Honors Calculus

Pete L. Clark

Contents

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 of their common feature: they all concern an arbitrary continuous function defined on a closed, bounded interval.

Theorem 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 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. This means that there exist 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 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

[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.

theorems are used to prove even more famous theorems in the course, e.g. the **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 defined operations of addition and multiplication, and now we have an additive identity, 0, as well as a multiplicative identity.

Let me remark that $\mathbb{Z}^+$ and $\mathbb{N}$ are clearly 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 Spivak uses $\mathbb{N}$ to denote the positive integers. I am perhaps 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 $0 - 17$. In the integers then 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 with respect to the basic operations of $+$ and $\cdot$ present already on $\mathbb{Z}^+$, the rational numbers are ideally suited: we have an additive identity 0, every element has an additive inverse, we have a multiplicative identity 1, and every nonzero elmeent has a multiplicative inverse.[3]

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 real number x reeks of contrivance: it seems to have been constructed only to make trouble. The mathematicians of the ancient Pythagorean school discovered a much more "real" irrational real number.

Theorem 4. *(Pythagoras) 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**. The strategy is simple: 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 in fact be true.

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

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

[3] Trust me for now that we would not want 0 to have a multiplicative inverse: the existence of a "number" x such that $0 \cdot x = 1$ has undesirable consequences.

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 this into the equation, we get

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

or

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

Thus b^2 is divisible by 2, so as above $b = 2B$ for some $B \in \mathbb{Z}$. Substitutin this into our last equation, 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 $[a, b]_{\mathbb{Q}} = \{x \in \mathbb{Q} \mid a \leq x \leq b\}$.

Example: 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: 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 arbitairly large.[4] 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 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, $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$.

[4] Once again, we will be much more precise about this sort of thing later on. This is just an overview.

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

Example: 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: 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 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 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 7. *In every system satisfying the field axioms, $(-1)^2 = 1$.*

Proof. By Proposition 7b), $(-1) \cdot (-1)$ is the additive inverse of -1, namely 1. □

Proposition 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 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 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 9. *In every number system satisfying the ordered field axioms:*
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: 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: 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 10. *For any* x, y, z *in a system satisfying the ordered field axioms:*
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 XXX. 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. Again, this implies $xy \neq 0$ 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 11. *For all a, b, c, d in a system satisfying the ordered field axioms:*
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 you as an exercise.
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

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

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 $n_0 + 1$.[5]

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 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 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 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 however prove part c) now, since this is true for elements of any system satisfying the ordered field axioms. 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 9c).

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

Here is a special case of Theorem 12 important enough to be recorded separately.

Corollary 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 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 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 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 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.[6] □

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 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

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

[6] Such **symmetry arguments** can often by used to reduce the number of cases considered in a proof.

by Proposition 14. So this gives a second proof of the Triangle Inequality.

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

PROPOSITION 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:[7] 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 17. *(Cauchy-Bunyakovsky-Schwarz Inequality, $n = 2$)*
a) For all numbers x_1, x_2, y_1, y_2,

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

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 establish the following **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 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. Again, the basic idea of showing that the right hand side minus the left hand side is equal to a sum of squares will work. Expanding out the right and left hand sides, we get

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

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

[7] All exercises appearing in these notes are informal and optional. However if you want to turn any exercise in and have me take a look at it, that can probably be arranged.

so

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

□

THEOREM 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, as an application of the theory of convex functions, we will present a signficant generalization of Theorem 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 certainly 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**, or abstract properties that 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:

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

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 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 20. *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 $1 = \frac{1(1+1)}{2}$.

[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.

Induction step: Let $n \in \mathbb{Z}^+$ and suppose that $1 + \ldots + n = \frac{n(n+1)}{2}$. Then

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

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

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 20 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 20.

For a very striking pictorial variation of the above argument, go to `http://mathoverflow.net/questions/8846/proofs-without-words` and scroll down to the first diagram.

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

Theorem 21. *(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 $kp_i = p_1 \cdots p_n + 1 = bp_i + 1$, 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: $p_1, \ldots, p_n, q$. □

Remark: 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 (no ancient Greek mathematician did), 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. (In particular Euclid *does not* verify the base case, and he must have regarded it as obvious that there is at least one prime number. And it is – but it should be included as part of the proof anyway!) 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 classic example of a proof by contradiction – which it isn't! Rather, Euclid's argument is perfectly contructive. 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.[3] 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.

[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.

[3] Many mathematical amateurs seem to have the idea that $N_n = p_1 \cdots p_n + 1$ is always prime, but clearly it isn't.

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 the lead in the wikipedia article on the *Parmenides*: "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 inductiive proof of Proposition 20 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 22. *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 that $P(n)$ 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 23. *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 22 from Proposition 20:

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

4. Inequalities

Proposition 24. *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 an arbitrary natural number and asusme $P(n)$: $2^n > n$. 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 $n \geq 2$ we have $2^n > n$. Then

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

since $n > 1$. □

Exercise 5: Use calculus to show that in fact $2^x > x$ for all real x. (To see what's going on, it will be very helpful to graph the two functions. Of course, merely drawing a picture will not be a sufficient proof.)

Proposition 25. *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 26. *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 $n \in \mathbb{Z}^+$ we have $1 + \frac{1}{4} + \ldots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$. 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 establish the inequality

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

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, at last, is a true inequality. 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 horses 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 27. *Let p be a prime number, and $a, b \in \mathbb{Z}^+$. If $p \mid ab$, $p \mid a$ or $p \mid b$.*

Later in these notes we will give a proof of Euclid's Lemma (yes, by induction!). For now we simply assume it to be true. Our point is that we can swiftly deduce the following useful generalization.

Proposition 28. *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 that it holds 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. Now 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 holds, we're done. If the former holds, then by our inductive hypothesis, $p \mid a_i$ for some $1 \leq i \leq n$, so we are also done. □

Comment: In this and other induction proofs of this type, it is the base case which is nontrivial, and the induction step is essentially the same argument every time.

Corollary 29. *Let p be a prime, $n \in \mathbb{Z}^+$, $a \in \mathbb{Z}^+$ such that $p \mid a^n$. Then $p \mid a$.*

Exercise 6: Use Corollary 29 to show that for any prime p, $p^{\frac{1}{n}}$ is irrational.

Proposition 30. *Let $n \geq 3$ be an integer, and let $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be differentiable functions. Then*

$$(f_1 \cdots f_n)' = f_1' f_2 \cdots f_n + f_1 f_2' \cdots f_n + \ldots + f_1 \cdots f_{n-1} f_n'.$$

Proof. We argue by induction on n.
Base case ($n = 2$): The assertion is $(f_1 f_2)' = f_1' f_2 + f_1 f_2'$, which is the product rule from differential calculus.

Induction step: We assume the result is true for any n differentiable functions. If $f_1, \ldots, f_{n+1}$ are all differentiable, then

$$(f_1 \cdots f_n f_{n+1})' = ((f_1 \cdots f_n) f_{n+1})' \overset{*}{=} (f_1 \cdots f_n)' f_{n+1} + f_1 \cdots f_n f_{n+1}' =$$

$$(f_1' f_2 \cdots f_n) f_{n+1} \overset{**}{=} f_1 f_2' f_3 \cdots f_n f_{n+1} + \ldots + f_1 \cdots f_{n-1} f_n' f_{n+1} + f_1 \cdots f_n f_{n+1}'.$$

Note that in the first starred equality we have applied the usual product rule and in the second starred equality we have applied the induction hypothesis. □

Corollary 31. *For any positive integer n, if $f(x) = x^n$, then $f'(x) = nx^{n-1}$.*

Proof. Exercise 7. □

When teaching freshman calculus, it is very frustrating not to be able to prove the power rule by this simple inductive argument!

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[4]

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

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 $P(1) \wedge P(2) \wedge \ldots \wedge P(n-1)$.[5]

Less obviously, one can use our previous PMI to prove PS/CI. To most mathematicians this is a comforting fact: one does not want to keep introducing additional "axioms" or "assumptions" in order to solve problems. Again the proof is not hard but slightly tricky. Suppose that we believe in PMI and we wish to prove PS/CI. Let $P(n)$ be any sentence with domain the positive integers and satisfying (i) and (ii) above. We wish to show that $P(n)$ is true for all positive integers n, 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 $Q(1) = P(1)$ and that (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 we also know $P(n+1)$, 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 ordinary mathematical induction, $Q(n)$ holds for all n, hence certainly $P(n)$ holds for all n.

Exercise 9: 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 32. *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! □

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

[5] I do admit that the underlying logical reasoning here is rather abstract and hence mildly confusing. If you want to follow along, give yourself some time and a quiet place to work it out!

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

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

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.

$$\frac{F_{201}}{F_{200}} = 1.6180339887498948482045868 34\ldots.$$

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

$$\frac{1-\sqrt{5}}{2} = 1 - \varphi = -.6180339887498948482045868 34\ldots.$$

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 $C_1r_1^n + C_2r_2^n$ for any constants C_1 and C_2. Therefore we propose $x_n = C_1r_1^n + C_2r_2^n$ 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 nothing more than high school algebra. But one cannot say that the calculation is much fun. It is always fun to find some clever way to circumvent a tedious calculation, so in that spirit I present the following alternate argument. Namely, instead of determining the constants by evaluating F_n at $n = 1$ and $n = 2$, it would be much easier algebraically to evaluate at $n = 1$ and $n = 0$, because 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 33. *(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 10: 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

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

What happened? The characteristic polynomial in this case is $r^2 - 2r + 1 = (r - 1)^2$, so that it has repeated roots. One solution is $C_1 1^n = C_1$ (i.e., x_n is a constant sequence). This occurs if and only if $x_2 = x_1$, 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 $x_{n+2} = 2r_0 x_{n+1} - r_0^2 x_n$ (which is the most general recurrence whose characteristic polynomial is $(r - r_0)^2$.

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 32.

Theorem 34. *(Well-Ordering Principle) Let S be any nonempty subset of the positive integers. 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 32 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**.

Upon examination, the two proofs of Proposition 32 are very close: the difference between a proof using strong induction and a proof using well ordering is more a matter of literary taste than mathematical 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 35. *(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 36. *(Fundamental Theorem of Arithmetic) The factorization of any integer $n > 1$ into primes is unique, up to the order of the factors. Explicitly,* *suppose that*

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

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* $1 \leq i \leq k$*.*

Let us say that a prime factorization $n = p_1 \cdots p_k$ is in **standard form** if, as above, we have $p_1 \leq \ldots \leq p_k$. Every prime factorization can be put in standard form by ordering the primes from least to greatest, and dealing with standard form factorizations is a convenient bookkeeping device, since otherwise our uniqueness statement would have to include a proviso "up to the order of the factors", which makes everything slightly more complicated.

Remark: When I teach number theory I state the *existence* of prime factorizations as the first part of the Fundamental Theorem of Arithmetic and the above uniqueness statement as the second part. Since we have already proven – twice! – that every integer greater than one may be factored into a product of primes, it doesn't seem necessary to restate it here. Anyway, the uniqueness of prime factorizations lies much deeper than the existence.

We wish to draw the reader's attention to the following important point: given Proposition 32 – i.e., the existence of prime factorizations, Theorems 35 and 36 are equivalent: each can be easily deduced from the other.

EL implies FTA: Assume Euclid's Lemma. As we have already seen, this implies the Generalized Euclid's Lemma (Proposition 28): 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 the Generalized Euclid's Lemma (Proposition 28), 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 now $\frac{n}{p_1}$ is strictly 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 we have $k = l$, $p_1 = q_j = q_1$ and $p_i = q_i$ for $2 \leq i \leq j$. But this means that in (7) the two factorizations are the same after all! Done.

FTA implies EL: Assume that every integer greater than one factors *uniquely* into a product of primes, and 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 that a and b are both greater than one 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 (employed, famously, by Euclid in his *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 route takes some time to develop – perhaps a week in an elementary number theory course. It is therefore remarkable that one can bypass all these intermediate steps and give direct inductive proofs of both EL and FTA. We will give both of these in turn (which is, to be sure, twice as much work as we need to do given the just proved equivalence of EL and FTA).

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

Here is a proof of Euclid's Lemma using the Well-Ordering Principle, following K. Rogers [**Ro63**].

As we saw earlier in the course, one can prove Euclid's Lemma for any particular prime p by consideration of cases. In particular we have already seen that Euclid's Lemma holds for all a and b when $p = 2$, and so forth. So suppose for a contradiction that there exists at least one prime such that Euclid's Lemma does not hold for that prime, and among all such primes, by WOP we consider the least one, say p. What this means that there exist $a, b \in \mathbb{Z}^+$ such that $p \mid ab$ but $p \nmid a$ and $p \nmid b$. Again we apply WOP to choose the least positive integer a such that there exists at least one positive integer b 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, certainly $a > 1$ – if $p \mid 1 \cdot b$, then indeed $p \mid b$! – so that a is divisible by at least one prime (a consequence of Proposition 32) 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. Therefore Euclid's Lemma holds for all primes p.

9.3. The Lindemann-Zermelo Inductive Proof of FTA.

Here is a proof of FTA using the Well-Ordering Principle, 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 elment, 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,we must have $p_1 \neq q_j$ 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 $2 \leq i \leq r$. But then multiplying back by $p_1 = q_1$ we see that we didn't have two different factorizations after all. (In fact this shows that for all i, j, $p_i \neq q_j$.)

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!

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 $C_a : \mathbb{R} \rightarrow \mathbb{R}$ 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 $I : \mathbb{R} \rightarrow \mathbb{R}$ 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ we have $I \circ f = f \circ I = f$.

Example: Let $m, b \in \mathbb{R}$, and consider the function $L : \mathbb{R} \rightarrow \mathbb{R}$ 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 37. *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 $\deg(fg) = \deg f + \deg g$.
c) We have $\deg(f \circ g) = \deg f \cdot \deg g$.

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 $\deg fg = m + n$.
c) Unlike the first two parts, this result is not actually very useful, so we leave the proof as an exercise for the interested reader. (To get the flavor of it, note that if $f = x^m$, $g = x^n$ then $f \circ g = f(x^n) = (x^n)^m = x^{mn}$. The general case is no harder than this but involves more notation.) □

From an algebraic perspective, the following result is the most important and fundamental property of polynomials.

Theorem 38. *(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) $\deg r(x) < \deg b(x)$.

Note that Theorem 38 is directly analogous to Integer Division with Remainder. In fact the only difference in the statement comes when we want to express the fact that the remainder polynomial $r(x)$ is "smaller" than the divisor polynomial $b(x)$. The set of polynomial expressions doesn't come with a natural "less than" relation, so we interpret smaller in terms of the degree.

Moreover the proof of Theorem 38 is analogous to that of Integer Division with Remainder: namely in junior / high school algebra we learned an explicit algorithm, the usual long division of polynomials, which when followed yields the quotient $q(x)$ and the remainder $r(x)$. What we probably didn't see in junior / high school is a *proof* that this algorithm always works and that the quotient and remainder polynomials are unique. We leave the writeup of this as an **extra credit** exercise.

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

Theorem 39. *(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 40. *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 (why not?).
Base case ($n = 0$): If f has degree 0 then it is a nonzero constant function, so it has no roots at all.
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 140, 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 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 41. *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 143 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 42. *(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 143, 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 146 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. Recall that we say 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 43. *(Generalized Euclid's Lemma) Let $x, y, z \in \mathbb{Z}^+$. Suppose that x, y are coprime and that yz is divisible by x. Then z is divisible by x.*

Extra Credit Exercise:
a) Prove Theorem 147 using the Fundamental Theorem of Arithmetic (uniqueness of prime factorizations).
b) Explain why Theorem 147 is indeed a generalization of Euclid's Lemma. (Hint: let $x = p$ be prime. What does it mean that p, y are coprime?)

Theorem 44. *(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. Well, we know

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

Multiplying through by c^n clears denominators, giving

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

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 147 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 147 a_0 is divisible by b. □

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

Corollary 45. *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? Well, here's what. Let p be any prime number, $n \geq 2$ any integer, and consider the polynomial

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

By Corollary 149, 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 *rational number* c 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!

2. Rational Functions

A rational function is a functtion which is a quotient of two polynomial functions: $f(x) = \frac{P(x)}{Q(x)}$, with $Q(x)$ not the zero function. Other than this definition and the remark that a natural domain of a rational function is the set of all real numbers except the finitely many roots of $Q(x)$, I do not believe that I said anything about rational functions at this point in the course. This is a happy coincidence, since I have nothing in this section of my notes!

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 46. *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 47. *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 48. *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 49. *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 50. *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 51. *(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 52. *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: now $|g(c)||f(x) - f(c)|$ is a constant times something which can be made arbitrarily small, so it 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 as this happens, by continuity of f, $|f(x)|$ gets arbitrarily close to $|f(c)|$: thus $|f(x)|$ is not constant but still *bounded*, and something which is bounded times something which can be made arbitarily small can be made arbitrarily small! Now formally:

Using Lemma 51a) 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

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

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 51b) 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 53. *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 52. □

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 54. *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 55. *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 56. *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 52 above, or (ii) adapting the proofs of Theorem 52 to the current context.

Now what about limits of composite functions? The natural analogue of Theorem 52e) 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 52e) 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 57. *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 57 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 58. *(Marjanović [**MaKa09**]) 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 57. 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 [**MaKa09**] gives some applications of the implication (ii)b) $\implies$ (i) of Theorem 58 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 59. *(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

$$\lim_{x \to 0} \frac{\sin x}{x} = 1. \tag{12}$$

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 (12). 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:

$$\lim_{x \to 0} \frac{1-\cos x}{x} = \lim_{x \to 0} \frac{\cos x - 1}{x} = 0. \tag{13}$$

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 60. *(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 61. *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 62. *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

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

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 63. *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

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

so

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

□

Remark about **linear continuity**...

The converse of Theorem 63 is far from being true: a function f which is continuous at a need not be differentiable at a. An easy example of this is $f(x) = |x|$ at $a = 0$.

But in fact the situation is even 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 64. *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.

PROOF. a) We have already proved this; it is restated for comparison with part b).
b) (i) $\implies$ (ii): Suppose first that f is differentiable at a and also that $f(a) \neq 0$. By Theorem 63 f is continuous at a and therefore there exists some $\delta > 0$ such that for all $x \in I = (a - \delta, a + \delta)$, f has the same sign at x as it does at a: in other words, if $f(a) > 0$ then $f(x)$ is positive for all $x \in I$ and if $f(a) < 0$ then $f(x)$ is negative for all $x \in I$. In the first case, upon restriction to I, $|f| = f$, so it is differentiable at a since f is. In the second case, upon restriction to I, $|f| = -f$, which is also differentiable at a since f is and hence also $-f$ is.

Now suppose that $f(a) = f'(a) = 0 \ldots$

□

2. Differentiation Rules

THEOREM 65. *(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 66. *(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 67. *(Linearity of the Derivative) For any differentiable functions f and g and any constants C_1, C_2, we have*

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

The proof is an immediate application of the Sum Rule followed by the Product Rule. The point here is that functions $L : V \to W$ with the property that $L(v_1 + v_2) = L(v_1) + L(v_2)$ and $L(Cv) = CL(v)$ are called **linear mappings**, and are extremely important across mathematics.[1] The study of linear mappings is the subject of **linear algebra**. That differentiation is a linear mapping (on the *infinite-dimensional vector space* of real functions) provides an important link between calculus and algebra.

THEOREM 68. *(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.

The generalized product rule: 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

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

Riding this train of thought a bit farther, here a rule for the product of any finite number $n \geq 2$ of differentiable functions.

THEOREM 69. *(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*

$$(f_1 \cdots f_n)'(a) = f_1'(a)f_2(a) \cdots f_n(a) + \ldots + f_1(a) \cdots f_{n-1}(a)f_n'(a). \tag{14}$$

[1] We are being purposefully vague here as to what sort of things V and W are...

Proof. By induction on n.
Base Case ($n = 2$): This is precisely the "ordinary" Product Rule (Theorem XX).
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 (14). 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

$$(f'_1(a)f_2(a) \cdots f_n(a) + \ldots + f_1(a) \cdots f_{n-1}(a)f'_n(a))\, f_{n+1}(a)+f_1(a) \cdots f_n(a)f'_{n+1}(a)$$
$$= f'_1(a)f_2(a) \cdots f_n(a)f_{n+1}(a) + \ldots + f_1(a) \cdots f_n(a)f'_{n+1}(a).$$

□

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 a positive integer. 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 $x' = 1$ multipled by $n - 1$ factors of x, so each term evalutes to x^{n-1}. Moreover we have n terms in all, so

$$(x^n)' = nx^{n-1}.$$

No need to mess around with binomial coefficients!
Example: More generally, for any differentiable function f and $n \in \mathbb{Z}^+$, the Generalized Product Rule shows that the function $f(x)^n$ is differentiable and $(f(x)^n)' = nf(x)^{n-1}$. (This sort of computation is more traditionally done using the Chain Rule...coming up soon!)

Theorem 70. *(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

$$(\frac{1}{g})'(a) = \lim_{h\to 0} \frac{1/g(a + h) - 1/g(a)}{h}$$

$$= \lim_{h\to 0} \frac{g(a) - g(a + h)}{hg(a)g(a + h)} = -\left(\lim_{h\to 0} \frac{g(a + h) - g(a)}{h}\right)\left(\lim_{h\to 0} \frac{1}{g(a)g(a + h)}\right) = \frac{-g'(a)}{g(a)^2}.$$

Above we have once again used the fact that g is differentiable at a implies 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

$$\left(\frac{f}{g}\right)'(a) = \left(f \cdot \frac{1}{g}\right)'(a) = f'(a)\frac{1}{g(a)} + f(a)\left(\frac{1}{g}\right)'(a)$$

$$= \frac{f'(a)}{g(a)} - f(a)\frac{g'(a)}{g(a)^2} = \frac{g(a)f'(a) - g'(a)f(a)}{g(a)^2}.$$

□

Lemma 71. *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. We leave this as an (assigned, this time!) exercise, with the following suggestion to the reader: suppose that $\lim_{x\to a} f(x) = M \neq L$, and derive a contradiction by taking ϵ to be small enough compared to $|M - L|$. □

Example: Consider, again, for $\alpha \in \mathbb{R}$, the function $f_\alpha : \mathbb{R} \to \mathbb{R}$ defined by $f_\alpha(x) = x^\alpha \sin(\frac{1}{x})$ for $x \neq 0$ and $f_\alpha(0) = 0$. Then f satisfies hypothesis (ii) of Lemma 71 with $L = 0$, since on any deleted interval around zero, the function $\sin(\frac{1}{x})$ takes the value 0 infinitely many times. According to Lemma 71 then, if $\lim_{x\to 0} f_\alpha(x)$ exists at all, then it must be 0. As we have seen, the limit exists iff $\alpha > 0$ and is indeed equal to zero in that case.

Theorem 72. *(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, the above argument has a gap in it: 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 $a = 0$.[2] This gap is often held to

[2] One should note that in order for a function to have this property it must be "highly oscillatory near a" as with the functions f_α above: indeed, f_α is essentially the simplest example of a function having this kind of behavior. In particular, most of the elementary functions considered in freshman calculus do not exhibit this highly oscillatory behavior near any point and therefore the above argument is *already* a complete proof of the Chain Rule for such functions. Of course

invalidate the proof, and thus the most common proof of the Chain Rule in honors calculus / basic analysis texts proceeds along (superficially, at least) different lines.

But in fact I maintain that the above gap may be rather easily filled to give a complete proof. The above argument is valid *unless* the following holds: for all $\delta > 0$, there exists x with $0 < |x - a| < \delta$ such that $f(x) - f(a) = 0$. So it remains to give a different proof of the Chain Rule in that case. First, observe that with the above hypothesis, the difference quotient $\frac{f(x)-f(a)}{x-a}$ is equal to 0 at points arbitarily close to $x = a$. It follows from Lemma 71 that if

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

exists at all, then it must be equal to 0. But we are *assuming* that the above limit exists, since we are assuming that f is differentiable at a. Therefore what we have seen is that in the remaining case we have $f'(a) = 0$, and therefore, since we are trying to show that $(g \circ f)'(a) = g'(f(a))f'(a)$, we are trying in this case to show that $(g \circ f)'(a) = 0$. So consider our situation: 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 the algebra

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

is justified, and the above argument shows that 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.[3] 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. ...

our business here is to prove the Chain Rule for *all* functions satisfying the hypotheses of the theorem, even those which are highly oscillatory!

[3] This is the same idea as in the proof of the Switching Theorem, although – to my mild disappointment – we are not able to simply apply the Switching Theorem directly, since one of our functions is not defined in a deleted interval around zero.

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, $f(x) > 0 = f(0)$.[4]

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 73. *Let $f : I \to \mathbb{R}$, and let a be an interior point of a. Suppose f is differentiable at a.*
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 through a, decreasing at a, or neither.

PROOF. a) The differentiability of f at a has an ϵ-δ interpretation, and the idea is to use this interpretation 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

[4] 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.

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.[5]
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 $x \in D$, $f(x) \geq m$.

Again a function clearly can have at most one minimum value but need not have any at all: 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 a function $f : D \to \mathbb{R}$ is **bounded above** if there exists a number B such that for all $x \in D$, $f(x) \leq B$. A function is **bounded below** if there exists a number b such that for all $x \in D$, $f(x) \geq b$. A function is **bounded** if it is both bounded above and bounded 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

[5] Suggestion: *either* go through the above proof flipping inequalities as appropriate, *or* 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).

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 the statement (but not yet the proof; sorry!) of one of the most important theorems in this or any course.

THEOREM 74. *(Extreme Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then f has a maximum and minimum value, and in particular is bounded above and below.*

Again this result is of paramount importance: 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 75. *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 X.X 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 X.X 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 76. *(Optimization Procedure) Let $f : [a, b] \to \mathbb{R}$ be continuous. Then the minimum and maximum values must each be attained at a point $x \in [a, b]$ which is either:*
(i) an endpoint: $x = a$ or $x = b$,
*(ii) a **stationary point**: $f'(a) = 0$, or*
*(iii) a **point of nondifferentiability**.*

Often one lumps cases (ii) and (iii) of Theorem XX 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 in fact not be *any* critical points, but that would only make our discussion 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 minimum of these values 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 XX out by finding the maximim 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 $-.619\ldots$, occurring at $\frac{9+\sqrt{17}}{8}$.

3.5. Remarks on finding roots of f'.

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 77. *(Mean Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists at least one c such that $a < c < b$ and*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Remark: If you will excuse a (vaguely) personal anecdote: 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 the class 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. Upon further reflection it became clear that one-sided differentiability at the endpoints was not in fact required, i.e., one of the pre-existing choices was the correct answer and not the one I had added. So I changed my answer and submitted my exam. As you can see from the statement above, my final answer was correct. **But** many students in the class figured that if I had successfully lobbied for an additional answer then this answer was probably the correct one, with the effect that they changed their answer from the correct answer to my incorrect added answer! They were not so thrilled with either me or the teacher, but in my opinion he at least behaved admirably: this was a real "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 certain checkpoints along an interstate highway in Georgia. One day you receive in the mail 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. The enclosed letter explains: 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 Extreme Value Theorem (which we have not yet proven, but all in good time...). However, it is convenient to first establish a special case.

THEOREM 78. *(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 the Extreme Value Theorem, f has a maximum M and a minimum m.
Case 1: Suppose $M > f(a) = f(b)$. Then the maximum value does not occur 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 $f(a) \leq m \leq M \leq f(a)$, which implies $m = M = f(a) = f(b)$, so f is constant. In this case $f'(c) = 0$ 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. Indeed we will meet such a generalization later – the **Cauchy Mean Value Theorem** – and use it to prove L'Hôpital's Rule – but at the moment it is, alas, easier to use a simple trick.

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 modest 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 certainly has its place: for instance it may be used to prove L'Hôpital's Rule.

Theorem 79. *(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*

$$(f(b) - f(a))g'(c) = (g(b) - g(a))f'(c). \tag{15}$$

Proof. Case 1: Suppose $g(a) = g(b)$. By Rolle's Theorem, there is $c \in (a,b)$ such that $g'(c) = 0$. With this value of 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 $h(a)=h(b)$.[6] 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?

5. Monotone Functions

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 increasing / weakly increasing / decreasing / weakly decreasing on an interval in terms of sign condition on f'.

Theorem 80. *(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 $x_1,x_2 inI$ 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 81. *(Zero Velocity Theorem) Let $f: I\to\mathbb{R}$ be a differentiable function with identically zero derivative. Then f is constant.*

[6] 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.

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: The strategy of the above proof is to deduce Corollary 81 from the Increasing Function Theorem. In fact if we argued directly from the Mean Value Theorem the proof would be significantly shorter: try it!

COROLLARY 82. *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 81, $h \equiv C$ and thus $f = g + h = g + C$. □

Remark: Corollary 82 can be viewed as the first "uniqueness theorem" for differential equations. Namely, suppose that $f : I \to \mathbb{R}$ is some function, and consider the set of all functions $F : I \to \mathbb{R}$ such that $F' = f$. Then Corollary **??** 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 $F + C$. In perhaps more familiar terms, this asserts that antiderivatives are unique up to an additive constant, when they exist.

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 in the second half of the course. (On the third hand, there are *some* discontinuous functions which have antiderivatives, but it is too soon to get into this...)

COROLLARY 83. *Let $n \in \mathbb{Z}^+$, and 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 84. *Let $f : [a, b] \to \mathbb{R}$ be a function. We suppose:*
(i) f is continuous at $x = a$ and $x = b$.
(ii) f is weakly increasing (resp. increasing, weakly decreasing, decreasing) on (a, b).

Then f is weakly increasing (resp. increasing, weakly decreasing, decreasing) on $[a, b]$.

Remark: The "resp." in the statement above is an abbreviation for "respectively". Use of respectively in this way is a shorthand way for writing out several cognate statements. In other words, we really should have four different statements, each one of the form "if f has property X on (a, b), then it also has property X on $[a, b]$", where X runs through weakly increasing, increasing, weakly decreasing, and decreasing. Use of "resp." in this way is not great mathematical writing, but it is sometimes seen as preferable to the tedium of writing out a large number of very similar statements. It certainly occurs often enough for you to get used to seeing and understanding it.

Proof. There are many similar statements here; let's prove some of them.
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 $f(a) < f(c) < f(b)$. □

Let us say that a function $f : I \to \mathbb{R}$ is **monotone** if it is either increasing on I or decreasing on I, and also that f is **weakly monotone** if it is either weakly increasing on I or weakly decreasing on I.

Theorem 85. *(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 weakly 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 weakly monotone. The following are equivalent:
(i) f is not *monotone.*
(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) $\Longrightarrow$ (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 84.
b) (i) $\implies$ (ii): Suppose f is weakly increasing on I but not increasing on I. By Theorem 84 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 86. *Let $f : I \to \mathbb{R}$ be differentiable. Suppose that $f'(x) \geq 0$ for all $x \in I$, and that $f'(x) > 0$ except at a finite set of points $x_1, \ldots, x_n$. Then f is increasing on I.*

Example: A typical application of Theorem 86 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 84 to quickly derive another staple of freshman calculus.

Theorem 87. *(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 $I \setminus \{a\}$ – i.e., differentiable at every point of I except possibly at $x = a$. 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 84 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 88. *(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 89. *(Intermediate Value Theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function defined on a closed, bounded interval. 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 (or **IVT**) is ubiquitously useful. As alluded to earlier, even such innocuous properties as every non-negative real number having a square root contain an implicit appeal to IVT. From the present point of view, it 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 1. *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 proved long ago and have used many times that 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 90. *(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 91. *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 92. *(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: In general this has nothing to do with $\frac{1}{f}$. Thus $\sin^{-1}(x) \neq \csc(x) = \frac{1}{\sin x}$. Because this is legitimately confusing, many calculus texts write the inverse sine function as $\arcsin x$. But in general one needs to get used to f^{-1} being used for the inverse function.)

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 93. *(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 $f(X)$.[7]

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 94. *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 95. *(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.*

[7] This is sometimes called the **range** of f, but sometimes not. It is safer to call it the image!

PROOF. At the moment we will give the proof only when $I = [a, b]$, i.e., is closed and bounded. The general case will be discussed later when we switchback to talk about least upper bounds. Now 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. It follows that the image $f([a, b])$ of f is a subset of the interval $[m, M]$. 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 $f([a, b]) = [m, M]$. □

Remark: Although we proved only a special case of the Interval Image Theorem, in this case we proved a stronger result: if f is a continuous function defined on a closed, bounded interval I, then $f(I)$ is again a closed, bounded interval. One might hope for analogues for other types of intervals, but in fact this is not true.

Exercise: Let I be a nonempty interval which is *not* of the form $[a, b]$. Let J be any nonempty interval in $\mathbb{R}$. Show that there is a continuous function $f : I \to \mathbb{R}$ with $f(I) = J$.

6.3. Monotone Functions and Invertibility.

Recall $f : I \to \mathbb{R}$ is **monotone** if it is either increasing or decreasing. Every monotone function is injective. (In fact, a weakly monotone function is monotone if and only if it is injective.) Therefore our dirty trick of codomain restriction works to show that if $f : I \to \mathbb{R}$ is monotone, $f : I \to f(I)$ is bijective, hence invertible. Thus in this sense we may speak of the inverse of any monotone function.

PROPOSITION 96. *Let $f : I \to f(I)$ be a 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 97. *(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 $f(a) > f(b) < f(c)$.*

PROOF. Exercise! □

THEOREM 98. *Let $f : I \to \mathbb{R}$ be continuous and injective. Then f 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 97 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 99. *(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 96, 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 with you, I don't find the above proof to be very enlightening. After reflecting a little bit on my dissatisfaction with this argument, I came up with an alternate proof, which in my opinion is conceptually simpler, but depends on the **Monotone Jump Theorem**, a characterization of the possible discontinuities of a weakly monotone function. The proof of this theorem uses the Dedekind completeness of the real numbers, so is postponed to the next part of the notes in which we discuss completeness head-on.[8]

[8] Nevertheless in my lectures I did state the Monotone Jump Theorem at this point and use it to give a second proof of the Continuous Inverse Function Theorem.

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 100. *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

$$(f^{-1})'(f(x))f'(x) = 1, \tag{16}$$

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 (16) 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 100 but end up deriving it again. Well, enough complaining: here goes.

THEOREM 101. *(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 $k_h \in I$.[9] 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:

$$(f^{-1})'(b) = \lim_{h\to 0} \frac{a + k_h - a}{f(a + k_h) - b} = \lim_{h\to 0} \frac{k_h}{f(a + k_h) - f(a)} = \lim_{h\to 0} \frac{1}{\frac{f(a+k_h)-f(a)}{k_h}}.$$

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 99 – 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 85 $f : \mathbb{R} \to \mathbb{R}$ is increasing. Moreover, we have $\lim_{x\to\pm\infty} f(x) = \pm\infty$.

[9] 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.

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 $x = 0$ (at which point there is a well-defined, but vertical, tangent line). It is typical to call this function $x^{\frac{1}{n}}$.[10]

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 the practice of **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 102. *For all $x, y \in (0, \infty)$, we have*

$$L(xy) = L(x) + L(y). \tag{17}$$

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

$$L(y) = 0 + L(y) + C_y,$$

[10] I'll let you think about why this is good notation: it has to do with the rules for exponentiation.

and thus $C_y = 0$, so $L(xy) = L(x) + L(y)$. □

Corollary 103. *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 104. *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

$$L(y_1 y_2) = L(y_1) + L(y_2).$$

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

$$E(x_1)E(x_2) = y_1 y_2 = E(L(y_1 y_2)) = E(L(y_1) + L(y_2)) = E(x_1 + x_2).$$

□

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 105. *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f'(x) = f(x)$ for all $x \in \mathbb{R}$. Then there is a constant C such that $f(x) = CE(x)$ for all $x \in \mathbb{R}$.*

Proof. Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined 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 106. *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 107. *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 ($n = 1$):
$\frac{d}{dx} e^{x^2} = 2xe^{x^2} = P_1(x)e^{x^2}$, 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 $\frac{d^{n+1}}{dx^{n+1}} e^{x^2} =$

$$\frac{d}{dx}\frac{d^n}{dx^n} e^{x^2} \overset{\text{IH}}{=} \frac{d}{dx} P_n(x)e^{x^2} = P_n'(x)e^{x^2} + 2xP_n(x)e^{x^2} = \left(P_n'(x) + 2xP_n(x)\right) e^{x^2}.$$

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

$$\arcsin : [-1, 1] \rightarrow [\frac{-\pi}{2}, \frac{\pi}{2}].$$

Being the inverse of anincreasing function, $\arcsin x$ is increasing. Moreover since the sine function has a nonzero derivative on $(\frac{-\pi}{2}, \frac{\pi}{2})$, $\arcsin x$ is differentiable there. As usual, to find the derivative we prefer to redo the implicit differentiation by hand: 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

$$\arccos : [-1, 1] \rightarrow [0, \pi].$$

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 the arccos function just as we did for arcsin above: differentiating the identity

$$\cos \arccos x = x$$

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

$$\arccos x + \arcsin x = \frac{\pi}{2}.$$

Thus the angle θ whose sine is x is complementary to the angle φ whose cosine is x. A little thought should convince you that this is a familiar fact.

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

$$\arctan : \mathbb{R} \to (\frac{-\pi}{2}, \frac{\pi}{2}).$$

Since the tangent function is differentiable with everywhere positive derivative, the same is true for $\arctan x$. In particular it is increasing, but not without bound: we have $\lim_{x\to\pm\infty} \arctan x = \pm\frac{\pi}{2}$. In other words the arctangent has horizontal asymptotes at $y = \pm\frac{\pi}{2}$.

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 108. *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 108 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 108 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 108 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 108 *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 109. *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 108a), $\mathbb{R}$ satisfies (LUB), and thus by part a) it satisfies (GLB). By Theorem 108b) $\mathbb{R}$ 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 109a).
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 109 it is reasonable to use the term **Dedekind completeness** to refer to either or both of (LUB), (GLB), and we shall do so.

Theorem 110. *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 111. *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 112. *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 111 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 113. *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 111 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 114. *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 114.

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 114: 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 114, 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 114. 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 114 to see what $\inf \varnothing$ and $\sup \varnothing$ should be. Namely, to define these quantities in such a way as to obey Proposition 114, 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 115. *Any nonempty convex subset D of $\mathbb{R}$ is an interval.*

Proof. We have already seen the most important insight for the proof: we *must* use the Dedekind-completeness of $\mathbb{R}$ in our argument. With this in mind the only remaining challenge is one of organization: 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 $I = D = \overline{I} = \mathbb{R}$.

[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.

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. □

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 116. *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 157 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 117. *(Strong Intermediate Value Theorem) If $f : I \to \mathbb{R}$ is continuous, then f satisfies the Intermediate Value Property. In particular, $f(I)$ 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: Finally, since $f : I \to \mathbb{R}$ satisfies the Intermediate Value Property, by Theorem 158 it maps every subinterval of I to an interval of $\mathbb{R}$. In particular $f(I)$ itself is an interval in $\mathbb{R}$. □

Remark: Theorem 159 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 118. *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 a function $f : F \to F$ 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 118 is continuous at every element of F.

3. The Monotone Jump Theorem

THEOREM 119. *(Monotone Jump) Let $f : I \to \mathbb{R}$ be weakly monotone, and let $c \in I$.*
a) Suppose c is an interior point of I. Then $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^+} f(x)$ both exist, and

$$\lim_{x\to c^-} f(x) \leq f(c) \leq \lim_{x\to c^+} f(x).$$

b) Suppose c is the left endpoint of I. Then $\lim_{x\to c^+} f(x)$ exists and is greater than or equal to $f(c)$.
c) Suppose c is the right endpoint of I. Then $\lim_{x\to c^-} f(x)$ exists and is less than or equal to $f(c)$.

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

$$\mathfrak{l} \leq f(c) \leq \mathfrak{r}. \tag{18}$$

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 (18) 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 120. *For a weakly 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) $\Longrightarrow$ (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) $\Longrightarrow$ (i): This follows immediately from Theorem 159. □

With Theorems 159 and 120 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 159, $f(I) = J$ 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 120, f^{-1} is continuous!

4. Real Induction

Theorem 121. *(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 122. *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 121.
(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 $x = \inf T$. □

Remark: Like Dedekind completeness, the notion of "Real Induction" depends only on the ordering relation $<$ and not on the field operations $+$ and $\cdot$. In fact, given an arbitrary 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 the Principle of Real Induction holds. In a recent note [**Cl11**], I proved that Theorem 122 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 123. *(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**:[5] 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 $f(x_M) = M$. □

[5] I am embarrassed to admit that the previous version of my lecture notes fell into exactly this trap. These notes were taken from a piece I wrote last year after giving a talk for math graduate students at UGA. They have never been formally published, but they are available on the web and have been read by several (dozen?) people, none of whom pointed out this mistake. Oh, well...

6. Uniform Continuity

6.1. 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.

6.2. 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 124. *(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 125. *(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 Patching 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 $x \in S(\epsilon)$. □

7. 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 126. *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 7.1: 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 7.2: If $S \subset T$ and x is a limit point of S, x is a limit point of T.

Example 7.4: No finite subset S of $\mathbb{R}$ has a limit point.

Example 7.5: 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 7.6: 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 127. *(Bolzano-Weierstrass) Every infinite subset A of a closed bounded interval $[a, b]$ has a limit point.*

Proof. Let S be the set of $x \in [a, b]$ such that *if* $A \cap [a, x]$ is infinite, then $A \cap [a, x]$ has a limit point. It suffices to show that $S = [a, b]$: then $A = A \cap [a, b]$ is infinite and thus has a limit point. We will use Real Induction.

(RI1): Since $A \cap [a, a] \subset \{a\}$ is finite, the "induction hypothesis" holds vacuously. (RI2): suppose $x \in S$. If $A \cap [a, x]$ is infinite, then by hypothesis $A \cap [a, x]$ has a limit point and hence so does $[a, b]$. So we may assume $A \cap [a, x]$ is finite. Now either there exists $\delta > 0$ such that $A \cap [a, x+\delta]$ is finite – which verifies our induction hypothesis – or every interval $[x, x+\delta]$ contains infinitely many points of A, in which case x is a limit point of A. (RI3): Suppose $[a, x) \subset S$. Then, if there exists some $y < x$ such that $A \cap [a, y]$ is infinite, then by hypothesis $A \cap [a, y]$ has a limit point and thus so does A. So we may assume that $A \cap [a, y]$ is finite for all $y < x$. As above, this means either that $A \cap [a, x]$ is finite, or that every interval $(x-\delta, x)$ intersects A, in which case x is a limit point of A. □

Remark: Theorem 127 is the "set version" of Bolzano-Weierstrass. There is a more common "sequence version" of Bolzano-Weierstrass: **every bounded sequence admits a convergent subsequence**. Sadly, we have not yet spoken of sequences, their convergence, and subsequences, but we will: this will be a major theme of the second half of the course. Once we acquire enough vocabulary to understand the above boldfaced statement, we will see almost immediately that it is equivalent to Theorem 127 above. We can (and will!) then use the Bolzano-Weierstrass Theorem for sequences to give new proofs of the Extreme Value Theorem and the Uniform Continuity Theorem which are quicker, cleaner and more conceptual.

CHAPTER 7

Differential Miscellany

1. L'Hôpital's Rule

1.1. The Cauchy Mean Value Theorem.

We present here a modest 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 certainly has its place: as we will soon see, it will be used to prove L'Hopital's Rule.

THEOREM 128. *(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 $g'(c) = 0$. With this value of 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 $h(a) = h(b)$.[1] 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?

[1] 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.

1.2. L'Hôpital's Rule.

We now 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 129. *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

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(t)}{g'(t)} < \beta \tag{20}$$

Suppose first that (i) holds. Then by letting x approach b in (20) 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 (20) and choose $c_1 \in (y, b)$ such that $c_1 < x < b$ implies $g(x) > g(y)$ and $g(x) > 0$. Multiplying (20) 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

$$\frac{f(x)}{g(x)} < \beta - \beta \frac{g(y)}{g(x)} + \frac{f(y)}{g(x)}. \tag{21}$$

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.[2] □

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!

[2] 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 $\lim_{x \to \infty}$.

Example: 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: ...

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 XX above without using L'Hôpital. There are any number of ways. For instance:

LEMMA 130. *(Racetrack Principle) Let $f, g : [a, \infty) \to \mathbb{R}$ be two differentiable functions, such that $f(a) \geq g(a)$ and $f'(x) \geq g'(x)$ for all $x \geq a$.*
a) Then $f(x) \geq g(x)$ for all $x \geq a$.
b) If $f(a) > g(a)$ or $f'(x) > g'(x)$ for all $x > a$, then $f(x) > g(x)$ for all $x > a$.

PROOF. Let $h = f - g$, so $h : [a, \infty) \to \mathbb{R}$ is differentiable with $h(a) \geq 0$ and $h'(x) \geq 0$ for all $x \geq a$. Then h is weakly increasing, so for all $x \geq a$ we have $h(x) \geq h(a) \geq 0$, i.e., $f(x) \geq g(x)$, establishing a). We leave it to the reader to check that under either of the additional hypotheses in part b) we can obtain strict inequality. □

PROPOSITION 131. *Let $a \in \mathbb{R}$ and let $f : [a, \infty) \to \mathbb{R}$ be a twice differentiable function such that $f'(x)$ and $f''(x)$ are both positive for all $x \geq a$. Then $\lim_{x \to \infty} f(x) = \infty$.*

PROOF. Let $g(x)$ be the tangent to F at $x = a$; by assumption it has slope $f'(a) > 0$, hence $\lim_{x \to \infty} g(x) = \infty$. Thus it suffices to show that $f(x) > g(x)$ for all $x > a$. Since $f'' > 0$, f' is increasing: in particular $f'(x) > f'(a) = g'(x)$ for all $x > a$. Now apply part b) of the Racetrack Principle. □

(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 131, $\log A = \infty$ 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 in fact rather complicated and easy to misremember/misapply. For instance, a classic blunder is to forget to check that condition (i) or (ii) applies: 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. There are some 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: 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,

$$\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} 2x \sin(\frac{1}{x}) - \cos(\frac{1}{x}) = - \lim_{x \to 0} \cos(\frac{1}{x}),$$

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...

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

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{22}$$

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:

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{2x_n^2 - (x_n^2 - 2)}{2x_n} = \frac{x_n^2 + 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right). \tag{23}$$

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[3] 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. This 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 (23) for successive approximations to $\sqrt{2}$ was known to the ancient Babylonians as a formula that worked well in computations, 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

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right),$$

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 100 decimal placees as part of your engineering research to build a raft that will carry you back to civilization...well, this is probably the method you should use.

2.3. Questioning Newton's Method.

Of course we haven't proven anything yet. Here are two natural questions:

QUESTION 2. *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 (including me) know little or nothing about this subject, which is a shame because the questions its treats are quite fundamental and closely related to many topics in pure mathematics.

2.4. Introducing Infinite Sequences.

We will give some answers to these questions – not the strongest possible answers, but a good start. 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.

[3] 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.

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 132. *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 132 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 132 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 133. *Let $f : I \to \mathbb{R}$ be differentiable, and suppose that 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 134. *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 $\mathbf{x}_1$ 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 $-\infty$.

Simple though this example was, there is plenty to learn from it. Here are three important observations:

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 $f : I \rightarrow I$. What about the last two observations? In fact they also hold quite generally.

LEMMA 135. *Let $f : I \rightarrow I$ be continuous, let $x_1 \in I$, and suppose that 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. b) Since f is continuous and $x_n \rightarrow L$, by Proposition 132 we have $f(x_n) \rightarrow T(L)$. But $f(x_n) = x_{n+1}$, and if $x_n \rightarrow L$ then certainly $x_{n+1} \rightarrow L$ as well. A sequence can have at most one limit, so $f(L) = L$. □

LEMMA 136. *(Contraction Lemma) For $c \in \mathbb{R}$ and $\delta > 0$, put $I = [c - \delta, c + \delta]$. Suppose that $f : I \rightarrow \mathbb{R}$ is a contraction with constant α and that $f(c) = c$.*
a) For all $x \in I$, $f(x) \in I$, i.e., $f : I \rightarrow I$.
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 $x_n \rightarrow c$.

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 $f : I \rightarrow I$.
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 $f : I \rightarrow \mathbb{R}$ 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 seems not so 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: we are assuming that $f'(c) \neq 0$, so perhaps $|f'(x)| > 1$ for all x. 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 137. *(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 $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_{n+1})}$ 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 136a), 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 136b), 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 137 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 137 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.

In fact, the right quantitative measure here is something called quadratic vonergence: we say that a sequence $\{x_n\}$ **quadratically converges** to L if if converges to L and there exists a constant C such that for all sufficiently large $n \in \mathbb{Z}^+$,

$$|x_{n+1} - L| \leq C|x_n - L|^2.$$

THEOREM 138. *Let $f : I \rightarrow \mathbb{R}$ be a C^2 function, i.e., f'' exists and is continuous. Let $c \in \mathbb{R}$ be such that $f(c) = 0$ and $f'(c) \neq 0$. Then there are $\delta, C > 0$ such that for all for all $x_1 \in [c - \delta, c + \delta]$ and all $n \in \mathbb{Z}^+$,*

$$|x_{n+1} - c| \leq C|x_n - c|^2.$$

The most natural way of proving this theorem requires the order two Taylor approximation, so we will defer the proof until we acquire that tool.

2.8. An example of nonconvergence of Newton's Method.

Consider the cubic function $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 and Q be distinct points in $\mathbb{R}^n$. There should be a unique line passing through P and Q. We can write this line down **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 $\bigcup_{i=1}^n \Omega_i$ – the set of all points lying in *at least one* Ω_i – need not be convex.

When $n = 1$, convex subsets are quite constrained. In fact, in the handout on completeness we proved the following theorem.

Theorem 139. *(Characterization of convex subsets of $\mathbb{R}$) For a subset $\Omega \subset \mathbb{R}$, the following are equivalent:*
(i) Ω is an interval.
(ii) Ω is convex.

The implication (i) $\implies$ (ii) is easy: for all $a < b$ in I and $\lambda \in [0,1]$, we have

$$a = (1-\lambda)a + \lambda a \leq (1-\lambda)a + \lambda b \leq (1-\lambda)b + \lambda b = b.$$

However the converse direction (ii) $\implies$ (i) cannot be so easy because it relies on the least upper bound axiom: for instance, working entirely in the rational numbers, the set $\Omega = \{x \in \mathbb{Q} \mid x^2 < 2\}$ is convex, but *as far as the rational numbers are concerned it is not an interval.* The idea is that it is “really” (pun intended?) the set of all rational numbers lying in $(-\sqrt{2}, \sqrt{2})$, but $\sqrt{2}$ is not a rational number, so living in the rational world we do not see an endpoint.

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

$$f((1-\lambda)a + \lambda b) \leq (1-\lambda)f(a) + \lambda f(b). \tag{24}$$

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 $a < x < b$,

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}. \tag{25}$$

A function $f : I \to \mathbb{R}$ satisfies the **two-secant inequality** if for all $a < x < b$,

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}. \tag{26}$$

Proposition 140. *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., $\operatorname{Epi}(f)$ *is a convex subset of* $\mathbb{R}^2$.

Proof. Strategy: 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 $P_1 = (x_1, y_1), P_2 = (x_2, y_2) \in \mathrm{Epi}(f)$. We want to show $\mathrm{Epi}(f)$ 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 $(1-\lambda_1)P_1 + \lambda_1 P_2 \notin \mathrm{Epi}(f)$: 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 $\mathrm{Epi}(f)$. Since $\mathrm{Epi}(f)$ is convex the line segment joining $(x, f(x))$ and $(y, f(y))$ lies inside $\mathrm{Epi}(f)$. 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. □

3.5. Continuity properties of convex functions.

Theorem 141. *Suppose $f : I \to \mathbb{R}$ is convex, and let $[a,b]$ be any subinterval of the interior of I. Then f is a* Lipschitz function *on $[a,b]$: there exists a constant C such that for all $x, y \in [a,b]$, $|f(x) - f(y)| \leq C|x-y|$.*

Corollary 142. *Let I be any* open *interval, and let $f : I \to \mathbb{R}$ be a convex function. Then f is continuous.*

Exercise: Prove Corollary 142. (Hint: If $f : I \to R$ is a function and $c \in I$ is such that there exists $\delta > 0$ such that $[c-\delta, c+\delta] \subset I$ and the restriction of f to $[c-\delta, c+\delta]$ is Lipschitz, then f is continuous at c.)

It is little more than a curiosity, but it turns out that a convex function need not be continuous on an endpoint of an interval on which it is defined (unless it is defined and convex on an open interval including that endpoint, of course). At least we can completely understand the situation.

Proposition 143. *Let $a \in \mathbb{R}$, and let I be an interval in $\mathbb{R}$ such that* $\inf I = a$ *but $a \notin I$.*[4] *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 $M = L$.

As we will not use Proposition 143 in what follows, we leave its proof as an exercise.

[4] In other words, I has one of the following forms: (a,b), $(a,b]$, (a,∞).

3.6. Differentiable convex functions.

THEOREM 144. *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): Applying the Mean Value Theorem on $[a, x]$ and on $[x, b]$, we get $c_1 \in [a, x]$ and $c_2 \in [x, b]$ such that

$$s(x) = \frac{f(x) - f(a)}{x - a} = f'(c_1) \leq f'(c_2) = \frac{f(b) - f(x)}{b - x} = S(x).$$

Thus $s(x) \leq S(x)$. This is the two-secant inequality, so f is convex. □

COROLLARY 145. *3.5.5 A differentiable convex function is C^1.*

PROOF. By Theorem 144, f' is weakly increasing and thus can have only jump discontinuities. But by Darboux's Theorem, a derivative cannot have jump discontinuities. So f' is continuous! □

COROLLARY 146. *(Kane Criterion) For twice differentiable $f : I \to \mathbb{R}$, TFAE:*
(i) f is convex.
(ii) $f'' \geq 0$.

PROOF. The function f' is weakly increasing iff $f'' \geq 0$. Now apply Theorem 144. □

3.7. An extremal property of convex functions.

THEOREM 147. *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. Since f is convex and differentiable, by Theorem 144 $f' : I \to \mathbb{R}$ 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 $\operatorname{Epi}(f)$ 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 148. *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 149. *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) $\Longrightarrow$ (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

$$0 = (\lambda_1+\lambda_2)f(\frac{\lambda_1}{\lambda_1+\lambda_2}(-\lambda_2\alpha) + \frac{\lambda_2}{\lambda_1+\lambda_2}(\lambda_1\alpha)) \leq \lambda_1 f(-\lambda_2\alpha)+\lambda_2 f(\lambda_1\alpha),$$

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) $\Longrightarrow$ (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 functins ℓ_c are certainly convex, f is the supremum of a family of convex functions, hence convex by Lemma 148. □

Theorem 150. *Let I be an open interval, $f : I \to \mathbb{R}$ be a convex function and $c \in I$.*
a) f is both left-differentiable and right-differentiable at c: $f'_-(c)$ and $f'_+(c)$ exist.
b) f'_-, $f'_+ : I \to \mathbb{R}$ are both weakly increasing functions.
c) We have $f'_-(c) \leq f'_+(c)$.
d) A line ℓ passing through $(c, f(c))$ is a supporting line for f iff its slope m satisfies

$$f'_-(c) \leq m \leq f'_+(c).$$

e) The following are equivalent:
(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.

Unfortunately the proof of Theorem 150 is a bit too much for us at this point: for instance, it involves the notion of a limit of a sequence of functions, which we will only discuss (perhaps briefly at that) at the end of the course. For now we refer the interested reader to [**Gr**, pp. 8-10].

Exercise: Use Theorem 150 to give a one-line proof of Theorem 147a).

3.9. Jensen's Inequality.

Theorem 151. *(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

$$f\left(\lambda_1 x_1 + \ldots + \lambda_n x_n \right) = f((1 - \lambda_{n+1})(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n) + \lambda_{n+1} x_{n+1} \right)$$

$$\leq (1 - \lambda_{n+1}) f \left(\frac{\lambda_1}{1 - \lambda_{n+1}} x_1 + \ldots + \frac{\lambda_n}{1 - \lambda_{n+1}} x_n \right) + \lambda_{n+1} f(x_{n+1}).$$

Now note that $\frac{\lambda_1}{1-\lambda_{n+1}}, \ldots, \frac{\lambda_n}{1-\lambda_{n+1}}$ are non-negative numbers that sum to 1, so 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 by Lemma XX 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.[5]

[5] 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.

THEOREM 152. *(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:*

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \ldots + \lambda_n x_n. \tag{27}$$

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 153. *(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*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{28}$$

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 154. *(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*

$$|x_1 y_1| + \ldots + |x_n y_n| \leq (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} (|y_1|^q + \ldots + |y_n|^q)^{\frac{1}{q}}. \tag{29}$$

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

$$x = \frac{|x_i|}{(|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}}, y = \frac{|y_i|}{(|y_1|^q + \ldots + |y_n|^q)^{\frac{1}{q}}},$$

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 155. *(Minkowski's Inequality) For* $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$ *and* $p \geq 1$*:*

$$(|x_1 + y_1|^p + \ldots + |x_n + y_n|^p)^{\frac{1}{p}} \leq (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} + (|y_1|^p + \ldots |y_n|^p)^{\frac{1}{p}} \tag{30}$$

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 156. *(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 157. *(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)

$$\mathcal{F}(x) - \mathcal{F}(c) = \int_a^x f - \int_a^c f = \int_c^x f. \tag{31}$$

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

$$|\int_A^B f| \leq M(B-A). \tag{32}$$

Now suppose $|x - c| < \delta = \frac{\epsilon}{M}$. Using (31) and then (32) 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 157 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 157b) to get that $\int_a^x f$ is an antiderivative of f.

Thus Theorem 157 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 "same 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

$$L(f, \mathcal{P}) \leq U(f, \mathcal{P}). \tag{33}$$

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 158. *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 159. *(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 160. *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 159 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 161. *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 160, 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 162. *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 163. *(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

$$U(f,\mathcal{P}) \leq U(f,\mathcal{P}_2) < \int_a^b f + \frac{\epsilon}{2} \tag{34}$$

and also

$$L(f,\mathcal{P}) \geq L(f,\mathcal{P}_1) > \int_a^b f - \frac{\epsilon}{2},$$

and thus

$$-L(f,\mathcal{P}) < \frac{\epsilon}{2} - \int_a^b f. \tag{35}$$

Adding (34) and (35) 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 161 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 164. *(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
(36)

$$U(f, \mathcal{P}_n) - L(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} (M_i(f) - m_i(f)) \left(\frac{b-a}{n} \right) \leq \left(\frac{b-a}{n} \right) \sum_{i=0}^{n-1} M_i(f) - m_i(f).$$

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

$$|M_i(f) - m_i(f)| = |f(d_i) - f(c_i)| < \frac{\epsilon}{b-a}. \tag{37}$$

Combining (36) and (37) 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 162, 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

$$(U(f, \mathcal{P}_1) - L(f, \mathcal{P}_1)) + (U(f, \mathcal{P}_2) - L(f, \mathcal{P}_1)) = (U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2)) - (L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2))$$
$$= U(f, \mathcal{P}_c) - L(f, \mathcal{P}_c) < \epsilon,$$

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

$$U(f, \mathcal{P}) - L(f, \mathcal{P}) = U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2) - (L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2))$$
$$= (U(f, \mathcal{P}_1) - L(f, \mathcal{P}_1)) + (U(f, \mathcal{P}_2) - L(f, \mathcal{P}_2)) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

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

$$L(f, \mathcal{P}) \leq L(f, \mathcal{P}_c) = L(f, \mathcal{P}_1) + L(f, \mathcal{P}_2) \leq \int_a^c f + \int_c^b f \leq U(f, \mathcal{P}_1) + U(f, \mathcal{P}_2)$$
$$= U(f, \mathcal{P}_c) \leq U(f, \mathcal{P}).$$

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 163 $\int_a^c f + \int_c^b f = \int_a^b f$.
c) This is immediate from Theorem 157 (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 165. *Let $f : [a,b] \to \mathbb{R}$ be a continuous function on a closed bounded interval. Then f is Darboux integrable.*

Proof. We will use Darboux's Integrability Criterion: we must show that for all $\epsilon > 0$, there exists 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 $b \in S(\epsilon)$, so it suffices to show $S(\epsilon) = [a,b]$. In fact it is necessary and sufficient: **observe** that if $x \in S(\epsilon)$ and $a \leq y \leq x$, then also $y \in S(\epsilon)$: you are asked to confirm this in an exercise. Our strategy – of course? – will be to show that $S(\epsilon) = [a,b]$ 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 $y \in S(\epsilon)$.

Remark: Spivak's uniform continuity-free proof of Theorem 165 [**S**, pp. 292-293] is really different from ours: inspired by the Fundamental Theorem of Calculus, 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 $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$, 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

$$\omega(f, J) \leq \omega(f, I). \tag{38}$$

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

$$\omega(f, c) = \lim_{\delta \rightarrow 0^+} \omega(f, [c-\delta, c+\delta]).$$

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 (38) 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 166. *Let I be an interval, $f : I \rightarrow \mathbb{R}$ 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 $\omega(f, c) = \lim_{\delta \rightarrow 0^+} \omega(f, [c-\delta, c+\delta]) = 0$. □

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 166 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 to 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 167. *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

[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 $\Delta(f, \mathcal{P})$!

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 $c \to a^+$. □

Remark: Of course there is an analogous result with the roles of a and b reversed.

Remark: 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 168. *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 167 (and its reflected version) to prove Corollary 168. (Suggestion: One can follow the ideas outlined above by establishing first the case of one discontinuity and then going by induction on the number of discontinuities. Alternately one can prove the result "directly" just by suitably partitioning $[a,b]$ into finitely many subintervals, each of which contains at most one discontinuity which must occur at one of the endpoints.)

THEOREM 169. *A weakly monotone function $f : [a,b] \to \mathbb{R}$ is Darboux integrable.*

PROOF. Recalling that $-f$ is weakly increasing iff f is weakly decreasing, 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 169 goes beyond Theorem 168 in that 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 170. *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

$$\Delta(f, \mathcal{P}_n) = \sum_{i=0}^{n-1} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i) < \frac{b-a}{n}. \tag{39}$$

Now (39) 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 166.
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 $f : [a,b] \rightarrow \mathbb{R}$ 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 171. *Let $f : [a,b] \rightarrow \mathbb{R}$ 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 171 different from Theorem 157c)? Only in a rather subtle way: if we try to apply Theorem 157c) to f', we see that we need it to be continuous, whereas in Theorem 171 we are assuming *only* that f' is Darboux integrable. Recall that every continuous function is Darboux integrable but as we saw in the previous section there are discontinuous Darboux integrable functions. But what about discontinuous, Darboux integrable *derivatives*? If you recall, 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 171 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'. In particular, for this we want an integration theory for which every derivative f' is an integrable function. As we have just seen, the Darboux integral is not such a theory. In graduate level real analysis, one learns of a more powerful and more 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 relatively 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 most of the familiar, elementary operations on integrable functions yields integrable functions. An important remark is that 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 172. *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

$$L(f, \mathcal{P}) + L(g, \mathcal{P}) \leq L(f + g, \mathcal{P}) \leq U(f + g, \mathcal{P}) \leq U(f, \mathcal{P}) + U(g, \mathcal{P}). \tag{40}$$

Moreover, subtracting the smallest quantity from the largest gives

$$0 \leq \Delta(f + g, \mathcal{P}) \leq \Delta(f, \mathcal{P}) + \Delta(g, \mathcal{P}),$$

and now since f and g are Darboux integrable, for $\epsilon > 0$, there exists $\mathcal{P}_1$ 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 $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, we get both inequalities and thus

$$\Delta(f + g, \mathcal{P}) \leq \Delta(f, \mathcal{P}) + \Delta(g, \mathcal{P}) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

By Darboux's Criterion, $f+g$ is Darboux integrable, and the inequalities (40) imply $\int_a^b f + g = \int_a^b f + \int_a^b g$. □

Remark: You may recall that I suggested this problem to you as the sort of integration problem you should be able to do on the final exam. But then I realized this was wrong: the argument above is harder than what I was envisioning. There is a reason behind my mistake: there is another, earlier, construction of the integral due to Riemann. Nine times out of ten Darboux's take on (what really is essentially Riemann's) integration theory comes out technically simpler than Riemann's. However this is an example of the tenth time: using the Riemann integral, part b) of the above result is no harder to prove than part a).

THEOREM 173. *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 173 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 3.4: The function $f : \mathbb{R} \to \mathbb{R}$ by $x \mapsto |x|$ is a Lipschitz function. Indeed, the reverse triangle inequalit 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 3.5: 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. (More precisely, show that the infimum C_0 of all Lipshitz contants for f is a Lipschitz constant for f.)

Proposition 174. *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 175. *Let $f : I \to [c, d]$ be bounded and $g : [c, d] \to \mathbb{R}$ a Lipshitz function with Lipschitz constant C. Then $\omega(g \circ f, I) \leq C\omega(f, I)$.*

Proof. We leave this as an exercise for the reader. □

We are now ready for our (much) easier proof of a weaker form of Theorem 172.

Theorem 176. *Let $f : [a, b] \to [c, d]$ be Darboux integrable, and let $g : [c, d] \to \mathbb{R}$ be a Lipschitz function 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 $[a, b]$ such that

$$\Delta(f, \mathcal{P}) = \sum_{i=0}^{n-1} \omega(f, [x_i, x_{i+1}])(x_{i+1} - x_i) < \frac{\epsilon}{C}.$$

Then by Lemma 175 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 177. *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 176 $g \circ f = |f|$ 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 178. *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 171, 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 176 applies to show that $(f_1+f_2)^2$ and $(f_1-f_2)^2$ are Darboux integrable, and then Theorem 172 appplies again to show that $f_1 f_2$ is Darboux integrable. □

Warning: Of course 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 proved that if g is continuous and f is Darboux integrable then (when defined) the composition $g \circ f$ is Darboux integrable, then we saw an example to show that if g and f are both merely Darboux integrable then $g \circ f$ 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*, but it takes some work to find an example: the easiest one I know is contained in a rather recent paper of Jitan Lu [**Lu99**]. (The example relies on the concept of *uniform convergence* of a sequence of functions, which is something we will cover towards the end of the course. Thus you would probably find [**Lu99**] tough going now, but if you are interested in such examples wait a few months and then take a look.) Lu also suggests the following positive result.

Theorem 179. *(Lu) Let $f : [a,b] \to [c,d]$ be a differentiable function such that f' is continuous and nonzero. Then for any Darboux integrable function $g : [c,d] \to \mathbb{R}$, $g \circ f : [a,b] \to \mathbb{R}$ is Darboux integrable.*

Lu leaves the proof to the reader...and so will I, at least for now.

4. Riemann Sums, Dicing, and the Riemann Integral

We now turn to the task of identifying G. Darboux's take on the integral – the one we have discussed so far – with B. Riemann's earlier work. One way to construe this is the following: Riemann gave an apparently different construction of an integral $\int_a^b f$ which also satisfies axioms (I0) through (I3). In virtue of the uniqueness of the integral of a continuous function, this Riemann integral $R\int_a^b f$ of a continuous function agrees with our previously constructed *Darboux integral* $\int_a^b f$. This leaves open however the question of how the class of "Riemann integrable functions" compares with the class of "Darboux integrable functions". It can be shown that – although the definitions look different – a function $f : [a, b] \to \mathbb{R}$ is Riemann integrable iff it is Darboux integrable and in this case the assigned numbers agree: $R \int_a^b f = \int_a^b f$. 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 $\mathcal{P}$,

$$L(f, \mathcal{P}) \leq R(f, \mathcal{P}, \tau) \leq U(f, \mathcal{P}). \tag{41}$$

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 $R(f, \mathcal{P}, \tau)$:

$$\sup_\tau R(f, \mathcal{P}, \tau) = U(f, \mathcal{P}), \ \inf_\tau R(f, \mathcal{P}, \tau) = L(f, \mathcal{P}). \tag{42}$$

Exercise 4.1: Show that (42) 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 (41) and (42) the following result follows almost immediately.

Theorem 180. *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 exists a partition $\mathcal{P}_\epsilon$ such that $\Delta(f, \mathcal{P}_\epsilon) = U(f, \mathcal{P}_\epsilon) - L(f, \mathcal{P}_\epsilon) < \epsilon$. As we have seen, 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 we have

$$L(f, \mathcal{P}) \leq \int_a^b f \leq U(f, \mathcal{P}).$$

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 $\int_a^b f$. □

Thus Theorem 180 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, one should not be too impressed by this: in the above proof we did little more than push around facts that we already know. On the other hand, Riemann himself 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 stronger convergence 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 $\max_{i=0}^{n-1} x_{i+1} - x_i$. In words, the mesh of $\mathcal{P}$ is the largest length of a subinterval in $\mathcal{P}$. One should think of the mesh of a partition as a better measure of its "size" than merely the number of points it contains: a partition of $[0, 1]$ could for instance contain one million points but no points in the subinterval $[\frac{1}{2}, 1]$, and a moment's thought shows that using such a partition will not give a good aproximation to the integral. One can also think of a kitchen assistant dicing carrots: just because you make a lot of knife cuts doesn't mean you have done a good dicing job: you might have some *tiny* pieces but a few remaining large pieces. A proper dicing involves making sure the mesh is sufficiently small.

Lemma 181. *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 182. *(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 $|\mathcal{P}| < \delta$,*

$$\underline{\int_a^b} f - L(f, \mathcal{P}) < \epsilon, \ U(f, \mathcal{P}) - \overline{\int_a^b} f < \epsilon. \tag{43}$$

Proof. (D. 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 $\mathcal{P}$. Choose $\delta > 0$ such that $2MN\delta < \frac{\epsilon}{2}$. We claim (43) 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

$$0 \leq \underline{\int_a^b} f - L(f, \mathcal{P}) = (\underline{\int_a^b} f - L(f, \mathcal{P}')) + (L(f, \mathcal{P}') - L(f, \mathcal{P})), \tag{44}$$

$$0 \leq U(f, \mathcal{P}) - \overline{\int_a^b} f = (U(f, \mathcal{P}') - \overline{\int_a^b} f) + (U(f, \mathcal{P}) - U(f, \mathcal{P}')). \tag{45}$$

We will establish the claim by showing that the two terms on the right hand side of (44) are each less than $\frac{\epsilon}{2}$ and, similarly, that the two terms on the right hand side of (45) are each less than $\frac{\epsilon}{2}$. Using the Refinement Lemma (Lemma 159), 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,

$$0 \leq L(f, \mathcal{P}') - L(f, \mathcal{P}) = \sum_{i=0}^{n-1} \left(L(f, \mathcal{P}'_i) - \inf(f, [x_i, x_{i+1}]) \right),$$

$$0 \leq U(f, \mathcal{P}) - U(f, \mathcal{P}') = \sum_{i=0}^{n-1} \left(\sup(f, [x_i, x_{i+1}]) - U(f, \mathcal{P}'_i) \right).$$

Becuase $\mathcal{P}'$ 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 in the above sums are zero. Further, each nonzero term in either sum above 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 (44) and (45) are each less than $\frac{\epsilon}{2}$. □

We can now deduce the main result of this section.

THEOREM 183. *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 182).
(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 180: 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 183 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 entire integration theory using the Riemann integral. Some contemporary texts take this approach as well, and indeed the last time I taught this material (early 2005, at McGill University) I started with the Riemann integral and developed the theory up to the Fundamental Theorem of Calculus. 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: for instance, I am not sure how to establish the *Riemann integrability* of monotone functions directly, i.e., without showing Darboux integrability instead and appealing to Theorem 183. Similarly, 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 slightly 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 180 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 much 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 textbook [**R**], W. Rudin gives the Darboux version of "the Riemann integral", but one of the exercises at the end of the chapter involves recognizing a certain limit as the limit of a sequence of Riemann sums and thereby equating it with the integral of a certain function: he's cheating, at least a little bit! 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 183, 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 183 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. Improper Integrals

5.1. Basic definitions and first examples.

5.2. 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 184. *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 184, 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

$$\int_1^\infty e^{-x^2} \leq \int_1^\infty e^{-x}.$$

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

[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."

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 185. *(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

$$\int_a^\infty f \leq \int_a^\infty g; \tag{46}$$

here each side is either a non-negative real number or ∞. From (46) both parts follow: if $\inf_a^\infty g < \infty$ then $\int_a^\infty f < \infty$, whereas if $\int_a^\infty f = \infty$ then $\int_a^\infty g = \infty$.[5] □

THEOREM 186. *(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.
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 $\int_a^\infty g < \infty$. □

Remark: Although from a strictly logical perspective part d) of Theorem 186 is the weakest, it is by far the most useful in practice.

[5] In fact parts a) and b) are contrapositives of each other, hence logically equivalent.

CHAPTER 9

Integral Miscellany

1. Some Antidifferentiation Techniques

1.1. Change of Variables.

1.2. Integration By Parts.

Proposition 187. *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

$$\int_0^\infty x^{n+1} e^{-x} dx = (n+1)x^n e^{-x}|_0^{oo} - \int_0^\infty (-e^{-x}(n+1)x^n) dx$$

$$= (0-0) + (n+1)\int_0^\infty x^n e^{-x} dx \overset{\text{IH}}{=} (n+1)n! = (n+1)!.$$

□

1.3. Reduction Formulas.

1.4. Integration of Rational Functions.

2. The Hermite-Hadamard Inequality

Theorem 188. *(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

$$\forall x \in [a,b],\ s(x) \leq f(x) \leq S(x). \tag{47}$$

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

$$\frac{\int_a^b s(x)}{b-a} = \frac{\int_a^b f(\frac{a+b}{2}) + m(x - \frac{a+b}{2})}{b-a} = f(\frac{a+b}{2}) + \frac{m}{b-a}\int_a^b (x - \frac{a+b}{2})$$

$$= f(\frac{a+b}{2}) + \frac{m}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2} - \frac{a+b}{2}(b-a)\right) = f(\frac{a+b}{2}) + \frac{m}{b-a} \cdot 0 = \frac{a+b}{2},$$

$$\frac{\int_a^b S(x)dx}{b-a} = \frac{\int_a^b (f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a))dx}{b-a} = \frac{f(a)+f(b)}{2}.$$

Substituting these evaluations of $\frac{\int_a^b s}{b-a}$ and $\frac{\int_a^b S}{b-a}$ into (47) 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 188 is not necessary: the inequality holds for any convex $f : [a,b] \rightarrow \mathbb{R}$.

3. Average Values

4. Jensen's Integral Equality

5. $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$

6. Applications to Probability

7. The Riemann-Lebesgue Lemma

Theorem 189. *(Riemann-Lebesgue Lemma) Let $f : [a,b] \rightarrow \mathbb{R}$ be Riemann integrable. Then:*
a) $\lim_{\lambda\rightarrow\infty} \int_a^b f(x)\cos(\lambda x)dx = 0.$
b) $\lim_{\lambda\rightarrow\infty} \int_a^b f(x)\sin(\lambda x)dx = 0.$

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 for the skeptical reader.

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. However, as λ increases the fact 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

$$|\int_a^b f(x)\cos(\lambda x)|dx \leq \int_a^b |f(x) - g(x)||\cos(\lambda x)|dx + |\int_a^b g(x)\cos(\lambda x)dx|$$

$$\leq \frac{\epsilon}{2} + |\sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} m_i \cos(\lambda x)dx| \leq \frac{\epsilon}{2} + |\sum_{i=0}^{n-1} \frac{m_i}{\lambda}(\sin(\lambda x_{i+1}) - \sin(\lambda x_i))|. \tag{48}$$

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 (48) 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

$$\lim_{x\to\infty} \frac{\log x}{x} = \frac{\infty}{\infty} \overset{\text{LH}}{=} \lim_{x\to\infty} \frac{\frac{1}{x}}{1} = \frac{1}{\infty} = 0.$$

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 one 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. Analogies Between the Continuous and the (Semi-)Discrete

2. Easy Facts

The following result collects some easy facts about convergence of infinite sequences.

THEOREM 190. *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 extremely familiar, and the ones that may not be are quite routine. In fact, every part of Theorem 190 holds verbatim for functions of a continuous variable approaching infinity. This means that 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. Really though you should just be honest with yourself: the parts of Theorem 190 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 190, but really this is almost obvous in any event.

PROPOSITION 191. *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 that we know that $L \in \mathbb{Z}$, 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 191 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. The following exercises follow up on this point.

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 think of 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 192. *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 193. *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 193.

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!

LEMMA 194. *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 195. *(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 195 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*[1], 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 196. *Let $\{a_n\}_{n=1}^\infty$ be a weakly increasing real 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 $L \in \mathbb{R}$, then the sequence converges to L.

[1] If you take this statement too seriously, you will end up regarding it as something like a definition of oscillation...

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 $a_n \rightarrow L$. 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 $a_n \rightarrow L$.
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 197. *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 $a_n \rightarrow L$.*

This is so important as to be worth spelling out very carefully. We get:

Theorem 198. *(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 199. *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 Theorem X.X, L must be 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 $n \in \mathbb{Z}^+$, $M_n = \min(M_{n-1}, y_n + \frac{1}{n})$. This is a decreasing sequence bounded below by any element of S, so by hypothesis it converges, say to M, and by Theorem X.X, M is the greatest lower bound of the set $\{M_n\}$. Moreover M must be the least upper bound 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

$$a_1, a_3, a_6, a_{100}, a_{103}, \ldots.$$

(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

$$-1, 1, -1, 1, \ldots \tag{49}$$

By selecting all the odd-numbered terms we get one subsequence:

$$-1, -1, -1, -1, \ldots \tag{50}$$

Similarly, by selecting all the even-numberd terms we get another subsequence:

$$1, 1, 1, 1, \ldots. \tag{51}$$

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 (49) is not convergent, due to oscillation. However, the subsequence (50) is constant, hence convergent to it constant value -1. Similarly, the subsequence (51) converges to its constant value 1.

Let's recap: we began with a sequence (51) 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

$$1, 2, 3, 4, \ldots \tag{52}$$

Here are some subsequences we can get:

$$1, 3, 5, 7, \ldots \tag{53}$$

$$2, 4, 6, 8, \ldots \tag{54}$$

$$1, 4, 9, 16, \ldots \tag{55}$$

$$1, 2, 4, 8, \ldots \tag{56}$$

And so forth. Indeed, the subsequences of (52) are precisely the increasing sequences with values in the positive integers. Note moreover that our sequence (52) 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 (53), (54), (55), (56), 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 200. *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 201. *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 weakly 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 202. *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. Because of its elegance, generality and usefulness, it has stayed with me through my entire adult mathematical career.

Lemma 203. *Every infinite sequence has a monotone subsequence.*

Proof. Let us say that $n \in \mathbb{N}$ is a **peak** of the sequence $\{a_n\}$ if for all $m < n$, $a_m < a_n$. 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 uevery 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 204. *(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**]) For all $1 \leq i \leq (r-1)(s-1)+1$, we label the ith term x_i of the sequence with the pair (a_i, b_i), where a_i is the longest length of a weakly increasing subsequence with final term x_i and b_i is the longest length of a weakly decreasing subsequence with final term x_i. Then every x_i gets a different label: indeed if $i < j$ and $x_i < x_j$ then $a_i < a_j$, whereas if $x_i > x_j$ then $b_i > b_j$. But the number of pairs (a_i, b_i) in which $a_i \leq r-1$ and $b_i \leq s-1$ is $(r-1)(s-1)$, and since we have one more term than this, we must have a pair labelled (a_i, b_i) with *either* $a_i \geq r$ – in which case we have a weakly increasing subsequence of length r – or $b_i \geq s$ – in which case we have a weakly decreasing subsequence of length s.
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 205. *(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 202, 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.[2]
(ii) Every bounded sequence admits a convergent subsequence.

6.3. Supplements to Bolzano-Weierstrass.

Theorem 206. *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 $\lim_{k\to\infty} x_{n_k} = \infty$. □

6.4. Applications of Bolzano-Weierstrass.

We will now give two quite different proofs of two of the three Interval Theorems.

Theorem 207. *(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 192 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 208. *(Uniform Continuity Theorem Again) Let $f : [a,b] \to \mathbb{R}$ be continuous. Then f is uniformly continuous.*

[2] Recall that we have already shown that this is equivalent to Dedekind completeness and thus holds *only* in the real numbers.

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. Applying the same result to the subsequence y_{n_k} there is a subsubsequence $y_{n_{k_l}}$ which is convergent. The notation for repeated subsequences is evidently rather cumbersome, so we drop it and simply write x_n and y_n, with the understanding that we have passed to a subsequence: notice that the inequalities $|x_n - y_n| < \frac{1}{n}$ and $|f(x_n) - f(y_n)| \geq \epsilon$ still hold. Now, suppose that $x_n \to L_1$ and $y_n \to L_2$. Then, since $|x_n - y_n| < \frac{1}{n}$ for all n, we have $L_1 - L_2 = \lim_{n \to \infty} x_n - y_n \to 0$, that is, $L_1 = L_2 = L$, say. But then $f(x_n) \to f(L)$ and $f(y_n) \to f(L)$, and since the sequences $f(x_n)$ and $f(y_n)$ have the same limit, we can make $|f(x_n) - f(y_n)| \leq |f(x_n) - L - (f(y_n)) - L)| \leq |f(x_n) - L| + |f(y_n) - L|$ arbitrarily small by taking n sufficient large. In particular, for sufficiently large n we have $|f(x_n) - f(y_n)| < \epsilon$: contradiction! □

Remark: Although I am very fond of the earlier Real Induction proof of the Uniform Continuity Theorem, I must admit that this argument seems shorter and easier to follow. 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.

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 209. *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 209. (Hint: this comes down to the fact that a subsequence of a subsequence is itself a subsequence.)

Theorem 210. *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 206a), ∞ is a partial limit. It $\{a_n\}$ is unbounded below, then by Theorem 206b) $-\infty$ 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 206, $\{a_n\}$ 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 209, 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 210b) 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 211. *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 211 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 212. *For any real sequence a_n, we have*

$$\overline{L} = \lim_{n\to\infty} \sup_{k \geq n} a_k \tag{57}$$

and

$$\underline{L} = \lim_{n\to\infty} \inf_{k \geq n} a_k. \tag{58}$$

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 208, $b_n \to L' \in [-\infty, \infty]$.

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 ±∞, 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 213. *A real sequence $\{a_n\}$ is convergent iff $\underline{L} = \overline{L} \in (-\infty, \infty)$.*

Exercise: Prove Corollary 213.

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 convergentsequence 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 $x \rightarrow c$" also has this feature: what we actually defined was $\lim_{x \rightarrow c} f(x) = L$. For continuity the issue is not pressing because we know that we want the limit to be $f(c)$.[3] 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 of integrals 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 halfway in between in that the definition includes a particular limiting value in it 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 what number* the integral should be. And that's

[3] By the way, I think this is a large part of the explanation for why the concept of continuity is actually simpler than the concept of limit!

what made Darboux's criterion so useful: we used it to show that every continuous function and every weakly 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 remedy this situation.)

The upshot is that 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 real 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.[4]

Lemma 214. *A subsequence of a convergent sequence is Cauchy.*

Proof. Left to the reader as an exercise. □

Proposition 215. *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 216. *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 217. *A Cauchy sequence which admits a convergent subsequence is itself convergent.*

Proof. Let $\{a_n\}$ be a Cauchy sequence in the ordered field F and suppose that there exists 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

$$|a_n - L| = |(a_n - a_{n_N}) - (a_{n_N} - L)| \leq |a_n - a_{n_N}| + |a_{n_N} - L| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

[4] You might notice that they are quite reminiscent of the elementary properties of *monotone* sequences we proved above. I must admit that I myself observed this only quite recently – in particular, more than 15 years after I first learned about monotone sequences and Cauchy sequences – and I wish I had a deeper understanding of the underlying source – if any! – of these commonalities.

In the above results we *have not* used the completeness axiom for real numbers, so in fact they hold in any ordered field.

THEOREM 218. *Any real Cauchy sequence is convergent.*

PROOF. Let $\{a_n\}$ be a real Cauchy sequence. By Proposition 216, $\{a_n\}$ is bounded. By Bolzano-Weierstrass there exists a convergent subeqeuence. Finally, by Proposition 217, this implies that $\{a_n\}$ is convergent. □

In contrast to the preceding results, Theorem 218 *does* crucially use the Dedekind completeness of the real numbers, in the form of the Bolzano-Weierstrass Theorem (which we showed holds in an ordered field iff that field is Dedekind complete).

8.3. Sequential completeness versus Dedekind completeness I: the Archimedean case.

Of course, just because one proof of a theorem uses a certain result does not mean that all proofs need to. To further investigate these matters, let us say that an ordered field is **sequentially complete** if every Cauchy sequence in that field is convergent. Then we have the following result.

THEOREM 219. *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 218, 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

$$0 \leq a_m - a_n = (a_m - a_{m+1}) + (a_{m+1} - a_{m+2}) + \ldots + (a_{n-1} - a_n)$$
$$\leq 2^{-(m+1)} + \ldots + 2^{-n} = 2^{-m}.$$

This shows that $\{a_n\}$ is a Cauchy sequence, 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 219 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**].

8.4. Sequential Completeness versus Dedekind completeness II: the non-Archimedean case.

We have just seen that in an *Archimedean* ordered field, all Cauchy sequences converge iff the field satisfies (LUB). This makes one wonder what happens if the hypothesis of "Archimedean" is dropped. In fact there are many non-Archimedean (hence not Dedekind complete) fields in which all Cauchy sequences converge. We will attempt to give 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 read on...

Example: Let $F = \mathbb{R}((t))$ be the field of **formal Laurent series** with $\mathbb{R}$-coefficients: an element of F 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. It is not so hard to show that K is actually a field: we skip this.

We need to equip K with an ordering; equivalently, we need to specify a set of positive elements. For every nonzero element $x \in F$, we put $v(x)$ 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 $x \in F$, exactly one of x and $-x$ is positive, so this gives an ordering on F 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 F. 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 F 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.

In the above example of a non-Archimedean sequentially complete ordered field, there were plenty of convergent sequences, but they all took a rather simple form that enabled us to show that the condition for convergence was the same as the Cauchy criterion. It is possible for a non-Archimedean field to be sequentially complete in a completely different way: there are non-Archimedean fields for which every Cauchy sequence is eventually constant. Certainly every eventually constant sequence is convergent, so such a field F must be sequentially complete!

In fact a certain property will imply that every Cauchy sequence is eventually constant. It is the following: **every sequence in F is bounded**. This is a “weird” property for an ordered field to have: certainly no Archimedean ordered field has this property, because by definition in an Archimedean field the sequence $x_n = n$ is unbounded. And there are plenty of non-Archimedean fields which *do not* have this property, for instance the field $\mathbb{R}((t)))$ discussed above, in which $\{t^{-n}\}$ is an unbounded sequence. Nevertheless there are such fields.

Suppose F is an ordered field in which every sequence is bounded, and let $\{x_n\}$

be a Cauchy sequence in F. Seeking a contradiction, we suppose that $\{x_n\}$ is *not* eventually constant. Then there is a subsequence which takes distinct values, i.e., for all $k \neq k'$, $x_{n_k} \neq x_{n_{k'}}$, and a subsequence of a Cauchy sequence is still a Cauchy sequence. Thus if there is a non-eventually constant Cauchy sequence, there is a noneventually constant Cauchy sequence with distinct terms, so we may assume from the start that $\{x_n\}$ has distinct terms. Now for $n \in \mathbb{Z}^+$, put $z_n = |x_{n+1} - x_n|$, so $z_n > 0$ for all n, hence so is $Z_n = \frac{1}{z_n}$. By our assumption on K, the sequence Z_n is bounded: there exists $\alpha > 0$ such that $Z_n < \alpha$ for all n. Now put $\epsilon = \frac{1}{\alpha}$. Taking reciprocals we find that for all $n \in \mathbb{Z}^+$,

$$|x_{n+1} - x_n| = z_n = \frac{1}{Z_n} > \frac{1}{\alpha} = \epsilon.$$

This shows that the sequence $\{x_n\}$ is not Cauchy and completes the argument.

It remains to construct an ordered field having the property that every sequence is bounded. At the moment the only constructions I know are wildly inappropriate for an undergraduate audience. For instance, one can start with the real numbers $\mathbb{R}$ and let K be an **ultrapower** of $\mathbb{R}$ corresponding to a nonprincipal ultrafilter on the positive integers. The details of this construction are wildly outside the scope of these notes.

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 arow never hits its target. Similarly he deduced that all motion is impossible.

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 for simplicity that 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 going on here: we have divided the total time of the trip into infinitely many parts, and the conclusion seems to be that

$$\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 1. \tag{59}$$

[1] Zeno of Elea, ca. 490 BC - ca. 430 BC.

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 (59)? 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 fist 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** (or sometimes **summable**, although this term will not be used here); 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

$$S_n = \frac{1}{2} + \ldots + \frac{1}{2^n}. \tag{60}$$

There is a standard trick for evaluating such finite sums. Namely, multiplying (60) by $\frac{1}{2}$ and subtracting it from (60) 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. Geometric Series.

Recall that a **geometric sequence** is a sequence $\{a_n\}_{n=0}^{\infty}$ of nonzero real numbers such that the ratio between successive terms $\frac{a_{n+1}}{a_n}$ is equal to some fixed number r, the **geometric ratio**. In other words, if we write $a_0 = A$, then for all $n \in \mathbb{N}$ we have $a_n = Ar^n$. A geometric sequence with geometric ratio r converges to zero if $|r| < 1$, converges to A if $r = 1$ and otherwise diverges.

We now define a **geometric series** to be an infinite series whose terms form a geometric sequence, thus a series of the form $\sum_{n=0}^{\infty} Ar^n$. Geometric series will play a singularly important role in the development of the theory of all infinite series, so we want to study them carefully here. In fact this is quite easily done.

Indeed, for $n \in \mathbb{N}$, put $S_n = a_0 + \ldots + a_n = A + Ar + \ldots + Ar^n$, the nth partial sum. It happens that we can give a closed form expression for S_n, for instance using a technique the reader has probably seen before. Namely, consider what happens when we multiply S_n by the geometric ratio r: it changes, but in a very clean way:

$$S_n = A + Ar + \ldots + Ar^n,$$

$$rS_n \quad = \qquad Ar + \ldots + Ar^n + Ar^{n+1}.$$

Subtracting the two equations, we get

$$(r-1)S_n = A(r^{n+1} - 1)$$

and thus

$$S_n = \sum_{k=0}^{n} Ar^k = A\left(\frac{1-r^{n+1}}{1-r}\right). \tag{61}$$

Note that the division by $r - 1$ is invalid when $r = 1$, but this is an especially easy case: then we have $S_n = A+A(1)+\ldots+A(1)^n = (n+1)A$, so that $|\lim_{n\to\infty} S_n| = \infty$ and the series diverges. For $r \neq 1$, we see immediately from (61) that $\lim_{n\to\infty} S_n$ exists if and only if $\lim_{n\to\infty} r^n$ exists if and only if $|r| < 1$, in which case the latter limit is 0 and thus $S_n \to \frac{A}{1-r}$. We record this simple computation as a theorem.

THEOREM 220. *Let A and r be nonzero real numbers, and consider the geometric series $\sum_{n=0}^{\infty} Ar^n$. Then the series converges if and only if $|r| < 1$; in which case the sum is $\frac{A}{1-r}$.*

Exercise: Show that for any $N \in \mathbb{Z}$, $\sum_{n=N}^{\infty} Ar^n = \frac{Ar^N}{1-r}$.

Exercise: Many ordinary citizens are uncomfortable with the identity

$$0.999999999999\ldots = 1.$$

Interpret it as a statement about geometric series, and show that it is correct.

1.3. 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 $n \in \mathbb{Z}^+$,

$$S_n = \sum_{k=1}^{n} \frac{1}{k^2+k} = \frac{n}{n+1}. \tag{62}$$

Indeed this is true when $n = 1$: both sides equal $\frac{1}{2}$. Now suppose that (62) holds for some $n \in \mathbb{Z}^+$; we wish to show that it also holds for $n+1$. But indeed we may just calculate:

$$S_{n+1} = S_n + \frac{1}{(n+1)^2+(n+1)} \overset{\text{IH}}{=} \frac{n}{n+1} + \frac{1}{n^2+3n+2} = \frac{n}{n+1} + \frac{1}{(n+1)(n+2)}$$

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

This proves the result.

As above, this is certainly a way to go, and the general technique 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 hope so, yes – recall from calculus that every rational function admits a **partial fraction 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.)[2]

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 3. *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 ±∞ 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 3b) – 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 3b) 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

[2] 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.

The simplest and most useful of these is simply that 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 $n \in \mathbb{Z}^+$

$$\left(\sum_{k=1}^{N} a_k\right) + T_n = a_1 + \ldots + a_N + a_{N+1} + \ldots + a_{N+n} = S_{N+n}.$$

It follows that if $\lim_{n\to\infty} T_n = \sum_{n=N+1}^{\infty} a_n$ exists, then so does $\lim_{n\to\infty} S_{N+n} = \lim_{n\to\infty} S_n = \sum_{n=1}^{\infty} a_n$ exists. 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 221. *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 seies $\sum_{n=1}^{\infty} a_n$ 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 222. *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 $\sum_{n=1}^{\infty} a_n = S$ is convergent, then so also is $\sum_{n=1}^{\infty} \alpha a_n$, with sum αS.

PROOF. a) Let $A_n^a = a_1 + \ldots + a_n$, $B_n^b = b_1 + \ldots + b_n$ and $C_n = a_1 + b_1 + \ldots + a_n + b_n$. By definition of convergence of infinite series we have $A_n \to S_a$ and $B_n \to B$. Thus for any $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ 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) We leave the case $\alpha = 0$ to the reader as an (easy) exercise. So suppose that $\alpha \neq 0$ and put $S_n = a_1 + \ldots + a_n$, and our assumption that $\sum_{n=1}^{\infty} a_n = S$ implies that for all $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|S_n - S| < \frac{\epsilon}{|\alpha|}$. It follows that

$$|\alpha a_1 + \ldots + \alpha a_n - \alpha S| = |\alpha||a_1 + \ldots + a_n - S| = |\alpha||S_n - S| < |\alpha| \left(\frac{\epsilon}{|\alpha|}\right) = \epsilon.$$

□

Exercise: Let $\sum_n a_n$ be an infinite series and $\alpha \in \mathbb{R}$.
a) If $\alpha = 0$, show that $\sum_n \alpha a_n = 0$.[3]
b) Suppose that $\alpha \neq 0$. Show that $\sum_n a_n$ converges iff $\sum_n \alpha a_n$ converges. Thus multiplying every term of a series by a nonzero real number does not affect its convergence.

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.

The following result is the first of the "convergence tests" that one encounters in freshman calculus.

THEOREM 223. *(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

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} S_n - S_{n-1} = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = S - S = 0.$$

□

The result is often applied in its contrapositive form: if $\sum_n a_n$ is a series such that $a_n \nrightarrow 0$ (i.e., either a_n converges to some nonzero number, or it does not converge), then the series $\sum_n a_n$ diverges.

Warning: The converse of Theorem 223 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, i.e., a quotient of polynomials with real coefficients (and, to be completely precise, such that Q is not the identically zero polynomial!). 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 the degree of $P(x)$ is at least at large as the degree of $Q(x)$, then $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ is divergent.

2.2. The Cauchy criterion for convergence.

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.

[3] This is a rare case in which we are interested in the sum of the series but the indexing does not matter!

Proposition 224. *(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 223). 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 condition 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.

Proposition 225. *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? Well, assume $a_n \geq 0$ for all $n \in \mathbb{Z}^+$ and consider the sequence of partial sums. We have

$$S_1 = a_1 \leq a_1 + a_2 = S_2 \leq a_1 + a_2 + a_3 = S_3,$$

and so forth. In general, we have that $S_{n+1} - S_n = a_{n+1} \geq 0$, so that the sequence of partial sums $\{S_n\}$ is increasing. Applying the Monotone Sequence Lemma we immediately get the following result.

Proposition 226. *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 227. *(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 with non-negative terms. (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. Here is one for starters.

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 $e = 2.718281828459045235360287 4714\ldots$, 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 228. *(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 228.

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 229. *(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 $n \geq N$, $0 \leq a_n \leq M b_n$. Then if $\sum_n b_n$ converges, $\sum_n a_n$ converges.*

Exercise: Prove Theorem 229.

Corollary 230. *(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 convege 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 three cases we deduce the result from the Limit Comparison Test (Theorem 229).
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 229 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 229 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 229 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 Theorem 230.

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 $\frac{1}{n^2}$. Indeed, the precies statement of this is that 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 230, 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 Direct Comparison Test does not 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 that $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$. (Recall from Exercise X.X our convention that we choose N to be larger than all the roots of $Q(x)$, so that every term of the series is well-defined.)

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:

$$(\sum_{n=0}^{\infty} a_n) \cdot (\sum_{n=0}^{\infty} b_n) \overset{??}{=} \sum_{n=0}^{\infty} a_n b_n.$$

In other words, given two sequences of terms $\{a_n\}$, $\{b_n\}$, we form a new sequence of terms $\{a_nb_n\}$ and then we form the associated series. In fact this is not a very useful candidate for the product. What we surely want to happen is that if $\sum_n a_n = A$ and $\sum_n b_n = B$ then our "product series" should 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 $\sum_{n=0}^{\infty} a_nb_n = \sum_{n=0}^{\infty} \frac{1}{4^n} = \frac{1}{1-\frac{1}{4}} = \frac{4}{3}$. Of course $\frac{4}{3} < 4$. 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 lager – 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.[4]

Let us try again at formally multiplying out a product of infinite series:

$$(a_0 + a_1 + \ldots + a_n + \ldots)(b_0 + b_1 + \ldots + b_n + \ldots)$$

$$= a_0b_0 + a_0b_1 + a_1b_0 + a_0b_2 + a_1b_1 + a_2b_0 + \ldots + a_0b_n + a_1b_{n-1} + \ldots + a_nb_0 + \ldots.$$

So it is getting a bit notationally complicated. In order to shoehorn the right hand side into a single infinite series, we need to either (i) choose some particular ordering to take the terms a_kb_ℓ on the right hand side, or (ii) collect some terms together into an nth term.

For the moment we choose the latter: we define for any $n \in \mathbb{N}$

$$c_n = \sum_{k=0}^{n} a_kb_{n-k} = a_0b_n + a_1b_{n-1} + \ldots + a_nb_n$$

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

$$\sum_{n=0}^{\infty} c_n = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_kb_{n-k}\right).$$

THEOREM 231. *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 $c_n = \sum_{k=0}^{n} a_kb_{n-k}$ 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. It is instructive to define yet another sequence, the **box product**, as follows: for all $N \in \mathbb{N}$,

$$\square_N = \sum_{0\leq i,j\leq N} a_ib_j = (a_0 + \ldots + a_N)(b_0 + \ldots + b_N) = A_NB_N.$$

Thus by the usual product rule for sequences, we have

$$\lim_{N\to\infty} \square_N = \lim_{N\to\infty} A_NB_N = AB.$$

[4] 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.

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:

$$\Box_N = a_0b_0 + a_0b_1 + \ldots + a_0b_N$$

$$+a_1b_0 + a_1b_1 + \ldots + a_1b_N$$

$$\vdots$$

$$+a_Nb_0 + aNb_1 + \ldots + a_Nb_N.$$

On the other hand, the terms of the Nth partial sum of the Cauchy product can naturally be arranged in a triangle:

$$C_N = \qquad a_0b_0$$

$$+a_0b_1 + a_1b_0$$

$$+ a_0b_2 + a_1b_1 + a_2b_0$$

$$+a_0b_3 + a_1b_2 + a_2b_1 + a_3b_0$$

$$\vdots$$

$$+a_0b_N + a_1b_{N-1} + a_2b_{N-2} + \ldots + a_Nb_0.$$

Thus while $\Box_N$ 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_ib_j with $i+j > N$ – as N gets large.

Here, since all the a_i's and b_j's are non-negative and $\Box_N$ contains all the terms of C_N and others as well, we certainly have

$$C_N \leq \Box_N = A_NB_N \leq AB.$$

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 $\Box_N$ is of the form a_ib_j with $0 \leq i, j \leq N$; thus $i + j \leq 2N$ so a_ib_j appears as a term in C_{2N}. It follows that $C_{2N} \geq \Box_N$ and thus

$$C = \lim_{N\to\infty} C_N = \lim_{N\to\infty} C_{2N} \geq \lim_{N\to\infty} \Box_N = \lim_{N\to\infty} A_NB_N = AB.$$

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 is in fact no loss of generality to assume that $a_n > 0$ for all n. Indeed, if not we have $a_N = 0$ for some N and then since the terms are assumed to be 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. Now observe that the power of $\frac{1}{2}$ which beings 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 such that the degree of the denominator is exactly one greater than the degree of the numerator. Show that $\sum_{n=N}^{\infty} \frac{P(n)}{Q(n)}$ diverges.

We deduce the following result.

Proposition 232. *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. Condensation Tests.

The apparently *ad hoc* argument used to prove the divergence of the harmonic series can be adapted to give the following useful test.

Theorem 233. *(Cauchy 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

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + \ldots$$

$$\leq a_1 + a_2 + a_2 + a_4 + a_4 + a_4 + a_4 + 8a_8 + \ldots = \sum_{n=0}^{\infty} 2^n a_{2^n}$$

$$= (a_1 + a_2) + (a_2 + a_4 + a_4 + a_4) + (a_4 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8) + (a_8 + \ldots)$$

$$\leq (a_1 + a_1) + (a_2 + a_2 + a_3 + a_3) + (a_4 + a_4 + a_5 + a_5 + a_6 + a_6 + a_7 + a_7) + (a_8 + \ldots)$$

$$= 2\sum_{n=1}^{\infty} a_n.$$

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**[5] $\sum_{n=1}^{\infty} \frac{1}{n^p}$. 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 234. *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

$$1 \leq \sum_{n=1}^{1024} \frac{1}{n^2} = 1.643957981030164240100762569\ldots.$$

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

$$\sum_{n=2^N+1}^{\infty} a_n \leq \sum_{n=0}^{\infty} 2^n a_{2^{n+N}}.$$

[5] Or sometimes: **hyperharmonic series**.

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 $\epsilon > 0$, $\sum_n \frac{1}{nn^{\epsilon}}$ 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 235. *(Integral Test) Let $f : [1,\infty) \to \mathbb{R}$ be a positive decreasing function, and for $n \in \mathbb{N}$ 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[6] [**Ma42**] and later in more modern form to A.L. Cauchy [**Ca89**]. I don't know why it is traditional to attach Cauchy's

[6] Colin Maclaurin, 1698-1746

name to the Condensation Test but not the Integral Test, but I have preserved the tradition nevetheless.

It happens that, at least among the series which arise naturally in calculus and undergraduate analysis, it is usually the case 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.

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.

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.[7]

LEMMA 236. *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. That $\sum_n a_n = \infty$ follows from the Limit Comparison Test. Now fix $\epsilon > 0$ and choose $K \in \mathbb{N}$ such that for all $n \geq K$ we have $a_n \leq (1+\epsilon)b_n$. Then for $N \geq K$,

$$\sum_{n=1}^N a_n = \sum_{n=1}^{K-1} a_n + \sum_{n=K}^N a_n \leq \sum_{n=1}^{K-1} a_n + \sum_{n=K}^N (1+\epsilon)b_n$$

$$= \left(\sum_{n=1}^{K_1} a_n - \sum_{n=1}^{K-1}(1+\epsilon)b_n\right) + \sum_{n=1}^N (1+\epsilon)b_n = C_{\epsilon,K} + (1+\epsilon)\sum_{n=1}^N b_n,$$

[7] Our treatment of the next two results owes a debt to K. Conrad's *Estimating the Size of a Divergent Sum.*

say, where $C_{\epsilon,K}$ is a constant independent of N. Dividing both sides by $\sum_{n=1}^N b_n$ and using the fact that $\lim_{N\to\infty}\sum_{n=1}^N b_n = \infty$, we find that the quantity $\frac{\sum_{n=1}^N a_n}{\sum_{n=1}^N b_n}$ 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 237. *Let $f : [1,\infty) \to \mathbb{R}$ be a positive monotone continuous function. 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

$$\int_n^{n+1} f(x)dx \sim f(n). \tag{63}$$

Applying Lemma 236 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

$$\lim_{N\to\infty}\frac{\int_1^{N+1} f(x)dx}{\int_1^N f(x)dx} = \frac{\infty}{\infty} \overset{*}{=} \lim_{N\to\infty}\frac{f(N+1)}{f(N)} = 1,$$

where in the starred equality we have applied L'Hopital's Rule and then the Fundamental Theorem of Calculus. We conclude

$$\int_1^N f(x)dx \sim \int_1^{N+1} f(x)dx \sim \sum_{n=1}^N f(n),$$

as desired.
Case 2: Suppose f is decreasing. Then for $n \leq x \leq n+1$, we have

$$f(n+1) \leq \int_n^{n+1} f(x)dx \leq f(n),$$

or

$$\frac{f(n+1)}{f(n)} \leq \frac{\int_n^{n+1} f(x)dx}{f(n)} \leq 1.$$

Once again, by our assumption that $f(n) \sim f(n+1)$ and the Squeeze Principle we get (63); 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 $a_n \geq 0$ 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 $0 \leq r < 1$, 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 $r \geq 1$, then the series diverges by comparison to the divergent geometric series $\sum_n Mr^n$.

5.1. The Ratio Test.

THEOREM 238. *(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 $n \geq N$, $\frac{a_{n+1}}{a_n} \leq r$. Then the series $\sum_n a_n$ converges.
b) Suppose there exists $N \in \mathbb{Z}^+$ and $r \geq 1$ such that for all $n \geq N$, $\frac{a_{n+1}}{a_n} \geq r$. 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 $n \geq N$, $\frac{a_{n+1}}{a_n} \leq r < 1$. Then $\frac{a_{n+2}}{a_n} = \frac{a_{n+2}}{a_{n+1}} \frac{a_{n+1}}{a_n} \leq r^2$. An easy induction argument shows that for all $k \in \mathbb{N}$,

$$\frac{a_{N+k}}{a_N} \leq r^k,$$

so

$$a_{N+k} \leq a_N r^k.$$

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 238.

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 sequence has the property that the nth root of the $n$rth term is equal to r. Suppose now that $\sum_n$ is a series with non-negative terms with the property 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 239. *(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 239.

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.[8]

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, $x_n \geq 1$.

[8] This is not a typo: we really mean the limsup both times, unlike in the previous exercise.

(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 239.

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 240. *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: Since for any sequence $\{x_n\}$ we have $\liminf x_n \leq \limsup x_n$, we certainly have $\underline{\theta} \leq \overline{\theta}$.
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. □

Exercise: Give the details of Step 3 in the proof of Proposition 240.

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 240, we have $\overline{\theta} \leq \overline{\rho} \leq 1$, so the Root Test also shows that the series is convegent. 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. In fact, turning these ideas around, Proposition 240 can be put to the following sneaky use.

Corollary 241. *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 240 we must also have $\theta = L$![9] □

Exercise: Use Corollary 241 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 when we do not have absolute convergence. As the reader will see, this leads to surprisingly delicate and intricate considerations: in practice, we very much hope that our series are absolutely convergent!)

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**.[10]

[9] There is something decidedly strange about this argument: to show something about a sequence $\{a_n\}$ we reason in terms of the corresponding infinite series $\sum_n a_n$. But it works!

[10] 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.

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 242. *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 a little 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

$$|\sum_{n=N}^{\infty} a_n| \leq \sum_{n=N}^{\infty} |a_n| < \epsilon,$$

and $\sum_n a_n$ converges by the Cauchy criterion. □

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.

As an example of how Theorem 242 may be combined with the previous tests to give tests for absolute convergence, we record the following result.

Theorem 243. *(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 238 (resp. Theorem 239) we find that $\sum_n |a_n|$ is convergent – and the point is that by Theorem 242, 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 $b_n \nrightarrow 0$. Thus under the hypotheses of parts b) and d) we have $|a_n| \nrightarrow 0$, hence also $a_n \nrightarrow 0$ so $\sum_n a_n$ diverges by the Nth Term Test (Theorem 223). □

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 244. *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 $|C_N - AB| < \epsilon$. □

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.[11]

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.

7.1. The Alternating Series Test.

Consider the **alternating harmonic series**

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

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 decreasing:

$$a_1 \geq a_2 \geq \ldots \geq a_n \geq \ldots.$$

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}:

$$S_{2n+1} = S_{2n-1} - a_{2n} + a_{2n+1} \leq S_{2n+1}.$$

It follows that the sequence of odd-numbered partial sums $\{S_{2n-1}\}$ is decreasing. Moreover,

$$S_{2n+1} = (a_1 - a_2) + (a_3 - a_4) + \ldots + (a_{2n-1}| - |a_{2n}) + a_{2n-1} \geq 0.$$

[11] 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.

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. By X.X, the full sequence $\{S_n\}$ converges iff $S_{\text{odd}} = S_{\text{even}}$. Now 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.

In fact these inequalities give something else: 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 245. *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of non-negative real numbers which is 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

$$E_N = |(\sum_{n=1}^{\infty}(-1)^{n+1}a_n) - (\sum_{n=1}^{N}(-1)^{n+1}a_n)|. \tag{64}$$

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 (64) 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 245b): 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. (Let us agree that to "compute a number α to k decimal place accuracy means to compute it with error less than 10^{-k}. A little thought shows that this is not quite enough to guarantee that the first k decimal places of the approximation are equal to the first k decimal places of α, but we do not want to occupy ourselves with such issues here.) By Theorem 245b), it is enough to find an $N \in \mathbb{Z}^+$ 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

$$\sum_{n=1}^{1000} \frac{(-1)^{n+1}}{n} = 0.6926474305598203096672310589\ldots.$$

Again, later we will show that the exact value of the sum is $\log 2$, which my software package tells me is[12]

$$\log 2 = 0.6931471805599453094172321214.$$

Thus the *actual error* in cutting off the sum after 1000 terms is

$$E_{1000} = 0.0004997500001249997500010625033.$$

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

[12] Yes, you should be wondering how it is computing this! More on this later.

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 245b) 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

$$\sum_{n=0}^{9} \frac{(-1)^n}{n!} = 0.3678791887125220458553791887.$$

In this case the exact value of the series is

$$\frac{1}{e} == 0.3678794411714423215955237701$$

so the true error is

$$E_9 = 0.0000002524589202757401445814516374,$$

which this time is only very slightly less than the guaranteed

$$\frac{1}{10!} = 0.0000002755731922398589065255731922.$$

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.[13]

Theorem 246. *(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 $n > m \geq N$,

$$|\sum_{k=m}^{n} a_k b_k| = |\sum_{k=m}^{n} a_k(B_k - B_{k-1})| = |\sum_{k=m}^{n} a_k B_k - \sum_{k=m-1}^{n-1} a_{k+1} B_k|$$

[13] Johann Peter Gustav Lejeune Dirichlet, 1805-1859

$$= |\sum_{k=m}^{n-1}(a_k - a_{k+1})B_k + a_nB_n - a_mB_{m-1}|$$

$$\leq \left(\sum_{k=m}^{n-1}|a_k - a_{k+1}||B_k|\right) + |a_n||B_n| + |a_m||B_{m-1}|$$

$$\leq M(\sum_{k=m}^{n-1}|a_k - a_{k+1}|) + |a_n| + |a_m| = M\left(\sum_{k=m}^{n-1}(a_k - a_{k+1}) + a_n + a_m\right)$$

$$= M(a_m - a_n + a_n + a_m) = 2Ma_m \leq 2Ma_N < \epsilon.$$

Therefore $\sum_n a_nb_n$ converges by the Cauchy criterion. □

In the preceding proof, without saying what we were doing, we used the technique of **summation by parts**.

LEMMA 247. *(Summation by Parts) Let $\{a_n\}$ and $\{b_n\}$ be two sequences. Then for all $m \leq n$ we have*

$$\sum_{k=m}^{n}a_k(b_{k+1} - b_k) = (a_{n+1}b_{n+1} - a_mb_m) + \sum_{k=m}^{n}(a_{k+1} - a_k)b_{k+1}.$$

PROOF.

$$\sum_{k=m}^{n}a_k(b_{k+1} - b_k) = a_mb_{m+1} + \ldots + a_nb_{n+1} - (a_mb_m + \ldots + a_nb_n)$$

$$= a_nb_{n+1} - a_mb_m - ((a_{m+1} - a_m)b_{m+1} + \ldots + (a_n - a_{n-1})b_n)$$

$$= a_nb_{n+1} - a_mb_n - \sum_{k=m}^{n-1}(a_{k+1} - a_k)b_{k+1}$$

$$= a_nb_{n+1} - a_mb_n + (a_{n+1} - a_n)b_{n+1} - \sum_{k=m}^{n}(a_{k+1} - a_k)b_{k+1}$$

$$= a_{n+1}b_{n+1} - a_mb_m - \sum_{k=m}^{n}(a_{k+1} - a_k)b_{k+1}.$$

□

Remark: Lemma 247 is a discrete analogue of the familiar integration by parts formula from calculus:

$$\int_a^b fg' = f(b)g(b) - f(a)g(a) - \int_a^b f'g.$$

(This deserves more elaboration than we are able to give at the moment.)

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 $\sum_n a_nb_n = \sum_n(-1)^{n+1}a_n$ 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 $\sum_n b_na_n$ converges.

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[14]

$$\cos n = \frac{\sin(n + \frac{1}{2}) - \sin(n - \frac{1}{2})}{2\sin(\frac{1}{2})}$$

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. 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:

[14] An instance of the **sum-product identities**. Yes, I hardly remember them either.

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 248. *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 $\sum_n a_n = \infty$.

Let us reflect for a moment on this state of affairs: in any nonabsolutely convergent series we have enough of a contribution from the positive terms to make the series diverge to ∞ and also enough of a contribution from the negative terms to make the series diverge to $-\infty$. Therefore if the series converges it is because of a subtle interleaving of the positive and negative terms, or, otherwise put, because lots of *cancellation* occurs between positive and negative terms. This suggests that the ordering of the terms in a nonabsolutely convergent series is rather important, and indeed in the next section we will see that changing the ordering of the terms of a nonabsolutely convergent series can have a dramatic effect.

8. Rearrangements

8.1. 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 $\sigma : \mathbb{N} \rightarrow \mathbb{N}$. We define the **rearrangement** of $\sum_{n=0}^{\infty} a_n$ by σ to be the series $\sum_{n=0}^{\infty} a_{\sigma(n)}$.

8.2. The Rearrangement Theorems of Weierstrass and Riemann.

The most basic questions on rearrangements of series are as follows.

Question 4. *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.

Theorem 249. *(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. For $N \in \mathbb{Z}^+$, define

$$N_\sigma = \max_{0 \leq k < N} \sigma^{-1}(k).$$

In other words, N_σ is the least natural number such that $\sigma(\{0, 1, \ldots, N_\sigma\}) \supseteq \{0, 1, \ldots, N-1\}$. Thus for $n > N_\sigma$, $\sigma(n) \geq N$. For all $\epsilon > 0$, by the Cauchy criterion for absolute convergence, there is $N \in \mathbb{N}$ with $\sum_{n=N}^{\infty} |a_n| < \epsilon$. Then

$$\sum_{n=N_\sigma+1}^{\infty} |a_{\sigma(n)}| \leq \sum_{n=N}^{\infty} |a_n| < \epsilon,$$

and the rearranged series is absolutely convergent by the Cauchy criterion. Let $A' = \sum_{n=0}^{\infty} a_{\sigma(n)}$. Then

$$|A - A'| \leq |\sum_{n=0}^{N_\sigma} a_n - a_{\sigma(n)}| + \sum_{n>N_\sigma} |a_n| + \sum_{n>N_\sigma} |a_{\sigma(n)}| < |\sum_{n=0}^{N_\sigma} a_n - a_{\sigma(n)}| + 2\epsilon.$$

Moreover, each term a_k with $0 \leq k \leq N$ appears in both $\sum_{n=0}^{N_\sigma} a_n$ and $\sum_{n=0}^{N_\sigma} a_{\sigma(n)}$, so we may make the very crude estimate

$$|\sum_{n=0}^{N_\sigma} a_n - a_{\sigma(n)}| \leq 2 \sum_{n>N} |a_n| < 2\epsilon$$

which gives

$$|A - A'| < 4\epsilon.$$

Since ϵ was arbitrary, we conclude $A = A'$. □

Exercise: 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 Theorem 249.

Theorem 250. *(Riemann Rearrangement Theorem) Let $\sum_{n=0}^{\infty} a_n$ be a nonabsolutely 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 248 which tells us that since the convergence is nonabsolute, $\sum_n p_n = \infty$, $\sum_n n_n = -\infty$. So we may specify a rearrangement 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 250 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 250 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 250 exposes the dark side of nonabsolutely convergent series: just by changing the order of the terms, we can make the series diverge to ±∞ 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 251. *(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[15] 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 $[a, A]$.

9. Power Series I: Power Series as Series

9.1. 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 5. *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:

$$\lim_{n\to\infty} \frac{(n+1)! x^{n+1}}{n! x^n} = \lim_{n\to\infty} (n+1)|x|.$$

[15] 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.

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 $x = 0$. Note that this is disappointing if we are interesteted 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 252. *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 253. *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 252 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 253.

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 253 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 the eminent turn of the century mathematician J.S. Hadamard, 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 indeed established by our most usual suspect: it was first proven by Cauchy in 1821 [**Ca21**] but apparently had been nearly forgotten.

THEOREM 254. *(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

$$\frac{1}{A} < A' < \frac{1}{R} = \overline{\theta}.$$

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 $a_n x^n \not\to 0$. □

Here is a useful criterion for the radius of convergence of a power series to be 1.

COROLLARY 255. *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 $a_n \not\to 0$, then $R \leq 1$.
c) Thus if $\{a_n\}$ is bounded but not convergent to zero, $R = 1$.

Exercise: Prove Corollary 255.

Finally we record the following result, which will be useful later on when we consider power series as functions and wish to differentiate and integrate them.

THEOREM 256. *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 the power series $\sum_{n\rightarrow\infty} n^k a_n x^n$ is also R.*

PROOF. Since $\lim_{n\rightarrow\infty} \frac{(n+1)^k}{n^k} = \lim_{n\rightarrow\infty} \left(\frac{n+1}{n}\right)^k = 1$, by Corollary 241,

$$\lim_{n\rightarrow\infty} (n^k)^{1/n} = \lim_{n\rightarrow\infty} n^{k/n} = 1.$$

(Alternately, one can of course compute this limit by the usual methods of calculus: take logarithms and apply L'Hôpital's Rule.) Therefore

$$\limsup_{n\rightarrow\infty} (n^k |a_n|)^{\frac{1}{n}} = \left(\lim_{n\rightarrow\infty} (n^k)^{\frac{1}{n}}\right) \left(\limsup_{n\rightarrow\infty} |a_n|^{\frac{1}{n}}\right) = \limsup_{n\rightarrow\infty} |a_n|^{\frac{1}{n}}.$$

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 $\rho = \lim_{n\rightarrow\infty} |\frac{a_{n+1}}{a_n}|$ exists and proving Theorem 256 under that additional assumption using the Ratio Test. This will be good enough for most of the power series encountered in practice.

Exercise: By Theorem 256, the radii of convergence of $\sum_n a_n x^n$ and $\sum_n n a_n x^n$ are equal. Show that the interval of convergence of $\sum_n n a_n x^n$ is contained in the interval of convergence of $\sum_n a_n x^n$, and give an example where a containment is proper. In other words, passage from $\sum_n a_n x^n$ to $\sum_n n a_n x^n$ does not change the radius of convergence, but convergence at one or both of the endpoints may be lost.

9.2. Recentered Power Series.

CHAPTER 12

Taylor Theorems, Taylor Polynomials, Taylor Series

1. Taylor Polynomials and Taylor Theorems

1.1. Taylor's Theorem (Without Remainder).

For $n \in \mathbb{N}$ and $c \in I$ (not an endpoint), we say that 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: If $0 \leq m \leq n$ 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 Exercise X.X, 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 258) that we call "Taylor's Theorem With Remainder", even though if I am not mistaken it is Theorem 257 and not Theorem 258 that was actually proved by Brook Taylor.

THEOREM 257. *(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.

[1] Brook Taylor, 1685 - 1731

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. We may also assume that $n \geq 2$, the cases $n = 0$ and $n = 1$ having been dealt with above.
(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):
claim There is a polynomial $P(x) = \sum_{k=0}^{n} a_k (x-c)^k$ of degree at most n such that $P(c) = h(c)$, $P'(c) = h'(c), \ldots, P^{(n)}(c) = h^{(n)}(c)$. This is easy to prove – but very important! – so we will save it until just after the proof of the theorem. Taking $f(x) = h(x)$, $g(x) = P(x)$, hypothesis (i) is satisfied, and thus by the already proven implication (i) $\implies$ (ii), we know that $h(x)$ and $P(x)$ agree to order n at $x = c$:

$$\lim_{x \to c} \frac{h(x) - P(x)}{(x-c)^n} = 0.$$

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

$$\lim_{x \to c} \frac{P(x)}{(x-c)^n} = 0. \tag{65}$$

Now it is easy to see – e.g. by L'Hôpital's Rule – that (65) can only hold if

$$a_0 = a_1 = \ldots = a_n = 0,$$

i.e., $P = 0$. Then for all $0 \leq k \leq n$, $h^{(k)}(c) = P^{(k)}(c) = 0$: (i) holds. □

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.

1.2. Taylor Polynomials.

Recall that we still need to establish the CLAIM made in the proof of Theorem 257. This is in fact more important than the rest of Theorem 257! So let

$$P(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + \ldots + a_n(x-c)^n$$

be a polynomial of degree at most n, let $f : I \to \mathbb{R}$ be a function which is n times differentiable at c, and let us see whether and in how many ways we may choose the coefficients $a_0, \ldots, a_n$ such that $f^{(k)}(c) = P^{(k)}(c)$ for all $0 \leq k \leq n$.

There is much less here than meets the eye. For instance, since $P(c) = a_0 + 0 + \ldots = a_0$, clearly we have $P(c) = f(c)$ iff

$$a_0 = f(c).$$

Moreover, since $P'(x) = a_1 + 2a_2(x-c) + 3a_3(x-c)^2 + \ldots + na_n(x-c)^{n-1}$, we have $P'(c) = a_1$ and thus $P'(c) = f(c)$ iff

$$a_1 = f'(c).$$

Since $P''(x) = 2a_2 + 3 \cdot 2a_3(x-c) + 4 \cdot 3a_4(x-c) + \ldots + n(n-1)a_n x^{n-2}$, we have $P''(c) = 2a_2$ and thus $P''(c) = f''(c)$ iff $a_2 = \frac{f''(c)}{2}$. And it proceeds in this way. (Just) a little thought shows that $P^{(k)}(c) = k!a_k$ – after differentiating k times the term $a_k(x-c)^k$ becomes the constant term – all higher terms vanish when we plug in $x = c$ – and since we have applied the power rule k times we pick up altogether a factor of $k \cdot (k-1) \cdots 1 = k!$. Therefore we must have

$$a_k = \frac{f^{(k)}(c)}{k!}.$$

In other words, no matter what the values of the derivatives of f at c are, there is a **unique** polynomial of degree at most k satisfying them, namely

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)(x-c)^k}{k!}.$$

$T_n(x)$ is called the **degree n Taylor polynomial for f at c**.

Exercise: Fix $c \in \mathbb{R}$. Show that every polynomial $P(x) = b_0 + b_1 x + \ldots + b_n x^n$ can be written in the form $a_0 + a_1(x-c) + a_2(x-c)^2 + \ldots + a_n(x-c)^n$ for unique $a_0, \ldots, a_n$. (Hint: $P(x+c)$ is also a polynomial.)

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 257 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 257 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 $a \geq b + 3$, in which case $f''_{a,b}(0) = 0$.
d) Deduce in particular that for any $n \geq 2$, $f_{n,n}$ vanishes to order n at $x = 0$ but is not twice differentiable – hence not n times differentiable – at $x = 0$.
e) Exactly how many times differentiable is $f_{a,b}$?

1.3. 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.

Theorem 258. *(Taylor's Theorem With Remainder) Let $n \in \mathbb{N}$ and $f : [a, b] \to \mathbb{R}$ be an $n + 1$ times differentiable function. Let $T_n(x)$ be the degree n Taylor polynomial for f at c, and let $x \in [a, b]$.*
a) There exists $z \in |[c, x]|$ such that

$$f(x) = T_n(x) + \frac{f^{(n+1)}(z)}{(n+1)!}(x - c)^{n+1}. \tag{66}$$

b) We have

$$R_n(x) = |f(x) - T_n(x)| \leq \frac{||f^{(n+1)}||}{(n+1)!}|x - c|^{n+1},$$

where $||f^{(n+1)}||$ is the supremum of $|f^{(n+1)}|$ on $|[c, x]|$.

Proof. a) [**R**, Thm. 5.15] Put

$$M = \frac{f(x) - T_n(x)}{(x - c)^{n+1}},$$

so

$$f(x) = T_n(x) + M(x - c)^{n+1}.$$

Thus our goal is to show that $(n + 1)!M = f^{(n+1)}(z)$ for some $z \in |[c, x]|$. To see this, we define an auxiliary function g: for $a \leq t \leq b$, put

$$g(t) = f(t) - T_n(t) - M(t - c)^{n+1}.$$

[2] Thanks to Didier Piau for a correction that led to this exercise.

Differentiating $n+1$ times, we get that for all $t \in (a,b)$,

$$g^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!M.$$

Therefore it is enough to show that there exists $z \in |[c,x]|$ such that $g^{(n+1)}(z) = 0$. By definition of T_n and g, we have $g^{(k)}(c) = 0$ for all $0 \leq k \leq n$. Moreover, by definition of M we have $g(x) = 0$. So in particular we have $g(c) = g(x) = 0$ and **Rolle's Theorem** applies to give us $z_1 \in |[c,x]|$ with $g'(z_1) = 0$ for some $z_1 \in |[c,x]|$. Now we iterate this argument: since $g'(c) = g'(z_1) = 0$, by Rolle's Theorem there exists $z_2 \in |[x,z_1]|$ such that $(g')'(z_2) = g''(z_2) = 0$. Continuing in this way we get a sequence of points $z_1, z_2, \ldots, z_{n+1} \in |[c,x]|$ such that $g^{(k)}(z_k) = 0$, so finally that $g^{(n+1)}(z_{n+1}) = 0$ for some $z_{n+1} \in |[c,x]|$. Taking $z = z_{n+1}$ completes the proof of part a).
Part b) follows immediately: we have $|f^{(n+1)}(z)| \leq ||f^{(n+1)}||$, so

$$|f(x) - T_n(x)| = |\frac{f^{(n+1)}(z)}{(n+1)!}(x-c)^{n+1}| \leq \frac{||f^{(n+1)}||}{(n+1)!}|x-c|^{n+1}.$$

□

Remark: There are in fact several different versions of "Taylor's Theorem With Remainder" corresponding to different ways of expressing the remainder $R_n(x) = |f(x) - T_n(x)|$. The particular expression derived above is due to Lagrange.[3]

Exercise: Show that Theorem 258 (Taylor's Theorem With Remainder) immediately implies Theorem 257 (Taylor's Theorem) under the additional hypothesis that $f^{(n+1)}$ exists on the interval $|[c,x]|$.

2. 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 by definition, $T(x) = \lim_{n\to\infty} T_n(x)$, where T_n is the degree n Taylor polynomial of x 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

$$T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)x^n}{n!},$$

since to get from this to the general case one merely has to make the change of variables $x \mapsto x - c$. It is somewhat traditional to call Taylor series centered around $c = 0$ **Maclaurin series**. But there is 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 in this work explicit attribution is made to Taylor's

[3] Joseph-Louis Lagrange, 1736-1813

work.[4] Using separate names for Taylor series centered at 0 and Taylor series centered at an arbitrary point 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 6. *Let $f : I \rightarrow \mathbb{R}$ 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 6a) 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 $\S X.X$. 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 $\frac{1}{R} = \limsup_{n \rightarrow \infty} |a_n|^{\frac{1}{n}}$, 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 there we certainly 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 6b): must $f(x) = T(x)$ for all $x \in (-A, A)$?

In fact the answer is **no**. Indeed, 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 $T(x) = \sum_{n=0}^{\infty} \frac{0x^n}{n!} = \sum_{n=0}^{\infty} 0 = 0$, i.e., it converges for all $x \in \mathbb{R}$ to the zero function. Of course $f(0) = 0$ (every function agrees with its Taylor series at $x = 0$), 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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.

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,

[4] For that matter, 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.

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. Although there are some tricks for this, in this case it is really no trouble to figure out exactly what $f^{(n)}(0)$ is for all non-negative integers n. Indeed, $f^{(0)}(0) = f(0) = e^0 = 1$, 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 258 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 259. *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 $n \in \mathbb{N}$,*

$$|f^{(n)}(x)| \leq M_A.$$

Then:
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 259.

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 259 and therefore is equal to its Taylor series expansion at $x = 0$ (or in fact, about any other point $x = c$).

Example: Let $f(x) = \sin x$. Then $f'(x) = \cos x$, $f''(x) = -\sin x$, $f'''(x) = -\cos x$, $f^{(4)}(x) = \sin x = f(x)$, so f has periodic derivatives. In particular the sequence of nth derivatives evaluated at 0 is $\{0, 1, 0, -1, 0, \ldots\}$. By Exercise X.X, we have

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

for all $x \in \mathbb{R}$. Similarly, we have

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}.$$

Exercise: Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function.
a) Suppose f is **odd**: $f(-x) = -f(x)$ for all $x \in \mathbb{R}$. Then the Taylor series expansion of f is of the form $\sum_{n=0}^{\infty} a_n x^{2n+1}$, i.e., only odd powers of x appear.
b) Suppose f is **even**: $f(-x) = f(x)$ for all $x \in \mathbb{R}$. Then the Taylor series expansion of f is of the form $\sum_{n=0}^{\infty} a_n x^{2n}$, i.e., only even powers of x appear.

Example: Let $f(x) = \log x$. Then f is defined and smooth on $(0, \infty)$, so in seeking a Taylor series expansion we must pick a point other than 0. It is traditional to set $c = 1$ instead. Then $f(1) = 0$, $f'(x) = x^{-1}$, $f''(x) = -x^{-2}$, $f'''(x) = (-1)^2 2! x^{-3}$, and in general $f^{(n)}(x) = (-1)^{n-1}(n-1)! x^{-n}$. Therefore the Taylor series expansion about $c = 1$ is

$$T(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}(n-1)!(x-1)^n}{n!} = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}(x-1)^n.$$

This power series is convergent when $-1 < x - 1 \leq 1$ or $0 < x \leq 2$. We would like to show that it is actually equal to $f(x)$ on $(0, 2)$. Fix $A \in (0, 1)$ and $x \in [1 - A, 1 + A]$. The functions $f^{(n)}$ are decreasing on this interval, so the maximum value of $|f^{(n+1)}|$ occurs at $1 - A$: $||f^{(n+1)}|| = n!(1 - A)^{-n}$. Therefore, by Theorem 258 we have

$$R_n(x) \leq \frac{||f^{(n+1)}||}{(n+1)!}|x - c|^{n+1} = \frac{|x-1|^{n+1}}{(1-A)^n(n+1)} \leq \frac{A}{n+1}\left(\frac{A}{1-A}\right)^n.$$

But now when we try to show that $R_n(x) \to 0$, we are in for a surprise: the quantity $\frac{A}{n+1}\left(\frac{A}{1-A}\right)^n$ tends to 0 as $n \to \infty$ iff $\frac{A}{1-A} \leq 1$ iff $A \leq \frac{1}{2}$. Thus we have shown that $\log x = T(x)$ for $x \in [\frac{1}{2}, \frac{3}{2}]$ only!

In fact $f(x) = T(x)$ for all $x \in (0, 2)$, but we need a different argument. Namely, we know that for all $x \in (0, 2)$ we have

$$\frac{1}{x} = \frac{1}{1-(1-x)} = \sum_{n=0}^{\infty} (1-x)^n.$$

As always for power series, the convergence is uniform on $[1 - A, 1 + A]$ for any $0 < A < 1$, so by Corollary 263 we may integrate termwise, getting

$$\log x = \sum_{n=0}^{\infty} \frac{-(1-x)^{n+1}}{n+1} = \sum_{n=0}^{\infty} \frac{(-1)^n (x-1)^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} (x-1)^n.$$

Moral: even if we can find an exact expression for $f^{(n)}$ and for $||f^{(n)}||$, the error bound given by Theorem 258b) may not be good enough to show that $R_n(x) \to 0$, even for rather elementary functions. This does not in itself imply that $T_n(x)$ does not converge to $f(x)$ on its interval of convergence: we may simply need to use other means. Often we try to exploit the Uniqueness Theorem for power series, which says that if we can – by any means necessary! – express $f(x)$ as a power series $\sum_n a_n (x - c)^n$ convergent in an interval around c of positive radius, then this power series must be the Taylor series expansion of f, i.e., $a_n = \frac{f^{(n)}(c)}{n!}$.

2.1. The Binomial Series.

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 $f : (-R, R) \rightarrow \mathbb{R}$. 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 $f : I \rightarrow \mathbb{R}$ is another function, we say f_n **converges to f pointwise** if for all $x \in I$, $f_n(x) \rightarrow f(x)$. (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 conveges 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

It turns out that the key hypothesis in most of our theorems is the notion of *uniform convergence.*

Let $\{f_n\}$ be a sequence of functions with domain I. We say f_n **converges uniformly** to f and write $f_n \overset{u}{\to} f$ 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.

Exercise: Show that there is a Cauchy Criterion for uniform convergence, namely: $f_n \overset{u}{\rightarrow} f$ on an interval I if and only if for all $\epsilon > 0$, there exists $N \in \mathbb{Z}^+$ such that for all $m, n \geq N$ and all $x \in I$, $|f_m(x) - f_n(x)| < \epsilon$.

The following result is the most basic one fitting under the general heading "uniform convegence justifies the exchange of limiting operations."

Theorem 260. *Let $\{f_n\}$ be a sequence of functions with common domain I, and let c be a point of I. Suppose that for all $n \in \mathbb{Z}^+$, $\lim_{x \to c} f_n = L_n$. Suppose moreover that $f_n \overset{u}{\rightarrow} f$. 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 261. *Let f_n be a sequence of continuous functions with common domain I and suppose that $f_n \overset{u}{\rightarrow} f$ on I. Then f is continuous on I.*

Since Corollary 261 is somewhat simpler than Theorem 260, as a service to the student 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

$$|f(x) - f(c)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

□

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 260 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 262. *Let $\{f_n\}$ be a sequence of Riemann integrable functions with common domain $[a, b]$. Suppose that $f_n \overset{u}{\to} f$. Then f is Riemann 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. Since we have not covered the Riemann integral in these notes, we are not in a position to give a full proof of Theorem 262. For this see [**R**, Thm. 7.16] or my McGill lecture notes. We will content ourselves with the special case in which each f_n is continuous, hence by Theorem 260 so is f. All continuous functions are Riemann integrable, so certainly f is Riemann integrable: what remains to be seen is that it is permisible to interchange the limit and the integral.

To see this, fix $\epsilon > 0$, and let $N \in \mathbb{Z}^+$ be such that for all $n \geq N$, $f(x) - \frac{\epsilon}{b-a} < f_n(x) \leq f(x) + \frac{\epsilon}{b-a}$. Then

$$\left(\int_a^n f\right) - \epsilon = \int_a^b \left(f - \frac{\epsilon}{b-a}\right) \leq \int_a^b f_n \leq \int_a^b \left(f + \frac{\epsilon}{b-a}\right) = \left(\int_a^b f\right) + \epsilon.$$

That is, $|\int_a^b f_n - \int_a^b f| < \epsilon$ and therefore $\int_a^b f_n \to \int_a^b f$. □

Exercise: It follows from Theorem 262 that the sequences in Examples 2 and 3 above are not uniformly convergent. Verify this directly.

Corollary 263. *Let $\{f_n\}$ be a sequence of continuous functions defined on the interval $[a, b]$ such that $\sum_{n=0}^{\infty} f_n \overset{u}{\to} f$. 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 $\sum_{n=0}^{\infty} F_n \overset{u}{\to} F$.*

Exercise: Prove Corollary 263.

Our next order of business is to discuss differentiation of sequences of functions. But let us take a look back at Example 4, which was of a bounded function $g : \mathbb{R} \to \mathbb{R}$ such that $\lim_{n\to\infty} g(x)$ does not exist and $f_n(x) = \frac{g(nx)}{n}$. Let M be such that $|g(x)| \leq M$ for all $\mathbb{R}$. Then for all $x \in \mathbb{R}$, $|f_n(x)| \leq \frac{M}{n}$. Since $\lim_{n\to\infty} \frac{1}{n} = 0$, for any $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$, $|f_n(x)| \leq \frac{M}{n} < \epsilon$. Thus $f_n \overset{u}{\to} 0$. In other words, we have the somewhat distressing fact that *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 264. *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions with common domain $[a, b]$. We suppose:*
(i) Each f_n is continuously differentiable, i.e., f_n' exists and is continuous,
(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 $f_n' \overset{u}{\to} g$ on $[a, b]$, certainly $f_n' \overset{u}{\to} g$ on $[a, x]$. Since each f_n' is assumed to be continuous, by 260 g is also continuous. Now applying Theorem 262 and (FTCb) we have

$$\int_a^x g = \lim_{n\to\infty} \int_a^x f_n' = \lim_{n\to\infty} f_n(x) - f_n(a) = f(x) - f(a).$$

Differentiating both sides and applying (FTCa), we get

$$g = (f(x) - f(a))' = f'.$$

□

COROLLARY 265. *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 $\sum_{n=0}^{\infty} f_n'(x) \overset{u}{\to} g$. Then f is differentiable and $f' = g$:*

$$(\sum_{n=0}^{\infty} f_n)' = \sum_{n=0}^{\infty} f_n'. \tag{67}$$

Exercise: Prove Corollary 265.

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 265 gives a condition under which a series of functions can be differentiated termwise.

Although Theorem 264 (or more precisely, Corollary 265) will be sufficient for our needs, we cannot help but record the following stronger version.

THEOREM 266. *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 $m, n \geq N$,

$$|g(x) - g(t)| = |x - t||g'(c)| \leq |x - t| \left(\frac{\epsilon}{2(b-a)} \right) \leq \frac{\epsilon}{2}. \tag{68}$$

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 (68) 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 260 on the interchange of limit operations:

$$f'(x) = \lim_{t \to x} \varphi(t) = \lim_{t \to x} \lim_{n \to \infty} \varphi_n(t) = \lim_{n \to \infty} \lim_{t \to x} \varphi_n(t) = \lim_{n \to \infty} f'_n(x).$$

□

2.1. 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 267. *(Weierstrass M-Test) Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions defined on an interval I. Let $\{M_n\}_{n=1}^{\infty}$ be a non-negative sequence such that $||f_n|| \leq M_n$ for all n and $M = \sum_{n=1}^{\infty} M_n < \infty$. Then $\sum_{n=1}^{\infty} f_n$ is uniformly convergent.*

PROOF. Put $S_n(x) = \sum_{k=1}^n f_k(x)$. Since the series $\sum_n M_n$ is convergent, it is Cauchy: for all $\epsilon > 0$ there exists $N \in \mathbb{Z}^+$ such that for all $n \geq N$ and $m \geq 0$ we have $M_{n+m} - M_n = \sum_{k=n+1}^{n+m} M_k < \epsilon$. But then for all $x \in I$ we have

$$|S_{n+m}(x) - S_n(x)| = |\sum_{k=n+1}^{n+m} f_k(x)| \leq \sum_{k=n+1}^{n+m} |f_k(x)| \leq \sum_{n=k+1}^{n+m} ||f_k|| \leq \sum_{k=n+1}^{n+m} M_k < \epsilon.$$

Therefore the series is uniformly convergent by the Cauchy criterion. □

3. Power Series II: Power Series as (Wonderful) Functions

THEOREM 268. *(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 $n \in \mathbb{N}$ we have $f^{(n)}(0) = (n!)a_n$.

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 260 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 265, 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 265 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 268d).

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 $f : \mathbb{R} \rightarrow \mathbb{R}$. 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 269. *(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 269 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, which brings us to the next section.

4. Applications to the Elementary Functions

5. The Stone-Weierstrass Theorem

6. A Continuous, Nowhere Differentiable Function

Bibliography

[Ac00] F. Acerbi, *Plato: Parmenides 149a7-c3. A Proof by Complete Induction?* Archive for History of the Exact Sciences 55 (2000), 57–76.

[A] T.M. Apostol, *Calculus. Volume One.*

[Ba98] B. Banaschewski, *On proving the existence of complete ordered fields.* Amer. Math. Monthly 105 (1998), 548–551.

[Bo71] R.P. Boas, Jr., *Signs of Derivatives and Analytic Behavior.* Amer. Math. Monthly 78 (1971), 1085–1093.

[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.

[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.

[Cu65] F. Cunningham, Jr., *Classroom Notes: The Two Fundamental Theorems of Calculus.* Amer. Math. Monthly 72 (1965), 406-407.

[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`

[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.

[Ka07] I. Kalantari, *Induction over the continuum.* Induction, algorithmic learning theory, and philosophy, 145–154, Log. Epistemol. Unity Sci., 9, Springer, Dordrecht, 2007.

[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.

[Ma42] C. Maclaurin, *Treatise of fluxions, 1.* Edinburgh (1742), 289–290.

[MaKa09] 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`

[Me72] F. Mertens, *Ueber die Multiplicationsregel für zwei unendliche Reihen.* Journal für die Reine und Angewandte Mathematik 79 (1874), 182–184.

[Mo50] R.K. Morley, *Classroom Notes: The Remainder in Computing by Series.* Amer. Math. Monthly 57 (1950), 550–551.

[M051] 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.

[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.

[Ro63] K. Rogers, *Classroom Notes: Unique Factorization.* Amer. Math. Monthly 70 (1963), no. 5, 547–548.

[R] W. Rudin, *Principles of mathematical analysis.* Third edition. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Dsseldorf, 1976.

[S] M. Schramm, *Introduction to Real Analysis.* Dover edition, 2008.

[Sc68] P. Schaefer, *Limit points of bounded sequences.*Amer. Math. Monthly 75 (1968), 51.

[Se59] A. Seidenberg, *A simple proof of a theorem of Erdős and Szekeres.* J. London Math. Soc. 34 (1959), 352.

[S] M. Spivak, *Calculus.* Fourth edition.

[Sp06] M.Z. Spivey, *The Euler-Maclaurin Formula and Sums of Powers.* Math. Magazine 79 (2006), 61–65.

[St] G. Strang, *Sums and Differences vs. Integrals and Derivatives.* College Math. Journal 21 (1990), 20–27.

[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.