence $\lim_{x\to\infty} f(x)$ exists iff the sequence $\{a_n\}$ is convergent, and if so the limits are the same.

The previous exercise shows that in principle *every* infinite sequence $\{a_n\}$ can be interpolated by a continuous function. The given f is **piecewise linear** but generally not differentiable at integer values. However, with only a little more trouble we could "round off the corners" and find a differentiable function f which interpolates $\{a_n\}$. But in practice this is only useful if the interpolating function f is simple enough to have a known limiting behavior at infinity. Many sequences which come up in practice cannot be interpolated *in a useful way*.

Example (Fibonacci Sequence)

Example (Newton's method sequences).

In fact we will be most interested in *sequences of finite sums*. For instance, let

$$a_n = \sum_{i=1}^{n} \frac{1}{i} = 1 + \frac{1}{2} + \ldots + \frac{1}{n},$$

and let

$$b_n = \sum_{i=1}^{n} \frac{1}{i^2} = 1 + \frac{1}{2^2} + \ldots + \frac{1}{n^2}.$$

What is the limiting behavior of a_n and b_n? In fact it turns out that $a_n \to \infty$ and $b_n \to \frac{\pi^2}{6}$: whatever is happening here is rather clearly beyond the tools we have at the moment! So we will need to develop new tools.

1. Summation by Parts

LEMMA 10.1. *(Summation by Parts) Let $\{a_n\}$ and $\{b_n\}$ be two sequences. Then for all $m \leq n$ we have*

$$\sum_{k=m}^{n} a_k(b_{k+1} - b_k) = (a_{n+1}b_{n+1} - a_m b_m) + \sum_{k=m}^{n} (a_{k+1} - a_k)b_{k+1}.$$

PROOF.

$$\sum_{k=m}^{n} a_k(b_{k+1} - b_k) = a_m b_{m+1} + \ldots + a_n b_{n+1} - (a_m b_m + \ldots + a_n b_n)$$

$$= a_n b_{n+1} - a_m b_m - ((a_{m+1} - a_m)b_{m+1} + \ldots + (a_n - a_{n-1})b_n)$$

$$= a_n b_{n+1} - a_m b_n - \sum_{k=m}^{n-1} (a_{k+1} - a_k)b_{k+1}$$

$$= a_n b_{n+1} - a_m b_n + (a_{n+1} - a_n)b_{n+1} - \sum_{k=m}^{n} (a_{k+1} - a_k)b_{k+1}$$

$$= a_{n+1}b_{n+1} - a_m b_m - \sum_{k=m}^{n} (a_{k+1} - a_k)b_{k+1}. \qquad \square$$

The proof of Lemma 10.1 could hardly be more elementary or straightforward: literally all we are doing is regrouping some finite sums. Nevertheless the human mind strives for understanding and rebels against its absence: without any further explanation, summation by parts would seem (to me at least) very mysterious.

The point is that Lemma 10.1 is a discrete analogue of integration by parts:

$$\int_a^b fg' = f(b)g(b) - f(a)g(a) - \int_a^b f'g.$$

Just as integration by parts is a fundamental tool in the study of integrals – and, after taking limits, improper integrals – summation by parts is a fundamental tool in the study of finite sums – and, after taking limits, infinite series.

Whereas for integration by parts there really is just the one identity, recalled above, for summation by parts, there are several alternate formulas and special cases, one of which is given in the following exercise. So really one should remember the idea – starting with a sum of the form $\sum_n a_n b_n$, there are several identities involving expressions in which we have discretely differentiated $\{a_n\}$ and discretely integrated $\{b_n\}$, or vice versa – rather than any one specific formula.

Exercise: Let $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ be sequences, and for $N \in \mathbb{Z}^+$, put $A_n = \sum_{i=1}^n a_n$, $B_n = \sum_{i=1}^n b_n$, $A_0 = B_0 = 0$. Show that

$$\sum_{n=1}^N a_n b_n = \sum_{n=1}^{N-1} A_n(b_n - b_{n+1}) + A_N b_N. \tag{63}$$

Proposition 10.2. *(Abel's Lemma) Let $\{a_n\}_{n=1}^\infty$ be a real sequence with the following property: there is an $M > 0$ such that $|a_1 + \ldots + a_N| \leq M$ for all positive integers N. Let $\{b_n\}_{n=1}^\infty$ be a real sequence such that $b_1 \geq b_2 \geq \ldots \geq b_n \ldots \geq 0$. Then, for all $N \in \mathbb{Z}^+$, $|\sum_{n=1}^N a_n b_n| \leq b_1 M$.*

Proof. Using (63), we have

$$|\sum_{n=1}^N a_b b_n| = |\sum_{n=1}^{N-1} A_n(b_n - b_{n+1}) + A_N b_N|$$

$$\leq \sum_{n=1}^{N-1}(b_n - b_{n+1})|A_N| + b_N|A_N| = b_1|A_N| \leq b_1 M.$$

□

We will put Abel's Lemma to use...but only much later on. For now though you might try to prove it *without* using the summation by parts variant (63): by doing so, you'll probably gain some appreciation that these formulas, though in some sense trivial, can be used in distinctly nontrivial ways.

2. Easy Facts

The following result collects some easy facts about convergence of infinite sequences.

Theorem 10.3. *Let $\{a_n\}$, $\{b_n\}, \{c_n\}$ be real infinite sequences.*
*a) If $a_n = C$ for all n – a **constant sequence** – then $a_n \to C$.*
b) The limit of a convergent sequence is unique: if for $L_1, L_2 \in \mathbb{R}$ we have $a_n \to L_1$

and $a_n \to L_2$, then $L_1 = L_2$.
c) If $a_n \to L$ and $b_n \to M$ then:
(i) For all $C \in \mathbb{R}$, $Ca_n \to CL$.
(ii) $a_n + b_n \to L + M$.
(iii) $a_n b_n \to LM$.
(iv) If $M \neq 0$, $\frac{a_n}{b_n} \to \frac{L}{M}$.
d) If $a_n \leq b_n$ for all n, $a_n \to L$ and $b_n \to M$, then $L \leq M$.
e) If $a \leq b$ are such that $a_n \in [a, b]$ for all n and $a_n \to L$, then $L \in [a, b]$.
f) (Three Sequence Principle) Suppose $c_n = a_n + b_n$. Then it is not possible *for exactly two of the three sequences $\{a_n\}, \{b_n\}, \{c_n\}$ to be convergent: if any two are convergent, then so is the third.*
g) Suppose there exists $N \in \mathbb{Z}^+$ such that $b_n = a_n$ for all $n \geq N$. Then for any $L \in [-\infty, \infty]$, $a_n \to L \iff b_n \to L$.

Most of these facts are qiute familiar, and the ones that may not be are routine. In fact, every part of Theorem 10.3 holds verbatim for functions of a continuous variable approaching infinity. Hence one method of proof would be to establish these for functions – or maintain that we have known these facts for a long time – and then apply the Sequence Interpolation Theorem. But be honest with yourself: for each part of Theorem 10.3 for which you have an iota of doubt as to how to prove, please take some time *right now* to write out a careful proof.

We say that a sequence $\{a_n\}$ is **eventually constant** if there is $C \in \mathbb{R}$ and $N \in \mathbb{Z}^+$ such that $a_n = C$ for all $n \geq N$. It is easy to see that if such a C exists then it is unique, and we call such a C the **eventual value** of the sequence. Of course an eventually constant sequence converges to its eventual value – e.g. by applying parts a) and g) of Theorem 10.3, but really this is almost obvous in any event.

PROPOSITION 10.4. *Let $\{a_n\}$ be an infinite sequence with values in* the integers *$\mathbb{Z}$. Then a_n is convergent iff it is eventually constant.*

PROOF. As above, it is clear that an eventually constant sequence is convergent. Conversely, suppose $a_n \to L \in \mathbb{R}$. First we claim that $L \in \mathbb{Z}$. If not, the distance from L to the nearest integer is a positive number, say ϵ. But since $a_n \to L$, there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|a_n - L| < \epsilon$. But the interval $(L - \epsilon, L + \epsilon)$ contains no integers: contradiction.

Now take $\epsilon = 1$ in the definition of convergence: there is $N \in \mathbb{Z}^+$ such that for $n \geq N$, we have $|a_n - L| < 1$, and since a_n and L are both integers this implies $a_n = L$. Thus the sequence is eventually constant with eventual value L. □

Proposition 10.4 goes a long way towards explaining why we have described a function from $\mathbb{Z}^+$ to $\mathbb{R}$ as *semi*-discrete. A function from $\mathbb{Z}^+$ to $\mathbb{Z}^+$ is "fully discrete", and thus the limiting behavior of such functions is very limited.

Exercise: A subset $S \subset \mathbb{R}$ is **discrete** if for all $x \in S$, there is $\epsilon > 0$ such that the only element of S which lies in $(x - \epsilon, x + \epsilon)$ is x.
a) Which of the following subsets of $\mathbb{R}$ are discrete?
(i) A finite set.
(ii) The integers $\mathbb{Z}$.
(iii) The rational numbers $\mathbb{Q}$.
(iv) The set $\{\frac{1}{n} \mid n \in \mathbb{Z}^+\}$ of reciprocals of positive integers.

(v) The set $\{\frac{1}{n} \mid n \in \mathbb{Z}^+\} \bigcup \{0\}$.
b) For a subset $S \subset \mathbb{R}$, show that the following are equivalent:
(i) S is discrete.
(ii) Every convergent sequence $\{a_n\}$ with $a_n \in S$ for all n is eventually constant.

It is often convenient to consider sequences whose initial term is something other than 1. It is certainly no problem to entertain sequences starting at any integer N_0: informally, they look like

$$a_{N_0}, a_{N_0+1}, \ldots.$$

Formally, instead of a function from $\mathbb{Z}^+$ to $\mathbb{R}$, we have a function from $\{n \in \mathbb{Z} \mid n \geq N_0\}$ to $\mathbb{R}$. This mild generalization changes nothing we have said so far or will say later. We leave it to the reader to make her own peace with this.

3. Characterizing Continuity

THEOREM 10.5. *Let I be an interval, let c be an interior point of I, and let $f : I \to \mathbb{R}$ be a function. The following are equivalent:*
(i) f is continuous at c.
(ii) For every infinite sequence $a_n \to c$, we have $f(a_n) \to f(c)$.

PROOF. (i) $\implies$ (ii): This argument is very similar to the (easy!) proof that a composition of continuous functions is continuous. Namely, fix $\epsilon > 0$. Since f is continuous at c, there is $\delta > 0$ such that $|x - c| < \delta \implies |f(x) - f(c)| < \epsilon$. Moreover, since $a_n \to c$, there exists $N \in \mathbb{Z}^+$ such that $n \geq N \implies |a_n - c| < \delta$. So if $n \geq N$ we have $|a_n - c| < \delta$ and thus $|f(a_n) - f(c)| < \epsilon$.
(ii) $\implies$ (i): We prove the contrapositive: that is, we will suppose that f is *not* continuous at c and find a sequence $a_n \to c$ but such that $f(a_n)$ does not converge to $f(c)$. If f is *not* continuous at c then there exists $\epsilon > 0$ such that for all $\delta > 0$, there is a_δ with $|a_\delta - c| < \delta$ and $|f(a_\delta) - f(c)| \geq \epsilon$. In particular, for $n \in \mathbb{Z}^+$ we may take $\delta = \frac{1}{n}$ and choose a_n with $|a_n - c| < \frac{1}{n}$ and $|f(a_n) - f(c)| \geq \epsilon$, and then indeed we have $a_n \to c$ but $f(a_n)$ does not approach $f(c)$. □

Spivak gives a slightly more general result, the proof of which (essentially the same as the one given above) we leave as an exercise.

THEOREM 10.6. *Let I be an interval, c an interior point of I, and let $f : I \setminus \{c\} \to \mathbb{R}$ be a function such that $\lim_{x \to c} f(x) = L$. Let $\{a_n\}_{n=1}^\infty$ be a real sequence such that for all $n \in \mathbb{Z}^+$, $a_n \in I \setminus \{c\}$ and $\lim_{n \to \infty} a_n = c$. Then*

$$\lim_{n \to \infty} f(a_n) = L.$$

Exercise: Prove Theorem 10.6.

4. Monotone Sequences

One commonality between sequences $a : \mathbb{Z}^+ \to \mathbb{R}$ and non-discrete functions $f : I \to \mathbb{R}$ is that – since indeed both the domain and codomain are subsets of $\mathbb{R}$ – we have a natural order structure $\leq$ and it makes sense to consider functions which systematically preserve – or systematically reverse! – this order structure. Thus the following definitions will be quite familiar.

A sequence $\{a_n\}$ is **weakly increasing** if for all $m, n \in \mathbb{Z}^+$, $m \leq n \implies a_m \leq a_n$.

A sequence $\{a_n\}$ is **increasing** if for all $m, n \in \mathbb{Z}^+$, $m < n \implies a_m < a_n$.

A sequence $\{a_n\}$ is **weakly decreasing** if for all $m, n \in \mathbb{Z}^+$, $m \leq n \implies a_m \geq a_n$.

A sequence $\{a_n\}$ is **decreasing** if for all $m, n \in \mathbb{Z}^+$, $m < n \implies a_m > a_n$.

A sequence is **monotone** if it is weakly increasing or weakly decreasing.

Remark: As before, we need to counsel the reader that the terminology here is not agreed upon: for instance, many use the term "strictly increasing" for sequences with $m < n \implies a_m < a_n$. If you are ever in doubt whether someone means $m \leq n \implies a_m \leq a_n$ or $m < n \implies a_m < a_n$, there is a simple remedy: ask!

Exercise: Let $\{a_n\}$ be a sequence.
a) Show that $\{a_n\}$ is increasing iff $a_n < a_{n+1}$ for all $n \in \mathbb{Z}^+$.
b) Formulate (and, if you feel it is worth your time, prove) analogues of part a) with *increasing* replaced by each of *weakly increasing*, *decreasing*, *weakly decreasing*.

Exercise: For a real sequence $\{a_n\}$, show that the following are equivalent:
(i) There is $n \in \mathbb{Z}^+$ such that either $a_n < a_{n+1} > a_{n+2}$ (Λ-**configuration**) or $a_n > a_{n+1} < a_{n+2}$ (**V-configuration**).
(ii) The sequence $\{a_n\}$ is *not monotone*.

LEMMA 10.7. *For a real sequence $\{a_n\}$, the following are equivalent:*
(i) $\{a_n\}$ is both weakly increasing and weakly decreasing.
(ii) $\{a_n\}$ is constant.

Exercise: Prove it.

LEMMA 10.8. *(Reflection Principle) Let $\{a_n\}$ be a real sequence. Then:*
a) $\{a_n\}$ is decreasing iff $\{-a_n\}$ is increasing.
b) $\{a_n\}$ is weakly increasing iff $\{-a_n\}$ is weakly increasing.

Exercise: Prove it.

Just as for functions of a continuous variable, Lemma 10.8 implies that whatever we can establish for increasing / weakly increasing sequences will carry over immediately to decreasing / weakly decreasing sequences (and conversely, of course).

What a monotone sequence lacks is *oscillation*, and as for functions of a continuous variable, this implies a much simpler limiting behavior. Especially, if $a_\bullet : \mathbb{Z}^+ \to \mathbb{R}$ is weakly increasing, there is really only *one way* for $\lim_{n\to\infty} a_n$ to fail to exist.

To speak of this it is convenient to make another definition: since a sequence a_n is really a function $\mathbb{Z}^+ \to \mathbb{R}$, it makes sense to consider its *image*, i.e., the set of all real numbers of the form a_n for some $n \in \mathbb{Z}^+$. Strangely, there is not really a standard name for this: I vaguely recall it being called the "trace" of the sequence, but the term "trace" has many other, unrelated meanings in mathematics so this is really not optimal. We avoid the problem by giving the concept a very clunky name: we define the **term set** of $\{a_n\}$ to be $A = \{a_n \in \mathbb{R} \mid n \in \mathbb{Z}^+\}$.

We are therefore able to speak of bounded sequences just as for bounded functions: i.e., in terms of the image...um, I mean the term set.

A sequence $a_\bullet : \mathbb{Z}^+ \to \mathbb{R}$ is **bounded above** if its term set is bounded above: that is, if there exists $M \in \mathbb{R}$ such that $a_n \leq M$ for all $n \in \mathbb{Z}^+$. Otherwise we say the sequence is **unbounded above**. Similarly, we say $a_\bullet$ is **bounded below** if its term set is bounded below: that is, if there exists $m \in \mathbb{R}$ such that $m \leq a_n$ for all $n \in \mathbb{Z}^+$. Otherewise we say the sequence is **unbounded below**. Finally, a sequenc is **bounded** if it is both bounded above and bounded below, and a sequence is **unbounded** if it is not bounded.

PROPOSITION 10.9. *Let $\{a_n\}_{n=1}^\infty$ be a weakly increasing sequence.*
a) If the sequence converges to $L \in \mathbb{R}$, then L is the least upper bound of the term set $A = \{a_n \mid n \in \mathbb{Z}^+\}$.
b) Conversely, if the term set A has an upper bound $L < \infty$, then $a_n \to L$.

PROOF. a) First we claim $L = \lim_{n\to\infty} a_n$ is an upper bound for the term set A. Indeed, suppose not: then there is $N \in \mathbb{Z}^+$ with $L < a_N$. But since the sequence is weakly increasing, this implies that for all $n \geq N$, $L < a_N \leq a_n$. Thus if we take $\epsilon = a_N - L$, then for no $n \geq N$ do we have $|a_n - L| < \epsilon$, contradicting our assumption that $a_n \to L$. Second we claim L is the least upper bound. Indeed, suppose not: then there is L' such that for all $n \in \mathbb{Z}^+$, $a_n \leq L' < L$. Let $\epsilon = L - L'$. For no n do we have $|a_n - L| < \epsilon$, contradicting our asumption that $a_n \to L$.
b) Let $\epsilon > 0$. We need to show that for all but finitely many $n \in \mathbb{Z}^+$ we have $-\epsilon < L - a_n < \epsilon$. Since L is the least upper bound of A, in particular $L \geq a_n$ for all $n \in \mathbb{Z}^+$, so $L - a_n \geq 0 > -\epsilon$. Next suppose that there are infinitely many terms a_n with $L - a_n \geq \epsilon$, or $L \geq a_n + \epsilon$. But if this inequality holds for ifninitely many terms of the sequence, then because a_n is increasing, it holds for all terms of the sequence, and this implies that $L - \epsilon \geq a_n$ for all n, so that $L - \epsilon$ is a smaller upper bound for A than L, contradiction. □

Remark: In the previous result we have *not* used the completeness property of $\mathbb{R}$, and thus it holds for sequences with values in the rationals $\mathbb{Q}$ (and where by *converges* we mean *converges to a rational number*!) or really in any ordered field. By combining this with the least upper bound axiom, we get a much stronger result.

THEOREM 10.10. *Let $\{a_n\}_{n=1}^\infty$ be a weakly increasing real sequence. Let $L \in (-\infty, \infty]$ be the least upper bound of the term set of A. Then $a_n \to L$.*

This is so important as to be worth spelling out very carefully. We get:

THEOREM 10.11. *(Monotone Sequence Theorem) a) Every bounded monotone real sequence is convergent. More precisely:*
b) Let $\{a_n\}$ be weakly increasing. Then if $\{a_n\}$ is bounded above, it converges to its least upper bound, and if it is unbounded above it diverges to ∞.
c) Let $\{a_n\}$ be weakly decreasing. Then if $\{a_n\}$ is bounded below, it converges to its greatest lower bound, and it is unbounded below it diverges to $-\infty$.

In fact, in proving the Monotone Sequence Theorem we did not just invoke the completeness of the real field: we used its full force, in the following sense.

THEOREM 10.12. *Let F be an ordered field in which every bounded monotone sequence converges. Then F is Dedekind complete: every nonempty bounded above subset has a least upper bound.*

PROOF. Step 1: We CLAIM that F is Archimedean.
PROOF OF CLAIM: Suppose not: then the sequence $x_n = n$ is increasing and bounded above. Suppose that it were convergent, say to $L \in F$. By Proposition 10.9, L is the least upper bound of $\mathbb{Z}^+$. But this is absurd: if $n \leq L$ for all $n \in \mathbb{Z}^+$ then $n+1 \leq L$ for all $n \in \mathbb{Z}^+$ and thus $n \leq L-1$ for all $n \in \mathbb{Z}^+$, so $L-1$ is a smaller upper bound for $\mathbb{Z}^+$.
Step 2: Let $S \subset \mathbb{R}$ be nonempty and bounded above by M_0.
CLAIM For all $n \in \mathbb{Z}^+$, there exists $y_n \in S$ such that for any $x \in S$, $x \leq y_n + \frac{1}{n}$.
Proof of claim: Indeed, first choose any element z_1 of S. If for all $x \in S$, $x \leq z_1 + \frac{1}{n}$, then we may put $y_n = z_1$. Otherwise there exists $z_2 \in S$ with $z_2 > z_1 + \frac{1}{n}$. If for all $x \in S$, $x \leq z_2 + \frac{1}{n}$, then we may put $y_n = z_2$. Otherwise, there exists $z_3 \in S$ with $z_3 > z_2 + \frac{1}{n}$. If this process continues infinitely, we get a sequence with $z_k \geq z_1 + \frac{k-1}{n}$. But by Step 1, F is Archimedean, so that for sufficiently large k, $z_k > M$, contradiction. Therefore the process musts terminate and we may take $y_n = z_k$ for sufficiently large k.
Now we define a sequence of upper bounds $\{M_n\}_{n=1}^{\infty}$ of S as follows: for all $n \in \mathbb{Z}^+$, $M_n = \min(M_{n-1}, y_n + \frac{1}{n})$. $\{M_n\}$ is decreasing and bounded below by any element of S, so by hypothesis it converges, say to M, and by the reflected version of Proposition 10.9, M is the infimum of the set $\{M_n\}$. Moreover M must be the supremum of S, since again by the Archimedean nature of the order, for any $m < M$, for sufficiently large n we have $m + \frac{1}{n} < M \leq M_n \leq y_n + \frac{1}{n}$ and thus $m < y_n$. □

5. Subsequences

A real infinite sequence $\{a_n\}$ is, informally, an orderest list of real numbers:

$$a_1, a_2, \ldots, a_n, \ldots$$

A subsequence of $\{a_n\}$, is – again informally, for a little while; a precise definition will be coming soon – obtained by selecting some infinitely many terms of the sequence, e.g

$$a_1, a_3, a_6, a_{100}, a_{703}, \ldots.$$

(There is not meant to be a clearly evident pattern as to which terms of the sequence we are choosing: the point is that there does not need to be.) To learn more about subsequences and what they are good for, let us look at some key examples.

Example: Let $a_n = (-1)^n$, so the sequence is

$$-1, 1, -1, 1, \ldots \tag{64}$$

By selecting all the odd-numbered terms we get one subsequence:

$$-1, -1, -1, -1, \ldots \tag{65}$$

Similarly, by selecting all the even-numberd terms we get another subsequence:

$$1, 1, 1, 1, \ldots. \tag{66}$$

There are other choices – *many* other choices. In fact, a real sequence can be obtained as a subsequence of $\{a_n\}$ iff it takes values in $\{\pm 1\}$.

But note that something very interesting happened in the passage from our original sequence to each of the first two subsequences. The original sequence (64) is not convergent, due to oscillation. However, the subsequence (65) is constant, hence convergent to it constant value -1. Similarly, the subsequence (66) converges to its constant value 1.

Let's recap: we began with a sequence (66) which did not converge due to oscillation. However, by choosing *appropriate* subsequences we were able to remove the oscillation, resulting – in this case, at least – in a convergent subsequence. (There are also lots of subsequences which are "inappropriate" for this purpose.)

Example: Let $a_n = n$, so the sequence is

$$1, 2, 3, 4, \ldots \tag{67}$$

Here are some subsequences we can get:

$$1, 3, 5, 7, \ldots \tag{68}$$

$$2, 4, 6, 8, \ldots \tag{69}$$

$$1, 4, 9, 16, \ldots \tag{70}$$

$$1, 2, 4, 8, \ldots \tag{71}$$

And so forth. Indeed, the subsequences of (67) are precisely the increasing sequences with values in the positive integers. Note moreover that our sequence (67) fails to converge, but *not* due to oscillation. It is an increasing sequence which is not unbounded above, and thus it diverges to infinity. For that matter, so do the subsequences (68), (69), (70), (71), and a little thought suggests that every subsequence will have this property. Thus, passing to a subsequence can cure divergence due to oscillation, but not divergence to infinity.

Example (subsequences of a convergent sequence):

We are now well-prepared for the formal definition. In fact, we practically saw it in the example above. Given a real sequence $\{a_n\}$, we view it as a function $a_\bullet : \mathbb{Z}^{\to}\mathbb{R}$, $n \mapsto a_n$. To obtain a subsequence we choose an increasing sequence $n_\bullet : \mathbb{Z}^+ \to \mathbb{Z}^+$, $k \mapsto n_k$ and form the composite function

$$a_\bullet \circ n_\bullet : \mathbb{Z}^+ \to \mathbb{R},\ k \mapsto a_{n_k}.$$

Less formally, we choose an increasing list $n_1 < n_2 < \ldots < n_k$ of positive integers and use this to tell us which terms of the sequence to take, getting

$$a_{n_1}, a_{n_2}, a_{n_3}, \ldots, a_{n_k}, \ldots.$$

Let's formalize these observations about what passing to subsequences does for us.

Exercise: Let $n_\bullet : \mathbb{Z}^+ \to \mathbb{Z}^+$ be increasing. Show that for all $k \in \mathbb{Z}^+$, $n_k \geq k$.

PROPOSITION 10.13. *Let $\{a_n\}$ be a real sequence, $L \in [-\infty, \infty]$, and suppose that $a_n \to L$. Then for any subsequence $\{a_{n_k}\}_{k=1}^{\infty}$, we have $a_{n_k} \to L$.*

PROOF. Case 1: Suppose that $L \in \mathbb{R}$. Since $a_n \to L$, for each $\epsilon > 0$ there exists $N = N(\epsilon) \in \mathbb{Z}^+$ such that for all $n \geq N$, $|a_n - L| < \epsilon$. Then, by Exercise X.X, for all $k \geq N$ we have $n_k \geq N$ and thus $|a_{n_k} - L| < \epsilon$.
Case 2: Suppose $L = \infty$. Since $a_n \to \infty$, for each $M \in \mathbb{R}$ there exists $N = N(M) \in \mathbb{Z}^+$ such that for all $n \geq N$, $a_n > M$. If $k \geq N$, then $n_k \geq k \geq N$ and thus $a_{n_k} > M$.
Case 3: We leave the case of $L = -\infty$ to the reader as a (not very challenging) exercise in modifying the argument of Case 2. □

PROPOSITION 10.14. *Every subsequence of a monotone sequence is monotone. More precisely:*
a) If $\{a_n\}$ is weakly increasing, then every subsequence a_{n_k} is weakly increasing.
b) If $\{a_n\}$ is increasing, then every subsequence a_{n_k} is increasing.
c) If $\{a_n\}$ is weakly decreasing, then every subsequence a_{n_k} is weakly decreasing.
d) If $\{a_n\}$ is decreasing, then every subsequence a_{n_k} is decreasing.
e) If $\{a_n\}$, is constant, then every subsequence a_{n_k} is constant.

PROOF. a) If $k_1 \leq k_2$, then by definition of a subsequence $n_{k_1} \leq n_{k_2}$, and then by definition of weakly increasing, $a_{n_{k_1}} \leq a_{n_{k_2}}$.
b),c),d),e) These may safely be left to the reader. □

COROLLARY 10.15. *For a montone sequence $\{a_n\}$, the following are equivalent:*
(i) $\{a_n\}$ is convergent.
(ii) $\{a_n\}$ is bounded.
(iii) Every subsequence of $\{a_n\}$ is convergent.
(iv) At least one subsequence of $\{a_n\}$ is convergent.

6. The Bolzano-Weierstrass Theorem For Sequences

6.1. The Rising Sun Lemma.

I learned of the following result from Mr. Evangelos Kobotis in late 1994, in my first quarter of college at the University of Chicago. Because of its elegance, generality and usefulness, it has stayed with me through my entire adult mathematical career. It seems that this argument goes back to a short note of Newman and Parsons [**NP88**].

LEMMA 10.16. *Every infinite sequence has a monotone subsequence.*

PROOF. Let us say that $m \in \mathbb{N}$ is a **peak** of the sequence $\{a_n\}$ if for all $n > m$, $a_n < a_m$. Suppose first that there are infinitely many peaks. Then any sequence of peaks forms a strictly decreasing subsequence, hence we have found a monotone subsequence. So suppose on the contrary that there are only finitely many peaks, and let $N \in \mathbb{N}$ be such that there are no peaks $n \geq N$. Since $n_1 = N$ is not a peak, there exists $n_2 > n_1$ with $a_{n_2} \geq a_{n_1}$. Similarly, since n_2 is not a peak, there exists $n_3 > n_2$ with $a_{n_3} \geq a_{n_2}$. Continuing in this way we construct an infinite (not necessarily strictly) increasing subsequence $a_{n_1}, a_{n_2}, \ldots, a_{n_k}, \ldots$. Done! □

Exercise: Show that every infinite sequence admits a subsequence which is (i) increasing, (ii) decreasing, or (iii) constant.

Remark: Although it is nothing to do with our development of *infinite sequences*, we cannot resist mentioning the following **finite analogue** of the Rising Sun Lemma.

THEOREM 10.17. *(Erdős-Szekeres [**ES35**]) Let $r, s \in \mathbb{Z}^+$.*
a) Consider a finite real sequence $x_1, \ldots, x_{(r-1)(s-1)+1}$ of length $(r-1)(s-1)+1$.
(i) There is a weakly increasing subsequence of length r, or
(ii) There is a weakly decreasing subsequence of length s.
b) The number $(r-1)(s-1)+1$ is best possible, in the sense that there are real sequences of length $(r-1)(s-1)$ without either a weakly increasing subsequence of length r or a weakly decreasing subsequence of length s.

PROOF. a) ([**Se59**], [**Er94**]) Seeking a contradiction, we suppose every weakly increasing subsequence has length at most $r-1$ and every weakly decreasing subsequence has length at most $s-1$. We define a function $F : \{1, \ldots, (r-1)(s-1)+1\} \to \{1, \ldots, r-1\} \times \{1, \ldots, s-1\}$ by $F(j) = (i_j, d_j)$, where i_j is the length of the longest weakly increasing subsequence beginning with x_j and d_j is the length of the longest weakly decreasing subsequence beginning with x_j. Since the domain of F has $(r-1)(s-1)+1$ elements and the codomain has $(r-1)(s-1)$ elements, there must be $j < k$ such that $F(j) = F(k)$: thus $i_j = i_k$ and $d_j = d_k$. But if $x_j \leq x_k$ then adding x_j to the beginning of the longest weakly increasing sequence starting with x_k gives a longer weakly increasing sequence starting at x_j, a contradiction; similarly, if $x_j \geq x_k$, then adding x_j to the beginning of the longest weakly decreasing sequence starting with x_k gives a longer weakly decreasing sequence starting at x_j, a contradiction. b) Consider the length $(r-1)(s-1)$ sequence

$$r, r-1, \ldots, 1, 2r, 2r-1, \ldots, r+1, 3r, 3r-1, \ldots, 2r+1, \ldots, sr, sr-1, \ldots, (s-1)r+1.$$

□

6.2. Bolzano-Weierstrass for Sequences.

THEOREM 10.18. *(Bolzano-Weierstrass) Every bounded real sequence admits a convergent subsequence.*

PROOF. Let $\{a_n\}$ be a sequence with $|a_n| \leq M$ for all $n \in \mathbb{Z}^+$. By the Rising Sun Lemma, there is a monotone subsequence a_{n_k}, which of course is also bounded: $|a_{n_k}| \leq M$ for all $k \in \mathbb{Z}^+$. By Corollary 10.15, the subsequence a_{n_k} is convergent. □

Exercise: Show that for an ordered field F, the following are equivalent:
(i) Every bounded monotone sequence converges.[1]
(ii) Every bounded sequence admits a convergent subsequence.

6.3. Supplements to Bolzano-Weierstrass.

THEOREM 10.19. *a) A real sequence which is unbounded above admits a subsequence diverging to ∞.*
b) A real sequence which is unbounded below admits a subsequence diverging to $-\infty$.

PROOF. We will prove part a) and leave the task of adapting the argument to prove part b) to the reader. Let $\{x_n\}$ be a real sequence which is unbounded above. Then for every $M \in \mathbb{R}$, there exists at least one n such that $x_n \geq M$. Let n_1 be the least positive integer such that $x_{n_1} > 1$. Let n_2 be the least positive integer such that $x_{n_2} > \max(x_{n_1}, 2)$. And so forth: having defined n_k, let n_{k+1} be the least positive integer such that $x_{n_{k+1}} > \max(x_{n_k}, k+1)$. Then $\lim_{k\to\infty} x_{n_k} = \infty$. □

[1]Recall that we have already shown that this is equivalent to Dedekind completeness.

6.4. Applications of Bolzano-Weierstrass.

We will now give two quite different proofs of two of the three Interval Theorems.

THEOREM 10.20. *(Extreme Value Theorem Again) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is bounded and attains its minimum and maximum values.*

PROOF. Seeking a contradiction we suppose f is *unbounded*: then, for each $n \in \mathbb{Z}^+$, there exists $x_n \in [a, b]$ with $|f(x_n)| > n$. The sequence $\{x_n\}$ takes values in $[a, b]$ so is bounded, so by Bolzano-Weierstrass there is a subsequence x_{n_k} such that $x_{n_k} \to L \in \mathbb{R}$. Now, on the one hand, since f is continuous we have by Theorem 10.5 that $f(x_{n_k}) \to f(L) \in \mathbb{R}$. On the other hand, we have for all k that $|f(x_{n_k})| > k$, so the sequence $f(x_{n_k})$ is divergent: contradiction!

We do not have a new argument for the attainment of the minimum and maximum values, but the original argument was sufficiently short and sweet that we don't mind repeating it here: let M be the supremum of the set of $f(x)$ for $x \in [a, b]$, so by what we just proved, $M < \infty$. If M is not attained then the function $f(x) - M$ is continuous and nowhere zero, so the function $g : [a, b] \to \mathbb{R}$ by $g(x) = \frac{1}{f(x)-M}$ is continuous on $[a, b]$, hence bounded. But this is impossible: by definition of M, $f(x)$ takes values arbitrarily close to M, so $f(x) - M$ takes values arbitrarily close to zero and thus $|g(x)|$ take arbitrarily large values. By reflection, f attains its minimum as well. □

THEOREM 10.21. *(Uniform Continuity Theorem Again) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is uniformly continuous.*

PROOF. Suppose not: then there exists $\epsilon > 0$ such that for all $n \in \mathbb{Z}^+$ we have $x_n, y_n \in [a, b]$ with $|x_n - y_n| < \frac{1}{n}$ but $|f(x_n) - f(y_n)| \geq \epsilon$. By Bolzano-Weierstrass, after passing to a subsequence we may assume x_{n_k} is convergent, say to L, Since $x_{n_k} - y_{n_k} \to 0$, by the Three Sequence Principle y_{n_k} is also convergent to L. But then $f(x_{n_k}) \to f(L)$ and $f(y_{n_k}) \to f(L)$, and since the sequences $f(x_{n_k})$ and $f(y_{n_k})$ have the same limit, we can make $|f(x_{n_k}) - f(y_{n_k})| \leq |f(x_{n_k}) - L - (f(y_{n_k})) - L)| \leq |f(x_{n_k}) - L| + |f(y_{n_k}) - L|$ arbitrarily small by taking n sufficiently large. In particular, for sufficiently large n we have $|f(x_{n_k}) - f(y_{n_k})| < \epsilon$: contradiction! □

Remark: Although I am very fond of the Real Induction proof of the Uniform Continuity Theorem, I must admit this argument seems shorter and easier. Very often in mathematics if one takes the time to develop additional, seemingly unrelated technology – in this case, the theory of infinite sequences – one is richly rewarded with the ability to come up with shorter, simpler (but more "high tech") proofs.

6.5. Bolzano-Weierstrass for Subsets Revisited.

Recall that earlier we proved a different result which we also called the Bolzano-Weierstrass Theorem: it concerned the notion of a limit point of a subset $X \subset \mathbb{R}$: a point $a \in \mathbb{R}$ is a **limit point** for X if for all $\epsilon > 0$, there is $x \in \mathbb{R}$ with $0 < |a - x| < \epsilon$. It was an easy observation that any finite subset of $\mathbb{R}$ has no limit points. We proved that the converse holds for bounded sets:

THEOREM 10.22. *(Bolzano-Weierstrass for Subsets) An infinite subset X of a closed, bounded interval $[a, b]$ has a limit point.*

Our task now is to explain why we have two different results called the Bolzano-Weierstrass Theorem. In fact they are really *equivalent* results, which means roughly that is much easier to deduce each from the other than it is to prove either one. Indeed:

Assume the Bolzano-Weierstrass Theorem for sequences, and let X be an infinite subset of $[a, b]$. Since X is infinite, we can choose a sequence $\{x_n\}_{n=1}^{\infty}$ of distinct elements of X. Since $x_n \in X \subset [a, b]$, we have $a \leq x_n \leq b$ for all n; in particular, $\{x_n\}$ is bounded, so by Bolzano-Weierstrass for sequences there is a subsequence x_{n_k} converging to some $L \in [a, b]$. We claim that L is a limit point for X: indeed, for any $\epsilon > 0$, there is $K \in \mathbb{Z}^+$ such that for all $k \geq K$, $|x_{n_k} - L| < \epsilon$: since the terms are distinct, at most one of them is equal to L and thus the interval $(L - \epsilon, L + \epsilon)$ contains infinitely many elements of X.

Assume the Bolzano-Weierstrass Theorem for subsets, and let $\{x_n\}$ be a bounded sequence: thus there are $a, b \in \mathbb{R}$ with $a \leq x_n \leq b$ for all $n \in \mathbb{Z}^+$. Let $X = \{x_n \mid n \in \mathbb{Z}^+\}$ be the term set of the sequence. If X is finite, then the sequence has a constant (hence convergent) subsequence. Otherwise X is infinite and we may apply Bolzano-Weierstrass for subsets to get a limit point L of X. This implies: there is $n_1 \in \mathbb{Z}^+$ such that $|x_{n_1} - L| < 1$; having chosen such an n_1, there is $n_2 \in \mathbb{Z}^+$ such that $n_2 > n_1$ and $|x_{n_2} - L| < \frac{1}{2}$: continuing in this way we build a subsequence $\{x_{n_k}\}$ such that for all $k \in \mathbb{Z}^+$, $|x_{n_k} - L| < \frac{1}{k}$, and thus $x_{n_k} \to L$.

7. Partial Limits; Limits Superior and Inferior

7.1. Partial Limits.

For a real sequence $\{a_n\}$, we say that an extended real number $L \in [-\infty, \infty]$ is a **partial limit** of $\{a_n\}$ if there exists a subsequence a_{n_k} such that $a_{n_k} \to L$.

LEMMA 10.23. *Let $\{a_n\}$ be a real sequence. Suppose that L is a partial limit of some subsequence of $\{a_n\}$. Then L is also a partial limit of $\{a_n\}$.*

Exercise: Prove Lemma 10.23. (Hint: this comes down to the fact that a subsequence of a subsequence is itself a subsequence.)

THEOREM 10.24. *Let $\{a_n\}$ be a real sequence.*
a) $\{a_n\}$ has at least one partial limit $L \in [-\infty, \infty]$.
b) The sequence $\{a_n\}$ is convergent iff it has exactly one partial limit L and L is finite*, i.e., $L \neq \pm\infty$.*
c) $a_n \to \infty$ iff ∞ is the only partial limit.
d) $a_n \to -\infty$ iff $-\infty$ is the only partial limit.

PROOF. a) If $\{a_n\}$ is bounded, then by Bolzano-Weierstrass there is a finite partial limit L. If $\{a_n\}$ is unbounded above, then by Theorem 10.19a), ∞ is a partial limit. It $\{a_n\}$ is unbounded below, then by Theorem 10.19b) $-\infty$ is a partial limit. Every sequence is either bounded, unbounded above or unbounded below (and the last two are not mutually exclusive), so there is always at least one partial limit.
b) Suppose that $L \in \mathbb{R}$ is the unique partial limit of $\{a_n\}$. We wish to show that $a_n \to L$. First observe that by Theorem 10.19, $\{a_n\}$ must be bounded above and below, for otherwise it would have an infinite partial limit. So choose $M \in \mathbb{R}$ such

that $|a_n| \leq M$ for all n.

Now suppose that a_n *does not converge* to L: then there exists $\epsilon > 0$ such that *it is not the case* that there exists $N \in \mathbb{N}$ such that $|a_n - L| < \epsilon$ for all $n \geq N$. What this means is that there are infinitely many values of n such that $|a_n - L| \geq \epsilon$. Moreover, since $|a_n - L| \geq \epsilon$ means *either* $-M \leq a_n \leq L - \epsilon$ or $L + \epsilon \leq a_n \leq M$, there must in fact be in infinite subset $S \subset \mathbb{N}$ such that either for all $n \in S$ we have $a_n \in [-M, L - \epsilon]$ or for all $n \in S$ we have $a_n \in [L + \epsilon, M]$.

Let us treat the former case. The reader who understands the argument will have no trouble adapting to the latter case. Writing the elements of S in increasing order as $n_1, n_2, \ldots, n_k$, we have shown that there exists a subsequence $\{a_{n_k}\}$ all of whose terms lie in the closed interval $[-M, L - \epsilon]$. Applying Bolzano-Weierstrass to this subsequence, we get a subsubsequence (!) $a_{n_{k_\ell}}$ which converges to some L'. We note right away that a subsubsequence of a_n is also a subsequence of a_n: we still have an infinite subset of $\mathbb{N}$ whose elements are being taken in increasing order. Moreover, since every term of $a_{n_{k_\ell}}$ is bounded above by $L - \epsilon$, its limit L' must satisfy $L' \leq L - \epsilon$. But then $L' \neq L$ so the sequence has a second partial limit L': contradiction.

c) Suppose $a_n \to \infty$. Then also every subsequence diverges to $+\infty$, so $+\infty$ is a partial limit and there are no other partial limits. We will prove the converse via its contrapositive (the inverse): suppose that a_n does not diverge to ∞. Then there exists $M \in \mathbb{R}$ and infinitely many $n \in \mathbb{Z}^+$ such that $a_n \leq M$, and from this it follows that there is a subsequence $\{a_{n_k}\}$ which is bounded above by M. This subsequence does not have $+\infty$ as a partial limit, hence by part a) it has some partial limit $L < \infty$. By Lemma 10.23, L is also a partial limit of the original sequence, so it is not the case that $+\infty$ is the only partial limit of $\{a_n\}$.

d) Left to the reader to prove, by adapting the argument of part c) or otherwise. □

Exercise: Let $\{x_n\}$ be a real sequence. Suppose that:
(i) Any two convergent subsequences converge to the same limit.
(ii) $\{x_n\}$ is bounded.
Show that $\{x_n\}$ is convergent. (Suggestion: Combine Theorem 10.24b) with the Bolzano-Weierstrass Theorem.)

Exercise: Let $\{x_n\}$ be a real sequence, and let $a \leq b$ be extended real numbers. Suppose that there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $a \leq x_n \leq b$. Show that $\{a_n\}$ has a partial limit L with $a \leq L \leq b$.

7.2. The Limit Supremum and Limit Infimum.

For a real sequence $\{a_n\}$, let $\mathcal{L}$ be the set of all partial limits of $\{a_n\}$.

We define the **limit supremum** $\overline{L}$ of a real sequence to be the least upper bound of the set of all partial limits of the sequence.

THEOREM 10.25. *For any real sequence $\{a_n\}$, $\overline{L}$ is a partial limit of the sequence and is thus the largest partial limit.*

PROOF. Case 1: The sequence is unbounded above. Then $+\infty$ is a partial limit, so $\overline{L} = +\infty$ is a partial limit.

Case 2: The sequence diverges to $-\infty$. Then $-\infty$ is the only partial limit and thus $\overline{L} = -\infty$ is the largest partial limit.

Case 3: The sequence is bounded above and does not diverge to $-\infty$. Then it has a finite partial L (it may or may not also have $-\infty$ as a partial limit), so $\overline{L} \in (-\infty, \infty)$. We need to find a subsequence converging to $\overline{L}$.

For each $k \in \mathbb{Z}^+$, $\overline{L} - \frac{1}{k} < \overline{L}$, so there exists a subsequence converging to some $L' > \overline{L} - \frac{1}{k}$. In particular, there exists n_k such that $a_{n_k} > \overline{L} - \frac{1}{k}$. It follows from these inequalities that the subsequence a_{n_k} cannot have any partial limit which is less than $\overline{L}$; moreover, by the definition of $\overline{L} = \sup \mathcal{L}$ the subsequence cannot have any partial limit which is strictly greater than $\overline{L}$: therefore by the process of elimination we must have $a_{n_k} \to \overline{L}$. □

Similarly we define the **limit infimum** $\underline{L}$ of a real sequence to be the infimum of the set of all partial limits. By reflection, the proof of Theorem 10.25 shows that $\underline{L}$ is a partial limit of the sequence, i.e., there exists a subsequence a_{n_k} such that $a_{n_k} \to \underline{L}$.

Here is a very useful characterization of the limit supremum of a sequence $\{a_n\}$ it is the unique extended real number L such that for any $M > L$, $\{n \in \mathbb{Z}^+ \mid a_n \geq M\}$ is finite, and such that for any $m < L$, $\{n \in \mathbb{Z}^+ \mid \ a_n \geq m\}$ is infinite.

Exercise:
a) Prove the above characterization of the limit supremum.
b) State and prove an analogous characterization of the limit infimum.

PROPOSITION 10.26. *For any real sequence a_n, we have*

$$\overline{L} = \lim_{n \to \infty} \sup_{k \geq n} a_k \tag{72}$$

and

$$\underline{L} = \lim_{n \to \infty} \inf_{k \geq n} a_k. \tag{73}$$

Because of these identities it is traditional to write $\limsup a_n$ *in place of* $\overline{L}$ *and* $\liminf a_n$ *in place of* $\underline{L}$.

PROOF. As usual, we will prove the statements involving the limit supremum and leave the analogous case of the limit infimum to the reader.

Our first order of business is to show that $\lim_{n \to \infty} \sup_{k \geq n} a_k$ exists as an extended real number. To see this, define $b_n = \sup_{k \geq n} a_k$. The key observation is that $\{b_n\}$ is decreasing. Indeed, when we pass from a set of extended real numbers to a subset, its supremum either stays the same or decreases. By Theorem 10.21, $b_n \to L' \in [-\infty, \infty]$.

Now we will show that $\overline{L} = L'$ using the characterization of the limit supremum stated above. First suppose $M > L'$. Then there exists $n \in \mathbb{Z}^+$ such that $\sup_{k \geq n} a_k < M$. Thus there are only finitely many terms of the sequence which are at least M, so $M \geq \overline{L}$. It follows that $L' \geq \overline{L}$.

On the other hand, suppose $m < L'$. Then there are infinitely many $n \in \mathbb{Z}^+$ such that $m < a_n$ and hence $m \leq \overline{L}$. It follows that $\overline{L} \leq L'$, and thus $\overline{L} = L'$. □

The merit of these considerations is the following: if a sequence converges, we have a number to describe its limiting behavior, namely its limit L. If a sequence diverges

to $\pm\infty$, again we have an "extended real number" that we can use to describe its limiting behavior. But a sequence can be more complicated than this: it may be highly oscillatory and therefore its limiting behavior may be hard to describe. However, to every sequence we have now associated two numbers: the limit infimum $\underline{L}$ and the limit supremum $\overline{L}$, such that

$$-\infty \leq \underline{L} \leq \overline{L} \leq +\infty.$$

For many purposes – e.g. for making upper estimates – we can use the limit supremum $\overline{L}$ in the same way that we would use the limit L if the sequence were convergent (or divergent to $\pm\infty$). Since $\overline{L}$ exists *for any sequence*, this is very powerful and useful. Similarly for $\underline{L}$.

COROLLARY 10.27. *A real sequence $\{a_n\}$ is convergent iff $\underline{L} = \overline{L} \in (-\infty, \infty)$.*

Exercise: Prove Corollary 10.27.

8. Cauchy Sequences

8.1. Motivating Cauchy sequences.

Let us look back at our definition of a convergent sequence: a real sequence $\{a_n\}$ is **convergent** if there is some $L \in \mathbb{R}$ such that $\{a_n\}$ converges to L.

There is something slightly curious about this definition: do you see it? It is this: we are defining the notion "convergent" in terms of the notion "convergent to L". One might think that things should go the other way around: shouldn't the concept of convergene should be somehow logically prior than "convergence to L"? Anyway, this is the way we tend to think of things in practice: if one is given an infinite sequence, *first* we want to know whether it is convergent or divergent, and then if we know it is convergent we can ask the more refined question: "To *which* real number L does our convergent sequence converge?"

This is a good opportunity to think back on our other notions of convergence.

If you think about it, our definition of "the limit of $f(x)$ as $x \to c$" also has this feature: what we actually defined was $\lim_{x\to c} f(x) = L$. For continuity the issue is not pressing because we know that we want the limit to be $f(c)$.[2] The concept of uniform continuity really does not mention a specific limiting value, i.e., a function is not "uniformly continuous to L at c".

Since derivatives are defined in terms of limits (and not in terms of continuity!) they have the same problem: to show that f is differentiable at c we have to show that the limiting slope of the secant line *is* some specific real number.

For integrals, the plot thickens. Our first, awkward, definition involved a fixed real number I. Then we gave a modified definition in terms of two quantities $\underline{\int_a^b f}$ and $\overline{\int_a^b f}$ which exist for any function, and our definition of integrability was that these two quantities are finite and equal. This is sort of in between in that the definition includes a particular limiting value but we don't have to "find" it: we have to show that an underestimate and an overestimate are actually the same. Moreover we have Darboux's criterion, which remedies the issue completely: this gives us a procedure to show that a function is integrable *without requiring any knowledge on*

[2]I think this partially explains why the concept of continuity is simpler than that of a limit.

what number the integral should be. And that's what made Darboux's criterion so useful: we used it to show that every continuous function and every monotone function is integrable, but of course without having to "find" in any sense the value of the integral. (This inexplicitness is not entirely a good thing, and the main point of our discussion of Riemann sums was to make the convergence more explicit.)

Upshot: it would be nice to have some way of expressing/proving that a sequence is convergent which doesn't have the limit of the sequence built into it. This is exactly what Cauchy sequences are for.

8.2. Cauchy sequences.

A sequence $\{a_n\}_{n=1}^{\infty}$ in is **Cauchy** if for all $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$, $|a_m - a_n| < \epsilon$.

Here are some elementary properties of Cauchy sequences.[3]

LEMMA 10.28. *A subsequence of a Cauchy sequence is Cauchy.*

PROOF. Left to the reader as an exercise. □

PROPOSITION 10.29. *A convergent sequence is Cauchy.*

PROOF. Suppose $a_n \to L$. Then there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|a_n - L| < \frac{\epsilon}{2}$. Thus for all $m, n \geq N$ we have

$$|a_n - a_m| = |(a_n - L) - (a_m - L)| \leq |a_n - L| + |a_m - L| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

PROPOSITION 10.30. *A Cauchy sequence is bounded.*

PROOF. Let $\{a_n\}$ be a Cauchy sequence. There exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$, $|a_m - a_n| < 1$. Therefore, taking $m = N$ we get that for all $n \geq N$, $|a_n - a_N| < 1$, so $|a_n| \leq |a_N| + 1 = M_1$, say. Moreover put $M_2 = \max_{1 \leq n \leq N} |a_n|$ and $M = \max(M_1, M_2)$. Then for all $n \in \mathbb{Z}^+$, $|a_n| \leq M$. □

PROPOSITION 10.31. *A Cauchy sequence which admits a convergent subsequence is itself convergent.*

PROOF. Let $\{a_n\}$ be a Cauchy sequence and suppose there is a subsequence a_{n_k} converging to $L \in F$. We claim that a_n converges to L. Fix any $\epsilon > 0$. Choose $N_1 \in \mathbb{Z}^+$ such that for all $m, n \geq N_1$ we have $|a_n - a_m| = |a_m - a_n| < \frac{\epsilon}{2}$. Further, choose $N_2 \in \mathbb{Z}^+$ such that for all $k \geq N_2$ we have $|a_{n_k} - L| < \frac{\epsilon}{2}$, and put $N = \max(N_1, N_2)$. Then $n_N \geq N$ and $N \geq N_2$, so

$$|a_n - L| = |(a_n - a_{n_N}) - (a_{n_N} - L)| \leq |a_n - a_{n_N}| + |a_{n_N} - L| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

[3]You might notice that they are reminiscent of the elementary properties of *monotone* sequences we proved above. But I myself noticed this only quite recently – more than 15 years after I first learned about monotone sequences and Cauchy sequences. I wish I had a deeper understanding of the underlying source – if any! – of these commonalities.

The above proofs *did not* use completeness, and thus the results hold in any ordered field. In contrast, the next result *does* crucially use the Dedekind completeness of the real numbers, in the form of the Bolzano-Weierstrass Theorem.

THEOREM 10.32. *(Cauchy Criterion) Any real Cauchy sequence is convergent.*

PROOF. Let $\{a_n\}$ be a real Cauchy sequence. By Proposition 10.30, $\{a_n\}$ is bounded. By Bolzano-Weierstrass there exists a convergent subequence. Finally, by Proposition 10.31, this implies that $\{a_n\}$ is convergent. □

It can be further shown that an *Archimedean* ordered field F in which every Cauchy sequence is convergent must be Dedekind complete. However, there are *non-Archimedean* – and thus necessarily *not* Dedekind complete – ordered fields in which every Cauchy sequence converges. (In fact there are non-Archimedean ordered fields in which the only Cauchy sequences are the eventually constant sequences!) But we had better not get into such matters here.

Exercise: Is there a "Cauchy criterion" for a function to be differentiable at a point? (Hint: yes. See [**Ma56**].)

9. Geometric Sequences and Series

A **geometric sequence** is a sequence $\{x_n\}_{n=0}^{\infty}$ of real numbers such that there is a fixed real number r with $x_{n+1} = rx_n$. We call r the **geometric ratio** since, if for all $n \in \mathbb{Z}^+$ $x_n \neq 0$ we have $\frac{x_{n+1}}{x_n} = r$.

THEOREM 10.33. *(Geometric Sequences) Let $\{x_n\}$ be a geometric sequence with geometric ratio r and $x_0 \neq 0$.*
a) We have $x_n = x_0 r^n$.
b) If $|r| < 1$, then $x_n \to 0$.
c) If $r = 1$, then $x_n \to x_0$.
d) If $r = -1$, then the sequence is $x_0, -x_0, x_0, -x_0, \ldots$, which diverges.
e) If $|r| > 1$, then $|x_n| \to \infty$.

PROOF. a) A simple induction argument which is left to the reader.

Suppose $x_n \to L \in [-\infty, \infty]$. Since $x_{n+1} = rx_n$, $L = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} rx_n = rL$. The solutions to $L = rL$ for $r \in \mathbb{R}, L \in [-\infty, \infty]$ are: $(1, L)$ for any L, $(r, 0)$ for any r, (r, ∞) for positive r and $(r, -\infty)$ for negative r. Now:
b) If $|r| < 1$, then $|x_{n+1}| = |r||x_n| < |x_n|$, so $\{|x_n|\}$ is decreasing and bounded below by 0. By the Montone Sequence Lemma, $|x_n|$ converges to a finite, non-negative number L, and by the above analysis $L = 0$. Since $|x_n| \to 0$, $x_n \to 0$.
c), d) These are immediate and are just recorded for easy reference.
e) If $|r| > 1$, then $|x_{n+1}| = |r||x_n| > |x_n|$, so the sequence $\{|x_n|\}$ is increasing and bounded below by $|x_0|$. By the Monotone Sequence Lemma, $|x_n| \to L \in (|x_0|, \infty]$, and by the above analysis we must have $L = \infty$. □

For $x_0, r \in \mathbb{R}$, we define the **finite geometric series**

$$S_n = x_0 + x_0 r + \ldots + x_0 r^n. \tag{74}$$

We claim that – quite luckily! – we are able to obtain a simple closed-form expression for S_n. We may dispose of the case $r = 1$: then $S_n = (n+1)$. When $r \neq 1$, we use a very standard *trick*: multiplying (75) by r gives

$$rS_n = \quad x_0 r + \ldots + x_0 r^n + x_0 r^{n+1}, \tag{75}$$

and subtracting (75) from (74) gives

$$(1-r)S_n = x_0(1-r^{n+1}.$$

Since $r \neq 1$, this yields

$$S_n = \sum_{k=0}^{n} x_0 r^k = \frac{1-r^{n+1}}{1-r}. \tag{76}$$

Having this closed form enables us to determine easily whether the sequence $\{S_n\}$ converges, and if so, to what.

THEOREM 10.34. *(The Geometric Series) For $x_0, r \in \mathbb{R}$, let*

$$S_n = \sum_{k=0}^{n} x_0 r^n = x_0 + x_0 r + \ldots + x_0 r^n.$$

a) If $|r| < 1$, then $\lim_{n\to\infty} S_n = \frac{1}{1-r}$.
b) If $|r| \geq 1$, then $\{S_n\}$ diverges.

Exercise: Use Theorem 10.33 to prove Theorem 10.34.

10. Contraction Mappings Revisited

The material of this section owes a debt to a paper of K. Conrad [**CdC**].

10.1. Preliminaries.

Recall that we last studied contractive mappings, fixed points and sequences of iterates in the context of Newton's method. Now, towards the end of a long chapter on infinite sequences, we have more tools at our disposal, so it is reasonable to take another look at this important circle of ideas. Let's take it again from the top:

A function $f : I \to \mathbb{R}$ is **contractive**, or a **contraction mapping**, if there is a constant $0 < C < 1$ such that:

$$\forall x, y \in I, |f(x) - f(y)| \leq C|x-y|. \tag{77}$$

Recall that $f : I \to \mathbb{R}$ is **Lipschitz** if (77) holds for some $C > 0$. If so, for any $\epsilon > 0$, we may – independently of x – choose $\delta < \frac{\epsilon}{C}$, and then $|x-y| \leq \delta \implies |f(x) - f(y)| \leq C\delta < \epsilon$. So a Lipschitz function is uniformly continuous. A contraction mapping is a Lipchitz mapping with Lipschitz constant $C < 1$.

A function $f : I \to \mathbb{R}$ is **weakly contractive** if:

$$\forall x \neq y \in I,\ |f(x) - f(y)| < |x-y|.$$

A function $f : I \to \mathbb{R}$ is a **short map** if 1 is a Lipschitz constant for f, i.e.,

$$\forall x, y \in I,\ |f(x) - f(y)| \leq |x-y|.$$

Thus contractive implies weakly contractive implies short map.

Exercise: Let $f : I \to I$ be differentiable.

a) Show that if $|f'(x)| \leq C$ for all $x \in I$, then C is a Lipschitz constant for f.
b) Deduce: if there is $C \in (0,1)$ with $|f'(x)| \leq C$ for all $x \in I$, f is contractive.
c) State and prove analogous sufficient conditions in terms of bounds on f' for f to be weakly contractive (resp. a short map).

Let X and Y be sets, and let $f : X \to Y$ be a function (or, synonymously, a "map" or "mapping"). An element $L \in X$ is a **fixed point** for f if $f(L) = L$. The study of fixed points of mappings is highly important throughout mathematics.

A fixed point L of f is, in particular, an element of both X and Y. Perhaps the sets X and Y have no elements in common: then of course no function between them can have a fixed point: e.g., looking for fixed points for mappings from the set X of real numbers to the set $Y = \{$Barack Obama, Lady Gaga, William Faulkner$\}$ would be an exercise in futility.

To alleviate this problem, we often restrict attention to mappings $f : X \to X$. Further, such a mapping can be **iterated**: that is, we can take the output and plug it back in as the input: $x \in X \mapsto f(x) \mapsto f(f(x))$. And we can keep going. Formally, starting from any **initial value** $x_0 \in X$, we get a **sequence of iterates** $\{x_n\}_{n=0}^{\infty}$ where for all $n \in \mathbb{N}$, $x_{n+1} = f(x_n)$. (Recall that Newton's method involved consideration of sequences of iterates of the function $T(x) = x - \frac{f(x)}{f'(x)}$.)

And now, back to earth: the methods of calculus allow us to analyze functions defined on subsets, especially intervals, of the real numbers. So suppose we have an interval $I \subset \mathbb{R}$ and a function $f : I \to I$. Here is what we want to know.

QUESTION 10.35. *Let $f : I \to I$ be a function.*
a) Does f have at least one fixed point?
b) Does f have exactly one fixed point?
c) If f has fixed points, how can we explicitly find them, at least approximately?

10.2. Some Easy Results.

Our first result is a very modest generalization of Lemma 7.7.

THEOREM 10.36. *A weakly contractive function has at most one fixed point.*

PROOF. Suppose there are distinct $L_1, L_2 \in I$ such that $f(L_1) = L_1$ and $f(L_2) = L_2$. Then $|f(L_1) - f(L_2)| = |L_1 - L_2|$, contradiction. □

Exercise: Show that a short map $f : I \to I$ can have more than one fixed point.

LEMMA 10.37. *Let $f : I \to I$ be a continuous function. Let $x_0 \in I$ and let $x_0, x_1 = f(x_1), x_2 = f(x_2), \ldots$ be the sequence of iterates of x_0 under f. Suppose that x_n converges to some $L \in I$. Then L is a fixed point of f.*

PROOF. Since f is continuous and $x_n \to L$, $f(x_n) \to f(L)$. But since $f(x_n) = x_{n+1}$, the sequence $\{f(x_n)\} = \{x_{n+1}\}$ is just a reindexed version of the sequence $\{x_n\}$, so it must have the same limit: $f(L) = L$. □

Exercise:[4] Consider the function $f : (0,1] \to (0,1]$ given by $f(x) = \frac{x}{2}$.
a) Show that f is contractive.

[4]This exercise and the next are easy bordering on insulting, but to properly appreciate later results it will be useful to have these examples on record.

b) Show that f has no fixed point on $(0, 1]$.

Exercise: Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x + 1$.
a) Show that f is a short map, hence continuous.
b) Show that f has no fixed point on $\mathbb{R}$.

Exercise: Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sqrt{x^2 + 1}$.
a) Show that f is weakly contractive.
b) Show that f has no fixed point on $\mathbb{R}$.
c) Show that f is not contractive.

When $f : I \to I$ has a fixed point L, we would like to be able to find L as the limit of a sequence of iterates. And we can: if $x_0 = L$, then $f(x_0) = L$, so the sequence of iterates is constantly L. But this observation is not very useful in finding L! The next, slightly less trivial, observation is that a convergent sequence of iterates of a continuous function yields a fixed point. To state this more precisely we introduce some further terminology.

We say $L \in I$ is an **attracting point** for f if for *every* $x_0 \in I$, the sequence of iterates $x_0, f(x_0), f(f(x_0)), \ldots$ converges to L. We say f is **attractive** if it has an attracting point.

LEMMA 10.38. *Suppose $f : I \to I$ is an attractive function.*
a) The attracting point L is unique.
b) f has at most one fixed point.
c) If f is continuous, then the unique attracting point is the unique fixed point.

PROOF. a) Suppose $L_1 \neq L_2$ are attractive points. Starting at any initial point $x_0 \in L$, the sequence of iterates of x_0 under f must converge to both L_1 and L_2. Since a sequence can have at most one limit, this is a contradiction.
b) If L is a fixed point for f, then the sequence of iterates of L under f is constantly L and thus converges to L. So if we had at least two fixed points, different sequences of iterates would have at least two different limits, contradicting attractivity.
c) This follows immediately from the first two parts and from Lemma 10.37. □

Thus attractive maps are precisely the maps with a unique fixed point which can be found as the limit of a sequence of iterates starting from *any* point x_0 in the domain. So the fixed point of an attractive map is easy to find: we just start anywhere, iterate, and we get sucked into it.

Exercise: Consider the function $f : [0, 1] \to [0, 1]$ such that

$$f(x) = \begin{cases} \frac{x}{2}, & 0 < x \leq 1 \\ 1, & x = 0 \end{cases}$$

a) Show that 0 is an attractive point of f but not a fixed point.
b) Why does this not contradict Lemma 10.38?

Remark: Note that, although our terminology suppresses this, whether L is an attracting fixed point for f depends very much on the interval of definition I. In

the literature, the term "attracting point" is often used for a weaker, "local property" which is studied in the following exercise.

Exercise: Let $f : I \to I$ be a function. A point $L \in I$ is a **locally attracting point** if there exists $\delta > 0$ such that f maps $[L - \delta, L + \delta]$ into itself and the restriction of f to $[L - \delta, L + \delta]$ has L as an attracting point.

Now let $f : I \to I$ be C^1, and let $L \in I^\circ$ be a fixed point of f.
a) Show that if $|f'(L)| < 1$, then L is a locally attracting point for f.
b) Show that if $|f'(L)| > 1$, then L is *not* a locally attracting point for f.
c) Exhibit $f : [-1, 1] \to \mathbb{R}$ such that $L = 0$ is a fixed point, $f'(0) = 1$, and 0 is a locally attracting point for f.
d) Exhibit $f : [-1, 1] \to \mathbb{R}$ such that $L = 0$ is a fixed point, $f'(0) = 1$ and 0 is *not* a locally attracting point for f.

10.3. The Contraction Mapping Theorem.

THEOREM 10.39. *(Contraction Mapping Theorem) Let $I \subset \mathbb{R}$ be a closed interval, and let $f : I \to I$ be contractive with constant $C \in (0, 1)$.*
a) Then f is attractive: there is a unique fixed point L, and for all $x_0 \in I$, the sequence $\{x_n\}$ of iterates of f under x_0 converges to L.
b) Explicitly, for all $x_0 \in I$ and all $n \in \mathbb{N}$,

$$|x_n - L| \leq \left(\frac{|x_0 - f(x_0)|}{1 - C}\right) C^n. \tag{78}$$

PROOF. Step 0: By Theorem 10.36, f has at most one fixed point.
Step 1: Let $x_0 \in I$, fix $\epsilon > 0$, let N be a large positive integer to be chosen (rather sooner than) later, let $n \geq N$ and let $k \geq 0$. Then

$$\begin{aligned} |x_{n+k} - x_n| &\leq |x_{n+k} - x_{n+k-1}| + |x_{n+k-1} - x_{n+k-2}| + \ldots + |x_{n+1} - x_n| \\ &\leq C^{n+k-1}|x_1 - x_0| + C^{n+k-2}|x_1 - x_0| + \ldots + C^n|x_1 - x_0| \\ = |x_1 - x_0|C^n\left(1 + C + \ldots + C^{k-1}\right) &= |x_1 - x_0|C^n\left(\frac{1 - C^k}{1 - C}\right) < \left(\frac{|x_1 - x_0|}{1 - C}\right) C^n. \end{aligned}$$

Since $|C| < 1$, $C^n \to 0$, so we may choose N such that for all $n \geq N$ and all $k \in \mathbb{N}$, $\left(\frac{|x_1 - x_0|}{1 - C}\right) C^n < \epsilon$. So $\{x_n\}$ is Cauchy and hence, by Theorem 10.32, convergent.
Step 2: By Step 1 and Lemma 10.37, for each $x_0 \in I$ the sequence of iterates of x_0 under f converges to a fixed point. By Step 0, f has at most one fixed point. So there must be a unique fixed point $L \in I$, which is the limit of every sequence of iterates: i.e., L is an attracting point for f. This completes the proof of part a).
Step 3: We leave the derivation of (78) as an exercise for the reader. □

Exercise: Complete the proof of Theorem 10.39 by establishing (78). (Suggestion: adapt the argument of Step 1 of the proof.)

Thus the Contraction Mapping Theorem asserts in particular that a contractive mapping on a closed interval is attractive. That's a fine result in and of itself. But we have more, namely the explicit estimate (78). It guarantees that starting at any point $x_0 \in I$ the sequence of iterates converges to the attracting point L *exponentially fast*, and is so explicit that it immediately gives an efficient algorithm for approximating L to any desired accuracy. Now that's a theorem!

Exercise: a) Show that there is a unique real number x such that $\cos x = x$.
b) Explain how to use (say) a handheld calculator to approximate x to as many decimal places of accuracy as your calculator carries.

Exercise: Let $f : [a, b) \to [a, b)$ be a contractive map with constant $C \in (0, 1)$.
a) Show that $\lim_{x \to b^-} f(x)$ exists (as a real number!), and that by defining f at b to be this limit, the extended function $f : [a, b] \to \mathbb{R}$ remains contractive with constant C.[5]
b) Use part a) to extend the Contraction Mapping Principle to contractions $f : [a, b) \to [a, b)$, with the proviso that the unique fixed point may be b.
c) Give an example of a contraction mapping on $[a, b)$ with fixed point b.
d) State and prove a version of the Contraction Mapping Principle valid for an arbitrary interval.

Exercise: Let $f : I \to I$ be a function. For $n \in \mathbb{Z}^+$, let $f^{\circ n} = f \circ \ldots \circ f$ be the nth iterate of f. Suppose that for some $N \in \mathbb{Z}^+$, $f^{\circ N}$ is contractive.
a) Show that any fixed point of f is a fixed point of $f^{\circ N}$.
b) Show that $f^{\circ N}$ has a unique fixed point $L \in \overline{I}$, and deduce that f has at most one fixed point in $\overline{I}$.
c) Show that $f(L)$ is also a fixed point for $f^{\circ N}$, and deduce that $f(L) = L$: thus f has a unique fixed point in $\overline{I}$.
d) Show that in fact L is an attracting point for f.
e) Consider the function $f : [0, 1] \to [0, 1]$ by

$$f(x) = \begin{cases} 0, & 0 \leq x \in [0, 1] \\ 1, & x \in (1, 2]. \end{cases}$$

Show that f is not a contraction but $f \circ f$ is a contraction.

10.4. Further Attraction Theorems.

Let I be an interval, and let $f : I \to I$ be continuous. Recall that $L \in I$ is attracting for f if for all $x_0 \in I$, the sequence of iterates of x_0 under f converges to L; above, we showed that if L is attracting for f then L is the unique fixed point of f. Let us say $\sup I$ is an attracting point for f if for all $x_0 \in I$, the sequence of iterates of x_0 under f approaches $\sup I$ ($\sup I = \infty$ iff I is unbounded above). Similarly, we say $\inf I$ is an attracting point for f if for all $x_0 \in I$, the sequence of iterates of x_0 under f approaches $\inf I$ ($\inf I = -\infty$ iff I is unbounded below).

THEOREM 10.40. *Let $I \subset \mathbb{R}$ be an interval, and let $f : I \to I$ be continuous.*
a) At least one of the following holds:
(i) f has a fixed point.
(ii) $\sup I$ is an attracting point for f.
(iii) $\inf I$ is an attracting point for f.
b) If $I = [a, b]$ then f has a fixed point.

PROOF. a) Define $g : I \to \mathbb{R}$ by $g(x) = f(x) - x$. A fixed point of f is precisely a root of g, so to prove part a) we may assume that g has no roots on I and show

[5]This is the hardest part of the problem. The result on extension will be much easier after you have read the next section, which treats similar but more general problems. You may wish to assume this part for now and come back to it later.

that either (ii) or (iii) holds. If the continuous function g has no roots then either (I) $f(x) > x$ for all $x \in I$ or (II) $f(x) < x$ for all $x \in I$. We will show (I) $\implies$ (ii); the very similar proof that (II) $\implies$ (iii) is left to the reader.

Suppose $f(x) > x$ for all $x \in I$. Then, for any $x_0 \in I$, the sequence of iterates is increasing. This sequence cannot converge to any element L of I, for by Lemma 10.37, L would then be a fixed point of f, contradiction. So x_n must approach $\sup I$.

b) Let $I = [a, b]$ and let $x_0 \in I$. If there were no fixed point, then by part a) the sequence of iterates of x_0 under f would approach either $\sup I = b$ or $\inf I = a$. But both are elements of I, so by Lemma 10.37 either a or b is a fixed point. □

THEOREM 10.41. *Let $I \subset \mathbb{R}$ be an interval, and let $f : I \to I$ be weakly contractive. Then f has an attracting point in* $[\inf I, \sup I]$.

PROOF. Step 0: If f has no fixed point in I, then by Theorem 10.40 either $\sup I$ or $\inf I$ is an attracting point for f. Thus we may assume that f has a fixed point $L \in I$, and our task is to show that L is attracting for f.

Step 1: Let $x_0 \in I$, and for $n \in \mathbb{N}$, put $d_n = |x_n - L|$. If for some N we have $x_N = L$, then $x_n = L$ for all $n \geq N$, so the sequence of iterates converges to L. So we may assume $d_n > 0$ for all $n \in \mathbb{N}$, and then by weak contractivity $\{d_n\}$ is decreasing. Since 0 is a lower bound, there is $d \geq 0$ such that $d_n \to d$. Observe that the desired conclusion that $x_n \to L$ is equivalent to $d = 0$.

Step 2: For all $n \in \mathbb{N}$, $x_n \in [L - d_0, L + d_0]$, so $\{x_n\}$ is bounded; by Bolzano-Weierstrass, there is a convergent subsequence, say $x_{n_k} \to y$. As $k \to \infty$ we have:

$$|x_{n_k} - L| \to |y - L|, \ |x_{n_k} - L| = d_{n_k} \to d,$$

$$|f(x_{n_k}) - L| = |x_{n_k+1} - L| = d_{n_k+1} \to d, \ |f(x_{n_k}) - L| \to |f(y) - L|,$$

so $|y - L| = d = |f(y) - L| = |f(y) - f(L)|$. By weak contractivity, $d = 0$. □

Remark: The case $I = \mathbb{R}$ of Theorem 10.41 is due to A. Beardon [**Be06**]. The case $I = [a, b]$ is an instance of a general result of M. Edelstein [**Ed62**] which is described in the next section. Our proof of Theorem 10.41 draws ideas from Beardon's proof and also from K. Conrad's treatment of Edelstein's Theorem in [**CdC**].

Exercise: Show that the function $f : [0, 1] \to [0, 1]$ defined by $f(x) = \frac{1}{1+x}$ is weakly contractive but not contractive.

10.5. Fixed Point Theorems in Metric Spaces.

Theorem 10.39 is a special case of a celebrated theorem of Banach,[6] itself called either the **Contraction Mapping Principle** or **Banach's Fixed Point Theorem**. The flavor of the generalization can perhaps be appreciated by looking back at what tools we used to prove Theorem 10.39. The general strategy of the proof was to show that any sequence of iterates of x_0 under f was a Cauchy sequence. To show this, we used nothing other than the definition of a contractive mapping and (repeatedly) the triangle inequality.

Now there is a general mathematical structure called a **metric space**: this is a set X equipped with a **metric function** $d : X \times X \to \mathbb{R}$. The intuition is that for $x, y \in X$, $d(x, y)$ measures the *distance* between x and y. When $X = \mathbb{R}$, this

[6] Stefan Banach (1892-1945)

function is nothing else than $d(x,y) = |x-y|$. In order for a function d to be a metric, it must satisfy three very familiar properties:

(M1) (Definiteness) For all $x, y \in X$, $d(x,y) \geq 0$, with equality iff $x = y$.
(M2) (Symmetry) For all $x, y \in X$, $d(x,y) = d(y,x)$.
(M3) (Triangle Inequality) For all $x, y, z \in X$, $d(x,z) \leq d(x,y) + d(y,z)$.

Thus $d(x,y) = |x-y|$ is certainly a metric on $\mathbb{R}$. It turns out that these three simple properties enable us to carry over a large portion of the theory of real infinite sequences to study sequences *with values in a metric space* X, i.e., functions $x_\bullet : \mathbb{Z}^+ \to X$. Essentially, wherever in any of our definitions we see $|x-y|$, we replace it with $d(x,y)$. For instance, we say that a sequence $\{x_n\}$ in a metric space X **converges** to an element $L \in X$ if for all $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $d(x_n, L) < \epsilon$. Similarly, one can directly carry over the definition of a Cauchy sequence: try it!

Similarly, if (X, d_X) and (Y, d_Y) are metric spaces, then a function $f : X \to Y$ is continuous at $x \in X$ if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $x' \in X$ with $d_X(x,x') < \delta$, $d_Y(f(x), f(x')) < \epsilon$. A function f is continuous if it is continuous at every point of X. (The reader should recognize this as a rather straightforward adaptation of our cherished ϵ-δ definition of continuity to this new context.)

One important class of examples of metric spaces are obtained simply by taking a subset $X \subset \mathbb{R}$ and defininining $d(x,y) : X \times X \to \mathbb{R}$ as before: $d(x,y) = |x-y|$. The point here is that for some subspaces of X, a Cauchy sequence with values in X need not converge to an element of X. This is not really new or surprising: for instance, consider the sequence $x_n = \frac{1}{n}$ as a sequence in $(0, \infty)$: it is a Cauchy sequence there, but it does not converge to an element of $(0, \infty)$; rather it converges to 0. Similarly, in a general metric space X a Cauchy sequence need not converge to an element of X: we say that a metric space X is **complete** if every Cauchy sequence in X is convergent in X. Now we can state the theorem.

THEOREM 10.42. *(Banach Fixed Point Theorem* [**Ba22**]*) Let X be a complete metric space, and let $f : X \to X$ be a contraction mapping: that is, there is $C \in (0,1)$ such that for all $x, y \in X$,*

$$d(f(x), f(y)) \leq C d(x,y).$$

Then:
a) There is a unique $L \in X$ with $f(L) = L$, i.e., a fixed point of f.
b) For every $x_0 \in X$, define the sequence of iterates of x_0 under f by $x_{n+1} = f(x_n)$ for $n \in \mathbb{N}$. Then for all $n \in \mathbb{N}$,

$$d(x_n, L) \leq C^n d(x_0, L).$$

The proof of Theorem 10.42 is exactly the same as that of Theorem 10.39.

For the last century, Banach's Fixed Point Theorem has been one of the most important and useful results in mathematical analysis: it gives a very general condition for the existence of fixed points, and a remarkable number of "existence theorems" can be reduced to the existence of a fixed point of some function on some metric space. For instance, if you continue on in your study of mathematics

you will surely learn about systems of differential equations, and the most important result in this area is that – with suitable hypotheses and precisions, of course – every system of differential equations has a unique solution. The now standard proof of this seminal result uses Banach's Fixed Point Theorem![7]

Let X be a metric space. Then the statement "Every sequence with values in X admits a convergent subsequence" is certainly meaningful, but – as is already the case with intervals on the real line! – whether it is true or false certainly depends on X. We say that a metric space is **compact** if every sequence with values in X admits a convergent subsequence.

In fact every compact metric space is complete, and the proof again requires no ideas other than the ones we have already developed: indeed, if $\{x_n\}$ is a Cauchy sequence in a compact metric space, then by definition of compactness it admits a subsequence x_{n_k} converging to some $L \in X$, and then we prove – exactly as we did before – that if a subsequence of a Cauchy sequence converges to L then the Cauchy sequence itself must converge to L.

Above we showed that a weakly contractive map on a closed, bounded (and thus compact) interval was attractive and attributed this to M. Edelstein. What Edelstein actually showed was the following result.

THEOREM 10.43. *(Edelstein* [**Ed62**]*) Let X be a compact metric space, and let $f : X \to X$ be a weakly contractive mapping. Then f is attractive:*
a) There is a unique fixed point L of f.
b) For all $x_0 \in X$, the sequence of iterates of x_0 under f converges to f.

PROOF. We follow [**CdC**].
Step 0: The Extreme Value Theorem has the following generalization to compact metric spaces: if X is a compact metric space, then any continuous function $f : X \to \mathbb{R}$ is bounded and attains its maximum and minimum values. Recall that we gave two proofs of the Extreme Value Theorem for $X = [a, b]$: one using Real Induction and one using the fact that every sequence in $[a, b]$ admits a convergent subsequence. Since by definition this latter property holds in a compact metric space, it is the second proof that we wish to carry over here, and we ask the interested reader to check that it does carry over with no new difficulties.
Step 1: We claim f has a fixed point. Here we need a new argument: the one we gave for $X = [a, b]$ used the Intermediate Value Theorem, which is not available in our present context. So here goes: let $g : X \to \mathbb{R}$ by $g(x) = d(x, f(x))$. Since f is continuous, so is g and thus the Extreme Value Theorem applies and in particular g attains a minimum value: there is $L \in X$ such that for all $y \in X$, $d(L, f(L)) \leq d(y, f(y))$. But if $f(L) \neq L$, then by weak contractivity we have $d(f(L), f(f(L))) < d(L, f(L))$, i.e., $g(f(L)) < g(L)$, contradiction. So L is a fixed point for f.
Step 2: The argument of Step 2 of the proof of Theorem 10.41 carries over directly to show that L is an attracting point for f. □

[7]In fact the title of [**Ba22**] indicates that applications to *integral equations* are being explicitly considered. An "integral equation" is very similar in spirit to a differential equation: it is an equating relating an unknown function to its integral(s).

11. Extending Continuous Functions

Let $I \subset \mathbb{R}$ be an interval.[8] A subset S of I is **dense in I** if for all $x \in I$ and all $\epsilon > 0$, there exists at least one element of s with $s \in (x - \epsilon, x + \epsilon)$.

For example, for every interval I the subset $I \cap \mathbb{Q}$ of rational numbers in I is dense in I, as is the subset $I \cap (\mathbb{R} \setminus \mathbb{Q})$ of irrational numbers in I.

Here we wish to consider the following problem: suppose we are given a dense subset S of I and a function $f : S \to \mathbb{R}$. Under which circumstances does f **extend** to a function $\tilde{f} : I \to \mathbb{R}$?

Stated this way the answer is easy: always. For instance, we may set $\tilde{f}(x) = 0$ for all $x \in I \setminus S$. So we were not careful enough. Here's what we really mean:

QUESTION 10.44. *Let S be a dense subset of the interval I, and suppose we are given a function $f : S \to \mathbb{R}$. Under what circumstances is there a* continuous *function $\tilde{f} : I \to \mathbb{R}$ extending f, i.e., such that $\tilde{f}(x) = f(x)$ for all $x \in S$?*

Here are some preliminary observations. First, the *uniqueness* problem is easier to solve than the *existence* problem, and it is helpful to nail that down first.

PROPOSITION 10.45. *(Uniqueness of Extension) Let S be a dense subset of an interval I, and let $f : S \to \mathbb{R}$ be a function. There is* at most *one continuus extension of f to I: that is, if $\tilde{f}_1, \tilde{f}_2 : I \to \mathbb{R}$ are two continuous functions such that $\tilde{f}_1(x) = \tilde{f}_2(x) = f(x)$ for all $x \in S$, then $\tilde{f}_1(x) = \tilde{f}_2(x)$ for all $x \in I$.*

PROOF. Step 1: First suppose $f \equiv 0$ on S. Since S is dense in I, for any $x \in I$ and $n \in \mathbb{Z}^+$, there is $x_n \in S$ with $|x_n - x| < \frac{1}{n}$. Thus we get a sequence $\{x_n\}_{n=1}^{\infty}$ taking values in S such that $x_n \to x$. Let $\tilde{f} : I \to \mathbb{R}$ be any continuous extension of f. Then since $x_n \to x$, $f(x_n) \to f(x)$. But $\{f(x_n)\} = 0, 0, 0, \ldots$, so it converges (uniquely!) to 0 and thus $f(x) = 0$.
Step 2: In general, suppose $\tilde{f}_1, \tilde{f}_2 : I \to \mathbb{R}$ are two continuous extensions of f, and put $\tilde{g} = \tilde{f}_1 - \tilde{f}_2 : I \to \mathbb{R}$. Then $\tilde{g}$ is continuous and for all $x \in S$, $\tilde{g}(x) = \tilde{f}_1(x) - \tilde{f}_2(x) = 0$. By Step 1, for all $x \in I$,

$$0 = \tilde{g}(x) = \tilde{f}_1(x) - \tilde{f}_2(x),$$

so $\tilde{f}_1(x) = \tilde{f}_2(x)$ for all $x \in I$. □

To go further, we need the definition of a continuous function $f : S \to \mathbb{R}$ for an arbitrary subset S of $\mathbb{R}$. There is nothing novel here: f is continuous at $s_0 \in S$ if for all $\epsilon > 0$, there is $\delta > 0$ such that for all $s \in S$ with $|s - s_0| < \delta$, $|f(s) - f(s_0)| < \epsilon$, $f : S \to \mathbb{R}$ is continuous if it is continuous at s for every $s \in S$.

PROPOSITION 10.46. *If f admits a continuous extension $\tilde{f}$ to I, then f is continuous on S.*

PROOF. Let $s_0 \in S$, and fix $\epsilon > 0$. Since $S \subset I$, $\tilde{f}$ is continuous at s_0, so there is $\delta > 0$ such that for all $x \in I$ with $|x - s_0| < \delta$, $|\tilde{f}(x) - \tilde{f}(s_0)| < \epsilon$. In particular then, for all $s \in S$ with $|s - s_0| < \delta$, $|f(s) - f(s_0)| = |\tilde{f}(s) - \tilde{f}(s_0)| < \epsilon$. □

[8]To evade trivialities, we assume that all of our intervals contain more than one point.

Example: $I = \mathbb{R}$, $S = \mathbb{R} \setminus \{\sqrt{2}\}$, $f(x) = 0$ for $x < \sqrt{2}$, $f(x) = 1$ for $x > \sqrt{2}$. Then f is continuous on S but admits no continuous extension to I.

Here is a key observation: our counterexample function f is continuous on S but not *uniformly continuous* there: there is no $\delta > 0$ such that for all $x_1, x_2 \in S$, $|x_1 - x_2| < \delta \implies |f(x_1) - f(x_2)| < 1$.

THEOREM 10.47. *(Extension Theorem) Let S be a dense subset of an interval I, and let $f : S \to \mathbb{R}$ be a function. If f is* uniformly continuous *on S, then it extends uniquely to a uniformly continuous function $\tilde{f} : I \to \mathbb{R}$.*

PROOF. As above, since S is dense in I, for each $x \in I$ there is a sequence $\{a_n\}$ in S with $a_n \to x$. This suggests that we define $\tilde{f}(x) = \lim_{n\to\infty} f(a_n)$. And we will, but there are several things we need to check:
(i) The sequence $f(a_n)$ is convergent.
(ii) For any other sequence $\{b_n\}$ in S with $b_n \to x$, $\lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} f(b_n)$.
(iii) The (now well-defined) extension function $\tilde{f}$ is continuous on I.
So let's do it:
(i) By Theorem 10.32, it suffices to show that $\{f(a_n)\}$ is Cauchy. Fix $\epsilon > 0$. Since f is **uniformly continuous** on S, there exists $\delta > 0$ such that $|s_1 - s_2| < \delta \implies |f(s_1) - f(s_2)| < \epsilon$. Since $a_n \to x$, there is $N \in \mathbb{Z}^+$ such that for $n \geq N$, $|a_n - x| < \frac{\delta}{2}$. Then for $m, n \geq N$, $|a_m - a_n| < \delta$, so $|f(a_m) - f(a_n)| < \epsilon$.
(ii) Suppose that $a_n \to x$ and $b_n \to x$. By (i), we know that $f(a_n) \to L$ and $f(b_n) \to M$, say, and we must show that $L = M$. Fix $\epsilon > 0$. By uniform continuity there exists $\delta > 0$ such that $|s_1 - s_2| < \delta \implies |f(s_1) - f(s_2)| < \epsilon$. Moreover for all sufficiently large n we have $|a_n - x| < \frac{\delta}{2}$, $|b_n - x| < \frac{\delta}{2}$, so $|a_n - b_n| < \delta$ and hence $|f(a_n) - f(b_n)| < \epsilon$. It follows that $|L - M| \leq \epsilon$. Since ϵ was arbitrary, $L = M$.
(iii) Let $x \in I$, and suppose that $\tilde{f}$ is *not* uniformly continuous on I: then there exists some $\epsilon > 0$ such that for all $\delta > 0$, there are $x_1, x_2 \in I$ with $|x_1 - x_2| < \frac{\delta}{3}$ but $|f(x_1) - f(x_2)| \geq \epsilon$. By (ii), there are $s_1, s_2 \in S$ such that

$$|s_1 - x_1|, |s_2 - x_2| < \frac{\delta}{3},\ |f(s_1) - f(x_1)|, |f(s_2) - f(x_2)| < \frac{\epsilon}{3}.$$

Then

$$|s_1 - s_2| \leq |s_1 - x_1| + |x_1 - x_2| + |x_2 - s_2| < \delta,$$

and

$$|f(x_1) - f(x_2)| \leq |f(x_1) - f(s_1)| + |f(s_1) - f(s_2)| + |f(s_2) - f(x_2)| \leq \frac{2\epsilon}{3} + |f(s_1) - f(s_2)|.$$

But since $f : S \to \mathbb{R}$ is uniformly continuous, we may take δ to be sufficiently small so that $|s_1 - s_2| < \delta \implies |f(s_1) - f(s_2)| < \frac{\epsilon}{3}$. For such a δ we get

$$|f(x_1) - f(x_2)| < \frac{2\epsilon}{3} + \frac{\epsilon}{3} = \epsilon,$$

contradiction.

□

Theorem 10.47 gives necessary and sufficient conditions for a function $f : S \to \mathbb{R}$ to admit a *uniformly* continuous extension to I. When I is closed and bounded, this solves our extension problem because uniform continuity is equivalent to continuity. However, for an interval which is not closed and bounded, being uniformly continuous is *much* stronger than being continuous. For instance, the only polynomial

functions which are uniformly continuous on all of $\mathbb{R}$ are the linear polynomials.

However there is an easy, but sneaky, way to "soup up" the Extension Theorem: if I is not closed and bounded, we don't have to extend f to I "all at once"; we can do it via an increasing sequence of closed, bounded subintervals of I.

COROLLARY 10.48. *Let S be a dense subset of $\mathbb{R}$, and let $f : S \to \mathbb{R}$. The following are equivalent:*
(i) For all $M > 0$, the restriction of f to $S \cap [-M, M]$ is uniformly continuous.
(ii) There is a unique extension of f to a continuous function $\tilde{f} : \mathbb{R} \to \mathbb{R}$.

PROOF. (i) $\implies$ (ii): Applying the Extension Theorem to f on $S \cap [-M, M]$, we get a unique continuous extension $\tilde{f}_M : [-M, M] \to \mathbb{R}$. Since the extensions are unique, for any $x \in \mathbb{R}$, we may choose any M with $M \geq |x|$ and define $\tilde{f}(x) = \tilde{f}_M(x)$: this does not depend on which M we choose. Moreover, since continuity at a point depends only on the behavior of the function in small intervals around the point, it is immediate that any function constructed from an expanding family of continuous functions in this way is continuous on all of $\mathbb{R}$. $\square$

CHAPTER 11

Infinite Series

1. Introduction

1.1. Zeno Comes Alive: a historico-philosophical introduction.

Humankind has had a fascination with, but also a suspicion of, infinite processes for well over two thousand years. Historically, the first kind of infinite process that received detailed infomation was the idea of adding together infinitely many quantitties; or, to put a slightly different emphasis on the same idea, to divide a whole into infinitely many parts.

The idea that any sort of infinite process can lead to a finite answer has been deeply unsettling to philosophers going back at least to Zeno,[1] who believed that a convergent infinite process was absurd. Since he had a sufficiently penetrating eye to see convergent infinite processes all around him, he ended up at the lively conclusion that many everyday phenomena are in fact absurd (so, in his view, illusory).

We will get the flavor of his ideas by considering just one paradox, Zeno's **arrow paradox**. Suppose that an arrow is fired at a target one stadium away. Can the arrow possibly hit the target? Call this event E. Before E can take place, the arrow must arrive at the halfway point to its target: call this event E_1. But before it does *that* it must arrive halfway to the halfway point: call this event E_2. We may continue in this way, getting infinitely many events $E_1, E_2, \ldots$ all of which must happen *before* the event E. That infinitely many things can happen before some predetermined thing Zeno regarded as absurd, and he concluded that the arrow never hits its target. Similarly he deduced that all motion is impossible.[2]

Nowadays we have the mathematical tools to retain Zeno's basic insight (that a single interval of finite length can be divided into infinitely many subintervals) without regarding it as distressing or paradoxical. Indeed, assuming the arrow takes one second to hit its target and (rather unrealistically) travels at uniform velocity, we know exactly when these events E_i take place: E_1 takes place after $\frac{1}{2}$ seconds, E_2 takes place after $\frac{1}{4}$ seconds, and so forth: E_n takes place after $\frac{1}{2^n}$ seconds. Nevertheless there is something interesting here: we have divided the total time of the trip into infinitely many parts, and the conclusion seems to be that

$$\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 1. \tag{79}$$

[1] Zeno of Elea, ca. 490 BC - ca. 430 BC.

[2] One has to wonder whether he got out much.

So now we have not a problem not in the philosophical sense but in the mathematical one: what meaning can be given to the left hand side of (79)? Certainly we ought to proceed with some caution in our desire to add infinitely many things together and get a finite number: the expression

$$1 + 1 + \ldots + 1 + \ldots$$

represents an infinite sequence of events, each lasting one second. Surely the aggregate of these events takes forever.

We see then that we dearly need a mathematical definition of an infinite series of numbers and also of its sum. Precisely, if $a_1, a_2, \ldots$ is a sequence of real numbers and S is a real number, we need to give a precise meaning to the equation

$$a_1 + \ldots + a_n + \ldots = S.$$

So here it is. We *do not* try to add everything together all at once. Instead, we form from our sequence $\{a_n\}$ an auxiliary sequence $\{S_n\}$ whose terms represent adding up the first n numbers. Precisely, for $n \in \mathbb{Z}^+$, we define

$$S_n = a_1 + \ldots + a_n.$$

The associated sequence $\{S_n\}$ is said to be the **sequence of partial sums** of the sequence $\{a_n\}$; when necessary we call $\{a_n\}$ the sequence of **terms**. Finally, we say that the **infinite series** $a_1 + \ldots + a_n + \ldots = \sum_{n=1}^{\infty} a_n$ **converges to S** – or **has sum S** – if $\lim_{n\to\infty} S_n = S$ in the familiar sense of limits of seqeunces. If the sequence of partial sums $\{S_n\}$ converges to some number S we say the infinite series is **convergent**; if the sequence $\{S_n\}$ diverges then the infinite series $\sum_{n=1}^{\infty} a_n$ is **divergent**.

Thus the trick of defining the infinite sum $\sum_{n=1}^{\infty} a_n$ is to do everything in terms of the associated sequence of partial sums $S_n = a_1 + \ldots + a_n$.

In particular by $\sum_{n=1}^{\infty} a_n = \infty$ we mean the sequence of partial sums diverges to ∞, and by $\sum_{n=1}^{\infty} a_n = -\infty$ we mean the sequence of partial sums diverges to $-\infty$. So to spell out the first definition completely, $\sum_{n=1}^{\infty} a_n = \infty$ means: for every $M \in \mathbb{R}$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $a_1 + \ldots + a_n \geq M$.

Let us revisit the examples above using the formal definition of convergence.

Example 1: Consider the infinite series $1 + 1 + \ldots + 1 + \ldots$, in which $a_n = 1$ for all n. Then $S_n = a_1 + \ldots + a_n = 1 + \ldots + 1 = n$, and we conclude

$$\sum_{n=1}^{\infty} 1 = \lim_{n\to\infty} n = \infty.$$

Thus this infinite series indeed diverges to infinity.

Example 2: Consider $\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots$, in which $a_n = \frac{1}{2^n}$ for all n, so

$$S_n = \frac{1}{2} + \ldots + \frac{1}{2^n}. \tag{80}$$

There is a standard trick for evaluating such finite sums. Namely, multiplying (80) by $\frac{1}{2}$ and subtracting it from (80) all but the first and last terms cancel, and we get

$$\frac{1}{2}S_n = S_n - \frac{1}{2}S_n = \frac{1}{2} - \frac{1}{2^{n+1}},$$

and thus

$$S_n = 1 - \frac{1}{2^n}.$$

It follows that

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \lim_{n\to\infty} (1 - \frac{1}{2^n}) = 1.$$

So Zeno was right!

Remark: It is not necessary for the sequence of terms $\{a_n\}$ of an infinite series to start with a_1. In our applications it will be almost as common to consider series starting with a_0. More generally, if N is any integer, then by $\sum_{n=N}^{\infty} a_n$ we mean the sequence of partial sums $a_N\ , a_N + a_{N+1},\ a_N + a_{N+1} + a_{N+2}, \ldots$.

1.2. Telescoping Series.

Example: Consider the series $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$. We have

$$S_1 = \frac{1}{2},$$
$$S_2 = S_1 + a_2 = \frac{1}{2} + \frac{1}{6} = \frac{2}{3},$$
$$S_3 = S_2 + a_3 = \frac{2}{3} + \frac{1}{12} = \frac{3}{4},$$
$$S_4 = S_3 + a_4 = \frac{3}{4} + \frac{1}{20} = \frac{4}{5}.$$

It certainly seems as though we have $S_n = 1 - \frac{1}{n+1} = \frac{n}{n+1}$ for all $n \in \mathbb{Z}^+$. If this is the case, then we have

$$\sum_{n=1}^{\infty} a_n = \lim_{n\to\infty} \frac{n}{n+1} = 1.$$

How to prove it?

First Proof: As ever, induction is a powerful tool to prove that an identity holds for all positive integers, *even if we don't really understand why the identity should hold!* Indeed, we don't even have to fully wake up to give an induction proof: we wish to show that for all positive integers n,

$$S_n = \sum_{k=1}^{n} \frac{1}{k^2+k} = \frac{n}{n+1}. \tag{81}$$

When $n = 1$, both sides equal $\frac{1}{2}$. Now suppose (81) holds for some $n \in \mathbb{Z}^+$. Then

$$S_{n+1} = S_n + \frac{1}{(n+1)^2 + (n+1)} \overset{\text{IH}}{=} \frac{n}{n+1} + \frac{1}{n^2+3n+2} = \frac{n}{n+1} + \frac{1}{(n+1)(n+2)}$$
$$= \frac{(n+2)n+1}{(n+1)(n+2)} = \frac{(n+1)^2}{(n+1)(n+2)} = \frac{n+1}{n+2}.$$

This approach will work whenever we have some reason to look for and successfully guess a simple closed form identity for S_n. But in fact, as we will see in the coming

sections, in practice it is *exceedingly rare* that we are able to express the partial sums S_n in a simple closed form. Trying to do this for each given series would turn out to be a discouraging waste of time. We need some *insight* into why the series $\sum_{n=1}^{\infty} \frac{1}{n^2+n}$ happens to work out so nicely.

Well, if we stare at the induction proof long enough we will eventually notice how convenient it was that the denominator of $\frac{1}{(n+1)^2+(n+1)}$ factors into $(n+1)(n+2)$. Equivalently, we may look at the factorization $\frac{1}{n^2+n} = \frac{1}{(n+1)(n)}$. Does this remind us of anything? I certainly hope so: this is a partial fractions decomposition. In this case, we know there are constants A and B such that

$$\frac{1}{n(n+1)} = \frac{A}{n} + \frac{B}{n+1}.$$

I leave it to you to confirm – in whatever manner seems best to you – that we have

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}.$$

This makes the behavior of the partial sums much more clear! Indeed we have

$$S_1 = 1 - \frac{1}{2}.$$

$$S_2 = S_1 + a_2 = (1 - \frac{1}{2}) + (\frac{1}{2} - \frac{1}{3}) = 1 - \frac{1}{3}.$$

$$S_3 = S_2 + a_3 = (1 - \frac{1}{3}) + (\frac{1}{3} - \frac{1}{4}) = 1 - \frac{1}{4},$$

and so on. This much simplifies the inductive proof that $S_n = 1 - \frac{1}{n+1}$. In fact induction is not needed: we have that

$$S_n = a_1 + \ldots + a_n = (1 - \frac{1}{2}) + (\frac{1}{2} - \frac{1}{3}) + \ldots + (\frac{1}{n} - \frac{1}{n+1}) = 1 - \frac{1}{n+1},$$

the point being that every term except the first and last is cancelled out by some other term. Thus once again $\sum_{n=1}^{\infty} \frac{1}{n^2+n} = \lim_{n\to\infty} 1 - \frac{1}{n+1} = 1$.

Finite sums which cancel in this way are often called **telescoping sums**, I believe after those old-timey collapsible nautical telescopes. In general an infinite sum $\sum_{n=1}^{\infty} a_n$ is telescoping when we can find an auxiliary sequence $\{b_n\}_{n=1}^{\infty}$ such that $a_1 = b_1$ and for all $n \geq 2$, $a_n = b_n - b_{n-1}$, for then for all $n \geq 1$ we have

$$S_n = a_1 + a_2 + \ldots + a_n = b_1 + (b_2 - b_1) + \ldots + (b_n - b_{n-1}) = b_n.$$

But looking at these formulas shows something curious: every infinite series is telescoping: we need only take $b_n = S_n$ for all n! Another, less confusing, way to say this is that if we start with any infinite sequence $\{S_n\}_{n=1}^{\infty}$, then there is a unique sequence $\{a_n\}_{n=1}^{\infty}$ such that S_n is the sequence of partial sums $S_n = a_1 + \ldots + a_n$. Indeed, the key equations here are simply

$$S_1 = a_1,$$

$$\forall n \geq 2,\ S_n - S_{n-1} = a_n,$$

which tells us how to define the a_n's in terms of the S_n's.

In practice all this seems to amount to the following: if you can find a simple closed form expression for the nth partial sum S_n (in which case you are very

lucky), then in order to prove it you do not need to do anything so fancy as mathematical induction (or fancier!). Rather, it will suffice to just compute that $S_1 = a_1$ and for all $n \geq 2$, $S_n - S_{n-1} = a_n$. This is the discrete analogue of the fact that if you want to show that $\int f dx = F$ – i.e., you already have a function F which you believe is an antiderivative of f – then you need not use any integration techniques whatsoever but may simply check that $F' = f$.

Exercise: Let $n \in \mathbb{Z}^+$. We define the **nth harmonic number** $H_n = \sum_{k=1}^n \frac{1}{k} = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n}$. Show that for all $n \geq 2$, $H_n \in \mathbb{Q} \setminus \mathbb{Z}$. (Suggestion: more specifically, show that for all $n \geq 2$, when written as a fraction $\frac{a}{b}$ in lowest terms, then the denominator b is divisible by 2.)[3]

Exercise: Let $k \in \mathbb{Z}^+$. Use the method of telescoping sums to give an exact formula for $\sum_{n=1}^\infty \frac{1}{n(n+k)}$ in terms of the harmonic number H_k of the previous exercise.

2. Basic Operations on Series

Given an infinite series $\sum_{n=1}^\infty a_n$ there are two basic questions to ask:

QUESTION 11.1. *For an infinite series $\sum_{n=1}^\infty a_n$:*
a) Is the series convergent or divergent?
b) If the series is convergent, what is its sum?

It may seem that this is only "one and a half questions" because if the series diverges we cannot ask about its sum (other than to ask whether it diverges to $\pm\infty$ or "due to oscillation"). However, later on we will revisit this missing "half a question": if a series diveges we may ask *how rapidly* it diverges, or in more sophisticated language we may ask for an **asymptotic estimate** for the sequence of partial sums $\sum_{n=1}^N a_n$ as a function of N as $N \to \infty$.

Note that we have just seen an instance in which we asked and answered both of these questions: for a geometric series $\sum_{n=N}^\infty cr^n$, we know that the series converges iff $|r| < 1$ and in that case its sum is $\frac{cr^N}{1-r}$. We should keep this success story in mind, both because geometric series are ubiquitous and turn out to play a distinguished role in the theory in many ways, *but* also because other examples of series in which we can answer Question 11.1b) – i.e., determine the sum of a convergent series – are much harder to come by. Frankly, in a standard course on infinite series one all but forgets about Question 11.1b) and the game becomes simply to decide whether a given series is convergent or not. In these notes we try to give a little more attention to the second question in some of the optional sections.

In any case, there is a certain philosophy at work when one is, for the moment, interested in determining the convergence / divergence of a given series $\sum_{n=1}^\infty a_n$ rather than the sum. Namely, there are certain operations that one can perform on an infinite series that will preserve the convergence / divergence of the series – i.e., when applied to a convergent series yields another convergent series and when applied to a divergent series yields another divergent series – but will in general

[3]This is a number theory exercise which has, so far as I know, nothing to do with infinite series. But I am a number theorist...

change the sum.

First, we may add or remove any finite number of terms from an infinite series without affecting its convergence. In other words, suppose we start with a series $\sum_{n=1}^{\infty} a_n$. Then, for any integer $N > 1$, consider the series $\sum_{n=N+1}^{\infty} a_n = a_{N+1} + a_{N+2} + \ldots$. Then the first series converges iff the second series converges. Here is one (among many) ways to show this formally: write $S_n = a_1 + \ldots + a_n$ and $T_n = a_{N+1} + a_{N+2} + \ldots + a_{N+n}$. Then for all $n \in \mathbb{Z}^+$,

$$\left(\sum_{k=1}^{N} a_k\right) + T_n = a_1 + \ldots + a_N + a_{N+1} + \ldots + a_{N+n} = S_{N+n}.$$

Thus if $\lim_{n\to\infty} T_n = \sum_{n=N+1}^{\infty} a_n$ exists, so does $\lim_{n\to\infty} S_{N+n} = \lim_{n\to\infty} S_n = \sum_{n=1}^{\infty} a_n$. Conversely if $\sum_{n=1}^{\infty} a_n$ exists, then so does $\lim_{n\to\infty} \sum_{k=1}^{N} a_k + T_n = \sum_{k=1}^{n} a_k + \lim_{n\to\infty} T_n$, hence $\lim_{n\to\infty} T_n = \sum_{n=N+1}^{\infty} a_n$ exists.

Similarly, if we are so inclined (and we will be, on occasion), we could add finitely many terms to the series, or for that matter *change* finitely many terms of the series, without affecting the convergence. We record this as follows.

PROPOSITION 11.2. *The addition, removal or altering of any finite number of terms in an infinite series does not affect the convergence or divergence of the series (though of course it may change the sum of a convergent series).*

As the reader has probably already seen for herself, reading someone else's formal proof of this result can be more tedious than enlightening, so we leave it to the reader to construct a proof that she finds satisfactory.

Because the convergence or divergence of a series $\sum_{n=1}^{\infty} a_n$ is not affected by changing the lower limit 1 to any other integer, we often employ a simplified notation $\sum_n a_n$ when discussing series only up to convergence.

PROPOSITION 11.3. *Let $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ be two infinite series, and let α be any real number.*
a) If $\sum_{n=1}^{\infty} a_n = A$ and $\sum_{n=1}^{\infty} b_n = B$ are both convergent, then the series $\sum_{n=1}^{\infty} a_n + b_n$ is also convergent, with sum $A + B$.
b) If $\sum_{n=1}^{\infty} a_n = A$ is convergent, then so also is $\sum_{n=1}^{\infty} \alpha a_n$, with sum αA.

PROOF. a) Let $A_n = a_1 + \ldots + a_n$, $B_n = b_1 + \ldots + b_n$ and $C_n = a_1 + b_1 + \ldots + a_n + b_n$. Then $A_n \to A$ and $B_n \to B$, so for any $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|A_n - A| < \frac{\epsilon}{2}$ and $|B_n - B| < \frac{\epsilon}{2}$. It follows that for all $n \geq N$,

$$|C_n - (A + B)| = |A_n + B_n - A - B| \leq |A_n - A| + |B_n - B| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

b) Left to the reader. □

Exercise: Let $\sum_n a_n$ be an infinite series and $\alpha \in \mathbb{R}$.
a) If $\alpha = 0$, show that $\sum_n \alpha a_n = 0$.
b) Suppose $\alpha \neq 0$. Show $\sum_n a_n$ converges iff $\sum_n \alpha a_n$ converges.

Exercise: Prove the **Three Series Principle**: let $\sum_n a_n$, $\sum_n b_n$, $\sum_n c_n$ be three infinite series with $c_n = a_n + b_n$ for all n. If any two of the three series $\sum_n a_n$, $\sum_n b_n$, $\sum_n c_n$ converge, then so does the third.

2.1. The Nth Term Test.

THEOREM 11.4. *(Nth Term Test) Let $\sum_n a_n$ be an infinite series. If $\sum_n a_n$ converges, then $a_n \to 0$.*

PROOF. Let $S = \sum_{n=1}^{\infty} a_n$. Then for all $n \geq 2$, $a_n = S_n - S_{n-1}$. Therefore

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} S_n - S_{n-1} = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = S - S = 0.$$

□

We will often apply the contrapositive form of Theorem 11.4: if for a series $\sum_n a_n$ we have $a_n \nrightarrow 0$,[4] then $\sum_n a_n$ diverges.

Warning: The converse of Theorem 11.4 is not valid! It may well be the case that $a_n \to 0$ but $\sum_n a_n$ diverges. Later we will see many examples. Still, when put under duress (e.g. while taking an exam) many students can will themselves into believing that the converse might be true. Don't do it!

Exercise: Let $\frac{P(x)}{Q(x)}$ be a rational function. The polynomial $Q(x)$ has only finitely many roots, so we may choose $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $Q(x) \neq 0$. Show that if $\deg P \geq \deg Q$, then $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ is divergent.

2.2. The Cauchy Criterion.

Recall that we proved that a sequence $\{x_n\}$ of real numbers is convergent if and only if it is Cauchy: that is, for all $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$ we have $|x_n - x_m| < \epsilon$.

Applying the Cauchy condition to the sequence of partial sums $\{S_n = a_1 + \ldots + a_n\}$ of an infinite series $\sum_{n=1}^{\infty} a_n$, we get the following result.

PROPOSITION 11.5. *(Cauchy criterion for convergence of series) An infinite series $\sum_{n=1}^{\infty} a_n$ converges if and only if: for every $\epsilon > 0$, there exists $N_0 \in \mathbb{Z}^+$ such that for all $N \geq N_0$ and all $k \in \mathbb{N}$, $|\sum_{n=N}^{N+k} a_n| < \epsilon$.*

Note that taking $k = 0$ in the Cauchy criterion, we recover the Nth Term Test for convergence (Theorem 11.4). It is important to compare these two results: the Nth Term Test gives a very weak *necessary* condition for the convergence of the series. In order to turn this into a necessary and sufficient condition we must require not only that $a_n \to 0$ but also $a_n + a_{n+1} \to 0$ and indeed that $a_n + \ldots + a_{n+k} \to 0$ for a k which is allowed to be (in a certain precise sense) arbitrarily large.

Let us call a sum of the form $\sum_{n=N}^{N+k} = a_N + a_{N+1} + \ldots + a_{N+k}$ a **finite tail** of the series $\sum_{n=1}^{\infty} a_n$. As a matter of notation, if for a fixed $N \in \mathbb{Z}^+$ and all $k \in \mathbb{N}$ we have $|\sum_{n=N}^{N+k} a_n| \leq \epsilon$, let us abbreviate this by

$$|\sum_{n=N}^{\infty} a_n| \leq \epsilon.$$

In other words the *supremum* of the absolute values of the finite tails $|\sum_{n=N}^{N+k} a_n|$ is at most ϵ. This gives a nice way of thinking about the Cauchy criterion.

[4]This means: $a_n \to L \neq 0$, or a_n diverges.

PROPOSITION 11.6. *An infinite series $\sum_{n=1}^{\infty} a_n$ converges if and only if: for all $\epsilon > 0$, there exists $N_0 \in \mathbb{Z}^+$ such that for all $N \geq N_0$, $|\sum_{n=N}^{\infty} a_n| < \epsilon$.*

In other (less precise) words, an infinite series converges iff by removing sufficiently many of the initial terms, we can make what remains arbitrarily small.

3. Series With Non-Negative Terms I: Comparison

3.1. The sum is the supremum.

Starting in this section we get down to business by restricting our attention to series $\sum_{n=1}^{\infty} a_n$ with $a_n \geq 0$ for all $n \in \mathbb{Z}^+$. This simplifies matters considerably and places an array of powerful tests at our disposal.

Why? Assume $a_n \geq 0$ for all $n \in \mathbb{Z}^+$ and consider the sequence of partial sums:

$$S_1 = a_1 \leq a_1 + a_2 = S_2 \leq a_1 + a_2 + a_3 = S_3,$$

and so forth. In general, $S_{n+1} - S_n = a_{n+1} \geq 0$, so the sequence of partial sums $\{S_n\}$ is increasing. Applying the Monotone Sequence Lemma we get:

PROPOSITION 11.7. *Let $\sum_n a_n$ be an infinite series with $a_n \geq 0$ for all n. Then the series converges if and only if the partial sums are bounded above, i.e., if and only if there exists $M \in \mathbb{R}$ such that for all n, $a_1 + \ldots + a_n \leq M$. Moroever if the series converges, its sum is precisely the least upper bound of the sequence of partial sums. If the partial sums are unbounded, the series diverges to ∞.*

Because of this, when dealing with series with non-negative terms we may express convergence by writing $\sum_n a_n < \infty$ and divergence by writing $\sum_n a_n = \infty$.

3.2. The Comparison Test.

Example: Consider the series $\sum_{n=1}^{\infty} \frac{1}{n2^n}$. Its sequence of partial sums is

$$T_n = 1 \cdot \left(\frac{1}{2}\right) + \frac{1}{2} \cdot \left(\frac{1}{4}\right) + \ldots + \frac{1}{n} \cdot \left(\frac{1}{2^n}\right).$$

Unfortunately we do not (yet!) know a closed form expression for T_n, so it is not possible for us to compute $\lim_{n\to\infty} T_n$ directly. But if we just want to decide whether the series converges, we can compare it with the geometric series $\sum_{n=1}^{\infty} \frac{1}{2^n}$:

$$S_n = \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n}.$$

Since $\frac{1}{n} \leq 1$ for all $n \in \mathbb{Z}^+$, we have that for all $n \in \mathbb{Z}^+$, $\frac{1}{n2^n} \leq \frac{1}{2^n}$. Summing these inequalities from $k = 1$ to n gives $T_n \leq S_n$ for all n. By our work with geometric series we know that $S_n \leq 1$ for all n and thus also $T_n \leq 1$ for all n. Therefore our given series has partial sums bounded above by 1, so $\sum_{n=1}^{\infty} \frac{1}{n2^n} \leq 1$. In particular, the series converges.

Example: conside the series $\sum_{n=1}^{\infty} \sqrt{n}$. Again, a closed form expression for $T_n = \sqrt{1} + \ldots + \sqrt{n}$ is not easy to come by. But we don't need it: certainly $T_n \geq 1 + \ldots + 1 = n$. Thus the sequence of partial sums is unbounded, so $\sum_{n=1}^{\infty} \sqrt{n} = \infty$.

THEOREM 11.8. *(Comparison Test) Let $\sum_{n=1}^\infty a_n$, $\sum_{n=1}^\infty b_n$ be two series with non-negative terms, and suppose that $a_n \leq b_n$ for all $n \in \mathbb{Z}^+$. Then*

$$\sum_{n=1}^\infty a_n \leq \sum_{n=1}^\infty b_n.$$

In particular: if $\sum_n b_n < \infty$ then $\sum_n a_n < \infty$, and if $\sum_n a_n = \infty$ then $\sum_n b_n = \infty$.

PROOF. There is really nothing new to say here, but just to be sure: write

$$S_n = a_1 + \ldots + a_n,\ T_n = b_1 + \ldots + b_n.$$

Since $a_k \leq b_k$ for all k we have $S_n \leq T_n$ for all n and thus

$$\sum_{n=1}^\infty a_n = \sup_n S_n \leq \sup_n T_n = \sum_{n=1}^\infty b_n.$$

The assertions about convergence and divergence follow immediately. □

3.3. The Delayed Comparison Test.

The Comparison Test is beautifully simple when it works. It has two weaknesses: first, given a series $\sum_n a_n$ we need to find some other series to compare it to. Thus the test will be more or less effective according to the size of our repertoire of known convergent/divergent series. (At the moment, we don't know much, but that will soon change.) Second, the requirement that $a_n \leq b_n$ for all $n \in \mathbb{Z}^+$ is rather inconveniently strong. Happily, it can be weakened in several ways, resulting in minor variants of the Comparison Test with a much wider range of applicability.

Example: Consider the series

$$\sum_{n=0}^\infty \frac{1}{n!} = 1 + 1 + \frac{1}{2} + \frac{1}{2 \cdot 3} + \frac{1}{2 \cdot 3 \cdot 4} + \ldots + \frac{1}{2 \cdot 3 \cdot \ldots \cdot n} + \ldots.$$

We would like to show that the series converges by comparison, but what to compare it to? Well, there is always the geometric series! Observe that the sequence $n!$ grows faster than any geometric r^n in the sense that $\lim_{n \to \infty} \frac{n!}{r^n} = \infty$. Taking reciprocals, it follows that for any $0 < r < 1$ we will have $\frac{1}{n!} < \frac{1}{r^n}$ – not necessarily for all $n \in \mathbb{Z}^+$, but at least for all sufficiently large n. For instance, one easily establishes by induction that $\frac{1}{n!} < \frac{1}{2^n}$ if and only if $n \geq 4$. Putting $a_n = \frac{1}{n!}$ and $b_n = \frac{1}{2^n}$ we cannot apply the Comparison Test because we have $a_n \geq b_n$ for all $n \geq 4$ rather than for all $n \geq 0$. But this objection is more worthy of a bureaucrat than a mathematician: certainly *the idea* of the Comparison Test is applicable here:

$$\sum_{n=0}^\infty \frac{1}{n!} = \sum_{n=0}^3 \frac{1}{n!} + \sum_{n=4}^\infty \frac{1}{n!} \leq 8/3 + \sum_{n=4}^\infty \frac{1}{2^n} = \frac{8}{3} + \frac{1}{8} = \frac{67}{24} < \infty.$$

So the series converges. More than that, we still retain a quantitative estimate on the sum: it is at most (in fact strictly less than, as a moment's thought will show) $\frac{67}{24} = 2.79166666\ldots$. (Perhaps this reminds you of $e = 2.71828182845904523536028747 14\ldots$, which also happens to be a bit less than $\frac{67}{24}$. It should! More on this later...)

We record the technique of the preceding example as a theorem.

THEOREM 11.9. *(Delayed Comparison Test) Let $\sum_{n=1}^\infty$, $\sum_{n=1}^\infty b_n$ be two series with non-negative terms. Suppose that there exists $N \in \mathbb{Z}^+$ such that for all $n > N$, $a_n \leq b_n$. Then*

$$\sum_{n=1}^\infty a_n \leq \left(\sum_{n=1}^N a_n - b_n\right) + \sum_{n=1}^\infty b_n.$$

In particular: if $\sum_n b_n < \infty$ then $\sum_n a_n < \infty$, and if $\sum_n a_n = \infty$ then $\sum_n b_n = \infty$.

Exercise: Prove Theorem 11.9.

Thus the Delayed Comparison Test assures us that we do not need $a_n \leq b_n$ for all n but only for all sufficiently large n. A different issue occurs when we wish to apply the Comparison Test and the inequalities do not go our way.

3.4. The Limit Comparison Test.

THEOREM 11.10. *(Limit Comparison Test) Let $\sum_n a_n$, $\sum_n b_n$ two series. Suppose that there exists $N \in \mathbb{Z}^+$ and $M \in \mathbb{R}^{\geq 0}$ such that for all $n \geq N$, $0 \leq a_n \leq Mb_n$. Then if $\sum_n b_n$ converges, $\sum_n a_n$ converges.*

Exercise: Prove Theorem 11.10.

COROLLARY 11.11. *(Calculus Student's Limit Comparison Test) Let $\sum_n a_n$ and $\sum_n b_n$ be two series. Suppose that for all sufficiently large n both a_n and b_n are positive and $\lim_{n\to\infty} \frac{a_n}{b_n} = L \in [0, \infty]$.*
a) If $0 < L < \infty$, the series $\sum_n a_n$ and $\sum_n b_n$ converge or diverge together (i.e., either both converge or both diverge).
b) If $L = \infty$ and $\sum_n a_n$ converges, then $\sum_n b_n$ converges.
c) If $L = 0$ and $\sum_n b_n$ converges, then $\sum_n a_n$ converges.

PROOF. In all cases we deduce the result from the Limit Comparison Test.
a) If $0 < L < \infty$, then there exists $N \in \mathbb{Z}^+$ such that $0 < \frac{L}{2} b_n \leq a_n \leq (2L)b_n$. Applying Theorem 11.10 to the second inequality, we get that if $\sum_n b_n$ converges, then $\sum_n a_n$ converges. The first inequality is equivalent to $0 < b_n \leq \frac{2}{L} a_n$ for all $n \geq N$, and applying Theorem 11.10 to this we get that if $\sum_n a_n$ converges, then $\sum_n b_n$ converges. So the two series $\sum_n a_n$, $\sum_n b_n$ converge or diverge together.
b) If $L = \infty$, then there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $a_n \geq b_n \geq 0$. Applying Theorem 11.10 to this we get that if $\sum_n$ converges, then $\sum_n b_n$ converges.
c) This case is left to the reader as an exercise. □

Exercise: Prove Corollary 11.11c).

Example: We will show that for all $p \geq 2$, the **p-series** $\sum_{n=1}^\infty \frac{1}{n^p}$ converges. In fact it is enough to show this for $p = 2$, since for $p > 2$ we have for all $n \in \mathbb{Z}^+$ that $n^2 < n^p$ and thus $\frac{1}{n^p} < \frac{1}{n^2}$ so $\sum_n \frac{1}{n^p} \leq \sum_n \frac{1}{n^2}$. For $p = 2$, we happen to know that

$$\sum_{n=1}^\infty \frac{1}{n^2+n} = \sum_{n=1}^\infty \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1,$$

and in particular that $\sum_n \frac{1}{n^2+n}$ converges. For large n, $\frac{1}{n^2+n}$ is close to $\frac{1}{n^2}$. More precisely, putting $a_n = \frac{1}{n^2+n}$ and $b_n = \frac{1}{n^2}$ we have $a_n \sim b_n$, i.e.,

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{n^2}{n^2+n} = \lim_{n\to\infty} \frac{1}{1+\frac{1}{n}} = 1.$$

Applying Theorem 11.11, we find that $\sum_n \frac{1}{n^2+n}$ and $\sum_n \frac{1}{n^2}$ converge or diverge together. Since the former series converges, we deduce that $\sum_n \frac{1}{n^2}$ converges, even though the Comparison Test does not directly apply.

Exercise: Let $\frac{P(x)}{Q(x)}$ be a rational function such that the degree of the denominator minus the degree of the numerator is at least 2. Show $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ converges.

Exercise: Determine whether each of the following series converges or diverges:
a) $\sum_{n=1}^{\infty} \sin \frac{1}{n^2}$.
b) $\sum_{n=1}^{\infty} \cos \frac{1}{n^2}$.

3.5. Cauchy products I: non-negative terms.

Let $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ be infinite series. Can we, in some sense, multiply them?

In order to forestall possible confusion, let us point out that many students are tempted to consider the following "product" operation on series:

$$\left(\sum_{n=0}^{\infty} a_n\right) \cdot \left(\sum_{n=0}^{\infty} b_n\right) \stackrel{??}{=} \sum_{n=0}^{\infty} a_n b_n.$$

In other words, given two sequences of terms $\{a_n\}$, $\{b_n\}$, we form a new sequence of terms $\{a_n b_n\}$ and then the associated series. In fact this is not a very useful candidate for the product. If $\sum_n a_n = A$ and $\sum_n b_n = B$, we want our "product series" to converge to AB. But for instance, take $\{a_n\} = \{b_n\} = \frac{1}{2^n}$. Then

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} b_n = \frac{1}{1-\frac{1}{2}} = 2,$$

so $AB = 4$, whereas

$$\sum_{n=0}^{\infty} a_n b_n = \sum_{n=0}^{\infty} \frac{1}{4^n} = \frac{1}{1-\frac{1}{4}} = \frac{4}{3}.$$

Unfortunately $\frac{4}{3} \neq 4$. What went wrong?

Plenty! We have ignored the laws of algebra for finite sums: e.g.

$$(a_0 + a_1 + a_2)(b_0 + b_1 + b_2) = a_0b_0 + a_1b_1 + a_2b_2 + a_0b_1 + a_1b_0 + a_0b_2 + a_1b_1 + a_2b_0.$$

The product is different and more complicated – and indeed, if all the terms are positive, strictly larger – than just $a_0b_0 + a_1b_1 + a_2b_2$. We have forgotten about the cross-terms which show up when we multiply one expression involving several terms by another expression involving several terms.[5]

Let us try again at formally multiplying out a product of infinite series:

$$(a_0 + a_1 + \ldots + a_n + \ldots)(b_0 + b_1 + \ldots + b_n + \ldots)$$

$$= a_0b_0 + a_0b_1 + a_1b_0 + a_0b_2 + a_1b_1 + a_2b_0 + \ldots + a_0b_n + a_1b_{n-1} + \ldots + a_nb_0 + \ldots.$$

[5]To the readers who did not forget about the cross-terms: my apologies. But it is a common enough misconception that it had to be addressed.

The notation is getting complicated. In order to shoehorn the right hand side into a single infinite series, we need to either (i) choose some particular ordering of the terms $a_k b_\ell$ or (ii) collect some terms together into an nth term.

For the moment we choose the latter: we define for any $n \in \mathbb{N}$

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} = a_0 b_n + a_1 b_{n-1} + \ldots + a_n b_n$$

and then we define the **Cauchy product** of $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ to be the series

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k b_{n-k} \right).$$

THEOREM 11.12. *Let $\{a_n\}_{n=0}^{\infty}$, $\{b_n\}_{n=0}^{\infty}$ be two series with non-negative terms. Let $\sum_{n=0}^{\infty} a_n = A$ and $\sum_{n=0}^{\infty} b_n = B$. Putting $c_n = \sum_{k=0}^{n} a_k b_{n-k}$ we have that $\sum_{n=0}^{\infty} c_n = AB$. In particular, the Cauchy product series converges iff the two "factor series" $\sum_n a_n$ and $\sum_n b_n$ both converge.*

PROOF. We define another sequence, the **box product**, as follows: for $N \in \mathbb{N}$,

$$\square_N = \sum_{0 \le i,j \le N} a_i b_j = (a_0 + \ldots + a_N)(b_0 + \ldots + b_N) = A_N B_N.$$

Thus by the usual product rule for sequences, we have

$$\lim_{N \to \infty} \square_N = \lim_{N \to \infty} A_N B_N = AB.$$

So the box product clearly converges to the product of the sums of the two series. This suggests that we compare the Cauchy product to the box product. The entries of the box product can be arranged to form a square, viz:

$$\begin{aligned} \square_N = & a_0 b_0 + a_0 b_1 + \ldots + a_0 b_N \\ & + a_1 b_0 + a_1 b_1 + \ldots + a_1 b_N \\ & \vdots \\ & + a_N b_0 + aN b_1 + \ldots + a_N b_N. \end{aligned}$$

On the other hand, the terms of the Nth partial sum of the Cauchy product can naturally be arranged in a triangle:

$$\begin{aligned} C_N = & \quad a_0 b_0 \\ & + a_0 b_1 + a_1 b_0 \\ & + a_0 b_2 + a_1 b_1 + a_2 b_0 \\ & + a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0 \\ & \vdots \\ & + a_0 b_N + a_1 b_{N-1} + a_2 b_{N-2} + \ldots + a_N b_0. \end{aligned}$$

Thus while $\square_N$ is a sum of $(N+1)^2$ terms, C_N is a sum of $1 + 2 + \ldots + N + 1 = \frac{(N+1)(N+2)}{2}$ terms: those lying on our below the diagonal of the square. Thus in considerations involving the Cauchy product, the question is to what extent one can neglect the terms in the upper half of the square – i.e., those with $a_i b_j$ with $i + j > N$ – as N gets large.

Here, since all the a_i's and b_j's are non-negative and $\square_N$ contains all the terms of C_N and others as well, we certainly have

$$C_N \leq \square_N = A_N B_N \leq AB.$$

Thus $C = \lim_{N\to\infty} C_N \leq AB$. For the converse, the key observation is that if we make the sides of the triangle twice as long, it will contain the box: that is, every term of $\square_N$ is of the form $a_i b_j$ with $0 \leq i, j \leq N$; thus $i + j \leq 2N$ so $a_i b_j$ appears as a term in C_{2N}. It follows that $C_{2N} \geq \square_N$ and thus

$$C = \lim_{N\to\infty} C_N = \lim_{N\to\infty} C_{2N} \geq \lim_{N\to\infty} \square_N = \lim_{N\to\infty} A_N B_N = AB.$$

Having shown both that $C \leq AB$ and $C \geq AB$, we conclude

$$C = \sum_{n=0}^{\infty} a_n = AB = \left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right).$$

□

4. Series With Non-Negative Terms II: Condensation and Integration

We have recently been studying criteria for convergence of an infinite series $\sum_n a_n$ which are valid under the assumption that $a_n \geq 0$ for all n. In this section we place ourselves under more restrictive hypotheses: that for all $n \in \mathbb{N}$, $a_{n+1} \geq a_n \geq 0$, i.e., that the sequence of terms is **non-negative** and **decreasing**.

Remark: It's no loss of generality to assume $a_n > 0$ for all n. If not, $a_N = 0$ for some N and, since the terms are assumed decreasing we have $0 = a_N = a_{N+1} = \ldots$ and our infinite series reduces to the finite series $\sum_{n=1}^{N-1} a_n$: this converges!

4.1. The Harmonic Series.

Consider $\sum_{n=1}^{\infty} \frac{1}{n}$, the **harmonic series**. Does it converge? None of the tests we have developed so far are up to the job: especially, $a_n \to 0$ so the Nth Term Test is inconclusive.

Let us take a computational approach by looking at various partial sums. S_{100} is approximately 5.187. Is this close to a familiar real number? Not really. Next we compute $S_{150} \approx 5.591$ and $S_{200} \approx 5.878$. Perhaps the partial sums never exceed 6? (If so, the series would converge.) Let's try a significantly larger partial sums: $S_{1000} \approx 7.485$, so the above guess is incorrect. Since $S_{1050} \approx 7.584$, we are getting the idea that whatever the series is doing, it's doing it rather slowly, so let's instead start stepping up the partial sums multiplicatively:

$$S_{100} \approx 5.878.$$
$$S_{10^3} \approx 7.4854.$$
$$S_{10^4} \approx 9.788.$$
$$S_{10^5} \approx 12.090.$$

Now there is a pattern for the perceptive eye to see: the difference $S_{10^{k+1}} - S_{10^k}$ appears to be approaching $2.30\ldots = \log 10$. This points to $S_n \approx \log n$. If this is so, then since $\log n \to \infty$ the series would diverge. I hope you notice that the relation between $\frac{1}{n}$ and $\log n$ is one of a function and its antiderivative. We ask the reader to hold this thought until we discuss the integral test a bit late on.

For now, we give the following brilliant and elementary argument due to Cauchy.

Consider the terms arranged as follows:

$$\left(\frac{1}{1}\right) + \left(\frac{1}{2} + \frac{1}{3}\right) + \left(\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}\right) + \ldots,$$

i.e., we group the terms in blocks of length 2^k. The power of $\frac{1}{2}$ which begins each block is larger than every term in the preceding block, so if we replaced every term in the current block the the first term in the next block, we would only decrease the sum of the series. But this latter sum is much easier to deal with:

$$\sum_{n=1}^{\infty} \frac{1}{n} \geq \left(\frac{1}{2}\right) + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \ldots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots = \infty.$$

Therefore the harmonic series $\sum_{n=1}^{\infty}$ diverges.

Exercise: Determine the convergence of $\sum_n \frac{1}{n^{1+\frac{1}{n}}}$.

Exercise: Let $\frac{P(x)}{Q(x)}$ be a rational function with $\deg Q - \deg P = 1$. Show that $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ diverges.[6]

We deduce the following result.

PROPOSITION 11.13. *For a rational function $\frac{P(x)}{Q(x)}$, the series $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ converges iff $\deg Q - \deg P \geq 2$.*

PROOF. Left to the reader. □

4.2. The Condensation Test.

The apparently *ad hoc* argument used to prove the divergence of the harmonic series can be adapted to give the following useful test, due to A.L. Cauchy.

THEOREM 11.14. *(Condensation Test) Let $\sum_{n=1}^{\infty} a_n$ be an infinite series such that $a_n \geq a_{n+1} \geq 0$ for all $n \in \mathbb{N}$. Then:*
a) We have $\sum_{n=1}^{\infty} a_n \leq \sum_{n=0}^{\infty} 2^n a_{2^n} \leq 2\sum_{n=1}^{\infty} a_n$.
*b) Thus the series $\sum_n a_n$ converges iff the **condensed series** $\sum_n 2^n a_{2^n}$ converges.*

PROOF. We have

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + \ldots$$

$$\leq a_1 + a_2 + a_2 + a_4 + a_4 + a_4 + a_4 + 8a_8 + \ldots = \sum_{n=0}^{\infty} 2^n a_{2^n}$$

$$= (a_1 + a_2) + (a_2 + a_4 + a_4 + a_4) + (a_4 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8) + (a_8 + \ldots)$$

$$\leq (a_1 + a_1) + (a_2 + a_2 + a_3 + a_3) + (a_4 + a_4 + a_5 + a_5 + a_6 + a_6 + a_7 + a_7) + (a_8 + \ldots)$$

$$= 2\sum_{n=1}^{\infty} a_n.$$

[6]Take N larger than any of the roots of $Q(x)$, so that every term in the sum is well defined.

This establishes part a), and part b) follows immediately. □

The Cauchy Condensation Test is, I think, an *a priori* interesting result: it says that, under the given hypotheses, in order to determine whether a series converges we need to know only a very sparse set of the terms of the series – whatever is happening in between a_{2^n} and $a_{2^{n+1}}$ is immaterial, *so long as the sequence remains decreasing*. This is a very curious phenomenon, and of couse without the hypothesis that the terms are decreasing, nothing like this could hold.

On the other hand, it may be less clear that the Condensation Test is of any practical use: after all, isn't the condensed series $\sum_n 2^n a_{2^n}$ more complicated than the original series $\sum_n a_n$? In fact the opposite is often the case: passing from the given series to the condensed series preserves the convergence or divergence but tends to exchange subtly convergent/divergent series for more obviously (or better: more rapidly) converging/diverging series.

Example: Fix a real number p and consider the **p-series**[7] $\sum_{n=1}^\infty \frac{1}{n^p}$. Our task is to find all values of p for which the series converges.

Step 1: The sequence $a_n = \frac{1}{n^p}$ has positive terms. The terms are decreasing iff the sequence n^p is increasing iff $p > 0$. So we had better treat the cases $p \leq 0$ separately. First, if $p < 0$, then $\lim_{n\to\infty} \frac{1}{n^p} = \lim_{n\to\infty} n^{|p|} = \infty$, so the p-series diverges by the nth term test. Second, if $p = 0$ then our series is simply $\sum_n \frac{1}{n^0} = \sum_n 1 = \infty$. So the p-series "obviously diverges" when $p \leq 0$.

Step 2: Henceforth we assume $p > 0$, so that the hypotheses of Cauchy's Condensation Test apply. We get that $\sum_n n^{-p}$ converges iff $\sum_n 2^n(2^n)^{-p} = \sum_n 2^n 2^{-np} = \sum_n (2^{1-p})^n$ converges. But the latter series is a geometric series with geometric ratio $r = 2^{1-p}$, so it converges iff $|r| < 1$ iff $2^{p-1} > 1$ iff $p > 1$.

Thus we have proved the following important result.

THEOREM 11.15. *For $p \in \mathbb{R}$, the p-series $\sum_n \frac{1}{n^p}$ converges iff $p > 1$.*

Example (p-series continued): Let $p > 1$. By applying part b) of Cauchy's Condensation Test we showed that $\sum_{n=1}^\infty \frac{1}{n^p} < \infty$. What about part a)? It gives an explicit upper bound on the sum of the series, namely

$$\sum_{n=1}^\infty \frac{1}{n^p} \leq \sum_{n=0}^\infty 2^n(2^n)^{-p} = \sum_{n=0}^\infty (2^{1-p})^n = \frac{1}{1-2^{1-p}}.$$

For instance, taking $p = 2$ we get

$$\sum_{n=1}^\infty \frac{1}{n^2} \leq \frac{1}{1-2^{1-2}} = 2.$$

Using a computer algebra package I get

$$1 \leq \sum_{n=1}^{1024} \frac{1}{n^2} = 1.643957981030164240100762569\ldots.$$

[7]Or sometimes: **hyperharmonic series**.

So it seems like $\sum_{n=1}^{\infty} \frac{1}{n^2} \approx 1.64$, whereas the Condensation Test tells us that it is at most 2. (Note that since the terms are positive, simply adding up any finite number of terms gives a lower bound.)

The following exercise gives a technique for using the Condensation Test to estimate $\sum_{n=1}^{\infty} \frac{1}{n^p}$ to arbitrary accuracy.

Exercise: Let N be a non-negative integer.
a) Show that under the hypotheses of the Condensation Test we have

$$\sum_{n=2^N+1}^{\infty} a_n \leq \sum_{n=0}^{\infty} 2^n a_{2^{n+N}}.$$

b) Apply part a) to show that for any $p > 1$,

$$\sum_{n=2^N+1} \frac{1}{n^p} \leq \frac{1}{2^{Np}\left(1 - 2^{1-p}\right)}.$$

Example: $\sum_{n=2}^{\infty} \frac{1}{n \log n}$. $a_n = \frac{1}{n \log n}$ is positive and decreasing (since its reciprocal is positive and increasing) so the Condensation Test applies. We get that the convergence of the series is equivalent to the convergence of

$$\sum_n \frac{2^n}{2^n \log 2^n} = \frac{1}{\log 2} \sum_n \frac{1}{n} = \infty,$$

so the series diverges. This is rather subtle: we know that for any $\epsilon > 0$, $\sum_n \frac{1}{n n^\epsilon}$ converges, since it is a p-series with $p = 1 + \epsilon$. But $\log n$ grows more slowly than n^ϵ for any $\epsilon > 0$, indeed slowly enough so that replacing n^ϵ with $\log n$ converts a convergent series to a divergent one.

Exercise: Determine whether the series $\sum_n \frac{1}{\log(n!)}$ converges.

Exercise: Let p, q, r be positive real numbers.
a) Show that $\sum_n \frac{1}{n(\log n)^q}$ converges iff $q > 1$.
b) Show that $\sum_n \frac{1}{n^p (\log n)^q}$ converges iff $p > 1$ or ($p = 1$ and $q > 1$).
c) Find all values of p, q, r such that $\sum_n \frac{1}{n^p (\log n)^q (\log \log n)^r}$ converges.

The pattern of Exercise X.X can be continued indefinitely, giving series which converge or diverge excruciatingly slowly, and showing that the difference between convergence and divergence can be arbitarily subtle.

4.3. The Integral Test.

THEOREM 11.16. *(Integral Test) Let $f : [1, \infty) \to \mathbb{R}$ be a positive decreasing function, and for $n \in \mathbb{Z}^+$ put $a_n = f(n)$. Then*

$$\sum_{n=2}^{\infty} a_n \leq \int_1^{\infty} f(x)dx \leq \sum_{n=1}^{\infty} a_n.$$

Thus the series $\sum_n a_n$ converges iff the improper integral $\int_1^{\infty} f(x)dx$ converges.

PROOF. This is a rare opportunity in analysis in which a picture supplies a perfectly rigorous proof. Namely, we divide the interval $[1, \infty)$ into subintervals $[n, n+1]$ for all $n \in \mathbb{N}$ and for any $N \in N$ we compare the integral $\int_1^N f(x)dx$ with the upper and lower Riemann sums associated to the partition $\{1, 2, \ldots, N\}$. From the picture one sees immediately that – since f is decreasing – the lower sum is $\sum_{n=2}^{N+1} a_n$ and the upper sum is $\sum_{n=1}^{N} a_n$, so that

$$\sum_{n=2}^{N+1} a_n \leq \int_1^N f(x)dx \leq \sum_{n=1}^{N} a_n.$$

Taking limits as $N \to \infty$, the result follows. □

Remark: The Integral Test is due to Maclaurin[8] [**Ma42**] and later in more modern form to Cauchy [**Ca89**].

Among series which arise naturally in undergraduate analysis, it usually holds that the Condensation Test can be successfully applied to determine convergence / divergence of a series if and only if the Integral Test can be successfully applied.[9]

Example: Let us use the Integral Test to determine the set of $p > 0$ such that $\sum_n \frac{1}{n^p}$ converges. Indeed the series converges iff the improper integral $\int_1^\infty \frac{dx}{x^p}$ is finite. If $p \neq 1$, then we have

$$\int_1^\infty \frac{dx}{x^p} = \frac{x^{1-p}}{1-p}|_{x=1}^{x=\infty}.$$

The upper limit is 0 if $p - 1 < 0 \iff p > 1$ and is ∞ if $p < 1$. Finally,

$$\int_1^\infty \frac{dx}{x} = \log x|_{x=1}^{\infty} = \infty.$$

So, once again, the p-series diverges iff $p > 1$.

Exercise: Verify that all of the above examples involving the Condensation Test can also be done using the Integral Test.

Given the similar applicability of the Condensation and Integral Tests, it is perhaps not so surprising that many texts content themselves to give one or the other. In calculus texts, one almost always finds the Integral Test, which is logical since often integration and then improper integation are covered earlier in the same course in which one studies infinite series. In elementary analysis courses one often develops sequences and series before the study of functions of a real variable, which is logical because a formal treatment of the Riemann integral is necessarily somewhat involved and technical. Thus many of these texts give the Condensation Test.

The Condensation Test and the Integral Test have a similar range of applicability: in most "textbook" examples, where one test succeeeds, so will the other. From an aesthetic standpoint, the Condensation Test is more appealing (to me). On the other hand, under a mild additional hypothesis the Integral Test can be used to give **asymptotic expansions** for divergent series. Our treatment of the

[8]Colin Maclaurin, 1698-1746

[9]Why this should be so is not clear to me: the observation is purely empirical.

next two results owes a debt to [**CdD**].

Let us first establish some notation: Suppose that $f, g : [a, \infty) \to \mathbb{R}$ and that $g(x) \neq 0$ for all sufficiently large x. We write $f \sim g$ and say **f is asymptotic to g** if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. Similarly, given sequences $\{a_n\}, \{b_n\}$ with $b_n \neq 0$ for all sufficiently large n, we write $a_n \sim b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$.

Exercise: Let f and g be nonzero polynomial functions. Show that $f \sim g$ iff f and g have the same degree and the same leading coefficient.

LEMMA 11.17. *Let $\{a_n\}$ and $\{b_n\}$ be two sequences of positive real numbers with $a_n \sim b_n$ and $\sum_n a_n = \infty$. Then $\sum_n b_n = \infty$ and $\sum_{n=1}^N a_n \sim \sum_{n=1}^N b_n$.*

PROOF. By the Limit Comparison Test, $\sum_n a_n = \infty$. Now fix $\epsilon > 0$ and choose $K \in \mathbb{Z}^+$ such that for all $n \geq K$ we have $a_n \leq (1+\epsilon)b_n$. Then for $N \geq K$,

$$\sum_{n=1}^N a_n = \sum_{n=1}^{K-1} a_n + \sum_{n=K}^N a_n \leq \sum_{n=1}^{K-1} a_n + \sum_{n=K}^N (1+\epsilon)b_n$$

$$= \left(\sum_{n=1}^{K_1} a_n - \sum_{n=1}^{K-1}(1+\epsilon)b_n\right) + \sum_{n=1}^N (1+\epsilon)b_n = C_{\epsilon,K} + (1+\epsilon)\sum_{n=1}^N b_n,$$

say, where $C_{\epsilon,K}$ does not depend on N. Dividing both sides by $\sum_{n=1}^N b_n$ and using $\lim_{N\to\infty} \sum_{n=1}^N b_n = \infty$, we find that $\frac{\sum_{n=1}^N a_n}{\sum_{n=1}^N b_n}$ is at most $1 + 2\epsilon$ for all sufficiently large N. Because our hypotheses are symmetric in $\sum_n a_n$ and $\sum_n b_n$, we also have that $\frac{\sum_{n=1}^N b_n}{\sum_{n=1}^N a_n}$ is at most $1 + 2\epsilon$ for all suffiicently large N. It follows that

$$\lim_{N\to\infty} \frac{\sum_{n=1}^N a_n}{\sum_{n=1}^N b_n} = 1.$$

□

THEOREM 11.18. *Let $f : [1, \infty) \to (0, \infty)$ be continuous and monotone. Suppose the series $\sum_n f(n)$ diverges and that as $x \to \infty$, $f(x) \sim f(x+1)$. Then*

$$\sum_{n=1}^N f(n) \sim \int_1^N f(x)dx.$$

PROOF. Case 1: Suppose f is increasing. Then, for $n \leq x \leq n+1$, we have $f(n) \leq \int_n^{n+1} f(x)dx \leq f(n+1)$, or

$$1 \leq \frac{\int_n^{n+1} f(x)dx}{f(n)} \leq \frac{f(n+1)}{f(n)}.$$

By assumption we have

$$\lim_{n\to\infty} \frac{f(n+1)}{f(n)} = 1,$$

so by the Squeeze Principle we have

$$\int_n^{n+1} f(x)dx \sim f(n). \tag{82}$$

Applying Lemma 11.17 with $a_n = f(n)$ and $b_n = \int_n^{n+1} f(x)dx$, we conclude

$$\int_1^{N+1} f(x)dx = \sum_{k=1}^{N} \int_k^{k+1} f(x)dx \sim \sum_{n=1}^{N} f(n).$$

Further, we have

$$\lim_{N\to\infty} \frac{\int_1^{N+1} f(x)dx}{\int_1^N f(x)dx} = \frac{\infty}{\infty} \stackrel{*}{=} \lim_{N\to\infty} \frac{f(N+1)}{f(N)} = 1,$$

where in the starred equality we have applied L'Hopital's Rule and then the Fundamental Theorem of Calculus. We conclude

$$\int_1^{N} f(x)dx \sim \int_1^{N+1} f(x)dx \sim \sum_{n=1}^{N} f(n).$$

Case 2: Suppose f is decreasing. Then for $n \le x \le n+1$, we have

$$f(n+1) \le \int_n^{n+1} f(x)dx \le f(n),$$

or

$$\frac{f(n+1)}{f(n)} \le \frac{\int_n^{n+1} f(x)dx}{f(n)} \le 1.$$

Once again, by our assumption that $f(n) \sim f(n+1)$ and the Squeeze Principle we get (82); the remainder of the proof proceeds as in the previous case. □

5. Series With Non-Negative Terms III: Ratios and Roots

We continue our analysis of series $\sum_n a_n$ with $a_n \ge 0$ for all n. In this section we introduce two important tests based on a very simple – yet powerful – idea: if for sufficiently large n a_n is bounded above by a non-negative constant M times r^n for $0 \le r < 1$, then the series converges by comparison to the convergent geometric series $\sum_n Mr^n$. Conversely, if for sufficiently large n a_n is bounded below by a positive constant M times r^n for $r \ge 1$, then the series diverges by comparison to the divergent geometric series $\sum_n Mr^n$.

5.1. The Ratio Test.

THEOREM 11.19. *(Ratio Test) Let $\sum_n a_n$ be a series with $a_n > 0$ for all n.*
a) Suppose there exists $N \in \mathbb{Z}^+$ and $0 < r < 1$ such that for all $n \ge N$, $\frac{a_{n+1}}{a_n} \le r$. Then the series $\sum_n a_n$ converges.
b) Suppose there exists $N \in \mathbb{Z}^+$ and $r \ge 1$ such that for all $n \ge N$, $\frac{a_{n+1}}{a_n} \ge r$. Then the series $\sum_n a_n$ diverges.
c) The hypothesis of part a) holds if $\rho = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}$ exists and is less than 1.
d) The hypothesis of part b) holds if $\rho = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}$ exists and is greater than 1.

PROOF. a) Our assumption is that for all $n \ge N$, $\frac{a_{n+1}}{a_n} \le r < 1$. Then $\frac{a_{n+2}}{a_n} = \frac{a_{n+2}}{a_{n+1}} \frac{a_{n+1}}{a_n} \le r^2$. An easy induction argument shows that for all $k \in \mathbb{N}$,

$$\frac{a_{N+k}}{a_N} \le r^k,$$

so

$$a_{N+k} \le a_N r^k.$$

Summing these inequalities gives

$$\sum_{k=N}^{\infty} a_k = \sum_{k=0}^{\infty} a_{N+k} \leq \sum_{k=0}^{\infty} a_N r^k < \infty,$$

so the series $\sum_n a_n$ converges by comparison.
b) Similarly, our assumption is that for all $n \geq N$, $\frac{a_{n+1}}{a_n} \geq r \geq 1$. As above, it follows that for all $k \in \mathbb{N}$,

$$\frac{a_{N+k}}{a_N} \geq r^k,$$

so

$$a_{N+k} \geq a_N r^k \geq a_N > 0.$$

It follows that $a_n \nrightarrow 0$, so the series diverges by the Nth Term Test.
We leave the proofs of parts c) and d) as exercises. □

Exercise: Prove parts c) and d) of Theorem 11.19.

Example: Let $x > 0$. We will show that the series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ converges. (Recall we showed this earlier when $x = 1$.) We consider the quantity

$$\frac{a_{n+1}}{a_n} = \frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}} = \frac{x}{n+1}.$$

It follows that $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = 0$. Thus the series converges for any $x > 0$.

5.2. The Root Test.

In this section we give a variant of the Ratio Test. Instead of focusing on the property that the geometric series $\sum_n r^n$ has constant ratios of consecutive terms, we observe that the nth root of the nth term is equal to r. Suppose now that $\sum_n a_n$ is a series with non-negative terms such that $a_n^{\frac{1}{n}} \leq r$ for some $r < 1$. Raising both sides to the nth power gives $a_n \leq r^n$, and once again we find that the series converges by comparison to a geometric series.

THEOREM 11.20. *(Root Test) Let $\sum_n a_n$ be a series with $a_n \geq 0$ for all n.*
a) Suppose there exists $N \in \mathbb{Z}^+$ and $0 < r < 1$ such that for all $n \geq N$, $a_n^{\frac{1}{n}} \leq r$. Then the series $\sum_n a_n$ converges.
b) Suppose that for infinitely many positive integers n we have $a_n^{\frac{1}{n}} \geq 1$. Then the series $\sum_n a_n$ diverges.
c) The hypothesis of part a) holds if $\rho = \lim_{n\to\infty} a_n^{\frac{1}{n}}$ exists and is less than 1.
d) The hypothesis of part b) holds if $\rho = \lim_{n\to\infty} a_n^{\frac{1}{n}}$ exists and is greater than 1.

Exercise: Prove Theorem 11.20.

5.3. Ratios versus Roots.

It is a fact – a piece of calculus folklore – that the Root Test is *stronger* than the Ratio Test. That is, whenever the ratio test succeeds in determining the convergence or divergence of a series, the root test will also succeed.

In order to explain this result we need to make use of the limit infimum and limit

supremum. First we recast the ratio and root tests in those terms.

Exercise: Let $\sum_n a_n$ be a series with positive terms. Put

$$\underline{\rho} = \liminf_{n\to\infty} \frac{a_{n+1}}{a_n},\ \overline{\rho} = \limsup_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

a) Show that if $\overline{\rho} < 1$, the series $\sum_n a_n$ converges.
b) Show that if $\underline{\rho} > 1$ the series $\sum_n a_n$ diverges.

Exercise: Let $\sum_n a_n$ be a series with non-negative terms. Put

$$\overline{\theta} = \limsup_{n\to\infty} a_n^{\frac{1}{n}}.$$

a) Show that if $\overline{\theta} < 1$, the series $\sum_n a_n$ converges.
b) Show that if $\overline{\theta} > 1$, the series $\sum_n a_n$ diverges.[10]

Exercise: Consider the following conditions on a real sequence $\{x_n\}_{n=1}^{\infty}$:
(i) $\limsup_{n\to\infty} x_n > 1$.
(ii) For infinitely many n, $x_n \geq 1$.
(iii) $\limsup_{n\to\infty} x_n \geq 1$.
a) Show that $(i) \implies (ii) \implies (iii)$ and that neither implication can be reversed.
b) Explain why the result of part b) of the previous Exercise is weaker than part b) of Theorem 11.20.
c) Give an example of a non-negative series $\sum_n a_n$ with $\overline{\theta} = \limsup_{n\to\infty} a_n^{\frac{1}{n}} = 1$ such that $\sum_n a_n = \infty$.

PROPOSITION 11.21. *For any series $\sum_n a_n$ with positive terms, we have*

$$\underline{\rho} = \liminf_{n\to\infty} \frac{a_{n+1}}{a_n} \leq \underline{\theta} = \liminf_{n\to\infty} a_n^{\frac{1}{n}} \leq \overline{\theta} \limsup_{n\to\infty} a_n^{\frac{1}{n}} \leq \overline{\rho} = \limsup_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

Exercise: Let A and B be real numbers with the following property: for any real number r, if $A < r$ then $B \leq r$. Show that $B \leq A$.

PROOF. Step 1: For any sequence $\{x_n\}$, $\liminf x_n \leq \limsup x_n$, hence $\underline{\theta} \leq \overline{\theta}$.
Step 2: We show that $\overline{\theta} \leq \overline{\rho}$. For this, suppose $r > \overline{\rho}$, so that for all sufficiently large n, $\frac{a_{n+1}}{a_n} \leq r$. As in the proof of the Ratio Test, we have $a_{n+k} < r^k a_n$ for all $k \in \mathbb{N}$. We may rewrite this as

$$a_{n+k} < r^{n+k} \left(\frac{a_n}{r^n}\right),$$

or

$$a_{n+k}^{\frac{1}{n+k}} < r \left(\frac{a_n}{r^n}\right)^{\frac{1}{n+k}}.$$

Now

$$\overline{\theta} = \limsup_{n\to\infty} a_n^{\frac{1}{n}} = \limsup_{k\to\infty} a_{n+k}^{\frac{1}{k}} \leq \limsup_{k\to\infty} r \left(\frac{a_n}{r^n}\right)^{\frac{1}{n+k}} = r.$$

By the preceding exercise, we conclude $\overline{\theta} \leq \overline{\rho}$.
Step 3: We must show that $\underline{\rho} \leq \underline{\theta}$. This is very similar to the argument of Step 2, and we leave it as an exercise. □

[10]This is not a typo: we really mean the limsup both times, unlike in the previous exercise.

Exercise: Give the details of Step 3 in the proof of Proposition 11.21.

Now let $\sum_n a_n$ be a series which the Ratio Test succeeds in showing is convergent: that is, $\overline{\rho} < 1$. Then by Proposition 11.21, we have $\overline{\theta} \leq \overline{\rho} \leq 1$, so the Root Test also shows that the series is convergent. Now suppose that the Ratio Test succeeds in showing that the series is divergent: that is $\underline{\rho} > 1$. Then $\overline{\theta} \geq \underline{\theta} \geq \underline{\rho} > 1$, so the Root Test also shows that the series is divergent.

Exercise: Consider the series $\sum_n 2^{-n+(-1)^n}$.
a) Show that $\underline{\rho} = \frac{1}{8}$ and $\overline{\rho} = 2$, so the Ratio Test fails.
b) Show that $\underline{\theta} = \overline{\theta} = \frac{1}{2}$, so the Root Test shows that the series converges.

Exercise: Construct further examples of series for which the Ratio Test fails but the Root Test succeeds to show either convergence or divergence.

Warning: The sense in which the Root Test is stronger than the Ratio Test is a theoretical one. For a given relatively benign series, it may well be the case that the Ratio Test is *easier to apply* than the Root Test, even though in theory whenever the Ratio Test works the Root Test must also work.

Example: Consider again the series $\sum_{n=0}^{\infty} \frac{1}{n!}$. In the presence of factorials one should always attempt the Ratio Test first. Indeed

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lim_{n\to\infty} \frac{1/(n+1)!}{1/n!} = \lim_{n\to\infty} \frac{n!}{(n+1)n!} = \lim_{n\to\infty} \frac{1}{n+1} = 0.$$

Thus the Ratio Test *limit* exists (no need for liminfs or limsups) and is equal to 0, so the series converges. If instead we tried the Root Test we would have to evaluate $\lim_{n\to\infty} (\frac{1}{n!})^{\frac{1}{n}}$. This is not so bad if we keep our head – e.g. one can show that for any fixed $R > 0$ and sufficiently large n, $n! > R^n$ and thus $(\frac{1}{n!})^{\frac{1}{n}} \leq (\frac{1}{R^n})^{\frac{1}{n}} = \frac{1}{R}$. Thus the root test limit is at most $\frac{1}{R}$ for any positive R, so it is 0. But this is elaborate compared to the Ratio Test computation, which was immediate.

Turning these ideas around, Proposition 11.21 can be put to the following sneaky use.

COROLLARY 11.22. *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers. Assume that $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} \to L \in [0, \infty]$. Then also $\lim_{n\to\infty} a_n^{\frac{1}{n}} = L$.*

PROOF. Indeed, the hypothesis gives that for the infinite series $\sum_n a_n$ we have $\rho = L$, so by Proposition 11.21 we must also have $\theta = L$. □

Exercise: Use Corollary 11.22 to evaluate the following limits:
a) $\lim_{n\to\infty} n^{\frac{1}{n}}$.
b) For $\alpha \in \mathbb{R}$, $\lim_{n\to\infty} n^{\frac{\alpha}{n}}$.
c) $\lim_{n\to\infty} (n!)^{\frac{1}{n}}$.

6. Absolute Convergence

6.1. Introduction to absolute convergence.

We turn now to the serious study of series with both positive and negative terms.

It turns out that under one relatively mild additional hypothesis, virtually all of our work on series with non-negative terms can be usefully applied in this case. In this section we study this wonderful hypothesis: absolute convergence. (In the next section we get *really* serious by studying series in the absence of absolute convergence. This will lead to surprisingly delicate and intricate considerations.)

A real series $\sum_n a_n$ is **absolutely convergent** if $\sum_n |a_n|$ converges. Note that $\sum_n |a_n|$ is a series with non-negative terms, so to decide whether it is convergent we may use all the tools of the last three sections. A series $\sum_n a_n$ which converges but for which $\sum_n |a_n|$ diverges is said to be **nonabsolutely convergent**.[11]

The terminology *absolutely convergent* suggests that the convergence of the series $\sum_n |a_n|$ is somehow "better" than the convergence of the series $\sum_n a_n$. This is indeed the case, although it is not obvious. But the following result already clarifies matters a great deal.

PROPOSITION 11.23. *Every absolutely convergent real series is convergent.*

PROOF. We shall give *two proofs* of this important result.
First Proof: Consider the three series $\sum_n a_n$, $\sum_n |a_n|$ and $\sum_n a_n + |a_n|$. Our hypothesis is that $\sum_n |a_n|$ converges. But we claim that this implies that $\sum_n a_n + |a_n|$ converges as well. Indeed, consider the expression $a_n + |a_n|$: it is equal to $2a_n = 2|a_n|$ when a_n is non-negative and 0 when a_n is negative. In particular the series $\sum_n a_n + |a_n|$ has non-negative terms and $\sum_n a_n + |a_n| \leq \sum_n 2|a_n| < \infty$. So $\sum_n a_n + |a_n|$ converges. By the Three Series Principle, $\sum_n a_n$ converges.
Second Proof: The above argument is clever – maybe too clever! Let's try something more fundamental: since $\sum_n |a_n|$ converges, for every $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $\sum_{n=N}^{\infty} |a_n| < \epsilon$. Therefore for all $k \geq 0$,

$$|\sum_{n=N}^{N+k} a_n| \leq \sum_{n=N}^{N+k} |a_n| \leq \sum_{n=N}^{\infty} |a_n| \leq \epsilon,$$

and $\sum_n a_n$ converges by the Cauchy criterion. □

Exercise: If $\sum_{n=0}^{\infty} a_n$ is absolutely convergent, show that $|\sum_{n=0}^{\infty} a_n| \leq \sum_{n=0}^{\infty} |a_n|$.

Exercise: Find a sequence $\{a_n\}_{n=1}^{\infty}$ of rational numbers such that $\sum_{n=1}^{\infty} |a_n|$ is a rational number but $\sum_{n=1}^{\infty} a_n$ is an irrational number.

The main idea is that when we try to extend our rich array of convergence tests for non-negative series to series with both positive and negative terms, a sufficient (and often necessary) hypothesis is that the series be not just convergent but absolutely convergent.

[11]We warn the reader that the more standard terminology is **conditionally convergent**. We will later on give a separate definition for "conditionally convergent" and then it will be a *theorem* that a real series is conditionally convergent if and only if it is nonabsolutely convergent. The reasoning for this – which we admit will seem abstruse at best to our target audience – is that in functional analysis one studies convergence and absolute convergence of series in a more general context, such that nonabsolute converge and conditional convergence may indeed differ.

Exercise: State and prove a Comparison Test and Limit Comparison Test for absolute convergence.

THEOREM 11.24. *(Ash [**As12**]) Let $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ be real sequences such that for all $n \in \mathbb{Z}^+$ we have $b_n \neq 0$. Suppose:*
(i) $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$, *and*
(ii) $\sum_{n=1}^\infty a_n$ *diverges and* $\sum_{n=1}^\infty b_n$ *converges.*
Then: $\sum_{n=1}^\infty |b_n| = \infty$ *and* $\lim_{n\to\infty} \frac{a_n - b_n}{b_n} = 0$.

PROOF. If $\sum_{n=1}^\infty |b_n| < \infty$, then applying the limit comparison test to the absolute series $\sum_{n=1}^\infty |a_n|$ and $\sum_{n=1}^\infty |b_n|$, we get that $\sum_{n=1}^\infty a_n$ is absolutely convergent and thus convergent, contradicting (ii). Moreover

$$\lim_{n\to\infty} \frac{a_n - b_n}{b_n} = \lim_{n\to\infty} \frac{a_n}{b_n} - 1 = 0.$$

□

Exercise: Show that if we weaken (i) to $\lim_{n\to\infty} |\frac{a_n}{b_n}| = L \in (0, \infty)$, then still $\sum_{n=1}^\infty b_n$ is not absolutely convergent.

As an example of how Theorem 11.23 may be combined with the previous tests to give tests for absolute convergence, we record the following result.

THEOREM 11.25. *(Ratio & Root Tests for Absolute Convergence) Let $\sum_n a_n$ be a real series.*
a) Assume $a_n \neq 0$ for all n. If there exists $0 \leq r < 1$ such that for all sufficiently large n, $|\frac{a_{n+1}}{a_n}| \leq r$, then the series $\sum_n a_n$ is absolutely convergent.
b) Assume $a_n \neq 0$ for all n. If there exists $r > 1$ such that for all sufficiently large n, $|\frac{a_{n+1}}{a_n}| \geq r$, the series $\sum_n a_n$ is divergent.
c) If there exists $r < 1$ such that for all sufficiently large n, $|a_n|^{\frac{1}{n}} \leq r$, the series $\sum_n a_n$ is absolutely convergent.
d) If there are infinitely many n for which $|a_n|^{\frac{1}{n}} \geq 1$, then the series diverges.

PROOF. Parts a) and c) are immediate: applying Theorem 11.19 (resp. Theorem 11.20) we find that $\sum_n |a_n|$ is convergent – and the point is that by Theorem 11.23, this implies that $\sum_n a_n$ is convergent.

There is something to say in parts b) and d), because in general just because $\sum_n |a_n| = \infty$ does not imply that $\sum_n a_n$ diverges. (We will study this subtlety later on in detail.) But recall that whenever the Ratio or Root tests establish the divergence of a non-negative series $\sum_n b_n$, they do so by showing that $b_n \not\to 0$. Thus under the hypotheses of parts b) and d) we have $|a_n| \not\to 0$, hence also $a_n \not\to 0$ so $\sum_n a_n$ diverges by the Nth Term Test (Theorem 11.4). □

In particular, for a real series $\sum_n a_n$ define the following quantities:

$$\rho = \lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| \text{ when it exists,}$$
$$\underline{\rho} = \liminf_{n\to\infty} |\frac{a_{n+1}}{a_n}|,$$
$$\overline{\rho} = \limsup_{n\to\infty} |\frac{a_{n+1}}{a_n}|,$$
$$\theta = \lim_{n\to\infty} |a_n|^{\frac{1}{n}} \text{ when it exists,}$$

$$\overline{\theta} = \limsup_{n\to\infty} |a_n|^{\frac{1}{n}},$$

and then all previous material on Ratio and Root Tests applies to all real series.

THEOREM 11.26. *Let $\sum_{n=0}^{\infty} a_n = A$ and $\sum_{n=0}^{\infty} b_n = B$ be two absolutely convergent series, and let $c_n = \sum_{k=0}^{n} a_k b_{n-k}$. Then the Cauchy product series $\sum_{n=0}^{\infty} c_n$ is absolutely convergent, with sum AB.*

PROOF. We have proved this result already when $a_n, b_n \geq 0$ for all n. We wish, of course, to reduce to that case. As far as the convergence of the Cauchy product, this is completely straightforward: we have

$$\sum_{n=0}^{\infty} |c_n| = \sum_{n=0}^{\infty} |\sum_{k=0}^{n} a_k b_{n-k}| \leq \sum_{n=0}^{\infty}\sum_{k=0}^{n} |a_k||b_{n-k}| < \infty,$$

the last inequality following from the fact that $\sum_{n=0}^{\infty}\sum_{k=0}^{n} |a_k||b_{n-k}|$ is the Cauchy product of the two non-negative series $\sum_{n=0}^{\infty} |a_n|$ and $\sum_{n=0}^{\infty} |b_n|$, hence it converges. Therefore $\sum_n |c_n|$ converges by comparison, so the Cauchy product series $\sum_n c_n$ converges.

We now wish to show that $\lim_{N\to\infty} C_N = \sum_{n=0}^{\infty} c_n = AB$. Recall the notation

$$\square_N = \sum_{0\leq i,j\leq N} a_i b_j = (a_0 + \ldots + a_N)(b_0 + \ldots + b_N) = A_N B_N.$$

We have

$$|C_N - AB| \leq |\square_N - AB| + |\square_N - C_N|$$

$$= |A_N B_N - AB| + |a_1 b_N| + |a_2 b_{N-1}| + |a_2 b_N| + \ldots + |a_N b_1| + \ \ldots + |a_N b_N|$$

$$\leq |A_N B_N - AB| + \left(\sum_{n=0}^{\infty} |a_n|\right)\left(\sum_{n\geq N} |b_n|\right) + \left(\sum_{n=0}^{\infty} |b_n|\right)\left(\sum_{n\geq N} |a_n|\right).$$

Fix $\epsilon > 0$; since $A_N B_N \to AB$, for sufficiently large N $|A_N B_N - AB| < \frac{\epsilon}{3}$. Put

$$\mathbb{A} = \sum_{n=0}^{\infty} |a_n|, \ \mathbb{B} = \sum_{n=0}^{\infty} |b_n|.$$

By the Cauchy criterion, for sufficiently large N we have $\sum_{n\geq N} |b_n| < \frac{\epsilon}{3\mathbb{A}}$ and $\sum_{n\geq N} |a_n| < \frac{\epsilon}{3\mathbb{B}}$ and thus $|C_N - AB| < \epsilon$. □

6.2. Cauchy products II: Mertens's Theorem.

While the proof of Theorem 11.26 may seem rather long, it is in fact a rather straightforward argument: one shows that the difference between the "box product" and the partial sums of the Cauchy product becomes negligible as N tends to infinity. In less space but with a bit more finesse, one can prove the following stronger result, a theorem of F. Mertens [**Me72**].[12]

THEOREM 11.27. *(Mertens' Theorem) Let $\sum_{n=0}^{\infty} a_n = A$ be an absolutely convergent series and $\sum_{n=0}^{\infty} b_n = B$ be a convergent series. Then the Cauchy product series $\sum_{n=0}^{\infty} c_n$ converges to AB.*

[12]Franz Carl Joseph Mertens, 1840-1927

PROOF. (Rudin [**R**, Thm. 3.50]): define (as usual)

$$A_N = \sum_{n=0}^{N} a_n,\ B_N = \sum_{n=0}^{N} b_n,\ C_N = \sum_{n=0}^{N} c_n$$

and also (for the first time)

$$\beta_n = B_n - B.$$

Then for all $N \in \mathbb{N}$,

$$C_N = a_0b_0 + (a_0b_1 + a_1b_0) + \ldots + (a_0b_N + \ldots + a_Nb_0)$$

$$= a_0B_N + a_1B_{N-1} + \ldots + a_NB_0$$

$$= a_0(B + \beta_N) + a_1(B + \beta_{N-1}) + \ldots + a_N(B + \beta_0)$$

$$= A_NB + a_0\beta_N + a_1\beta_{N-1} + \ldots + a_N\beta_0 = A_NB + \gamma_N,$$

say, where $\gamma_N = a_0\beta_N + a_1\beta_{N-1} + \ldots + a_n\beta_0$. Since our goal is to show that $C_N \to AB$ and we know that $A_NB \to AB$, it suffices to show that $\gamma_N \to 0$. Now, put $\alpha = \sum_{n=0}^{\infty} |a_n|$. Since $B_N \to B$, $\beta_N \to 0$, and thus for any $\epsilon > 0$ we may choose $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ we have $|\beta_n| \leq \frac{\epsilon}{2\alpha}$. Put

$$M = \max_{0 \leq n \leq N_0} |\beta_n|.$$

By the Cauchy criterion, for all sufficiently large N, $M \sum_{n \geq N - N_0} |a_n| \leq \epsilon/2$. Then

$$|\gamma_N| \leq |\beta_0a_N + \ldots + \beta_{N_0}a_{N-N_0}| + |\beta_{N_0+1}a_{N-N_0-1} + \ldots + \beta_Na_0|$$

$$\leq |\beta_0a_N + \ldots + \beta_{N_0}a_{N-N_0}| + \frac{\epsilon}{2} \leq M\left(\sum_{n \geq N-N_0} |a_n|\right) + \frac{\epsilon}{2} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

So $\gamma_N \to 0$. □

7. Non-Absolute Convergence

We say that a real series $\sum_n a_n$ is **nonabsolutely convergent** if the series converges but $\sum_n |a_n|$ diverges, thus if it is convergent but not absolutely convergent.[13]

A series which is nonabsolutely convergent is a more delicate creature than any we have studied thus far. A test which can show that a series is convergent but nonabsolutely convergent is necessarily subtler than those of the previous section. In fact the typical undergraduate student of calculus / analysis learns exactly one such test, which we give in the next section.

[13]One therefore has to distinguish between the phrases "not absolutely convergent" and "nonabsolutely convergent": the former allows the possibility that the series is divergent, whereas the latter does not. In fact our terminology here is not completely standard. We defend ourselves grammatically: "nonabsolutely" is an adverb, so it must modify "convergent", i.e., it describes *how* the series converges.

7.1. The Alternating Series Test.

Consider the **alternating harmonic series**

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \dots.$$

Upon taking the absolute value of every term we get the usual harmonic series, which diverges, so the alternating harmonic series is *not* absolutely convergent. However, some computations with partial sums suggests that the alternating harmonic series *is* convergent, with sum $\log 2$. By looking more carefully at the partial sums, we can find a pattern that allows us to show that the series does indeed converge. (Whether it converges to $\log 2$ is a different matter, of course, which we will revisit much later on.)

It will be convenient to write $a_n = \frac{1}{n}$, so that the alternating harmonic series is $\sum_n \frac{(-1)^{n+1}}{n+1}$. We draw the reader's attention to three properties of this series:

(AST1) The terms alternate in sign.
(AST2) The nth term approaches 0.
(AST3) The sequence of absolute values of the terms is weakly decreasing:

$$a_1 \geq a_2 \geq \ldots \geq a_n \geq \dots.$$

These are the clues from which we will make our case for convergence. Here it is: consider the process of passing from the first partial sum $S_1 = 1$ to $S_3 = 1 - \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$. We have $S_3 \leq 1$, and this is no accident: since $a_2 \geq a_3$, subtracting a_2 and then adding a_3 leaves us no larger than where we started. But indeed this argument is valid in passing from any S_{2n-1} to S_{2n+1}:

$$S_{2n+1} = S_{2n-1} - (a_{2n} - a_{2n+1}) \leq S_{2n-1}.$$

Thus the sequence of odd-numbered partial sums $\{S_{2n-1}\}$ is decreasing. Moreover,

$$S_{2n+1} = (a_1 - a_2) + (a_3 - a_4) + \ldots + (a_{2n-1}| - |a_{2n}) + a_{2n-1} \geq 0.$$

Therefore all the odd-numbered terms are bounded below by 0. By the Monotone Sequence Lemma, the sequence $\{S_{2n+1}\}$ converges to its greatest lower bound, say S_{odd}. On the other hand, just the opposite sort of thing happens for the even-numbered partial sums:

$$S_{2n+2} = S_{2n} + a_{2n+1} - a_{2n+2} \geq S_{2n}$$

and

$$S_{2n+2} = a_1 - (a_2 - a_3) - (a_4 - a_5) - \ldots - (a_{2n} - a_{2n+1}|) - a_{2n+2} \leq a_1.$$

Therfore the sequence of even-numbered partial sums $\{S_{2n}\}$ is increasing and bounded above by a_1, so it converges to its least upper bound, say S_{even}. Thus we have split up our sequence of partial sums into two complementary subsequences and found that each of these series converges. Now the sequence $\{S_n\}$ converges iff $S_{\text{odd}} = S_{\text{even}}$, and the inequalities

$$S_2 \leq S_4 \leq \ldots \leq S_{2n} \leq S_{2n+1} \leq S_{2n-1} \leq \ldots \leq S_3 \leq S_1$$

show that $S_{\text{even}} \leq S_{\text{odd}}$. Moreover, for any $n \in \mathbb{Z}^+$ we have

$$S_{\text{odd}} - S_{\text{even}} \leq S_{2n+1} - S_{2n} = a_{2n+1}.$$

Since $a_{2n+1} \to 0$, we conclude $S_{\text{odd}} = S_{\text{even}} = S$, i.e., the series converges.

Further, since for all n we have $S_{2n} \leq S_{2n+2} \leq S \leq S_{2n+1}$, it follows that

$$|S - S_{2n}| = S - S_{2n} \leq S_{2n+1} - S_{2n} = a_{2n+1}$$

and similarly

$$|S - S_{2n+1}| = S_{2n+1} - S \leq S_{2n+1} - S_{2n+2} = a_{2n+2}.$$

Thus the error in cutting off the infinite sum $\sum_{n=1}^{\infty}(-1)^{n+1}|a_n|$ after N terms is in absolute value at most the absolute value of the next term: a_{N+1}.

Of course in all this we never used that $a_n = \frac{1}{n}$ but only that we had a series satisfying (AST1) (i.e., an alternating series), (AST2) and (AST3). Therefore the preceding arguments have in fact proved the following more general result, due originally due to Leibniz.

THEOREM 11.28. *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of non-negative real numbers which is weakly decreasing and such that $\lim_{n\to\infty} a_n = 0$. Then:*
*a) The associated **alternating series** $\sum_n (-1)^{n+1} a_n$ converges.*
b) For $N \in \mathbb{Z}^+$, put

$$E_N = |(\sum_{n=1}^{\infty}(-1)^{n+1}a_n) - (\sum_{n=1}^{N}(-1)^{n+1}a_n)|. \tag{83}$$

*Then we have the **error estimate***

$$E_N \leq a_{N+1}.$$

Exercise: Let $p \in \mathbb{R}$: Show that the **alternating p-series** $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^p}$ is:
(i) divergent if $p \leq 0$,
(ii) nonabsolutely convergent if $0 < p \leq 1$, and
(iii) absolutely convergent if $p > 1$.

Exercise: Let $\frac{P(x)}{Q(x)}$ be a rational function. Give necessary and sufficient conditions for $\sum_n (-1)^n \frac{P(x)}{Q(x)}$ to be nonabsolutely convergent.

For any convergent series $\sum_{n=1}^{\infty} a_n = S$, we may define E_N as in (83) above:

$$E_N = |S - \sum_{n=1}^{N} a_n|.$$

Then because the series converges to S, $\lim_{N\to\infty} E_N = 0$, and conversely: in other words, to say that the error goes to 0 is a rephrasing of the fact that the partial sums of the series converge to S. Each of these statements is (in the jargon of mathematicians working in this area) **soft**: we assert that a quantity approaches 0 and $N \to \infty$, so that in theory, given any $\epsilon > 0$, we have $E_N < \epsilon$ for all suffuciently large N. But as we have by now seen many times, it is often possible to show that $E_N \to 0$ without coming up with an *explicit* expression for N in terms of ϵ. But this stronger statement is exactly what we have given in Theorem 11.28b): we have

given an *explicit* upper bound on E_N as a function of N. This type of statement is called a **hard** statement or an **explicit error estimate**: such statements tend to be more difficult to come by than soft statements, but also more useful to have. Here, as long as we can similarly make explicit how large N has to be in order for a_N to be less than a given $\epsilon > 0$, we get a completely explicit error estimate and can use this to actually compute the sum S to arbitrary accuracy.

Example: We compute $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$ to three decimal place accuracy. By Theorem 11.28b), it is enough to find $N \in \mathbb{Z}^+$ such that $a_{N+1} = \frac{1}{N+1} < \frac{1}{1000}$. We may take $N = 1000$. Therefore

$$|\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} - \sum_{n=1}^{1000} \frac{(-1)^{n+1}}{n}| \leq \frac{1}{1001}.$$

Using a software package, we find that

$$\sum_{n=1}^{1000} \frac{(-1)^{n+1}}{n} = 0.69264743055982030966723105 89\ldots.$$

Again, later we will show that the exact value of the sum is $\log 2$, which my software package tells me is[14]

$$\log 2 = 0.69314718055994530941723212 14.$$

Thus the *actual error* in cutting off the sum after 1000 terms is

$$E_{1000} = 0.00049975000012499975000106250 33.$$

It is important to remember that this and other error estimates only give upper bounds on the error: the true error could well be much smaller. In this case we were guaranteed to have an error at most $\frac{1}{1001}$ and we see that the true error is about half of that. Thus the estimate for the error is reasonably accurate.

Note well that although the error estimate of Theorem 11.28b) is very easy to apply, if a_n tends to zero rather slowly (as in this example), it is not especially efficient for computations. For instance, in order to compute the true sum of the alternating harmonic series to six decimal place accuracy using this method, we would need to add up the first million terms: that's a lot of calculation. (Thus please be assured that this is *not* the way that a calculator or computer would compute $\log 2$.)

Example: We compute $\sum_{n=0}^{\infty} \frac{(-1)^n}{n!}$ to six decimal place accuracy. Thus we need to choose N such that $a_{N+1} = \frac{1}{(N+1)!} < 10^{-6}$, or equivalently such that $(N+1)! > 10^6$. A little calculation shows $9! = 362,880$ and $10! = 3,628,800$, so that we may take $N = 9$ (but not $N = 8$). Therefore

$$|\sum_{n=0}^{\infty} \frac{(-1)^n}{n!} - \sum_{n=0}^{9} \frac{(-1)^n}{n!}| < \frac{1}{10!} < 10^{-6}.$$

Using a software package, we find

$$\sum_{n=0}^{9} \frac{(-1)^n}{n!} = 0.36787918871252204585537918 87.$$

[14]Yes, you should be wondering how it is computing this! More on this later.

In this case the exact value of the series is

$$\frac{1}{e} == 0.36787944117144232159552377 01$$

so the true error is

$$E_9 = 0.000000252458920275740144581451 6374,$$

which this time is only very slightly less than the guaranteed

$$\frac{1}{10!} = 0.000000275573192239858906525573 1922.$$

7.2. Dirichlet's Test.

What lies beyond the Alternating Series Test? We present one more result, an elegant (and useful) test due originally to Dirichlet.[15]

THEOREM 11.29. *(Dirichlet's Test) Let $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ be two infinite series. Suppose that:*
(i) The partial sums $B_n = b_1 + \ldots + b_n$ are bounded.
(ii) The sequence a_n is decreasing with $\lim_{n\to\infty} a_n = 0$. Then $\sum_{n=1}^{\infty} a_n b_n$ is convergent.

PROOF. Let $M \in \mathbb{R}$ be such that $B_n = b_1 + \ldots + b_n \leq M$ for all $n \in \mathbb{Z}^+$. For $\epsilon > 0$, choose $N > 1$ such that $a_N \frac{\epsilon}{2M}$. Then for $n > m \geq N$,

$$|\sum_{k=m}^{n} a_k b_k| = |\sum_{k=m}^{n} a_k(B_k - B_{k-1})| = |\sum_{k=m}^{n} a_k B_k - \sum_{k=m-1}^{n-1} a_{k+1} B_k|$$

$$= |\sum_{k=m}^{n-1} (a_k - a_{k+1}) B_k + a_n B_n - a_m B_{m-1}|$$

$$\leq \left(\sum_{k=m}^{n-1} |a_k - a_{k+1}||B_k|\right) + |a_n||B_n| + |a_m||B_{m-1}|$$

$$\leq M(\sum_{k=m}^{n-1} |a_k - a_{k+1}|) + |a_n| + |a_m| = M\left(\sum_{k=m}^{n-1} (a_k - a_{k+1}) + a_n + a_m\right)$$

$$= M(a_m - a_n + a_n + a_m) = 2Ma_m \leq 2Ma_N < \epsilon.$$

Therefore $\sum_n a_n b_n$ converges by the Cauchy criterion. □

In the preceding proof, without saying what we were doing, we used the technique of **summation by parts**.

If we take $b_n = (-1)^{n+1}$, then $B_{2n+1} = 1$ for all n and $B_{2n} = 0$ for all n, so $\{b_n\}$ has bounded partial sums. Applying Dirichlet's Test with a sequence a_n which decreases to 0 and with this sequence $\{b_n\}$, we find that the series $\sum_n a_n b_n = \sum_n (-1)^{n+1} a_n$ converges. We have recovered the Alternating Series Test!

In fact Dirichlet's Test yields the following **Almost Alternating Series Test**: let $\{a_n\}$ be a sequence decreasing to 0, and for all n let $b_n \in \{\pm 1\}$ be a "sign sequence" which is **almost alternating** in the sense that the sequence of partial sums $B_n = b_1 + \ldots + b_n$ is bounded. Then the series $\sum_n b_n a_n$ converges.

[15] Johann Peter Gustav Lejeune Dirichlet, 1805-1859

Exercise: Show that Dirichlet's generalization of the Alternating Series Test is "as strong as possible" in the following sense: if $\{b_n\}$ is a sequence of elements, each ± 1, such that the sequence of partial sums $B_n = b_1 + \ldots + b_n$ is *unbounded*, then there is a sequence a_n decreasing to zero such that $\sum_n a_n b_n$ diverges.

Exercise:
a) Use the trigonometric identity

$$\cos n = \frac{\sin(n + \frac{1}{2}) - \sin(n - \frac{1}{2})}{2\sin(\frac{1}{2})}$$

to show that the sequence $B_n = \cos 1 + \ldots + \cos n$ is bounded.
b) Apply Dirichlet's Test to show that the series $\sum_{n=1}^{\infty} \frac{\cos n}{n}$ converges.
c) Show that $\sum_{n=1}^{\infty} \frac{\cos n}{n}$ is not absolutely convergent.

Exercise: Show that $\sum_{n=1}^{\infty} \frac{\sin n}{n}$ is nonabsolutely convergent.

Remark: Once we know about series of complex numbers and Euler's formula $e^{ix} = \cos x + i \sin x$, we will be able to give a "trigonometry-free" proof of the preceding two exercises.

Dirichlet himself applied his test to establish the convergence of a certain class of series of a mixed algebraic and number-theoretic nature. The analytic properties of these series were used to prove his celebrated theorem on prime numbers in arithmetic progressions. To give a sense of how influential this work has become, in modern terminology Dirichlet studied the analytic properties of **Dirichlet series** associated to nontrivial **Dirichlet characters**. For more information on this work, the reader may consult (for instance) [**DS**].

7.3. Cauchy Products III: A Divergent Cauchy Product.

Let us give an example – due to Cauchy! – of a Cauchy product of two nonabsolutely convergent series which fails to converge. Take $\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} \frac{(-1)^n}{\sqrt{n+1}}$. The nth term in the Cauchy product is

$$c_n = \sum_{i+j=n} (-1)^i (-1)^j \frac{1}{\sqrt{i+1}} \frac{1}{\sqrt{j+1}}.$$

Now $(-1)^i(-1)^j = (-1)^{i+j} = (-1)^n$, so c_n is equal to $(-1)^n$ times a sum of positive terms. Since $i, j \leq n$, $\frac{1}{\sqrt{i+1}}, \frac{1}{\sqrt{j+1}} \geq \frac{1}{\sqrt{n+1}}$, and thus each term in c_n has absolute value at least $(\frac{1}{\sqrt{n+1}})^2 = \frac{1}{n+1}$. Since we are summing from $i = 0$ to n there are $n+1$ terms, all of the same size, we find $|c_n| \geq 1$ for all n. Thus the general term of $\sum_n c_n$ does not converge to 0, so the series diverges.

7.4. Decomposition into positive and negative parts.

For a real number r, we define its **positive part**

$$r^+ = \max(r, 0)$$

and its **negative part**

$$r^- = -\min(r, 0).$$

Exercise: Let r be a real number. Show:
a) $r = r^+ - r^-$.
b) $|r| = r^+ + r^-$.

For any real series $\sum_n a_n$ we have a decomposition

$$\sum_n a_n = \sum_n a_n^+ - \sum_n a_n^-,$$

at least if all three series converge. Let us call $\sum_n a_n^+$ and $\sum_n a_n^-$ the **positive part** and **negative part** of $\sum_n a_n$. Let us now suppose that $\sum_n a_n$ converges. By the Three Series Principle there are two cases:

Case 1: Both $\sum_n a_n^+$ and $\sum_n a_n^-$ converge. Hence $\sum_n |a_n| = \sum_n (a_n^+ + a_n^-)$ converges: i.e., $\sum_n a_n$ is absolutely convergent.

Case 2: Both $\sum_n a_n^+$ and $\sum_n a_n^-$ diverge. Hence $\sum_n |a_n| = \sum_n a_n^+ + a_n^-$ diverges: indeed, if it converged, then adding and subtracting $\sum_n a_n$ we would get that $2\sum_n a_n^+$ and $2\sum_n a_n^-$ converge, contradiction. Thus:

PROPOSITION 11.30. *If a series $\sum_n$ is absolutely convergent, both its positive and negative parts converge. If a series $\sum_n$ is nonabsolutely convergent, then both its positive and negative parts diverge.*

Exercise: Let $\sum_n a_n$ be a real series.
a) Show that if $\sum_n a_n^+$ converges and $\sum_n a_n^-$ diverges then $\sum_n a_n = -\infty$.
b) Show that if $\sum_n a_n^+$ diverges and $\sum_n a_n^-$ converges then $\sum_n a_n = \infty$.

8. Power Series I: Power Series as Series

8.1. Convergence of Power Series.

Let $\{a_n\}_{n=0}^\infty$ be a sequence of real numbers. Then a series of the form $\sum_{n=0}^\infty a_n x^n$ is called a **power series**. Thus, for instance, if we had $a_n = 1$ for all n we would get the *geometric series* $\sum_{n=0}^\infty x^n$ which converges iff $x \in (-1, 1)$ and has sum $\frac{1}{1-x}$.

The nth partial sum of a power series is $\sum_{k=0}^n a_k x^k$, a **polynomial** in x. One of the major themes of Chapter three will be to try to view power series as "infinite polynomials": in particular, we will regard x as a variable and be interested in the propeties – continuity, differentiability, integrability, and so on – of the function $f(x) = \sum_{n=0}^\infty a_n x^n$ defined by a power series.

However, if we want to regard the series $\sum_{n=0}^\infty a_n x^n$ as a function of x, what is its domain? The natural domain of a power series is the set of all values of x for which the series converges. Thus the basic question about power series that we will answer in this section is the following.

QUESTION 11.31. *For a sequence $\{a_n\}_{n=0}^\infty$ of real numbers, for which values of $x \in \mathbb{R}$ does the power series $\sum_{n=0}^\infty a_n x^n$ converge?*

There is one value of x for which the answer is trivial. Namely, if we plug in $x = 0$ to our general power series, we get

$$\sum_{n=0}^{\infty} a_n 0^n = a_0 + a_1 \cdot 0 + a_2 \cdot 0^2 = a_0.$$

So every power series converges at least at $x = 0$.

Example 1: Consider the power series $\sum_{n=0}^{\infty} n! x^n$. We apply the Ratio Test:

$$\lim_{n\to\infty} \frac{(n+1)!x^{n+1}}{n!x^n} = \lim_{n\to\infty} (n+1)|x|.$$

The last limit is 0 if $x = 0$ and otherwise is $+\infty$. Therefore the Ratio Test shows that (as we already knew!) the series converges absolutely at $x = 0$ and diverges at every nonzero x. So it is indeed possible for a power series to converge *only* at $x = 0$. This is disappointing if we are interested in $f(x) = \sum_n a_n x^n$ as a function of x, since in this case it is just the function from $\{0\}$ to $\mathbb{R}$ which sends 0 to a_0. There is nothing interesting going on here.

Example 2: Consider $\sum_{n=0}^{\infty} \frac{x^n}{n!}$. We apply the Ratio Test:

$$\lim_{n\to\infty} |\frac{x^{n+1}}{(n+1)!}||\frac{n!}{x^n}| = \lim_{n\to\infty} \frac{|x|}{n+1} = 0.$$

So the power series converges for all $x \in \mathbb{R}$ and defines a function $f : \mathbb{R} \to \mathbb{R}$.

Example 3: Fix $R \in (0, \infty)$; consider $\sum_{n=0}^{\infty} \frac{1}{R^n} x^n$. This is a geometric series with geometric ratio $\rho = \frac{x}{R}$, so it converges iff $|\rho| = |\frac{x}{R}| < 1$, i.e., iff $x \in (-R, R)$.

Example 4: Fix $R \in (0, \infty)$; consider $\sum_{n=1}^{\infty} \frac{1}{nR^n} x^n$. We apply the Ratio Test:

$$\lim_{n\to\infty} \frac{nR^n}{(n+1)R^{n+1}} \frac{|x|^{n+1}}{|x|^n} = |x| \lim_{n\to\infty} \frac{n+1}{n} \cdot \frac{1}{R} = \frac{|x|}{R}.$$

Therefore the series converges absolutely when $|x| < R$ and diverges when $|x| > R$. We must look separately at the case $|x| = R$ – i.e., when $x = \pm R$. When $x = R$, the series is the harmonic series $\sum_n \frac{1}{n}$, hence divergent. But when $x = -R$, the series is the alternating harmonic series $\sum_n \frac{(-1)^n}{n}$, hence (nonabsolutely) convergent. So the power series converges for $x \in [-R, R)$.

Example 5: Fix $R \in (0, \infty)$; consider $\sum_{n=1}^{\infty} \frac{(-1)^n}{nR^n} x^n$. We may rewrite this series as $\sum_{n=1}^{\infty} \frac{1}{nR^n}(-x)^n$, i.e., the same as in Example 4 but with x replaced by $-x$ throughout. Thus the series converges iff $-x \in [-R, R)$, i.e., iff $x \in (-R, R]$.

Example 6: Fix $R \in (0, \infty)$; consider $\sum_{n=1}^{\infty} \frac{1}{n^2 R^n} x^n$. We apply the Ratio Test:

$$\lim_{n\to\infty} \frac{n^2 R^n}{(n+1)^2 R^{n+1}} \frac{|x|^{n+1}}{|x|^n} = |x| \lim_{n\to\infty} \left(\frac{n+1}{n}\right)^2 \cdot \frac{1}{R} = \frac{|x|}{R}.$$

So once again the series converges absolutely when $|x| < R$, diverges when $|x| > R$, and we must look separately at $x = \pm R$. This time plugging in $x = R$ gives $\sum_n \frac{1}{n^2}$ which is a convergent p-series, whereas plugging in $x = -R$ gives $\sum_n \frac{(-1)^n}{n^2}$: since

the p-series with $p = 2$ is convergent, the alternating p-series with $p = 2$ is absolutely convergent. Therefore the series converges (absolutely, in fact) on $[-R, R]$.

Thus the convergence set of a power series can take any of the following forms:

- the single point $\{0\} = [0, 0]$.
- the entire real line $\mathbb{R} = (-\infty, \infty)$.
- for any $R \in (0, \infty)$, an open interval $(-R, R)$.
- for any $R \in (0, \infty)$, a half-open interval $[-R, R)$ or $(-R, R]$
- for any $R \in (0, \infty)$, a closed interval $[-R, R]$.

In each case the set of values is an interval containing 0 and with a certain **radius**, i.e., an extended real number $R \in [0, \infty)$ such that the series definitely converges for all $x \in (-R, R)$ and definitely diverges for all x outside of $[-R, R]$. Our goal is to show that this is the case for *any* power series.

This goal can be approached at various degrees of sophistication. At the calculus level, we have already said what is needed: we use the Ratio Test to see that the convergence set is an interval around 0 of a certain radius R. Namely, taking a general power series $\sum_n a_n x^n$ and applying the Ratio Test, we find

$$\lim_{n\to\infty} \frac{|a_{n+1}x^{n+1}|}{|a_n x^n|} = |x| \lim_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

So if $\rho = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}$, the Ratio Test tells us that the series converges when $|x|\rho < 1$ – i.e., iff $|x| < \frac{1}{\rho}$ – and diverges when $|x|\rho > 1$ – i.e., iff $|x| > \frac{1}{\rho}$. That is, the radius of convergence R is precisely the reciprocal of the Ratio Test limit ρ, with suitable conventions in the extreme cases, i.e., $\frac{1}{0} = \infty$, $\frac{1}{\infty} = 0$.

So what more is there to say or do? The issue here is that we have *assumed* that $\lim_{n\to\infty} \frac{a_{n+1}}{a_n}$ exists. Although this is usually the case in simple examples of interest, it is certainly does not happen in general (we ask the reader to revisit §X.X for examples of this). This we need to take a different approach in the general case.

LEMMA 11.32. *Let $A > 0$ and let $\sum_n a_n x^n$ be a power series. If $\sum_n a_n A^n$ converges, then $\sum_n x^n$ converges absolutely for all $x \in (-A, A)$.*

PROOF. Let $0 < B < A$. It is enough to show $\sum_n a_n B^n$ is absolutely convergent, for then so is $\sum_n a_n (-B)^n$. Since $\sum_n a_n A^n$ converges, $a_n A^n \to 0$: by omitting finitely many terms, we may assume $|a_n A^n| \leq 1$ for all n. Since $0 < \frac{B}{A} < 1$,

$$\sum_n |a_n B^n| = \sum_n |a_n A^n| \left(\frac{B}{A}\right)^n \leq \sum_n \left(\frac{B}{A}\right)^n < \infty.$$

□

THEOREM 11.33. *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series.*
a) There exists $R \in [0, \infty]$ such that:
(i) For all x with $|x| < R$, $\sum_n a_n x^n$ converges absolutely and
(ii) For all x with $|x| > R$, $\sum_n a_n x^n$ diverges.
b) If $R = 0$, then the power series converges only at $x = 0$.
c) If $R = \infty$, the power series converges for all $x \in \mathbb{R}$.

d) If $0 < R < \infty$, the convergence set of the power series is either $(-R, R)$, $[-R, R)$, $(-R, R]$ or $[-R, R]$.

PROOF. a) Let R be the least upper bound of the set of $x \geq 0$ such that $\sum_n a_n x^n$ converges. If y is such that $|y| < R$, then there exists A with $|y| < A < R$ such that $\sum_n a_n A^n$ converges, so by Lemma 11.32 the power series converges absolutely on $(-A, A)$, so in particular it converges absolutely at y. Thus R satisfies property (i). Similarly, suppose there exists y with $|y| > R$ such that $\sum_n a_n y^n$ converges. Then there exists A with $R < A < |y|$ such that the power series converges on $(-A, A)$, contradicting the definition of R.
We leave the proof of parts b) through d) to the reader as a straightforward exercise. □

Exercise: Prove parts) b), c) and d) of Theorem 11.33.

Exercise: Let $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$ be two power series with positive radii of convergence R_a and R_b. Let $R = \min(R_a, R_b)$. Put $c_n = \sum_{k=0}^{n} a_k b_{n-k}$. Show that the "formal identity"

$$\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} b_n x^n\right) = \sum_{n=0}^{\infty} c_n x^n$$

is valid for all $x \in (-R, R)$. (Suggestion: use past results on Cauchy products.)

The drawback of Theorem 11.33 is that it does not give an explicit description of the radius of convergence R in terms of the coefficients of the power series, as is the case when the ratio test limit $\rho = \lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$ exists. In order to achieve this in general, we need to appeal instead to the Root Test and make use of the limit supremum. The following elegant result is generally attributed to J.S. Hadamard,[16] who published it in 1888 [**Ha88**] and included it in his 1892 PhD thesis. This seems remarkably late in the day for a result which is so closely linked to (Cauchy's) Root Test. It turns out that the result was established by our most usual suspect: it was first proven by Cauchy in 1821 [**Ca21**] but apparently had been nearly forgotten.

THEOREM 11.34. *(Cauchy-Hadamard) Let $\sum_n a_n x^n$ be a power series, and put*

$$\overline{\theta} = \limsup_{n\to\infty} |a_n|^{\frac{1}{n}}.$$

Then the radius of convergence of the power series is $R = \frac{1}{\overline{\theta}}$: that is, the series converges absolutely for $|x| < R$ and diverges for $|x| > R$.

PROOF. We have $\limsup_{n\to\infty} |a_n x^n|^{\frac{1}{n}} = |x| \limsup_{n\to\infty} |a_n|^{\frac{1}{n}} = |x|\overline{\theta}$. Put $R = \frac{1}{\overline{\theta}}$. If $|x| < R$, choose A such that $|x| < A < R$ and then A' such that

$$\overline{\theta} = \frac{1}{R} < A' < \frac{1}{A}.$$

Then for all sufficiently large n, $|a_n x^n|^{\frac{1}{n}} \leq A'A < 1$, so the series converges absolutely by the Root Test. Similarly, if $|x| > R$, choose A such that $R < |x| < A$ and then A' such that

$$\frac{1}{A} < A' < \frac{1}{R} = \overline{\theta}.$$

[16] Jacques Salomon Hadamard (1865-1963)

Then there are infinitely many non-negative integers n such that $|a_n x^n|^{\frac{1}{n}} \geq A' A > 1$, so the series $\sum_n a_n x^n$ diverges: indeed $a_n x^n \nrightarrow 0$. □

Here is a useful criterion for the radius of convergence of a power series to be 1.

COROLLARY 11.35. *Let $\{a_n\}_{n=0}^{\infty}$ be a sequence of real numbers, and let R be the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n x^n$.*
a) If $\{a_n\}$ is bounded, then $R \geq 1$.
b) If $a_n \nrightarrow 0$, then $R \leq 1$.
c) Thus if $\{a_n\}$ is bounded but not convergent to zero, $R = 1$.

Exercise: Prove Corollary 11.35.

THEOREM 11.36. *Let $\sum_n a_n x^n$ be a power series with radius of convergence R. Then, for any $k \in \mathbb{Z}$, the radius of convergence of $\sum_n n^k a_n x^n$ is also R.*

PROOF. Since $\lim_{n\to\infty} \frac{(n+1)^k}{n^k} = \lim_{n\to\infty} \left(\frac{n+1}{n}\right)^k = 1$, by Corollary 11.22,

$$\lim_{n\to\infty} (n^k)^{1/n} = \lim_{n\to\infty} n^{k/n} = 1.$$

(Alternately, take logarithms and apply L'Hôpital's Rule.) Therefore

$$\limsup_{n\to\infty} (n^k |a_n|)^{\frac{1}{n}} = \left(\lim_{n\to\infty} (n^k)^{\frac{1}{n}}\right) \left(\limsup_{n\to\infty} |a_n|^{\frac{1}{n}}\right) = \limsup_{n\to\infty} |a_n|^{\frac{1}{n}}.$$

The result now follows from the Cauchy-Hadamard Formula. □

Remark: For the reader who is less than comfortable with limits infimum and supremum, we recommend simply assuming that the Ratio Test limit $\rho = \lim_{n\to\infty} |\frac{a_{n+1}}{a_n}|$ exists and proving Theorem 11.36 under that additional assumption using the Ratio Test. This will be good enough for most of the power series encountered in practice.

Exercise: Let $\sum_n a_n x^n$ be a power series with interval of convergence I. Let J be the interval of convergence of $\sum_n n a_n x^{n-1}$.
a) Show that $J \subset I$.
b) Give an example in which $J \neq I$.

CHAPTER 12

Taylor Taylor Taylor Taylor

1. Taylor Polynomials

Recall that early on we made a distinction between *polynomial expressions* and *polynomial functions*. A polynomial expression is something of the form $P(t) = \sum_{n=0}^N a_n t^n$ with $N \in \mathbb{N}$ and $a_0, \ldots, a_N \in \mathbb{R}$. To every polynomial expression we associated a polynomial function $P(x)$ from $\mathbb{R}$ to $\mathbb{R}$, given by $x \mapsto \sum_{n=0}^N a_n x^n$. Then we showed that the process is reversible in the following sense: if two expressions $\sum_{n=0}^N a_n t^n$ and $\sum_{n=0}^N b_n t^n$ define the same polynomial function, then $a_n = b_n$ for all $0 \leq n \leq N$. (This came down to showing that that if a polynomial expression $\sum_{n=0}^\infty a_n t^n$ defines the zero function, then $a_n = 0$ for all n.) And after that – thankfully! – we have spoken only of polynomials.

We wish – briefly! – to revisit this formalism in a slightly generalized context. Namely we want to consider polynomial c-expansions and their associated functions. A **polynomial c-expansion** is a formal expression $\sum_{n=0}^N a_n (t-c)^n$ with $N \in \mathbb{N}$ and $c, a_n \in \mathbb{R}$. A polynomial c-expansion induces a function in an evident manner: we map x to $\sum_{n=0}^N a_n (x-c)^n$. Now we wish to establish the following simple – indeed, almost trivial – but nevertheless important result.

THEOREM 12.1. *Let P be a polynomial of degree at most N, and let $c \in \mathbb{R}$.*
a) There is a unique polynomial c-expansion $\sum_{n=0}^N b_n (t-c)^n$ such that for all $x \in \mathbb{R}$, $P(x) = \sum_{n=0}^N b_n (x-c)^n$.
b) Explicitly, we have $b_n = \frac{P^{(n)}(c)}{n!}$ for all $0 \leq n \leq N$.

PROOF. a) We first discuss the *existence* of c-expansions. Perhaps this is most cleanly handled by induction on the degree of P. If P is the zero polynomial or is a constant polynomial, then there is nothing to show: a constant a is certainly of the form $\sum_n b_n (x-c)^n$: take $b_0 = a$ and $b_n = 0$ for all $n > 0$. Suppose now that $N > 0$ and we have established that every polynomial function of degree less than N has a polynomial c-expansion, and consider a degree d polynomial expression $P(t) = a_0 + a_1 t + \ldots + a_{N-1} t^{N-1} + a_N t^N$. Then we may write

$$P(t) = a_N (t-c)^N + Q(t),$$

where, since the leading term of $a_N (t-c)^N$ is $a_N t^n$, $Q(t)$ is a polynomial of degree less than N and thus by induction can be written as

$$Q(t) = \sum_{n=0}^{N-1} b_n (t-c)^n.$$

Then

$$P(t) = b_0 + b_1 (t-c) + \ldots + b_{N-1} (t-c)^{N-1} + a_N (t-c)^N.$$

Alternately and perhaps more directly, simply write

$$a_0+a_1t+a_2t^2+\ldots+a_Nt^N = a_0+a_1((t-c)+c)+a_2((t-c)+c)^2+\ldots+a_N((t-c)+c)^N$$

and use the binomial theorem to expand each $((t-c)+c)^n$ in powers of $t-c$.

As for the uniqueness, suppose that we have two c-expansions $\sum_{n=0}^N a_n(t-c)^n$, $\sum_{n=0}^N b_n(t-c)^n$ which induce the same function. By an argument almost identical to that just given, we may write $\sum_{n=0}^N (a_n-b_n)(t-c)^n$ as $\sum_{n=0}^N d_nt^n$, with highest degree term $d_N = a_N - b_N$. Since the corresponding function $x \mapsto \sum_{n=0}^N d_nx^n$ is identically zero, by what we showed about polynomial functions we have $d_n = 0$ for all n. In particular $0 = d_N = a_N - b_N$, so that $\sum_{n=0}^N (a_n - b_n)(t-c)^n = \sum_{n=0}^{N-1}(a_n-b_n)(t-c)^n = \sum_{n=0}^{N-1} d_nt^n$. Reasoning as above we find that $d_{N-1} = a_{N-1} - b_{N-1} = 0$, and so forth: continuing in this way we find that $a_n = b_n$ for all $0 \neq n \leq N$, and thus the c-expansion is unique.

b) Consider the identity

$$P(x) = \sum_{n=0}^N b_n(x-c)^n = b_0 + b_1(x-c) + b_2(x-c)^2 + \ldots + b_N(x-c)^N. \tag{84}$$

Evaluating (84) at $x = c$ gives $P(c) = b_0$. Differentiating (84) gives

$$P'(x) = b_1 + 2b_2(x-c) + 3b_3(x-c)^2 + \ldots + Nb_N(x-c)^{N-1},$$

and evaluating at $x = c$ gives $P'(c) = b_1$. Differentiating (84) again and evaluating at $x = c$ gives $P''(c) = 2b_2$, so $b_2 = \frac{P''(c)}{2}$. Continuing in this manner one finds that for $0 \leq k \leq N$, $P^{(k)}(c) = k!b_k$, so $b_k = \frac{P^{(k)}(c)}{k!}$. □

Corollary 12.2. *Let I be an interval and c an interior point of I, and let $f : I \to \mathbb{R}$ be a function which is N times differentiable at c. Let*

$$T_N(x) = \sum_{k=0}^N \frac{f^{(k)}(c)}{k!}(x-c)^k.$$

Then $T_N(x)$ is the unique polynomial function of degree at most N such that $T_N^{(k)}(c) = f^{(k)}(c)$ for all $0 \leq k \leq N$.

Proof. Applying Theorem 0 to $T_N(x) = \sum_{k=0}^N \frac{f^{(k)}(c)}{k!}(x-c)^k$ gives $\frac{f^{(k)}(c)}{k!} = \frac{T_N^{(k)}(c)}{k!}$, hence $T_N^{(k)}(c) = f^{(k)}(c)$ for all $0 \leq k \leq N$. As for the uniqueness: let $P(x)$ be any polynomial of degree at most N such that $P^{(k)}(c) = f^{(k)}(c)$ for $0 \leq k \leq N$, and let $Q = T_N - P$. Then Q is a polynomial of degree at most N such that $Q^{(k)}(c) = 0$ for $0 \leq k \leq N$; applying Theorem 12.1 we get $Q(x) = \sum_{k=0}^N \frac{Q^{(k)}(c)}{k!}(x-c)^k = \sum_{k=0}^N 0 \cdot (x-c)^k = 0$, i.e., Q is the zero polynomial and thus $P = T_N$. □

By definition, $T_N(x)$ is the **degree N Taylor polynomial of f at c**.

2. Taylor's Theorem Without Remainder

For $n \in \mathbb{N}$ and $c \in I^\circ$, we say two functions $f, g : I \to \mathbb{R}$ **agree to order n at c** if

$$\lim_{x \to c} \frac{f(x) - g(x)}{(x-c)^n} = 0.$$

Exercise: Show: if $m \leq n$ and f and g agree to order n at c, then f and g agree to order m at c.

Example 0: We claim that two continuous functions f and g agree to order 0 at c if and only if $f(c) = g(c)$. Indeed, suppose that f and g agree to order 0 at c. Since f and g are continuous, we have

$$0 = \lim_{x \to c} \frac{f(x) - g(x)}{(x-c)^0} = \lim_{x \to c} f(x) - g(x) = f(c) - g(c).$$

The converse, that if $f(c) = g(c)$ then $\lim_{x \to c} f(x) - g(x) = 0$, is equally clear.

Example 1: We claim that two differentiable functions f and g agree to order 1 at c if and only if $f(c) = g(c)$ and $f'(c) = g'(c)$. Indeed, by the exercise above both hypotheses imply $f(c) = g(c)$, so we may assume that, and then we find

$$\lim_{x \to c} \frac{f(x) - g(x)}{x - c} = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} - \frac{g(x) - g(c)}{x - c} = f'(c) - g'(c).$$

Thus assuming $f(c) = g(c)$, f and g agree to order 1 at c if and only $f'(c) = g'(c)$.

The following result gives the expected generalization of these two examples. It is generally attributed to Taylor,[1] probably correctly, although special cases were known to earlier mathematicians. Note that **Taylor's Theorem** often refers to a later result (Theorem 12.4) that we call "Taylor's Theorem With Remainder," even though it is Theorem 12.3 (only) that was proved by Brook Taylor.

THEOREM 12.3. *(Taylor) Let $n \in \mathbb{N}$ and $f, g : I \to \mathbb{R}$ be two n times differentiable functions. Let c be an interior point of I. The following are equivalent:*
(i) We have $f(c) = g(c), f'(c) = g'(c), \ldots, f^{(n)}(c) = g^{(n)}(c)$.
(ii) f and g agree to order n at c.

PROOF. Set $h(x) = f(x) - g(x)$. Then (i) holds iff $h(c) = h'(c) = \ldots = h^{(n)}(c) = 0$ and (ii) holds iff $\lim_{x \to c} \frac{h(x)}{(x-c)^n} = 0$. So we may work with h instead of f and g. Since we dealt with $n = 0$ and $n = 1$ above, we may assume $n \geq 2$.
(i) $\implies$ (ii): $L = \lim_{x \to c} \frac{h(x)}{(x-c)^n}$ is of the form $\frac{0}{0}$, so L'Hôpital's Rule gives

$$L = \lim_{x \to c} \frac{h'(x)}{n(x-c)^{n-1}},$$

provided the latter limit exists. By our assumptions, this latter limit is still of the form $\frac{0}{0}$, so we may apply L'Hôpital's Rule again. We do so iff $n > 2$. In general, we apply L'Hôpital's Rule $n - 1$ times, getting

$$L = \lim_{x \to c} \frac{h^{(n-1)}(x)}{n!(x-c)} = \frac{1}{n!} \left(\lim_{x \to c} \frac{h^{(n-1)}(x) - h^{(n-1)}(c)}{x - c} \right),$$

provided the latter limit exists. But the expression in parentheses is nothing else than the derivative of the function $h^{(n-1)}(x)$ at $x = c$ – i.e., it is $h^{(n)}(c) = 0$ (and, in particular the limit exists; only now have the $n - 1$ applications of L'Hôpital's Rule been unconditionally justified), so $L = 0$. Thus (ii) holds.
(ii) $\implies$ (i): Let $T_n(x)$ be the degree N Taylor polynomial to h at c. By Corollary

[1] Brook Taylor, 1685 - 1731

12.2, f and T_n agree to order n at c, so by the just proved implication (i) $\implies$ (ii), $h(x)$ and $T_n(x)$ agree to order n at $x = c$:

$$\lim_{x \to c} \frac{h(x) - T_n(x)}{(x-c)^n} = 0.$$

Moreover, by assumption $h(x)$ agrees to order n with the zero function:

$$\lim_{x \to c} \frac{h(x)}{(x-c)^n} = 0.$$

Subtracting these limits gives

(85)
$$\lim_{x \to c} \frac{T_n(x)}{(x-c)^n} = \lim_{x \to c} \frac{h(c) + h'(c)(x-c) + \frac{h''(c)}{2}(x-c)^2 + \ldots + \frac{h^{(n)}(c)}{n!}(x-c)^n}{(x-c)^n} = 0.$$

Clearly $\lim_{x \to c} T_n(x) = T_n(c)$, so if $T_n(c) \neq 0$, then $\lim_{x \to c} \frac{T_n(x)}{(x-c)^n}$ would not exist, so we must have $T_n(c) = h(c) = 0$. Therefore

$$\lim_{x \to c} \frac{T_n(x)}{(x-c)^n} = \lim_{x \to c} \frac{h'(c) + \frac{h''(c)}{2}(x-c) + \ldots + \frac{h^{(n)}(c)}{n!}(x-c)^{n-1}}{(x-c)^{n-1}} = 0.$$

As above, we have the limit of a quotient of continuous functions which we know exists such that the denominator approaches 0, so the numerator must also approach zero (otherwise the limit would be infinite): evaluating the numerator at c gives $h'(c) = 0$. And so forth: continuing in this way we find that the existence of the limit in (85) implies that $h(c) = h'(c) = \ldots = h^{(n-1)}(c) = 0$, so (85) simplifies to

$$0 = \lim_{x \to c} \frac{\frac{h^{(n)}(c)}{n!}(x-c)^n}{(x-c)^n} = \frac{h^{(n)}(c)}{n!},$$

so $h^{(n)}(c) = 0$. □

Remark: Above we avoided a subtle pitfall: we applied L'Hôpital's Rule $n-1$ times to $\lim_{x \to c} \frac{h(x)}{(x-c)^n}$, but the final limit we got was still of the form $\frac{0}{0}$ – so why not apply L'Hôpital one more time? The answer is if we do we get that

$$L = \lim_{x \to c} \frac{h^{(n)}(x)}{n!},$$

assuming this limit exists. But to assume this last limit exists and is equal to $h^{(n)}(0)$ is to assume that nth derivative of h is *continuous* at zero, which is slightly more than we want (or need) to assume.

For $n \in \mathbb{N}$, a function $f : I \to \mathbb{R}$ **vanishes to order n** at c if $\lim_{x \to c} \frac{f(x)}{(x-c)^n} = 0$. Note that this concept came up prominently in the proof of Theorem 12.3 in the form: f and g agree to order n at c iff $f - g$ vanishes to order n at c.

Exercise: Let f be a function which is n times differentiable at $x = c$, and let T_n be its degree n Taylor polynomial at $x = c$. Show that $f - T_n$ vanishes to order n at $x = c$. (This is just driving home a key point of the proof of Theorem 12.3 in our new terminology.)

Exercise:

a) Show that for a function $f : I \to \mathbb{R}$, the following are equivalent:
(i) f is differentiable at c.
(ii) We may write $f(x) = a_0 + a_1(x-c) + g(x)$ for a function $g(x)$ vanishing to order 1 at c.
b) Show that if the equivalent conditions of part a) are satisfied, then we must have $a_0 = f(c)$, $a_1 = f'(c)$ and thus the expression of a function differentiable at c as the sum of a linear function and a function vanishing to first order at c is unique.

Exercise:[2] Let $a, b \in \mathbb{Z}^+$, and consider the following function $f_{a,b} : \mathbb{R} \to \mathbb{R}$:

$$f_{a,b}(x) = \begin{cases} x^a \sin\left(\frac{1}{x^b}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

a) Show that $f_{a,b}$ vanishes to order $a-1$ at 0 but does not vanish to order a at 0.
b) Show that $f_{a,b}$ is differentiable at $x = 0$ iff $a \geq 2$, in which case $f'_{a,b}(0) = 0$.
c) Show that $f_{a,b}$ is twice differentiable at $x = 0$ iff $a \geq b+3$.
d) Deduce that for any $n \geq 2$, $f_{n,n}$ vanishes to order n at $x = 0$ but is not twice differentiable at $x = 0$.

3. Taylor's Theorem With Remainder

To state the following theorem, it will be convenient to make a convention: real numbers a, b, by $|[a,b]|$ we will mean the interval $[a,b]$ if $a \leq b$ and the interval $[b,a]$ if $b < a$. So $|[a,b]|$ is the set of real numbers lying between a and b.

For a function $f : |[a,b]| \to \mathbb{R}$, we put $||f|| = \sup_{x \in [a,b]} |f(x)|$.

THEOREM 12.4. *(Taylor's Theorem With Remainder) Let $n \in \mathbb{N}$, let I be an interval, and let $f : I \to \mathbb{R}$ be defined and $n+1$ times differentiable. Let $c \in I^\circ$ and $x \in I$. Moreover, let*

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)(x-c)^k}{k!}$$

be the degree n Taylor polynomial of f at c, and let

$$R_n(x) = f(x) - T_n(x)$$

be the remainder. Then:
a) There is $z \in |[c,x]|$ such that

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-z)^n(x-c). \tag{86}$$

b) There exists $z \in |[c,x]|$ such that

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-c)^{n+1}. \tag{87}$$

c) If $f^{(n+1)}$ is integrable on $|[c,x]|$, then

$$R_n(x) = \int_c^x \frac{f^{n+1}(t)(x-t)^n dt}{n!}.$$

[2]Thanks to Didier Piau for a correction that led to this exercise.

d) We have

$$|R_n(x)| = |f(x) - T_n(x)| \leq \frac{||f^{(n+1)}||}{(n+1)!}|x-c|^{n+1}.$$

PROOF. We closely follow [**S**, Thm. 20.4].

a) First note that, notwithstanding the notation chosen, the expression $R_n(x) = f(x) - T_n(x)$ certainly depends upon the "expansion point" $c \in I^\circ$: thus it would be more accurate to write it as $R_{n,c}(x)$. In fact this opens the door to a key idea: for any $t \in I^\circ$, we may consider the Taylor polynomial $T_{n,t}$ to f centered at t and thus the remainder function

$$R_{n,t}(x) = f(x) - T_{n,t}(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!}(x-t)^k. \tag{88}$$

To emphasize the dependence on t, we define a new function $S : |[c,x]| \to \mathbb{R}$ by $S(t) = R_{n,t}(x)$. At first thought (*my* first thought, at least) consideration of $S(t)$ just seems to make things that much more complicated. But in fact the assumed differentiability properties of f easily translate into differentiability properties of $S(t)$, and these can used to our advantage in a rather beautiful way. Indeed, since f is $n+1$ times differentiable on I, S is differentiable on $|[c,x]|$. Differentiating both sides of (88) *with respect to* t gives

$$S'(t) = -f'(t) - \sum_{k=1}^{n} \left(\frac{f^{(k+1)}(t)}{k!}(x-t)^k + \frac{f^{(k)}(t)}{(k-1)!}(x-t)^{k-1} \right) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n.$$

We apply the Mean Value Theorem to S on $|[c,x]|$: there is $z \in |(c,x)|$ such that

$$\frac{S(x) - S(c)}{x-c} = S'(z) = \frac{-f^{(n+1)}(z)}{n!}(x-z)^n.$$

Noting that $S(x) = R_{n,x}(x) = 0$ and $S(c) = R_{n,c}(x) = R_n(x)$, this gives

$$R_n(x) = S(c) - S(x) = \frac{f^{(n+1)}(z)}{n!}(x-z)^n(x-c).$$

b) Apply the Cauchy Mean Value Theorem to $S(t)$ and $g(t) = (x-t)^{n+1}$ on $|[c,x]$: there is $z \in |(c,x)|$ such that

$$\frac{R_n(x)}{(x-c)^{n+1}} = \frac{S(x) - S(c)}{g(x) - g(c)} = \frac{S'(z)}{g'(z)} = \frac{\frac{-f^{(n+1)}(z)}{n!}(x-z)^n}{-(n+1)(x-z)^n} = \frac{f^{(n+1)}(z)}{(n+1)!},$$

so

$$R_n(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-c)^{n+1}.$$

c) If $f^{(n+1)}$ is integrable on $|[c,x]|$, then

$$R_n(x) = -(S(x) - S(c)) = -\int_c^x S'(t) = \int_c^x \frac{f^{(n+1)}(t)(x-t)^n dt}{n!}.$$

d) This follows almost immediately from part b); the proof is left to the reader. □

Exercise: Show that Theorem 12.4 (Taylor's Theorem With Remainder) implies Theorem 12.3 (Taylor's Theorem) under the additional hypothesis that $f^{(n+1)}$ exists and is continuous[3] on the interval $|[c,x]|$.

[3]Thanks to Nick Fink for pointing out the hypothesis of continuity seems to be needed here.

4. Taylor Series

4.1. The Taylor Series. Let $f : I \to \mathbb{R}$ be an infinitely differentiable function, and let $c \in I$. We define the **Taylor series** of f at c to be

$$T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)(x-c)^n}{n!}.$$

Thus $T(x) = \lim_{n\to\infty} T_n(x)$, where T_n is the degree n Taylor polynomial at c. In particular $T(x)$ is a power series, so all of our prior work on power series applies.

Just as with power series, it is no real loss of generality to assume that $c = 0$, in which case our series takes the simpler form

$$T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)x^n}{n!};$$

indeed, to get from this to the general case one merely has to make the change of variables $x \mapsto x - c$. It is traditional to call Taylor series centered around $c = 0$ **Maclaurin series**. But I know no good reason for this – Taylor series were introduced by Taylor in 1721, whereas Colin Maclaurin's *Theory of fluxions* was not published until 1742 and makes explicit attribution is made to Taylor's work.[4] Using separate names for Taylor series centered at 0 and Taylor series centered at c often suggests – misleadingly! – to students that there is some conceptual difference between the two cases. So we will not use the term "Maclaurin series" here.

Exercise: Define a function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = e^{\frac{-1}{x^2}}$ for $x \neq 0$ and $f(0) = 0$. Show that f is infinitely differentiable and in fact $f^{(n)}(0) = 0$ for all $n \in \mathbb{N}$.

When dealing with Taylor series there are two main issues.

QUESTION 12.5. *Let $f : I \to \mathbb{R}$ be an infinitely differentiable function and $T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)x^n}{n!}$ be its Taylor series.*
a) For which values of x does $T(x)$ converge?
b) If for $x \in I$, $T(x)$ converges, do we have $T(x) = f(x)$?

Notice that Question 12.5a) is simply asking for which values of $x \in \mathbb{R}$ a power series is convergent, a question to which we worked out a very satisfactory answer in §X.X. Namely, the set of values x on which a power series converges is an interval of radius $R \in [0, \infty]$ centered at 0. More precisely, in theory the value of R is given by Hadamard's Formula $\frac{1}{R} = \limsup_{n\to\infty} |a_n|^{\frac{1}{n}}$, and in practice we expect to be able to apply the Ratio Test (or, if necessary, the Root Test) to compute R.

If $R = 0$ then $T(x)$ only converges at $x = 0$ and we have $T(0) = f(0)$: this is a trivial case. Henceforth we assume that $R \in (0, \infty]$ so that f converges (at least) on $(-R, R)$. Fix a number A, $0 < A \leq R$ such that $(-A, A) \subset I$. We may then move on to Question 12.5b): must $f(x) = T(x)$ for all $x \in (-A, A)$?

The answer is *no*: consider the function $f(x)$ of Exercise X.X. $f(x)$ is infinitely differentiable and has $f^{(n)}(0) = 0$ for all $n \in \mathbb{N}$, so its Taylor series is

[4]Special cases of the Taylor series concept were well known to Newton and Gregory in the 17th century and to the Indian mathematician Madhava of Sangamagrama in the 14th century.

$T(x) = \sum_{n=0}^{\infty} \frac{0x^n}{n!} = \sum_{n=0}^{\infty} 0 = 0$, i.e., it converges for all $x \in \mathbb{R}$ to the zero function. Of course $f(0) = 0$, but for $x \neq 0$, $f(x) = e^{\frac{-1}{x^2}} \neq 0$. Therefore $f(x) \neq T(x)$ in any open interval around $x = 0$.

There are plenty of other examples. Indeed, in a sense that we will not try to make precise here, "most" infinitely differentiable functions $f : \mathbb{R} \to \mathbb{R}$ are not equal to their Taylor series expansions in any open interval about any point. That's the bad news. However, one could interpret this to mean that we are not really interested in "most" infinitely differentiable functions: the **special functions** one meets in calculus, advanced calculus, physics, engineering and analytic number theory are almost invariably equal to their Taylor series expansions, at least in some small interval around any given point $x = c$ in the domain.

4.2. Easy Examples.

In any case, if we wish to *try* to show that a $T(x) = f(x)$ on some interval $(-A, A)$, we have a tool for this: Taylor's Theorem With Remainder. Indeed, since $R_n(x) = |f(x) - T_n(x)|$, we have

$$f(x) = T(x) \iff f(x) = \lim_{n\to\infty} T_n(x)$$

$$\iff \lim_{n\to\infty} |f(x) - T_n(x)| = 0 \iff \lim_{n\to\infty} R_n(x) = 0.$$

So it comes down to being able to give upper bounds on $R_n(x)$ which tend to zero as $n \to \infty$. According to Taylor's Theorem with Remainder, this will hold whenever we can show that the norm of the nth derivative $||f^{(n)}||$ does not grow too rapidly.

Example: We claim that for all $x \in \mathbb{R}$, the function $f(x) = e^x$ is equal to its Taylor series expansion at $x = 0$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

First we compute the Taylor series expansion: $f^{(0)}(0) = f(0) = e^0 = 1$, and $f'(x) = e^x$, hence every derivative of e^x is just e^x again. We conclude that $f^{(n)}(0) = 1$ for all n and thus the Taylor series is $\sum_{n=0}^{\infty} \frac{x^n}{n!}$, as claimed. Next note that this power series converges for all real x, as we have already seen: just apply the Ratio Test. Finally, we use Taylor's Theorem with Remainder to show that $R_n(x) \to 0$ for each fixed $x \in \mathbb{R}$. Indeed, Theorem 12.4 gives us

$$R_n(x) \leq \frac{||f^{(n+1)}||}{(n+1)!}|x - c|^{n+1},$$

where $||f^{(n+1)}||$ is the supremum of the the absolute value of the $(n+1)$st derivative on the interval $|[0, x]|$. But – lucky us – in this case $f^{(n+1)}(x) = e^x$ for all n and the maximum value of e^x on this interval is e^x if $x \geq 0$ and 1 otherwise, so in either way $||f^{(n+1)}|| \leq e^{|x|}$. So

$$R_n(x) \leq e^{|x|} \left(\frac{x^{n+1}}{(n+1)!} \right).$$

And now we win: the factor inside the parentheses approaches zero with n and is being multiplied by a quantity which is *independent of* n, so $R_n(x) \to 0$. In fact a

moment's thought shows that $R_n(x) \to 0$ *uniformly* on any bounded interval, say on $[-A, A]$, and thus our work on the general properties of uniform convergence of power series (in particular the M-test) is not needed here: everything comes from Taylor's Theorem With Remainder.

Example continued: we use Taylor's Theorem With Remainder to compute $e = e^1$ accurate to 10 decimal places.

A little thought shows that the work we did for $f(x) = e^x$ carries over verbatim under somewhat more general hypotheses.

THEOREM 12.6. *Let $f(x) : \mathbb{R} \to \mathbb{R}$ be a smooth function. Suppose that for all $A \in [0, \infty)$ there exists a number M_A such that for all $x \in [-A, A]$ and all $n \in \mathbb{N}$,*

$$|f^{(n)}(x)| \leq M_A.$$

a) The Taylor series $T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)x^n}{n!}$ converges absolutely for all $x \in \mathbb{R}$.
b) For all $x \in \mathbb{R}$ we have $f(x) = T(x)$: that is, f is equal to its Taylor series expansion at 0.

Exercise: Prove Theorem 12.6.

Exercise: Suppose $f : \mathbb{R} \to \mathbb{R}$ is a smooth function with **periodic derivatives**: there exists some $k \in \mathbb{Z}^+$ such that $f = f^{(k)}$. Show that f satisfies the hypothesis of Theorem 12.6 and therefore is equal to its Taylor series expansion at $x = c$.

4.3. The Binomial Series.

Even for familiar, elementary functions, using Theorem 12.4 to show $R_n(x) \to 0$ may require nonroutine work. We give a case study: the **binomial series**.

Let $\alpha \in \mathbb{R}$. For $x \in (-1, 1)$, we define

$$f(x) = (1 + x)^{\alpha}.$$

Case 1: Suppose $\alpha \in \mathbb{N}$. Then f is just a polynomial; in particular f is defined and infinitely differentiable for all real numbers.
Case 2: Suppose α is positive but not an integer. Depending on the value of α, f may or may not be defined for $x < -1$ (e.g. it is for $\alpha = \frac{2}{3}$ and it is not for $\alpha = \frac{3}{2}$), but in any case f is only $\langle \alpha \rangle$ times differentiable at $x = -1$.
Case 3: Suppose $\alpha < 0$. Then $\lim_{x \to -1^+} f(x) = \infty$.

The upshot of this discussion is that if α is not a positive integer, then f is defined and infinitely differentiable on $(-1, \infty)$ and on no larger interval than this.

For $n \in \mathbb{Z}^+$, $f^{(n)}(x) = (\alpha)(\alpha - 1) \cdots (\alpha - (n - 1))(1 + x)^{\alpha - n}$, so $f^{(n)}(0) = (\alpha)(\alpha - 1) \cdots (\alpha - (n - 1))$. Of course we have $f^{(0)}(0) = f(0) = 1$, so the Taylor series to f at $c = 0$ is

$$T(x) = 1 + \sum_{n=1}^{\infty} \frac{(\alpha)(\alpha - 1) \cdots (\alpha - (n - 1))}{n!} x^n.$$

If $\alpha \in \mathbb{N}$, we recognize the nth Taylor series coefficient as the *binomial coefficient* $\binom{\alpha}{n}$, and this ought not to be surprising because for $\alpha \in \mathbb{N}$, expanding out $T(x)$ simply gives the binomial theorem:

$$\forall \alpha \in \mathbb{N},\ (1+x)^\alpha = \sum_{n=0}^{\alpha} \binom{\alpha}{n} x^n.$$

So let's extend our definition of binomial coefficients: for *any* $\alpha \in \mathbb{R}$, put

$$\binom{\alpha}{0} = 1,$$

$$\forall n \in \mathbb{Z}^+, \binom{\alpha}{n} = \frac{(\alpha)(\alpha-1)\cdots(\alpha-(n-1))}{n!}.$$

Exercise: For any $\alpha \in \mathbb{R}, n \in \mathbb{Z}^+$, show

$$\binom{\alpha}{n} = \binom{\alpha-1}{n-1} + \binom{\alpha-1}{n}. \tag{89}$$

Finally, we rename the Taylor series to $f(x)$ as the **binomial series**

$$B(\alpha, x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n.$$

The binomial series is as old as calculus itself, having been studied by Newton in the 17th century.[5] It remains one of the most important and useful of all power series. For us, our order of business is the usual one when given a Taylor series: first, for each fixed α we wish to find the interval I on which the series $B(\alpha, x)$ converges. Second, we would like to show – if possible! – that for all $x \in I$, $B(\alpha, x) = (1+x)^\alpha$.

THEOREM 12.7. *Let $\alpha \in \mathbb{R} \setminus \mathbb{N}$, and consider the* ***binomial series***

$$B(\alpha, x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n = 1 + \sum_{n=1}^{\infty} \binom{\alpha}{n} x^n.$$

a) For all such α, the radius of convergence of $B(\alpha, x) = 1$.
b) For all $\alpha > 0$, the series $B(\alpha, 1)$ and $B(\alpha, -1)$ are absolutely convergent.
c) If $\alpha \in (-1, 0)$, the series $B(\alpha, 1)$ is nonabsolutely convergent.
d) If $\alpha \leq -1$, then $B(\alpha, -1)$ and $B(\alpha, 1)$ are divergent.

PROOF. a) We apply the Ratio Test:

$$\rho = \lim_{n\to\infty} \left| \frac{\binom{\alpha}{n+1}}{\binom{\alpha}{n}} \right| = \lim_{n\to\infty} \left| \frac{\alpha - n}{n+1} \right| = 1,$$

so the radius of convergence is $\frac{1}{\rho} = 1$.

b) Step 0: Let $a > 1$ be a real number and $m \in \mathbb{Z}^+$. Then we have

$$\left(\frac{a}{a-1}\right)^m = \left(1 + \frac{1}{a-1}\right)^m \geq 1 + \frac{m}{a-1} > 1 + \frac{m}{a} = \frac{a+m}{a} > 0,$$

where in the first inequality we have just taken the first two terms of the usual (finite!) binomial expansion. Taking reciprocals, we get

$$\left(\frac{a-1}{a}\right)^m < \frac{a}{a+m}.$$

[5] In fact it is older. For an account of the early history of the binomial series, see [**Co49**].

Step 1: Suppose $\alpha \in (0,1)$. Choose an integer $m \geq 2$ such that $\frac{1}{m} < \alpha$. Then

$$|\binom{\alpha}{n}| = \frac{\alpha(1-\alpha)\cdots(n-1-\alpha)}{n!} < 1(1-\frac{1}{m})\cdots(n-1-\frac{1}{m})\frac{1}{n!}$$

$$= \frac{m-1}{m}\frac{2m-1}{2m}\cdots\frac{(n-1)m-1}{(n-1)m}\frac{1}{n} = a_n\frac{1}{n},$$

say. Using Step 0, we get

$$a_n^{m-1} < \frac{m}{2m-1}\frac{2m}{3m-1}\cdots\frac{(n-1)m}{nm-1}$$

$$= \frac{m}{m-1}\cdots\frac{2m}{2m-1}\cdots\frac{(n-1)m}{(n-1)m-1}\frac{m-1}{nm-1} \leq \frac{1}{a_n}\frac{1}{n}$$

It follows that $a_n < \frac{1}{n^{\frac{1}{m}}}$, so $|\binom{\alpha}{n}| < \frac{1}{n^{1+\frac{1}{m}}}$, so

$$\sum_n |\binom{\alpha}{n}| \leq \sum_n \frac{1}{n^{1+\frac{1}{m}}} < \infty.$$

This shows that $B(\alpha,1)$ is absolutely convergent; since $|\binom{\alpha}{n}(-1)^n| = |\binom{\alpha}{n}|$, it also shows that $B(\alpha,-1)$ is absolutely convergent.
Step 2: Using the identity (89), we find

$$S(\alpha,x) = 1+\sum_{n=1}^{\infty}\binom{\alpha}{n}x^n = 1+\sum_{n=1}^{\infty}\left(\binom{\alpha-1}{n-1}+\binom{\alpha-1}{n}\right)x^n = (1+x)S(\alpha-1,x).$$

Thus for any fixed x, if $S(\alpha-1,x)$ (absolutely) converges, so does $S(\alpha,x)$. By an evident induction argument, if $S(\alpha,x)$ (absolutely) converges, so does $S(\alpha+n,x)$ for all $n \in \mathbb{N}$. Since $S(\alpha,-1)$ and $S(\alpha,1)$ are absolutely convergent for all $\alpha \in (0,1)$, they are thus absolutely convergent for all non-integers $\alpha > 0$.
c) If $\alpha \in (-1,0)$ and $n \in \mathbb{N}$, then

$$\binom{\alpha}{n+1}/\binom{\alpha}{n} = \frac{\alpha-n}{n+1} \in (-1,0);$$

this shows simultaneously that the sequence of terms of $B(\alpha,1) = \sum_{n=0}^{\infty}\binom{\alpha}{n}$ is decreasing in absolute value and alternating in sign. Further, write $\alpha = \beta - 1$, so that $\beta \in (0,1)$. Choose an integer $m \geq 2$ such that $\frac{1}{m} < \beta$. Then

$$|\binom{\alpha}{n}| = \frac{(1-\beta)(2-\beta)\cdots(n-1-\beta)}{(n-1)!}\frac{n-\beta}{n} = b_n\frac{n-\beta}{n}.$$

Arguing as in Step 1 of part b) shows that $b_n < \frac{1}{n^{\frac{1}{m}}}$, and hence

$$\lim_{n\to\infty}|\binom{\alpha}{n}| = \lim_{n\to\infty} b_n \cdot \lim_{n\to\infty}\frac{n-\beta}{n} = 0\cdot 1 = 0.$$

Therefore the Alternating Series Test applies to show that $S(\alpha,1)$ converges.
d) The absolute value of the nth term of both $B(\alpha,-1)$ and $B(\alpha,1)$ is $|\binom{\alpha}{n}|$. If $\alpha \leq -1$, then $|\alpha - n| \geq n+1$ and thus

$$|\binom{\alpha}{n+1}/\binom{\alpha}{n}| = |\frac{\alpha-n}{n+1}| \geq 1,$$

and thus $\binom{\alpha}{n} \not\to 0$. By the Nth term test, $S(\alpha,-1)$ and $S(\alpha,1)$ diverge. □

Exercise*: Show that for $\alpha \in (-1,0)$, the binomial series $B(\alpha,-1)$ diverges.

Remark: As the reader has surely noted, the convergence of the binomial series $S(\alpha,x)$ at $x = \pm 1$ is a rather delicate and tricky enterprise. In fact most texts at this level – even [**S**] – do not treat it. We have taken Step 1 of part b) from [**Ho66**].

Remark: There is an extension of the Ratio Test due to J.L. Raabe which simplifies much the of the above analysis, including the preceding exercise.

THEOREM 12.8. *Let $\alpha \in \mathbb{R} \setminus \mathbb{N}$; let $f(x) = (1+x)^\alpha$, and consider its Taylor series at zero, the* ***binomial series***

$$B(\alpha,x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n.$$

a) For all $x \in (-1,1)$, $f(x) = B(\alpha,x)$.
b) If $\alpha > -1$, $f(1) = B(\alpha,1)$.
c) If $\alpha > 0$, $f(-1) = B(\alpha,-1)$.

PROOF. [**La**] Let $T_{n-1}(x)$ be the $(n-1)$st Taylor polynomial for f at 0, so

$$B(\alpha,x) = \lim_{n\to\infty} T_{n-1}(x)$$

is the Taylor series expansion of f at zero. As usual, put $R_{n-1}(x) = f(x) - T_{n-1}(x)$.
a) By Theorem 12.4b),

$$R_{n-1}(x) = \int_0^x \frac{f^n(t)(x-t)^{n-1}dt}{(n-1)!} = \frac{1}{(n-1)!}\int_0^x \alpha(\alpha-1)\cdots(\alpha-n+1)(1+t)^{\alpha-n}(x-t)^{n-1}dt.$$

By the Mean Value Theorem for Integrals, there is $\theta \in (0,1)$ such that

$$R_{n-1}(x) = \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{(n-1)!}(1+\theta x)^{\alpha-n}(x-\theta x)^{n-1}(x-0).$$

Put

$$t = \frac{1-\theta}{1+\theta x}, c_n(s) = \binom{\alpha-1}{n-1}s^{n-1}.$$

Then

$$(1+s)^{\alpha-1} = \sum_{n=1}^{\infty} c_n(s)$$

and

$$R_{n-1}(x) = c_n(xt)\alpha x(1+\theta x)^{\alpha-1}.$$

Since $x \in (-1,1)$, we have $t \in (0,1)$, so $|xt| < 1$. It follows that $\sum_{n=1}^{\infty} c_n(xt)$ converges, so by the nth term test $c_n(xt) \to 0$ as $n \to \infty$ and thus $R_{n-1}(x) \to 0$.
b) The above argument works verbatim if $x = 1$ and $\alpha > -1$.
c) If $\alpha > 0$, then by Theorem 12.7b), $S(\alpha,-1)$ is convergent. Moreover, $\alpha - 1 > -1$, so $\sum_{n=1}^{\infty} c_n(1)$ converges and thus $c_n(1) \to 0$. But $|c_n(-1)| = |c_n(1)|$, so also $c_n(-1) \to 0$ and thus $R_{n-1}(-1) \to 0$. □

5. Hermite Interpolation

It is a well-known fact that "two points determine a line". One version of this is: given real numbers $x_1 < x_2$ and real numbers f_1, f_2, there is a unique linear function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x_1) = f_1$ and $f(x_2) = f_2$. In a similar way, "three points determine a parabola": given real numbers $x_1 < x_2 < x_3$ and real numbers f_1, f_2, f_3, there is a unique quadratic polynomial $P(x) = ax^2 + bx + c$ such that $P(x_i) = f_i$ for $i = 1, 2, 3$. We will give a generalization.

We work with respect to a fixed sequence

$$x_\bullet = x_1 \leq x_2 \leq \ldots \leq x_n \leq \ldots$$

of real numbers. To include this dependence explicitly in our notation would be cumbersome, so we will leave it implicit.

We define a sequence $\{h_n\}_{n=0}^\infty$ of polynomials, as follows: $h_0 = 1$ and for all $n \in \mathbb{Z}^+$,

$$h_n(t) = (t - x_1) \cdots (t - x_n).$$

For polynomials $a(t), b(t)$ and a nonzero polynomial $c(t)$, we write $a(t) \equiv b(t) \pmod{c(t)}$ if there is a polynomial $q(t)$ such that $a(t) - b(t) = q(t)c(t)$. Equivalently, $a(t)$ and $b(t)$ leave the same remainder upon division by $c(t)$.

EXERCISE 12.1. *Suppose $a(t), b(t)$ are polynomials and $c_1(t) \mid c_2(t)$ are nonzero polynomials. Show: if $a \equiv b \pmod{c_2}$, then $a \equiv b \pmod{c_1}$.*

EXERCISE 12.2. *Suppose $a(t), b(t), c(t)$ are polynomials with $\deg(a), \deg(b) < \deg(c)$ and $a \equiv b \pmod{c}$. Show: $a = b$.*

EXERCISE 12.3. *Let $P_1(t), \ldots, P_n(t)$ be nonzero polynomials such that $\deg(P_1) < \ldots < \deg(P_n)$. Suppose that $A_1P_1(t) + \ldots + A_nP_n(t) = 0$ (i.e., the zero polynomial). Show that $A_1 = \ldots = A_n = 0$.*

THEOREM 12.9. *Let $P(t)$ be a polynomial, and let $n \in \mathbb{Z}^+$.*
a) There are unique real numbers $A_0, \ldots, A_n$ such that

$$P(t) \equiv A_0 + A_1h_1(t) + \ldots + A_nh_n(t) \pmod{h_{n+1}(t)}.$$

b) Suppose $n = \deg(P)$. Then there are unique real numbers $A_0, \ldots, A_n$ such that

$$P(t) = A_0 + A_1h_1(t) + \ldots + A_nh_n(t).$$

PROOF. a) Existence: After replacing $P(t)$ by its remainder upon division by $h_{n+1}(t)$, we may assume $\deg(P) < \deg(h_{n+1})$, i.e., $\deg(P) \leq n$. Now divide P by h_n, getting $q(t)$ and $r(t)$ such that $P(t) = q(t)h_n(t) + r(t)$ and $\deg(r) < \deg(h_n) = n$. In fact, because we are dividing a polynomial $P(t)$ of degree at most n by a polynomial $h_n(t)$ of degree n, the quotient $q(t)$ is a constant, i.e., a real number: call it A_n. Then $P(t) - A_nh_n$ is a polynomial of degree at most $n-1$, and we divide it by h_{n-1}. Continuing in this way, we arrive at the desired result.
Uniquness: If also $P(t) \equiv B_0 + B_1h(t) + \ldots + B_nh(t) \pmod{h_{n+1}(t)}$, then

$$h_{n+1}(t) \mid (A_0 - B_0) + \ldots + (A_n - B_n)h_n(t).$$

Since the left hand side has degree $n+1$ and the right hand side has degree n, this means that $(A_0 - B_0) + (A_1 - B_1)h_1 + \ldots + (A_n - B_n)h_n(t)$ is the zero polynomial, contradicting Exercise X.X.
b) This follows from part a) and Exercise X.X. □

We had better hasten to explain why we care about whether two polynomials are congruent modulo $h_n(t)$: what is the significance of this? The answer is that if $f \equiv g \pmod{h_n(t)}$ then g interpolates f in a precise sense. In particular, this implies that $f(x_i) = g(x_i)$ for all $1 \leq i \leq n$. Indeed, $f \equiv g \pmod{h_n}$ means there is a polynomial q such that $f - g = qh_n(t) = q\prod_{i=1}^n (t - x_i)$, and plugging in $t = x_i$ gives $f(x_i) - g(x_i) = 0$.

Next I claim that in the special case in which $x_n < x_{n+1}$ for all n ("Lagrange Interpolation"), conversely if $f(x_i) = g(x_i)$ for all $1 \leq i \leq n$ then $f \equiv g \pmod{h_n}$. Indeed, $f - g$ vanishes at x_1, so by the Root Factor Theorem is divisible by $t - x_1$:

$$f(t) - g(t) = q_1(t)(t - x_1).$$

Now evaluate the above equation at $t = x_2$:

$$0 = f(x_1) - g(x_2) = q_1(x_2)(x_2 - x_1).$$

Since $x_2 > x_1$, we find that $q_1(x_2) = 0$, so by Root Factor we may write $q_1(t) = q_2(t)(t - x_2)$ and thus $f(t) - g(t) = (t - x_1)(t - x_2)q_2(t)$. And similarly we find that $q_2(x_3) = 0$, and so forth: in the end we get $f(t) - g(t) = (t - x_1) \cdots (t - x_n)q_n(t) = h_n(t)q_n(t)$, so $f \equiv g \pmod{h_n(t)}$.

However, whenever $x_{n-1} = x_n$, the condition $f \equiv g \pmod{h_n}$ is strictly stronger than $f(x_i) = g(x_i)$ for all $1 \leq i \leq n$: the idea is that $f \equiv g \pmod{h_n}$ always imposes "n independent linear conditions", whereas when $x_{n-1} = x_n$ then the constraints $f(x_{n-1}) = g(x_{n-1})$ and $f(x_n) = g(x_n)$ are redundant. For a simple example, suppose $x_1 = x_2 = 0$. Then $h_2 = t^2$, and if two polynomials have the same constant term, then their diference need not be divisible by t^2.

Well, let's explore a bit more: suppose $f \equiv g \pmod{t^2}$, i.e., $f(t) - g(t) = r(t)t^2$. The fundamental observation here is that not only do we have $f(0) = g(0)$ but also $f'(0) = g'(0)$. Indeed, differentiating gives

$$f'(t) - g'(t) = r'(t)t^2 + 2tr(t) = t(tr't) + 2r(t),$$

so plugging in $t = 0$ we get $f'(0) = g'(0)$. Once we have the idea, arriving at the following generalization is not difficult.

LEMMA 12.10. *Let $c \in \mathbb{R}$ and $n \in \mathbb{Z}^+$. For polynomials $f(t)$ and $g(t)$, the following are equivalent:*
(i) We have $f \equiv g \pmod{(t - c)^n}$.
(ii) For all $0 \leq k < n$, we have $f^{(k)}(c) = g^{(k)}(c)$.

PROOF. (i) $\implies$ (ii): We claim that if $f \equiv g \pmod{(t - c)^n}$, then $f' \equiv g' \pmod{(t - c)^{n-1}}$. Indeed, writing $f(t) - g(t) = (t - c)^n r(t)$ and differentiating, we get

$$f'(t) - g'(t) = n(t - c)^{n-1}r(t) + (t - c)^{n-1}r'(t) = (t - c)^{n-1}(nr(t) + r'(t),$$

establishing the claim. From this it follows that for all $0 \leq k < n - 1$, we have $f^{(k)} \equiv g^{(k)} \pmod{(t - c)^{n-k}}$, so by Exercise X.X $f^{(k)} \equiv g^{(k)} \pmod{t - c}$, hence $f^{(k)}(c) = g^{(k)}(c)$.
(ii) $\implies$ (i): By considering the polynomial $h = f - g$, it is enough to show that if $h^{(k)}(c) = 0$ for all $0 \leq k < n$ then $(t - c)^n \mid h$. By polynomial division we may write $h = q(t)(t - c)^n + r(t)$ with $\deg(r) < n$. By the just-proved implication (i)

$\implies$ (ii), we know that for all $0 \leq k < n$, the kth derivative of $(t-c)^n$ at c is 0, so (by the linearity of derivatives) for all $0 \leq k < n$ also $r^{(k)}(c) = 0$. We claim that, in conjunction with $\deg(r) < n$, this implies $r = 0$, hence $(t-c)^n \mid h$. Indeed, suppose not, and write $r = (t-c)^m s$ with $s(c) \neq 0$. Then

$$r' = (t-c)^{m-1}(ms + (t-c)s').$$

Plugging $t = c$ into $ms + (t-c)s'$, the first term is not zero and the second term is, so c is not a root of $ms + (t-c)s'$. Continuing in this manner we find that $r'' = (t-c)^{m-2}s_2$ with $s_2(c) \neq 0$, and so forth: eventually we get that $r^{(m)}(c) \neq 0$. Since $m \leq \deg(r) < n$, this is a contradiction. □

The general meaning of $f \equiv g \pmod{h_n}$ is clear: it enforces equality of not only the values of f and g at $x_1, \ldots, x_n,$, but when there are repetitions in the sequence, equality of some of the derivatives of f and g. A precise enunciation is unfortunately rather technical. Here it goes: for $n \in \mathbb{Z}^+$ and $c \in \mathbb{R}$, let $x_\bullet(c,n)$ be the number of times the real number c occurs in the finite sequence $x_1, \ldots, x_n$. Then:

PROPOSITION 12.11. *For polynomials f and g and $n \in \mathbb{Z}^+$, the following are equivalent:*
(i) We have $f \equiv g \pmod{h_n}$.
(ii) For all $c \in \mathbb{R}$ and all $0 \leq k < x_\bullet(c,n)$, we have $f^{(k)}(c) = g^{(k)}(c)$.

PROOF. First observe that in the case $x_1 = \ldots = x_n$, then $x_\bullet(c,n) = n$ and this reduces to the previous result. Suppose that after removing repetitions from the sequence $x_1, \ldots, x_m$ we get the increasing sequence $c_1 < \ldots < c_r$, and for $1 \leq i \leq r$, let us write m_i for $x_\bullet(c_i, n)$.
(i) $\implies$ (ii): Since for each $1 \leq i \leq r$ we have $(t-c_i)^{m_i} \mid h_n(t)$, condition (i) implies $f \equiv g \pmod{(t-c_i)^{m_i}}$, and then the previous result gives condition (ii).
(ii) $\implies$ (i): Condition (ii) and the previous result tell us that $f \equiv g \pmod{(t-c_i)^{m_i}}$ for all $1 \leq i \leq r$. So for starters, $f - g = (t-c_1)^{m_1} h_1$. We may write $h_1 = (t-c_2)^{d_2} j_1(t)$ with $d_2 \geq 0$ and $j_1(c_2) \neq 0$. Then

$$f - g = (t-c_2)^{d_2}\left((t-c_1)^{m_1} j_1(t)\right) = (t-c_2)^{d_2} k_2(t),$$

where (since $c_1 \neq c_2$) $k_2(c_2) \neq 0$. On the other hand, we can write

$$(t-c_2)^{m_2} q_2 = f - g = (t-c_2)^{d_2} k_2(t),$$

so if $d_2 < m_2$ we get

$$k_2(t) = (t-c_2)^{m_2 - d_2} q_2(t),$$

and plugging in $t = c$ gives a contradiction. So $d_2 \geq m_2$ and thus

$$f - g = (t-c_1)^{m_1}(t-c_2)^{m_2} h_3.$$

Continuing in this manner, eventually we get that $h_n(t) = \prod_{i=1}^r (t-c_i)^{m_i} \mid f - g$. □

In summary, we have proved the following result.

THEOREM 12.12. *(Generalized Polynomial Interpolation) Let $r, m_1, \ldots, m_r \in \mathbb{Z}^+$, and put $n+1 = m_1 + \ldots + m_r$. Let $c_1 < \ldots, c_r$ be real numbers, and for each $1 \leq i \leq r$ and $1 \leq j \leq m_r$, let f_{ij} be a real number. Then there is a unique polynomial P of degree at most n such that for all $1 \leq i \leq r$ and $0 \leq j \leq m_i - 1$,*

$$P^{(j)}(c_i) = f_{ij}.$$

THEOREM 12.13. *(Lagrange Interpolation Formula) Let $n \in \mathbb{Z}^+$, let $x_0 < \ldots < x_n$ be real numbers, and let $y_0, \ldots, y_n$ be real numbers. For $0 \leq j \leq n$, define*

$$\ell_j(x) = \prod_{0 \leq i \leq n,\ i \neq j} \frac{x - x_n}{x_j - x_n} = \frac{x - x_0}{x_j - x_0} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_n}{x_j - x_n}.$$

*Then an explicit formula for the unique polynomial $P(x)$ of degree at most n such that $P(x_i) = y_i$ for all $0 \leq i \leq n$ of Theorem **??** is*

$$P(x) = \sum_{j=0}^{n} y_j \ell_j(x). \tag{90}$$

PROOF. Each $\ell_j(x)$ is a product of $n + 1 - 1 = n$ linear polynomials, so is a polynomial of degree n. Therefore $P(x)$ is a polynomial of degree at most n. To establish (90) the key observation is that for all $1 \leq i, j \leq n$, $\ell_j(x_i)$ is equal to 1 if $i = j$ and 0 if $i \neq j$. We ask the reader to just stop and think about this: the formula looks a little complicated at first, so it is natural to worry that it may not be so easy to see this...but in fact it is immediate. The equation (90) follows, also immediately. □

The polynomials $\ell_j(x)$ are called **Lagrange basis polynomials**. One of the merits of the formula is that they do not depend on the values $y_0, \ldots, y_n$ – or, if you like, the function $y = f(x)$ that we are interpolating – but only on the values $x_0, \ldots, x_n$ – which in the lingo of this field are often called the **interpolation nodes**.

Now look back at (**??**): it is reminiscent of the degree n Taylor polynomial, except instead of taking powers of $x - x_0$ for a fixed center x_0 we are taking different points x_i. Just brainstorming then, might it be possible to generalize Theorem **??** by allowing some of the x_i's to coincide, i.e., replacing the condition $x_0 < \ldots < x_n$ with $x_0 \leq x_1 \leq \ldots \leq x_n$? At first glance no: if for instance $x_1 = x_2$ then the conditions $f(x_1) = f_1$ and $f(x_2) = f_2$ are inconsistent unless $f_1 = f_2$, in which case they are simply redundant. But there is a way of salvaging the situation. The key observation is that multiple roots of a polynomial function show up as roots of the derivative. More precisely, if a polynomial f can be written as $(x - c)^m g(x)$ with $g(c) \neq 0$, then

$$f' = m(x - c)^{m-1} g(x) + (x - c)^m g'(x) = (x - c)^{m-1} \left(m g(x) + (x - c) g'(x) \right).$$

Observe that when we plug in $x = c$ the expression $mg(c) + (c - c)g'(c) = mg(c) \neq 0$. This shows that if a polynomial f has a root of multiplicity $m \geq 1$ at a point c, then the polynomial f' has a root of multiplicity $m - 1$ at c.

A reasonable way of repairing "$f(x_1) = f_1$, $f(x_2) = f_2$" when $x_1 = x_2$ then is to take the second condition as a condition on the derivative of f at x_2: i.e., $f'(x_2) = f_2$. But it is probably clearer to switch to a different notation: suppose that we have r distinct real numbers $x_1 < \ldots < x_r$, and each occurs with a certain **multiplicity** m_i. Thus our list of numbers contains m_1 instances of x_1, m_2 instances of x_2, up to m_r instances of x_r, thus $m_1 + \ldots + m_r$ entries altogether. Let's put $n + 1 = m_1 + \ldots + m_r$. Here is a more general interpolation theorem.
Let us now try to switch back to the old notation: we give ourselves $n + 1$ real numbers "with multiplicity": $x_0 \leq x_1 \leq \ldots \leq x_n$, a and $n + 1$ real numbers $f_0, \ldots, f_n$. We write the interpolation problem as above as $f(x_i) = f_i$, but with the understanding that when a root is repeated more than once, the further conditions

are conditions on the derivatives of f. In this case we claim that the interpolation polynomial can be taken in the same form as above: namely, there are unique real numbers $A_0, \ldots, A_n$ such that

$$f = f(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \ldots + A_n(x - x_0) \cdots (x - x_{n-1}).$$

At the moment we will prove this by linear algebraic considerations (which is cheating: we are not supposed to be assuming any knowledge of linear algebra in this text!). Namely, since we have already shown the existence of an interpolating polynomial f of degree at most n, it suffices to show that the set of polynomials

$$\mathcal{S} = \{1, x - x_0, (x - x_0)(x - x_1), \ldots, (x - x_0) \cdots (x - x_{n-1})\}$$

spans the $\mathbb{R}$-vector space $\mathcal{P}_n$ of all polynomials of degree at most n. The set $\mathcal{S}$ is linearly independent: indeed, the polynomials have distinct degrees, so a nontrivial linear indeendence relationship would allow us to write a nonzero polynomial as a linear combination of polynomials of smaller degree, which is absurd. Further, $\#\mathcal{S} = n + 1$. But $\mathcal{P}_n$ has dimension n, so $\mathcal{S}$ must be a basis for $\mathcal{P}_n$: in paticular $\mathcal{S}$ spans $\mathcal{P}_n$.

Having billed Theorem 12.12 as generalizing Theorem **??**, let us now call attention to the other extreme: suppose $x_0 = \ldots = x_n = c$, say. Then the interpolating polynomial P is precisely the degree n Taylor polynomial at c to any n times differentiable function f with $f^{(j)}(c) = f_j$ for $0 \leq j \leq n$. This brings up a key idea in general: let I be an interval containing the points $x_0 \leq \ldots \leq x_n$, and let $f : I \to \mathbb{R}$ be an n times differentiable function. We define the **Hermite interpolation polynomial** $P(x)$ to be the unique polynomial of degree at most n such that for all $0 \leq i \leq n$, $P(x_i) = f(x_i)$: here we are using the above slightly shady convention that when the x_i's occur with multiplicity greater than 1, the conditions $P(x_i) = f(x_i)$ are actually conditions on the derivatives of P and f at x_i.

Let us define the **remainder function**: for $x \in I$,

$$R(x) = f(x) - P(x).$$

Following [**CJ**], we will now give an expression for R which generalizes one form of Taylor's Theorem With Remainder. We begin with one preliminary result.

THEOREM 12.14. *(Generalized Rolle's Theorem) Let $f : I \to \mathbb{R}$ be n times differentiable, and assume that f has at least $n+1$ roots on I, counted with multiplicity. Then there is $\zeta \in I$ with $f^{(n)}(\zeta) = 0$.*

Exercise: Prove Theorem 12.14.

THEOREM 12.15. *(Hermite With Remainder) Let $x_0 \leq \ldots \leq x_n \in I$, and let $f : I \to \mathbb{R}$ be $(n + 1)$ times differentiable. Let P be the Hermite Interpolation Polynomial for f. Then, for all $x \in I$, there is $\zeta \in I$ – in fact, lying in any closed interval containing $x, x_0, \ldots, x_n$ – such that*

$$R(x) = f(x) - P(x) = \frac{(x - x_0) \cdots (x - x_n)}{(n + 1)!} f^{(n+1)}(\zeta). \tag{91}$$

PROOF. If $x = x_i$ for some i, then both sides of (91) are 0, so equality holds. We may thus assume $x \neq x_i$ for any i. Let $c \in \mathbb{R}$, and consider

$$K(x) = R(x) - c(x - x_0) \cdots (x - x_n).$$

There is a unique value of c such that $K(x) = 0$: namely,

$$c = \frac{R(x)}{(x - x_0) \cdots (x - x_n)}.$$

The function $K : I \to \mathbb{R}$ thus vanishes at least $n + 2$ times on I with multiplicity, so by the Generalized Rolle's Theorem there is $\zeta \in I$ such that

$$0 = K^{(n+1)}(\zeta) = f^{(n+1)}(\zeta) - P^{(n+1)}(\zeta) - (n+1)!c = f^{(n+1)}(\zeta) - (n+1)!c,$$

so

$$c = \frac{f^{(n+1)}(\zeta)}{(n+1)!}.$$

It follows that

$$R(x) = \frac{(x - x_0) \cdots (x - x_n)}{(n+1)!} f^{(n+1)}(\zeta).$$

□

Remark: Restricting to Taylor polynomials, our earlier argument for the existence of the interpolating polynomial is certainly easier: recall this consisted of simply writing down the answer and checking that it was correct. However this proof of part b) of Taylor's Theorem with Remainder seems easier.

Exercise: a) Let $x_0 \leq \ldots \leq x_n$, and let $m \leq n$. Let

$$P_n(x) = A_0 + A_1(x - x_0) + \ldots + A_n(x - x_0) \cdots (x - x_{n-1})$$

be the Hermite interpolation polynomial for a function f. Show that the Hermite interpolation polynomial for f with respect to the approximation points $x_0 \leq \ldots \leq x_k$ is

$$P_m(x) = A_0 + A_1(x - x_0) + \ldots + A_m(x - x_0) \cdots (x - x_{m-1}).$$

b) Suppose $x_{n-1} \neq x_n$. Show that there is $\zeta \in [x_0, x_n]$ such that

$$A_n = \frac{f^{(n)}(\zeta)}{n!}.$$

c) Show that there is a sequence $\{\zeta_k\}$ taking values in I such that

$$A_n = \lim_{k \to \infty} \frac{f^{(n)}(\zeta_k)}{n!}.$$

d) Suppose that f has a continuous $(n+1)$st derivative. Use part c) to recover the formula or the nth Taylor series coefficient.

CHAPTER 13

Sequences and Series of Functions

1. Pointwise Convergence

All we have to do now is take these lies and make them true somehow. – G. Michael[1]

1.1. Pointwise convergence: cautionary tales.

Let I be an interval in the real numbers. A **sequence of real functions** is a sequence $f_0, f_1, \ldots, f_n, \ldots$, with each f_n a function from I to $\mathbb{R}$.

For us the following example is all-important: let $f(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence $R > 0$. So f may be viewed as a function $f : (-R, R) \to \mathbb{R}$. Put $f_n = \sum_{k=0}^{n} a_k x^k$, so each f_n is a polynomial of degree at most n; therefore f_n makes sense as a function from $\mathbb{R}$ to $\mathbb{R}$, but let us restrict its domain to $(-R, R)$. Then we get a sequence of functions $f_0, f_1, \ldots, f_n, \ldots$.

As above, our stated goal is to show that the function f has many desirable properties: it is continuous and indeed infinitely differentiable, and its derivatives and antiderivatives can be computed term-by-term. Since the functions f_n have all these properties (and more – each f_n is a polynomial), it seems like a reasonable strategy to define some sense in which the sequence $\{f_n\}$ **converges** to the function f, in such a way that this converges process *preserves* the favorable properties of the f_n's.

The previous description perhaps sounds overly complicated and mysterious, since in fact there is an evident sense in which the sequence of functions f_n converges to f. Indeed, to say that x lies in the open interval $(-R, R)$ of convergence is to say that the sequence $f_n(x) = \sum_{k=0}^{n} a_k x^k$ converges to $f(x)$.

This leads to the following definition: if $\{f_n\}_{n=1}^{\infty}$ is a sequence of real functions defined on some interval I and $f : I \to \mathbb{R}$ is another function, we say f_n **converges to f pointwise** if for all $x \in I$, $f_n(x) \to f(x)$. (We also say f is the **pointwise limit** of the sequence $\{f_n\}$.) In particular the sequence of partial sums of a power series converges pointwise to the power series on the interval I of convergence.

Remark: There is similarly a notion of an infinite series of functions $\sum_{n=0}^{\infty} f_n$ and of pointwise convergence of this series to some limit function f. Indeed, as in the case of just one series, we just define $S_n = f_0 + \ldots + f_n$ and say that $\sum_n f_n$ converges pointwise to f if the sequence S_n converges pointwise to f.

[1] George Michael, 1963–

The great mathematicians of the 17th, 18th and early 19th centuries encountered many sequences and series of functions (again, especially power series and Taylor series) and often did not hesitate to assert that the pointwise limit of a sequence of functions having a certain nice property itself had that nice property.[2] The problem is that statements like this unfortunately need not be true!

Example 1: Define $f_n = x^n : [0,1] \to \mathbb{R}$. Clearly $f_n(0) = 0^n = 0$, so $f_n(0) \to 0$. For any $0 < x \leq 1$, the sequence $f_n(x) = x^n$ is a geometric sequence with geometric ratio x, so that $f_n(x) \to 0$ for $0 < x < 1$ and $f_n(1) \to 1$. It follows that the sequence of functions $\{f_n\}$ has a pointwise limit $f : [0,1] \to \mathbb{R}$, the function which is 0 for $0 \leq x < 1$ and 1 at $x = 1$. Unfortunately the limit function is discontinuous at $x = 1$, despite the fact that each of the functions f_n are continuous (and are polynomials, so really as nice as a function can be). Therefore **the pointwise limit of a sequence of continuous functions need not be continuous**.

Remark: Example 1 was chosen for its simplicity, not to exhibit maximum pathology. It is possible to construct a sequence $\{f_n\}_{n=1}^{\infty}$ of polynomial functions converging pointwise to a function $f : [0,1] \to \mathbb{R}$ that has infinitely many discontinuities! (On the other hand, it turns out that it is not possible for a pointwise limit of continuous functions to be discontinuous at *every* point. This is a theorem of R. Baire. But we had better not talk about this, or we'll get distracted from our stated goal of establishing the wonderful properties of power series.)

One can also find assertions in the math papers of old that if f_n converges to f pointwise on an interval $[a,b]$, then $\int_a^b f_n dx \to \int_a^b f dx$. To a modern eye, there are in fact two things to establish here: first that if each f_n is Riemann integrable, then the pointwise limit f must be Riemann integrable. And second, that *if* f is Riemann integrable, its integral is the limit of the sequence of integrals of the f_n's. In fact *both* of these are false!

Example 2: Define a sequence $\{f_n\}_{n=0}^{\infty}$ with common domain $[0,1]$ as follows. Let f_0 be the constant function 1. Let f_1 be the function which is constantly 1 except $f(0) = f(1) = 0$. Let f_2 be the function which is equal to f_1 except $f(1/2) = 0$. Let f_3 be the function which is equal to f_2 except $f(1/3) = f(2/3) = 0$. And so forth. To get from f_n to f_{n+1} we change the value of f_n at the finitely many rational numbers $\frac{a}{n}$ in $[0,1]$ from 1 to 0. Thus each f_n is equal to 1 except at a finite set of points: in particular it is bounded with only finitely many discontinuities, so it is Riemann integrable. The functions f_n converges pointwise to a function f which is 1 on every irrational point of $[0,1]$ and 0 on every rational point of $[0,1]$. Since every open interval (a,b) contains both rational and irrational numbers, the function f is not Riemann integrable: for any partition of $[0,1]$ its upper sum is 1 and its lower sum is 0. Thus a pointwise limit of Riemann integrable functions need not be Riemann integrable.

[2]This is an exaggeration. The precise definition of convergence of real sequences did not come until the work of Weierstrass in the latter half of the 19th century. Thus mathematicians spoke of functions f_n "approaching" or "getting infinitely close to" a fixed function f. Exactly what they meant by this – and indeed, whether even they knew exactly what they meant (presumably some did better than others) is a matter of serious debate among historians of mathematics.

Example 3: We define a sequence of functions $f_n : [0,1] \to \mathbb{R}$ as follows: $f_n(0) = 0$, and $f_n(x) = 0$ for $x \geq \frac{1}{n}$. On the interval $[0, \frac{1}{n}]$ the function forms a "spike": $f(\frac{1}{2n}) = 2n$ and the graph of f from $(0,0)$ to $(\frac{1}{2n}, 2n)$ is a straight line, as is the graph of f from $(\frac{1}{2n}, 2n)$ to $(\frac{1}{n}, 0)$. In particular f_n is piecewise linear hence continuous, hence Riemann integable, and its integral is the area of a triangle with base $\frac{1}{n}$ and height $2n$: $\int_0^1 f_n dx = 1$. On the other hand this sequence converges pointwise to the zero function f. So

$$\lim_{n\to\infty} \int_0^1 f_n = 1 \neq 0 = \int_0^1 \lim_{n\to\infty} f_n.$$

Example 4: Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded differentiable function such that $\lim_{n\to\infty} g(n)$ does not exist. (For instance, we may take $g(x) = \sin(\frac{\pi x}{2})$.) For $n \in \mathbb{Z}^+$, define $f_n(x) = \frac{g(nx)}{n}$. Let M be such that $|g(x)| \leq M$ for all $x \in \mathbb{R}$. Then for all $x \in \mathbb{R}$, $|f_n(x)| \leq \frac{M}{n}$, so f_n converges pointwise to the function $f(x) \equiv 0$ and thus $f'(x) \equiv 0$. In particular $f'(1) = 0$. On the other hand, for any fixed nonzero x, $f_n'(x) = \frac{ng'(nx)}{n} = g'(nx)$, so

$$\lim_{n\to\infty} f_n'(1) = \lim_{n\to\infty} g'(n) \text{ does not exist.}$$

Thus

$$\lim_{n\to\infty} f_n'(1) \neq (\lim_{n\to\infty} f_n)'(1).$$

A common theme in all these examples is the **interchange of limit operations**: that is, we have some other limiting process corresponding to the condition of continuity, integrability, differentiability, integration or differentiation, and we are wondering whether it changes things to perform the limiting process on each f_n individually and then take the limit versus taking the limit first and then perform the limiting process on f. As we can see: in general it does matter! This is not to say that the interchange of limit operations is something to be systematically avoided. On the contrary, it is an essential part of the subject, and in "natural circumstances" the interchange of limit operations is probably valid. But we need to develop theorems to this effect: i.e., under *some specific additional hypotheses*, interchange of limit operations is justified.

2. Uniform Convergence

Most of the above pathologies vanish if we consider a stronger notion of convergence.

Let $\{f_n\}$ be a sequence of functions with domain I. We say f_n **converges uniformly** to f and write $f_n \stackrel{u}{\to} f$ if for all $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$ and all $x \in I$, $|f_n(x) - f(x)| < \epsilon$.

How does this definition differ from that of pointwise convergence? Let's compare: $f_n \to f$ pointwise if for all $x \in I$ and all $\epsilon > 0$, there exists $n \in \mathbb{Z}^+$ such that for all $n \geq N$, $|f_n(x) - f(x)|\epsilon$. The only difference is in the order of the quantifiers: in pointwise convergence we are first given ϵ and x and then must find an $N \in \mathbb{Z}^+$: that is, the N is allowed to depend both on ϵ and the point $x \in I$. In the definition of uniform convergence, we are given $\epsilon > 0$ and must find an $N \in \mathbb{Z}^+$ which

works simultaneously (or "uniformly") for all $x \in I$. Thus uniform convergence is a stronger condition than pointwise convergence, and in particular if f_n converges to f uniformly, then certainly f_n converges to f pointwise.

LEMMA 13.1. *(Cauchy Criterion For Uniform Convergence) Let $f_n : I \to \mathbb{R}$ be a sequence of functions. The following are equivalent:*
(i) For all $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$ and all $x \in I$, $|f_m(x) - f_n(x)| < \epsilon$.
(ii) For all $\epsilon > 0$, there is $N \in \mathbb{Z}^+$ such that for all $n \geq N$, all $k \geq 0$ and all $x \in I$, $|f_{n+k}(x) - f_n(x)| < \epsilon$.
(iii) $f_n \overset{u}{\to} f$.

Exercise: Prove Lemma 13.1.

2.1. Uniform Convergence and Inherited Properties.

The following result is the most basic one fitting under the general heading "uniform convergence justifies the exchange of limiting operations."

THEOREM 13.2. *Let $\{f_n\}$ be a sequence of functions with common domain I, and let c be a point of I. Suppose that for all $n \in \mathbb{Z}^+$, $\lim_{x \to c} f_n(x) = L_n$. Suppose moreover that $f_n \overset{u}{\to} f$. Then the sequence $\{L_n\}$ is convergent, $\lim_{x \to c} f(x)$ exists and we have equality:*

$$\lim_{n \to \infty} L_n = \lim_{n \to \infty} \lim_{x \to c} f_n(x) = \lim_{x \to c} f(x) = \lim_{x \to c} \lim_{n \to \infty} f_n(x).$$

PROOF. Step 1: We show that the sequence $\{L_n\}$ is convergent. Since we don't yet have a real number to show that it converges *to*, it is natural to try to use the Cauchy criterion, hence to try to bound $|L_m - L_n|$. Now comes the trick: for all $x \in I$ we have

$$|L_m - L_n| \leq |L_m - f_m(x)| + |f_m(x) - f_n(x)| + |f_n(x) - L_n|.$$

By the Cauchy criterion for uniform convergence, for any $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$ and all $x \in I$ we have $|f_m(x) - f_n(x)| < \frac{\epsilon}{3}$. Moreover, the fact that $f_m(x) \to L_m$ and $f_n(x) \to L_n$ give us bounds on the first and last terms: there exists $\delta > 0$ such that if $0 < |x - c| < \delta$ then $|L_n - f_n(x)| < \frac{\epsilon}{3}$ and $|L_m - f_m(x)| < \frac{\epsilon}{3}$. Combining these three estimates, we find that by taking $x \in (c - \delta, c + \delta)$, $x \neq c$ and $m, n \geq N$, we have

$$|L_m - L_n| \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

So the sequence $\{L_n\}$ is Cauchy and hence convergent, say to the real number L.
Step 2: We show that $\lim_{x \to c} f(x) = L$ (so in particular the limit exists!). Actually the argument for this is very similar to that of Step 1:

$$|f(x) - L| \leq |f(x) - f_n(x)| + |f_n(x) - L_n| + |L_n - L|.$$

Since $L_n \to L$ and $f_n(x) \to f(x)$, the first and last term will each be less than $\frac{\epsilon}{3}$ for sufficiently large n. Since $f_n(x) \to L_n$, the middle term will be less than $\frac{\epsilon}{3}$ for x sufficiently close to c. Overall we find that by taking x sufficiently close to (but not equal to) c, we get $|f(x) - L| < \epsilon$ and thus $\lim_{x \to c} f(x) = L$. □

COROLLARY 13.3. *Let f_n be a sequence of continuous functions with common domain I and suppose that $f_n \overset{u}{\to} f$ on I. Then f is continuous on I.*

Since Corollary 13.3 is easier than Theorem 13.2, we include a separate proof.

PROOF. Let $x \in I$. We need to show that $\lim_{x \to c} f(x) = f(c)$, thus we need to show that for any $\epsilon > 0$ there exists $\delta > 0$ such that for all x with $|x - c| < \delta$ we have $|f(x) - f(c)| < \epsilon$. The idea – again! – is to trade this one quantity for three quantities that we have an immediate handle on by writing

$$|f(x) - f(c)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(c)| + |f_n(c) - f(c)|.$$

By uniform convergence, there exists $n \in \mathbb{Z}^+$ such that $|f(x) - f_n(x)| < \frac{\epsilon}{3}$ for all $x \in I$: in particular $|f_n(c) - f(c)| = |f(c) - f_n(c)| < \frac{\epsilon}{3}$. Further, since $f_n(x)$ is continuous, there exists $\delta > 0$ such that for all x with $|x - c| < \delta$ we have $|f_n(x) - f_n(c)| < \frac{\epsilon}{3}$. Consolidating these estimates, we get

$$|f(x) - f(c)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \qquad \square$$

Exercise: Consider again $f_n(x) = x^n$ on the interval $[0, 1]$. We saw in Example 1 above that f_n converges pointwise to the discontinuous function f which is 0 on $[0, 1)$ and 1 at $x = 1$.
a) Show directly from the definition that the convergence of f_n to f *is not* uniform.
b) Try to pinpoint exactly where the proof of Theorem 13.2 breaks down when applied to this non-uniformly convergent sequence.

Exercise: Let $f_n : [a, b] \to \mathbb{R}$ be a sequence of functions. Show TFAE:
(i) $f_n \overset{u}{\to} f$ on $[a, b]$.
(ii) $f_n \overset{u}{\to} f$ on $[a, b)$ and $f_n(b) \to f(b)$.

THEOREM 13.4. *Let $\{f_n\}$ be a sequence of (Riemann-Darboux) integrable functions with common domain $[a, b]$. Suppose that $f_n \overset{u}{\to} f$. Then f is integrable and*

$$\lim_{n \to \infty} \int_a^b f_n = \int_a^b \lim_{n \to \infty} f_n = \int_a^b f.$$

PROOF. Step 1: We prove the integrability of f. Fix $\epsilon > 0$; since $f \overset{u}{\to} f$, there is $N \in \mathbb{Z}^+$ such that for all $n \geq N$ and all $x \in [a, b]$, $|f_n(x) - f(x)| < \epsilon$; it follows that for any subinterval $[c, d] \subset [a, b]$,

$$|\sup(f_n, [c, d]) - \sup(f, [c, d])| \leq \epsilon, |\inf(f_n, [c, d]) - \inf(f, [c, d])| \leq \epsilon.$$

So for any partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a, b]$ and $n \geq N$,

$$|U(f_n, \mathcal{P}) - U(f, \mathcal{P})| \leq \sum_{i=0}^{n-1} |\sup(f_n, [x_n, x_{n+1}]) - \sup(f, [x_n, x_{n+1}]|(x_{i+1} - x_i)|$$

$$\leq \sum_{i=0}^{n-1} \epsilon(x_{i+1} - x_i) = (b - a)\epsilon,$$

and similarly,

$$|L(f_n, \mathcal{P}) - L(f, \mathcal{P})| \leq (b - a)\epsilon.$$

Since f_N is integrable, by Darboux's Criterion there is a partition $\mathcal{P}$ of $[a, b]$ such that $U(f_N, \mathcal{P}) - L(f_N, \mathcal{P}) < \epsilon$. Thus

$$|U(f, \mathcal{P}) - L(f, \mathcal{P})| \leq |U(f, \mathcal{P}) - U(f_n, \mathcal{P})| + |U(f_n, \mathcal{P}) - L(f_n, \mathcal{P})| + |L(f_n, \mathcal{P}) - L(f, \mathcal{P})|$$

$$\leq (b - a)\epsilon + \epsilon + (b - a)\epsilon = (2(b - a) + 1)\epsilon.$$

Since $\epsilon > 0$ was arbitrary, Darboux's Criterion shows f is integrable on $[a,b]$.
Step 2: If $f, g : [a,b] \to \mathbb{R}$ are integrable and $|f(x) - g(x)| \leq \epsilon$ for all $x \in [a,b]$, then

$$|\int_a^b f - \int_a^b g| = |\int_a^b f - g| \leq \int_a^b |f - g| \leq (b-a)\epsilon.$$

From this simple observation and Step 1 the fact that $f_n \stackrel{u}{\to} f$ implies $\int_a^b f_n \to \int_a^b f$ is almost immediate. The details are left to you. □

Exercise: It follows from Theorem 13.4 that the sequences in Examples 2 and 3 above are not uniformly convergent. Verify this directly.

COROLLARY 13.5. *Let $\{f_n\}$ be a sequence of continuous functions defined on the interval $[a,b]$ such that $\sum_{n=0}^{\infty} f_n \stackrel{u}{\to} f$. For each n, let $F_n : [a,b] \to \mathbb{R}$ be the unique function with $F_n' = f_n$ and $F_n(a) = 0$, and similarly let $F : [a,b] \to \mathbb{R}$ be the unique function with $F' = f$ and $F(a) = 0$. Then $\sum_{n=0}^{\infty} F_n \stackrel{u}{\to} F$.*

Exercise: Prove Corollary 13.5.

Our next order of business is to discuss differentiation of sequences of functions. For this we should reconsider Example 4: let $g : \mathbb{R} \to \mathbb{R}$ be a bounded differentiable function such that $\lim_{n\to\infty} g(n)$ does not exist, and let $f_n(x) = \frac{g(nx)}{n}$. Let M be such that $|g(x)| \leq M$ for all $\mathbb{R}$. Then for all $x \in \mathbb{R}$, $|f_n(x)| \leq \frac{M}{n}$, so $f_n \stackrel{u}{\to} 0$. But as we saw above, $\lim_{n\to\infty} f_n'(1)$ does not exist.

Thus we have shown the following somewhat distressing fact: *uniform convergence of f_n to f does not imply that f_n' converges.*

Well, don't panic. What we want is true *in practice*; we just need suitable hypotheses. We will give a relatively simple result sufficient for our coming applications.

THEOREM 13.6. *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions on $[a,b]$. We suppose:*
(i) Each f_n is continuously differentiable on $[a,b]$,
(ii) The functions f_n converge pointwise on $[a,b]$ to some function f, and
(iii) The functions f_n' converge uniformly on $[a,b]$ to some function g.
Then f is differentiable and $f' = g$, or in other words

$$(\lim_{n\to\infty} f_n)' = \lim_{n\to\infty} f_n'.$$

PROOF. Let $x \in [a,b]$. Since $f_n' \stackrel{u}{\to} g$ on $[a,b]$, certainly $f_n' \stackrel{u}{\to} g$ on $[a,x]$. Since each f_n' is continuous, by Corollary 13.3 g is continuous. Now applying Theorem 13.4 and the Fundamental Theorem of Calculus we have

$$\int_a^x g = \int_a^x \lim_{n\to\infty} f_n' = \lim_{n\to\infty} \int_a^x f_n' = \lim_{n\to\infty} f_n(x) - f_n(a) = f(x) - f(a).$$

Differentiating and applying the Fundamental Theorem of Calculus, we get

$$g = (f(x) - f(a))' = f'.$$ □

COROLLARY 13.7. *Let $\sum_{n=0}^{\infty} f_n(x)$ be a series of functions converging pointwise to $f(x)$. Suppose that each f_n' is continuously differentiable and $\sum_{n=0}^{\infty} f_n'(x) \stackrel{u}{\to} g$. Then f is differentiable and $f' = g$:*

$$(\sum_{n=0}^{\infty} f_n)' = \sum_{n=0}^{\infty} f_n'. \tag{92}$$

Exercise: Prove Corollary 13.7.

When for a series $\sum_n f_n$ it holds that $(\sum_n f_n)' = \sum_n f_n'$, we say that the series can be differentiated **termwise** or **term-by-term**. Thus Corollary 13.7 gives a condition under which a series of functions can be differentiated termwise.

Although Theorem 13.6 (or more precisely, Corollary 13.7) will be sufficient for our needs, we cannot help but record the following stronger version.

THEOREM 13.8. *Let $\{f_n\}$ be differentiable functions on the interval $[a,b]$ such that $\{f_n(x_0)\}$ is convergent for some $x_0 \in [a,b]$. If there is $g : [a,b] \to \mathbb{R}$ such that $f_n' \overset{u}{\to} g$ on $[a,b]$, then there is $f : [a,b] \to \mathbb{R}$ such that $f_n \overset{u}{\to} f$ on $[a,b]$ and $f' = g$.*

PROOF. [**R**, pp.152-153]
Step 1: Fix $\epsilon > 0$, and choose $N \in \mathbb{Z}^+$ such that $m,n \geq N$ implies $|f_m(x_0) - f_n(x_0)| \frac{\epsilon}{2}$ and $|f_m'(t) - f_n'(t)| < \frac{\epsilon}{2(b-a)}$ for all $t \in [a,b]$. The latter inequality is telling us that the derivative of $g := f_m - f_n$ is small on the entire interval $[a,b]$. Applying the Mean Value Theorem to g, we get a $c \in (a,b)$ such that for all $x,t \in [a,b]$ and all $m,n \geq N$,

$$|g(x) - g(t)| = |x-t||g'(c)| \leq |x-t| \left(\frac{\epsilon}{2(b-a)} \right) \leq \frac{\epsilon}{2}. \tag{93}$$

It follows that for all $x \in [a,b]$,

$$|f_m(x) - f_n(x)| = |g(x)| \leq |g(x) - g(x_0)| + |g(x_0)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

By the Cauchy criterion, f_n is uniformly convergent on $[a,b]$ to some function f.
Step 2: Now **fix** $x \in [a,b]$ and define

$$\varphi_n(t) = \frac{f_n(t) - f_n(x)}{t - x}$$

and

$$\varphi(t) = \frac{f(t) - f(x)}{t - x},$$

so that for all $n \in \mathbb{Z}^+$, $\lim_{x \to t} \varphi_n(t) = f_n'(x)$. Now by (93) we have

$$|\varphi_m(t) - \varphi_n(t)| \leq \frac{\epsilon}{2(b-a)}$$

for all $m,n \geq N$, so once again by the Cauchy criterion φ_n converges uniformly for all $t \neq x$. Since $f_n \overset{u}{\to} f$, we get $\varphi_n \overset{u}{\to} \varphi$ for all $t \neq x$. Finally we apply Theorem 13.2 on the interchange of limit operations:

$$f'(x) = \lim_{t \to x} \varphi(t) = \lim_{t \to x} \lim_{n \to \infty} \varphi_n(t) = \lim_{n \to \infty} \lim_{t \to x} \varphi_n(t) = \lim_{n \to \infty} f_n'(x).$$

□

2.2. The Weierstrass M-test.

We have just seen that uniform convergence of a sequence of functions (and possibly, of its derivatives) has many pleasant consequences. The next order of business is to give a useful general criterion for a sequence of functions to be uniformly convergent.

For a function $f : I \to \mathbb{R}$, we define

$$||f|| = \sup_{x \in I} |f(x)|.$$

In (more) words, $||f||$ is the least $M \in [0, \infty]$ such that $|f(x)| \leq M$ for all $x \in I$.

THEOREM 13.9. *(Weierstrass M-Test) Let $\{f_n\}_{n=0}^{\infty}$ be a sequence of functions defined on an interval I. Let $\{M_n\}_{n=0}^{\infty}$ be a non-negative sequence such that $||f_n|| \leq M_n$ for all n and $M = \sum_{n=0}^{\infty} M_n < \infty$. Then $\sum_{n=0}^{\infty} f_n$ is uniformly convergent.*

PROOF. Let $S_N(x) = \sum_{n=0}^{N} f_n(x)$. Since $\sum_n M_n < \infty$, for each $\epsilon > 0$ there is $N_0 \in \mathbb{Z}^+$ such that for all $N \geq N_0$, $\sum_{n>N} M_n < \epsilon$. For $x \in I$, $N \geq N_0$ and $k \in \mathbb{N}$,

$$|S_{N+k}(x) - S_N(x)| = \left| \sum_{n=N+1}^{N+k} f_n(x) \right| \leq \sum_{n>N} |f_n(x)| \leq \sum_{n>N} M_n < \epsilon.$$

Therefore the series is uniformly convergent by the Cauchy criterion. □

3. Power Series II: Power Series as (Wonderful) Functions

THEOREM 13.10. *(Wonderful Properties of Power Series) Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence $R > 0$. Consider $f(x) = \sum_{n=0}^{\infty} a_n x^n$ as a function $f : (-R, R) \to \mathbb{R}$. Then:*
a) f is continuous.
b) f is differentiable. Morever, its derivative may be computed termwise*:*

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

c) Since the power series f' has the same radius of convergence $R > 0$ as f, f is in fact infinitely differentiable.
d) For all $n \in \mathbb{N}$, $f^{(n)}(0) = (n!)a_n$.

PROOF.

a) Let $0 < A < R$, so f defines a function from $[-A, A]$ to $\mathbb{R}$. We claim that the series $\sum_n a_n x^n$ converges to f uniformly on $[-A, A]$. Indeed, as a function on $[-A, A]$, we have $||a_n x^n|| = |a_n| A^n$, and thus $\sum_n ||a_n x^n|| = \sum_n |a_n| A^n < \infty$, because power series converge absolutely on the interior of their interval of convergence. Thus by the Weierstrass M-test f is the uniform limit of the sequence $S_n(x) = \sum_{k=0}^{n} a_k x^k$. But each S_n is a polynomial function, hence continuous and infinitely differentiable. So by Theorem 13.2 f is continuous on $[-A, A]$. Since any $x \in (-R, R)$ lies in $[-A, A]$ for some $0 < A < R$, f is continuous on $(-R, R)$.
b) According to Corollary 13.7, in order to show that $f = \sum_n a_n x^n = \sum_n f_n$ is differentiable and the derivative may be compuited termwise, it is enough to check that (i) each f_n is continuously differentiable and (ii) $\sum_n f_n'$ is uniformly convergent. But (i) is trivial, since $f_n = a_n x^n$ – of course monomial functions are continuously differentiable. As for (ii), we compute that $\sum_n f_n' = \sum_n (a_n x^n) = \sum_n n a_{n-1} x^{n-1}$. By X.X, this power series also has radius of convergence R, hence by the result of part a) it is uniformly convergent on $[-A, A]$. Therefore Corollary 13.7 applies to show $f'(x) = \sum_{n=0}^{\infty} n a_n x^{n-1}$.
c) We have just seen that for a power series f convergent on $(-R, R)$, its derivative f' is also given by a power series convergent on $(-R, R)$. So we may continue in

this way: by induction, derivatives of all orders exist.
d) The formula $f^{(n)}(0) = (n!)a_n$ is simply what one obtains by repeated termwise differentiation. We leave this as an exercise to the reader. □

Exercise: Prove Theorem 13.10d).

Exercise: Show that if $f(x) = \sum_{n=0}^{\infty} a_n x^n$ has radius of convergence $R > 0$, then $F(x) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1}$ is an anti-derivative of f.

The following exercise drives home that uniform convergence of a sequence or series of functions on all of $\mathbb{R}$ is a very strong condition, often too much to hope for.

Exercise: Let $\sum_n a_n x^n$ be a power series with infinite radius of convergence, hence defining a function $f : \mathbb{R} \to \mathbb{R}$. Show that the following are equivalent:
(i) The series $\sum_n a_n x^n$ is uniformly convergent on $\mathbb{R}$.
(ii) We have $a_n = 0$ for all sufficiently large n.

Exercise: Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series with $a_n \geq 0$ for all n. Suppose that the radius of convergence is 1, so that f defines a function on $(-1, 1)$. Show that the following are equivalent:
(i) $f(1) = \sum_n a_n$ converges.
(ii) The power series converges uniformly on $[0, 1]$.
(iii) f is bounded on $[0, 1)$.

The fact that for any power series $f(x) = \sum_n a_n x^n$ with positive radius of convergence we have $a_n = \frac{f^{(n)}(0)}{n!}$ yields the following important result.

Corollary 13.11. *(Uniqueness Theorem) Let $f(x) = \sum_n a_n x^n$ and $g(x) = \sum_n b_n x^n$ be two power series with radii of convergence R_a and R_b with $0 < R_a \leq R_b$, so that both f and g are infinitely differentiable functions on $(-R_a, R_a)$. Suppose that for some δ with $0 < \delta \leq R_a$ we have $f(x) = g(x)$ for all $x \in (-\delta, \delta)$. Then $a_n = b_n$ for all n.*

Exercise: Suppose $f(x) = \sum_n a_n x^n$ and $g(x) = \sum_n b_n x^n$ are two power series each converging on some open interval $(-A, A)$. Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of elements of $(-A, A) \setminus \{0\}$ such that $\lim_{n\to\infty} x_n = 0$. Suppose that $f(x_n) = g(x_n)$ for all $n \in \mathbb{Z}^+$. Show that $a_n = b_n$ for all n.

The upshot of Corollary 13.11 is that the only way that two power series can be equal as functions – even in some very small interval around zero – is if all of their coefficients are equal. This is not obvious, since in general $\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} b_n$ does not imply $a_n = b_n$ for all n. Another way of saying this is that the only power series a function can be equal to on a small interval around zero is its Taylor series.

CHAPTER 14

Serial Miscellany

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

Let $p \in \mathbb{R}$. Recall that the *p-series* $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges iff $p > 1$. It is another matter entirely to determine the sum of the series exactly. In this section we devote our attention to the $p = 2$ case.

THEOREM 14.1. *(Euler)* $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

Euler's original argument is brilliant but not fully rigorous by modern standards. Since then several branches of mathematical analysis have been founded which give systematic tools for finding sums of this and similar series. In particular if one learns about **Fourier series** or **complex analysis** then very natural proofs can be given, but both of these topics are beyond the scope of an honors calculus course.

On the other hand, in the intervening centuries literally hundreds of proofs of Theorem 14.1 have been given, some of which use only tools we have developed (or indeed, no tools beyond standard freshman calculus). Among these we give here a particularly nice argument due to D. Daners [**Da12**] following Y. Matsuoka [**Ma61**]. In fact this argument barely uses notions from infinite series! Rather, it gives an upper bound on $\frac{\pi^2}{6} - \sum_{n=1}^{N} \frac{1}{n^2}$ in terms of N which approaches 0 as $N \to \infty$, and this certainly suffices. Precisely, we will show the following result.

THEOREM 14.2. *For all positive integers N,*

$$0 \leq \frac{\pi^2}{6} - \sum_{n=1}^{N} \frac{1}{n^2} \leq \frac{\pi^2}{4(N+1)}.$$

The proof will exploit a family of trigonometric integrals. For $n \in \mathbb{Z}^+$, we define

$$A_n = \int_0^{\frac{\pi}{2}} \cos^{2n} x dx, \ B_n = \int_0^{\frac{\pi}{2}} x^2 \cos^{2n} x dx.$$

Exercise: Show that $A_n, B_n > 0$ for all $n \geq 1$.

LEMMA 14.3. *a) For all positive integers n, we have:*

$$\int_0^{\frac{\pi}{2}} \sin^2 x \cos^{2(n-1)} x dx = \frac{A_n}{2n-1} = \frac{A_{n-1}}{2n}. \tag{94}$$

b) For all positive integers n, we have

$$A_n = (2n-1)nB_{n-1} - 2n^2 B_n. \tag{95}$$

PROOF. a) Integrating by parts gives

$$A_n = \int_0^{\frac{\pi}{2}} \cos x \cos^{2n-1} x dx$$

$$= (\sin x)(\cos^{2n-1} x) \,|_0^{\frac{\pi}{2}} - \int_0^{\frac{\pi}{2}} (\sin x)((2n-1)\cos^{2(n-1)} x)(-\sin x)dx$$

$$= (2n-1)\int_0^{\frac{\pi}{2}} \sin^2 x \cos^{2(n-1)} x dx = (2n-1)\int_0^{\frac{\pi}{2}} (1-\cos^2 x)\cos^{2(n-1)} x dx$$

$$= (2n-1)(A_{n-1} - A_n).$$

Thus

$$\int_0^{\frac{\pi}{2}} \sin^2 x \cos^{2(n-1)} x dx = \frac{A_n}{2n-1} = A_{n-1} - A_n,$$

and

$$\frac{A_{n-1}}{2n} = \frac{1}{2n}\left(A_n + \frac{A_n}{2n-1}\right) = \frac{1}{2n}\left(\frac{(2n-1)A_n + A_n}{2n-1}\right) = \frac{A_n}{2n-1}.$$

b) Integrating by parts twice gives[1]

$$A_n = \int_0^{\frac{\pi}{2}} 1 \cdot \cos^{2n} x dx = 2n\int_0^{\frac{\pi}{2}} x \sin x \cos^{2n-1} x dx$$

$$= n\int_0^{\frac{\pi}{2}} x^2 \left(\cos x \cos^{2n-1} x - (2n-1)\sin^2 x \cos^{2(n-1)} x\right) dx$$

$$= -nB_n + n(2n-1)\int_0^{\frac{\pi}{2}} x^2(1-\cos^2 x)\cos^{2(n-1)} x dx$$

$$= (2n-1)nB_{n-1} - 2n^2 B_n. \qquad \square$$

Dividing (95) by $n^2 A_n$ and using (94), we get

$$\frac{1}{n^2} = \frac{(2n-1)B_{n-1}}{nA_n} - \frac{2B_n}{A_n} = \frac{2B_{n-1}}{A_{n-1}} - \frac{2B_n}{A_n}.$$

Thus for all $n \in \mathbb{Z}^+$ we have

$$\sum_{n=1}^N \frac{1}{n^2} = \sum_{n=1}^N \left(\frac{2B_{n-1}}{A_{n-1}} - \frac{2B_n}{A_n}\right) = \frac{2B_0}{A_0} - \frac{2B_N}{A_N}.$$

Since

$$A_0 = \int_0^{\frac{\pi}{2}} dx = \frac{\pi}{2},\ B_0 = \int_0^{\frac{\pi}{2}} x^2 dx = \frac{\pi^3}{24},$$

we have

$$\frac{2B_0}{A_0} = \frac{\pi^2}{6}.$$

and thus for all $N \in \mathbb{Z}^+$,

$$\sum_{n=1}^N \frac{1}{n^2} = \frac{\pi^2}{6} - \frac{2B_N}{A_N}.$$

Equivalently

$$\frac{\pi^2}{6} - \sum_{n=1}^N \frac{1}{n^2} = \frac{2B_N}{A_N} > 0. \tag{96}$$

[1]This time we leave it to the reader to check that the boundary terms $uv\,|_0^{\frac{\pi}{2}}$ evaluate to 0.

LEMMA 14.4. *For all $x \in [0, \frac{\pi}{2}]$, we have:*

$$\left(\frac{2}{\pi}\right) x \leq \sin x$$

and thus

$$x^2 \leq \left(\frac{\pi}{2}\right)^2 \sin^2 x.$$

Exercise: Prove Lemma 14.4. (Hint: use convexity!)

Using Lemma 14.4 and Lemma 14.3a) with $N = n - 1$ we get

$$0 < \frac{\pi^2}{6} - \sum_{n=1}^{N} \frac{1}{n^2} = \frac{2B_N}{A_N} = \frac{2}{A_N} \int_0^{\frac{\pi}{2}} x^2 \cos^{2N} x dx$$

$$\leq \frac{2}{A_N} \left(\frac{\pi}{2}\right)^2 \int_0^{\frac{\pi}{2}} \sin^2 x \cos^{2N} x dx = \frac{2}{A_N} \left(\frac{\pi}{2}\right)^2 \frac{A_N}{2(N+1)} = \frac{\pi^2}{4(N+1)},$$

which proves Theorem 14.2.

2. Rearrangements and Unordered Summation

2.1. The Prospect of Rearrangement.

In this section we systematically investigate the validity of the "commutative law" for infinite sums. Namely, the definition we gave for convergence of an infinite series

$$a_1 + a_2 + \ldots + a_n + \ldots$$

in terms of the limit of the sequence of partial sums $A_n = a_1 + \ldots + a_n$ makes at least apparent use of the *ordering* of the terms of the series. Note that this is somewhat surprising even from the perspective of infinite sequences: the statement $a_n \to L$ can be expressed as: for all $\epsilon > 0$, there are only finitely many terms of the sequence lying outside the interval $(L - \epsilon, L + \epsilon)$, a description which makes clear that convergence to L will not be affected by any *reordering* of the terms of the sequence. However, if we reorder the terms $\{a_n\}$ of an infinite *series* $\sum_{n=1}^{\infty} a_n$, the corresponding change in the sequence A_n of partial sums is *not* simply a reordering, as one can see by looking at very simple examples. For instance, if we reorder

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots + \frac{1}{2^n} + \ldots$$

as

$$\frac{1}{4} + \frac{1}{2} + \frac{1}{8} + \ldots + \frac{1}{2^n} + \ldots$$

Then the first partial sum of the new series is $\frac{1}{4}$, whereas every nonzero partial sum of the original series is at least $\frac{1}{2}$.

Thus there is some evidence to fuel suspicion that reordering the terms of an infinite series may not be so innocuous an operation as for that of an infinite seuqence. All of this discussion is mainly justification for our setting up the "rearrangement problem" carefully, with a precision that might otherwise look merely pedantic.

Namely, the formal notion of rearrangement of a series $\sum_{n=0}^{\infty} a_n$ begins with a

permuation σ of $\mathbb{N}$, i.e., a bijective function $\sigma : \mathbb{N} \to \mathbb{N}$. We define the **rearrangement** of $\sum_{n=0}^{\infty} a_n$ by σ to be the series $\sum_{n=0}^{\infty} a_{\sigma(n)}$.

2.2. The Rearrangement Theorems of Weierstrass and Riemann.

The most basic questions on rearrangements of series are as follows.

QUESTION 14.5. *Let $\sum_{n=0}^{\infty} a_n = S$ is a convergent infinite series, and let σ be a permutation of $\mathbb{N}$. Then:*
a) Does the rearranged series $\sum_{n=0}^{\infty} a_{\sigma(n)}$ converge?
b) If it does converge, does it converge to S?

As usual, the special case in which all terms are non-negative is easiest, the case of absolute convergence is not much harder than that, and the case of nonabsolute convergence is where all the real richness and subtlety lies.

Indeed, suppose that $a_n \geq 0$ for all n. In this case the sum $A = \sum_{n=0}^{\infty} a_n \in [0, \infty$ is simply the supremum of the set $A_n = \sum_{n=0}^{k} a_k$ of finite sums. More generally, let $S = \{n_1, \ldots, n_k\}$ be any finite subset of the natural numbers, and put $A_S = a_{n_1} + \ldots + a_{n_k}$. Now every finite subset $S \subset \mathbb{N}$ is contained in $\{0, \ldots, N\}$ for some $N \in \mathbb{N}$, so for all S, $A_S \leq A_N$ for some (indeed, for all sufficiently large) N. This shows that if we define

$$A' = \sup_S A_S$$

as S ranges over all finite subsets of $\mathbb{N}$, then $A' \leq A$. On the other hand, for all $N \in \mathbb{N}$, $A_N = a_0 + \ldots + a_N = A_{\{0,\ldots,N\}}$: in other words, each partial sum A_N arises as A_S for a suitable finite subset S. Therefore $A \leq A'$ and thus $A = A'$.

The point here is that the description $\sum_{n=0}^{\infty} a_n = \sup_S A_S$ is manifestly unchanged by rearranging the terms of the series by any permutation σ: taking $S \mapsto \sigma(S)$ gives a bijection on the set of all finite subsets of $\mathbb{N}$, and thus

$$\sum_{n=0}^{\infty} a_n = \sup_S A_S = \sup_S A_{\sigma(S)} = \sum_{n=0}^{\infty} a_{\sigma(n)}.$$

The case of absolutely convergent series follows rather easily from this.

LEMMA 14.6. *Let $\sum_n a_n$ be a real series with $a_n \geq 0$ for all n and sum $A \in [0, \infty]$. Then A is the supremum of the set of all finite sums $A_S = \sum_{n \in S} a_n$ as S ranges over all nonempty finite subsets of $\mathbb{N}$.*

PROOF. Let $\mathcal{A}$ be the supremum of the finite sums A_S. For $N \in \mathbb{N}$, let $A_N = \sum_{n=0}^{N} a_n$. Since $a_n \geq 0$ for all n, the sequence A_N is increasing, so $A = \lim_{N \to \infty} A_N = \sup_N A_N$. Since $A_N = A_{\{0,\ldots,N\}}$, we have $A = \sup_N A_N \leq \sup_S A_S = \mathcal{A}$. On the other hand, for any nonempty finite subset S of $\mathbb{N}$, let N be the largest element of S. Then $S \subset \{0, \ldots, N\}$ so $A_S = \sum_{n \in S} a_n \leq \sum_{n=0}^{N} a_N = A_N \leq A$, so $\mathcal{A} = \sup A_S \leq A$. Thus $\mathcal{A} = A$. □

The point of Lemma 14.6 is that we have expressed the sum of a series with non-negative terms in a way which is manifestly independent of the ordering of the terms:[2] for any bijection σ of $\mathbb{N}$, as $S = \{n_1, \ldots, n_k\}$ ranges over all finite subsets

[2]This is a small preview of "unordered summation", the subject of the following section.

of $\mathbb{N}$, so does $\sigma(S) = \{\sigma(n_1), \ldots, \sigma(n_k)\}$. It follows that

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} a_{\sigma(n)} \in [0, \infty],$$

i.e., rearrangement of a series with non-negative terms does not disturb the convergence/divergence or the sum.

THEOREM 14.7. *(Weierstrass) Let $\sum_{n=0}^{\infty} a_n$ be an absolutely convergent series with sum A. Then for every permutation σ of $\mathbb{N}$, the rearranged series $\sum_{n=0}^{\infty} a_{\sigma(n)}$ converges to A.*

PROOF. Put $A = \sum_{n=0}^{\infty} a_n$. Fix $\epsilon > 0$ and let $N_0 \in \mathbb{N}$ be such that $\sum_{n=N_0}^{\infty} |a_n| < \epsilon$. Let $M_0 \in \mathbb{N}$ be sufficiently large so that the terms $a_{\sigma(0)}, \ldots, a_{\sigma(M_0)}$ include all the terms $a_0, \ldots, a_{N_0-1}$ (and possibly others). Then for all $M \geq M_0$,

$$|\sum_{n=0}^{M} a_{\sigma(n)} - A| = |\sum_{n=0}^{M} a_{\sigma(n)} - \sum_{n=0}^{\infty} a_n| \leq \sum_{n=N_0}^{\infty} |a_n| < \epsilon.$$

Indeed: by our choice of M we know that the terms $a_0, \ldots, a_{N_0-1}$ appear in both $\sum_{n=0}^{M} a_{\sigma(n)}$ and $\sum_{n=0}^{\infty} a_n$ and thus get cancelled; some further terms may or may not be cancelled, but by applying the triangle inequality and summing the absolute values we get an upper bound by assuming no further cancellation. This shows $\sum_{n=0}^{\infty} a_{\sigma(n)} = \lim_{M\to\infty} \sum_{n=0}^{M} a_{\sigma(n)} = A$. □

Exercise: a) Give a proof of Step 1 of Theorem 14.7 that bypasses Lemma 14.6. (Suggestion: by reasoning as in Step 2, argue that for each $\epsilon > 0$ and all sufficiently large N, $\sum_{n=N}^{\infty} |a_{\sigma(n)}| < \epsilon$.
b) Use the decomposition of $\sum_n a_n$ into its series of positive parts $\sum_n a_n^+$ and negative parts $\sum_n a_n^-$ to give a second proof of Step 2 of Theorem 14.7.

THEOREM 14.8. *(Riemann Rearrangement Theorem) Let $\sum_{n=0}^{\infty} a_n$ be a non-absolutely convergent series. For any $B \in [-\infty, \infty]$, there exists a permutation σ of $\mathbb{N}$ such that $\sum_{n=0}^{\infty} a_{\sigma(n)} = B$.*

PROOF.
Step 1: Since $\sum_n a_n$ is convergent, we have $a_n \to 0$ and thus that $\{a_n\}$ is bounded, so we may choose M such that $|a_n| \leq M$ for all n. We are not going to give an explicit "formula" for σ; rather, we are going to describe σ by a certain process. For this it is convenient to imagine that the sequence $\{a_n\}$ has been sifted into a disjoint union of two subsequences, one consisting of the positive terms and one consisting of the negative terms (we may assume without loss of generality that there $a_n \neq 0$ for all n). If we like, we may even imagine both of these subseqeunce ordered so that they are decreasing in absolute value. Thus we have two sequences

$$p_1 \geq p_2 \geq \ldots \geq p_n \geq \ldots \geq 0,$$

$$n_1 \leq n_2 \leq \ldots \leq n_n \leq \ldots \leq 0$$

so that together $\{p_n, n_n\}$ comprise the terms of the series. The key point here is Proposition 11.30 which tells us that since the convergence is nonabsolute, $\sum_n p_n = \infty$, $\sum_n n_n = -\infty$. So we may specify a rearangement as follows: we specify a choice of a certain number of positive terms – taken in decreasing order – and then a choice of a certain number of negative terms – taken in order of decreasing absolute value

– and then a certain number of positive terms, and so on. As long as we include a finite, positive number of terms at each step, then in the end we will have included every term p_n and n_n eventually, hence we will get a rearrangement.
Step 2 (diverging to ∞): to get a rearrangement diverging to ∞, we proceed as follows: we take positive terms $p_1, p_2, \ldots$ in order until we arrive at a partial sum which is at least $M+1$; then we take the first negative term n_1. Since $|n_1| \leq M$, the partial sum $p_1 + \ldots + p_{N_1} + n_1$ is still at least 1. Then we take at least one more positive term p_{N_1+1} and possibly further terms until we arrive at a partial sum which is at least $M+2$. Then we take one more negative term n_2, and note that the partial sum is still at least 2. And we continue in this manner: after the kth step we have used at least k positive terms, at least k negative terms, and *all* the partial sums from that point on will be at least k. Therefore every term gets included eventually and the sequence of partial sums diverges to $+\infty$.
Step 3 (diverging to $-\infty$): An easy adaptation of the argument of Step 2 leads to a permutation σ such that $\sum_{n=0}^{\infty} a_{\sigma(n)} = -\infty$. We leave this case to the reader.
Step 4 (converging to $B \in \mathbb{R}$): if anything, the argument is simpler in this case. We first take positive terms $p_1, \ldots, p_{N_1}$, stopping when the partial sum $p_1 + \ldots + p_{N_1}$ is greater than B. (To be sure, we take at least one positive term, even if $0 > B$.) Then we take negative terms $n_1, \ldots, n_{N_2}$, stopping when the partial sum $p_1 + \ldots + p_{N_1} + n_1 + \ldots + n_{N_2}$ is less than B. Then we repeat the process, taking enough positive terms to get a sum strictly larger than B then enough negative terms to get a sum strictly less than B, and so forth. Because both the positive and negative parts diverge, this construction can be completed. Because the general term $a_n \to 0$, a little thought shows that the absolute value of the difference between the partial sums of the series and B approaches zero. □

The conclusion of Theorem 14.8 holds under somewhat milder hypotheses.

Exercise: Let $\sum_n a_n$ be a real series such that $a_n \to 0$, $\sum_n a_n^+ = \infty$ and $\sum_n a_n^- = -\infty$. Show that the conclusion of Theorem 14.8 holds: for any $A \in [-\infty, \infty]$, there exists a permutation σ of $\mathbb{N}$ such that $\sum_{n=0}^{\infty} a_{\sigma(n)} = A$.

Exercise: Let $\sum_n a_n$ be a real series such that $\sum_n a_n^+ = \infty$.
a) Suppose that the sequence $\{a_n\}$ is bounded. Show that there exists a permutation σ of $\mathbb{N}$ such that $\sum_n a_{\sigma(n)} = \infty$.
b) Does the conclusion of part a) hold without the assumption that the sequnece of terms is bounded?

Theorem 14.8 exposes the dark side of nonabsolutely convergent series: just by changing the order of the terms, we can make the series diverge to $\pm\infty$ or converge to any given real number! Thus nonabsolute convergence is necessarily of a more delicate and less satisfactory nature than absolute convergence. With these issues in mind, we define a series $\sum_n a_n$ to be **unconditionally convergent** if it is convergent and every rearrangement converges to the same sum, and a series to be **conditionally convergent** if it is convergent but not unconditionally convergent. Then much of our last two theorems may be summarized as follows.

THEOREM 14.9. *(Main Rearrangement Theorem) A convergent real series is unconditionally convergent if and only if it is absolutely convergent.*

Many texts do not use the term "nonabsolutely convergent" and instead *define* a series to be conditionally convergent if it is convergent but not absolutely convergent. Aside from the fact that this terminology can be confusing to students to whom this rather intricate story of rearrangements has not been told, it seems correct to make a distinction between the following two *a priori* different phenomena:

- $\sum_n a_n$ converges but $\sum_n |a_n|$ does not, versus
- $\sum_n a_n$ converges to A but some rearrangement $\sum_n a_{\sigma(n)}$ does not.

As we have seen, these two phenomena are equivalent for real series. However the notion of an infinite series $\sum_n a_n$, absolute and unconditional convergence makes sense in other contexts, for instance[3] for series with values in an **infinite-dimensional Banach space** or with values in a **p-adic field**. In the former case it is a celebrated theorem of Dvoretzky-Rogers [**DR50**] that there exists a series which is unconditionally convergent but not absolutely convergent, whereas in the latter case one can show that *every* convergent series is unconditionally convergent whereas there exist nonabsolutely convergent series.

Exercise: Let $\sum_{n=0}^{\infty} a_n$ be any nonabsolutely convergent real series, and let $-\infty \leq a \leq A \leq \infty$. Show that there exists a permutation σ of $\mathbb{N}$ such that the set of partial limits of $\sum_{n=0}^{\infty} a_{\sigma(n)}$ is the closed interval $[a, A]$.

2.3. Unordered summation.

It is very surprising that the ordering of the terms of a nonabsolutely convergent series affects both its convergence and its sum – it seems fair to say that this phenomenon was undreamt of by the founders of the theory of infinite series.

Armed now, as we are, with the full understanding of the implications of our definition of $\sum_{n=0}^{\infty} a_n$ as the limit of a sequence of partial sums, it seems reasonable to ask: is there an alternative definition for the sum of an infinite series, one in which the ordering of the terms is *a priori* immaterial?

The answer to this question is **yes** and is given by the theory of **unordered summation**.

To be sure to get a definition of the sum of a series which does not depend on the ordering of the terms, it is helpful to work in a context in which no ordering is present. Namely, let S be a nonempty set, and define an **S-indexed sequence of real numbers** to be a function $a_\bullet : S \to \mathbb{R}$. The point here is that we recover the usual definition of a sequence by taking $S = \mathbb{N}$ (or $S = \mathbb{Z}^+$) but whereas $\mathbb{N}$ and $\mathbb{Z}^+$ come equipped with a natural ordering, the "naked set" S does not.

We wish to define $\sum_{s \in S} a_s$, i.e., the "unordered sum" of the numbers a_s as s ranges over all elements of S. Here it is: for every *finite* subset $T = \{s_1, \ldots, s_N\}$ of S, we define $a_T = \sum_{s \in T} a_s = a_{s_1} + \ldots + a_{s_N}$. (We also define $a_\varnothing = 0$.) Finally, for $A \in \mathbb{R}$, we say that the unordered sum $\sum_{s \in S} a_s$ **converges to A** if: for all

[3]Both of these are well beyond the scope of these notes, i.e., you are certainly not expected to know what I am talking about here.

$\epsilon > 0$, there exists a finite subset $T_0 \subset S$ such that for all finite subsets $T_0 \subset T \subset S$ we have $|a_T - A| < \epsilon$. If there exists $A \in \mathbb{R}$ such that $\sum_{s \in S} a_s = A$, we say that $\sum_{s \in S} a_s$ is **convergent** or that the S-indexed sequence $a_\bullet$ is **summable**. (When $S = \mathbb{Z}^{\geq N}$ we already have a notion of summability, so when we need to make the distinction we will say **unordered summable**.)

Notation: because we are often going to be considering various finite subsets T of a set S, we allow ourselves the following time-saving notation: for two sets A and B, we denote the fact that A is a finite subset of B by $A \subset_f B$.

Exercise: Suppose S is finite. Show that every S-indexed sequence $a_\bullet : S \to \mathbb{R}$ is summable, with sum $a_S = \sum_{s \in S} a_s$.

Exercise: If $S = \varnothing$, there is a unique function $a_\bullet : \varnothing \to \mathbb{R}$, the "empty function". Convince yourself that the most reasonable value to assign $\sum_{s \in \varnothing} a_s$ is 0.

Exercise: Give reasonable definitions for $\sum_{s \in S} a_s = \infty$ and $\sum_{s \in S} a_s = -\infty$.

Confusing Remark: We say we are doing "unordered summation", but our sequences take values in $\mathbb{R}$, in which the absolute value is derived from the order structure. One could also consider unordered summation of S-indexed sequences with values in an arbtirary **normed abelian group** $(G, |\ |)$.[4] A key feature of $\mathbb{R}$ is the positive-negative part decomposition, or equivalently the fact that for $M \in \mathbb{R}$, $|M| \geq A$ implies $M \geq A$ or $M \leq -A$. In other words, there are *exactly two ways* for a real number to be large in absolute value: it can be very positive or very negative. At a certain point in the theory considerations like this *must be used* in order to prove the desired results, but we delay such arguments for as long as possible.

The following result holds without using the positive-negative decomposition.

THEOREM 14.10. *(Cauchy Criteria for Unordered Summation) Let $a_\bullet : S \to \mathbb{R}$ be an S-index sequence, and consider the following assertions:*
(i) The S-indexed sequence $a_\bullet$ is summable.
(ii) For all $\epsilon > 0$, there exists a finite subset $T_\epsilon \subset S$ such that for all finite subsets T, T' of S containing T_ϵ,

$$|\sum_{s \in T} a_s - \sum_{s \in T'} a_s| < \epsilon.$$

(iii) For all $\epsilon > 0$, there exists $T_\epsilon \subset_f S$ such that: for all $T \subset_f S$ with $T \cap T_\epsilon = \varnothing$, we have $|a_T| = |\sum_{s \in T} a_s| < \epsilon$.
(iv) There exists $M \in \mathbb{R}$ such that for all $T \subset_f S$, $|a_T| \leq M$.

PROOF. (i) $\implies$ (ii) is immediate from the definition.
(ii) $\implies$ (i): We may choose, for each $n \in \mathbb{Z}^+$, a finite subset T_n of S such that $T_n \subset T_{n+1}$ for all n and such that for all finite subsets T, T' of S containing T_n, $|a_T - a_{T'}| < \frac{\epsilon}{2}$. It follows that the real sequence $\{a_{T_n}\}$ is Cauchy, hence convergent, say to A. We claim that $a_\bullet$ is summable to A: indeed, for $\epsilon > 0$, choose $n > \frac{2}{\epsilon}$.

[4] And it's not totally insane to do so: these arise in functional analysis and number theory.

Then, for any finite subset T containing T_n we have

$$|a_T - A| \leq |a_T - a_{T_n}| + |a_{T_n} - A| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

(ii) $\implies$ (iii): Fix $\epsilon > 0$, and choose $T_0 \subset_f$ as in the statement of (ii). Now let $T \subset_f S$ with $T \cap T_0 = \varnothing$, and put $T' = T \cup T_0$. Then T' is a finite subset of S containing T_0 and we may apply (ii):

$$|\sum_{s\in T} a_s| = |\sum_{s\in T'} a_s - \sum_{s\in T_0} a_s| < \epsilon.$$

(iii) $\implies$ (ii): Fix $\epsilon > 0$, and let $T_\epsilon \subset_f S$ be such that for all finite subsets T of S with $T \cap T_\epsilon = \varnothing$, $|a_T| < \frac{\epsilon}{2}$. Then, for any finite subset T' of S containing T_ϵ,

$$|a_{T'} - a_{T_\epsilon}| = |a_{T'\setminus T_\epsilon}| < \frac{\epsilon}{2}.$$

From this and the triangle inequality it follows that if T and T' are two finite subsets containing T_ϵ,

$$|a_T - a_{T'}| < \epsilon.$$

(iii) $\implies$ (iv): Using (iii), choose $T_1 \subset_f S$ such that for all $T' \subset_f S$ with $T_1 \cap T' = \varnothing$, $|a_{T'}| \leq 1$. Then for any $T \subset_f S$, write $T = (T \setminus T_1) \cup (T \cap T_1)$, so

$$|a_T| \leq |\sum_{s\in T\setminus T_1} a_s| + |\sum_{s\in T\cap T_1} a_s| \leq 1 + \sum_{s\in T_1} |a_s|,$$

so we may take $M = 1 + \sum_{s\in T_1} |a_s|$. □

Confusing Example: Let $G = \mathbb{Z}_p$ with its standard norm. Define $a_n : \mathbb{Z}^+ \to G$ by $a_n = 1$ for all n. Because of the non-Archimedean nature of the norm, we have for any $T \subset_f S$ $|a_T| = |\#T| \leq 1$. Therefore $a_\bullet$ satisfies condition (iv) of Theorem 14.10 above but not condition (iii): given any finite subset $T \subset \mathbb{Z}^+$, there exists a finite subset T', disjoint from T, such that $|a_{T'}| = 1$: indeed, we may take $T' = \{n\}$, where n is larger than any element of T and prime to p.

Although we have no reasonable expectation that the reader will be able to make any sense of the previous example, we offer it as motivation for delaying the proof of the implication (iv) $\implies$ (i) above, which uses the positive-negative decomposition in $\mathbb{R}$ in an essential way.

THEOREM 14.11. *An S-indexed sequence $a_\bullet : S \to \mathbb{R}$ is summable iff the finite sums ae uniformly bounded: i.e., there exists $M \in \mathbb{R}$ such that for all $T \subset_f S$, $|a_T| \leq M$.*

PROOF. In Theorem 14.10 above we showed that if $a_\bullet$ is summable, the finite sums are uniformly bounded. Now suppose $a_\bullet$ is not summable, so by Theorem 14.10 there is $\epsilon > 0$ with the following property: for any $T \subset_f S$ there is $T' \subset_f S$ with $T \cap T' = \varnothing$ and $|a_{T'}| \geq \epsilon$. Of course, if we can find such a T', we can also find a T'' disjoint from $T \cup T'$ with $|a_{T''}| \geq \epsilon$, and so forth: there will be a sequence $\{T_n\}_{n=1}^\infty$ of pairwise disjoint finite subsets of S such that for all n, $|a_{T_n}| \geq \epsilon$. But now decompose $T_n = T_n^+ \cup T_n-$, where T_n^+ consists of the elements s such that $a_s \geq 0$ and T_n^- consists of the elements s such that $a_s < 0$. It follows that

$$a_{T_n} = |a_{T_n^+}| - |a_{T_n^-}|$$

hence

$$\epsilon \leq |a_{T_n}| \leq |a_{T_n^+}| + |a_{T_n^-}|,$$

from which it follows that $\max |a_{T_n^+}, a_{T_n -}| \geq \frac{\epsilon}{2}$, so we may define for all n a subset $T_n' \subset T_n$ such that $|a_{T_n'}| \geq \frac{\epsilon}{2}$ and the sum $a_{T_n'}$ consists either entirely of non-negative elements of entirely of negative elements. If we now consider $T_1', \ldots, T_{2n-1}'$, then by the Pigeonhole Principle there must exist $1 \leq i_1 < \ldots < i_n \leq 2n-1$ such that all the terms in each T_i' are non-negative *or* all the terms in each T_i' are negative. Let $\mathcal{T}_n = \bigcup_{j=1}^n T_{i_j}'$. Then we have a disjoint union and no cancellation, so $|\mathcal{T}_n| \geq \frac{n\epsilon}{2}$: the finite sums a_T are not uniformly bounded. □

PROPOSITION 14.12. *Let $a_\bullet : S \to \mathbb{R}$ be an S-indexed sequence with $a_s \geq 0$ for all $s \in S$. Then*

$$\sum_{s \in S} a_s = \sup_{T \subset_f S} a_T.$$

PROOF. Let $A = \sup_{T \subset_f S} a_T$.

We first suppose that $A < \infty$. By definition of the supremum,, for any $\epsilon > 0$, there exists a finite subset $T \subset S$ such that $A - \epsilon < a_T \leq A$. Moreover, for any finite subset $T' \supset T$, we have $A - \epsilon a_T \leq a_{T'} \leq A$, so $a_\bullet \to A$.

Next suppose $A = \infty$. We must show that for any $M \in \mathbb{R}$, there exists a subset $T_M \subset_f S$ such that for every finite subset $T \supset T_M$, $a_T \geq M$. But the assumption $A = \infty$ implies there exists $T \subset_f S$ such that $a_T \geq M$, and then non-negativity gives $a_{T'} \geq a_T \geq M$ for all finite subsets $T' \supset T$. □

THEOREM 14.13. *(Absolute Nature of Unordered Summation) Let S be any set and $a_\bullet : S \to \mathbb{R}$ be an S-indexed sequence. Let $|a_\bullet|$ be the S-indexed sequence $s \mapsto |a_s|$. Then $a_\bullet$ is summable iff $|a_\bullet|$ is summable.*

PROOF. Suppose $|a_\bullet|$ is summable. Then for any $\epsilon > 0$, there exists T_ϵ such that for all finite subsets T of S disjoint from T_ϵ, we have $||a|_T| < \epsilon$, and thus

$$|a_T| = |\sum_{s \in T} a_s| \leq |\sum_{s \in T} |a_s|| = ||a|_T| < \epsilon.$$

Suppose $|a_\bullet|$ is *not* summable. Then by Proposition 14.12, for every $M > 0$, there exists $T \subset_f S$ such that $|a|_T \geq 2M$. But as in the proof of Theorem 14.11, there must exist a subset $T' \subset T$ such that (i) $a_{T'}$ consists entirely of non-negative terms or entirely of negative terms and (ii) $|a_{T'}| \geq M$. Thus the partial sums of $a_\bullet$ are not uniformly bounded, and by Theorem 14.11 $a_\bullet$ is not summable. □

THEOREM 14.14. *For $a_\bullet : \mathbb{N} \to \mathbb{R}$ an ordinary sequence and $A \in \mathbb{R}$, TFAE:*
(i) The unordered sum $\sum_{n \in \mathbb{Z}^+} a_n$ is convergent, with sum A.
(ii) The series $\sum_{n=0}^\infty a_n$ is unconditionally convergent, with sum A.

PROOF. (i) $\implies$ (ii): Fix $\epsilon > 0$. Then there exists $T_\epsilon \subset_f S$ such that for every finite subset T of $\mathbb{N}$ containing T_ϵ we have $|a_T - A| < \epsilon$. Put $N = \max_{n \in T_\epsilon} n$. Then for all $n \geq N$, $\{0, \ldots, n\} \supset T_\epsilon$ so $|\sum_{k=0}^n a_k - A| < \epsilon$. It follows that the infinite series $\sum_{n=0}^\infty a_n$ converges to A in the usual sense. Now for any permutation σ of $\mathbb{N}$, the unordered sum $\sum_{n \in \mathbb{Z}^+} a_{\sigma(n)}$ is manifestly the same as the unordered sum $\sum_{n \in \mathbb{Z}^+} a_n$, so the rearranged series $\sum_{n=0}^\infty a_{\sigma(n)}$ also converges to A.
(ii) $\implies$ (i): We will prove the contrapositive: suppose the unordered sum $\sum_{n \in \mathbb{N}} a_n$ is divergent. Then by Theorem 14.11 for every $M \geq 0$, there exists $T \subset S$ with $|a_T = \sum_{s \in T} a_s| \geq M$. Indeed, as the proof of that result shows, we can choose T to be disjoint from any given finite subset. We leave it to you to check that we can therefore build a rearrangement of the series with unbounded partial sums. □

Exercise: Fill in the missing details of (ii) $\implies$ (i) in the proof of Theorem 14.14.

Exercise*: Can one prove Theorem 14.14 without appealing to the fact that $|x| \geq M$ implies $x \geq M$ or $x \leq -M$? For instance, does Theorem 14.14 holds for S-indexed sequences with values in any Banach space? Any complete normed abelian group?

Comparing Theorems 14.12 and 14.13 we get a second proof of the portion of the Main Rearrangement Theorem that says that a real series is unconditionally convergent iff it is absolutely convergent. Recall that our first proof of this depended on the Riemann Rearrangement Theorem, a more complicated result.

On the other hand, if we allow ourselves to use the previously derived result that unconditional convergence and absolute convergence coincide, then we can get an easier proof of (ii) $\implies$ (i): if the series $\sum_n a_n$ is unconditionally convergent, then $\sum_n |a_n| < \infty$, so by Proposition 14.10 the unordered sequence $|a_\bullet|$ is summable, hence by Theorem 14.12 the unordered sequence $a_\bullet$ is summable.

To sum up (!), when we apply the very general definition of unordered summability to the classical case of $S = \mathbb{N}$, we recover precisely the theory of absolute (= unconditional) convergence. This gives us a clearer perspective on exactly what the usual, order-dependent notion of convergence is buying us: namely, the theory of conditionally convergent series. It may perhaps be disappointing that such an elegant theory did not gain us anything new.

However when we try to generalize the notion of an infinite series in various ways, the results on unordered summability become very helpful. For instance, often in nature one encounters **biseries**

$$\sum_{n=-\infty}^{\infty} a_n$$

and **double series**

$$\sum_{m,n \in \mathbb{N}} a_{m,n}.$$

We may treat the first case as the unordered sum associated to the $\mathbb{Z}$-indexed sequence $n \mapsto a_n$ and the second as the unordered sum associated to the $\mathbb{N} \times \mathbb{N}$-indexed sequence $(m,n) \mapsto a_{m,n}$ and we are done: there is no need to set up separate theories of convergence. Or, if we prefer, we may shoehorn these more ambitiously indexed series into conventional $\mathbb{N}$-indexed series: this involves choosing a bijection b from $\mathbb{Z}$ (respectively $\mathbb{N} \times \mathbb{N}$) to $\mathbb{N}$. In both cases such bijections exist, in fact in great multitude: if S is any countably infinite set, then for any two bijections $b_1,\ b_2 : S \to \mathbb{N}$, $b_2 \circ b_1^{-1} : \mathbb{N} \to \mathbb{N}$ is a permutation of $\mathbb{N}$. Thus the discrepancy between two chosen bijections corresponds precisely to a rearrangement of the series. By Theorem 14.13, if the unordered sequence is summable, then the choice of bijection b is immaterial, as we are getting an unconditionally convergent series.

The theory of products of infinite series comes out especially cleanly in this unordered setting (which is not surprising, since it corresponds to the case of absolute convergence, where Cauchy products are easy to deal with).

Exercise: Let S_1 and S_2 be two sets, and let $a_\bullet : S_1 \to \mathbb{R}$, $b_\bullet : S_2 \to \mathbb{R}$. We

assume the following nontriviality condition: there exists $s_1 \in S_1$ and $s_2 \in S_2$ such that $a_{s_1} \neq 0$ and $a_{s_2} \neq 0$. We define $(a,b)_\bullet : S_1 \times S_2 \to \mathbb{R}$ by

$$(a,b)_s = (a,b)_{(s_1,s_2)} = a_{s_1} b_{s_2}.$$

a) Show that $a_\bullet$ and $b_\bullet$ are both summable iff $(a,b)_\bullet$ is summable.
b) Assuming the equivalent conditions of part a) hold, show

$$\sum_{s \in S_1 \times S_2} (a,b)_s = \left(\sum_{s_1 \in S_1} a_{s_1} \right) \left(\sum_{s_2 \in S_2} b_{s_2} \right).$$

c) When $S_1 = S_2 = \mathbb{N}$, compare this result with the theory of Cauchy products we have already developed.

Exercise: Let S be an uncountable set,[5] and let $a_\bullet : S \to \mathbb{R}$ be an S-indexed sequence. Show that if $a_\bullet$ is summable, then $\{s \in S \mid a_s \neq 0\}$ is countable.

3. Abel's Theorem

3.1. Statement and Proof.

THEOREM 14.15. *(Abel's Theorem)*
Let $\sum_{n=0}^\infty a_n$ be a convergent series.
a) The series $\sum_{n=0}^\infty a_n x^n$ is uniformly convergent on $[0,1]$.
b) We have $\lim_{x \to 1^-} \sum_{n=0}^\infty a_n x^n = \sum_{n=0}^\infty a_n$.

PROOF. a) ([**C**, p. 47]) Since $\sum_{n=0}^\infty a_n 1^n = \sum_{n=0}^\infty a_n$, convergence at $x = 1$ is our hypothesis. By our work on power series – specifically Lemma 11.32 – convergence at 1 implies convergence on $(-1,1)$, and thus $\sum_{n=0}^\infty a_n x^n$ converges pointwise on $[0,1]$. Since we have convergence at $x = 1$, it suffices to show uniform convergence on $[0,1)$. Fix $\epsilon > 0$; because $\sum_n a_n$ converges, there is $N \in \mathbb{Z}^+$ such that $|\sum_{n=N}^\infty a_n| < \epsilon$. Now we apply **Abel's Lemma** (Proposition 10.2) with "a_n sequence" $a_N, a_{N+1}, \ldots$, with "b_n sequence" $x^N, x^{N+1}, \ldots$ (note this is positive and decreasing) and with $M = |\sum_{n=N}^\infty a_n|$. The conclusion is that for any $k \in \mathbb{N}$,

$$|\sum_{n=N}^{N+k} a_n x^n| \leq x^N M \leq x^N \epsilon < \epsilon.$$

By the Cauchy Criterion (Lemma 13.1), $\sum_{n=0}^N a_n x^n \stackrel{u}{\to} f$ on $[0,1)$.
b) By part a), $\sum_{n=0}^\infty a_n x^n$ is a uniform limit of continuous functions (indeed, of polynomials) on $[0,1]$, so by Theorem 13.3, it is continuous on $[0,1]$. In particular $\sum_{n=0}^\infty a_n x^n$ is continuous at $x = 1$: $\lim_{x \to 1^-} \sum_{n=0}^\infty a_n x^n = \sum_{n=0}^\infty a_n$. □

Remark: Usually "Abel's Theorem" means part b) of the above result: $\sum_{n=0}^\infty a_n = \lim_{x \to 1^-} a_n x^n$. But this is an immediate consequence of the uniformity of the convergence on $[0,1]$, so having this statement be part of Abel's Theorem gives a stronger and also more conceptually transparent result.

[5]Here we are following our usual convention of allowing individual exercises to assume knowledge that we do not want to assume in the text itself. Needless to say, there is no need to attempt this exercise if you do not already know and care about uncountable sts.

Above we followed an exercise in the text [**C**] of Cartan.[6] For comparison, here is a different proof of Theorem 14.15b) from a famous text of Rudin.

PROOF. [**R**, Thm. 8.2] Since $\sum_n a_n$ converges, $\{a_n\}$ is bounded, so by Corollary 11.35 the radius of convergence of $f(x) = \sum_n a_n x^n$ is at least 1. Put $A_{-1} = 0$; for $n \geq 0$, put $A_n = a_0 + \ldots + a_n$; and put $A = \lim_{n\to\infty} A_n = \sum_{n=0}^{\infty} a_n$. Then

$$\sum_{n=0}^{N} a_n x^n = \sum_{n=0}^{N} (A_n - A_{n-1}) x^n = (1-x) \sum_{n=0}^{N-1} A_n x^n + A_N x^N.$$

For each fixed $x \in [0,1)$, we let $N \to \infty$ to get

$$f(x) = \sum_{n=0}^{\infty} a_n x^n = (1-x) \sum_{n=0}^{\infty} A_n x^n.$$

Now fix $\epsilon > 0$, and choose N such that $n \geq N$ implies $|A - A_n| < \epsilon$. Then, since

$$(1-x) \sum_{n=0}^{\infty} x^n = 1 \tag{97}$$

for all $x \in [0,1)$, we get

$$|f(x) - A| = |(1-x) \sum_{n=0}^{\infty} A_n x^n - A(1-x) \sum_{n=0}^{\infty} x^n| = |(1-x) \sum_{n=0}^{\infty} (A_n - A) x^n|$$

$$\leq (1-x) \sum_{n=0}^{N} |A_n - A| x^n + \left(\frac{\epsilon}{2}\right)(1-x) \sum_{n=N+1}^{\infty} x^n \leq (1-x) \sum_{n=0}^{N} |A_n - A| x^n + \epsilon.$$

The last quantity above approaches ϵ as x approaches 1 from the left. Since ϵ was arbitrary, this shows $\lim_{x\to 1^-} f(x) = A$. □

As you can see, Rudin's proof uses much less than Cartan's: rather than relying on Abel's Lemma, a bit of partial summation is done on the fly. Moreover, most of the appeals to the theory of power series and uniform convergence are replaced by a clever introduction of the geometric series! Nevertheless I must say that, although Rudin's argument is easy enough to follow line by line, in terms of "what's going on in the proof" I find it absolutely impenetrable.

The rest of this section is an extended exercise in "Abel's Theorem appreciation". First of all, it may help to restate the result in a form which is slightly more general and moreover makes more clear exactly what has been established.

THEOREM 14.16. *(Abel's Theorem Mark II) Let $f(x) = \sum_n a_n (x-c)^n$ be a power series with radius of convergence $R > 0$, hence convergent at least for all $x \in (c-R, c+R)$.*
a) Suppose that the power series converges at $x = c+R$. Then the function $f : (c-R, c+R] \to \mathbb{R}$ is continuous at $x = c+R$: $\lim_{x\to(c+R)^-} f(x) = f(c+R)$.
b) Suppose that the power series converges at $x = c-R$. Then the function $f : [c-R, c+R) \to \mathbb{R}$ is continuous at $x = c-R$: $\lim_{x\to(c-R)^+} f(x) = f(c-R)$.

[6]Henri Cartan (1904-2008) was one of the leading mathematicians of the 20th century.

Exercise: Prove Theorem 14.16.

Exercise: Consider $f(x) = \frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$, which converges for all $x \in (-1,1)$. Show $\lim_{x \to -1^+} f(x)$ exists, so f extends to a continuous function on $[-1,1)$. Nevertheless $f(-1) \neq \lim_{x \to -1^+} f(x)$. Why doesn't this contradict Abel's Theorem?

3.2. An Application to the Cauchy Product.

As our first application, we round out our treatment of Cauchy products by showing that the Cauchy product never "wrongly converges".

THEOREM 14.17. *Let $\sum_{n=0}^{\infty} a_n$ be a series converging to A and $\sum_{n=0}^{\infty} b_n$ be a series converging to B. As usual, we define $c_n = \sum_{k=0}^{n} a_k b_{n-k}$ and the Cauchy product series $\sum_{n=0}^{\infty} c_n$. Suppose that $\sum_{n=0}^{\infty} c_n$ converges to C. Then $C = AB$.*

PROOF. Put $f(x) = \sum_{n=0}^{\infty} a_n x^n$, $g(x) = \sum_{n=0}^{\infty} b_n x^n$ and $h(x) = \sum_{n=0}^{\infty} c_n x^n$. By assumption, $f(x)$, $g(x)$ and $h(x)$ all converge at $x = 1$, so by Lemma 11.32 the radii of convergence of $\sum_n a_n x^n$, $\sum_n b_n x^n$ and $\sum_n c_n x^n$ are all at least one. Now all we need to do is apply Abel's Theorem:

$$C = h(1) \stackrel{\text{AT}}{=} \lim_{x \to 1^-} h(x) = \lim_{x \to 1^-} f(x)g(x)$$

$$= \left(\lim_{x \to 1^-} f(x) \right) \left(\lim_{x \to 1^-} g(x) \right) \stackrel{\text{AT}}{=} f(1)g(1) = AB. \qquad \square$$

3.3. Two Identities Justified By Abel's Theorem.

Here are two surprising and beautiful identities.

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots = \log 2. \tag{98}$$

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots = \frac{\pi}{4}. \tag{99}$$

The identity (99) was "known" to Leibniz (as would be logical, given that the convergence of both series follows from Leibniz's Alternating Series Test), where the quotation marks are meant to suggest that Leibniz probably did not have an argument we would accept as a rigorous proof. Suppose someone shows you such identities, as I now have: what would you make of them?

A good first reaction is to attempt numerical verification. Even this is not as easy as one might expect, because the convergence of both series is rather slow. (In particular, among all ways one might try to numerically compute π, (99) is one of the *worst* I know.) Like any series which is shown to be convergent by the Alternating Series Test, there is a built in error estimate for the sum: if we cut off after N terms the error is in absolute value at most $|a_{N+1}|$. The problem with this is that the Nth terms of these series tend to zero quite slowly! So for instance,

$$S_{10^4} = \sum_{n=1}^{10^4} \frac{(-1)^{n+1}}{n} = 0.693097183059945296917232371 4\ldots$$

(Using a software package I asked for the *exact sum* of the series, which is of course a rational number, but a very complicated one: it occupies more than one full screen on my computer. The amount of time spent to compute this rational number was small but not instantaneous. In fact the reason I chose 10^4 is that the software managed this but had trouble with S_{10^5}. Then I converted the fraction to a decimal. Of course a much better way to do this would be to convert the fractionals to decimals as we go along, but this needs to be done carefully to prevent rounding errors: to be serious about this sort of thing one needs to know some **numerical analysis**.)

By the Alternating Series Test, the difference between the infinite sum and the finite sum S_{10^4} is at most $\frac{1}{1001}$, so we are guaranteed (roughly) four decimal places of accuracy. For comparison, computing $\log 2$ – e.g. by writing it as $-\log(\frac{1}{2})$ and using the Taylor series for $\log(1+x)$ – we get

$$\log 2 = 0.693147180559945309417232121 4\ldots.$$

So indeed the identity (98) holds true at least up to four decimal places. If we wanted to do much more numerical verification than this, we would probably have to do something a little more clever.

Similarly, we have

$$\sum_{n=0}^{10^4} \frac{(-1)^n}{2n+1} = 0.78542316\ldots,$$

and by the Alternating Series test this approximates the infinite sum $\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$ to at least four decimal places of accuracy, whereas

$$\frac{\pi}{4} = 0.78539816339744830961566084 58\ldots,$$

which shows that (99) holds true at least up to four decimal places.

Okay, so these identities are probably true: how do we prove them?!?

We exploit our knowledge of power series. First, take $f(x) = \log(1+x)$. Then

$$f'(x) = \frac{1}{1+x} = \frac{1}{1-(-x)} = \sum_{n=0}^{\infty} (-x)^n = \sum_{n=0}^{\infty} (-1)^n x^n. \tag{100}$$

Integrating termwise gives

$$\log(1+x) = f(x) = f(0) + \sum_{n=1}^{\infty} \frac{(-1)^n x^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{(-1)^n x^{n+1}}{n+1}. \tag{101}$$

(This was "known" to Newton and Leibniz.) So – aha! – we plug in $x = 1$ to get

$$\log(2) = \log(1+1) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n+1}.$$

But not so fast. Although you will find this explanation in many freshman calculus books, it is not yet justified. We were being sloppy in our above work by writing down power series expansions and not keeping track of the interval of convergence. The identity (100) holds for $x \in (-1,1)$, and that's really the best we can do, since the power series on the right hand side does not converge for any other values of x.

Integrating this term by term, we find that (101) holds for $x \in (-1,1)$. Above, in our excitement, we plugged in $x = 1$: so close, but out of bounds. Too bad!

But don't despair: it's Abel's Theorem for the win! Indeed, because the series converges at $x = 1$ and the function $\log x$ is defined and continuous at $x = 1$,

$$\log(2) = \log(1+1) = \lim_{x \to 1^-} \log x = \lim_{x \to 1^-} \sum_{n=1}^{\infty} \frac{(-1)^n x^{n+1}}{n+1} \stackrel{\text{AT}}{=} \sum_{n=1}^{\infty} \frac{(-1)^n}{n+1}.$$

Exercise: Establish (99) similarly, starting with the function $f(x) = \arctan x$.

3.4. Abel Summability.

Abel's Theorem gives rise to a **summability method**: a way to extract numerical values out of certain divergent series $\sum_n a_n$ "as though they converged". Instead of forming the sequence of partial sums $A_n = a_0 + \ldots + a_n$ and taking the limit, suppose instead we look at $\lim_{x \to 1^-} \sum_{n=0}^{\infty} a_x x^n$. We say the series $\sum_n a_n$ is **Abel summable** if this limit exists, in which case we write it as $\mathcal{A} \sum_{n=0}^{\infty} a_n$, the **Abel sum** of the series. The point of this is that by Abel's theorem, if a series $\sum_{n=0}^{\infty} a_n$ is actually convergent, say to A, then it is also Abel summable and its Abel sum is also equal to A. However, there are series which are divergent yet Abel summable.

EXAMPLE 14.1. *Consider the series $\sum_{n=0}^{\infty} (-1)^n$. As we saw, the partial sums alternate between 0 and 1 so the series does not diverge. We mentioned earlier that (the great) L. Euler believed that nevertheless the right number to attach to the series $\sum_{n=0}^{\infty} (-1)^n$ is $\frac{1}{2}$. Since the two partial limits of the sequence of partial sums are 0 and 1, it seems vaguely plausible to split the difference.*

Abel's Theorem provides a much more convincing argument. The power series $\sum_n (-1)^n x^n$ converges for all x with $|x| < 1$, and moreover for all such x we have

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

and thus

$$\lim_{x \to 1^-} \sum_{n=0}^{\infty} (-1)^n x^n = \lim_{x \to 1^-} \frac{1}{1+x} = \frac{1}{2}.$$

That is, the series $\sum_n (-1)^n$ is divergent but Abel summable, with Abel sum $\frac{1}{2}$.

So Euler's idea was better than we gave him credit for.

Exercise: Suppose that $\sum_{n=0}^{\infty} a_n$ is a series with $a_n \geq 0$ for all n. Show the converse of Abel's Theorem: if $\lim_{x \to 1^-} \sum_{n=0}^{\infty} a_n x^n = L$, then $\sum_{n=0}^{\infty} a_n = L$.

4. The Peano-Borel Theorem

4.1. Statement.

The Taylor series of a smooth (i.e., infinitely differentiable) function is surely one of the most useful concepts of basic calculus. It is only reasonable to concentrate first on "the good news" here: most naturally occurring functions, including all of the elementary functions one meets in precalculus mathematics, are **analytic**: at every point c in their domain I there is a $\delta_c > 0$ such that the Taylor series $T_{f,c}$

converges to f on $(c - \delta_c, c + \delta_c)$.

This good news builds up a strong intuition that smooth functions should have well-behaved Taylor series. In fact this is far from the case. To fix ideas, consider smooth functions $f : \mathbb{R} \to \mathbb{R}$. Then there are two ways for f to fail to be analytic:

(i) At some $c \in \mathbb{R}$, $T_{f,c}$ converges only at 0.
(ii) At some $c \in \mathbb{R}$, $T_{f,c}$ has positive radius of convergence, but there is no interval $(c - \delta_c, c + \delta_c)$ about c on which $T_{f,c} = f$.

Both of these pathologies occur. In the second case, an example was constructed by Cauchy in 1823 and remains well known to this day.

PROPOSITION 14.18. *Let* $f(x) = \begin{cases} e^{-x^2} & x \neq 0 \\ 0 & x = 0 \end{cases}$. *Then:*

a) $f : \mathbb{R} \to \mathbb{R}$ *is smooth.*
b) For all $n \in \mathbb{N}$, $f^{(n)}(0) = 0$.
c) Thus the Taylor series expansion of f *at* 0 *is the identically zero function. It has infinite radius of convergence and is equal to* f *only at the central point* $x = 0$.

Exercise: Prove Proposition 14.18.

That a smooth function can have a Taylor series which converges only at the central point – which we may certainly assume to be 0 – is more subtle. The first example of such a function was given by du Bois-Reymond in 1876 [**dBR76**], [**dBR83**].

It is no problem to construct a *power series* $\sum a_n x^n$ which converges only at 0. By the Cauchy-Hadamard formula we need only select a sequence with $\limsup |a_n|^{\frac{1}{n}} = \infty$, which as we have seen is implied by $\rho = \lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \infty$, so e.g. $\sum n! x^n$ works. But is this power series a Taylor series? It is if and only if we can find a smooth function $f : \mathbb{R} \to \mathbb{R}$ with $\frac{f^{(n)}(0)}{n!} = n!$, i.e., that

$$\forall n \in \mathbb{N},\ f^{(n)}(0) = (n!)^2.$$

Thus the essential problem is to construct a smooth function $f : \mathbb{R} \to \mathbb{R}$ whose sequence of derivatives at 0 grows (at least along a subsequence) sufficiently rapidly, and in particular much more rapidly than $n!$.

It is not so auspicious to "search in nature" for a function with such rapidly growing derivatives. It turns out to be simpler solve a more general problem: show that any real sequence whatsoever can serve as the sequence of derivatives at 0 of a smooth function $f : \mathbb{R} \to \mathbb{R}$. This is the content of the following remarkable result.

THEOREM 14.19. *(Peano-Borel) Let* $\{c_n\}_{n=0}^{\infty}$ *be a real sequence. Then there is an infinitely differentiable function* $f : \mathbb{R} \to \mathbb{R}$ *such that for all* $n \in \mathbb{N}$ *we have* $f^{(n)}(0) = c_n$.

Exercise: Show that the Peano-Borel Theorem is equivalent to the statement that every power series is a Taylor series. More precisely, for any $c \in \mathbb{R}$ and any real sequence $\{a_n\}_{n=0}^{\infty}$, show there is an infinitely differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that the Taylor series expansion of f at c is $\sum_{n=0}^{\infty} a_n (x - c)^n$.

Exercise: Prove du Bois-Reymond's Theorem that there is an infinitely differentiable function $f : \mathbb{R} \to \mathbb{R}$ whose Taylor series expansion at 0 converges only at $x = 0$.

Exercise: Let $\{a_n\}_{n=0}^{\infty}$ be a real sequence.
a) Show that

$$\mathcal{S} = \{\text{infinitely differentiable } f : \mathbb{R} \to \mathbb{R} \mid \text{the Taylor series of } f \text{ at } 0 \text{ is } \sum_{n=0}^{\infty} a_n x^n\}$$

is infinite.
b) (This part is for those who know some linear algebra.) Let $f_0 \in \mathcal{S}$. Show: $\{f - f_0 \mid f \in \mathcal{S}\}$ is a real vector space of infinite dimension.

Theorem 14.19 is generally known as "Borel's Theorem". It was proved – along with the Heine-Borel Theorem – in the doctoral thesis of É. Borel [**Bo95**]. But according to a recent article of Á. Besenyei [**Be14**] the result was in fact first proved by G. Peano in 1884, so we speak of the Peano-Borel Theorem.

In fact Besenyei's article is the source of this entire section, both the historical material and the account of Peano's proof that we will now give.

4.2. Proof.

The following argument uses a bit of complex numbers, which are presented in more detail in a later chapter.

Step 1: Let $\{a_n\}_{n=0}^{\infty}$, $\{b_n\}_{n=0}^{\infty}$ be real sequences with $b_n \geq 0$ for all $n \in \mathbb{N}$. Put

$$f(x) = \sum_{k=0}^{\infty} \frac{a_k x^k}{1 + b_k x^2}.$$

Suppose f converges on all of $\mathbb{R}$ and that for all $n \in \mathbb{N}$ and $x \in \mathbb{R}$ we have

$$f^{(n)}(x) = \sum_{k=0}^{\infty} \left(\frac{a_k x^k}{1 + b_k x^2} \right)^{(n)}.$$

Later we will show how this can be achieved with suitable choices of the sequences $\{a_n\}$ and $\{b_n\}$. Now we have

$$|b_k x^2| < 1 \implies \frac{a_k x^k}{1 + b_k x^2} = a_k x^k \sum_{j=0}^{\infty} (-1)^j b_k^j x^{2j} = \sum_{j=0}^{\infty} (-1)^j a_k b_k^j x^{2j+k}.$$

Thus

$$\left(\frac{a_k x^k}{1 + b_k x^2} \right)^{(n)} (0) = \begin{cases} n!(-1)^j a_{n-2j} b_{n-2j}^j & \text{if } k = n - 2j \text{ for some } j \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$f(0) = a_0, \ f'(0) = a_1, \text{ and} \tag{102}$$

$$\forall n \geq 2,\ \frac{f^{(n)}(0)}{n!} = a_n + \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} (-1)^j a_{n-2j} b_{n-2j}^j. \tag{103}$$

From (102) and (103) it follows that given $\{b_n\}$ and $\{c_n\}$, there is a uniquely determined sequence $\{a_n\}$ such that $f^{(n)}(0) = c_n$ for all $n \in \mathbb{N}$.
Step 2: Let $b > 0$ and consider

$$\frac{x^k}{b^2 + x^2} = \frac{x^{k-1}}{2}\left(\frac{1}{x + bi} + \frac{1}{x - bi}\right).$$

By the Generalized Leibniz Rule, we have

$$\left(\frac{x^k}{b^2 + x^2}\right)^{(n)} = \frac{1}{2}\sum_{j=0}^{n} \binom{n}{j} (k-1)\cdots(k-1-n+j) x^{k-1-n+j}$$

$$\cdot \left(\frac{(-1)^j j!}{(x + bi)^{j+1}} + \frac{(-1)^j j!}{(x - bi)^{j+1}}\right)$$

$$= \frac{n!}{2} x^{k-n-2} \sum_{j=0}^{n} (-1)^j \frac{(k-1)\cdots(k-1-n+j)}{(n-j)!} \left(\frac{x^{j+1}}{(x + bi)^{j+1}} + \frac{x^{j+1}}{(x - bi)^{j+1}}\right).$$

We have (see the following exercise)

$$\left|\frac{x^{j+1}}{(x + bi)^{j+1}} + \frac{x^{j+1}}{(x - bi)^{j+1}}\right| \leq 2. \tag{104}$$

It follows that

$$\forall k \geq n+2,\ \left|\left(\frac{a_k x^k}{1 + b_k x^2}\right)^{(n)}\right| \leq \frac{(n+1)! k! |a_k|}{b_k} |x|^{k-n-2}. \tag{105}$$

Taking $b_k = (k!)^2 |a_k|$, we get

$$\sum_{k \geq n+2} \left|\left(\frac{a_k x^k}{1 + b_k x^2}\right)^{(n)}\right| \leq (n+1)! \sum_{k \geq n+2} \frac{|x|^{k-n-2}}{k!}. \tag{106}$$

The right hand side of (106) is uniformly convergent on every bounded interval, hence for every $n \in \mathbb{N}$ the convergence of

$$\sum_{k=0}^{\infty} \left(\frac{a_k x^k}{1 + b_k x^2}\right)^{(n)}$$

follows from the Weierstrass M-Test. Applying Theorem 13.6 establishes the fact that the given series of nth derivatives converges to the nth derivative of f. This completes the proof.

Exercise: a) Let $w \in \mathbb{C}$. Show that if $|w| \leq 1$, then for all $n \in \mathbb{Z}^+$, $|w^n + \overline{w}^n| \leq 2$.
b) Verify equation (104).

5. The Weierstrass Approximation Theorem

5.1. Statement of Weierstrass Approximation.

THEOREM 14.20. *(Weierstrass Approximation Theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function and ϵ any positive number. Then there exists a polynomial function P such that for all $x \in [a, b]$, $|f(x) - P(x)| < \epsilon$. In other words, any continuous function defined on a closed, bounded interval is the uniform limit of a sequence of polynomials.*

Exercise: For each $n \in \mathbb{Z}^+$, let $P_n : \mathbb{R} \to \mathbb{R}$ be a polynomial function. Suppose that there is $f : \mathbb{R} \to \mathbb{R}$ such that $P_n \stackrel{u}{\to} f$ on all of $\mathbb{R}$. Show that the sequence of functions $\{P_n\}$ is eventually constant: there exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$, $P_n(x) = P_m(x)$ for all $x \in \mathbb{R}$.

It is interesting to compare Theorem 14.20 with Taylor's theorem, which gives conditions for a function to be equal to its Taylor series. Note that any such function must be C^∞ (i.e., it must have derivatives of all orders), whereas in the Weierstrass Approximation Theorem we can get *any* continuous function. An important difference is that the Taylor polynomials $T_N(x)$ have the property that $T_{N+1}(x) = T_N(x) + a_{N+1}x^N$, so that in passing from one Taylor polynomial to the next, we are not changing any of the coefficients from 0 to N but only adding a higher order term. In contrast, for the sequence of polynomials $P_n(x)$ uniformly converging to f in Theorem 1, $P_{n+1}(x)$ is not required to have any simple algebraic relationship to $P_n(x)$.

Theorem 14.20 was first established by Weierstrass in 1885. To this day it is one of the most central and celebrated results of mathematical analysis. Many mathematicians have contributed novel proofs and generalizations, notably S.J. Bernstein [**Be12**] and M.H. Stone [**St37**], [**St48**]. But – more than any result of undergraduate mathematics I can think of except the **quadratic reciprocity law** – the passage of time and the advancement of mathematical thought have failed to single out any one preferred proof. We have decided to follow an argument given by Noam Elkies.[7] This argument is reasonably short and reasonably elementary, although as above, not definitively more so than certain other proofs. However it unfolds in a logical way, and every step is of some intrinsic interest. Best of all, at a key stage we get to apply our knowledge of Newton's binomial series!

5.2. Piecewise Linear Approximation.

A function $f : [a, b] \to \mathbb{R}$ is **piecewise linear** if it is a continuous function made up out of finitely many straight line segments. More formally, there exists a partition $P = \{a = x_0 < x_1 \ldots < x_n = b\}$ such that for $1 \leq i \leq n$, the restriction of f to $[x_{i-1}, x_i]$ is a linear function. For instance, the absolute value function is piecewise linear. In fact, the general piecewise can be expressed in terms of absolute values of linear functions, as follows.

[7] http://www.math.harvard.edu/∼elkies/M55b.10/index.html

LEMMA 14.21. *Let $f : [a, b] \to \mathbb{R}$ be a piecewise linear function. Then there is $n \in \mathbb{Z}^+$ and $a_1, \ldots, a_n, m_1, \ldots, m_n, b \in \mathbb{R}$ such that*

$$f(x) = b + \sum_{i=1}^{n} \pm |m_i x_i + b_i|.$$

PROOF. We leave this as an elementary exercise. Some hints:
(i) If $a_j \leq a$, then as functions on $[a, b]$, $|x - a_j| = x - a_j$.
(ii) The following identities may be useful:

$$\max(f, g) = \frac{|f + g|}{2} + \frac{|f - g|}{2}$$

$$\min(f, g) = \frac{|f + g|}{2} - \frac{|f - g|}{2}.$$

(iii) One may, for instance, go by induction on the number of "corners" of f. □

Now every continuous function $f : [a, b] \to \mathbb{R}$ may be uniformly approximated by piecewise linear functions, and moreover *this is very easy to prove.*

PROPOSITION 14.22. *(Piecewise Linear Approximation) Let $f : [a, b] \to \mathbb{R}$ be a continuous function and $\epsilon > 0$ be any positive number. Then there exists a piecewise linear function P such that for all $x \in [a, b]$, $|f(x) - P(x)| < \epsilon$.*

PROOF. Step 1: Let $f : [c, d] \to \mathbb{R}$ be a continuous function, and put $M = \omega(f, [c, d]) = \max(f, [c, d]) - \min(f, [c, d])$. Let $L : [c, d] \to \mathbb{R}$ be the unique linear function with $L(c) = f(c)$ and $L(d) = f(d)$. We CLAIM that for all $x \in [c, d]$, $|f(x) - L(x)| \leq M$. To establish the CLAIM we may argue (for instance) as follows: for all x, either $f(x) \geq L(x)$ or $f(x) \leq L(x)$. In the former case,

$$|f(x) - L(x)| \leq \max(f, [c, d]) - \min(L, [c, d]) = \max(f, [c, d]) - \min(L(c), L(d))$$
$$= \max(f, [c, d]) - \min(f(c), f(d)) \leq \max(f, [c, d]) - \min(f, [c, d]) = M,$$

and the argument in the latter case is very similar.
Step 2: By the Uniform Continuity Theorem, there is $\delta > 0$ such that whenever $|x - y| < \delta$, $|f(x) - f(y)| < \epsilon$. Choose n large enough so that $\frac{b-a}{n} < \delta$, and consider the partition $P_n = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n\}$ of $[a, b]$ into n subintervals of equal length $\frac{b-a}{n}$. Let $P : [a, b] \to \mathbb{R}$ be piecewise linear, linear on $[x_i, x_{i+1}]$ and such that $P(x_i) = f(x_i)$ for all $0 \leq i \leq n$. Then for all $0 \leq i \leq n$, $\omega(f, [x_i, x_{i+1}]) \leq \epsilon$, so by Step 1 $|f(x) - P(x)| \leq \epsilon$ for all $x \in [x_i, x_{i+1}]$ and thus for all $x \in [a, b]$. □

5.3. A Very Special Case.

LEMMA 14.23. *(Elkies) Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series. We suppose:*
(i) The sequence of signs *of the coefficients a_n is eventually constant.*
(ii) The radius of convergence is 1.
(iii) $\lim_{x \to 1^-} f(x) = L$ exists.
Then $\sum_{n=0}^{\infty} a_n = L$, and the convergence of the series to the limit function is uniform on $[0, 1]$.

Exercise: Prove Lemma 14.23. Two suggestions:
(i) Reduce to the case in which $a_n \geq 0$ for all $n \in \mathbb{N}$.
(ii) Use the Weierstrass M-Test.

PROPOSITION 14.24. *For any $\alpha > 0$, the function $f(x) = |x|$ on $[-\alpha, \alpha]$ can be uniformly approximated by polynomials.*

PROOF. Step 1: Suppose that for all $\epsilon > 0$, there is a polynomial function $P : [-1,1] \to \mathbb{R}$ such that $|P(x) - |x|| < \epsilon$ for all $x \in [-1,1]$. Put $x = \frac{y}{\alpha}$. Then for all $y \in [-\alpha, \alpha]$ we have

$$|\alpha P(\frac{y}{\alpha}) - \alpha|\frac{y}{\alpha}|| = |Q(y) - |y|| < \alpha\epsilon,$$

where $Q(y) = \alpha P(\frac{y}{\alpha})$ is a polynomial function of y. Since $\epsilon > 0$ was arbitrary: if $x \mapsto |x|$ can be uniformly approximated by polynomials on $[-1,1]$, then it can be uniformly approximated by polynomials on $[-\alpha, \alpha]$. So we are reduced to $\alpha = 1$.
Step 2: Let $T_N(y)$ be the degree N Taylor polynomial at zero for the function $f(y) = \sqrt{1-y}$. By the Binomial Theorem,

$$(1+y)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} y^n$$

valid for all $\alpha \in \mathbb{R}$ and all $y \in (-1,1)$. Taking $\alpha = \frac{1}{2}$ and substituting $-y$ for y, we see that the degree N Taylor polynomial for $\sqrt{1-y}$ at zero is

$$T_N(y) = \sum_{n=0}^{N} (-1)^n \binom{\frac{1}{2}}{n} y^n,$$

and $\lim_{N\to\infty} T_N(y) = \sqrt{1-y}$ for $y \in [0,1)$. Further, $(-1)^n \binom{\frac{1}{2}}{n} < 0$ for $n \geq 1$, and

$$\lim_{y\to 1^-} f(y) = \lim_{y\to 1^-} \sqrt{1-y} = 0.$$

Thus we may apply Elkies' Lemma to get $T_N(y) \stackrel{u}{\to} \sqrt{1-y}$ on $[0,1]$. For $x \in [-1,1]$, $y = 1 - x^2 \in [0,1]$, so making this substitution we find that on $[-1,1]$,

$$T_N(1-x^2) = \sum_{n=0}^{N} (-1)^n \binom{\frac{1}{2}}{n} (1-x^2)^n \stackrel{u}{\to} \sqrt{1-(1-x^2)} = \sqrt{x^2} = |x|.$$

□

5.4. Proof of the Weierstrass Approximation Theorem.

For $a < b \in \mathbb{R}$, let $\mathcal{C}[a,b]$ be the set of all continuous functions $f : [a,b] \to \mathbb{R}$, and let $\mathcal{P}$ be the set of all polynomial functions $f : [a,b] \to \mathbb{R}$. Let $\mathrm{PL}([a,b])$ denote the set of piecewise linear functions $f : [a,b] \to \mathbb{R}$.

For a subset $S \subset \mathcal{C}[a,b]$, we define the **uniform closure** of S to be the set $\overline{S}$ of all $f \in \mathcal{C}[a,b]$ which are uniform limits of sequences in S: precisely, for which there is a sequence of functions $f_n : [a,b] \to \mathbb{R}$ with each $f_n \in S$ and $f_n \stackrel{u}{\to} f$.

LEMMA 14.25. *For any subset $S \subset \mathcal{C}[a,b]$, we have $\overline{\overline{S}} = \overline{S}$.*

PROOF. Simply unpacking the notation is at least half of the battle here. Let $f \in \overline{\overline{S}}$, so that there is a sequence of functions $g_i \in \overline{S}$ with $g_i \stackrel{u}{\to} f$. Similarly, since each $g_i \in \overline{S}$, there is a sequence of continuous functions $f_{ij} \stackrel{u}{\to} g_i$. Fix $k \in \mathbb{Z}^+$: choose n such that $||g_n - f|| < \frac{1}{2k}$ and then j such that $||f_{nj} - g_n|| < \frac{1}{2k}$; then

$$||f_{nj} - f|| \leq ||f_{nj} - g_n|| + ||g_n - f|| < \frac{1}{2k} + \frac{1}{2k} = \frac{1}{k}.$$

Thus if we put $f_k = f_{nj}$, then for all $k \in \mathbb{Z}^+$, $||f_k - f|| < \frac{1}{k}$ and thus $f_k \stackrel{u}{\to} f$. □

Observe that the Piecewise Linear Approximation Theorem is

$$\overline{\mathrm{PL}[a,b]} = \mathcal{C}[a,b],$$

whereas the Weierstrass Approximation Theorem is

$$\overline{\mathcal{P}} = \mathcal{C}[a,b].$$

Finally, the point: it's enough to show that every piecewise linear function can be uniformly approximated by polynomial functions, for then $\overline{\mathcal{P}} \supset \mathrm{PL}[a,b]$, so

$$\overline{\mathcal{P}} = \overline{\overline{\mathcal{P}}} \supset \overline{\mathrm{PL}[a,b]} = \mathcal{C}[a,b].$$

Thus the following result completes the proof of Theorem 14.20.

PROPOSITION 14.26. *We have* $\overline{\mathcal{P}} \supset \mathrm{PL}[a,b]$.

PROOF. Let $f \in \mathrm{PL}[a,b]$. By Lemma 14.21, we may write

$$f(x) = b + \sum_{i=1}^{n} \pm |m_i x + b_i|.$$

Choose $\alpha > 0$ such that for all $1 \leq i \leq n$, if $x \in [a,b]$, then $m_i x + b_i \in [-\alpha, \alpha]$. For each $1 \leq i \leq n$, by Lemma 14.24 there is a polynomial P_i such that for all $x \in [a,b]$, $|P_i(m_i x + b_i) - |m_i x + b_i|| < \frac{\epsilon}{n}$. Let $P : [a,b] \to \mathbb{R}$ by $P(x) = b + \sum_{i=1}^{n} \pm P_i(m_i x + b_i)$. Then $P \in \mathcal{P}$ and for all $x \in [a,b]$,

$$|P(x) - f(x)| = \left| \sum_{i=1}^{n} \pm (P_i(m_i x + b_i) - |m_i x + b_i|) \right| < \sum_{i=1}^{n} \frac{\epsilon}{n} = \epsilon.$$

□

6. A Continuous, Nowhere Differentiable Function

We are going construct a function $f : \mathbb{R} \to \mathbb{R}$ with the following striking property: for all $x_0 \in \mathbb{R}$, f is continuous at x_0 but f is *not* differentiable at x_0. In short, we say that f is *continuous but nowhere differentiable*.

The first such construction (accompanied by a complete, correct proof) was given in a seminal 1872 paper of Weierstrass. Weierstrass's example was as follows: let $\alpha \in (0,1)$, and let b be a positive odd integer such that $\alpha b > 1 + \frac{3\pi}{2}$. Then the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \sum_{n=0}^{\infty} \alpha^n \cos(b^n \pi x) \tag{107}$$

is continuous on $\mathbb{R}$ but not differentiable at any $x \in \mathbb{R}$.

Exercise: Show that the function defined by (107) above is continuous.

Unfortunately the proof that f is nowhere differentiable is not so easy, as indicated by the rather specific conditions given on the parameters α, b. (For less carefully chosen α, b the function f can have a "small" set of points of differentiability.) Thus, as with most other contemporary treatments, we will switch to a different function for which the nowhere differentiability calculation is more straightforward. More specifically, we will switch from trigonometric functions to our new friends the piecewise linear functions, so first we interpose the following exercise nailing

down some further (simple) properties of these functions.

Exercise:
a) Let $f : [a, b] \to \mathbb{R}$ be a piecewise linear function with slopes $m_1, \ldots, m_n$. Show that f is Lipschitz, and the smallest possible Lipschitz constant is $C = \max_i |m_i|$.
b) Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Suppose that there is $C > 0$ such that for every closed subinterval $[a, b]$ of $\mathbb{R}$, C is a Lipschitz constant for the restriction of f to $[a, b]$. Show that C is a Lipschitz constant for f.
c) Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise linear function with "corners" at the integers – i.e., f is differentiable on $(n, n + 1)$ for all $n \in \mathbb{Z}^+$ and is not differentiable at any integer n. For $n \in \mathbb{Z}$, let m_n be the slope of f on the interval $(n, n + 1)$. Let $C = \sup_{n\in\mathbb{Z}} m_n$. Show that f is Lipschitz iff $C < \infty$, in which case C is the smallest Lipschitz contant for f.

Now we begin our construction with the "sawtooth function" $S : \mathbb{R} \to \mathbb{R}$: the unique piecewise linear function with corners at the integers and such that $S(n) = 0$ for every even integer n and $S(n) = 1$ for every odd integer n. The slopes of S are all ± 1, so by the preceding exercise S is Lipschitz (hence continuous):

$$\forall x, y \in \mathbb{R}, |S(x) - S(y)| \le |x - y|.$$

Also S is 2-periodic: for all $x \in \mathbb{R}$, $S(x + 2) = S(x)$. For $k \in \mathbb{N}$, define

$$f_k : \mathbb{R} \to \mathbb{R},\ f_k(x) = \left(\frac{3}{4}\right)^k S(4^k x).$$

We suggest that the reader sketch the graphs of the functions f_k: roughly speaking they are sawtooth functions which, as k increases, oscillate more and more rapidly but with smaller amplitude: indeed $||f_k(x)|| = \left(\frac{3}{4}\right)^k$. We define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \sum_{k=0}^{\infty} f_k(x) = \sum_{k=0}^{\infty} \left(\frac{3}{4}\right)^k S(4^k x).$$

Since $\sum_{k=0}^{\infty} ||f_k|| = \sum_{k=0}^{\infty} \left(\frac{3}{4}\right)^k < \infty$, the series defining f converges uniformly by the Weierstrass M-Test. This also gives that f is continuous, since f is a uniform limit of a sequence of continuous functions. We claim however that f is nowhere differentiable. To see this, fix $x_0 \in \mathbb{R}$. We will define a sequence $\{\delta_n\}$ of nonzero real numbers such that $\delta_n \to 0$ and the sequence

$$D_n = \frac{f(x_0 + \delta_n) - f(x_0)}{\delta_n}$$

is divergent. This implies that f is not differentiable at x_0.

Let's do it. First suppose that the fractional part of x_0 lies in $[0, \frac{1}{2})$, so that the interval $(x_0, x_0 + \frac{1}{2})$ contains no integers. In this case we put

$$\delta_n = \frac{4^n}{2},$$

and the reason for our choice is that the interval $(4^n x_0, 4^n(x_0 + \delta_n))$ contains no integers. Let $k, n \in \mathbb{N}$. We claim the following inequalities:

$$\forall k > n,\ |S(4^k x_0 + 4^k \delta_n) - S(4^k x_0)| = 0. \tag{108}$$

$$\forall k = n,\ |S(4^k x_0 + 4^k \delta_n) - S(4^k x_0)| = \frac{1}{2}. \tag{109}$$

$$]\forall k < n,\ |S(4^k x_0 + 4^k \delta_n) - S(4^k x_0)| \leq |4^k \delta_n|. \tag{110}$$

Indeed: (108) holds because if $k > n$, $4^k x_0 + 4^k \delta_n - 4^k x_0 = 4^k \delta_n = \frac{4^{k-n}}{2}$ is a multiple of 2 and S is a 2-periodic function; (109) holds because if $k = n$, $4^k x_0 + 4^k \delta_n = 4^k x_0 + \frac{1}{2}$, so by our choice of δ_n, the function S is linear on $[4^k x_0, 4^k x_0 + \frac{1}{2}]$ of slope ± 1, hence the difference between its values at the endpoints is $\frac{\pm 1}{2}$. Finally, (110) holds because 1 is a Lipschitz constant for S. Using these results and the Reverse Triangle Inequality gives

$$\left|\frac{f(x_0 + \delta_n) - f(x_0)}{\delta_n}\right| = \left|\sum_{k=0}^{n} \left(\frac{3}{4}\right)^k \frac{S(4^k x_0 + 4^k \delta_n) - S(4^k x_0)}{\delta_n}\right|$$

$$\geq \left(\frac{3}{4}\right)^n 4^n - \sum_{k=0}^{n-1} \left(\frac{3}{4}\right)^k \cdot \left|\frac{S(4^k x_0 + 4^k \delta_n) - S(4^k x_0)}{\delta_n}\right|$$

$$\geq 3^n - \sum_{k=0}^{n-1} 3^k = 3^n - \frac{3^n - 1}{2} \geq \frac{3^n}{2}.$$

Thus $D_n \to \infty$, so f is not differentiable at x_0.

We're not quite done: recall that we assumed that the fractional part of x_0 lay in $[0, \frac{1}{2})$, with the consequence that S was linear on the interval $[(4^n x_0, 4^n(x_0 + \delta_n)]$. What to do if the fractional part of x_0 lies in $[\frac{1}{2}, 1)$? In this case we take $\delta_n = \frac{-4^n}{2}$ so that the interval $(4^n(x_0 + \delta_n), 4^n x_0)$ contains no integers so S is linear on the interval $[4^n(x_0 + \delta_n), 4^n x_0]$, and the rest of the proof goes through as above.

So, albeit with a different function, we have proved Weierstrass's Theorem.

THEOREM 14.27. *(Weierstrass, 1872) There is a function $f : \mathbb{R} \to \mathbb{R}$ which is continuous at every point of $\mathbb{R}$ but differentiable at no point of $\mathbb{R}$.*

Notice that if we restrict f to some closed interval, say $[0, 2]$, then by the Weierstrass Approximation Theorem f is – like any continuous function on $[0, 2]$ – a uniform limit of polynomials. Thus even a uniform limit of polynomials on a closed, bounded interval need not have any good differentiability properties whatsoever!

7. The Gamma Function

There are many real functions. Some are "elementary": i.e., built up from power functions, trigonometric functions, exponentials and logarithms. Most are not, like the Gaussian error function $E(x) = \int_{-\infty}^{x} e^{-t^2} dt$. As mathematical analysis developed from Newton and Leibniz in the 17th century on down, eventually it became clear that certain non-elementary functions arise again and again in a variety of contexts, and satisfy some remarkable, beautiful identities.

Such functions are called **special functions**. There is no precise mathematical definition of a special function. Rather, the appellation is historical and cultural. Unfortunately the days when the majority of students and practitioners of mathematics naturally learned about special functions in the context of their work are past, since indeed it is no longer the case that that the majority of students and practitioners of mathematics are deeply concerned with real function theory. In our day your smart friend the physicist or the engineer probably knows more about

special functions than you do...and this may not be a good thing.

In this section we will study one of the most ubiquitous special functions of them all, the **Gamma function**.

7.1. Definition and Basic Properties.

For $x \in (0, \infty)$ we define

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

As this is an improper integral, thre is of course something to check.

PROPOSITION 14.28. *For $x \in (0, \infty)$, the integral defining $\Gamma(x)$ is convergent.*

Exercise: Prove Proposition 14.28. Some suggestions:
(i) First deal with the case $x \geq 1$, since in this case the function $t \mapsto t^{x-1}$ is continuous on $[0, \infty)$ and the integral is "improper only at ∞. To show the convergence use the fact that exponentials grow faster than polynomials.
(ii) When $0 < x < 1$, the integral is also improper at 0, so it should be split into two pieces, say $\int_0^1$ and $\int_1^\infty$. The argument of part a) will handle the latter integral. Reduce the former integral to $\int_0^1 \frac{dx}{x^p}$.

Exercise: a) Show that the improper integral $\int_0^\infty t^{x-1} e^{-t} dt$ is divergent for $x \leq 0$.
b) Show that $\Gamma(x) > 0$ for all $x \in (0, \infty)$.
b) Show that $\lim_{x \to 0^+} \Gamma(x) \lim_{x \to \infty} \Gamma(x) = \infty$.
c) Show that $\Gamma(1) = \Gamma(2) = 1$.

Exercise: a) By making the change of variables $t = s^2$, show that for all $x \in (0, \infty)$,

$$\Gamma(x) = 2 \int_0^\infty s^{2x-1} e^{-s^2} ds. \tag{111}$$

b) Deduce that

$$\Gamma(\frac{1}{2}) = 2 \int_0^\infty e^{-x^2} dx = \int_{-\infty}^\infty e^{-x^2} dx. \tag{112}$$

THEOREM 14.29. *a) For all $x \in (0, \infty)$ we have*

$$\Gamma(x+1) = x\Gamma(x). \tag{113}$$

b) For all $n \in \mathbb{N}$, $\Gamma(n+1) = n!$.

Exercise: Prove Theorem 14.29. Suggestions:
For part a), integrate by parts, much as in XXXX.
For part b), use part a), the previous exercise, and induction.

Thus the Gamma function is a continuous interpolation of the factorial function (with a slight shift in the argument that everyone finds a little distressing at first but eventually learns to live with). Of course there are infinitely many continuous (even infinitely differentiable) functions which interpolate any real sequence. Is there a sense in which the Gamma function is the "right" interpolation?

Yes, although it is rather curious.

THEOREM 14.30. *(Bohr-Mollerup* [**BM22**]*)*
a) The function $\log\Gamma : (0,\infty) \to \mathbb{R}$ *is convex.*
b) Conversely, let $f : (0,\infty) \to (0,\infty)$ *satisfy:*
(i) $f(x+1) = xf(x)$ *for all* $x \in (0,\infty)$*;*
(ii) $f(1) = 1$*; and*
(iii) $\log f$ *is convex.*
Then $f = \Gamma$*.*

PROOF. a) Let $f = \log\Gamma$. We will verify the secant-graph inequality: for all $0 < x < y < \infty$ and $\lambda \in (0,1)$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y),$$

and for this we will use Hölder's Integral Inequality with $p = \frac{1}{\lambda}$ and $q = \frac{1}{1-\lambda}$:

$$f((1-\lambda)a + \lambda b) = \log \int_0^\infty t^{\lambda x + (1-\lambda)y - 1} e^{-t} dt$$

$$= \log\left(\int_0^\infty (t^{x-1}e^{-t})^\lambda (t^{y-1}e^{-t})^{1-\lambda} dt\right)$$

$$\leq \log\left(\left(\int_0^\infty t^{x-1}e^{-t}dt\right)^\lambda \left(\int_0^\infty t^{y-1}e^{-t}dt\right)^{1-\lambda}\right)$$

$$= \lambda \log\Gamma(x) + (1-\lambda)\log\Gamma(y).$$

b) ... □

So it seems interesting to investigate non-integer values of the Gamma function. In some sense the first order of business is to evaluate $\Gamma(\frac{1}{2})$. Then using (113) we can evaluate $\Gamma(\frac{n}{2})$ for all integers n. Moreover, by (112) we know that $\Gamma(\frac{1}{2}) = \int_{-\infty}^\infty e^{-x^2} dx$, an improper integral that is easily seen to be convergent but whose precise value is not so easy to determine.

To evaluate $\Gamma(\frac{1}{2})$ we introduce a second special function and relate it back to the Gamma function using the Bohr-Mollerup Theorem.

7.2. The Beta Function.

For $x, y \in (0,\infty)$, we define the **Beta function**

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt.$$

When $x, y \geq 1$, this is a "proper integral", and there are no convergence issues. However, when one of x, y is less than 1, there is something to check.

Exercise: Show that the integral defining $B(x,y)$ is convergent for all $x, y > 0$.

THEOREM 14.31. *For all* $x, y > 0$*, we have*

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

PROOF. ... □

Making the substitution $t = \sin^2\theta$ in the integral defining $B(x,y)$ and applying Theorem 14.31 we get

$$2\int_0^{\frac{\pi}{2}} (\sin\theta)^{2x-1}(\cos\theta)^{2y-1}\,d\theta = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{114}$$

Taking $x = y = \frac{1}{2}$ in (114) we deduce:

THEOREM 14.32.

$$\Gamma\left(\frac{1}{2}\right) = \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}. \tag{115}$$

7.3. Interlude: A Dominated Convergence Theorem.

THEOREM 14.33. *Let $f_n, f, g : (0,\infty) \to \mathbb{R}$. We suppose:*
(i) f_n and g are Riemann integrable on every closed, bounded subinterval of $(0,\infty)$;
(ii) $f_n \to f$ pointwise, and the convergence is uniform on each closed, bounded subinterval of $[0,\infty)$.
(iii) $|f_n(x)| \le g(x)$ for all $x \in (0,\infty)$; and
(iv) $\int_0^\infty g(x)dx < \infty$.
Then $\lim_{n\to\infty}\int_0^\infty f_n(x)dx = \int_0^\infty f(x)dx$.

PROOF. Since $|f_n(x)| \le g(x)$ for all x and $f_n \to f$, also $|f(x)| \le g(x)$ for all x. Fix $\epsilon > 0$. Since g is non-negative and $\int_0^\infty g(x)dx < \infty$, there is $N \in \mathbb{Z}^+$ with

$$\int_0^{\frac{1}{N}} g(x)dx + \int_N^\infty g(x)dx < \epsilon,$$

Thus for all $n \in \mathbb{Z}^+$,

$$\left|\int_0^{\frac{1}{N}} f_n(x)dx\right| + \left|\int_N^\infty f_n(x)dx\right| \le \int_0^{\frac{1}{N}} |f_n(x)|dx + \int_N^\infty |f_n(x)|dx < \epsilon,$$

and similarly

$$\left|\int_0^{\frac{1}{N}} f(x)dx\right| + \left|\int_N^\infty f(x)dx\right| \le \int_0^{\frac{1}{N}} |f(x)|dx + \int_N^\infty |f(x)|dx < \epsilon.$$

Since $f_n \to f$ uniformly on $[\frac{1}{N}, N]$,

$$\lim_{n\to\infty}\int_{\frac{1}{N}}^N f_n(x)dx = \int_{\frac{1}{N}}^N f(x)dx.$$

It follows that for all sufficiently large n,

$$\left|\int_0^\infty f_n(x)dx - \int_0^\infty f(x)dx\right|$$

$$\le \left(\int_0^{\frac{1}{N}} + \int_N^\infty\right)|f_n(x)|dx + \left(\int_0^{\frac{1}{N}} + \int_N^\infty\right)|f(x)|dx + \left|\int_{\frac{1}{N}}^N (f_n(x) - f(x))dx\right|$$

$$\epsilon + \epsilon + \epsilon = 3\epsilon.$$

Since ϵ was arbitrary, the proof is complete. □

Remark: Theorem 14.33 is given as an exercise in W. Rudin's text [**R**, p. 167]. It is a weak form of what is perhaps the most important and useful single result of graduate level real variable theory, the **Lebesgue Dominated Convergence Theorem**. In Lebesgue's version, the integrals are taken in his more permissive sense (which we certainly have not discussed). More significantly, the hypothesis of uniform convergence on closed bounded subintervals is dropped entirely: all that is needed is that $f_n \to f$ pointwise.

Since uniform convergence on bounded subintervals is a much stronger hypothesis than pointwise convergence, Rudin's version of the Dominated Convergence Theorem is significantly weaker than Lebesgue's. (For a version of the Dominated Convergence Theorem which uses only pointwise convergence but stays in the context of the Riemann integral, see [**Ke70**]. But I must warn you that even this paper requires somewhat more background than we have developed here.) Nevertheless Theorem 14.33 has some useful applications.

7.4. Stirling's Formula.

THEOREM 14.34.

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{(x/e)^x\sqrt{2\pi x}} = 1. \tag{116}$$

PROOF. ... □

CHAPTER 15

Several Real Variables and Complex Numbers

1. A Crash Course in the Honors Calculus of Several Variables

Let $M, N \in \mathbb{Z}^+$. In **multivariable calculus** one studies functions $f : \mathbb{R}^N \to \mathbb{R}^M$, that is "vector valued functions of several variables". Here we will be briefly interested in several aspects of the case $M = 1$, i.e., of functions $f : \mathbb{R}^N \to \mathbb{R}$ and also of sequences and series with values in $\mathbb{R}^N$.

First, recall that by $\mathbb{R}^N$ we mean the set of all ordered N-tuples $x = (x_1, \ldots, x_N)$ of real numbers. To define things like continuity and convergence, need a way to measure the distance between $x, y \in \mathbb{R}^N$. For this, we define

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

Notice that, at least when $1 \leq N \leq 3$, this is the distance from the origin $0 = (0, \ldots, 0)$ to the vector x that we learn about in high school geometry. The basic fact that allows estimates to be made is the **triangle inequality**:

$$\forall x, y \in \mathbb{R}^N, |x + y| \leq |x| + |y|.$$

We proved this at the end of Chapter 8 as an application of Jensen's Inequality.

Here are the two key definitions: a sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathbb{R}^N$ is given by a function $n \mapsto x_n$ from $\mathbb{Z}^+$ to $\mathbb{R}^N$. We say that the sequence x_n **converges** to $x \in \mathbb{R}^N$ if for all $\epsilon > 0$, there exists $N_0 \in \mathbb{Z}^+$ such that for all $n \geq N_0$, $|x_n - x| < \epsilon$. A sequence is **convergent** if it converges to some element of $\mathbb{R}^N$ and otherwise **divergent**.

Let $D \subset \mathbb{R}^N$, and let $x \in D$. A function $f : D \to \mathbb{R}$ is **continuous at x** if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $y \in D$ with $|x - y| < \delta$, $|f(x) - f(y)| < \epsilon$. A function $f : D \to \mathbb{R}$ is **continuous** if it is continuous at every point of D.

Exercise 15.1: Let $x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in \mathbb{R}^N$. Show that for all $1 \leq i \leq N$, $|x_i - y_i| \leq |x - y|$.

Example 15.2: For $1 \leq i \leq N$, let $\pi_i : \mathbb{R}^N \to \mathbb{R}$ be the function which maps $(x_1, \ldots, x_N)$ to x_i. Then π_i is continuous: indeed, for any $\epsilon > 0$, we may take $\delta = \epsilon$, since then if $|x - y| < \delta$, $|\pi_i(x) - \pi_i(y)| = |x_i - y_i| \leq |x - y| < \epsilon$.

Exercise 15.3: Fix $x_0 \in \mathbb{R}^N$, and consider the function $d_{x_0} : \mathbb{R}^N \to \mathbb{R}$, $d_{x_0}(x) = |x - x_0|$. Show that d_{x_0} is continuous.

PROPOSITION 15.1. *Let $f_1, f_2 : D \subset \mathbb{R}^N \to \mathbb{R}$ be two functions, and let $\alpha, \beta \in \mathbb{R}$.*

a) If f_1 and f_2 are continuous, so is $\alpha f_1 + \beta f_2$.
b) If f_1 and f_2 are continuous, so is $f_1 f_2$.
c) If f_1 and f_2 are continuous and for all $x \in D$, $f_2(x) \neq 0$, then $\frac{f_1}{f_2}$ is continuous.

Exercise 15.4: Prove Proposition 15.1. (Suggestion: adapt the $N = 1$ case.)

Exercise 15.5: a) Give a reasonably careful definition of a polynomial function $f : \mathbb{R}^N \to \mathbb{R}$. (Hint: the functions π_i defined above are relevant.)
b) Show that every polynomial function is continuous.

PROPOSITION 15.2. *Let $\{x_n\}$ be a sequence in $\mathbb{R}^N$, and let $x \in \mathbb{R}^N$. TFAE:*
(i) $x_n \to x$.
(ii) For all $1 \leq i \leq n$, $\pi_i(x_n) \to \pi_i(x)$.

PROOF. (i) $\implies$ (ii): Fix $\epsilon > 0$. By assumption, there is $N_0 \in \mathbb{Z}^+$ such that for all $n \geq N_0$, $|x_n - x| < \epsilon$. Hence by Exercise 15.1, for all $1 \leq i \leq N$, $|\pi_i(x_n) - \pi_i(x)| < \epsilon$.
(ii) $\implies$ (i): Fix $\epsilon > 0$. For each $1 \leq i \leq N$, there exists $N_i \in \mathbb{Z}^+$ such that for all $n \geq N_i$, $|\pi_i(x_n) - \pi_i(x)| < \epsilon$. Take $N_0 = \max(N_1, \ldots, N_N)$; for $n \geq N_0$, $|x_n - x| = \sqrt{(x_{n,1} - x_1)^2 + \ldots + (x_{n,N} - x_N)^2} < \sqrt{N\epsilon^2} = \sqrt{N}\epsilon$. Good enough. □

PROPOSITION 15.3. *Let $f : D \subset \mathbb{R}^N \to \mathbb{R}$ be a continuous function, and let $\{x_n\}$ be a sequence in D. if $x_n \to x$, then $f(x_n) \to f(x)$.*

Exercise 15.6: Prove Proposition 15.3.

We now press on to give analogues of the fundamental **Interval Theorems** of honors calculus. Our first order of business is to find a suitable analogue in $\mathbb{R}^N$ of the notion of a closed, bounded interval in $\mathbb{R}$. It turns out that we have some leeway here. Let us consider two different kinds of sets.

For $1 \leq i \leq N$, choose real numbers $a_i \leq b_i$. We define the **closed coordinate box** $[a, b]$ to be the set of all $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$ such that $a_i \leq x_i \leq b_i$ for all $1 \leq i \leq N$. Further, for $x \in \mathbb{R}^N$ and $r \geq 0$, we define the **open ball** $B_r(x)$ to be the set of all $y \in \mathbb{R}^N$ such that $|x - y| < r$ and the **closed ball** $\overline{B}_r(x)$ to be the set of all $y \in \mathbb{R}^N$ such that $|x - y| \leq r$.

Exercise 15.7: For a subset D of $\mathbb{R}^N$, show that the following are equivalent:
(i) There is some closed coordinate box $[a, b]$ with $D \subset [a, b]$.
(ii) There are $x \in \mathbb{R}^N$ and $r \geq 0$ with $D \subset B_r(x)$.
(iii) There is $r \geq 0$ with $D \subset B_0(x)$.
A subset D satisfying these equivalent properties is **bounded**.

Exercise 15.8: Let $D \subset \mathbb{R}^N$ be a bounded subset. For all $1 \leq i \leq N$, show thaqt $\pi_i(D)$ is a bounded subset of $\mathbb{R}$.

Exercise 15.9: Show that all closed coordinate boxes and all open and closed balls are convex subsets of $\mathbb{R}^N$.

THEOREM 15.4. *Let $D \subset \mathbb{R}^N$ be a nonempty convex subset, and let $f : D \to \mathbb{R}$ be a continuous function. Then $f(D) = \{f(x) \mid x \in D\}$ is an interval in $\mathbb{R}$.*

PROOF. By Theorem 8.1, it is enough to show $f(D)$ is convex, i.e., if $y_1, y_2 \in f(D)$, then the entire interval from y_1 to y_2 is in $f(D)$. Suppose $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Define a function $g : [0,1] \to \mathbb{R}$ by $g : t \mapsto f((1-t)x_1 + tx_2)$. The function g is well-defined since D is convex; it is continuous since f is continuous, and $g([0,1]) \subset f(D)$. By the usual Intermediate Value Theorem, $g([0,1])$ is an interval in $\mathbb{R}$. Since $g(0) = f(x_1) = y_1$ and $g(1) = f(x_2) = y_2$, this implies that every point between y_1 and y_2 lies in $f(D)$. □

THEOREM 15.5. *(Multivariate Bolzano-Weierstrass Theorem) Every bounded sequence in $\mathbb{R}^N$ admits a convergent subsequence.*

PROOF. When $N = 1$ this is the usual Bolzano-Weierstrass Theorem. We will reduce to this case by repeated passage to subsequences. Let $\{x_n\}$ be a bounded sequence in $\mathbb{R}^N$. By Exercise 15.8, $\pi_1(x_n)$ is a bounded sequence of real numbers. By the usual Bolzano-Weierstrass Theorem, there is a subsequence n_k such that $\pi_1(x_{n_k})$ converges. Since $\{x_n\}$ is bounded, so is its subequence x_{n_k}, and hence so is $\pi_2(x_{n_k})$. Applying Bolzano-Weierstrass again, we get a subsubsequence $x_{n_{k_l}}$ such that $\pi_2(x_{n_{k_l}})$ converges. Continuing in this manner – i.e., passing again to a subsequence for each coordinate in turn – we eventually get a subsub...subsequence (N "subs" in all!) all of whose coordinates converge. This is a subsequence of the original sequence which converges by Proposition 15.2. □

A subset $D \subset \mathbb{R}^N$ is **closed** if for all sequences $\{x_n\}$ with values in D, if $x_n \to x \in \mathbb{R}^N$, then $x \in D$.

Exercise 15.10: Show that a closed interval $I \subset \mathbb{R}$ is a closed subset of $\mathbb{R}$.

PROPOSITION 15.6. *Closed coordinate boxes $[a,b]$ and closed balls $\overline{B}_r(x)$ are closed subsets of $\mathbb{R}^N$.*

PROOF. Let $D = [a,b] = \prod_{i=1}^N [a_i, b_i]$ be a closed coordinate box, and let x_n be a sequence in D converging to $x \in \mathbb{R}^N$. Then for $1 \le i \le N$, $\pi_i(x_n) \to \pi_i(x)$. Since $\pi_i(x_n) \in [a_i, b_i]$, $\pi_i(x) \in [a_i, b_i]$. It follows that $x \in [a,b]$ and thus $[a,b]$ is closed.

Let $D = \overline{B}_r(x_0)$ be a closed ball and let x_n be a sequence in D converging to $x \in \mathbb{R}^N$. By Exercise 15.3, $y \mapsto |x_0 - y|$ is continuous, and thus $|x_0 - x_n| \to |x_0 - x|$. Since $|x_0 - x_n| \le r$ for all n, $|x_0 - x| \le r$, i.e., $x \in \overline{B}_r(x_0)$: $\overline{B}_r(x_0)$ is closed. □

THEOREM 15.7. *(Multivariable Extreme Value and Uniform Continuity) Let $D \subset \mathbb{R}^N$ be closed and bounded.*
a) Every sequence in D admits a subsequence converging to an element of D.
b) Every continuous function $f : D \to \mathbb{R}$ is bounded and attains its maximum and minimum values.
c) Eery continuous function $f : D \to \mathbb{R}$ is uniformly continuous.

PROOF. a) Since D is bounded, by Theorem 15.5, every sequence in D admits a subsequence converging to $x \in \mathbb{R}^N$. Since D is closed, x must lie in D.
b),c) The proofs of Theorems 10.20 and 10.21 which treat the case $D = [a,b]$ adapt easily to the general case. Details are left to the reader. □

2. Complex Numbers and Complex Series

It turns out that the sine, cosine and exponential functions are all very closely related to each other, provided we are willing to work with complex values. In

this section we provide a review of complex numbers and the rudiments of complex power series. This theory can be developed to an amazing extent – further than the theory of real power series, in fact! – but such qualitatively different developments are the subject of another course. Here we just want to develop the theory enough so that we can make sense of plugging complex numbers into power series.

Recall that a complex number is an expression of the form $z = a + bi$. Here a and b are real numbers and i is a formal symbol having the property that $i^2 = -1$.

For many years people had "philosophical difficulties" with complex numbers; indeed, numbers of the form ib were called "imaginary," and the prevailing view was that although they did not exist, they were nevertheless very useful.

From a modern point of view this is neither acceptable (we cannot work with things that don't exist, no matter how useful they may be!) nor necessary: we can define the complex numbers entirely in terms of the real numbers. Namely, we may identify a complex number $a + bi$ with the ordered pair (a, b) of real numbers, and we will define addition and multiplication. Since we would want $(a+bi)+(c+di) = (a+c)+(b+d)i$, in terms of ordered pairs this is just $(a, b) + (c, d) = (a + c, b + d)$. In other words, this is the usual addition of vectors in the plane. The multiplication operation is more interesting but still easy enough to write down in terms of only real numbers: to compute $(a + bi)(c + di)$, we would want to use the distributive law of multiplication over addition and the relation $i^2 = -1$. In other words, we would like $(a + bi) \cdot (c + di) = ac + bci + adi + bdi^2 = (ac - bd) + (ad + bc)i$. Thus in terms of ordered pairs we *define* a multiplication operation

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

Note that with this convention, we may identify real numbers a (i.e., those with $b = 0$) with pairs of the form $(a, 0)$; moreover, what we were calling i corresponds to $(0, 1)$, and now any ordered pair $a + bi$ can be expressed as $(a, 0) + (b, 0) \cdot (0, 1)$.

Exercise: Show that the above operations of addition and multiplication on ordered pairs satisfy all the *field axioms* (P0) through (P9). The resulting structure is called the **complex numbers** and denoted $\mathbb{C}$.

Exercise: Show that because of the relation $i^2 = -1$, $\mathbb{C}$ cannot be endowed with the structure of an ordered field.

Two other important operations on the complex numbers are conjugation and taking the modulus. For any complex number $z = a + bi$, we define its complex conjugate to be $\overline{z} = a - bi$. Conjugation fits in nicely with the rest of the algebraic structure: one has $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ and $\overline{(z_1 z_2)} = \overline{z_1}\ \overline{z_2}$.

For any complex number $z = a + bi$, we define its **modulus** (or norm, or absolute value) to be $|z| = \sqrt{a^2 + b^2}$. This is just the usual norm of an element of $\mathbb{R}^N$ specialized to the case $N = 2$. In particular, we have the triangle inequality

$$\forall z_1, z_2 \in \mathbb{C}, |z_1 + z_2| \leq |z_1| + |z_2|.$$

However, the norm also behaves nicely with respect to the multiplicative structure.

PROPOSITION 15.8. *a) For all $z_1, z_2 \in \mathbb{C}$, $|z_1 z_2| = |z_1||z_2|$.*
b) For all $z \in \mathbb{C}$, $z\overline{z} = |z|^2$.

Exercise: Prove Proposition 15.8.

Finally, we mention that the theory of complex series, and especially, of complex power series, works (at least) as well as the theory of real series. Namely, if $\sum_{n=0}^{\infty} a_n z^n$ is a power series with complex coefficients, then defining $\rho = \limsup |a_n|^{\frac{1}{n}}$, we find that ρ is the radius of convergence of the complex power series in the sense that the series converges for all z with $|z| < \rho$ and diverges for all z with $|z| > \rho$. Especially, if $\sum_n a_n x^n$ is a power series with real coefficients and infinite radius of convergence, then because for a real number x, its absolute value $|x|$ is the same as the modulus of the complex number $x + 0i$, then the power seres $\sum_n a_n z^n$ must converge for all complex numbers z.

3. Elementary Functions Over the Complex Numbers

3.1. The complex exponential function.

Consider the following complex power series:

$$E(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Because the ratio test limit is $\lim_{n\to\infty} \frac{\frac{1}{(n+1)!}}{\frac{1}{n!}} = \lim_{n\to\infty} \frac{1}{n+1} = 0$, the radius of convergence is infinite: the series converges for all complex numbers z.

PROPOSITION 15.9. *For all complex numbers z and w, $E(z+w) = E(z)E(w)$.*

PROOF. Since the series representations of $E(z)$ and $E(w)$ are absolutely convergent, we know that $E(z)E(w)$ is given by the Cauchy product, namely

$$E(z)E(w) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{z^k w^{n-k}}{k!(n-k)!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} z^k w^{n-k} = \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!} = E(z+w).$$

□

Since $E(0) = 1$, we have for all z that

$$E(z)E(-z) = E(z-z) = E(0) = 1,$$

or $E(-z) = \frac{1}{E(z)}$. Note in particular that $E(z)$ is never zero. Restricting attention to real values, since $E : x \mapsto E(x)$ is a continuous function which is never zero and such that $E(0) = 1$, we conclude $E(x) > 0$ for all real x.

3.2. The trigonometric functions.

Let us now turn to the functions $\sin x$ and $\cos x$. Recall that we have already shown that any pair of differentiable functions $S(x)$ and $C(x)$ such that $S'(x) = C(x)$, $C'(x) = -S(x)$, $S(0) = 0$ and $C(0) = 1$ must be equal to their Taylor series and given by the following expansions:

$$S(x) := \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}.$$

$$C(x) := \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}.$$

Of course we would like to say $S(x) = \sin x$ and $C(x) = \cos x$, but we do not want to have to resort to discussions involving angles, lengths of arcs and other such things. We want to see how much can be derived directly from the power series expansions themselves. For instance, we would like the show that $C^2 + S^2 = 1$. Unfortunately, although this identity does hold, showing it directly from the power series expansions involves some rather unpleasant algebra (try it and see).

This is where complex numbers come in to save the day:

PROPOSITION 15.10. *For all real x, we have the following identities:*

$$C(x) = \frac{1}{2}(E(ix) + E(-ix)),$$
$$S(x) = \frac{1}{2i}(E(ix) - E(-ix)),$$
$$E(ix) = C(x) + iS(x).$$

Exercise: Prove Proposition 15.10.

Now we are in business: since the coefficients of $E(z)$ are real, we have $\overline{E(ix)} = E(\overline{ix}) = E(-ix)$ for all real x, hence

$$C(x)^2 + S(x)^2 = |E(ix)|^2 = E(ix)\overline{E(ix)} = E(ix)E(-ix) = E(ix - ix) = E(0) = 1.$$

We're not done yet: we'd like to prove that $S(x)$ and $C(x)$ are periodic functions, whose period is a mysterious number approximately equal to $2 \cdot 3.141592653\ldots$. This can also be worked out from the power series expansions, with some cleverness:

We first claim that there exists $x_0 > 0$ such that $C(x_0) = 0$. Otherwise, since $C(0) = 1 > 0$, we'd have $C(x) > 0$ for all x, hence $S'(x) = C(x) > 0$ for all x, hence S would be strictly increasing on the entire real line. Since $S(0) = 0$, it follows that $S(x) > 0$ for all $x > 0$. Now, if $0 < x < y$, we have

$$S(x)(y - x) < \int_x^y S(t)dt = C(x) - C(y) \leq 2.$$

But now for fixed x and $y > x + \frac{2}{S(x)}$, this gives a contradiction.

LEMMA 15.11. *Let $f : [0, \infty) \to \mathbb{R}$ be continuous such that $f(0) > 0$ and $f(x) = 0$ for some $x > 0$. Then there is a* least *positive number x_0 such that $f(x_0) = 0$.*

PROOF. Left to the reader. □

Now we *define* the number π by $\pi = 2x_0$. where x_0 is the lesat positive number x such that $C(x) = 0$. The relation $C(x)^2 + S(x)^2 = 1$ together with $C(\frac{\pi}{2}) = 0$ shows that $S(\frac{\pi}{2}) = \pm 1$. On the other hand, since $C(x) = S'(x)$ is non-negative on $[0, \frac{\pi}{2}]$, $S(x)$ is increasing on this interval, so it must be that $S(\frac{\pi}{2}) = 1$. Thus $E(\frac{\pi i}{2}) = i$. Using the addition formula for $E(z)$ we recover Euler's amazing identity

$$e^{i\pi} = \left(e^{\frac{i\pi}{2}}\right)^2 = -1,$$

and also $e^{2\pi i} = 1$. In general, $e^{z+2\pi i} = e^z e^{2\pi i} = e^z$, so E is periodic with period $2\pi i$.

Using the periodicity of E and the formula of Proposition 2, we get that for all x $C(x + 2\pi) = C(x)$ and $S(x + 2\pi) = S(x)$.

Since for all real t, $|e^{it}| = 1$, the parameterized curve

$$r(t) = e^{it} = C(t) + iS(t) \iff (C(t), S(t)) = (x(t), y(t))$$

has image contained in the unit circle. We claim that every point on the unit circle is of the form e^{it} for a unique $t \in [0, 2\pi)$. To see this, start at the point $1 = e^{i\cdot 0}$, and consider $t \in [0, \frac{\pi}{2}]$. The function $C : [0, \frac{\pi}{2}] \to \mathbb{R}$ is continuous and decreasing, hence injective, with $C(0) = 1$ and $C(\frac{\pi}{2}) = 0$. By the Intermediate Value Theorem, all values in $[0, 1]$ are assumed for a (necessarily unique, by the injectivity) $t \in [0, \frac{pi}{2}]$, and every point in the first quadrant of the unit circle is of the form (x, y) for a unique $x \in [0, 1]$. By making similar arguments in the intervals $[\frac{\pi}{2}, \pi]$, $[\pi, \frac{3\pi}{2}]$, $[\frac{3\pi}{2}, 2\pi]$ we establish the claim.

Finally, if we grant that by the arclength of the parameterized curve $r(t) = (x(t), y(t))$ from $t = a$ to $t = b$ we mean the integral

$$\int_{t=a}^{b} \sqrt{|\frac{dx}{dt}|^2 + |\frac{dy}{dt}|^2} dt$$

it is easy to show that $C(x) = \cos x$ and $S(x) = \sin x$. Indeed, for $r(t) = (C(t), S(t))$, the arclength integral is

$$\int_{t=0}^{\theta} S^2(t) + C^2(t) dt = \theta,$$

so the point $r(\theta) = (C(\theta), S(\theta))$ really is the point that we arrive at by starting at the point $(1, 0)$ on the unit circle and traversing θ units of arc.

Exercise: Show that for all $x \in (0, \pi/2)$, $S(x) < x < \frac{S(x)}{C(x)}$.
(Show this directly from the power series definitions: no pictures involving the unit circle! Hint: use the power series representation for $\arctan x$.)

LEMMA 15.12. *(DeMoivre) Let $k \in \mathbb{Z}^+$ and $z \in \mathbb{C}$. Then there is $w \in \mathbb{C}$ such that $w^k = z$.*

PROOF. If $z = 0$ we may take $w = 0$. Otherwise, $r := |z| > 0$, so $\frac{z}{r}$ lies on the unit circle and thus $\frac{z}{r} = e^{i\theta}$ for a unique $\theta \in [0, 2\pi)$. We well know (as a consequence of the Intermediate Value Theorem) that every positive real number has a positive kth root, denoted $r^{\frac{1}{k}}$. Thus if $w = r^{\frac{1}{k}} e^{i\frac{\theta}{7}k}$,

$$w^k = (r^{\frac{1}{k}} e^{i\frac{\theta}{7}k})^k = r(e^{i\frac{\theta}{k}})^k = re^{i\theta} = z. \qquad \square$$

Exercise: Let z be a nonzero complex number and k a positive integer. Show that there are preciesly k complex numbers w such that $w^k = z$.

4. The Fundamental Theorem of Algebra

4.1. The Statement and Some Consequences.

THEOREM 15.13. *(Fundamental Theorem of Algebra) Let*

$$P(z) = a_n z^n + \ldots + a_1 z + z_0$$

be a polynomial with complex coefficients and positive degree. Then P has a root in the complex numbers: there is $z_0 \in \mathbb{C}$ such that $P(z_0) = 0$.

Theorem 15.13 is *not* easy to prove, and we defer the proof until the next section. For now we give some important consequences of this seminal result.

COROLLARY 15.14. *Every nonconstant polynomial with complex coefficients factors as a product of linear polynomials. More precisely, let*

$$P(z) = a_n z^n + \ldots + a_1 z + a_0, \text{ with } a_n \neq 0.$$

Then there are $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ (not necessarily distinct) such that

$$P(z) = a_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n). \tag{117}$$

PROOF. First observe that if the result holds for $P(z)$ then it holds for $\beta P(z)$ for any $\beta \in \mathbb{C} \setminus \{0\}$. It is therefore no loss of generality to assume that the leading coefficient a_n of $P(z)$ is equal to 1. Let us do so.

We now prove the result by induction on n, the degree of the polynomial P.
Base Case ($n = 1$): A degree one polynomial with leading coefficient 1 is of the form $z + a_0 = z - \alpha_1$, with $\alpha_1 = -a_0$.
Induction Step: Let $n \in \mathbb{Z}^+$, suppose the result holds for all polynomials of degree n, and let $P(z)$ be a polynomial of degree $n + 1$. By Theorem 15.13, there is $z_0 \in \mathbb{C}$ such that $P(z_0) = 0$. By the Root-Factor Theorem, we may write $P(z) = (z - z_0)Q(z)$, with $Q(z)$ a polynomial of degree n and leading coefficient 1. By induction, $Q(z) = (z - \alpha_1) \cdots (z - \alpha_n)$, so putting $\alpha_{n+1} = z_0$ we get

$$P(z) = (z - \alpha_1) \cdots (z - \alpha_n)(z - \alpha_{n+1}). \qquad \square$$

More generally, let F be a field: that is, a set endowed with two binary operations denoted $+$ and $\cdot$ and satisfying the field axioms (P0) through (P9) from Chapter 1. We say that F is **algebraically closed** if every nonconsant polynomial $P(x)$ with coefficients in F has a root in F, i.e., there is $x_0 \in F$ such that $P(x_0) = 0$. In this terminology, the Fundamental Theorem of Algebra asserts precisely that the complex field $\mathbb{C}$ is algebraically closed.

Exercise: Let F be an algebraically closed field. Show that the conclusion of Corollary 15.14 holds for F: that is, every nonconstant polynomial factors as a product of linear polynomials.

Since every real number is, in particular, a complex number, Corollary 15.14 applies in particular to polynomials over $\mathbb{R}$: if $P(x) = a_n x^n + \ldots + a_1 x + a_0$ with $a_n \neq 0$, then there are complex numbers $\alpha_1, \ldots, \alpha_n$ such that

$$P(x) = a_n(x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n).$$

But since the coefficients of P are real, it is natural to ask whether or to what extent some or all of the roots α_i must also be real. Recall that we need not have any real roots (that is, the field $\mathbb{R}$ is *not* algebraically closed), for any $n \in \mathbb{Z}^+$, the polynomial $P_n(x) = (x^2 + 1)^n$ is positive for all real x, so has no real roots. And indeed, its factorization over $\mathbb{C}$ is $(x^2 + 1)^n = (x + i)^n(x - i)^n$.

However, the polynomials P_n all had even degree. Recall that, as a consequence

of the Intermediate Value Theorem, every polynomial of odd degree has at least one real root (and need not have more than one, as the family of examples $xP_n(x)$ shows). So there is some relation between the *parity* (i.e., the evenness or oddness) of the degree of a real polynomial and its real and complex roots. This observation can be clarified and sharpened in terms of the operation $z = a + bi \mapsto a - bi = \overline{z}$ of **complex conjugation**.

LEMMA 15.15. *a) For $z, a_0, \ldots, a_n \in \mathbb{C}$, we have*

$$\overline{a_n z^n + \ldots + a_1 z + a_0} = \overline{a_n}\overline{z}^n + \ldots + \overline{a_1}\overline{z} + \overline{a_0}.$$

b) Thus, if $a_0, \ldots, a_n \in \mathbb{R}$, we have

$$\overline{a_n z^n + \ldots + a_1 z + a_0} = a_n \overline{z}^n + \ldots + a_1 \overline{z} + a_0.$$

c) Let $a_0, \ldots, a_n \in \mathbb{R}$ and put $P(z) = a_n z^n + \ldots + a_1 z + a_0$. If $z_0 \in \mathbb{C}$ is such that $P(z_0) = 0$, then also $P(\overline{z_0}) = 0$.

PROOF. We have already observed that for any $z_1, z_2 \in \mathbb{C}$, $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ and $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$. Keeping these identities in mind, the proof becomes a straightforward exercise which we leave to the reader. □

It is part c) of Lemma 15.15 that is important for us: this well-known result often goes by the description "The complex roots of a real polynomial occur in conjugate pairs." To see why this is relevant, consider the following extremely simple – but extremely important – result.

LEMMA 15.16. *For a complex number z, $z \in \mathbb{R} \iff \overline{z} = z$.*

PROOF. If $z \in \mathbb{R}$, then $z = z + 0i$, so $\overline{z} = z - 0i = z$. Conversely, let $z = a + bi$. IF $a - bi = \overline{z} = z = a + bi$, then $(2i)b = 0$. Multiplying through by $(2i)^{-1} = \frac{-i}{2}$, we get $b = 0$, so $z = a + 0b \in \mathbb{R}$. □

Let us say that $\alpha \in \mathbb{C}$ is **properly complex** if it is not a real number. If $P(x)$ is a polynomial with real coefficients and has the properly complex number α as a root, then by Lemma 15.16 it also has the (distinct!) properly complex number $\overline{\alpha}$ as a root. By the Root-Factor Theorem, we may write $P(z) = (z - \alpha)(z - \overline{\alpha})P_2(z)$. Now **wake up!** – something interesting is about to happen. Namely, if we write

$$P_1(z) = (z - \alpha)(z - \overline{\alpha}),$$

then we CLAIM both $P_1(z)$ and $P_2(z)$ have real coefficients, so we have obtained a factorization of real polynomials

$$P(x) = P_1(x)P_2(x).$$

For $P_1(z)$ we need only write $\alpha = a + bi$ and multiply it out:

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

For $P_2(z)$ we have to argue a bit more abstractly. Namely, for polynomials over $\mathbb{R}$ (and really, over any field) we can always perform division with remainder: there are unique real polynomials $Q(x)$, $R(x)$ such that

$$P(x) = Q(x)P_1(x) + R(x), \deg R(x) < \deg P_1(x). \tag{118}$$

We claim $R(x) \equiv 0$ (i.e., it is the zero polynomial). To see this, we use the *uniqueness* part of the division algorithm in a slightly sneaky way: namely, consider $P(x)$

and $P_1(x)$ as polynomials with complex coefficients and perform the division algorithm there: there are unique complex polynomials, say $Q_{\mathbb{C}}(x)$ and $R_{\mathbb{C}}(x)$, such that

$$P(x) = Q_{\mathbb{C}}(x)P_1(x) + R_{\mathbb{C}}(x), \deg R_{\mathbb{C}} < \deg Q_{\mathbb{C}}. \tag{119}$$

Here's the point: on the one hand, real polynomials *are* complex polynomials, so comparing (118) and (119) we deduce $Q(x) = Q_{\mathbb{C}}(x)$ and $R(x) = R_{\mathbb{C}}(x)$. On the other hand, the identity $P(x) = P_1(x)P_2(x)$ of complex polynomials shows that we may take $Q_{\mathbb{C}}(x) = P_2(x)$ and $R_{\mathbb{C}}(x) \equiv 0$. Putting these together, we get $Q(x) = P_2(x)$ and $R(x) \equiv 0$, so indeed $P(x) = P_1(x)P_2(x)$ is a factorization of real polynomials.

A positive degree polynomial $P(x)$ over a field F is called **irrreducible** if it cannot be factored as $P(x) = P_1(x)P_2(x)$ with $\deg P_1, \deg P_2 < \deg P$. (This last condition is there to prevent trivial factorizations like $x^2 + 1 = (2) \cdot (\frac{1}{2}x^2 + \frac{1}{2}.)$ A polynomial of positive degree which is not irreducible is called **reducible**.

Exercise: Let F be a field, and let $P(x)$ be a polynomial with coefficients in F.
a) Show that if P is irreducible, it has no roots in F.
b) Suppose P has degree 2 and no roots in F. Show that P is irreducible.
c) Suppose P has degree 3 and no roots in F. Show that P is irreducible.
d) For each $n \geq 4$, exhibit a degree n polynomial with coefficients in $\mathbb{R}$ which is reducible, but has no real roots.

THEOREM 15.17. *Let $P(x)$ be a real polynomial of degree $n \geq 1$.*
a) There are natural numbers $r, s \in \mathbb{N}$ such that $r + 2s = n$, linear polynomials $L_1(x), \ldots, L_r(x)$ and irreducible quadratic polynomials $Q_1(x), \ldots, Q_s(x)$ such that

$$P(x) = L_1(x) \cdots L_r(x)Q_1(x) \cdots Q_s(x).$$

b) If n is odd, then P has at least one real root.

PROOF. It's harmless to assume that P has leading coefficient 1, and we do so.
a) We go by strong induction on n. When $n = 1$, $P(x)$ is a linear polynomial, so we may take $r = n = 1$, $s = 0$ and $P(x) = L(x)$. Suppose $n \geq 2$ and the result holds for all polynomials of degree less than n, and let $P(x)$ be a real polynomial of degree n. By the Fundamental Theorem of Algebra, $P(x)$ has a complex root α. If $\alpha \in \mathbb{R}$, then by the Root-Factor Theorem $P(x) = (x - \alpha)P_2(x)$ and we are done by induction. If α is properly complex, $Q(x) = (x - \alpha)(x - \overline{\alpha})$ is a real, irreducible quadratic polynomial and $P(x) = Q(x)P_2(x)$, and again we are done by induction.
b) Since $r + 2s = n$, if n is odd we cannot have $r = 0$. Thus P has at least one linear factor and thus at least one real root. □

The partial fractions decomposition rests on the foundation of Theorem 15.17.

Exercise:[1] Let $f(z) = a_n z^n + \ldots + a_1 z + a_0$ be a complex polynomial with positive degree. Suppose that for all $z \in \mathbb{C}$, we have $z \in \mathbb{R} \iff f(z) \in \mathbb{R}$.
a) Show: $a_0, \ldots, a_n \in \mathbb{R}$.
b) Show: $n = 1$.

[1]Taken from `http://math.stackexchange.com/questions/1332641/`

4.2. Proof of the Fundamental Theorem of Algebra.

We now give a proof of Theorem 15.13, closely following W. Rudin [**R**, Thm. 8.8].

Let $P(z)$ be a polynomial with complex coefficients, degree $n \geq 1$, and leading coefficient a_n. We want to show that $P(z)$ has a complex root; certainly this holds iff $\frac{1}{a_n}P(z)$ has a complex root, so it is no loss of generality to assume that the leading coefficient is 1 and thus

$$P(z) = 1 + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0, a_i \in \mathbb{C}.$$

Let

$$\mu = \inf_{z\in\mathbb{C}} |P(z)|.$$

Thus μ is a non-negative real number and our job is to show (i) that μ is actually attained as a value of M and (ii) $\mu = 0$.
Step 1: Since for $z \neq 0$,

$$|P(z)| = |z|^n \left(1 + \frac{|a_{n-1}|}{|z|} + \ldots + \frac{|a_0|}{|z|^n}\right),$$

it follows that

$$\lim_{z\to\infty} |P(z)| = \infty \cdot 1 = \infty.$$

Thus there is $R > 0$ such that for all $z \in \mathbb{C}$ with $|z| > R$, $|P(z)| > |P(0)|$. By Theorem 15.7, the continuous function $|P(z)|$ assumes a minimum value on the closed, bounded set $\{z \in \mathbb{C} \mid |z| \leq R\}$, say at z_0. But R was chosen so that $|P(z)| > |P(0)| \geq |P(z_0)|$ for all z with $|z| > R$, so altogether $|P(z)| \geq |P(z_0)|$ for all $z \in \mathbb{C}$ and thus $\mu = \inf_{z\in\mathbb{C}} |P(z)| = |P(z_0)|$.
Step 2: Seeking a contradiction, we suppose $\mu > 0$. Define $Q : \mathbb{C} \to \mathbb{C}$ by $Q(z) = \frac{P(z+z_0)}{P(z_0)}$. Thus Q is also a degree n polynomial function, $Q(z) = 1$, and by minimality of z_0, $|Q(z)| \geq 1$ for all $z \in \mathbb{C}$. We may write $Q(z) = 1+b_k z^k+\ldots+b_n z^n$ with $b_k \neq 0$ for some $1 \leq k \leq n$. Let $w \in \mathbb{C}$ be such that

$$w^k = \frac{-|b_k|}{b_k};$$

the existence of such a w is guaranteed by Lemma 15.12. Then for $r \in (0, \infty)$,

$$Q(rw) = 1 + b_k r^k w^k + b_{k+1} r^{k+1} w^{k+1} + \ldots + b_n r^n w^n$$

$$= 1 - r^k \left(|b_k| - r w^{k+1} b_{k+1} - \ldots - r^{n-k} w^n b_n\right) = 1 - r^k \left(|b_k| + C(r)\right),$$

where we have set $C(r) = -rw^{k+1}b_{k+1} - \ldots - r^{n-k}w^b b_n$. Thus

$$|Q(rw)| = |1 - r^k(|b_k| + C(r))| \leq |1 - r^k|b_k|| + |r^k C(r)|.$$

As r approaches 0 from the right, $r^k|b_k|$ and $C(r)$ both approach 0. Thus for sufficiently small r, we have $r^k|b_k| < 1$ and $|C(r)| < |b_k|$ and then

$$|Q(rw)| \leq |1 - r^k|b_k|| + |r^k C(r)| = 1 - r^k(|b_k| - |C(r)|) < 1.$$

This contradicts the fact that $\min_{z\in\mathbb{C}} |Q(z)| = 1$ and completes the proof.

CHAPTER 16

Foundations Revisited

The reader should picture a street mime juggling non-existent balls. As the mime continues, the action of juggling slowly brings the balls into existence, at first in dim outline and then into solid reality. — T.W. Körner[1]

An ordered field F is **Dedekind complete** if every nonempty subset which is bounded above has a least upper bound (or "supremum").

Exercise 16.0: Show that an ordered field is Dedekind complete iff every nonempty subset which is bounded below has a greatest upper bound (or "infimum").

Our initial definition of $\mathbb{R}$ was precisely that it was a Dedekind complete ordered field. Practically speaking, this is a great foundation for honors calculus and real analysis, because it contains all the information we need to know about $\mathbb{R}$.

In other words, we have put a neat little black box around our foundational problems. Real analysis works perfectly well without ever having to look in the box. But curiosity is a fundamental part of mathematics, and at some point most of us will want to look in the box. This chapter is for those who have reached that point, i.e., who want to understand a proof of the following theorem.

THEOREM 16.1. *(Black Box Theorem)*
a) There is a Dedekind complete ordered field.
b) If F_1 and F_2 are Dedekind complete ordered fields, they are isomorphic*: that is, there is a bijection $\Phi : F_1 \to F_2$ such that:*
(i) For all $x, y \in F_1$, $\Phi(x + y) = \Phi(x) + \Phi(y)$.
(ii) For all $x, y \in F_1$, $\Phi(xy) = \Phi(x)\Phi(y)$.
(iii) For all $x, y \in F_1$, $x \leq y \iff \Phi(x) \leq \Phi(y)$.
c) The isomorphism Φ of part b) is unique*: there is exactly one such map between any two Dedekind complete ordered fields.*

The Black Box Theorem explains why we never needed any further axioms of $\mathbb{R}$ beyond the fact that it is a Dedekind complete ordered field: there is exactly one such structure, up to isomorphism.[2]

[1]Thomas William Körner, 1946–

[2]The student unfamiliar with the notion of "isomorphism" should think of it as nothing else than a relabelling of the points of $\mathbb{R}$. For instance consider the x-axis $\mathbb{R}_x = \{(a, 0) \mid a \in \mathbb{R}\}$ and the y-axis $\mathbb{R}_y = \{(0, a) \mid a \in \mathbb{R}\}$ in the plane. These are two copies of $\mathbb{R}$. Are they "the same"? Not in a hard-nosed set-theoretic sense: they are different subsets of the plane. But they are essentially the same: the bijection Φ which carries $(a, 0) \mapsto (0, a)$ preserves the addition, multiplication and order relation. So really we have two slightly different presentations of the same

We will prove the Black Box Theorem...eventually. But rather than taking the most direct possible route we broaden our focus to a study of the structure of ordered fields, not just $\mathbb{Q}$ and $\mathbb{R}$.

1. Ordered Fields

1.1. Basic Definitions.

In this section we revisit the considerations of §1.2 from a somewhat different perspective. Before we listed certain **ordered field axioms**, but the perspective there was that we were collecting true, and basic, facts about the real numbers for use in our work with them. This time our perspective is to study and understand the collection of all ordered fields. One of our main goals is to *construct* the real numbers $\mathbb{R}$ in terms of the rational numbers $\mathbb{Q}$ and to understand this in terms of a more general process, **completion**, which can be applied in any ordered field.

A **field** is a set F endowed with two binary operations $+$ and $\cdot$ which satisfy all of the field axioms (P0) through (P9). To a first approximation, these axioms simply encode the usual rules of addition, subtraction, multiplication and division of numbers, so any field can be thought of as a kind of "generalized number system". The most important basic examples are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. But there are other examples which seem farther removed from the "usual numbers": e.g. finite fields like $\mathbb{F}_2 = \{0, 1\}$ are smaller than what we normally think of as a number system, whereas the set $\mathbb{R}(t)$ of all rational functions (with real coefficients) is a field whose elements are naturally regarded as functions, not as numbers.

Field theory is an active branch of mathematical research, with several texts and thousands of papers devoted to it (e.g. [**FT**]). Nevertheless the very simple properties of fields established in § 1.2.1 will be sufficient for our needs here, in part because we are not interested in fields *per se* but rather **ordered fields**. An ordered field is a field equipped with the additional structure of a **total order relation**, namely a binary relation $\leq$ which satisfies:

(TO1) Reflexivity: for all $x \in F$, $x \leq x$.
(TO2) Antisymmetry: for all $x, y \in F$, if $x \leq y$ and $y \leq x$, then $x = y$.
(TO3) Transitivity: for all $x, y, z \in F$, if $x \leq y$ and $y \leq z$ then $x \leq z$.
(TO4) Totality: for all $x, y \in F$, either $x \leq y$ or $y \leq x$.

Given a total order, we may define a **strict ordering** $x < y$ by $x \leq y$ but $x \neq y$. This is (as we well know!) natural and useful. From an abstract perspective the key remark is that it is essentially equivalent to a total ordering. Namely, any strict ordering coming from a total ordering satisfies the following modified versions of the above axioms:

(TR) Trichotomy: for all $x, y \in F$, exactly one of the following holds:

$$x < y; \ x = y; \ \text{or } x > y.$$

---

essential structure. An arbitrary isomorphism is no more than this, except that the "relabelling map" Φ might be more complicated.

(TO3): Transitivity: for all $x, y, z \in F$, if $x < y$ and $y < z$ then $x < z$.

Exercise 16.1: Let $(F, \leq)$ be a total order relation on the set F.
a) Show that the associated strict ordering satisfies (TR) and (TO3).
b) Let $<$ be a binary relation on F satisfying (TR) and (TO3). Define $x \leq y$ by $x < y$ or $x = y$. Show that this gives a total ordering on F.
c) Show that the processes of passing from a total ordering to its strict ordering and from a strict ordering to its total ordering are mutually inverse: doing one followed by the other, in either order, brings us back where we started.

In light of the preceding exercise we may (and shall) in fact pass from a total ordering to its strict ordering and back without explicit comment.

An **ordered field** is a field $(F, +, \cdot)$ equipped with a total ordering $\leq$ which is **compatible** with the field structure in the sense of satisfying the following two familiar axioms:

(P11) If $x, y > 0$, then $x + y > 0$.
(P12) If $x, y > 0$, then $xy > 0$.

We refer the reader back to § to 1.2.2 for the most elementary consequences of these axioms, e.g. Proposition 1.11.

Now we introduce a new idea. Let F and F' be ordered fields. (To be formally corrrect, we should speak of "the ordered field $(F, +, \cdot, <)$", and similarly for F'. In practice this extra notation weighs us down without any advantages in clarity or precision.) An **ordered field homomorphism** is a map $f : F \to F'$ satisfying all of the following:

(FH1) For all $x, y \in F$, $f(x + y) = f(x) + f(y)$.
(FH2) For all $x, y \in F$, $f(xy) = f(x)f(y)$.
(FH3) $f(1) = 1$.
(OFH) For all $x \in F$, if $x > 0$, then $f(x) > 0$.

Exercise 16.2: Show that for any homomorphism of ordered fields, $f(0) = 0$.

Because of Exercise 16.2 it is natural to wonder whether we really need (FH3) or whether in fact $f(1) = 1$ follows automatically from the other axioms. The answer is a resounding *no*. Indeed, in the absence of (FH3), for any ordered fields F and F' we could get a homomorphism betweeen them simply by mapping every $x \in F$ to 0. This is not an interesting map,[3] and as we shall see the theory is cleaner for not allowing it.

If F and F' are fields, then a **field homomorphism** is a map $f : F \to F'$ sayisfying (FH1) through (FH3).

PROPOSITION 16.2. *Every field homomorphism $f : F \to F'$ is injective.*

[3]Nothing can come of nothing. – King Lear

PROOF. Assume to the contrary that there are $x \neq y \in F$ with $f(x) = f(y)$. Then $f(x-y) = f(x) - f(y) = 0$. Since $x \neq y$, $x - y \neq 0$, and thus we have a multiplicative inverse $\frac{1}{x-y}$. Then

$$1 = f(1) = f((x-y)\frac{1}{x-y}) = f(x-y)f(\frac{1}{x-y}) = 0 \cdot f(\frac{1}{x-y}) = 0,$$

contradicting axiom (P0). Note the use of (FH3)! □

A **field isomorphism** is a field homomorphism $f : F \to F'$ which has an inverse: i.e., there is a field homomorphism $f' : F' \to F$ such that $f' \circ f = 1_F$, $f \circ f' = 1_{F'}$.

Exercise 16.3: Show that for a field homomorphism $f : F \to F'$, TFAE:
(i) f is a field isomorphism.
(ii) f is bijective.[4]
(iii) f is surjective.

Exercise 16.4: Let $f : F \to F'$ be a field homomorphism. Show that f induces a field isomorphism $F \to f(F)$.

A **subfield** of a field K is a subset of K which is a field under the binary operations $+$ and $\cdot$ on K. For example, $\mathbb{Q}$ is a subfield of $\mathbb{R}$ and $\mathbb{R}$ is a subfield of $\mathbb{C}$.

Exercise 16.5: Let K be a field and F a subset of K. Show that F is a subfield iff all of the following hold:
(SF1) $0, 1 \in F$. (In particular F is nonempty.)
(SF2) For all $x, y \in F$, $x + y, x - y, xy \in F$.
(SF3) If $x \in F$ and $x \neq 0$, then $\frac{1}{x} \in F$.

Thus whenever we have a field homomorphism $f : F \to F'$, we get an isomorphic copy of F as a subfield of F', namely $f(F)$. Because of this it is safe to think of of F itself as a subfield of F', and this perspective is often useful.

Exercise 16.6: Let K be an ordered field, and let $F \subset K$ be a subfield. Show that restricting the total ordering $\leq$ on K to F endows F with the structure of an ordered field in such a way that the inclusion map $\iota : F \to K$ is a homomorphism of ordered fields.

We say that an element a of a field F is **a sum of squares** if there is $n \in \mathbb{Z}^+$ and $x_1, \ldots, x_n \in F$ such that $a = x_1^2 + \ldots + x_n^2$. Notice that in any field 0 and 1 are sums of squares.

Exercise 16.7: Let F be a field.
a) Let $x \in F$ be a sum of squares. Show that for any compatible ordering $\leq$ on F,

[4]Some philistines take this as the definition of a field isomorphism. Without getting into too much detail about it, this is really the wrong definition. It happens to be equivalent to the right definition for fields, but it has an analogue for other types of isomorphisms of mathematical structure which is not always true. E.g. this definition suggests that an isomorphism of topological spaces should be a continuous bijection, and this is truly weaker than the correct definition, namely a continuous map between topological spaces which admits a continuous inverse.

$x \geq 0$. b) Suppose that there is a nonzero element $x \in F$ such that both x and $-x$ are sums of squares. Show that there *does not* exist an ordering $\leq$ on F compatible with the field structure.
c) Deduce that the field $\mathbb{C}$ of complex numbers admits no compatible ordering.

Exercise 16.8: Let F be a field.
a) Suppose that F has the following property: for each nonzero $x \in F$, exactly one of x and $-x$ is a sum of squares. Show that F admits a unique compatible ordering: namely $x \geq 0$ iff x is a sum of squares.
b) Show that the standard ordering on $\mathbb{R}$ is the only compatible ordering.
c) Show that the standard ordering on $\mathbb{Q}$ is the only compatible ordering.

Exercise 16.9: A field F is **formally real** if -1 is not a sum of squares in F.
a) Suppose that F admits a compatible ordering. Show that F is formally real. (Hint: show the contrapositive.)
b) We suppose that $1 + 1 \neq 0 \in F$. (If not, then $-1 = 1$ is a sum of squares, so by part a) F admits no compatible ordering.) remarkable theorem of Artin-Schreier asserts for $x \in F$, the following are equivalent:
(i) For every compatible ordering $\leq$ on F, $x \leq 0$.
(ii) x is a sum of squares.
Deduce from this theorem that if F is formally real then it admits at least one ordering. (Hint: (i) holds vacuously if there are no compatible orderings!)

PROPOSITION 16.3. *For every ordered field F, there is a unique ordered field homomorphism $\iota : \mathbb{Q} \to F$.*

PROOF. Step 1: Since F can be ordered, the sum of positive elements is positive: in particular the sum of positive elements is not zero. Thus for any positive integer n, $1 + \ldots + 1$ (n times) is not zero. The map ι which sends 0 to 0 and each positive integer to $\iota(n) = 1 + \ldots + 1$ (n times) therefore gives an injective map $\mathbb{N} \to F$. We can extend this map to negative integers by mapping $-n$ to the additive inverse of $1 + \ldots + 1$ (n times). Further, for each positive integer n, since $\iota(n) \neq 0$, it has a multiplicative inverse $\frac{1}{\iota(n)}$, and we may map $\frac{m}{n} \mapsto \iota(m)\frac{1}{\iota(n)}$. It is now straightforward (but not completely trivial) to check that this map $\iota : \mathbb{Q} \to F$ is a field homomorphism: we leave this verification to the reader.
Step 2: In fact we had no choice in the matter: the map ι is the *unique* field homomorphism from $\mathbb{Q}$ to F. We leave this to the reader.
Step 3: We must check that if $x = \frac{m}{n} \in \mathbb{Q}$ is positive, then so is $\iota(a) = \iota(m)\frac{1}{\iota(n)}$. But this is easy: $\iota(m)$ is $1 + \ldots + 1$, so it is a sum of positive elements and is thus positive. Similarly $\iota(n)$ is positive, and thus so is its reciprocal $\frac{1}{\iota(n)}$. Finally, $\iota(\frac{m}{n}) = \iota(m)\frac{1}{\iota(n)}$ is a product of two positive elements, hence positive. □

In light of this result, for any ordered field F, we may view $\mathbb{Q}$ as a subfield.

1.2. Some Topology of Ordered Fields.

Let $(K, \leq)$ be a linearly ordered set, and let $F \subset K$ be a subset. We say that F is **cofinal** in K if for all $x \in K$, there is $y \in F$ with $y > x$. In fact this is precisely the concept that in the case of subsets of $\mathbb{R}$ we call **unbounded above**. We use this terminology as an indication that this property can behave a bit differently

in a non-Archimedean field.

By definition, an ordered field K is Archimedean if $\mathbb{Z}$ is cofinal in K. Equivalently, K is Archimedean iff $\mathbb{Q}$ is cofinal in K.

Let F be a subfield of the ordered field F. We say that F is **dense** in K if for all $x < y \in K$, there is $z \in F$ such that $x < z < y$.

LEMMA 16.4. *Let K be an ordered field, and let F be a subfield of K. If F is dense in K, then F is cofinal in K.*

PROOF. We show the contrapositive: suppose F is not cofinal in K: there is $x \in K$ such that for all $y \in F$, $y \leq x$. Then the interval $(x, x+1)$ contains no points of F, so F is not dense in K. □

More generally, let $f : F \to F'$ be a homomorphism of ordered fields. We say that f is **cofinal** if the image $f(F)$ is a cofinal subfield of F'. We say that f is **dense** if the image $f(F)$ is a dense subfield of F'.

Exercise 16.10: Let K be a subfield of F.
a) Suppose that for every $\alpha \in F$, there is a sequence $\{x_n\}$ of elements of K such that $x_n \to \alpha$. Show that K is a dense subfield of F.
b) Does the converse of part a) hold? (Hint: no, but counterexamples are not so easy to come by.)

LEMMA 16.5. *For a homomorphism $f : F \to F'$ of ordered fields, TFAE:*
(i) f is cofinal.
(ii) For every positive $\epsilon' \in F'$, there is a positive $\delta \in F$ such that $f(\delta) < \epsilon'$.

Exercise 16.11: Prove Lemma 16.5. (Hint: take reciprocals!)

LEMMA 16.6. *For an ordered field F, the following are equivalent:*
(i) F is Archimedean.
(ii) $\mathbb{Q}$ is a dense subfield of F.

PROOF. (i) $\implies$ (ii): Suppose F is Archimedean and let $x < y \in F$. Let $n \in \mathbb{Z}^+$ be such that $\frac{1}{y-x} < n$; then $0 < \frac{1}{n} < y - x$, so

$$x < x + \frac{1}{n} < y.$$

(ii) $\implies$ (i): If $\mathbb{Q}$ is dense in F then by Lemma 16.4, $\mathbb{Q}$ is cofinal in F. □

Let S be a subset of an ordered field F. For $s \in S$, we say that S is **discrete at s** if there is a positive $\epsilon \in F$ such that $(s - \epsilon, s + \epsilon) \cap S = \{s\}$. We say that S is **discrete** if it is discrete at s for all $s \in S$. An ordered field homomorphism $f : F \to F'$ is **discrete** if $f(F)$ is a discrete subset of F'.

PROPOSITION 16.7. *(Dorais's Dichotomy) Let $f : F \to F'$ be an ordered field homomorphism. Then* exactly one *of the following holds:*
(i) f is cofinal.
(ii) f is discrete.

PROOF. We will use the characterization of cofinality from Lemma 16.5.

Suppose f is cofinal. Then $0 \in f(F)$ and for every positive $\epsilon' \in F'$, there is a

positive $\delta \in F$ such that $f(\delta) \in (-\epsilon, \epsilon)$. It follows that $f(F)$ is not discrete.

Suppose f is not cofinal: there is $\epsilon' \in F'$ such that $(-\epsilon', \epsilon') \cap f(F) = \{0\}$: $f(F)$ is discrete at 0. Now let $x \in F$. Suppose there is $y \in F$ such that $x - \epsilon < y < x + \epsilon$. Then $y - x \in (\epsilon, \epsilon)$. By what we've just seen we must have $y - x = 0$, i.e., $y = x$. Thus $f(F)$ is discrete at each of its elements, so it is discrete. □

For any element x in an ordered field, we can define $|x|$ in the usual way, i.e., x if $x \geq 0$ and $-x$ if $x \leq 0$.

Exercise 16.12: Let $f : F \to F'$ be an ordered field homomorphism. Show that for all $x \in F$, $|f(x)| = f(|x|)$. (The absolute value on the left hand side is taking place in F'; the one on the right hand side is taking place in F.)

It is now possible to carry over our definitions of convergent sequencess, Cauchy sequences and continuous functions to the context of ordered fields and homomorphisms between them. We repeat the basic definitions. In many cases the proofs are exactly the same as in the case $F = \mathbb{R}$ treated in loving detail in this text; when this is so we leave the proofs to the reader. However, there are some things which *do not carry over* to the context of all ordered fields, and we treat these in some detail.

First recall that for any set S, a sequence $\{x_n\}$ in S is given by a function $f : \mathbb{Z}^+ \to S$; we write x_n in place of $f(n)$.

Let $\{x_n\}$ be a sequence in an ordered field F and let $x \in F$. We say that x_n **converges to** x and write $x_n \to x$ if: for all positive $\epsilon \in F$, there is $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|x_n - x| < \epsilon$. We say that a sequence $\{x_n\}$ is **convergent** if $x_n \to x$ for some $x \in F$.

We observe that this is verbatim the same as the definition in $\mathbb{R}$. At the same itme, the fact that our "small ϵ" is now an element of the ordered field F rather than (necessarily) a real number has certain surprising implications. The following exercises exhibits one.

Exercise 16.13: For an ordered field F, show that the following are equivalent:
(i) The sequence $\{\frac{1}{n}\}_{n=1}^{\infty}$ is convergent in F.
(ii) F is Archimedean.

LEMMA 16.8. *Let $\{x_n\}$ be a sequence in an ordered field F, and let $x, y \in F$. If $x_n \to x$ and $x_n \to y$, then $x = y$.*

PROOF. Left to the reader. □

Thus if $x_n \to x$ we may call x **the limit** of the sequence: it is unique.

A sequence $\{x_n\}$ in an ordered field F is **Cauchy** if for all positive $\epsilon \in F$ there is $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$, $|x_m - x_n| < \epsilon$.

PROPOSITION 16.9. *A Cauchy sequence in an ordered field which admits a convergent subsequence is itself convergent.*

PROOF. Let F be an ordered field, let $\{x_n\}$ be a Cauchy sequence in F, and let $\{x_{n_k}\}$ be a subsequence converging to $x \in F$. Let $\epsilon > 0$. Since $\{x_n\}$ is Cauchy,

there is $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$, $|x_m - x_n| < \frac{\epsilon}{2}$. Also there is $K \geq N$ such that for all $k \geq K$, $|x_{n_k} - x| < \frac{\epsilon}{2}$. If $n \geq K$, then since $n_K \geq K$ we have

$$|x_n - x| \leq |x_n - x_{n_K}| + |x_{n_K} - x| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so $x_n \to x$. □

LEMMA 16.10. *A convergent sequence in an ordered field is Cauchy.*

PROOF. Left to the reader. □

On the other hand, in an arbitrary ordered field a Cauchy sequence need not be convergent. For instance the "Babylonian sequence" $\{x_n\}$ of §VII.2.2 is a sequence of rational numbers converging to $\sqrt{2} \in \mathbb{R}$. Since $\sqrt{2} \notin \mathbb{Q}$ and limits in $\mathbb{R}$ are unique, viewed as a sequence *in the ordered field* $\mathbb{Q}$, the sequence $\{x_n\}$ is not convergent. It is however Cauchy. More generally we have the following result.

PROPOSITION 16.11. *Let $f : F \to F'$ be an ordered field homomorphism, and let $\{x_n\}$ be a sequence in F. If the sequence $\{f(x_n)\}$ is Cauchy in F', then $\{x_n\}$ is Cauchy in F.*

PROOF. We show the contrapositive: suppose $\{x_n\}$ is *not* Cauchy in F. Then there is some positive $\epsilon \in F$ such that for all $N \in \mathbb{Z}^+$ there are $m, n \geq N$ such that $|x_m - x_n| \geq \epsilon$. Then

$$f(\epsilon) \leq f(|x_m - x_n|) = |f(x_m - x_n)| = |f(x_m) - f(x_n)|.$$

Since $\epsilon' = f(\epsilon)$ is a positive element of F', this shows that $\{f(x_n)\}$ is not Cauchy. □

Thus we get a method for producing nonconvergent Cauchy sequences in an ordered field F: find a sequence $\{x_n\}$ in F and a homomorphism $f : F \to F'$ such that $f(x_n)$ converges to an element of $F' \setminus f(F)$. In fact every nonconvergent Cauchy sequence arises this way, as we will see later on.

Let F and F' be ordered fields, and let $f : F \to F'$ be a map between them. We say that f is **continuous at** $c \in F$ if for all positive $\epsilon' \in F'$, there is a positive $\delta \in F$ such that for all $x \in F$ with $|x - c| < \delta$, $|f(x) - f(c)| < \epsilon'$. We say that $f : F \to F'$ is **continuous** if it is continuous at every $c \in F$. We say that $f : F \to F'$ is **uniformly continuous** if for all positive $\epsilon' \in F'$ there is a positive $\delta \in F$ such that for all $x, x' \in F$, if $|x - x'| < \delta$ then $|f(x) - f(x')| < \epsilon'$.

LEMMA 16.12. *Let $f : F \to F'$ be a function.*
a) If f is continuous and x_n is a sequence in F which converges to $x \in F$, then $f(x_n) \to f(x)$ in F'.
b) If f is uniformly continuous and $\{x_n\}$ is a Cauchy sequence in F, then $\{f(x_n)\}$ is a Cauchy sequence in F'.

PROOF. Left to the reader. □

Warning: When $F = F' = \mathbb{R}$, the converse of Lemma 16.6a) holds (Theorem 10.5): a map which preserves limits of convergent sequences is necessarily continuous. This *does not hold* in general; but unfortunately counterexamples lie beyond the scope of this text.

THEOREM 16.13. *(F. Dorais) Let $f : F \to F'$ be a homomorphism of ordered fields. The following are equivalent:*
(i) f is uniformly continuous.
(ii) f is continuous.
(iii) f is continuous at 0.
(iv) f is cofinal.

PROOF. (i) $\implies$ (ii) $\implies$ (iii) is immediate.
(iii) $\iff$ (iv): By Lemma 16.5, $f(F)$ is cofinal in F' iff for every positive ϵ' in F', there is a positive δ in F such that $f(\delta) < \epsilon'$. Suppose this holds, and let ϵ' be positive in F', and choose a positive $\delta \in F$ as above. Then for $x \in F$, if $|x| = |x - 0| < \delta$, then

$$|f(x)| = f(|x|) < f(\delta) = |f(\delta) - f(0)| < \epsilon'.$$

This shows f is continuous at 0. Conversely, if f is continuous at 0, then for each positive $\epsilon' \in F$ there is a positive $\delta \in F$ such that if $|x| < \delta$, $|f(x)| < \epsilon'$. Thus $\frac{\delta}{2}$ is a positive element of F such that $\frac{f(\delta)}{2} < \epsilon'$, so $f(F')$ is cofinal in F by Lemma 16.5.
(iii) $\implies$ (i): Let ϵ' be a positive element of F'; since f is continuous at 0 there is a positive $\delta \in F$ such that for all $x \in F$ with $|x| < \delta$, $|f(x)| < \epsilon'$. So, if $x, y \in F$ are such that $|x - x'| < \delta$, then

$$|f(x) - f(y)| = |f(x - y)| < \epsilon'.$$

□

1.3. A Non-Archimedean Ordered Field.

Let $K = \mathbb{R}((t))$ be the field of **formal Laurent series** with $\mathbb{R}$-coefficients: an element of K is a formal sum $\sum_{n \in \mathbb{Z}} a_n t^n$ where there exists $N \in \mathbb{Z}$ such that $a_n = 0$ for all $n < N$. We add such series term by term and multiply them in the same way that we multiply polynomials.

Exercise 16.14: Show that $K = \mathbb{R}((t))$ is a field.

We need to equip K with an ordering; equivalently, we need to specify a set of positive elements. For every nonzero element $x \in K$, we take $v(x)$ to be the smallest $n \in \mathbb{Z}$ such that $a_n \neq 0$. Then we say that x is positive if the coefficient $a_{v(x)}$ of the smallest nonzero term is a positive real number. It is straightforward to see that the sum and product of positive elements is positive and that for each nonzero $x \in K$, exactly one of x and $-x$ is positive, so this gives an ordering on K in the usual way: we decree that $x < y$ iff $y - x$ is positive.

We observe that this ordering is non-Archimedean. Indeed, the element $\frac{1}{t}$ is positive – its one nonzero coefficient is 1, which is a positive real number – and infinitely large: for any $n \in \mathbb{Z}$, $\frac{1}{t} - n$ is still positive – recall that we look to the smallest degree coefficient to check positivity – so $\frac{1}{t} > n$ for all n.

Next we observe that the set $\{\frac{1}{t^n}\}$ is unbounded in K. Taking reciprocals, it follows that the sequence $\{t^n\}$ converges to 0 in K: explicitly, given any $\epsilon > 0$ – here ϵ is not necessarily a real number but any positive element of F! – for all sufficiently large n we have that $\frac{1}{t^n} > \frac{1}{\epsilon}$, so $|t^n| = t^n < \epsilon$. We will use this fact to give a simple explicit description of all convergent sequences in F. First, realize that a sequence

in K consists of, for each $m \in \mathbb{Z}^+$ a formal Laurent series $x_m = \sum_{n \in \mathbb{Z}} a_{m,n} t^n$, so in fact for each $n \in \mathbb{Z}$ we have a real sequence $\{a_{m,n}\}_{m=1}^{\infty}$. Now consider the following conditions on a sequence $\{x_m\}$ in K:

(i) There is an integer N such that for all $m \in \mathbb{Z}^+$ and $n < N$, $a_{m,n} = 0$, and
(ii) For each $n \in \mathbb{Z}$ the sequence $a_{m,n}$ is eventually constant: i.e., for all sufficiently large m, $a_{m,n} = C_n \in \mathbb{R}$. (Because of (i) we must have $C_n = 0$ for all $n < N$.)

Then condition (i) is equivalent to boundedness of the sequence.

I claim that if the sequence converges – say $x_m \to x = \sum_{n=N}^{\infty} a_n t^n \in F$ – then (i) and (ii) both hold. Indeed convergent sequences are bounded, so (i) holds. Then for all $n \geq N$, $a_{m,n}$ is eventually constant in m iff $a_{m,n} - a_n$ is eventually constant in m, so we may consider $x_m - x$ instead of x_m and thus we may assume that $x_m \to 0$ and try to show that for each fixed n, $a_{m,n}$ is eventually equal to 0. As above, this holds iff for all $k \geq 0$, there exists M_k such that for all $m \geq M_k$, $|x_m| \leq t^k$. This latter condition holds iff the coefficient $a_{m,n}$ of t^n in x_n is zero for all $N < k$. Thus, for all $m \geq M_k$, $a_{m,-N} = a_{m,-N+1} = \ldots = a_{m,k-1} = 0$, which is what we wanted to show.

Conversely, suppose (i) and (ii) hold. Then since for all $n \geq N$ the sequence $a_{m,n}$ is eventually constant, we may define a_n to be this eventual value, and an argument very similar to the above shows that $x_m \to x = \sum_{n \geq N} a_n t^n$.

Next I claim that if a sequence $\{x_n\}$ is Cauchy, then it satisfies (i) and (ii) above, hence is convergent. Again (i) is immediate because every Cauchy sequence is bounded. The Cauchy condition here says: for all $k \geq 0$, there exists M_k such that for all $m, m' \geq M_k$ we have $|x_m - x'_m| \leq t^k$, or equivalently, for all $n < k$, $a_{m,n} - a_{m',n} = 0$. In other words this shows that for each fixed $n < k$ and all $m \geq M_k$, the sequence $a_{m,n}$ is constant, so in particular for all $n \geq N$ the sequence $a_{m,n}$ is eventually constant in m, so the sequence x_m converges.

Exercise 16.15: Show that the subfield $\mathbb{R}((t^2))$ of $\mathbb{R}((t))$ is cofinal but not dense.

2. The Sequential Completion

2.1. Sequentially Complete Fields.

An ordered field F is **sequentially complete** if every Cauchy sequence in F converges to an element of F. We have seen that a Dedekind complete ordered field is sequentially complete. Here we wish to examine the converse.

THEOREM 16.14. *For an* Archimedean *ordered field F, TFAE:*
*(i) F is **Dedekind complete**.*
*(ii) F is **sequentially complete**: every Cauchy sequence converges.*

PROOF. The implication (i) $\implies$ (ii) is the content of Theorem 10.32, since the Bolzano-Weierstrass Theorem holds in any ordered field satisfying (LUB).
(ii) $\implies$ (i): Let $S \subset F$ be nonempty and bounded above, and write $\mathcal{U}(S)$ for the set of least upper bounds of S. Our strategy will be to construct a decreasing Cauchy sequence in $\mathcal{U}(S)$ and show that its limit is $\sup S$.

Let $a \in S$ and $b \in \mathcal{U}(S)$. Using the Archimedean property, we choose a negative integer $m < a$ and a positive integer $M > b$, so

$$m < a \leq b \leq M.$$

For each $n \in \mathbb{Z}^+$, we define

$$S_n = \{k \in \mathbb{Z} \mid \frac{k}{2^n} \in \mathcal{U}(A) \text{ and } k \leq 2^n M\}.$$

Every element of S_n lies in the interval $[2^n m, 2^n M]$ and $2^n M \in S_n$, so each S_n is finite and nonempty. Put $k_n = \min S_n$ and $a_n = \frac{k_n}{2^n}$, so $\frac{2k_n}{2^{n+1}} = \frac{k_n}{2^n} \in \mathcal{U}(S)$ wihle $\frac{2k_n-2}{2^{n+1}} = \frac{k_n-1}{2^n} \notin \mathcal{U}(S)$. It follows that we have either $k_{n+1} = 2k_n$ or $k_{n+1} = 2k_n - 1$ and thus either $a_{n+1} = a_n$ or $a_{n+1} = a_n - \frac{1}{2^{n+1}}$. In particular $\{a_n\}$ is decreasing. For all $1 \leq m < n$ we have

$$\begin{aligned} 0 \leq a_m - a_n &= (a_m - a_{m+1}) + (a_{m+1} - a_{m+2}) + \ldots + (a_{n-1} - a_n) \\ &\leq 2^{-(m+1)} + \ldots + 2^{-n} = 2^{-m}. \end{aligned}$$

Thus $\{a_n\}$ is Cauchy, hence by our assumption on F $a_n \to L \in F$.

We CLAIM $L = \sup(S)$. Seeking a contradiction we suppose that $L \notin \mathcal{U}(S)$. Then there exists $x \in S$ such that $L < x$, and thus there exists $n \in \mathbb{Z}^+$ such that

$$a_n - L = |a_n - L| < x - L.$$

It follows that $a_n < x$, contradicting $a_n \in \mathcal{U}(S)$. So $L \in \mathcal{U}(S)$. Finally, if there exists $L' \in \mathcal{U}(S)$ with $L' < L$, then (using the Archimedean property) choose $n \in \mathbb{Z}^+$ with $\frac{1}{2^n} < L - L'$, and then

$$a_n - \frac{1}{2^n} \geq L - \frac{1}{2^n} > L',$$

so $a_n - \frac{1}{2^n} = \frac{k_n-1}{2^n} \in \mathcal{U}(S)$, contradicting the minimality of k_n. □

The proof of (ii) $\implies$ (i) in Theorem 16.14 above is taken from [**HS**] by way of [**Ha11**]. It is rather unexpectedly complicated, but I do not know a simpler proof at this level. However, if one is willing to introduce the notion of convergent and Cauchy **nets**, then one can show first that in an Archimedean ordered field, the convergence of all Cauchy sequences implies the convergence of all Cauchy nets, and second use the hypothesis that all Cauchy nets converge to give a proof which is (in my opinion of course) more conceptually transparent. This is the approach taken in my (more advanced) notes on Field Theory [**FT**].

In fact there are (many!) non-Archimedean sequentially complete ordered fields. We will attempt to describe two very different examples of such fields here. We hasten to add that this is material that the majority of working research mathematicians are happily unfamiliar with, and which is thus extremely rarely covered in undergraduate courses. Only the exceptionally curious need the next section.

2.2. Sequential Completion I: Statement and Applications.

We will now establish one of our main results: for every ordered field F, there is a sequentially complete ordered field $\mathcal{R}$ and a homomorphism $f : F \to \mathcal{R}$.

In fact we can, and will prove, even more than this. The point is that there will be many (nonisomorphic) sequentially complete fields into which any given ordered

field embeds. For example, when we construct the real numbers we will have an embedding $\mathbb{Q} \hookrightarrow \mathbb{R}$. But we also have an embedding $\mathbb{R} \hookrightarrow \mathbb{R}((t))$, so taking the composite gives an embedding $\mathbb{Q} \to \mathbb{R}((t))$. (There is no way that $\mathbb{R}$ and $\mathbb{R}((t))$ are isomorphic, since the former is Archimedean and the latter is not.)

We would like a general definition which allows us to prefer the embedding $\mathbb{Q} \hookrightarrow \mathbb{R}$ to the embedding $\mathbb{Q} \hookrightarrow \mathbb{R}((t))$. The key observation is that, since $\mathbb{R}$ is Archimedean, the embedding of $\mathbb{Q}$ into $\mathbb{R}$ is dense, whereas since $\mathbb{R}((t))$ is not Archimedean, the embedding of $\mathbb{Q}$ into $\mathbb{R}$ is not dense. This leads to the following important definition.

A **sequential completion** of an ordered field F is a dense embedding $F \hookrightarrow \mathcal{R}$ into a sequentially complete ordered field.

LEMMA 16.15. *For an ordered field F, the following are equivalent.*
(i) F is Dedekind complete.
(ii) The inclusion $\iota : \mathbb{Q} \hookrightarrow F$ makes F into a sequential completion of $\mathbb{Q}$.

PROOF. (i) $\implies$ (ii): By Theorem 16.14, F is sequentially complete and Archimedean. By Lemma 16.6, $\mathbb{Q} = \iota(\mathbb{Q})$ is a dense subfield of F, and it follows that F is a sequential completion of $\mathbb{Q}$. □

We will prove that every ordered field admits a sequential completion. And again, we will in fact prove a bit more.

THEOREM 16.16. *Let F be an ordered field.*
a) F admits a sequential completion $\iota : F \to \mathcal{R}$.
b) If L is any sequentially complete ordered field and $f : F \to L$ is a cofinal ordered field homomorphism, then there is a unique ordered field homomorphism $g : \mathcal{R} \to L$ such that $f = g \circ \iota$.

COROLLARY 16.17. *Two sequential completions of the same ordered field are isomorphic.*

PROOF. Let $\iota_1 : F \to \mathcal{R}_1$ and $\iota_2 : F \to \mathcal{R}_2$ be two sequential completions. Applying Theorem 16.16 with $\mathcal{R} = \mathcal{R}_1$ and $f = \iota_2 : F \to \mathcal{R}_2$, we get a unique homomorphism $g : \mathcal{R}_1 \to \mathcal{R}_2$ such that $\iota_2 = g \circ \iota_1$. Interchanging the roles of $\mathcal{R}_1$ and $\mathcal{R}_2$ we also get a unique homomorphism $g' : \mathcal{R}_2 \to \mathcal{R}_1$ such that $\iota_1 = g' \circ \iota_2$.

Now consider $g' \circ g : \mathcal{R}_1 \to \mathcal{R}_\infty$. We have

$$(g' \circ g) \circ \iota_1 = g' \circ (g \circ \iota_1) = g' \circ \iota_2 = \iota_1.$$

Applying Theorem 16.16 with $L = \mathcal{R} = \mathcal{R}_1$ we get that there is a *unique* homomorphism $G : \mathcal{R}_\infty \to \mathcal{R}_1$ such that $G \circ \iota_1 = \iota_1$, but clearly the identity map $1_{\mathcal{R}_1}$ has this property. Thus we must have $g' \circ g = 1_{\mathcal{R}_1}$. Similarly considering $g \circ g' : \mathcal{R}_2 \to \mathcal{R}_2$, then in view of

$$(g \circ g') \circ \iota_2 = g \circ (g' \circ \iota_2) = g \circ \iota_1 = \iota_2,$$

we deduce that $g \circ g' = 1_{\mathcal{R}_2}$. In other words, g and g' are mutually inverse isomorphisms...so $\mathcal{R}_1$ and $\mathcal{R}_2$ are isomorphic. □

Applying Theorem 16.16 to the ordered field $\mathbb{Q}$, we get a sequential completion $\mathcal{R}$ of $\mathbb{Q}$. Since $\mathcal{R}$ is Archimedean and sequentially complete, by Theorem 16.14, $\mathcal{R}$ is Dedekind complete. Conversely, by Lemma 16.15any Dedekind complete ordered

field $\mathcal{R}'$ is isomorphic to $\mathcal{R}$. Thus the existence and uniqueness statements of Theorem 16.16 imply the existence and uniqueness up to isomorphism of a Dedekind complete ordered field.

The uniqueness statement can be strengthened: let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two Dedekind complete ordered fields. We claim that not only are they isomorphic, but that the isomorphism between them is unique. Indeed, for $i = 1, 2$ let $\iota_i : \mathbb{Q} \to \mathcal{R}_i$ be the inclusion maps. We saw above that there is a unique map $g : \mathcal{R}_1 \to \mathcal{R}_2$ such that $g \circ \iota_1 = \iota_2$ and this g is an isomorphism. But any isomorphism $h : \mathcal{R}_1 \to \mathcal{R}_2$ will satisfy $h \circ \iota_1 = \iota_2$, since in fact there is exactly one embedding from $\mathbb{Q}$ into any ordered field. Thus whereas in general there is an isomorphim g between two sequential completions of a given ordered field F which is unique such that blah blah blah (more precisely, such that $g \circ \iota_1 = \iota_2$), in this case the "blah blah blah" is vacuous and the isomorphism is unique full stop.

In abstract mathematics, uniqueness up to a *unique* isomorphism is as close to identical as we can reasonably ask for two structures to be. (Even the "horizontal copy of $\mathbb{R}$" and the "vertical copy of $\mathbb{R}$" are different sets, but the obvious isomorphism between them is the only isomorphism, so no trouble can arise by identifying the two.) We denote this unique field by $\mathbb{R}$ and call it the real numbers...of course.

2.3. Sequential Completion II: The Proof.

Now we are properly motivated to roll up our sleeves and endure the rather lengthy, technical proof of Theorem 16.16. The essential idea (which is indeed due to A.L. Cauchy) is to build the sequential completion directly from the set $\mathcal{C}$ of all Cauchy sequences in F.

We can observe that $\mathcal{C}$ itself has some structure reminiscent of an ordered field but that things do not quite work out: it is somehow *too large* to itself be an ordered field. Namely, it makes perfectly good sense to add, subtract and multiply Cauchy sequences in F. For that matter, it makes perfectly good sense to add, subtract and multiply arbitrary sequences in F: we simply put

$$\{x_n\} + \{y_n\} = \{x_n + y_n\},$$
$$\{x_n\} - \{y_n\} = \{x_n - y_n\},$$
$$\{x_n\} \cdot \{y_n\} = \{x_n \cdot y_n\}.$$

It remains to check that these operations take Cauchy sequences to Cauchy sequences. At the very beginning of our study of sequences we showed this for *convergent* sequences (in $\mathbb{R}$, but the proofs certainly did not use any form of the completeness axiom). It is no more difficult to establish the analogue for Cauchy sequences in F.

LEMMA 16.18. *Let F be any ordered field, and let $a_\bullet$, $b_\bullet$ be Cauchy sequences. Then $a_\bullet + b_\bullet$ and $a_\bullet \cdot b_\bullet$ are both Cauchy.*

PROOF. Since $a_\bullet$ and $b_\bullet$ are both Cauchy, for $\epsilon > 0$ there is $N \in \mathbb{Z}^+$ such that for $m, n \geq N$, $|a_m - a_n| < \epsilon$ and and $|b_m - b_n| < \epsilon$. Then

$$|(a_m + b_m) - (a_n + b_n)| \leq |a_m - a_n| + |b_m - b_n| < 2\epsilon.$$

Further, since the sequences are Cauchy, they are bounded: there are $M_a, M_b \in F$ such that $|a_n| \leq M_a$ and $|b_n| \leq M_b$ for all $n \in \mathbb{Z}^+$. Then for $m, n \geq N$,

$$|a_m b_m - a_n b_n| \leq |a_m - a_n||b_m| + |a_n||b_m - b_n| \leq (M_a + M_b)\epsilon.$$

□

So does this addition and multiplication endow $\mathcal{C}$ with the structure of a field? There is an additive identity, namely the sequence with $x_n = 0$ for all n. There is also a multiplicative identity, namely the sequence with $x_n = 1$ for all n. It all works well until we get to multiplicative inverses.

Exercise 16.16: Let $\{x_n\}$ be a sequence in the ordered field F.
a) Show that there is a sequence $\{y_n\}$ with $\{x_n\} \cdot \{y_n\} = \{1\}$ if and only if for all $n \in \mathbb{Z}^+$, $x_n \neq 0$.
b) Show that if $\{x_n\}$ is Cauchy and $x_n \neq 0$ for all n, then its inverse $\{\frac{1}{x_n}\}$ is again a Cauchy sequence.

Thus there are plenty of Cauchy sequences other than the constantly zero sequence which do not have multiplicative inverses: e.g. $(0, 1, 1, 1, \ldots)$, or indeed any constant sequence which takes the value 0 at least once. Thus $\mathcal{C}$ has many good algebraic properties, but it is not the case that every nonzero element has a multiplicative inverse, so it is not a field.[5]

We also have some order structure on $\mathcal{C}$. For instance, it is tempting to define $\{x_n\} > \{y_n\}$ if $x_n > y_n$ for all n. This turns out not to be a good definition in the sense that it does not lead to a trichotomy: there will be unequal Cauchy sequences $\{x_n\}$ and $\{y_n\}$ for which neither is less than the other, e.g.

$$\{x_n\} = \{0, 1, 1, \ldots\},\ \{y_n\} = \{1, 0, 0, \ldots\}.$$

As in the definition of convergence, it is more fruitful to pay attention to what a Cauchy sequence is doing *eventually*. Exploiting this idea we can get a sort of trichotomy result.

LEMMA 16.19. *(Cauchy Trichotomy) For a Cauchy sequence $\{x_n\}$ in an ordered field F, exactly one of the following holds:*
(i) There is a positive element $\epsilon \in F$ and $N \in \mathbb{Z}^+$ such that $x_n \geq \epsilon$ for all $n \geq N$.
(ii) There is a positive element $\epsilon \in F$ and $N \in \mathbb{Z}^+$ such that $x_n \leq -\epsilon$ for all $n \geq N$.
(iii) x_n converges to 0.

PROOF. It is easy to see that the conditions are mutually exclusive. Let us suppose that (iii) does not hold: thus there is $\epsilon > 0$ and a subsequence $\{x_{n_k}\}$ such that $|x_{n_k}| \geq \epsilon$ for all $k \in \mathbb{Z}^+$. By passing to a further subsequence we may assume either that $x_{n_k} \geq \epsilon$ for all k or $x_{n_k} \leq -\epsilon$ for all k. Let us suppose that the former holds and show that this implies (i): if so, replacing x by $-x$ shows that the latter alternative implies (ii). Since $\{x_n\}$ is Cauchy, there is $N \in \mathbb{Z}^+$ such that $|x_m - x_n| \leq \frac{\epsilon}{2}$ for all $m, n \geq N$. Putting these two conditions together we get $x_n \geq \frac{\epsilon}{2}$ for all $n \geq N$. □

[5]for those who know some abstract algebra: what we've shown is that $(\mathcal{C}, +, \cdot)$ is a **commutative ring**. There is a very general algebraic method which, when given a commutative ring, will yield a collection of fields defined in terms of that ring. The present construction is indeed an instance of this.

Unfortunately this is not quite the kind of trichotomy which defines a total order relation: we have some elements that we regard as positive – case (i) above, some elements that we regard as negative – case (ii) above – but for an order relation only the zero element should be neither positive nor negative, whereas case (iii) above includes the much larger collection of elements *converging* to zero.

Lemma 16.19 suggests that if we could somehow "squash down" the subset of Cauchy sequences which converge to 0 to a single point, then we would actually get a total order relation. This business of "squashing subsets to a point" is formalized in mathematics (more so in algebra and topology than the part of mathematics we've been studying for most of this text!) by an **equivalence relation**. Rather than providing a logically complete but pedagogically useless whirlwind tour of equivalence relations, we will simply assume that the reader is familiar with them.[6] Namely, we regard any two Cauchy sequences which converge to 0 as equivalent. We are left with the question of when to regard two Cauchy sequences which do not converge to zero as equivalent. We could simply "not squash them", i.e., declare two such sequences to be equivalent exactly when they are equal. But a little exploration shows that this won't work: we'll get a total order relation but it won't interact well with the algebraic structure. For instance, consider the Cauchy sequences

$$\{x_n\} = (0, 0, 1, 1, 1, \ldots), \{y_n\} = (1, 1, 1, 1, \ldots).$$

The difference $\{x_n\} - \{y_n\}$ converges to 0 so is getting identified with 0. Thus we should identify $\{x_n\}$ and $\{y_n\}$ as well. This leads to the following key definitions.

Let $\mathcal{Z}$ be the set of all sequences in F which converge to 0; convergent sequences are Cauchy, so certainly $\mathcal{Z} \subset \mathcal{C}$. For two Cauchy sequences $a_\bullet, b_\bullet \in \mathcal{C}$, we put

$$a_\bullet \sim b_\bullet \iff a_\bullet - b_\bullet \in \mathcal{Z}.$$

In words, two Cauchy sequences are equivalent iff their difference converges to zero.

Exercise 16.17 (if you know abstract algebra): Show that $\mathcal{Z}$ is a maximal ideal in the commutative ring $\mathcal{C}$. Why is this an exciting sign that we're on the right track?

Exercise 16.18: Let $\{x_n\}$ be a Cauchy sequence in F, and let $\{x_{n_k}\}$ be any subsequence. Show that $\{x_n\} \sim \{x_{n_k}\}$.

Now we define $\mathcal{R}$ as $\mathcal{C}/\sim$, that is, the set of equivalence classes of Cauchy sequences. This will be the underlying set of our sequential completion. It remains to endow it with all the rest of the structure. The idea here is that when we pass to a quotient by an equivalence relation we can try to simply carry over the structure we already had, but at every step we must check that the operations are **well-defined**, meaning they are independent of the chosen equivalence class. At no point are these verifications difficult, but we admit they can be somewhat tedious.

Let us check the addition and multiplication induced well-defined operations on

[6] At UGA they are covered in the "transitional" Math 3200 course. The reader who has made it through most of this text will have no problem learning this concept.

the set $\mathcal{R}$ of equivalence classes. This means: if we have four Cauchy sequences $a_\bullet, b_\bullet, c_\bullet, d_\bullet$ and $a_\bullet \sim c_\bullet$, $b_\bullet \sim d_\bullet$, then

$$a_\bullet + b_\bullet \sim c_\bullet + d_\bullet,\ a_\bullet b_\bullet \sim c_\bullet d_\bullet.$$

All right: since $a_\bullet \sim c_\bullet$ and $b_\bullet \sim d_\bullet$, $a_\bullet - c_\bullet \to 0$ and $b_\bullet - d_\bullet \to 0$, so

$$(a_\bullet + b_\bullet - (c_\bullet + d_\bullet) = (a_\bullet - c_\bullet) + (b_\bullet - d_\bullet) \to 0 + 0 = 0,$$

so $a_\bullet + b_\bullet \sim c_\bullet + d_\bullet$. Similarly,

$$a_\bullet b_\bullet - c_\bullet d_\bullet = (a_\bullet - c_\bullet)b_\bullet + (b_\bullet - d_\bullet)c_\bullet,$$

and this converges to 0 because $a_\bullet - c_\bullet, b_\bullet - d_\bullet \to 0$ and $b_\bullet, c_\bullet$ are bounded. Thus we have equipped our set $\mathcal{R}$ with two binary operations $+$ and $\cdot$.

PROPOSITION 16.20. *$(\mathcal{R}, +, \cdot)$ is a field.*

PROOF. Most of these axioms are completely straightforward (but, yes, somewhat tedious) to verify and are left to the reader as exercises. Let us single out:
(P3) The additive identity is $[0_\bullet]$, the class of the constant sequence 0.
(P7) The multiplicative identity is $[1_\bullet]$, the class of the constant sequence 1.
(P8) Suppose that $x \in \mathcal{R} \setminus \{[0_\bullet]\}$, and let $x_\bullet$ be any Cauchy sequence representing x. Then we must have $x_n \neq 0$ for all sufficiently large n: indeed, otherwise we would have $0_\bullet = (0, 0, 0, \ldots)$ as a subsequence, and if a subsequence of a Cauchy sequence converges to 0, then the Cauchy sequence itself converges to 0, contradiction. Suppose $x_n \neq 0$ for all $n > N$. Then define $y_\bullet$ by $y_n = 0$ for all $1 \leq n \leq N$ (or whatever you want: it doesn't matter) and $y_n = \frac{1}{x_n}$ for $n > N$. Then $x_n y_n = 1$ for all $n > N$, so $x_\bullet y_\bullet$ differs from $1_\bullet$ by a sequence which is convergent to zero: $[x_\bullet][y_\bullet = [1_\bullet] = 1$, so $y = [y_\bullet]$ is the multiplicative inverse of $x = [x_\bullet]$. □

We now equip $\mathcal{R}$ with an ordering. For $a_\bullet, b_\bullet \in \mathcal{C}$, we put $[a_\bullet] > [b_\bullet]$ if there is a positive element ϵ in F such that $a_n - b_n \geq \epsilon$ for all sufficiently large n. We claim that this is well-defined independent of the representatives $a_\bullet$ and $b_\bullet$ chosen. Indeed, if $x_\bullet$ and $y_\bullet$ converge to zero, then for sufficiently large n we have $|x_n - y_n| < \frac{\epsilon}{2}$ and then

$$a_n + x_n) - (b_n + y_n) = (a_n - b_n) + (x_n - y_n) \geq \frac{\epsilon}{2}.$$

THEOREM 16.21. *The strict ordering $<$ on $\mathcal{R}$ endows it with the structure of an ordered field.*

PROOF. Step 1: The trichotomy property follows from Lemma 16.19. The transitive property is easy: if $[a_\bullet] < [b_\bullet]$ and $[b_\bullet] < [c_\bullet]$, then there is $N \in \mathbb{Z}^+$ and $\epsilon_1, \epsilon_2 > 0$ such that for all $n \geq N$, $b_n - a_n \geq \epsilon_1$ and $c_n - b_n \geq \epsilon_2$, hence $c_n - a_n = (c_n - b_n) + (b_n - a_n) \geq \epsilon_1 + \epsilon_2$. Thus $<$ is a strict ordering on $\mathcal{R}$.
Step 2: If $[a_\bullet], [b_\bullet] > 0$, then as above there is $N \in \mathbb{Z}^+$ and $\epsilon_1, \epsilon_2 > 0$ such that for all $n \geq N$, $a_n \geq \epsilon_1$ and $b_n \geq \epsilon_2$. Then for all $n \geq N$, $a_n + b_n \geq \epsilon_1 + \epsilon_2$ and $a_n b_n \geq \epsilon_1 \epsilon_2$. □

PROPOSITION 16.22. *For an ordered field F, let $\iota : F \to \mathcal{R}$ by sending $x \in F$ to the class of the constant sequence $(x, x, \ldots)$. Then ι is a dense homomorphism of ordered fields.*

PROOF. Step 1: It is immediate that ι is a field homomorphism. Further, if $x > 0$, then $(x, x, \ldots) > 0$: indeed, we can take $\epsilon = x$. So ι is a homomorphism of ordered fields.

Step 2: Let $x = [a_\bullet]$ in $\mathcal{R}$. We claim that $\iota(a_n) \to x$. Indeed, if $\epsilon = [\epsilon_\bullet]$ is a positive element of $\mathcal{R}$, then there is $e > 0$ in F such that $\epsilon_n \geq \epsilon$ for all sufficiently large n. Since $\{a_n\}$ is Cauchy in F, there is $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$,

$$|a_n - a_m| < \frac{e}{2}.$$

Then for sufficiently large m,

$$\epsilon_m - |a_n - a_m| \geq e - |a_n - a_m| \geq \frac{e}{2},$$

which shows that for all $n \geq N$, $|\iota(a_n) - x| < \epsilon$. Step 3: It remains to show that ι is dense, so let $a = [a_\bullet] < b = [b_\bullet] \in \mathcal{R}$. Let $m = \frac{a+b}{2}$ and let $\epsilon = \frac{b-a}{2}$; thus the interval of width ϵ centered at m has as its endpoints a and b. By Step 2, there is a sequence $\{x_n\} \in F$ such that $\iota(x_n) \to m$. By definition then, for all sufficiently large n we have$|\iota(x_n) - m| < \epsilon$, and thus $\iota(x_n) \in (a, b)$. □

Finally, we prove Theorem 16.16 in the following explicit form.

THEOREM 16.23. *For any ordered field F, the homomorphism $\iota : F \to \mathcal{R} = \mathcal{C}/\mathcal{Z}$ is a sequential completion. Moreover, if L is a sequentially complete field and $f : F \to L$ is a cofinal ordered field homomorphism, then there is a unique ordered field homomorphism $g : \mathcal{R} \to L$ such that $f = g \circ \iota$.*

PROOF. We already know that $\iota : F \to \mathcal{R}$ is a dense ordered field homomorphism. It remains to check first that $\mathcal{R}$ is sequentially complete and second that ι satisfies the "universal property" of Theorem 16.16b).
Step 1: Let $\{x_n\}$ be a Cauchy sequence in $\mathcal{R}$; we must show that it is convergent. By Proposition 16.9 it suffices to check this after passing to a subsequence. If $\{x_n\}$ has only finitely many distinct points, we have a constant subsequence, which is certainly convergent. Otherwise $\{x_n\}$ has infinitely many distinct points, so that after passage to a subsequence we may assume they are all distinct. Since by Proposition 16.22 ι is a dense homomorphism, for all $n \in \mathbb{Z}^+$ there is $y_n \in F$ such that y_n lies strictly between x_n and x_{n+1}. It follows that for any $n, k \in \mathbb{Z}^+$, y_{n+k} lies in any interval containing $x_n, x_{n+1}, \ldots, x_{n+k+1}$. Since $\{x_n\}$ is Cauchy, this forces $\{y_n\}$ be to Cauchy, and thus $y = [\{y_n\}] \in \mathcal{R}$. We also have $x_n - y_n \to 0$, so $\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = y$.
Step 2: Let L be a sequentially complete ordered field, and let $f : F \to L$ be a cofinal ordered field homomorphism. We must show that there is a unique ordered field homomorphism $g : \mathcal{R} \to L$ such that $f = g \circ \iota$.
Uniqueness: Let $g : \mathcal{R} \to L$ be such that $f = g \circ \iota$. We will see explicitly what g must be on each $x \in \mathcal{R}$. Write $x = [a_\bullet]$. Since $\{a_n\}$ is Cauchy in F and f is cofinal, f is uniformly continuous, so $\{f(a_n)\}$ is Cauchy in L, hence convergent, say to $y \in L$. Since f is cofinal, is g, hence g is uniformly continuous. Since $\iota(a_n) \to x$, $f(a_n) = g(\iota(a_n)) \to g(x)$. Since $f(a_n) \to y$ and limits are unique, we must have

$$g(x) = y = \lim_{n\to\infty} f(a_n).$$

Existence: We must show that putting $g(x) = y$ as above defines an ordered field homomorphism from $\mathcal{R}$ to L. If $x_1 = [a_n]$ and $x_2 = [b_n]$, let $y_1 = \lim_{n\to\infty} f(a_n)$ and $y_2 = \lim_{n\to\infty} f(b_n)$. Then $a_n + b_n \to x_1 + x_2$ and $f(a_n + b_n) = f(a_n) + f(b_n) \to y_1 + y_2$, so $g(x_1 + x_2) = y_1 + y_2 = g(x_1) + g(x_2)$. Similarly, $a_n b_n \to x_1 x_2$ and $f(a_n b_n) = f(a_n)f(b_n) \to y_1 y_2$, so $g(x_1 x_2) = y_1 y_2 = g(x_1)g(x_2)$. Finally, if

$x = [a_\bullet] > 0$, then there is a positive $\epsilon \in F$ such that $a_n \geq \epsilon$ for all sufficiently large n, so

$$g(x) = \lim_{n\to\infty} f(a_n) \geq \lim_{n\to\infty} f(\epsilon) = f(\epsilon) > 0.$$

□

Bibliography

[A] T.M. Apostol, *Calculus. Volume One.*

[Ab84] Y.S. Abu-Mostafa, *A Differentiation Test for Absolute Convergence*. Math. Mag. 57 (1984), 228-231.

[Ac00] F. Acerbi, *Plato: Parmenides 149a7-c3. A Proof by Complete Induction?* Archive for History of the Exact Sciences 55 (2000), 57–76.

[ADC12] S. Ali and M. Deutsche Cohen, *The phi-ratio tests*. Elem. Math. 67 (2012), 164-168.

[Al08] S.A. Ali, *The mth ratio test: new convergence tests for series*. Amer. Math. Monthly 115 (2008), 514-524.

[An95] P. Andrews, *Where not to find the critical points of a polynomialvariation on a Putnam theme.* Amer. Math. Monthly 102 (1995), 155-158.

[Ap73] T.M. Apostol, *Another Elementary Proof of Euler's Formula for* $\zeta(2n)$. Amer. Math. Monthly 80 (1973), 425–431.

[As12] J.M. Ash, *The Limit Comparison Test Needs Positivity*. Math. Magazine 85 (2012), 374–375.

[Ba96] R.G. Bartle, *Return to the Riemann integral*. Amer. Math. Monthly 103 (1996), 625-632.

[B] J.C. Bowman, *Honours Calculus*, online lecture notes.

[Ba70] R. Baer, *Dichte, Archimedizität und Starrheit geordneter Körper*. Math. Ann. 188 (1970), 165-205.

[Ba22] S. Banach, *Sur les opérations dans les ensembles abstraits et leur applications aux équations intégrales*. Fund. Math. 3 (1922), 133–181.

[Ba98] B. Banaschewski, *On proving the existence of complete ordered fields*. Amer. Math. Monthly 105 (1998), 548–551.

[Be06] A.F. Beardon, *Contractions of the Real Line*. Amer. Math. Monthly 113 (2006), 557–558.

[Be12] S.J. Bernstein, *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités*. Communications of the Kharkov Mathematical Society 13 (1912), 1–2.

[Be13] Á. Besenyei, *Lebesgue's Road to Antiderivatives*. Math. Mag. 86 (2013), 255–260.

[Be14] Á. Besenyei, *Peano's unnoticed proof of Borel's theorem*. Amer. Math. Monthly 121 (2014), 69-72.

[BM22] H. Bohr and J. Mollerup, Lœrebog i matematisk Analyse, vol. 3, Jul. Gjellerups Forlag, Copenhagen, 1922.

[Bo71] R.P. Boas, Jr., *Signs of Derivatives and Analytic Behavior*. Amer. Math. Monthly 78 (1971), 1085–1093.

[Bo95] É. Borel, *Sur quelque points de la théorie des fonctions*. Ann. Sci. lÉcole Norm. Sup. 12 (1895), 9–55.

[BS] R.G. Bartle and D.R. Sherbert, *Introduction to real analysis*. Second edition. John Wiley & Sons, Inc., New York, 1992.

[C] H. Cartan, *Elementary theory of analytic functions of one or several complex variables*. Dover Publications, Inc., New York, 1995.

[Ca21] A.L. Cauchy, *Analyse algébrique*, 1821.

[Ca89] A.L. Cauchy, *Sur la convergence des séries*, in *Oeuvres complètes Sér. 2*, Vol. 7, Gauthier-Villars (1889), 267–279.

[Ch01] D.R. Chalice, *How to Differentiate and Integrate Sequences*. Amer. Math. Monthly 108 (2001), 911–921.

[CJ] R. Courant and F. John, *Introduction to Calculus and Analysis*.

[Cl10] P.L. Clark, *Real induction*. `http://math.uga.edu/~pete/realinduction.pdf`

[Cl11] P.L. Clark, *Induction and completeness in ordered sets*. `http://math.uga.edu/~pete/induction_completeness_brief.pdf`

[Co77] G.L. Cohen, *Is Every Absolutely Convergent Series Convergent?* The Mathematical Gazette 61 (1977), 204–213.

[CdC] K. Conrad, *The contraction mapping theorem.* `http://www.math.uconn.edu/~kconrad/blurbs/analysis/contractionshort.pdf`

[CdD] K. Conrad, *Estimating the size of a divergent sum.* `http://www.math.uconn.edu/~kconrad/blurbs/analysis/sumest.pdf`

[Co49] J.L. Coolidge, *The story of the binomial theorem.* Amer. Math. Monthly 56 (1949), 147-157.

[Cu65] F. Cunningham, Jr., *Classroom Notes: The Two Fundamental Theorems of Calculus.* Amer. Math. Monthly 72 (1965), 406-407.

[Da12] D. Daners, *A Short Elementary Proof of* $\sum \frac{1}{k^2} = \frac{\pi^2}{6}$. Math. Mag. 85 (2012), 361–364.

[dBR76] P. du Bois-Reymond, *Über den Gültigkeitschbereich der Taylor'schen Reihenentwickelung.* Sitzungb. k. Bayer. Akad. Wiss., math.-phys. Klasse (1876), 225–237.

[dBR83] P. du Bois-Reymond, *Über den Gültigkeitschbereich der Taylor'schen Reihenentwickelung.* Math. Ann. 21 (1883), 107–119.

[DC] P.L. Clark, *Discrete calculus.* In preparation. Draft available on request.

[DR50] A. Dvoretzky and C.A. Rogers, *Absolute and unconditional convergence in normed linear spaces.* Proc. Nat. Acad. Sci. USA 36 (1950), 192–197.

[DS] *Dirichlet series*, notes by P.L. Clark, available at `http://math.uga.edu/~pete/4400dirichlet.pdf`

[Ed62] M. Edelstein, *On fixed and periodic points under contractive mappings.* J. London Math. Soc. 37 (1962), 74-79.

[Eh94] P. Ehrlich, *Dedekind cuts of Archimedean complete ordered abelian groups.* Algebra Universalis 37 (1997), 223-234.

[Er94] M. Erickson, *An introduction to combinatorial existence theorems.* Math. Mag. 67 (1994), 118-123.

[ES35] P. Erdős and G. Szekeres, *A combinatorial problem in geometry.* Compositio Math. 2 (1935), 463-470.

[FT] *Field Theory*, notes by P.L. Clark, available at `http://www.math.uga.edu/~pete/FieldTheory.pdf`

[Go] R. Gordon, *Real Analysis: A First Course.* Second Edition, Addison-Wesley, 2001.

[Gr] P.M. Gruber, *Convex and discrete geometry.* Grundlehren der Mathematischen Wissenschaften 336. Springer, Berlin, 2007.

[Ha88] J. Hadamard, *Sur le rayon de convergence des séries ordonnées suivant les puissances d'une variable.* C. R. Acad. Sci. Paris 106 (1888), 259–262.

[Ha11] J.F. Hall, *Completeness of Ordered Fields.* 2011 arxiv preprint.

[Ha50] H.J. Hamilton, *A type of variation on Newton's method.* Amer. Math. Monthly 57 (1950), 517-522.

[H] G.H. Hardy, *A course of pure mathematics.* Centenary edition. Reprint of the tenth (1952) edition with a foreword by T. W. Körner. Cambridge University Press, Cambridge, 2008.

[Ha02] F. Hartmann, *Investigating Possible Boundaries Between Convergence and Divergence.* College Math. Journal 33 (2002), 405–406.

[Ha11] D. Hathaway, *Using Continuity Induction.* College Math. Journal 42 (2011), 229–231.

[HS] E. Hewitt and K. Stromberg, *Real and abstract analysis. A modern treatment of the theory of functions of a real variable.* Third printing. Graduate Texts in Mathematics, No. 25. Springer-Verlag, New York-Heidelberg, 1975.

[Ho66] A. Howard, *Classroom Notes: On the Convergence of the Binomial Series.* Amer. Math. Monthly 73 (1966), 760-761.

[Ho95] M.E. Hoffman, *Derivative polynomials for tangent and secant.* Amer. Math. Monthly 102 (1995), 23-30.

[Ka07] I. Kalantari, *Induction over the continuum.* Induction, algorithmic learning theory, and philosophy, 145–154, Log. Epistemol. Unity Sci., 9, Springer, Dordrecht, 2007.

[Ke70] H. Kestelman, *Riemann Integration of Limit Functions.* Amer. Math. Monthly 77 (1970), 182–187.

[Kn80] W.J. Knight, *Functions with zero right derivatives are constant.* Amer. Math. Monthly 87 (1980), 657-658.

[Kö91] T.W. Körner, *Differentiable functions on the rationals*. Bull. London Math. Soc. 23 (1991), 557-562.

[La] J. Labute, *Math 255, Lecture 22: Power Series: The Binomial Series*. http://www.math.mcgill.ca/labute/courses/255w03/L22.pdf

[L] S. Lang, *Undergraduate analysis*. Second edition. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1997.

[Li33] F.A Lindemann, *The Unique Factorization of a Positive Integer*. Quart. J. Math. 4, 319–320, 1933.

[LV06] M. Longo and V. Valori, *The Comparison Test – Not Just for Nonnegative Series*. Math. Magazine 79 (2006), 205–210.

[Lu99] J. Lu, *Is the Composite Function Integrable?* Amer. Math. Monthly 106 (1999), 763–766.

[Ma56] A.M. Macbeath, *A criterion for differentiability*. Edinburgh Math. Notes (1956), 8-11.

[Ma42] C. Maclaurin, *Treatise of fluxions, 1*. Edinburgh (1742), 289–290.

[Ma61] Y. Matsuoka, *An elementary proof of the formula* $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. Amer. Math. Monthly 68 (1961), 485–487.

[MK09] M.M. Marjanović and Z. Kadelburg, *Limits of composite functions*. The Teaching of Mathematics, XII (2009), 1–6. http://elib.mi.sanu.ac.rs/files/journals/tm/22/tm1211.pdf

[Me08] A. Melman, *Bounds on the zeros of the derivative of a polynomial with all real zeros*. Amer. Math. Monthly 115 (2008), 145-147.

[Me72] F. Mertens, *Ueber die Multiplicationsregel für zwei unendliche Reihen*. Journal für die Reine und Angewandte Mathematik 79 (1874), 182–184.

[MS22] E.H. Moore and H.L. Smith, *A General Theory of Limits*. Amer. J. of Math. 44 (1922), 102–121.

[Mo50] R.K. Morley, *Classroom Notes: The Remainder in Computing by Series*. Amer. Math. Monthly 57 (1950), 550–551.

[Mo51] R.K. Morley, *Classroom Notes: Further Note on the Remainder in Computing by Series*. Amer. Math. Monthly 58 (1951), 410–412.

[Mo57] T.E. Mott, *Newton's method and multiple roots*. Amer. Math. Monthly 64 (1957), 635-638.

[Mu63] A.A. Mullin, *Recursive function theory (A modern look at a Euclidean idea)*. Bulletin of the American Mathematical Society 69 (1963), 737.

[Ne03] Nelsen, R.B. *An Impoved Remainder Estimate for Use with the Integral Test*. College Math. Journal 34 (2003), 397–399.

[Ne81] D.J. Newman, *Differentiation of asymptotic formulas*. Amer. Math. Monthly 88 (1981), 526-527.

[NP88] D.J. Newman and T.D. Parsons, *On monotone subsequences*. Amer. Math. Monthly 95 (1988), 44-45.

[No52] M.J. Norris, *Integrability of continuous functions*. Amer. Math. Monthly 59 (1952), 244-245.

[Ol27] L. Olivier, Journal für die Reine und Angewandte Mathematik 2 (1827), 34.

[PFD] P.L. Clark, *Partial Fractions Via Linear Algebra*, http://www.math.uga.edu/∼pete/partialfractions.pdf.

[Ro63] K. Rogers, *Classroom Notes: Unique Factorization*. Amer. Math. Monthly 70 (1963), 547–548.

[R] W. Rudin, *Principles of mathematical analysis*. Third edition. International Series in Pure and Applied Mathematics. McGraw-Hill, New York-Auckland-Düsseldorf, 1976.

[Sa95] H. Samelson, *More on Kummer's test*. Amer. Math. Monthly 102 (1995), 817-818.

[Sc] M. Schramm, *Introduction to Real Analysis*. Dover edition, 2008.

[SP88] D. Scott and D.R. Peeples, *The Teaching of Mathematics: A Constructive Proof of the Partial Fraction Decomposition*. Amer. Math. Monthly 95 (1988), 651-653.

[Se59] A. Seidenberg, *A simple proof of a theorem of Erdős and Szekeres*. J. London Math. Soc. 34 (1959), 352.

[Sm86] R.S. Smith, *Rolle over Lagrange – Another Shot at the Mean Value Theorem*. College Math. J. 17 (1986), 403–406.

[S] M. Spivak, *Calculus*. Fourth edition.

[St95] S.K. Stein, *Do Estimates of an Integral Really Improve as n Increases?* Math. Mag. 68 (1995), 16-26.

[St37] M.H. Stone, *Applications of the theory of Boolean rings to general topology*. Trans. Amer. Math. Soc. 41 (1937), 375–481.

[St48] M.H. Stone, *The generalized Weierstrass approximation theorem*. Math. Mag. 21 (1948), 167–184.

[St67] J.H. Staib, *A Sequence-Approach to Uniform Continuity*. Math. Mag. 40 (1967), 270–273.

[St90] G. Strang, *Sums and Differences vs. Integrals and Derivatives*. College Math. Journal 21 (1990), 20–27.

[Ta55] A. Tarski, *A lattice-theoretical fixpoint theorem and its applications*. Pacific J. Math. 5 (1955), 285–309.

[To94] J.C. Tong, *Kummer's test gives characterizations for convergence or divergence of all positive series*. Amer. Math. Monthly 101 (1994), 450–452.

[T] J.W. Tukey, *Convergence and Uniformity in Topology*. Annals of Mathematics Studies, no. 2. Princeton University Press, 1940.

[Ve02] D.J. Velleman, *Partial fractions, binomial coefficients, and the integral of an odd power of* $\sec\theta$. Amer. Math. Monthly 109 (2002), 746–749.

[Wa48] H.S. Wall, *A modification of Newton's method*. Amer. Math. Monthly 55 (1948), 90–94.

[Wa95] J.A. Walsh, *The Dynamics of Newton's Method for Cubic Polynomials*. The College Mathematics Journal 26 (1995), 22–28.

[Wa36] M. Ward, *A Calculus of Sequences*. Amer. J. Math. 58 (1936), 255–266.

[Ze34] E. Zermelo, *Elementare Betrachtungen zur Theorie der Primzahlen*. Nachr. Gesellsch. Wissensch. Göttingen 1, 43–46, 1934.