Course Material for

MTH 327H: Honors Intro to Analysis

(Fall 2020)

Willie WY Wong

Preface

The following are course material that I used for the fall 2020 edition of Michigan State University's MTH 327H course. This is a "first" course in rigorous analysis for the Advanced Track students in the mathematics major, most of them having previously taken a semi-rigorous honors calculus sequence that introduced a bit of ϵ-δ arguments. Students are expected to have taken also our introduction-to-proofs course MTH 299, as well as a proof-based linear algebra course. That is to say, students are expected to have some familiarity with what are proofs and what are mathematical statements.

Traditionally this course has been taught using Walter Rudin's *Principles* —, which is indisputably an excellent text. However, given that the students have had a semi-rigorous calculus course under their belts, I wish the course to include something interesting that the students will not have encountered before, that is nevertheless fundamental and broadly applicable. One of the modern analysis concepts that have not been included in Rudin's treatise is the concept of *nets*, which encapsulates the many modes of limits and convergences beyond the basic sequences. It also serves as a good organizing principle for many of the analytical topics defined using limits, such as continuity and integration.

In this I admit I am heavily influenced by Eric Schechter, whose *Handbook of Analysis and its Foundations* served as an inspiration. That monumental resource has always been intended as a multi-year self-study guide for a student of analysis. My goal with these notes can be regarded as an attempt to condense from it a one-semester course for the advanced undergraduate student, that covers most of the aspects of single-variable advanced calculus in an entirely rigorous fashion, with a somewhat different organization compared to Rudin.

I feel that I have mostly succeeded: the main omission is a section on sequence and series of functions. These are originally intended to be included, but certain logistical issues due to the 2020 coronavirus pandemic demand that I effectively shorten the course by one week. Perhaps in a later installation I will restore that section.

So what are the main differences between these notes with, say, the textbook by Rudin?

- The textbook is *nets-forward*. In these notes convergence concepts are developed, from the very get-go, from a nets perspective with sequences listed only as special cases. This requires a bit more discussion of general order theory beyond chain orders (so we do lots of posets and directed sets).
- Numerical sequences and sequences in metric spaces are replaced by nets; topology of metric spaces are *defined* using net convergences. For most analysis practices this is a bit of an overkill; but for students who eventually study topology of non-metrisizable spaces this can be some useful background.
- The discussion of numerical series is flipped, with summation over an arbitrary index set done first and numerical series as a special case. This allows Riemann's rearrangement theorem to be recast as a statement about accumulation points of a non-convergent net.
- Continuity is defined in terms of commuting with convergent nets, and then characterized in terms of openness and closedness of the power-set maps. This definition feels more natural to me and is in agreement with how topology is presented earlier in the notes.
- In Rudin single variable differentiation is defined by limits of difference quotients, while higher dimensional derivatives are described by linear approximations. In these notes we directly define *tangency* of two mappings between arbitrary metric spaces using the idea of the order of approximation. This is a sort of precursor to defining the higher jets of smooth mappings between smooth manifolds, and also the more algebraic definitions of the tangent and cotangent

spaces of a manifold. Differentiability is then described as tangency against affine functions, showing where the affine structure of Euclidean spaces come into play. (This idea also underlies some modern work of Cheeger's extending the derivative concepts to metric measure spaces.) While this definition is certainly more complicated, I think it better prepares students for modern analysis on manifolds and in more general spaces.

- The idea of strong differentiability is explored briefly in exercises and problems.
- Finally, Riemann integration is defined using the convergence of an appropriate net of Riemann sums. The definitions are written in such a way that extensions to both the Henstock (gauge) and Stieltjes integrals become transparent. Henstock integrals are introduced and shown to be described by subnets of the Riemann one, which helps organize the understanding of their relationship with the Riemann version.

In Fall of 2020 this course was taught using a *flipped* instructional style, in part due to the pandemic. The lecture notes are distributed to students as "readings" on the *Perusall* website; Perusall was designed to support engaged and active reading, and provided collaborative commenting features that allows the instructor and the students to interact with each other by posting and answering questions. Its built-in LaTeX support makes it very usable for mathematics texts.

To support this "weekly reading" format, the subjects are chunked into 12-to-15-page-long segments. Weeks 5 and 10 are shorter: I used it as a pressure release valve for this stressful semester, to give both the students and myself a mental break. These notes can be used for an intensive 10-week (or 11-week) course (with no extra time allotted for assessments[1]), or a 12-week course with midterm assessment. For a 13-to-14 week course I would augment this with a discussion on sequences (nets) and series of functions. This would take about one to one-and-a-half weeks, and cover point-wise convergence, uniform convergence, and equicontinuity.

Finally, I would like to thank all the students that took the course in Fall 2020. I had effectively 17 pairs of very critical eyes helping me improve the notes and, importantly, catch typos big and small. Their infectious enthusiasm really helped lift my spirits and made the whole project worthwhile.

W. 2020.11

[1] This is essentially how I ran it in Fall 2020. The university's original coronavirus response plan involved having students on campus for the first twelve weeks of semester, sending students home after Thanksgiving. I gave no midterm exams, and concluded the course with one-on-one oral interviews.

Contents

The material for each week includes:

- Lecture notes with incorporated exercises;
- A separate exercise sheet for additional in-class group work (except weeks 5 and 10); and
- A problem set to be written up individually, and graded by the instructor.

The page numbering resets for each "item"; one can easily split up the document and distribute to students "in pieces".

Reading Assignment 1

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We will recall basic notions from set theory, including that of functions and relations. The natural numbers $\mathbb{N}$ will be taken as given, from this we define the integers $\mathbb{Z}$ and the rational numbers $\mathbb{Q}$. For the time being the real numbers $\mathbb{R}$ will remain undefined, and we will leave an interpretation of the reals to your prior experience working with them. We will return to this point at a later date. The algebraic properties of these special sets will be axiomatized. We will define the cardinality of sets and examine the cardinalities of $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.

Contents

I will assume that you have a basic familiarity with the concept of sets and classical logic (such as those taught in MTH 299 at MSU). If you feel you can use a refresher, you can read through Chapter 1 of Schechter's book. Specifically, I will assume you are familiar with the methods of Proof by Contradiction (we will accept the law of excluded middle in this class) and of Proof by Mathematical Induction. We begin with a rapid review of the most important aspects that we will need for this course.

§1.1 Review of set theory

We will not give an axiomatic treatment of set theory. We shall treat the notion of *sets* naively and intuitively as a collection of objects, which in this course will typically be numbers (though

sometimes also sequences, series, and functions). The members of a set are called *elements*. We indicate membership using the symbol $\in$; so $a \in X$ says that a is an element of the set X. We say that $a \notin X$ if X does not contain a as an element.[1]

A convenient notation for defining sets is by listing all its elements within braces, e.g. $\{a, b, c\}$. The listing is *unordered* and *unique*, in that $\{a, b, c\}$ and $\{c, b, b, a\}$ refer to the same set. In other words, membership in a set is counted without multiplicity.

A special set is the *empty set*, denoted by $\{\}$ or by $\emptyset$. The empty set contains no elements. Sets can be included within other sets as an element. The set $\{\emptyset\}$ is a set containing one element, which is the empty set. In fact, this construction is the basis of Peano arithmetic, which allows a construction of the natural numbers; this is discussed in Chapter 14 of Schechter's book and is outside of the scope of this class. In this course, we shall takes as given the set of numbers called the *natural numbers*, denoted by $\mathbb{N}$, which contains

$$\mathbb{N} := \{1, 2, 3, 4, \ldots\}. \tag{1.1}$$

The notation := indicates we are defining the object to its left to equal the expression on the right. Note we start counting from 1, and not zero.

A useful way of constructing new sets out of existing ones is via the set-builder notation. The construction usually takes the form

$$\{x \in X : \text{ condition(s) satisfied by } x\}.$$

For example, $\{n \in \mathbb{N} : n = 2m \text{ for some } m \in \mathbb{N}\}$ will denote the set of even natural numbers.

We say that a set A is a *subset* of a set B if every element of A is also an element of B. We denote this by $A \subseteq B$. Quite clearly $B \subseteq B$ by this definition. Two sets are said to be equal, denoted $A = B$, if either is a subset of the other $A \subseteq B$ and $B \subseteq A$. At times we will need to speak of *proper subsets*. The set A is a *proper subset* of B if it is a subset of B but it is not equal to B. We denote this by $A \subsetneq B$.[2]

The empty set $\emptyset$ is a subset of any set.

The intersection of two sets A, B, denoted by $A \cap B$, is the set that contains elements that are in both A and B. The union of two sets A, B, denoted by $A \cup B$, is the set that contains elements that are in at least one of A or B. The intersections and unions can be extended in the obvious way to any finite number of sets $A_1, A_2, A_3, \ldots, A_k$. As a short hand notation we will write $\cup_{i=1}^{k} A_i$ for the set of elements that appear in at least one of the sets A_1 through A_k, and similarly $\cap_{i=1}^{k} A_i$ for the set of elements that appear in all of A_1 through A_k.

Definition 1.1. *Given a set $\mathcal{S}$ of sets, we say $\mathcal{S}$ is* pairwise disjoint *if for $S, T \in \mathcal{S}$,*

$$S \neq T \implies S \cap T = \emptyset.$$

There are two ways of taking "differences" of sets. Given X, Y sets, we have

$$X \setminus Y = \{x \in X : x \notin Y\},$$
$$X \triangle Y = (X \cup Y) \setminus (X \cap Y).$$

[1] The slash notation for negation will be used frequently; since the usage and meaning is usually self evident, we will generally not make further comment on it.

[2] For this course I will avoid using the ambiguous notation $\subset$.

The latter is called the *symmetric difference* of the sets. The set difference operation $\setminus$ is not associative (like the subtraction of integers).

Exercise 1.1. Prove that the symmetric difference of sets is associative; that is $(X \triangle Y) \triangle Z = X \triangle (Y \triangle Z)$. Given a finite collection of sets $X_1, \ldots, X_n$, describe the set $X_1 \triangle X_2 \triangle \cdots \triangle X_n$.

Even more generally, let $\mathscr{S}$ denote a set of sets, not necessarily finite, we write

$$\bigcup \mathscr{S} := \{x : x \in S \text{ for some } S \in \mathscr{S}\}, \qquad \bigcap \mathscr{S} := \{x : x \in S \text{ for every } S \in \mathscr{S}\}. \tag{1.2}$$

Food for Thought 1.2. Based on the definition above, what is $\bigcup \emptyset$? How about $\bigcap \emptyset$?

Given a set X, we can form a new set containing all of the possible subsets of X, which we will call the *power set* of X and denote by 2^X.

Exercise 1.3. If $X = \{1, a, \{2, D\}\}$, what is 2^X? *(Note: don't forget the empty set.)*

Definition 1.2. *Given a set X and a set $\mathscr{S}$ whose elements are sets.*

1. *We say that $\mathscr{S}$* covers *X if $X \subseteq \bigcup \mathscr{S}$.*
2. *We say that $\mathscr{S}$* partitions *X if $X = \bigcup \mathscr{S}$, the elements of $\mathscr{S}$ are non-empty, and $\mathscr{S}$ is pairwise disjoint.*

§1.1.1 Tuples, products, and functions.—We will call on also the intuitive notion of an ordered pair[3]. An ordered pair is an ordered list of two elements, each of which can be an arbitrary mathematical object and may or may not be the same, and is written as (x_1, x_2). More generally, for any $n \in \mathbb{N}$, an n-tuple is an ordered list of n elements, written as $(x_1, \ldots, x_n)$.

Given two sets X and Y, its *Cartesian product* $X \times Y$ is the set of *all* ordered pairs (x, y) with $x \in X$ and $y \in Y$. More generally, given any number $n \in \mathbb{N}$, the Cartesian product of sets $X_1, X_2, \ldots, X_n$, denoted by

$$X_1 \times X_2 \times \cdots \times X_n \quad \text{or} \quad \prod_{i=1}^{n} X_i$$

is the set of *all* n-tuples $(x_1, x_2, \ldots, x_n)$ where each element $x_i \in X_i$.

Food for Thought 1.4. What is the set $X \times \emptyset$?

The ordering in Cartesian products matter, given unequal sets X, Y, the sets $X \times Y$ and $Y \times X$ are not the same. When we take the Cartesian product of n copies of the set X, we will use the shorthand X^n to refer to the resulting set of n-tuples.

Definition 1.3. *The* diagonal *of the product X^n is the subset $\{(x_1, \ldots, x_n) \in X^n : x_1 = x_2 = \cdots = x_n\}$.*

Definition 1.4. *Given two sets X, Y, we say that f is a* function *with domain X and codomain Y, which we write in notations using $f : X \to Y$, if f is a subset of $X \times Y$ such that every element of X appears as the first component of exactly one element of f. We use the notation $f(x)$ to refer to the element y such that (x, y) is the unique ordered pair in f that corresponds to the element $x \in X$.*

Definition 1.5. *Given $f : X \to Y$, the* range *of f is the subset $\{y \in Y : y = f(x) \text{ for some } x \in X\}$.*

[3] There's a way to make this precise using only sets, but we will omit that discussion; see paragraph 1.32 in Schechter.

Definition 1.6. *If $X_1, \ldots, X_n$ is a collection of sets, for any $1 \leq j \leq n$, the function $\pi_j : \prod_{i=1}^n X_i \to X_j$ is called the* projection to the jth component, *and is defined by*

$$\pi_j = \left\{ \left((x_1, \ldots, x_n), y \right) \in \left(\prod_{i=1}^{n} X_i \right) \times X_j : y = x_j \right\}.$$

Exercise 1.5. Parse the definition to check that it corresponds to what our naive understanding of what a projection should be. The definition is given as an illustration of the notations introduced thus far. Check the definition to see that π_j is indeed a function.

Exercise 1.6. How many functions are there with domain $\{1, 2, \ldots, m\}$ and codomain $\{1, 2, \ldots, n\}$?

Exercise 1.7. Based on the definitions: how many functions have domain $\{1, 2, \ldots, m\}$ and codomain $\emptyset$? How many functions have domain $\emptyset$ and codomain $\{1, 2, \ldots, m\}$?

Definition 1.7. *The* identity function *with domain and codomain both the set X, written as $\mathbf{1}_X : X \to X$, is the function corresponding to the diagonal of X^2.*

Definition 1.8. *Given $f : X \to Y$ and $g : W \to Z$ with $Y \subseteq W$, the composition $g \circ f : X \to Z$ is the function satisfying $g \circ f(x) = g(f(x))$.*

Definition 1.9. *We say that a function $f : X \to Y$ is injective if $f(x) = f(y) \implies x = y$. We say that a function $f : X \to Y$ is surjective if the range of f equals Y. We say that a function is bijective if it is both injective and surjective.*

Exercise 1.8. Show that

1. If X is non-empty, a function $f : X \to Y$ is injective if and only if it has a *left inverse*, that is, there exists a surjective function $g : Y \to X$ such that $g \circ f = \mathbf{1}_X$.
2. A function $f : X \to Y$ is surjective if and only if it has a *right inverse*, that is, there exists an injective function $h : Y \to X$ such that $f \circ h = \mathbf{1}_Y$.

(Note that the empty function is automatically injective; how do we handle the case where X is empty in the second statement?).

Definition 1.10. *If $S \subseteq X$, the* inclusion map *$\iota : S \to X$ is the map given by $\iota(s) = s$ for every $s \in S$.*

Definition 1.11. *Suppose $X \subsetneq Y$, and let ι denote the inclusion map.*

- *Given $f : Y \to Z$, its* restriction *to X is the function $g = f \circ \iota$.*
- *Given $f : X \to Z$, we say that h is an* extension *of f to Y if $f = h \circ \iota$.*

Definition 1.12. *Given $f : X \to Y$, it induces a function $2^X \to 2^Y$ between the power sets, which by common abuse of notation we denote again by f. This function is defined by*

$$\{(S, T) \in 2^X \times 2^Y : t \in T \iff \exists s \in S, f(s) = t\}.$$

The function f also induces a function $2^Y \to 2^X$ between the power sets, which by common abuse of notation we denote by f^{-1}. This function is defined by

$$\{(T, S) \in 2^Y \times 2^X : s \in S \iff \exists t \in T, f(s) = t\}.$$

Exercise 1.9. Show, with explicit example, that in spite of the notational abuse, the two power set functions f and f^{-1} described in the previous definition are generally *not* inverses of each other.

Food for Thought 1.10. A good test of one's understanding of propositional calculus, specifically the rule that $\neg P \implies (P \implies Q)$ for any Q can be had by trying to process what the definitions of the power set functions $f : 2^X \to 2^Y$ and $f^{-1} : 2^Y \to 2^X$ state, when X is the empty set.

§1.1.2 Relations and equivalence relations.—

Definition 1.13. *A* relation *R on a given set X is a subset of X^2. Conventionally instead of writing $(x, y) \in R$ we write xRy.*

The inverse relation[4] *of R is the set $R^{-1} := \{(x, y) \in X^2 : (y, x) \in R\}$.*

Given a relation $R \subseteq X^2$ and $Y \subsetneq X$, the restriction *of R to Y is $R \cap Y^2$.*

Example 1.14. The relation "is equals to", typically denoted by $=$, is the same as the diagonal of X^2. ■

We will take as understood the standard $<$, $\leq$, and $=$ relations on $\mathbb{N}$.

Definition 1.15. *The following are common properties of relations.*

- *A relation R is said to be* transitive[5] *if xRy and yRz implies xRz.*
- *The order matters in a relation in general. A relation R is said to be*
  - symmetric[6] *if $xRy \iff yRx$.*
  - antisymmetric *if xRy and yRx implies $x = y$.*
- *A relation R is said to be* connex[7] *if for every $x, y \in X$, at least one of xRy and yRx hold.*
- *A relation R is said to be* reflexive[8] *if xRx always.*

Exercise 1.11. For each of the standard $<$, $\leq$, and $=$ relations on $\mathbb{N}$, check which of the properties in the definition above are satisfied.

Definition 1.16. *An* equivalence relation *on a set X is a relation that is reflexive, symmetric, and transitive.*

If $\sim$ is an equivalence relation on X, the equivalence class *of $x \in X$, is the set $[x] := \{y \in X : x \sim y\}$. We denote by $X/\sim$ the set $\{[x] : x \in X\}$ of all equivalence classes. We say an element $z \in X$ is a* representative *of the equivalence class $[x]$ if $z \in [x]$.*

Exercise 1.12. Consider the relation R on $\mathbb{N}$ given by $\{(n, m) \in \mathbb{N}^2 : n = 3k + m$ for some $k \in \mathbb{N}$ or $m = 3k + n$ for some $k \in \mathbb{N}$ or $m = n\}$.

1. Prove that R is an equivalence relation.
2. Describe the set $\mathbb{N}/R$.

§1.2 The sets $\mathbb{Z}$ and $\mathbb{Q}$

We will take as given that $\mathbb{N}$ is closed under the two binary operations: addition (denoted by $+$), and multiplication (denoted by $\cdot$). While I perfectly expect you to understand exactly what the sets $\mathbb{Z}$ and $\mathbb{Q}$ are, in this section we will show that it is in fact possible to define the integers and the rationals starting just from the naturals. *We will not, in this course, actually interpret $\mathbb{Z}$ and $\mathbb{Q}$ using*

[4] Depending on author, this is also called the converse, or the transpose.
[5] Noun form: transitivity
[6] Noun form: symmetry
[7] Noun form: connexity
[8] Noun form: reflexivity

this construction. The material in this section is included to indicate that it is possible to reduce our understanding of "numbers" down to very basic levels (just the naturals); in later parts of this course we will work with the more intuitive "usual" descriptions of $\mathbb{Z}$ and $\mathbb{Q}$. ($\mathbb{R}$, on the other hand, is a different story.)

§1.2.1 The algebraic properties of $\mathbb{N}$.—The natural numbers $\mathbb{N}$ with its addition $+$ and multiplication $\cdot$ forms a *commutative semiring*[9].

Definition 1.17. *A* commutative semiring *is a set R with a binary operation $+ : R \times R \to R$ called addition and a binary operation $\cdot : R \times R \to R$ called multiplication such that*

- *$(R, +)$ is a commutative semigroup (i.e. addition is associative and commutative).*
- *$(R, \cdot)$ is a commutative semigroup (i.e. multiplication is associative and commutative).*
- *Multiplication distributes over addition $a \cdot (b + c) = ab + ac$.*

Notice that, given $r \in R$, the functions $f, g : R \to R$ given by $f(x) = r + x$ and $g(x) = r \cdot x$ are in general *not invertible*.

The usual ordering of the natural numbers can be described by

$$m < n \iff \exists k \in \mathbb{N} \text{ such that } m + k = n. \tag{1.3}$$

This ordering has the properties that

- it respects addition: $m < n \implies m + k < n + k$; $m < n$ and $l < k$ implies $m + l < k + n$.
- it respects multiplication: $m < n \implies k \cdot m < k \cdot n$; $m < n$ and $l < k$ implies $m \cdot l < k \cdot n$.

Proposition 1.18. *Every non-empty set $S \subseteq \mathbb{N}$ has a minimum element; in other words, there exists $s_0 \in S$ such that every $s \in S$ satisfies $s_0 \leq s$.*

Proof. We prove the contrapositive: if $S \subseteq \mathbb{N}$ is such that S has no minimum element, then S is empty.

We argue by mathematical induction. First, since $1 \leq n$ for any $n \in \mathbb{N}$, this means $1 \leq s$ for any $s \in S \subseteq \mathbb{N}$. This means $1 \notin S$; for if it were, it would be the minimum element.

Now supposing the induction hypothesis that $\{1, \ldots, n-1\} \cap S = \emptyset$, we must have $n - 1 < s$ for every $s \in \mathbb{S}$. This implies $n \leq s$, and hence $n \notin S$; for else n would be its minimum element. And hence $\{1, \ldots, n\} \cap S = \emptyset$. By induction we see that $\mathbb{N} \cap S = \emptyset$ as claimed. □

Food for Thought 1.13. A similar proof can be used to argue that there can be no surprise quizzes: Suppose the surprise quiz is scheduled for the last day of class, but by the end of the second to last day of class, as it has not happened for any of the prior days, the students would know that the quiz is coming on the following day, making it no longer a surprise. By the same induction argument as in the previous proof, neither can the quiz take place during any day of class.

§1.2.2 Construction of $\mathbb{Z}$ and its properties.—We can extend $\mathbb{N}$ to make addition invertible. One way to do so is to build a new set, the set of integers $\mathbb{Z}$, by the following definition.

Definition 1.19. *The integers $\mathbb{Z}$ is defined as set of equivalence classes $\mathbb{N}^2/\sim$, where the equivalence relation is*

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

[9] Some authors require semi-rings to also contain the additive unit 0 and the multiplicative unit 1; I don't.

Exercise 1.14. Answer the following about $\sim$

1. Check that $\sim$ is indeed an equivalence relation.
2. Check that if $(a,b) \sim (a',b')$, and $(c,d) \sim (c',d')$, then $(a+c, b+d) \sim (a'+c', b'+d')$, and hence we can define addition on $\mathbb{Z}$ by $[(a,b)] + [(c,d)] = [(a+c, b+d)]$.
3. Prove that $[(a,a)]$ is the unique additive identity in $\mathbb{Z}$ (this is the element we usually call 0): that is $[(a,a)] + [(b,c)] = [(b,c)]$. This means that the additive inverse of $[(b,c)]$ can be identified as $[(c,b)]$.
4. How can we define multiplication on $\mathbb{Z}$? In particular, what is the formula for $[(a,b)] \cdot [(c,d)]$? Prove that this multiplication is commutative and distributes over addition.
5. Based on your answer to the previous part, prove that the additive identity annihilates $\mathbb{Z}$, meaning $[(a,a)] \cdot [(b,c)] = [(a,a)]$.
6. Identify the element that is the multiplicative identity (the element we usually call 1).

(Remark: there are two inequivalent ways of defining multiplication, which result in two different elements being identified as the corresponding multiplicative identity. Can you find them both?)

Exercise 1.15. Continuing from the previous exercise: the identification of the multiplicative identity element allows us to embed $\mathbb{N}$ into $\mathbb{Z}$, in a way such that multiplication and addition on $\mathbb{N}$ is the restriction of the multiplication and addition on $\mathbb{Z}$. Prove that there is a unique extension of the ordering $<$ from $\mathbb{N}$ to $\mathbb{Z}$ that respects addition in $\mathbb{Z}$. Describe the relations $<$.

The constructed set $\mathbb{Z}$ now satisfies the properties of a commutative ring (which is a commutative semiring that possesses (1) an additive identity (2) a multiplicative identity and (3) additive inverses for every element).

§1.2.3 Construction of $\mathbb{Q}$ and its properties.—We can further extend $\mathbb{Z}$ to make multiplication by non-zero numbers invertible.

Definition 1.20. *The rational numbers $\mathbb{Q}$ is defined as the set of equivalent classes $(\mathbb{Z} \times \mathbb{N})/\sim$, where the equivalence relation is given by*

$$(a,n) \sim (b,m) \iff am = bn.$$

Note: following Exercise 1.15, we have an embedding of $\mathbb{N}$ into $\mathbb{Z}$ and so we can multiply the integers with natural numbers.

We extend addition and multiplication to $\mathbb{Q}$ as follows:

$$[(a,n)] + [(b,m)] = [(am + bn, mn)],$$
$$[(a,n)] \cdot [(b,m)] = [(a \cdot b, m \cdot n)].$$

Exercise 1.16. Check that the definitions of addition and multiplication respects the equivalence relation; namely that if $(a,n) \sim (a'n')$ and $(b,m) \sim (b',m')$, then $(am + bn, mn) \sim (a'm' + b'n', m'n')$, and $(a \cdot b, m \cdot n) \sim (a' \cdot b', m' \cdot n')$.

Exercise 1.17. Show that $\mathbb{Z}$ embeds into $\mathbb{Q}$ by identifying $a \in \mathbb{Z}$ with the equivalence class $[(a,1)]$, such that addition and multiplication on $\mathbb{Z}$ is the restriction of those for $\mathbb{Q}$.

Definition 1.21. *A* field *is a commutative ring (see final paragraph in the previous section) F with addition $+$ and multiplication $\cdot$, such that every element except the additive identity 0 has a multiplicative inverse (i.e., corresponding to each $x \in F$ there is an element y such that $x \cdot y$ is equal to the multiplicative identity 1).*

Exercise 1.18. Check that the definition of $\mathbb{Q}$ makes it a field.

As you may have already noticed, the equivalence class $[(a,n)] \in \mathbb{Q}$ with $(a,n) \in \mathbb{Z} \times \mathbb{N}$, represents the rational number (as we typically understand it) a/n.

Exercise 1.19. Apply Proposition 1.18 to show that among the equivalence class $[(a,n)] \subseteq \mathbb{Z} \times \mathbb{N}$, there is a *unique* element (a_0, n_0) for which n_0 is the minimum possible. We refer to this representative as the representation of $[(a,n)] \in \mathbb{Q}$ *in lowest terms*.

Exercise 1.20. Define the relation $\leq$ on $\mathbb{Q}$ by requiring $[(a,n)] \leq [(b,m)] \iff am \leq bn$ (as integers). Prove that $\leq$, as defined, is connex, antisymmetric, and transitive.

§1.3 Cardinality of sets

From this point forward, we will refer to elements of $\mathbb{Z}$ and $\mathbb{Q}$ in their more comfortable standard forms, and not in the form of equivalence classes.

In this section, we will consider the "sizes" of sets. We begin with a statement of the Pigeonhole Principle.

Proposition 1.22 (Pigeonhole Principle)**.** *Suppose $n < m$, then there does not exist an injective function $\{1,2,\ldots,m\} \to \{1,2,\ldots,n\}$. (By Exercise 1.8 this also implies that there does not exist a surjective function $\{1,2,\ldots,n\} \to \{1,2,\ldots,m\}$.)*

Proof. It suffices to prove the proposition for $m = n+1$: for if $f : \{1,2,\ldots,m\} \to \{1,2,\ldots,n\}$ were an injective function, and if $m > n+1$, then we can consider the restriction of f to the subset $\{1,2,\ldots,n+1\}$, which is also an injection.

We now argue by induction on n. When $n = 1$, the only function $f : \{1,2\} \to \{1\}$ is the constant function $f(1) = f(2) = 1$, which is not injective.

Suppose, as induction hypothesis, that there are no injective functions from $\{1,\ldots,n\}$ to $\{1,\ldots,n-1\}$. Suppose further for contradiction that there exists an injection $f : \{1,\ldots,n+1\} \to \{1,\ldots,n\}$. Consider the value $m = f(n+1)$. Let g be the permutation of $\{1,\ldots,n\}$ that swaps n and m leaving the other elements fixed. As a permutation it is a bijection, and hence $g \circ f$ is an injection from $\{1,\ldots,n+1\}$ to $\{1,\ldots,n\}$ with $g \circ f(n+1) = n$. Since $g \circ f$ is an injection, we see that its restriction to $\{1,\ldots,n\}$ is also an injection, furthermore, the range of the restriction cannot contain $g \circ f(n+1) = n$, and hence is a subset of $\{1,\ldots,n-1\}$. This means that $g \circ f$ restricts to an injection $\{1,\ldots,n\} \to \{1,\ldots,n-1\}$, in contradiction to the induction hypothesis. Hence such f cannot exist.

Thus by induction, for every $n \in \mathbb{N}$, there can be no injection from $\{1,\ldots,n+1\} \to \{1,\ldots,n\}$. □

Food for Thought 1.21. One has to be careful with the "remove an element" proofs like above. Here is a proof that all horses have the same color. For base case, the claim "a set of horses all have the same color" obviously holds for any singleton set of horses. By induction, given a set of n horses, if we remove one horse, the remaining set is a set of $n-1$ horses and contains horses of the same color by the induction hypothesis. Since the horse to be removed is arbitrary, this implies any set of n horses must only contains horses of the same color.

From an intuitive point of view, the Pigeonhole Principle holds because you cannot injectively map a set with more elements to a set with fewer elements (if you have more lodgers than your hotel have rooms, then some of your lodgers must share a room). This actually gives us a pretty convenient way to define the sizes of sets.

Definition 1.23. *Given two sets X, Y, we write*[10]

- $\mathrm{card}(X) = \mathrm{card}(Y)$ *if there exists a bijective function from* $X \to Y$.
- $\mathrm{card}(X) \leq \mathrm{card}(Y)$ *if there exists an injective function from* $X \to Y$.
- $\mathrm{card}(X) \geq \mathrm{card}(Y)$ *if there exists a surjective function from* $X \to Y$*; or if* Y *is empty.*

By Exercise 1.8 we see that $\mathrm{card}(X) \leq \mathrm{card}(Y) \iff \mathrm{card}(Y) \geq \mathrm{card}(X)$ as one may expect. And we can also interpret $\mathrm{card}(X) < \mathrm{card}(Y)$ to mean, appropriately, that there exist injections from $X \to Y$, but none of them are also surjections. However, it is not entirely obvious that

$$\mathrm{card}(X) \leq \mathrm{card}(Y) \quad \text{and} \quad \mathrm{card}(Y) \leq \mathrm{card}(X) \quad \iff \quad \mathrm{card}(X) = \mathrm{card}(Y). \tag{1.4}$$

Theorem 1.24 (Cantor-Bernstein-Schröder)**.** *The bi-implication* (1.4) *is true.*

Proof. Omitted, since it is slightly technical and tangential to the point of this course. For a sketch of proof see Paragraph 2.19 in Schechter's book. □

Exercise 1.22. Prove that the cardinality comparison is transitive. Namely, prove that if $\mathrm{card}(X) \leq \mathrm{card}(Y)$ and $\mathrm{card}(Y) \leq \mathrm{card}(Z)$, then $\mathrm{card}(X) \leq \mathrm{card}(Z)$.

Exercise 1.23. Let m, n be natural numbers. Prove that

$$\mathrm{card}(\{1, \ldots, n\} \times \{1, \ldots, m\}) = \mathrm{card}(\{1, \ldots, mn\}).$$

Definition 1.25. *Here are some terminology regarding cardinality of sets: A set X is said to be ...*

finite *if there exists some* $n \in \mathbb{N}$ *such that* $\mathrm{card}(X) \leq \mathrm{card}(\{1, \ldots, n\})$.
infinite *if* $\mathrm{card}(\mathbb{N}) \leq \mathrm{card}(X)$.
countable *if* $\mathrm{card}(X) \leq \mathrm{card}(\mathbb{N})$.
countably infinite *if* $\mathrm{card}(X) = \mathrm{card}(\mathbb{N})$.
uncountable *if* $\mathrm{card}(X) > \mathrm{card}(\mathbb{N})$.

The definition above is entirely based on comparison of the cardinality against that of the natural numbers. But notice that we have not established that the cardinality of any pair of sets is comparable! Could there exist X and Y such that there are neither an injection $X \to Y$ nor an injection $Y \to X$? Such a situation turns out to be impossible, provided we accept the **Axiom of Choice**.

A detour to Axiom(s) of Choice

The various forms of the Axiom of Choice states typically that certain constructions involving "infinities" can actually be performed. Let $\mathcal{S}$ be a set, whose elements are non-empty sets. A *choice function* on $\mathcal{S}$ is a function $f : \mathcal{S} \to \bigcup \mathcal{S}$ such that $f(X) \in X$ for every $X \in \mathcal{S}$. Colloquially a choice function is a function that, when given a set $X \in \mathcal{S}$, "chooses"

[10] In spite of the suggestive notation, card(—) is not a function, at least in this class. This is for several reasons. First, the collection of all sets is *not* a set; in most notions of set theory this collection is too large. Were one to allow this collection to be a set one easily runs into Russell's paradox concerning "the set of all sets which do not contain themselves as an element". Secondly, it is also not entirely clear what the codomain should be. For finite sets it may make sense to let card(—) spit out a number, but as we shall see that the comparison in this definition extends also to infinite sets.

an element $s \in X$. The Axiom of Choice states: for any set $\mathcal{S}$ whose elements are non-empty sets, there exists a choice function.

When $\mathcal{S}$ is a finite collection, there is no doubt that such a construction can be done (in fact this can be established as a theorem in many forms of set theory). It turns out that when $\mathcal{S}$ is *infinite*, standard ways of axiomatizing set theory *cannot prove* that such a construction is possible, without assuming something else that is logically equivalent. In other words, frequently the existence of choice functions need to be taken as an axiom of set theory, and not a theorem. As an example: it turns out that if we assume the Axiom of Choice, then we can prove that given any pair of sets X, Y, then either $\text{card}(X) \leq \text{card}(Y)$ or vice versa. But on the other hand, if we assume the statement that "the cardinalities of any pair of sets are always comparable", then we can derive the Axiom of Choice as a consequence!

A detailed discussion of the Axiom of Choice is way beyond the scope of this course. Luckily, Schechter is exhaustive in documenting his use of Choice in his textbook; for an introduction to the subject, see the final section of Chapter 1, as well as Chapter 6 in the book. There is in fact an entire book by Paul Howard and Jean Rubin, titled *Consequences of the Axiom of Choice*, that states various forms of the Axiom of Choice and what they imply and what they are equivalent to. **For this course, we will assume Axiom of Choice holds in its most general form**, and will not comment too much on its application.

Finally: one does not need to always take the whole shebang when it comes to the Axiom of Choice. The most generous version of the Axiom of Choice allows that choice functions exist for all sets $\mathcal{S}$ of non-empty sets. There are versions of the Axiom that only apply to more restrictive classes of $\mathcal{S}$. A lot of real analysis can in fact be done without appealing to any versions of the Axiom of Choice, and much of those that cannot be done, can be done by assuming the choice function exists whenever $\mathcal{S}$ is countable (note, not necessarily so for the sets making up its elements); this is called the *Axiom of Countable Choice*. The following theorem holds assuming this Axiom; observe how its proof requires making countably infinitely many distinct, but arbitrary choices.

Theorem 1.26. *Every set X is either finite or infinite. (In other words, the cardinality of every set is either $\leq$ or $\geq$ that of $\mathbb{N}$.)*

Proof. We will prove the following statement: if for every $n \in \mathbb{N}$, and every function $f : \{1, \ldots, n\} \to X$, we have that f is not surjective, then there exists an injective function $g : \mathbb{N} \to X$.

We start by choosing an arbitrary element x_1 of X, and set $g(1) = x_1$. Now, assuming we have specified the values of $g(1), \ldots, g(n-1)$ such that they are all different, the restriction of g to $\{1, \ldots, n-1\}$ is an injective function. By assumption this function cannot be surjective, which means that $X \setminus g(\{1, \ldots, n-1\}) \neq \emptyset$. And therefore we can choose x_n from among the remaining elements and set $g(n) = x_n$. Notice that by this choice the restriction of g to $\{1, \ldots, n\}$ is injective. Running this construction recursively we obtain our function g. □

Example 1.27. The integers $\mathbb{Z}$ is countably infinite through the following bijection $f : \mathbb{N} \to \mathbb{Z}$:

$$f(n) = \begin{cases} 0 & n = 1 \\ m & n = 2m, m \in \mathbb{N} \\ -m & n = 2m+1, m \in \mathbb{N} \end{cases}$$

■

Example 1.28. The rationals $\mathbb{Q}$ is countably infinite. There is an obvious injection $\mathbb{N} \to \mathbb{Q}$ from inclusion of natural numbers within the rationals as a subset. It suffices to give a surjection (by the Cantor-Bernstein-Schröder Theorem).

First, notice that the function $\mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ given by $(a, n) \mapsto a/n$ is a surjection. It then suffices to provide a surjection from $\mathbb{N} \to \mathbb{Z} \times \mathbb{N}$. Next, let Q_k, for $k \in \mathbb{N}$, be the subset

$$Q_k = \{(a, n) \in \mathbb{Z} \times \mathbb{N} : |a| + n = k\}.$$

The set Q_k has exactly $2k - 1$ elements, and hence $\cup_{k=1}^{j} Q_k$ has j^2 elements. List the elements of $\mathbb{Z} \times \mathbb{N}$ by listing first the elements of Q_1, then the element of Q_2, then Q_3, and so on. For each Q_k, list the elements in it in increasing order. Let $f : \mathbb{N} \to \mathbb{Z} \times \mathbb{N}$ map m to the mth entry in this list. This map is surjective. ■

Exercise 1.24. Prove that $\text{card}(X^2) \geq \text{card}(X)$ for any set X. Make sure your proof also works when X is the empty set.

At this point one typically proves that the real numbers are uncountable. However, we haven't really defined what the real numbers are! For the time being, we will do the next best thing: we will prove that the set of all *decimal numbers* is uncountable. Later on we can connect this back to the real numbers once we have defined them.

Example 1.29. By a *decimal number*, we refer to an ordered pair (E, m) where $E \in \mathbb{Z}$ and $m : \mathbb{N} \to \{0, 1, 2, \ldots, 9\}$ satisfies

- $m(1) \neq 0$.
- for every $j \in \mathbb{N}$, there exists $k > j$ such that $m(k) \neq 9$.

(The mapping to the reals, appealing to our common sense, will be $(E, m) \mapsto \sum_{j=1}^{\infty} m(j) \cdot 10^{E-j}$. But this is not important for this example.) We claim that the set $\mathscr{D}$ of decimal numbers is uncountable.

First, the set is clearly infinite. Let m_1 be the constant function $m_1(j) = 1$, then the mapping $n \mapsto (n, m_1)$ is an injective map from $\mathbb{N}$ to $\mathscr{D}$. To prove that it is uncountable, it suffices to show that there does not exist any surjective map $\mathbb{N} \to \mathscr{D}$. Let $\mathscr{F}$ denote the set of functions with domain $\mathbb{N}$ and codomain $\{1, 2, 3, \ldots, 8\}$. It further suffices to show that there does not exist any surjective map $\mathbb{N} \to \mathscr{F}$ (why?).

To prove this final assertion, let $f : \mathbb{N} \to \mathscr{F}$. We will show it is not surjective by constructing an explicit element σ that is not in the range of f. Define (note that $f(k)$ is a function, and $f(k)(k)$ is the output of the function $f(k)$ applied to the input k)

$$\sigma(k) = \text{minimum element of } \{1, \ldots, 8\} \setminus \{f(k)(k)\}.$$

The function $\sigma \in \mathscr{F}$. However, for every $k \in \mathbb{N}$, since $\sigma(k) \neq f(k)(k)$, the functions σ and $f(k)$ differ, and therefore σ is not in the range of f. ■

The method used in the previous example is called a "diagonalization argument", and appears frequently in various forms in mathematical analysis.

Exercise Sheet: Week 1

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 1.1. In the readings, the proof of the countability of the rationals $\mathbb{Q}$ is based on the following statement:

"Let $\mathscr{S}$ be a countably infinite set, whose elements are finite sets. Then $\bigcup \mathscr{S}$ is countably infinite."

1. Sketch the proof of the statement above (this gives an abstraction of the proof of the countability of the rationals).

To apply the statement, it remains to construct the countable set $\mathscr{S}$; specifically, we need a set that covers $\mathbb{Q}$. In the proof in the readings, we did so by building a function $f : \mathbb{Q} \to \mathbb{N}$ and setting

$$\mathscr{S} = \{f^{-1}(\{n\}) : n \in \mathbb{N}\}.$$

In the expression above f^{-1} is interpreted as the power set map of Definition 1.12.

2. Prove that given $f : X \to Y$, the set $\{f^{-1}(\{y\}) : y \in Y\}$ is a cover of X.
3. Prove that if $f : X \to Y$ is surjective, the set $\{f^{-1}(\{y\}) : y \in Y\}$ is a partition of X.

To conclude the proof, we need also to show that elements of $\mathscr{S}$ is finite.

4. Using the same method, prove that the set $\{M \subsetneq \mathbb{N} : M \text{ is finite }\}$ is countably infinite.

Question 1.2. This question concerns Definition 1.2 in the notes.

1. How many partitions are there of the set $\{1, 2, 3, 4\}$?
2. Quite clearly from the definition, if X is a set and $\mathscr{S}$ a partition thereof, then $\mathscr{S} \subseteq 2^X$. Is it the case that $\mathscr{S} \subsetneq 2^X$ always? *(There is one set that you may want to worry a little bit about.)*

Question 1.3. Let $X_1, \ldots, X_n$ be sets. Consider the projection maps $\pi_k : \prod_{j=1}^n X_j \to X_k$.

1. Is π_k always a surjection? Why or why not?
2. Is π_k ever an injection?

Question 1.4. Denote, by $\mathsf{F}[X, Y]$ the set of all functions from $X \to Y$. Prove that $\text{card}(\mathsf{F}[X, Y] \times \mathsf{F}[X, Z]) = \text{card}(\mathsf{F}[X, Y \times Z])$ by establishing a bijection between the two sets.

Problem Set 1

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 1.1. This question concerns the power set functions induced by $f : X \to Y$, as defined in Definition 1.12 in the notes. Prove that:

1. If $f : X \to Y$ is injective, then $f : 2^X \to 2^Y$ is injective and $f^{-1} : 2^Y \to 2^X$ is surjective.
2. if $f : X \to Y$ is surjective, then $f : 2^X \to 2^Y$ is surjective and $f^{-1} : 2^Y \to 2^X$ is injective.

(*Comment: a consequence of these statements is that* $\operatorname{card}(X) \leq \operatorname{card}(Y) \implies \operatorname{card}(2^X) \leq \operatorname{card}(2^Y)$.)

Problem 1.2. Prove *Cantor's Theorem*: $\operatorname{card}(2^X) > \operatorname{card}(X)$ for any set X.
Hint: follow these steps.

1. *First show the easier statement* $\operatorname{card}(2^X) \geq \operatorname{card}(X)$ *for any set* X.
2. *Thus it suffices to show that no bijection* $X \to 2^X$ *is possible; suppose for contradiction there is a bijection* f.
3. *Let* $R = \{x \in X : x \notin f(x)\}$. *Answer the question: is* $f^{-1}(R) \in R$*?*

Problem 1.3. The proofs of countability of $\mathbb{N}$ and $\mathbb{Q}$ extend to the following more general statements. Let $X_1, X_2, \ldots, X_n$ be a finite collection of countable (not necessarily infinite) sets. Prove that both

$$\bigcup_{i=1}^{n} X_i \quad \text{and} \quad \prod_{i=1}^{n} X_i$$

are countable.

Problem 1.4. Let X be an arbitrary set, and let R be an antisymmetric relation on X.

1. Prove that $\operatorname{card}(R) \leq \operatorname{card}(2^X)$. *(Hint: try to construct an explicit injection.)*
2. Show that if we drop the antisymmetric assumption, the claim no longer holds in general. *(Find an explicit counterexample.)*

Reading Assignment 2

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We begin by studying a special kind of "relation" called a *partial order*. Many concepts about inequalities on the real line can be generalized to partially ordered sets; examples include increasing and decreasing functions, and intervals. For a set X equipped with a partial order, given a subset $S \subseteq X$, we can ask whether it possesses certain special elements: the maximum/minimum, the maximal/minimal elements, and the supremum/infimum. Definitions will be given for these terms. Finally we return to our discussion of the real number line (more precisely, we focus first on $\mathbb{Q}$) by introducing the notion of a *total order*, and the related notion of a *totally ordered field*.

Contents

This set of readings will focus on the concept of *order*, which in the mathematical context means how we can "compare" two mathematical objects. For natural numbers, the ordering $<$ was defined in (1.3); this was extended to an ordering of $\mathbb{Z}$ in Exercise 1.15, and then to $\mathbb{Q}$ in Exercise 1.20. The focus on order is not incidental: our understanding of what the real number line $\mathbb{R}$ is, especially in contrast to the rational numbers $\mathbb{Q}$, rests on the ordering. Additionally, the order properties of $\mathbb{R}$ is what ultimately pins down the precise topological and geometric properties of the real number line.

§2.1 Partial Orders

For the rational number $\mathbb{Q}$, the comparison between two numbers is connex[1]. In many mathematical settings, we have to admit objects which are incomparable. This leads us to the notion of a *partial order*.

[1] See Definition 1.15 for terminology.

Definition 2.1. *A relation $\leqq$ on a set X is said to be a* partial order *if it is transitive, reflexive, and* antisymmetric.

A set X equipped with a partial order is called a partially ordered set, *or* poset.

Two elements x, y of the poset X is said to be comparable *if either $(x,y) \in \leqq$ or $(y,x) \in \leqq$.*

To verbally distinguish partial orders, which may not be connex, from total orders (which we will introduce later), when $x \leqq y$ instead of saying that x is less than or equal to y, we frequently say that x "precedes" y. Similarly, when $x \geqq y$ we say that x "succeeds" y.

Food for Thought 2.1. Compare the definition of a partial order to that of an equivalence relation (Definition 1.16).

Exercise 2.2. Prove that

1. The only partial order that is also an equivalence relation is the equality / diagonal relation.
2. The inverse relation to a partial order is also a partial order.
3. The restriction of a partial order on X to any subset Y is a partial order.

Example 2.2. The $\leq$ relation on $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{Q}$ are partial orders. ■

Exercise 2.3. We can define a relation $\leqq$ on 2^X, where the subsets $Y, Z \subseteq X$ are said to be $Y \leqq Z \iff Y \subseteq Z$. Check that this relation is a partial order. Check also that in general, this partial order is not connex.

Example 2.3. *(This example draws a bit on linear algebra.)*

Let M_2 denote the set of *symmetric* 2×2 matrices with real entries, which we can interpret as the set of *quadratic forms* on $\mathbb{R}^2$. Observe that the sum and difference of two elements of M_2 are still in M_2.

Recall that $A \in M_2$ is said to be *positive semi-definite* if for every vector $v \in \mathbb{R}^2$ the value $v^T A v \geq 0$. We can use this to define a partial order. More precisely, we can define $\leqq$ by $A \leqq B \iff B - A$ is positive semi-definite. Let's check that this indeed defines a partial order.

- For reflexivity: notice that $A - A$ is the 0 matrix, which satisfies $v^T 0 v = 0 \geq 0$ and hence $A \leqq A$.
- For transitivity: if $A \leqq B$ and $B \leqq C$, we can write $C - A = (C - B) + (B - A)$. So by distributive property of multiplication,

$$v^T (C - A) v = \underbrace{v^T (C - B) v}_{\geq 0} + \underbrace{v^T (B - A) v}_{\geq 0} \geq 0,$$

  showing that $A \leqq C$.
- For antisymmetry: if both $B - A$ and $A - B$ are positive semi-definite, we have

$$0 \leq v^T (B - A) v = -v^T (A - B) v \leq 0$$

  and hence $v^T (B - A) v = 0$ for every vector v. This implies that $B - A$ is the zero matrix, or that $B = A$.

This is an example of a poset that is *not* connex. Consider the matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. If $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ we have $v^T A v = 1 > 0$. But if $w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, then $w^T A w = -1 < 0$. Since this matrix is neither positive semi-definite

nor negative semi-definite, it neither precedes nor succeeds the zero matrix, and hence we have demonstrated a pair of *incomparable* elements of M_2. ■

Example 2.4. Let $(Y, \leqq)$ be a poset, and let X be a set. The set $\digamma$ of all functions from $X \to Y$ can be equipped with the partial order $\trianglelefteq$:

$$f \trianglelefteq g \iff \forall x \in X, f(x) \leqq g(x).$$

When the poset Y is the real numbers with the usual ordering, this partial order is equivalent to requiring that the function $g - f$ takes only non-negative values. ■

Example 2.5. Let X, Y be sets. Consider the set $\{f : S \to Y : S \subseteq X\}$ of all functions whose domain is a subset of X and whose codomain is Y. We can define the partial order $\leqq$ where $f \leqq g \iff f \subseteq g$ (remember that a function $S \to Y$ is a subset of $S \times Y$, which is in turn a subset of $X \times Y$). This is a special case of Exercise 2.3.

This partial order can be interpreted as saying $f \leqq g$ if and only if f is a restriction of g if and only if g is an extension of f. (See Definition 1.11.) ■

Definition 2.6. *Let $(X, \leqq)$ and $(Y, \trianglelefteq)$ be posets. We say that a function $f : X \to Y$ is...*

increasing *if $x_1 \leqq x_2 \implies f(x_1) \trianglelefteq f(x_2)$;*
decreasing *if $x_1 \leqq x_2 \implies f(x_2) \trianglelefteq f(x_1)$;*
monotone *if it is either increasing or decreasing;*
strictly monotone *(strictly increasing/strictly decreasing) if it is monotone (increasing/decreasing) and injective;*
an order isomorphism *if f is invertible and both f and f^{-1} are increasing.*

Food for Thought 2.4. We usually choose fairly suggestive notations, so that the partial order comes marked with either an "equals sign" or a "line segment" at the bottom (or sometimes, top) of the symbol; and the removal of that portion of the symbol usually refers to "strict inequality", in the sense that $x \prec y$ means $x \leqq y$ and $x \neq y$. (We say in this case "x strictly precedes y".) Try using Definition 2.6 to prove that a strictly increasing function satisfies $x_1 \prec x_2 \implies f(x_1) \triangleleft f(x_2)$.

Exercise 2.5. Let X, Y be sets, and let $f : X \to Y$ be a function. Prove that the power set maps $f : 2^X \to 2^Y$ and $f^{-1} : 2^Y \to 2^X$ (see Definition 1.12) are both increasing with respect to the $\subseteq$ partial ordering.

Example 2.7. It turns out that the subset partial order on 2^X, as described in Exercise 2.3, encompasses all possible posets.

Given a poset $(X, \leqq)$, we can define the function $\downarrow : X \to 2^X$ by setting $\downarrow(x) = \{y \in X : y \leqq x\}$. We claim that this mapping is an order isomorphism from X to its range.

First we show that $\downarrow$ is injective. Suppose $\downarrow(x) = \downarrow(y)$, by reflexivity of $\leqq$ we have that $x \in \downarrow(x)$ and hence also $x \in \downarrow(y)$, which in turn implies $x \leqq y$. The same argument also yields $y \leqq x$, and so by antisymmetry $x = y$. This shows that $\downarrow$ is a bijection to its range.

Next we show that $\downarrow$ is increasing: this follows by the transitivity of $\leqq$: if $x \leqq y$ and $z \in \downarrow(x)$, then $z \leqq x$ by definition of $\downarrow$ and hence $z \leqq y$, which in turn implies $z \in \downarrow(y)$. This shows $\downarrow(x) \subseteq \downarrow(y)$.

Finally, we show that $\downarrow^{-1}$ is increasing. Since $\downarrow$ is a bijection to its image, it suffices to show that $\downarrow(x) \subseteq \downarrow(y) \implies x \leqq y$. But as $x \in \downarrow(x)$ by reflexivity, our hypothesis implies $x \in \downarrow(y)$ and hence by definition $x \leqq y$. ■

One may ask why, in Definition 2.6 that strictly monotone and order isomorphic are two distinct properties. Referring back to our experience in Calculus: letting I be a subset of $\mathbb{R}$ and $f : I \to \mathbb{R}$ is a strictly increasing function, we have that f is in fact a bijection to its range, and the corresponding inverse is also increasing. So strictly monotone, at least for functions on the real line, implies that the function is an order isomorphism (provided we appropriately restrict the codomain).

That this holds for the real numbers boils down to the fact that the $\leq$ comparison on $\mathbb{R}$ is connex. For general posets, this need not be the case.

Exercise 2.6. Given an example of posets $(X, \leqq)$ and $(Y, \trianglelefteq)$, together with a bijection $f : X \to Y$, such that f is increasing but f^{-1} is not.

Definition 2.8. *Let $(X, \leqq)$ be a poset. Define the two functions $\uparrow, \downarrow : X \to 2^X$ by*

$$\downarrow(x) := \{y \in X : y \leqq x\}, \quad \uparrow(x) := \{y \in X : x \leqq y\}. \tag{2.1}$$

We say a subset $S \subseteq X$ is a lower set *or* downward-closed *if $s \in S \implies \downarrow(s) \subseteq S$. Dually, we say it is an* upper set *or* upward-closed *if $s \in S \implies \uparrow(s) \subseteq S$.*

A lower (upper) set S is said to be principal[2] *if there exists an element $x_0 \in X$ such that $S = \downarrow(x_0)$ (resp. $\uparrow(x_0)$). A lower (upper) set is* proper *if it is a proper subset of X.*

Given $a, b \in X$, the (closed) interval $[a, b]$ *is defined to be $\uparrow(a) \cap \downarrow(b)$.*

Example 2.9. In Reading Assignment 1, we frequently used the subset $\{1, \ldots, n\}$ of $\mathbb{N}$. With respect to the $\leq$ partial order on $\mathbb{N}$, these sets can be written as $\downarrow(n)$. ■

Exercise 2.7. Prove that in a poset, if $a \neq b$, at most one of $[a, b]$ and $[b, a]$ can be non-empty. (What does it mean when they are both empty?)

Exercise 2.8. Let $(X, \leqq)$ be a poset. Prove that the relation $\leqq$ is connex if and only if for every $a \in X$, $\uparrow(a) \cup \downarrow(a) = X$.

Definition 2.10. *Let $(X, \leqq)$ be a poset, and let $S \subseteq X$, and $z \in X$.*

- *We say that z* is an upper bound *of S if $S \subseteq \downarrow(z)$. The set S is said to be* bounded above *if it has an upper bound in X.*
- *We say that z* is a lower bound *of S if $S \subseteq \uparrow(z)$. The set S is said to be* bounded below *if it has a lower bound in X.*
- *We say that S is* order bounded *(or, when the context is clear, simply* bounded*), if it is bounded both above and below.*

Pay attention that upper (lower) bounds of the subset S can be any element in X, and need *not* be an element of S. The condition of being order bounded is equivalent to being a subset of an interval.

Example 2.11. Consider $\mathbb{N}$ with the usual $\leq$ order. The set $\downarrow(n)$ is contained in (in fact equal to) the interval $[1, n]$, and hence is bounded. Any number $m \geq n$ is an upper bound, and 1 is the only lower bound. The set $\uparrow(n)$ is bounded below; any number in $[1, n]$ is a lower bound of $\uparrow(n)$. But $\uparrow(n)$ has no upper bound in $\mathbb{N}$: ■

Example 2.12. (Another exercise in logic.) The empty set $\emptyset$ is bounded as a subset of any non-empty poset X. In fact, *any* element $z \in X$ is *both* an upper bound and a lower bound to $\emptyset$, going by the definitions.

[2] Please be careful with the spelling of "principal" versus "principle".

On the other hand, $\emptyset$ is *unbounded* as a subset of itself; as $\emptyset$ is empty, its set of upper bounds is also empty.

Provided $S \subseteq X$ is not empty, then we do in fact have (based on the transitive property of $\leqq$) that a lower bound a of S and an upper bound b of S must satisfy $a \leqq b$. ■

Exercise 2.9. Let S be a subset of a poset X. Prove that the set of all lower bounds of S is a lower set.

Example 2.13. The lack of connexity of a partial order can lead to some counterintuitive statements. For example, in a general poset, the union of two bounded sets need not be bounded. This can happen when the set of all upper bounds of the set A ends up being disjoint from the set of all upper bounds of the set B. An explicit example is given by considering the equality relation as a partial order on a set X. (The diagonal of X^2 is reflexive, antisymmetric, and transitive.) Then any singleton set $\{x\}$ for $x \in X$ is bounded (in fact, in any poset the singleton subsets are always bounded), both above and below, by x itself. However, if $x \neq y$, the two-element set $\{x, y\}$ has no upper nor lower bounds: our choice of partial order is such that any two unequal elements are not comparable. ■

The previous example motivates the following definition.

Definition 2.14. *Given a poset $(X, \leqq)$, a subset $S \subseteq X$ is said to be ...*

downward directed *if every* finite *subset has a lower bound $z \in S$.*
upward directed *if every* finite *subset has an upper bound $z \in S$.*

Exercise 2.10. Prove that on $\mathbb{Q}$ with the usual $\leq$ ordering, any subset is both downward and upward directed.

Exercise 2.11. It turns out that in Definition 2.14, it is not necessary to have "every finite subset". Fix the poset X. Let $P(S)$, where S is a subset, be the statement

> If x, y are distinct elements of S, then there exists $z \in S$ such that $z \leqq x$ and $z \leqq y$.

Prove that $P(S)$ is true implies S is downward directed. *(In other words, in practice, to check that a set is downward directed it suffices to only check for lower bounds for subsets of cardinality* 2*.)*

Exercise 2.12. Let $(X, \leqq)$ be a poset, such that X itself is downward (upward) directed. Prove the statement: Suppose $X_1, \ldots, X_n$ is a finite list of subsets of X, all of which bounded below (above), then $\cup_{i=1}^{n} X_i$ is also bounded below (above).

Example 2.15 (Continuation of Example 2.3)**.** The set M_2 with $\leqq$ is both downward and upward directed. We focus on the upward case here.

Let $\{A_1, \ldots, A_k\} \subsetneq M_2$ be a finite subset. By the spectral theorem for symmetric matrices, there exists a list of $2k$ numbers $\lambda_{1,1}, \lambda_{1,2}, \lambda_{2,1}, \lambda_{2,2}, \ldots, \lambda_{k,1}, \lambda_{k,2}$ corresponding to the (real) eigenvalues of the matrices (counted with multiplicity). As this is a finite list of numbers, we can let λ be the largest one. The matrix $\lambda I - A_i$ has eigenvalues $(\lambda - \lambda_{i,1})$ and $(\lambda - \lambda_{i,2})$, both non-negative, and hence is positive semi-definite. This means that λI is an upper bound for $\{A_1, \ldots, A_k\}$. ■

§2.2 Special elements

This section concentrates on the notions of *maximum*, *minimum*, *maximal*, *minimal*, *supremum*, and *infimum*.

§2.2.1 Max and min.—Within a general poset, there is a distinction between "bigger than all other elements" and "not less than any element"; this is again due to the lack of connexity of the order.

Definition 2.16. *Let $(X, \leqq)$ be a poset, and let $S \subseteq X$. We say that an element $s_0 \in S$ is*

the maximum of S *(also denoted $s_0 = \max S$) if $S \subseteq \downarrow(s_0)$;*
the minimum of S *(also denoted $s_0 = \min S$) if $S \subseteq \uparrow(s_0)$;*
a maximal element of S *if, for $s \in S$, the statement $s_0 \in \downarrow(s)$ implies $s_0 = s$;*
a minimal element of S *if, for $s \in S$, the statement $s_0 \in \uparrow(s)$ implies $s_0 = s$.*

Furthermore, we say that $s_0 \in S$ is an extremum *of S if it is either the maximum or the minimum, and we say it is an* extremal element *of S if it is either an maximal or minimal element.*

The choice of the article "the" and "a" in the definition are deliberate.

Proposition 2.17. *Let $(X, \leqq)$ be a poset, and let S be a subset. The maximum (minimum) of S, if it exists, is unique, and is the unique maximal (minimal) element of S.*

Proof. We first show that any maximum element is maximal: let s_0 be a maximum of S, and let $s \in S$ be arbitrary. Since s_0 is the maximum, we know $s \leqq s_0$. Now suppose $s_0 \in \downarrow(s)$; this implies $s_0 \leqq s$. By antisymmetry of $\leqq$ we have $s_0 = s$, proving that the maximum s_0 is maximal.

Next we show that if a maximum element exists, then there can be no other maximal elements. Together with the first part this shows that in this case both the maximum and the maximal are unique. Suppose t_0 is a maximal element of S. Since s_0 is the maximum, by definition, $t_0 \in \downarrow(s_0)$. But by the definition of maximal elements, this implies $t_0 = s_0$. □

Exercise 2.13. Prove that if $S \subseteq X$, and $\min S$ exists, then the set of lower bounds of S is exactly $\downarrow(\min S)$.

Exercise 2.14. Consider the statement "if S has a unique minimal element s_0, then s_0 is the minimum of S." Is this statement true? Why or why not?

Example 2.18. When S doesn't have a maximum, it may have multiple maximal elements. Let $(X, \leqq)$ be a poset for which $\leqq$ is not connex; then there exists $a, b \in X$ incomparable. The set $\{a, b\}$ have no maximum. But both a and b are maximal elements.

Furthermore, a set S may have an upper bound *without* having a maximum. For an example, consider the set M_2 discussed in Examples 2.3 and 2.15. We have shown that this set contains pairs of incomparable elements, but also that this set is upward directed, and hence any pair of elements is bounded above.

Finally, a set S may have an upper bound, yet have no maximal elements at all. Define a formal symbol called ∞. Let $X = \mathbb{N} \cup \{\infty\}$, and extend the ordering $\leq$ of $\mathbb{N}$ such that $n \leq \infty$ for any $n \in \mathbb{N}$. It is easy to check that with this ordering X is a poset. The subset $\mathbb{N}$ is bounded above by ∞, but $\mathbb{N}$ itself doesn't contain any maximal element. (For any $n \in \mathbb{N}$, $n + 1 \geq n$.) ■

We will see during next week's readings a lot more examples of bounded sets with no maximal elements.

Food for Thought 2.15. Unfortunately both "maximum" and "maximal" start with "max". The notation $\max S$ refers to the *maximum* (if it exists). Try not to get confused. Additionally, since the

maximum element may fail to exist, each time you assert $z = \max S$ you should ask yourself (and prove to the reader) "does the maximum exist?"

Food for Thought 2.16. One point of frequent confusion concerns the question "which of the notions we've defined is intrinsic, and which is extrinsic"?

In our context, an "intrinsic" statement concerning a subset S of a poset X is a statement that can be studied by forgetting about X and thinking of S itself as a poset with the order restricted from X. An example of such a statement is $z = \max S$. The interpretation of this statement does not refer to the external set X at all.

An "extrinsic" statement is one that depends on the external set being considered. An example of an extrinsic statement is "S is bounded".

It is worth going through all the statements that involve subsets of posets in this set of readings, and make sure you understand whether each statement is intrinsic or extrinsic.

Exercise 2.17. Prove that the following three statements are equivalent concerning a poset.

1. Every finite subset contains its maximum.
2. Every finite subset contains its minimum.
3. The partial order is connex.

(This exercise should be contrasted against Definition 2.14.)

Example 2.19 (Continuation of Example 2.4)**.** Let $S \subseteq \mathsf{F}$, there are two ways of taking the maximum.

- We can set $\max S$ to be the maximum of S under the partial order $\trianglelefteq$, when it exists.
- We can also consider the function given by $x \mapsto \max\{f(x) : f \in S\}$. This function (which only exists provided that for every $x \in X$, the set $\{f(x) : f \in S\} \subseteq Y$ has a maximum) is called *the pointwise maximum* of the family S.

In typical analysis contexts, the latter meaning is much more frequently used. ■

Exercise 2.18. Consider the functions $f, g : \mathbb{Z} \to \mathbb{Z}$, given by $f(x) = x$ and $g(x) = 2 \cdot x$. Does $\max\{f, g\}$ exist in the first sense of the previous example? If it does, describe the function. Does the function $x \mapsto \max\{f(x), g(x)\}$ exist in the second sense? If it does, describe the function.

Example 2.20 (Continuation of Example 2.15)**.** Coming back to the set M_2, let $A_+ = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $A_- = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$. These two elements are not comparable. Now consider the set $S = \uparrow(A_+) \cap \uparrow(A_-)$. What does the set S have in terms of minimal elements?

Let's try to first describe the set S. A matrix $B \in M_2$ can be parametrized linearly by $\begin{pmatrix} t+x & y \\ y & t-x \end{pmatrix}$. The condition that $B \in \uparrow(A_+)$ requires $B - A_+$ to be positive semi-definite, which is the same as requiring both its trace and determinant to be non-negative. This means that[3]

$$B \in \uparrow(A_+) \iff t \geq \sqrt{(x-1)^2 + y^2}; \qquad B \in \uparrow(A_-) \iff t \geq \sqrt{(x+1)^2 + y^2}.$$

[3] A here I tip my hand. On the space M_2, the determinant is a quadratic form on the coefficients of the matrix. What we have just shown here is the fact that M_2 with $\leqq$ is in fact order isomorphic to the Minkowski space with 2 space and 1 time dimension, with the causal ordering. The same question is much harder if you look at $n \times n$ matrices with $n \geq 3$.

Now, if B is such that $t > \max(\sqrt{(x-1)^2+y^2}, \sqrt{(x+1)^2+y^2})$, then we can choose a $t' < t$ such that the matrix B' corresponding to the triple (t', x, y) is still in $\uparrow(A_+) \cap \uparrow(A_-)$. And $B - B' = (t - t')I$ is positive definite. This shows that B cannot be minimal. Therefore minimal elements must satisfy $t = \sqrt{(x-1)^2+y^2}$ or $t = \sqrt{(x+1)^2+y^2}$.

If $t = \sqrt{(x-1)^2+y^2} > \sqrt{(x+1)^2+y^2}$, we must have $(x-1)^2 > (x+1)^2$, which in turn implies $x < 0$. Consider for some $\lambda > 0$, the triple

$$(t', x', y') = (\lambda t, 1 + \lambda(x-1), \lambda y).$$

This satisfies still $t' = \sqrt{(x'-1)^2 + (y')^2}$, and we see

$$(x'+1)^2 + (y')^2 = (x'-1)^2 + (y')^2 + 4x' = (t')^2 + 4 + 4\lambda(x-1).$$

If we choose $\lambda = \frac{1}{1-x} > 0$, we see then $t' = \sqrt{(x'+1)^2 + (y')^2}$, and so the matrix B' corresponding to the triple (t', x', y') also belongs to S.

But now, consider the matrix $B - B'$, a little bit of algebra shows its entries are

$$\frac{1}{1-x}\begin{pmatrix} -x(t+x-1) & -xy \\ -xy & -x(t+1-x) \end{pmatrix}.$$

And a computation of its trace and determinant shows that $B - B'$ is positive semi-definite. A similar computation with the role of A_+ and A_- swapped gives us finally that minimal elements must satisfy $t = \sqrt{(x-1)^2+y^2} = \sqrt{(x+1)^2+y^2}$. This implies $x = 0$ and $t = \sqrt{1+y^2}$, or that

$$B = \begin{pmatrix} \sqrt{1+y^2} & y \\ y & \sqrt{1+y^2} \end{pmatrix}.$$

I claim that this necessary condition is also sufficient. It suffices to show that if K is positive semi-definite, that $B - K \notin S$. A direct computation shows that $B - A_-$ and $B - A_+$ both have vanishing determinant, and hence are singular. The kernel of $B - A_+$ is the span of $v_+ = (1 + \sqrt{1+y^2}, -y)$, while the kernel of $B - A_-$ is the span of $v_- = (-y, 1 + \sqrt{1+y^2})$. Since v_+ and v_- are linearly independent, if K is a non-zero, positive semi-definite matrix, at least one of $v_+^T K v_+$ and $v_-^T K v_-$ is positive, and correspondingly at least one of

$$v_+^T(B - A_+ - K)v_+ \quad \text{and } v_-^T(B - A_- - K)v_-$$

is negative, showing that $B - K \notin S$. ∎

Food for Thought 2.19. The linear algebraic details of the above example is not particularly important; I included it for completeness of the discussion. The take away message of the example should be "how does one try to study the set of minimal elements of a set S"? The method illustrated is a very general approach for studying any class of mathematical objects: you want to try to describe necessary conditions for membership in that class, and you want to try to describe sufficient conditions for membership in that class. In certain situations, the two sets of conditions meet and you obtain a "characterization" of the class in question.

§2.2.2 Sup and inf.—Sometimes, a subset S of a poset X can come very close to having a minimum. Consider the set $\mathbb{Q}$ with the normal $\leq$. And let $S = \{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots\}$. This set does not have a minimal element: given any $\frac{1}{n}$, the element $\frac{1}{n+1}$ is also in S and is smaller. But intuitively, we see that the number 0 is almost as good as a minimum. In fact, it satisfies $\downarrow(0) = \{q \in \mathbb{Q} : q \text{ is a lower bound for } S\}$ (see also Exercise 2.13). As we will see, for a lot of analysis applications it is enough to think about elements of X for which this property holds, without needing the element to actually belong to S.

Definition 2.21. *Given a poset $(X, \leqq)$ and a subset S. We say that an element $z \in X$ is ...*

the supremum of S *(denoted $z = \sup S$) if $z = \min\{y \in X : S \subseteq \downarrow(y)\}$.*
the infimum of S *(denoted $z = \inf S$) if $z = \max\{y \in X : S \subseteq \uparrow(y)\}$.*

Exercise 2.20. Prove that

1. a lower set is principal if and only if it has a maximum; and
2. consequently, a set S has an infimum if and only if its set of lower bounds is principal.

Exercise 2.21. Prove that

1. if S has a supremum, and $\sup S \in S$, then $\sup S = \max S$.
2. if S has a maximum, then $\sup S = \max S$.

And so we see that the only practical difference between S having a minimum and S having an infimum is that whether the element $\inf S$ belongs to S or not.

Example 2.22 (Continuation of Example 2.12)**.** It is instructive, when dealing with these kinds of definitions, to ask: "what about the empty set"? By Example 2.12 we see that the set of upper bounds and lower bounds of the empty set in a poset X is always the entire set. Thus we get that $\sup \emptyset$ exists and equals $\min X$ if and only if $\min X$ exists. Dually, $\inf \emptyset$ exists and equals $\max X$ if and only if $\max X$ exists. When X is bounded in itself, there exists $a, b \in X$ such that $X \subseteq [a, b]$, which implies that $a = \min X$ and $b = \max X$. When X is not the singleton set, then we have the counterintuitive statement that $\sup \emptyset \lneqq \inf \emptyset$. ■

Intuition is restored when S is non-empty: by transitivity of partial orders, $\inf S \leqq \sup S$ whenever $S \neq \emptyset$ and $\inf S, \sup S$ exist.

So when can suprema and infima fail to exist? Examining the definition we see that there are essentially only two ways that a subset S in a poset X can have no suprema/infima.

1. S can fail to be (upper/lower) bounded. In this case its set of upper/lower bounds is empty, and has no minimum/maximum element.
2. S can be bounded, while the set(s) of upper/lower bounds fail to have corresponding extrema.

For the purposes of analysis, the second case is more interesting.

Example 2.23. Let $X = \{1, 2, 3\}$ and consider 2^X with the $\subseteq$ partial order. The set $S \subseteq 2^X$ given by $S = \{\{1\}, \{2\}, \{3\}\}$ satisfies

- each of $\{1\}$, $\{2\}$, and $\{3\}$ is both a minimal and a maximal element of S.
- neither $\min S$ nor $\max S$ exist.
- $\inf S = \emptyset$, being the unique lower bound of S; $\sup S = X$, being the unique upper bound of S.

More generally, letting Y be any set, and consider 2^Y with the $\subseteq$ partial order. Given any $\mathscr{S} \subseteq 2^Y$, we

have in fact

$$\sup \mathscr{S} = \bigcup \mathscr{S}, \qquad \inf \mathscr{S} = \begin{cases} \bigcap \mathscr{S} & \mathscr{S} \neq \emptyset \\ Y & \mathscr{S} = \emptyset \end{cases}$$

(It may be worth thinking back to Food for Thought 1.2 for what happens when $\mathscr{S}$ is empty.) ■

§2.3 Total Order

A total order is a special kind of partial order; a partial order is so named because some elements may not be comparable (a similar terminology is that of a *partial function*; a partial function from X to Y is just a function from some subset of X to Y).

Definition 2.24. *A total order is a partial order which is connex.*

In these notes we've already treated some cases where connexity matters; see e.g. Exercise 2.8, Example 2.13, and Exercise 2.17. For total orders, things are generally much simpler. In the following proposition we show that in total orders, finite unions of bounded sets are bounded; and there is no longer a distinction between maxim*um* and maxim*al*.

Proposition 2.25. *Let $(X, \leqq)$ be a total order.*[4]

1. *If $A, B \subseteq X$ are bounded above (below) then $A \cup B$ is bounded above (below).*
2. *If $A \subseteq X$ has a minimal (maximal) element a, then $a = \min A$ (resp. $\max A$).*

Proof. We will only prove that statements not in the parentheses; the other follows with almost the exact same proof.

1. By definition there exists $y, z \in X$ such that $A \subseteq \downarrow(y)$ and $B \subseteq \downarrow(z)$. Since X is a total order, either $y \leqq z$ or $z \leqq y$, and so $\max\{y, z\}$ exists. By transitivity we have $A \cup B \subseteq \downarrow(\max\{y, z\})$ and so $A \cup B$ is bounded above.
2. Let $a \in A$ be minimal. Let $b \in A$. By connexity either $a \leqq b$ or $b \leqq a$. We treat the two cases separately.
   If $a \leqq b$, then $b \in \uparrow(a)$.
   If $b \leqq a$, since a is minimal, by definition we have $a = b$, and hence also $b \in \uparrow(a)$.
   Hence we conclude that $A \subseteq \uparrow(a)$ which is the definition of $a = \min A$.

□

Food for Thought 2.22. You may ask why we fudged around so much with partial orders when we know (or at least I asserted) that the ordering on $\mathbb{Q}$ and $\mathbb{R}$ are total. The reason is two-fold:

1. While the theory for convergence on $\mathbb{R}$ can be largely stated using only total orders, we can define a theory of convergence based on much more general types of orderings. For those of you who intend to study classical point-set topology and its modern applications, this method of define convergence, through "nets", turns out to be what is needed to understand general topological spaces. The introduction of this theory, and the application of this theory to $\mathbb{R}$, only requires a tiny bit more effort than the total order case, so it is good to start early and get used to things now. This notion of convergence also has the advantage of making the theory of Riemann integration (and its cousins) much cleaner to state.

[4] Other names for the same concept: linear order, chain order, chain.

2. Order theory, specifically concerning partial orders, turns out to crop up in many places in mathematics. Topology and the theory of convexity systems (an abstraction of convex analysis) can be captured through studying "closure operators" defined on the posets; the Legendre-Fenchel transform for convex functions can be understood as a special case of so-called "Galois connections" between certain posets; and in mathematical general relativity a very powerful idea is to concentrate on understanding causal geometry, which places a partial order on space-time events based on whether signals can be sent from one to the other.

All of the sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ that we have defined so far, and subsets thereof, are total orders. An interesting property enjoyed by $\mathbb{N}$ is that every nonempty subset of it has a minimum (Proposition 1.18); the same is not true for $\mathbb{Z}$ or $\mathbb{Q}$. Sometimes you will see in the mathematical literature the notion of a *well-ordered set*; while we will not make use it, for culture, I will include its definition here.

Definition 2.26. *A well-ordered set is a totally ordered set where every non-empty subset has a minimum.*

The interest in well-ordered sets primarily lies in the **principle of transfinite induction**: that mathematical induction can be carried out relative to any well-ordered set. More precisely: let $P(w)$ be a family of mathematical statements indexed by a well-ordered set W. If (a) $P(\min W)$ is true, and (b) the statement "if $P(w)$ holds on a lower set Ω, then $P(\min(W \setminus \Omega))$" also holds, then $P(w)$ is true for all $w \in W$. This principle is useful because the principle of mathematical induction, being based on $\mathbb{N}$, can only handle up to countably infinitely many statements at one time. By using well-ordered sets of larger cardinality, you can now use induction to prove larger collections of statements at once. Of course, one may ask whether there exists any well-ordered sets of larger cardinality. One particular form of the Axiom of Choice is, in fact, the positive assertion that any set can be equipped with a partial order that makes it a well-ordered set. (Note: the usual ordering on $\mathbb{R}$ is *not* well-ordered.) To learn more about well-orderings, please look at pp. 72–77 in Schechter.

The set $\mathbb{Q}$ is special too, because it satisfies both the *field* axioms and the *total order* axioms, and the two sets of axioms are compatible. We introduce a name for such objects.

Definition 2.27. *A* totally ordered field *is a set F together with a partial order $\leqq$ such that*

1. *F is a field, with additive identity 0 and multiplicative identity 1.*
2. *$\leqq$ is a total order.*
3. *$\leqq$ respects addition: $a \leqq b \implies a + c \leqq b + c$.*
4. *$\leqq$ respects positive multiplication: $0 \leqq a$ and $0 \leqq b$ implies $0 \leqq a \cdot b$.*

Exercise 2.23. Prove the following concerning totally ordered fields.

1. A totally ordered field F is *not* bounded (either above or below) within itself.
2. If x, y are elements in a totally ordered field, then $0 \lneqq x \leqq y \implies 0 \lneqq \frac{1}{y} \leqq \frac{1}{x}$.

Example 2.28. Every totally ordered field contains a copy of $\mathbb{Q}$.

A sketch of the inclusion. We know that 1 is in F, and so is $1 + 1$, and $1 + 1 + 1$, and so on. By the fact that $\leqq$ respects addition, and the fact that $1 \neq 0$, we have that all of these numbers 1, $1 + 1$, ... are distinct. This gives us a copy of $\mathbb{N}$. Adding in also the additive inverses of $\mathbb{N}$ gives us a copy of $\mathbb{Z}$ in F. Then we can perform division of elements of $\mathbb{Z}$ by $\mathbb{N}$, the fact that the ordering respects multiplication allows us to say that the resulting object is in fact a copy of $\mathbb{Q}$. ∎

Food for Thought 2.24. It is obvious that there is a mapping from $\mathbb{N}$ into F that preserves addition

and multiplication, for *any* field F. What is not always the case is the injectivity of this function. (If you have taken abstract algebra, think finite fields.) The main part of this proof above is captured in the fact that compatibility of the total order with addition implies that this obvious function from $\mathbb{N}$ into F is an *injection*. Once you've established the existence of a copy of $\mathbb{N}$, to go up to $\mathbb{Q}$ is basically just pushing symbols around on a piece of paper.

We will therefore abuse notation and say that $\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subseteq F$ for any totally ordered field.

The condition that F is a totally ordered field is a fairly restrictive one. For example, the following theorem, one of whose consequences is that it is impossible to make the set of complex numbers $\mathbb{C}$ into a totally ordered field, is true.

Theorem 2.29. *Totally ordered fields cannot contain imaginary numbers. More precisely, if F is a totally ordered field, there does not exist any $x \in F$ satisfying $x^2 = -1$.* (Here we interpret -1 to be the additive inverse of the multiplicative identity.)

Proof. We will show first that if $x \in F$, then $0 \leqq x^2$. If $0 \leqq x$, then this follows from the fact that the order respects positive multiplication.

Suppose then $x \leqq 0$. Denote by its additive inverse y. Then $0 = x + y \leqq 0 + y = y$. And hence $0 \leqq y^2$, or $(x+y)^2 \leqq y^2$. Distributing we get $x^2 + 2xy + y^2 \leqq y^2$, and adding the additive inverse of y^2 to both sides we get $x^2 + 2xy \leqq 0$. Adding x^2 to both sides we finally get $2x^2 + 2xy = 2x(x+y) = 2x \cdot 0 = 0 \leqq x^2$.

Since $-1 \lneqq 0$ and $0 \leqq x^2$ for any $x \in F$, we conclude that $x^2 \neq -1$. □

By Exercise 2.8, any element of a total order splits it into two halves. In a totally ordered field, by the field axioms there is a special element 0.

Definition 2.30. *In a totally ordered field, the set of* positive[5] *elements is $\uparrow(0) \setminus \{0\}$. The set of* negative *elements is $\downarrow(0) \setminus \{0\}$.*

This allows us to define, for any totally ordered field, an *absolute value* function.

Definition 2.31. *Given a totally order field $(F, \leqq)$. The absolute value function $F \to F$, denoted by $x \mapsto |x|$, is*

$$|x| := \begin{cases} x & 0 \leqq x \\ -x & x \lneqq 0 \end{cases}$$

Proposition 2.32. *The following properties of the absolute value function hold.*

1. $|a \cdot b| = |a| \cdot |b|$;
2. $|a - b| \leq |a| + |b|$. *(The "triangle inequality".)*

Proof. Both statements should be proven by cases.

1. If $a, b \geqq 0$, the statement follows from the fact that the order respects positive multiplication. If one of a, b is negative (say a), and the other not, then as $-a$ is positive, we have $(-a) \cdot b \geqq 0$ which implies $a \cdot b \leqq 0$. And hence $|a \cdot b| = -(a \cdot b) = (-a) \cdot b = |a| \cdot |b|$.

[5] This is a rare moment I wish I were French. In French *positif* is defined to mean what we in English say "non-negative", and what English calls "positive" the French call *positif strictement*. So this definition in French would not include the set substraction.

If both are negative, then as $0 \leqq (-a) \cdot (-b) = a \cdot b$ the equality follows.

2. Since at least one of $a - b$ and $b - a$ is non-negative, and they have the same absolute value, we can assume $a - b \geqq 0$ (otherwise swap the symbols). So the goal is to prove that $a - b \leqq |a| + |b|$. But as $|x| \geqq x$ by construction, we have

$$a - b \leqq |a| - b \leqq |a| + |-b| = |a| + |b|$$

with the first two inequalities by the fact $\leqq$ respects addition.

□

Exercise Sheet: Week 2

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below.*

Question 2.1. Consider the relation $\leqq$ on $\mathbb{Z}^2$ given by

$$(x_1, x_2) \leqq (y_1, y_2) \iff 5 \cdot (y_2 - x_2) \geq |y_1 - x_1|.$$

1. Prove that $\leqq$ is a partial order. *(You may find the triangle inequality useful.)*
2. Prove that $\leqq$ is upward directed. (It is also downward directed, the proof is the same.)
3. Prove that the subset $\{(x, 0) : x \in \mathbb{Z}\}$ is neither bounded below, nor bounded above.
4. Prove that the subset $\{(x_1, x_2) : x_2 = -x_1^2\}$ is bounded above, but not below. *(Give an explicit upper bound.)*
5. Prove that the subset $\{(x_1, x_2) : x_2 \leq 0\}$ is a lower set.
6. Prove that the subset $\{(x_1, x_2) : x_2 \geq -x_1^2\}$ is *not* an upper set.
7. Prove that every interval is finite; can you give an estimate of the number of elements in $[(x_1, x_2), (y_1, y_2)]$?
8. Prove that the function $f : \mathbb{Z}^2 \to \mathbb{Z}^2$ (with the same partial order on domain and codomain) given by $f(x_1, x_2) = (x_1 + 1, 2 \cdot x_2)$ is strictly increasing, but is not an order isomorphism to its range.

For the next few questions, consider the set $S = \uparrow((0,0)) \cap \uparrow((1,0))$.

9. Prove that S is bounded below.
10. Prove that S has no infimum.
11. Identify all of the minimal elements of S, if any exists.

Finally one more question,

12. Let $f : \mathbb{Z} \to \mathbb{Z}$; interpreting $f \subseteq \mathbb{Z}^2$ as a subset, formulate a sufficient condition so that the restriction of $\leqq$ to f is a total order.

Problem Set 2

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 2.1. Let $[0,1]_{\mathbb{Q}}$ denote the closed interval between 0 and 1 in the poset $\mathbb{Q}$, with the usual $\leq$. Consider the set $X = \{f : [0,1]_{\mathbb{Q}} \to \mathbb{Q}\}$, ordered by $\trianglelefteq$ according[1] to Example 2.4. For $n \in \mathbb{N}$, denote by $f_n \in X$ the function $f_n(x) = x^n$; and let $F = \{f_1, f_2, \ldots\}$.

For each of $\max F$, $\min F$, $\sup F$, $\inf F$, indicate whether it exists (with proof), and if it does, identify the corresponding function. (For sup and inf, the set F is to be considered as a subset of the poset $(X, \trianglelefteq)$.)

Problem 2.2. Consider the natural numbers $\mathbb{N}$ with the relation[2] $\preceqq$:

$$n \preceqq m \iff \exists k \in \mathbb{N}, k \cdot n = m.$$

1. Consider the subset $\mathbb{N} \setminus \{1\}$, what are the minimal elements?
2. Let $S \subsetneq \mathbb{N}$ be a non-empty finite set, prove that $\inf S$ and $\sup S$ both exist; describe them.
3. Construct a non-trivial (meaning, not the identity map) order isomorphism of $\mathbb{N}$ to itself.

The following two questions concerns the definition:

> Let $(X, \preceqq)$ be a poset. An *anti-chain* is a subset $A \subseteq X$ such that any two distinct elements of A are not comparable.

Denote by[3] $\mathscr{A}_X$ the set of all anti-chains of X. Define the relation $\trianglelefteq$ on $\mathscr{A}_X$ by:

> $A \trianglelefteq A'$ if and only if $\forall x \in A$ there exists $x' \in A'$ such that $x \preceqq x'$.

Problem 2.3. Verify that given the poset $(X, \preceqq)$, the pair $(\mathscr{A}_X, \trianglelefteq)$ is a poset.

Problem 2.4. Prove that: when X is a finite set, $\max \mathscr{A}_X$ is exactly the set of all maximal elements of X.

[1] The LaTeX code for $\trianglelefteq$ is `\trianglelefteq`.
[2] The LaTeX code for $\preceqq$ is `\preceqq`.
[3] The LaTeX code for $\mathscr{A}$ is `\mathscr{A}`.

Reading Assignment 3

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We begin by continuing our story about totally ordered sets. We introduce the notion of "cuts" and use it to give a fairly intuitive definition of "completeness" of a totally ordered set. In this abstract setting we prove two equivalent, characterising properties of complete, totally order sets: the first being the possession of the least upper bound property, the second being the possession of the Heine-Borel property. While the "cut" definition of completeness is easier to understand intuitively, the least upper bound property and the Heine-Borel property are much easier to wield when studying properties of the real numbers and their functions. Next we show that the rationals fail to be Dedekind complete, and provide a procedure of constructing the real numbers from the rationals by the procedure of Dedekind cuts, showing that there does exist such a thing as a Dedekind complete totally ordered field. For the remainder of the course we will not use the specific details of this construction, and will just consider $\mathbb{R}$ as axiomatically a totally ordered field that is Dedekind complete, and study its properties based on this description. This point of view is why we have chosen to prove the fundamental Heine-Borel and least upper bound properties in the setting that we did. This set of readings close out with some first properties of the real numbers. More will be discussed starting next week.

Contents

§3.1 More about Total Orders

We pick up where we left off last week, about total orders. In this section we are no longer enforcing the algebraic condition of "being a field". The following proposition is useful:

Proposition 3.1. *Let $(X, \leqq)$ be a total order. Let A and B be both upper (both lower) sets (see Definition 2.8). Then either $A \subseteq B$ or $B \subseteq A$.*

Proof. If either A or B is empty, the conclusion follows trivially. So we assume that neither is empty.

We prove by contradiction. If neither A nor B are a subset of the other, then both $A \setminus B$ and $B \setminus A$ are non-empty. Choose $\alpha \in A \setminus B$ and $\beta \in B \setminus A$. Since $\leqq$ is connex, either $\alpha \leqq \beta$ or $\beta \leqq \alpha$. But as A and B are both upper (lower), in the first case this would require $\beta \in \uparrow(\alpha) \subseteq A$ (resp. $\alpha \in \downarrow(\beta) \subseteq B$), contradicting the assumption that $\beta \notin A$ (resp. $\alpha \notin B$). The second case is similar and omitted. □

Our goal this section is to define the notion of order completeness through the concept of cuts of a total order. We will also discuss two of the most important consequences of order completeness, namely, the least upper bound property and the Heine-Borel property.

§3.1.1 Cuts and completeness.—The starting point of the definition of cuts is Exercise 2.8, where we showed that the poset is totally ordered if and only if given any element a, its up and down sets $\uparrow(a)$ and $\downarrow(a)$ cover X. Our mental picture should be that a totally ordered set is one which can be "disconnected" by the removal of any one element.

Definition 3.2. *Let $(X, \leqq)$ be a totally ordered set. We say that pair of subsets (X_-, X_+) form a* cut *of X if*

1. *$\{X_-, X_+\}$ is a partition of X; and*
2. *X_- is a lower set and X_+ is an upper set.*

Exercise 3.1. Prove that if (X_-, X_+) is a cut, then every pair of elements $(x_-, x_+) \in X_- \times X_+$ satisfy $x_- \lneqq x_+$.

In plain English, a cut is a decomposition of X into an upper set and a lower set. Notice that given $a \in X$, the following two are *both* cuts of X: the pair $(\downarrow(a), \uparrow(a) \setminus \{a\})$, as well as the pair $(\downarrow(a) \setminus \{a\}, \uparrow(a))$. Notice that at most one of the pair can contain the end point a. It is, however, possible for a totally ordered set to admit a cut for which *neither* of the components is principal.

Example 3.3. Let $X = \mathbb{Q} \setminus \{0\}$. Let X_+ be the set of all positive rationals; this is clearly an upper set. Let X_- be the set of all negative rationals; this is a lower set. Their union is X, and hence together (X_-, X_+) is a cut. But neither X_- nor X_+ is principal: by Exercise 2.20 upper and lower sets are principal if and only if they contain their minimum and maximum respectively. And it is easy to see that neither X_- nor X_+ contains an extremum. ■

This motivates our definition of completeness for total orders:

Definition 3.4. *A totally ordered set $(X, \leqq)$ is said to be* Dedekind complete *if in every cut (X_-, X_+) of X, at least one of X_- and X_+ is principal.*

The mental picture should be that a totally ordered set should be a collection of points arranged on a line. A cut can be imagined as a slice by a hypothetical knife, breaking the collection in twain. The set is said to be complete if, no matter how you cut it, you can always identify at least one point that is "right next to the slice"; this point could be on either the left of the right of the knife. In other words, a totally ordered set is Dedekind complete if every cut can be identified as "immediately below x_0" or "immediately above x_0" for some x_0.

Not all total orders are complete (see Example 3.3).

Example 3.5. Any well-ordering is Dedekind complete. If $(X, \leqq)$ is well-ordered, and (X_-, X_+) is a cut, then X_+ is a non-empty subset and by definition has a minimum, and hence is principal. In particular $\mathbb{N}$ is Dedekind complete. ■

Exercise 3.2. Prove that $\mathbb{Z}$ with the usual ordering is Dedekind complete.

Exercise 3.3. Prove that the set $X = \{1, \frac{1}{2}, \frac{1}{3}, \ldots\}$ with the usual ordering is Dedekind complete. But denoting by $-X = \{q \in \mathbb{Q} : -q \in X\}$, the set $X \cup (-X)$ is not Dedekind complete.

The above exercise shows that Dedekind completeness is a *global* property and it is not sufficient to just prove Dedekind completeness for some collection of subsets. It is, however, sufficient if you choose your subsets wisely.

Exercise 3.4. Let $(X, \leqq)$ be a total order. Let $a, b \in X$ with $a \lneqq b$. Prove that, if $(\uparrow(a), \leqq)$ and $(\downarrow(b), \leqq)$ are both Dedekind complete, then so is X. *(Hint: if (X_-, X_+) is a cut of X, start by showing that either $(X_- \cap \uparrow(a), X_+ \cap \uparrow(a))$ is a cut of $\uparrow(a)$, or $(X_- \cap \downarrow(b), X_+ \cap \downarrow(b))$ is a cut of $\downarrow(b)$; why doesn't this argument work with Exercise 3.3?)*

While there are some potential problems with *extending* Dedekind completeness from subsets to the whole set, restrictions are not an issue.

Proposition 3.6. *Let $(X, \leqq)$ be a Dedekind complete total order. Then the total order restricted to $\uparrow(a)$ and $\downarrow(a)$ for any $a \in X$ is also Dedekind complete.*

Proof. We only include the proof for the case $\uparrow(a)$; the other is similar.

Let (A_-, A_+) be a cut of $\uparrow(a)$. Then noting that $A_- \cup \downarrow(a)$ is a lower set in X, we see that setting $X_- = A_- \cup \downarrow(a)$ and $X_+ = A_+$ gives us a cut of X. Then either X_+ has a minimum (in which case A_+ has a minimum), or X_- has a maximum (in which case since any element of A_- is an upper bound of $\downarrow(a)$, the maximum is also the maximum of A_-). □

Food for Thought 3.5. Make sure you can fill in all the suppressed details in the above proof. Can you extend the proposition to also $\uparrow(a) \setminus \{a\}$?

A final word on the terminology: the word "complete" is very heavily overloaded in mathematics. Even in order theory there is a distinction between the two concepts[1] "Dedekind completeness" and "order completeness", to say nothing of the notions of completeness in logic, metric geometry, integrals curves of vector fields, just to name a few. In the limited scope of this course we will for the most part take completeness to be synonymous with Dedekind completeness, but please be aware of the need to be more specific outside of this course.

§3.1.2 Least upper bound property.—A consequence of Dedekind completeness is the *least upper bound property*.

Definition 3.7. *A poset $(X, \leqq)$ is said to possess the* least upper bound property *if every nonempty subset that is bounded above has a supremum; similarly, the poset is said to possess the* greatest lower bound property *if every nonempty subset that is bounded below has a infimum.*

Theorem 3.8. *For a totally ordered set $(X, \leqq)$, the three statements*

1. *X is Dedekind complete;*
2. *X has the least upper bound property;*
3. *X has the greatest lower bound property;*

[1] We will not be using order completeness in this course. For total orders, the difference is that Dedekind completeness requires $\{X_-, X_+\}$ to partition X and hence neither are empty; order completeness allows one of the two to be empty. As a consequence we see that a order complete total order is nothing more than a Dedekind complete total order that is also bounded in itself.

are equivalent.

Proof. (1 $\implies$ 2) Let S be a nonempty subset that is bounded above. Consider first the case that S is finite, then by Exercise 2.17 the set attains is maximum, and hence by Exercise 2.21 has a supremum. (Note that this does not depend on completeness.)

Suppose now that S is infinite; this means S contains at least 2 distinct elements. Let X_+ be the set of all upper bounds of S, this is non-empty by assumption and an upper set by Exercise 2.9 (after taking the inverse order). Let X_- be $X \setminus X_+$. This set is lower because, if $a, b \in X$ and $a < b$, then $a \in X_+ \implies b \in X_+$ since X_+ is upper; thus the contrapositive $b \in X_- \implies a \in X_-$ proves X_- is lower. X_- is also non-empty since S contains distinct elements x_1, x_2; since $\leqq$ is connex we see that $\min\{x_1, x_2\}$ exists and cannot be an upper bound of S, and so must belong to X_-. And hence (X_-, X_+) is a cut.

Dedekind completeness thus implies that either X_+ has a minimum (in which case it is by definition the supremum of S), or X_- has a maximum. Supposing the latter, since $\max X_- \in X_-$ it is *not* an upper bound of S, and hence there exists $s_0 \in S$ such that $s_0 > \max X_-$. I claim that no other element of S is greater than $\max X_-$: for if $s_1 \in S$ were also greater than $\max X_-$, then $\min\{s_0, s_1\}$ is in X_- and greater than $\max X_-$, contradicting the requirement that $\max X_-$ is the maximum. Thus s_0 is greater than $\max X_-$, which is greater than any other element of S. Hence $s_0 = \max S$ and is also the supremum of S.

(1 $\implies$ 3) via essentially the same argument with minor changes.

(2 $\implies$ 1) Let (X_-, X_+) be a cut of X, then X_- is bounded above by any element of X_+. The least upper bound property implies that $\sup X_-$ exists. Since $\{X_-, X_+\}$ partitions X, $\sup X_-$ lives in exactly one of the two sets. If $\sup X_- \in X_+$, then since every element of X_+ is an upper bound of X_-, we see that necessarily $\sup X_- = \min X_+$. If on the other hand $\sup X_- \in X_-$, then by Exercise 2.21 again we have $\sup X_- = \max X_-$. And hence at least one of X_- and X_+ is principal.

(3 $\implies$ 1) via essentially the same argument with minor changes. □

Food for Thought 3.6. Definition 3.4, while making more intuitive sense for totally ordered sets, cannot be easily generalized to partial orders. Hence in the modern literature one generally uses the *possession of the least upper bound property* as the definition of Dedekind completeness for *general posets*. For general posets, however, since Definition 3.4 is no longer operable, the above proof does not prove the equivalence of the least upper bound property and the greatest lower bound property. You should try to prove the equivalence of the two directly. (*Hint: given an upper bounded nonempty set S, the set of its upper bounds is a nonempty set that is bounded below.*)

Exercise 3.7. Let $(X, \leqq)$ be a Dedekind complete total order. Suppose we are given a family of non-empty intervals indexed by $\mathbb{N}$, which we write as $[a_1, b_1]$, $[a_2, b_2]$, $[a_3, b_3]$, and so on. Prove that if the family of intervals are *nested*, in the sense that $[a_{i+1}, b_{i+1}] \subseteq [a_i, b_i]$, then $\cap_{i=1}^{\infty} [a_i, b_i] \neq \emptyset$. (*This result was original due to Cantor. Hint: consider the sets $A = \{a_1, a_2, \ldots\}$ and $B = \{b_1, b_2, \ldots\}$.*)

Exercise 3.8. Give an example of a total order which is *not* Dedekind complete, and a family of non-empty nested closed intervals, such that the intersection of the family is empty.

§3.1.3 Heine-Borel property.—Another super useful consequence of Dedekind complete total orders is the Heine-Borel property. To introduce the property we need a preliminary definition. For simplicity of notation we only describe the case for total orders that are bounded neither above nor below; one can modify the definitions to also deal with total orders with max or min, but the proofs gain additional

complications that obscure the main ideas.

Definition 3.9. *Given $(X, \leqq)$ a partial order, and elements $a, b \in X$, the* open interval (a, b) *is defined to be* $\uparrow(a) \cap \downarrow(b) \setminus \{a, b\}$.

Definition 3.10. *Given $(X, \leqq)$ a total order with no max and no min, an* entourage mapping *is a function $f : X \to 2^X$ such that $f(x)$ is an open interval that contains x.*

We should imagine the entourage mapping as outputting an interval of X whose elements we declare to be "well approximated" by x. (Approximation will be a recurring theme in this course.) We note that if the total order has either a max or a min, then with our definition the set of entourage mappings is empty.

Food for Thought 3.9. In the case where the total order has a max or a min, additional exceptions have to be written into the definition. We need to modify the definition of an entourage mapping so that when $x = \min X$ we allow $f(x)$ to be an interval of the form $[x, b)$, and when $x = \max X$ we allow $f(x)$ to be an interval of the form $(a, x]$ (with the obvious definitions). It is a good (and *entirely optional*) exercise to see how the remaining statements of this section should be modified to account for the max and min.

Definition 3.11. *Given $(X, \leqq)$ a total order with no max and no min, we say that it possesses the* Heine-Borel property *if, for every closed interval $[a, b]$ and every entourage mapping f, there exists a finite subset $S \subseteq [a, b]$ such that $f(S)$ covers $[a, b]$.*

Returning to the analogy above, this says that the local structure of a total order with the Heine-Borel property cannot be too complicated (or that $[a, b]$ cannot be too long). For any definition of "well approximation", the property asserts that we can well approximate any element of $[a, b]$ using just a finite subset of points.

Now we turn to the main theorem in this section, which shows the equivalence of the Heine-Borel property with Dedekind completeness. In the case of the real numbers, that its Dedekind completeness implies the Heine-Borel property is often called the "Borel covering lemma".

Theorem 3.12. *Suppose $(X, \leqq)$ is a total order with no max and no min. Then it is Dedekind complete if and only if it possesses the Heine-Borel property.*

For clarity of exposition, we divide the proof into separate parts for the two implications. We first prove the reverse implication.

Lemma 3.13. *Suppose $(X, \leqq)$ is a total order with no max and no min, and that it is* not *Dedekind complete. Then there exists an interval $[a, b]$ and an entourage mapping f such that for any finite $S \subseteq [a, b]$, $[a, b] \setminus \cup f(S) \neq \emptyset$.*

Proof. Since X is not Dedekind complete, there exists a cut (X_-, X_+) of X such that X_- has no maximum and X_+ has no minimum. Since neither X_- nor X_+ is empty, we can choose $a \in X_-$ and $b \in X_+$. Since X has no max and no min, there exists $a' \lneqq a$ and $b' \gneqq b$. Define the entourage mapping f as follows.[2]

- If $x \lneqq a$, set $f(x) = (x', a)$ for some $x' \lneqq x$ (which exists because X has no minimum).
- If $x \gneqq b$, set $f(x) = (b, x')$ for some $x' \gneqq x$ (which exists because X has no maximum).

[2] The first two items are not really necessary to the argument, which doesn't really depend on how f is defined outside of $[a, b]$. But we include a precise description just to make the function f concrete.

- If $x \in [a,b] \cap X_-$, set $f(x) = (a', x')$ for some $x' \gneqq x$ (which exists because X_- has no maximum).
- If $x \in [a,b] \cap X_+$, set $f(x) = (x', b')$ for some $x' \lneqq x$ (which exists because X_+ has no minimum).

Let now S be any finite subset of $[a,b]$. If S is empty clearly $f(S)$ cannot cover $[a,b]$. If S is non-empty, then at least one of $X_- \cap S$ and $X_+ \cap S$ is finite and non-empty. We shall assume it is X_- (the other case can be argued similarly).

From the definition we have $f(S \cap X_+) \subseteq X_+$ and hence is disjoint from X_-. It suffices then to show that $f(S \cap X_-)$ cannot cover $X_- \cap [a,b]$. List the elements of $S \cap X_-$ as $\{s_1, \dots, s_n\}$. Corresponding to each s_i there is an element $t_i \in X_-$ such that $f(s_i) = (a', t_i)$. Since the list $\{t_1, \dots, t_n\}$ is finite, it contains a maximum which we call t'. Then by definition $\cup f(S \cap X_-) = (a', t')$. But $X_- \cap [a,b] \setminus (a', t')$ contains t' as an element; this shows the result. □

For the forward implication, we will use a continuous version of mathematical induction.[3]

Lemma 3.14 (Continuity argument)**.** *Let $(X, \leqq)$ be a Dedekind complete total order with a minimum x_0. Suppose $S \subseteq X$ satisfies the following properties:*

(Non-empty) $x_0 \in S$.
(Initial) $x \in S \implies [x_0, x] \subseteq S$.
(Continuation) *If* $\max S$ *exists, then* $\max X$ *exists and equals* $\max S$.
(Closure) *If* $\sup S$ *exists, it is in* S.

Then $S = X$.

Proof. S is either bounded or unbounded.

Suppose S is bounded. The Dedekind completeness of X implies by Theorem 3.8 that it has the least upper bound property. Since by assumption $x_0 \in S$ and it is non-empty, this implies that $\sup S$ exists. By the closure property of S we then know that $\sup S \in S$, which by Exercise 2.21 we see means $\sup S = \max S$. By the continuation property of S this requires $\max S = \max X$, and by the initial property this means $S = [x_0, \max X] = X$.

Suppose S is unbounded. We suppose further for contradiction that $S \neq X$. Then there exists $y \in X \setminus S$. The unboundedness of S requires there to be $z \in S$ with $z \geqq y$. But the initial property of S requires $[x_0, z] \subseteq S$ and hence $y \in S$, a contradiction. □

With the continuity argument, we will prove the forward implication.

Lemma 3.15 (Borel covering lemma)**.** *Suppose $(X, \leqq)$ is a Dedekind complete total order with no max and no min. Given an interval $[a,b]$ and an entourage mapping f, then there exists a finite set $S \subseteq [a,b]$ such that $f(S)$ covers $[a,b]$.*

Proof. Applying Proposition 3.6 twice, we find that $[a,b]$ with restricted ordering $\leqq$ is again a Dedekind complete total order. Let Z be defined as

$$Z \coloneqq \{x \in [a,b] : \exists S_x \subseteq [a,b] \text{ finite, such that } \cup f(S) \supseteq [a,x]\}.$$

[3] This, order theoretic, version of the continuous induction is closely related to a version in topology, which states that if (X, τ) is a connected topological space, and $S \subseteq X$ is a non-empty subset that is both open and closed, then $S = X$.

In view of Lemma 3.14, it suffices to check that Z has the four listed properties.

It is non-empty since $a \in Z$; we can set $S_a = \{a\}$. And it is initial by definition: if $x \in Z$ with corresponding finite set S_x, then for any $y \in [a, x]$ we can set $S_y = S_x$ to show that $y \in Z$.

We next claim that $x \in Z \implies \min\{b, \sup f(x)\} \in Z$. We prove by cases. Suppose first that $b \lneqq \sup f(x)$, then since $x \leqq b$ we have that $b \in f(x)$. And hence $S_x \cup \{x\}$ which is still a finite set, has an image under f that covers $[a, b]$. Suppose next that $b \geqq \sup f(x)$. I claim the set $S_x \cup \{x, \sup f(x)\}$ gives rise to a covering of $[a, \sup f(x)]$. It suffices to show that every $y \in [x, \sup f(x)]$ belong to either $f(x)$ or $f(\sup f(x))$. But if $y \lneqq \sup f(x)$ by definition it belongs to $f(x)$; and if $y = \sup f(x)$ we have that $y \in f(\sup f(x))$.

Since $\sup f(x) \gneqq x$ by definition, the above claim implies that Z can have *no maximum* except for b, showing the continuity property of Z.

To finish the proof we therefore must show that any supremum of Z is in fact an element of Z. Set $z_0 = \sup Z$. If $z_0 = a$ then as already established $z_0 \in Z$. Suppose thus that $z_0 \gneqq a$. If y is such that $a \leqq y \lneqq z_0$, then y is not an upper bound of Z, and hence there exists $y' \in Z$ with $y' \geqq y$. By the initial property of Z this implies $y \in Z$. Denote by $z_1 = \max\{a, \inf f(z_0)\}$. Since $a \leqq z_1 \lneqq z_0$, we have $z_1 \in Z$, hence there exists a finite set S_{z_1} such that $f(S_{z_1})$ covers $[a, z_1]$. On the other hand, by construction $[a, z_0] \setminus [a, z_1] \subseteq f(z_0)$, and we see that the finite set $S_{z_1} \cup \{z_0\}$ shows $z_0 \in Z$. □

§3.2 The Reals

And now, three weeks into the course, we finally get to our main player, the real numbers. In this section we will first discuss some properties of the rational numbers, including its *incompleteness*. Then we will show how to construct the real numbers by "filling in the holes" left by the rationals.

§3.2.1 The rationals.—Let's talk about the rationals. Drawing on our "intuitive" knowledge of the rationals, we know that it is holey, since it is missing all the irrational numbers. But exactly how holey is the rationals?

Proposition 3.16 (Archimedean property of $\mathbb{Q}$)**.** *If $p, q \in \mathbb{Q}$ are both positive, then there exists $n \in \mathbb{N}$ such that $np > q$.*

Proof. We can write $p = a/b$ and $q = c/d$ with $a, b, c, d \in \mathbb{N}$. Then choosing $n = 2bc$ we find $np = 2ac > c \geq c/d = q$. The first inequality because $2a \geq 2 > 1$ for $a \in \mathbb{N}$, the second because $d \geq 1$. □

Corollary 3.17. *If $p, q \in \mathbb{Q}$ and $p < q$, then there is $r \in \mathbb{Q}$ with $p < r < q$.*

Proof. Let $u = q - p > 0$, it is rational. Applying the Archimedean property to the pair $(u, 1)$, we see that there exists $n \in \mathbb{N}$ such that $nu > 1$. Rearranging we get $q > p + 1/n > p$ so declaring $r = p + 1/n$ we are done. □

A particular consequence is that for any positive rational q, there must be a smaller positive rational r; this means that there are no such things as "infinitesimals" in $\mathbb{Q}$: that the positive rationals does not have a minimal element. Another consequence is that there are not sizable gaps among the rationals: if there is a positive difference between any two rational numbers, there is something else that goes between them. And hence any holes among the elements of $\mathbb{Q}$ must have "width zero".

As everyone knows, there are plenty of irrational numbers formed by repeated radicals, such as $\sqrt{2}$ or $\sqrt{3+\sqrt[3]{15}}$. So we know that some of these holes that we are looking for must be filled by such numbers. A natural question to ask is: why not just fill the holes with this construction? More precisely, consider polynomials with rational coefficients, many of their roots are not rational. Why not just add those numbers to our line? Isn't that enough?

There are two problems with this argument.

1. The addition of roots of rational polynomials will necessarily force us to confront $\sqrt{-1}$; as we saw in the last set of readings, imaginary numbers cannot slot into the number line, as they cannot be placed consistently within the ordering we have established for $\mathbb{Q}$. So this process of taking "algebraic closure" will necessarily give us *too many new numbers*.
2. On the other hand, we also get *too few new numbers* by this process. First, if x is a root of the polynomial $p(x)$, then it also is a root of the polynomial $a \cdot p(x)$ for any $a \in \mathbb{N}$. Thus by taking common denominators we find that any root of a rational polynomial is also a root of a polynomial with integer coefficients. Appealing a bit to the fundamental theorem of algebra, we note that a polynomial of degree d can have at most d roots. So considering the function $\nu : \{\text{Integer coefficient polynomials}\} \to \mathbb{N}$ where $\nu(p)$ is set to equal

$$\nu(p) = \text{ degree of } p + \text{ sum of absolute values of the coefficients of } p,$$

we see that each integer coefficient polynomial p corresponds to a finite number $\nu(p)$. Given $k \in \mathbb{N}$, let P_k be the set of integer coefficient polynomials satisfying $\nu(p) = k$. We see that elements of P_k have degree at most k and coefficients at most k in absolute value, so there are no more than $(2k+1)^k$ polynomials in P_k, representing no more than $k \cdot (2k+1)^k$ possible roots. So the set of numbers (these are called the "algebraic numbers") that you can construct from this process of adding roots of polynomials is no more in size then the size of the union of countably infinitely many finite sets. As you saw on the Exercise Sheet for Week 1, this means that the set of "algebraic numbers" is countably infinite, and hence far from containing all the other real numbers. (In fact it fails to include well-known numbers such as π or e.)

A next question is then: does using the order structure help? Can the order structure in fact detect these kinds of holes?

Example 3.18. $\mathbb{Q}$ is *not* Dedekind complete.

To prove this we need to exhibit a cut for which neither X_- nor X_+ is principal. Define

$$X_- = \{q \in \mathbb{Q} : q^2 < 2 \text{ or } q \leq 0\},$$
$$X_+ = \{q \in \mathbb{Q} : q^2 > 2 \text{ and } q > 0\}.$$

First, we need to prove (X_-, X_+) is a cut of $\mathbb{Q}$.

- The defining conditions of X_- and X_+ are mutually exclusive, so they are disjoint. To show that they form the partition, we need to show that they are non-empty (obvious) and that their union is $\mathbb{Q}$. For the latter, it suffices to show that every element $q \in \mathbb{Q}$ belongs to one of the two sets. We treat by cases: if $q \leq 0$, then $q \in X_-$. If $q > 0$, then $q^2 > 0$, and we ask how it compares to 2. If it is above it goes in X_+, if below it goes into X_-. What if $q^2 = 2$? This is impossible as $\sqrt{2}$ is not rational.
- That X_- is a lower set: let $p \in X_-$ and $q \in \mathbb{Q}$ satisfy $q < p$. If $q \leq 0$ then there's nothing to prove. If $q > 0$, then so does p, and hence $q < p \implies q^2 < p^2$ and hence by transitivity $q^2 < 2$ and $q \in X_-$.

- That X_+ is an upper set: let $p \in X_+$ and $q \in \mathbb{Q}$ with $p < q$. Since $p > 0$ we have by transitivity $q > 0$. And since they are both positive, $p < q \implies p^2 < q^2$ which implies $2 < q^2$ and hence $q \in X_+$.

Next, we need to show that neither X_- nor X_+ is principal.

- For X_-, it suffices to show it has no maximum. Let $q \in X_-$. If $q \leq 0$ then noticing $q < 1$ and $1 \in X_-$, we see q cannot be the maximum. So we next consider $q > 0$ and $q^2 < 2$. This implies $\frac{1}{q} - \frac{q}{2} > 0$, as well as $1 - \frac{q^2}{2} > 0$. Hence we can find natural numbers n, m (using the Archimedean property) such that
$$\frac{1}{n} < \frac{1}{2q} - \frac{q}{4}, \qquad \frac{1}{m^2} < \frac{1}{m} < 1 - \frac{q^2}{2}.$$
Letting $n_0 = \max\{m, n\}$, we see that
$$(q + \frac{1}{n_0})^2 = q^2 + \frac{2q}{n_0} + \frac{1}{n_0^2} < q^2 + 2 - q^2 = 2.$$
And hence $q + \frac{1}{n_0} \in X_-$ and shows q is not the maximum.
- For X_+, it suffices to show it has no minimum. Let $q \in X_+$, then it is positive and $q^2 > 2$. And hence $\frac{q}{2} - \frac{1}{q} > 0$ and by the Archimedean property there is some natural n such that $\frac{1}{n} < \frac{q}{2} - \frac{1}{q}$. Rearranging we find
$$(q - \frac{1}{n})^2 = q^2 - \frac{2q}{n} + \frac{1}{n^2} > q^2 - \frac{2q}{n} > 2.$$
Furthermore, as $\frac{1}{n} < \frac{q}{2}$ by construction, we have $q - \frac{1}{n} > 0$. So $q - \frac{1}{n} \in X_+$ and q cannot be a minimum.

The argument can be modified to show that the cut corresponding to any irrational algebraic number (meaning that can be written as the root of a polynomial with rational coefficients) would have neither half principal, and thereby showing that $\mathbb{Q}$ has a "hole" right there. ■

Before we proceed to construct the reals, we need the following proposition, which is a consequence of the Archimedean property.

Proposition 3.19. *Let (X_-, X_+) be a cut of $\mathbb{Q}$. Then* at most *one X_- and X_+ can have an end point.*

Proof. Suppose $\max X_-$ and $\min X_+$ both exist. Since X_- is a lower set and X_+ is an upper set, and they are disjoint, we see $\max X_- < \min X_+$. Then by Corollary 3.17 there exists $r \in \mathbb{Q}$ with $\max X_- < r < \min X_+$; but this is absurd as such $r \in \mathbb{Q} \setminus (X_+ \cup X_-)$, contradicting the hypothesis that (X_-, X_+) forms a cut. □

§3.2.2 The Dedekind construction.—Having shown that that Dedekind completeness is sensitive enough to see at least the holes left by at least the algebraic irrational numbers, let's now fill in the holes left by cuts. By Proposition 3.19, any cut of $\mathbb{Q}$ has at most one of the two halves in possession of an end point; further more if $\max X_-$ exists, then $(X_- \setminus \{\max X_-\}, X_+ \cup \{\max X_-\})$ is another cut. So we can define

Definition 3.20. *The set $\mathbb{R}$ is defined to be the set of all cuts (X_-, X_+) of $\mathbb{Q}$, such that X_- has no maximum.*

Obviously, we will embed $\mathbb{Q}$ into $\mathbb{R}$ by identifying $q \in \mathbb{Q}$ with the cut whose $X_+ = \uparrow(q)$. We can equip $\mathbb{R}$ with an ordering.

Definition 3.21. *We equip* $\mathbb{R}$ *with an ordering* $(X_-, X_+) \leq (Y_-, Y_+) \iff X_- \subseteq Y_-$.

By virtue of Proposition 3.1 we see that this ordering is connex, antisymmetric, reflexive, and transitive, and hence is a total order. That $\mathbb{R}$ is Dedekind complete follows from the following: suppose $S \subset \mathbb{R}$ is bounded above, then there exists an element $q \in \mathbb{Q}$ such that if $(X_-, X_+) \in S$, then $X_- \not\ni q$. Now, let $Y_- = \cup\{X_- \subseteq \mathbb{Q} : (X_-, X_+) \in S\}$. It is easy to see that Y_- is a nonempty lower set. Let $Y_+ = \mathbb{Q} \setminus Y_-$; it is an upper set. And since $q \in Y_+$ it is non-empty. And hence (Y_-, Y_+) is a cut of $\mathbb{Q}$. Y_- has no maximum since, had it one, $\max Y_- \in X_-$ for some X_-, and by definition must be its maximum, contradicting the assumption that $(X_-, X_+) \in \mathbb{R}$, and we see $(Y_-, Y_+) \in \mathbb{R}$. Finally we shows that $(Y_-, Y_+) = \sup S$: it is clear that by definition it is an upper bound. Suppose $Z_- \subsetneq Y_-$ is a lower set, then there exists $r \in Y_- \setminus Z_-$ and by the definition of Y_- there exists some $(X_-, X_+) \in S$ with $X_- \supseteq {\downarrow}(r) \supsetneq Z_-$ showing that (Z_-, Z_+) cannot be an upper bound of S.

Having discussed the order structure of $\mathbb{R}$, we also need to sketch the arithmetic properties of $\mathbb{R}$. Addition is fairly simple. If (X_-, X_+) and (Y_-, Y_+) are in $\mathbb{R}$, we set their sum to be the cut (Z_-, Z_+) where Z_- contains elements which can be written as sums $x + y$ with $x \in X_-$ and $y \in Y_-$. It is easy to check that such Z_- can contain no maximum.

Food for Thought 3.10. You should try to convince yourself quickly that this definition of addition restricts to that of $\mathbb{Q}$ under the identification described above, and that every element of $\mathbb{R}$ has an additive inverse under this definition.

Multiplication is a tiny bit more difficult. We will use that multiplication respects the order if the factors are both positive. Now suppose $x = (X_-, X_+)$ and $y = (Y_-, Y_+)$ are in $\mathbb{R}$.

- If either x or y is equal to 0, then set the product to 0.
- If exactly one of x and y is negative, replace it by its additive inverse, and set the product to be the additive inverse of the resulting product of positive numbers.
- If both x and y is negative, set the product to be the product of their additive inverses.
- If both x and y are positive: let their product be $z = (Z_-, Z_+)$ where $\zeta \in Z_-$ if there exists $\xi \in X_- \cap {\uparrow}(0)$ and $\eta \in Y_- \cap {\uparrow}(0)$ with $\zeta < \xi \cdot \eta$. Since neither X_- nor Y_- contain their maxima, this definition ensures that Z_- also has no maximum.

Exercise 3.11. *(This is a very lengthy exercise! I don't encourage you to do it unless you have nothing better to do for an evening.)* Check that the multiplication operations restricts to that of $\mathbb{Q}$ appropriately, that every non-zero element has a multiplicative inverse, and that multiplication distributes over addition. In other words, check that the field axioms hold for $\mathbb{R}$.

Having constructed the real numbers from the rationals, we will proceed to forget the entire Dedekind construction (the same way we will forget about how $\mathbb{Z}$ and $\mathbb{Q}$ are made up from $\mathbb{N}$) and instead use our prior intuition of what $\mathbb{R}$ feels like when trying to visualize it, and use the fact that

> "$\mathbb{R}$ is a Dedekind complete totally ordered set, equipped with an addition and a multiplication that satisfy the field axioms and are compatible with the ordering"

when writing our proofs. The content of this section can be largely described as proving the *existence* of such an object that is $\mathbb{R}$. *A priori* it is possible that the conditions imposed when defining $\mathbb{R}$ can be mutually exclusionary leaving no such object to exist.

> *An aside*: the question whether the conditions are mutually exclusionary, and whether the

> real numbers exist, is not purely academic. We would like such a thing as a Dedekind complete totally ordered field to exist. But to form such an object we need to add new points to $\mathbb{Q}$; and each time we add a point we have to take care to also add all of its sums against existing numbers, all of its multiples and their sums against existing number, all of its powers and their multiples etc. Whether such a process can terminate is not guaranteed! The fact that we have a way of actually filling in all the holes is already a minor miracle.
>
> On top of that, since our method of filling in the hole is order theoretic in nature, it is natural to worry whether the numbers we added can still coherently integrate the field axioms as part of their structure. Seen from this angle, the real numbers turns out to be a sort of Goldilocks solution. Things turn out to work just right for the real numbers to exist, and be essentially unique. To see how delicate this is: if you try to stick *even more* numbers into the real number line, it turns out that between Dedekind completeness and the field axioms, *at least one* will have to give.

Of course, it is very likely that you are not content with our construction above. Specifically: you may have two questions.

1. Earlier we harped about how just adding the algebraic irrationals to $\mathbb{Q}$ is not adding *enough* numbers, how do we know that the Dedekind construction is adding enough numbers?
2. The discussion in the previous section showed that real algebraic irrationals correspond to bad cuts of $\mathbb{Q}$, and hence will appear as elements of $\mathbb{R}$. But how do I know my favourite non-algebraic (a.k.a. transcendental) irrational number (say π or e) is represented in $\mathbb{R}$?

For the first question, you will get an answer in the next section. (Or rather, you get that assigned as an exercise in the next section.)

For the second, the answer is that you are asking the wrong question. The correct question to ask is: suppose I have a description of a number, how do I prove that it is real? For most everyday household names (like π or e), it turns out one can usually find an approximation of such numbers either by rationals or by algebraic numbers. In this case based on what we will discuss later about sequences we can realize the number as a point in $\mathbb{R}$. In fact, if there is a way to compute its decimal expansion, then there is a way to prove that this number is in $\mathbb{R}$.

However, it is also possible to describe a number by asserting that it is the solution to a particular problem, or that it is a number with such and such properties. In these cases the situation becomes more subtle. We will not touch upon such deep questions in this course, but some keywords to search for if you want to do extracurricular reading are "effective descriptive set theory", "computability theory", and "arithmetic hierarchy."

§3.3 Some first properties of the reals

The statements described in this section can *all* be proven directly in a very hands-on manner by using the Dedekind construction. However, *we will not do it like this*; and instead we will prove them by starting with $\mathbb{R}$ as *axiomatically* a totally ordered field that is Dedekind complete. The reason is to emphasize that these statements are true *regardless* of how you constructed $\mathbb{R}$ to start with, and are therefore universal to any model of the real numbers. We will freely use the fact that every totally ordered field contains a copy of $\mathbb{Q}$.

Earlier we asserted the Archimedean property of $\mathbb{Q}$. Now let us prove it for $\mathbb{R}$.

Theorem 3.22. *If x, y are positive real numbers, then there exists n a natural number so that $n \cdot x > y$.*

Proof. The statement is equivalent to showing that the set S of positive integer multiples of x is not bounded above for any positive real x. Suppose for contradiction that S is bounded above, then since $\mathbb{R}$ is Dedekind complete S has a supremum, which we denote by s_0. The element $s_0 - x < s_0$, and hence is not an upper bound of S, and hence there exists $n \in \mathbb{N}$ such that $n \cdot x > s_0 - x$. But rearranging we find $(n+1)x > s_0$, and as $n+1$ is still a natural number, this shows that s_0 cannot be an upper bound of S after all. □

Notice that in addition to the Dedekind completeness, this proof also required the compatibility of the arithmetic operations of addition and multiplication with the order $\leq$, as well as the field axioms (distributive law).

Exercise 3.12. Using the Archimedean property of $\mathbb{R}$ and the axiomatic description of $\mathbb{R}$, prove the following statements.

1. Given any positive real x, there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < x < n$.
2. The set of *irrational* real numbers (so $\mathbb{R} \setminus \mathbb{Q}$) is unbounded in $\mathbb{R}$.

Example 3.23. A consequence of the Archimedean property of the real numbers is that $\mathbb{Q}$ is order-dense in $\mathbb{R}$. We will in fact prove the following statement: "Let $x, y \in \mathbb{R}$ with $x < y$, and let $\alpha > 0$ be a real number, then there exists a non-zero rational number q such that $q \cdot \alpha \in (x, y)$."

Note that we can assume x and y are both non-negative or both non-positive: Suppose x and y have different signs so $x < 0 < y$. Then applying the statement to $x' = 0$ and $y > 0$ we obtain a rational multiple of α between $(0, y)$, and hence also between (x, y). Without loss of generality we can assume $0 \leq x < y$; else we replace x, y, and q by their negatives.

By the Archimedean property, there exists a natural number n such that $n \cdot (y - x) > \alpha$, so that $\alpha/n < y - x$. Let $\beta = \alpha/n$. Notice that since x is non-negative, $\beta < y$. By the Archimedean property again, there exists a natural number m such that $m \cdot \beta > y$. Consider the set $\{k \in \mathbb{N} : k \cdot \beta < y\}$. This set is non-empty since it contains 1. This set is bounded since all of its elements are no more than m. Hence this set is finite. Let k_0 be its maximum (recall that any non-empty finite set in a totally order set has a maximum). I claim that $k_0 \cdot \beta > x$.

Suppose the claim is false, and that $k_0 \cdot \beta \leq x$, then since $\beta < y - x$, we have that $(k_0 + 1) \cdot \beta < y$; but this contradicts the choice of k_0 as the maximum. Unwrapping the definitions we see that setting $q = k_0/n$ we get our desired rational number.

Setting $\alpha = 1$ we obtain that between any two distinct real numbers there exists a rational. (This conclusion is the statement that $\mathbb{Q}$ is order dense in $\mathbb{R}$.) Setting α to be your favourite irrational number (say, the Golden Ratio), you obtain that between any two distinct real numbers there exists a irrational number. ■

Exercise 3.13. The Dedekind completeness of $\mathbb{R}$ turns out to be enough to show that it is uncountable. (This was in fact Cantor's *first* proof of this statement.) I'll sketch the proof for you: see if you can figure out how to finish it.

1. It suffices to show that any function $f : \mathbb{N} \to \mathbb{R}$ cannot be surjective. (Why?)
2. Let $a_1 = f(1)$. If $f(n) \leq a_1$ for all $n > 1$, then we are done. (Why?)

3. Else, there exists a smallest n such that $f(n) > a_1$, call this value of $f(n) = b_1$. (Why is there a smallest n?)
4. Next, consider the smallest n (if it exists) such that $f(n) \in (a_1, b_1)$, and call this value of $f(n) = a_2$. (What if n doesn't exist?)
5. Next, consider the smallest n (if it exists) such that $f(n) \in (a_2, b_1)$, and call this value of $f(n) = b_2$. (Do you see the pattern? What should the next steps in the construction be?)
6. (How would you use the numbers $a_1, a_2, a_3, \ldots$ and $b_1, b_2, b_3, \ldots$ to show that f cannot be a surjection?)

Exercise 3.14. Prove that given a, b in $\mathbb{R}$ with $a < b$, the open interval (a, b) satisfies $\sup(a, b) = b$ and $\inf(a, b) = a$.

Food for Thought 3.15. The above exercise seems trivial, but requires several of the properties of $\mathbb{R}$. For example, $\mathbb{N}$ is also a Dedekind complete total order, but it is not true that $\sup(a, b) = b$ or $\inf(a, b) = a$ for *any* open interval (as defined in Definition 3.9). The fact that $\mathbb{R}$ has the Archimedean property (which is a consequence of the field axioms) is crucial in the exercise.

Definition 3.24. *We say that a subset $U \subseteq \mathbb{R}$ is* open *if for every $x \in U$, there is an open interval $(a_x, b_x) \subseteq U$ with $x \in (a_x, b_x)$.*

We say that a subset $K \subseteq \mathbb{R}$ is closed *if $\mathbb{R} \setminus K$ is open.*

Exercise 3.16. Let $[a, b]$ be a bounded interval in $\mathbb{R}$. Prove that given any set $\mathscr{S}$ of *open* subsets of $\mathbb{R}$, such that $\cup\mathscr{S} \supseteq [a, b]$, there exists a finite subset $\mathscr{D} \subseteq \mathscr{S}$ such that $\cup\mathscr{D} \supseteq [a, b]$. *(In plain English, this states that if $[a, b]$ is a bounded closed interval, then every open cover of $[a, b]$ has a finite subcover.)*

Exercise Sheet: Week 3

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 3.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a function, so that $f(x) > 0$ for all x. Let $[a, b]$ be a closed, bounded interval. Prove that there exists a partition of $[a, b]$ into finitely many intervals (the intervals may be closed, open, or have one side open and the other side closed), such that each interval I in the partition contains a point $p \in I$ such that $p - \inf I$ and $\sup I - p$ are both less than $f(p)$.

Question 3.2. Let $[a, b] \subseteq \mathbb{R}$ be a bounded closed interval. Suppose $f : [a, b] \to [a, b]$ is an increasing function, then there exists $x \in [a, b]$ such that $f(x) = x$. *(Note, f is not assumed to be continuous, so you cannot use the intermediate value theorem. Instead, try to use the least upper bound property.)*

Question 3.3. *(To emphasize the fact that we are not using continuity: when f is continuous, it doesn't matter whether f is increasing or not. But without continuity it matters a lot.)* Given an example of a decreasing function $f : [a, b] \to [a, b]$ such that $f(x) \neq x$ for any x.

Question 3.4. Let X be the non-negative real numbers. Suppose $f : X \to \mathbb{R}$ satisfies $\sup f([0, x)) \geq f(x)$. Must $f(X)$ be bounded above? Why or why not?

Question 3.5. Let X be the non-negative real number, and $f : X \to \mathbb{R}$. Suppose for every $x \in \mathbb{R}$, there exists a $y > x$ such that the restriction of f to $[x, y]$ is a decreasing function. Must $f(X)$ be bounded above? Why or why not?

Question 3.6. If $\{[a_1, b_1], \ldots, [a_n, b_n]\}$ is a finite family of closed intervals, prove that their union (with the restricted ordering) is Dedekind complete. What if the intervals are open instead?

Problem Set 3

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 3.1. Let S and T be bounded, non-empty subsets of $\mathbb{R}$. Denote by

$$U = \{x \in \mathbb{R} : x = s - t,\ s \in S,\ t \in T\}.$$

Prove that:

1. $\sup U = \sup S - \inf T$.
2. $\max U$ exists if and only if *both* $\max S$ and $\min T$ exists.

Problem 3.2. Suppose $K \subseteq \mathbb{R}$ is a closed, bounded set (not necessarily a simple interval). Let $\mathscr{S}$ be a collection of *open* subsets of $\mathbb{R}$ that covers K. Prove that there is a finite subset $\mathscr{D} \subseteq \mathscr{S}$ that also covers K.

(Hint: you may use the statement of the final exercise of the lecture notes.)

Problem 3.3. Denote by X the set of non-negative real numbers. Suppose $f : X \to \mathbb{R}$ is a function that satisfies the following three properties:

1. $f(0) = 0$,
2. $\sup f([0, x)) \geq f(x)$,
3. if $f(x) \leq 1$ then there exists $y > x$ such that $f(z) < 1$ for all $z \in (x, y)$.

Prove that $\sup f(X) \leq 1$. *(Hint: use the continuity argument.)*

Problem 3.4. Consider the set M of functions with domain $\mathbb{N}$ and codomain $\mathbb{Z}$. Equip M with the ordering $\leqq$ where $m_1 \lneqq m_2 \iff m_1 \neq m_2$ and the first non-zero output of the function $m_2 - m_1$ is positive.

1. Check that $(M, \leqq)$ is a total order.
2. Prove that $(M, \leqq)$ is *not* Dedekind complete.

Reading Assignment 4

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We begin by defining the notion of directed sets, and via it, the notion of nets. Nets are a generalization of the sequences that you may have encountered in a previous calculus course, and is of sufficient generality (as opposed to sequences) to study convergence properties in general topological settings. The behavior of nets "in the limit" can be captured by the description of whether it is "eventually" in a set, "frequently" in a set, or "infrequently" in a set; these are discussed in detail. Using these notions, we next define limits and accumulation points of nets. Basic properties of limits and accumulation points are discussed, and a connection is drawn from these notions to the idea of open and closed sets, as described in the previous reading. We next discuss the Monotone Convergence Theorem, and the idea of Cauchy nets, showing that for real-valued nets Cauchy and converging are equivalent notions. We finish by defining the limits superior and inferior, discussing their properties, and proving the Bolzano-Weierstrass Theorem.

Contents

Having constructed the real numbers, now let's talk about its structure. In this set of notes you will learn about the concept of "nets" and their convergence. Before we start, let us introduce some useful notations:

Definition 4.1. *Let A, B be subsets of $\mathbb{R}$. We write*

- $A + B = \{x \in \mathbb{R} : x = a + b, a \in A, b \in B\}$;
- *for $c \in \mathbb{R}$, the set $cA = \{x \in \mathbb{R} : x = ca, a \in A\}$.*

Observe that for intervals:

$$(a, b) + (c, d) = (a + c, b + d) \tag{4.1}$$

$$c(a, b) = \begin{cases} (ca, cb) & c > 0 \\ (cb, ca) & c < 0 \\ \{0\} & c = 0 \end{cases} \tag{4.2}$$

§4.1 Nets

Nets generalize the notion of sequences. Unfortunately the name is not very suggestive of its features (one of many instances where the nomenclature could've been improved way back when). In this section we will define nets and examine some basic properties.

To get us started, we first loosen the notion of a poset a little bit. This makes some of the arguments simpler later on.

Definition 4.2. *By a* directed set *we refer to a pair* $(X, \leqq)$*, where* X *is a set,* $\leqq$ *is a relation on* X*, such that*

- $\leqq$ *is reflexive;*
- $\leqq$ *is transitive;*
- $\leqq$ *is upward directed: given* $x, y \in X$ *there exists* $z \in X$ *such that* $x \leqq z$ *and* $y \leqq z$*.*

Note that by convention, when we speak of a directed set without specifying the direction, it is assumed to be upward directed. You may wish to compare this to Definitions 2.1 and 2.14. The difference between a directed set and an upward directed poset is that in a poset we assume the relation is antisymmetric. While for the most part, the directed sets that will make appearances in this course are posets, there are a few instances where antisymmetry doesn't necessary hold, hence we leave open the more general possibility.

The distinguishing feature of being "upward directed", has the implication that for any x, y in a directed set $(X, \leqq)$, the principal upper sets $\uparrow(x)$ and $\uparrow(y)$ have non-trivial intersection.

Example 4.3. The universal ordering $\leqq = X^2$ is directed. Since any pair of elements are comparable, the three properties are trivially satisfied. ■

Example 4.4. If $(X, \leqq_X)$ and $(Y, \leqq_Y)$ are directed sets, the relation $\leqq_Z$ on $Z = X \times Y$ defined by

$$(x_1, y_1) \leqq_Z (x_2, y_2) \iff x_1 \leqq_X x_2 \text{ and } y_1 \leqq_Y y_2$$

(which we call the "product ordering") makes $(Z, \leqq_Z)$ another directed set. ■

Exercise 4.1. Subsets of a directed set, with the restricted ordering, is not always a directed set. Give an example.

The previous exercise notwithstanding, *some* subsets of a directed set are still directed.

Lemma 4.5. *Let* $(X, \leqq)$ *be a directed set. Let* $\alpha \in X$*. Then* $\uparrow(\alpha)$*, equipped with the restricted ordering, is a directed set.*

Proof. Reflexivity and transitivity are always preserved under restrictions. So we only need to check whether the restriction is still directed. Suppose $x, y \in \uparrow(\alpha)$, then we know that by the directedness of X there exists $z \in X$ that succeeds both x and y. But by transitivity since both x, y succeeds α, we see that $z \in \uparrow(\alpha)$ also. Hence $\uparrow(\alpha)$ is directed. □

Example 4.6. Any total order is directed. So $\mathbb{R}, \mathbb{Q}, \mathbb{N}, \mathbb{Z}$ are all directed sets. Any upward directed poset is a directed set. ■

Example 4.7. Let $x \in \mathbb{R}$ and let $\mathbb{I}_x$ denote the set of all open intervals of $\mathbb{R}$ that contain the point x. Equip $\mathbb{I}_x$ with the ordering $I_1 \leqq I_2 \iff I_1 \supseteq I_2$ (that is, the inverse ordering of the subset order on $2^{\mathbb{R}}$). Since $\leqq$ is the inverse of the restriction of a partial ordering, it is also a partial order. It is also

directed: if $I_1, I_2 \in \mathbb{I}_x$, then their intersection $I_1 \cap I_2$ is also an open interval, it contains the point x, and is included as a subset of both I_1 and I_2. ■

Exercise 4.2. Consider the set $\mathbb{I}_x$ from the previous example. Prove that it has no maximum. (*Hint: use some version of the Archimedean property.*)

Definition 4.8 (Nets)**.** *A* net *(otherwise known as a* Moore-Smith sequence*) in a set X is a function from a directed set to X. More precisely, a* net *x with values in X indexed by the directed set $(A, \leqq)$ is a function $x : A \to X$.*

Notationally, instead of $x(\alpha)$, frequently the element in X corresponding to $\alpha \in A$ through the function x is denoted by x_α.

Example 4.9. A sequence of real numbers $x_1, x_2, x_3, \ldots$ is a special case of a net taking values in $\mathbb{R}$. The directed set that forms the domain of the sequence is $\mathbb{N}$. ■

In view of the previous example, we should think of nets as a generalization of the notion of sequences. The key property that is preserved under the generalization is the "sense of direction". In a sequence we can move from an "earlier term" to a "later term". The same is available in a net by moving from x_α to x_β with $\beta \geqq \alpha$.

Example 4.10. Returning to Example 4.7, since $\mathbb{I}_x$ has as its elements intervals, any choice function $c : \mathbb{I}_x \to \mathbb{R}$ (recall that a choice function would satisfy $c(I) \in I$) is a net in the reals.

One such net is the mapping $c_I = x$ for all I. This is a constant net.

By our definition, we can also find a net such that $c_I \neq x$ for any I. Suppose that I is an open interval containing x, then $I = (a, b)$ for $a < x < b$. By the Archimedean property of $\mathbb{R}$, there exists $r \in (a, x)$. (In other words, for any open interval I containing x, the subset $I \setminus \{x\}$ is non-empty, and hence we can choose a point from it.) ■

Here are some more terminology used to discuss nets.

Definition 4.11. *Given $x : A \to X$ a net, we denote by*

$$x_{\uparrow(\alpha_0)} \coloneqq \{x_\alpha : \alpha \geqq \alpha_0\}.$$

Sets of this form are called tail sets *of x.*

Definition 4.12. *Given $x : A \to X$ a net, and $S \subseteq X$:*

- *We say that x is* eventually in *S if there exists $\alpha \in A$ such that $x_{\uparrow(\alpha)} \subseteq S$.*
- *We say that x is* frequently in *S if for every $\alpha \in A$, the intersection $x_{\uparrow(\alpha)} \cap S \neq \emptyset$.*
- *We say that x is* infrequently in *S if there exists $\alpha \in A$ such that $x_{\uparrow(\alpha)}$ is disjoint from S.*

Notice that by definition, if the net x is eventually (or frequently) in a set S, it is also eventually (or frequently) is any superset of S; dually, if the net is infrequently in a set S, it is also infrequently in any subset of S.

Example 4.13. Continuing from 4.10, and consider $c : \mathbb{I}_x \to \mathbb{R}$ one such net.

Let J_1 be any open interval that contains x. Then we have that c is eventually in J_1: since $J_1 \in \mathbb{I}_x$ we see that $c_{\uparrow(J_1)} \subseteq J_1$ by definition. (If $I \geqq J_1$, this means $I \subseteq J_1$, and hence $c_I \in J_1$.)

Let $J_2 = [x + 1, x + 2]$, then c is infrequently in J_2. This is because $I_2 = (x - 1, x + 1) \in \mathbb{I}_x$, and for any

$I \geqq I_2$, the element $c_I \in I_2$ and hence cannot be in J_2.

Now suppose further that we know c is such that $c_I \neq x$ for any I. I claim that c is frequently in at least one of $(x-1,x)$ and $(x,x+1)$. Suppose not, this means that there exists I_3, I_4 such that $c_{\uparrow(I_3)}$ is disjoint from $(x,x+1)$, and $c_{\uparrow(I_4)}$ is disjoint from $(x-1,x)$. But as $\mathbb{I}_x$ is directed, there is I_5 that follows both I_3 and I_4. So we have $c_{\uparrow(I_5)}$ is disjoint from both $(x-1,x)$ and $(x,x+1)$. Finally, let $I_6 = I_5 \cap (x-1,x+1)$. Then $c_{\uparrow(I_6)} \subseteq I_6 \subseteq (x-1,x+1)$, but is disjoint from both $(x-1,x)$ and $(x,x+1)$. This means that $c_{\uparrow(I_6)} = \{x\}$, which contradicts our initial supposition. ∎

Exercise 4.3. Let X be a set and $x : A \to X$ be a non-empty net. Fix a subset S. Prove that:

1. The net x is either frequently in S or infrequently in S.
2. The net x is eventually in S if and only if x is infrequently in $X \setminus S$.
3. If x is eventually in S, then it is frequently in S.

Exercise 4.4 (A sort of Pigeonhole Principle)**.** Let $x : A \to X$ be a net. Given a *finite* collection $\mathscr{S}$ of subsets of X, prove that if x is frequently in $\cup\mathscr{S}$, then there exists $S \in \mathscr{S}$ such that x is frequently in S. (Does the same statement hold if "frequently" is replaced by "eventually"?)

Example 4.14. Even though nets have "a sense of direction", one has to be careful about generalizing sequence-based reasoning to nets.

For sequences: suppose x is a sequence $\mathbb{N} \to X$, and let $S \subseteq X$ be a set such that x is eventually in S. Our naive expectation holds: if we start with an arbitrary x_i, and keep moving by increasing the index, then "eventually" we will end up inside S.

For nets this doesn't hold: it is possible to keep moving by increasing the index yet never reach inside an eventual set. One example is this: consider $\mathbb{N}$ with the "divisibility partial order" described in Problem Set 2. ($n \leqq m$ if and only if n divides m.) You proved on Problem Set 2 that any finite set has a supremum, which implies that this poset is directed. Using this directed set as the index set, we can define the net with values in $\mathbb{N}$ by the identity function $i : \mathbb{N} \to \mathbb{N}$.

Let S be the set of all numbers divisible by 17. Then i is eventually in S. This is because by definition $i_{\uparrow(17)} = S$. However, it is possible to move around, increasing the index the whole time, while avoiding S. For example, consider the sequence of values $1, 2, 4, 8, 16, \ldots, 2^k, \ldots$. This sequence moves in increasing order within $\mathbb{N}$ with the divisibility partial order, but it never intersects $\uparrow(17)$. ∎

In view of the previous example, perhaps it is useful to interpret the notion of "frequently" and "eventually". The net x being "frequently" in a set S means that starting from any point in the net x, if we only move in the direction of increasing the index, there's always the opportunity of moving into S. The net x being "eventually" in a set S requires additionally that there is a point of no-return: at some point once you enter S you cannot leave S.

§4.2 Limits of Real-valued Nets

In this section we will discuss the notion of convergence for real-valued nets. Much of what we discuss can easily be generalizable to arbitrary topological spaces, but we will restrict to the real numbers for simplicity and clarity.

Definition 4.15. *A real-valued net x is said to* converge *to the real number z if for every open interval I containing z, the net x is eventually in I. When x converges to z we say that z is* the limit of *x and write $z = \lim x$.*

Notice that since the definition of the net is as a function, and a function's definition includes its domain, and the domain is a directed set for a net, it is *not necessary* to specify, as you may have learned in you calculus course, that you are taking $\lim_{n\to\infty} x_n$ by indicating the index and where it tends to.

Later, however, when we discuss functions and their continuity, the subscript notation for limits will again be useful.

The use of the article "the" is intentional in the definition above.[1]

Proposition 4.16. *A real-valued net x can have at most one limit.*

We will defer the proof of this proposition; it follows as an immediately corollary of Proposition 4.19 below. This proposition should be contrasted against Example 4.14 above.

Example 4.17. Consider the net (sequence) $x : \mathbb{N} \to \mathbb{R}$ where $\mathbb{N}$ is given the usual order structure. Let $x_n = \frac{1}{n}$. Then $\lim x = 0$.

This is because for any open interval (a, b) with $a < 0 < b$, by the Archimedean property we can find m such that $\frac{1}{m} < b$. Thus $x_{\uparrow(m)} \subseteq (0, \frac{1}{m}] \subseteq (a, b)$ showing that x must be eventually in (a, b). Since this holds for a, b arbitrary, this shows that x is eventually in any open interval containing 0 and thus converges to 0. ■

Exercise 4.5. Consider the function $x : \mathbb{N} \to \mathbb{R}$ where $x_n = \frac{1}{n}$. This time, however, $\mathbb{N}$ is equipped with the divisibility partial order as described in Example 4.14. Prove that $\lim x = 0$.

Exercise 4.6. Consider a net c such as described in Example 4.10. Prove that $\lim c = x$.

Definition 4.18. *A real-valued net x is said to* accumulate *(or* cluster*) at the real number z if for every open interval I containing z, the net x is frequently in I. When x accumulates at z we say that z is* an accumulation point of x.

Proposition 4.19. *If a real-valued net x converges to z, then z is its unique accumulation point.*

Proof. That z is an accumulation point follows from Exercise 4.3, part 3. We will focus on proving its uniqueness.

Suppose $z' \neq z$. Then either $z' > z$ or $z' < z$. By the Archimedean property of the real numbers we can find two distinct reals between z and z', and hence two distinct open intervals I and I', satisfying $I \ni z$, $I' \ni z'$, and $I \cap I' = \emptyset$. Since x converges to z, we see that x is eventually in I, and so by Exercise 4.3, part 2, it is infrequently in $\mathbb{R} \setminus I$. As I' is a subset of $\mathbb{R} \setminus I$, this means that x is also infrequently in I', showing that z' cannot be an accumulation point. □

Food for Thought 4.7. Is it the case that nets with a unique accumulation point must converge to it?

Example 4.20. Consider the net (sequence) $x : \mathbb{N} \to \mathbb{R}$ where $x_n = (-1)^n$, where $\mathbb{N}$ has the standard ordering. Then x does not converge. We will show that this is the case by showing that, for every $z \in \mathbb{R}$, there is an open interval $I_z \ni z$ such that x is frequently in $\mathbb{R} \setminus I_z$.

[1] This proposition is not entirely trivial. In a course in point-set topology you will learn that it is possible, for general topological spaces, to have nets with multiple limits. Most spaces that you will run into in *analysis* however will be nice enough that a version of the proposition holds.

Let $I_z = (z - 1/2, z + 1/2)$. Since $1 = (-1) + 2$, at most one of 1 and -1 can belong to I_z. Since x is frequently in both $\{1\}$ and $\{-1\}$, we see that x must be frequently in $\mathbb{R} \setminus I_z$.

In fact, we have that both 1 and -1 are accumulation points of x: let I be any open interval around 1, let $n \in \mathbb{N}$, then either n is even in which case $x_n \in I$, or n is odd in which case $x_{n+1} \in I$, showing that x is frequently in I. A similar argument shows that x is also frequently in any open interval containing -1. ■

Example 4.21. Consider the net $x : \mathbb{N} \to \mathbb{R}$ where $x_n = (-1)^n$, and where $\mathbb{N}$ is equipped with the divisibility partial order as in Example 4.14. I claim that this net converges. (Compare this to the previous example.)

Consider the set $x_{\uparrow(2)}$. Since by definition all element of $\uparrow(2)$ are even numbers, this means that the set $x_{\uparrow(2)} = \{1\}$. Thus this shows that for any open interval I containing 1, we have x is eventually in I. This shows that $\lim x = 1$. ■

Exercise 4.8. Below several functions $x : \mathbb{N} \to \mathbb{R}$ are given. For each function, consider (A) the net (sequence) where the index set $\mathbb{N}$ is equipped with the standard ordering, and (B) the net where the index set $\mathbb{N}$ is equipped with the divisibility partial order as in Example 4.14. For each of the two cases, determine whether the net converges, its limit if it does, and its set of accumulation points if it doesn't.

1. Primality testing: $x(n) = 1$ if n is prime, $x(n) = 0$ if n is composite.
2. Prime counting: $x(n) =$ the number of distinct prime factors (counted without multiplicity) of n.
3. $x(n) = \cos(n\pi/6)$.
4. $x(n) = n/2$ if n is even, and $x(n) = 0$ if n is odd.

§4.2.1 Arithmetic of limits.—The process of taking limits plays well with the arithmetics of the real numbers.

Proposition 4.22. *If x, y are real valued nets with the same domain and $\lim x = \bar{x}$ and $\lim y = \bar{y}$, then*

$$\lim(x + y) = \bar{x} + \bar{y}, \qquad \textit{and} \quad \lim(x \cdot y) = \bar{x} \cdot \bar{y}.$$

Proof. Sum: let $(a, b) \ni \bar{x} + \bar{y}$. Choose a_x, a_y such that $a_x + a_y = a$, $a_x < \bar{x}$, $a_y < \bar{y}$. Choose b_x, b_y such that $b_x + b_y = b$, $b_x > \bar{x}$, $b_y > \bar{y}$. Since x converges to $\bar{x}$, there exists α_x such that $x_{\uparrow(\alpha_x)} \subseteq (a_x, b_x)$. Similarly there exists α_y such that $y_{\uparrow(\alpha_y)} \subseteq (a_y, b_y)$. Since the index set is directed, choose α that succeeds both α_x and α_y. We then have $(x + y)_{\uparrow(\alpha)} \subseteq (a, b)$, showing convergence.

Product: Let $(a, b) \ni \bar{x} \cdot \bar{y}$. Choose

$$w_x = \frac{1}{3(|\bar{y}| + 1)} \min\{\bar{x} + \bar{y} - a, b - \bar{x} - \bar{y}, 1\},$$
$$w_y = \frac{1}{3(|\bar{x}| + 1)} \min\{\bar{x} + \bar{y} - a, b - \bar{x} - \bar{y}, 1\}.$$

Observe that if $z_x \in (\bar{x} - w_x, \bar{x} + w_x)$ and $z_y \in (\bar{y} - w_y, \bar{y} + w_y)$, then using

$$z_x \cdot z_y - \bar{x} \cdot \bar{y} = (z_x - \bar{x}) \cdot (z_y - \bar{y}) + (z_x - \bar{x}) \cdot \bar{y} + (z_y - \bar{y}) \cdot \bar{x}$$

and the triangle inequality, we obtain

$$|z_x \cdot z_y - \bar{x} \cdot \bar{y}| \leq |w_x| \cdot |w_y| + |w_x| \cdot |\bar{y}| + |w_y| \cdot |\bar{x}| \leq \frac{7}{9} \min\{\bar{x} + \bar{y} - a, b - \bar{x} - \bar{y}\}.$$

This implies that

$$z_x \in (\bar{x} - w_x, \bar{x} + w_x), z_y \in (\bar{y} - w_y, \bar{y} + w_y) \implies z_x \cdot z_y \in (a, b).$$

Using that $\lim x = \bar{x}$ we can find α_x such that $x_{\uparrow(\alpha_x)} \subseteq (\bar{x} - w_x, \bar{x} + w_x)$; similarly we can find α_y such that $y_{\uparrow(\alpha_y)} \subseteq (\bar{y} - w_y, \bar{y} + w_y)$. Using that the index set is directed, we can find α succeeding both α_x and α_y. For this α we see that $(x \cdot y)_{\uparrow(\alpha)} \subseteq (a, b)$. Since the choice of interval is arbitrary, this shows that $x \cdot y$ converges to $\bar{x} \cdot \bar{y}$. □

Exercise 4.9. Given x a real-valued net whose range doesn't contain 0. Suppose that x converges and $\lim x \neq 0$, prove that $\lim \frac{1}{x} = \frac{1}{\lim x}$.

Similar statements do *not* hold for accumulation points.

Example 4.23. Let x be the real-valued sequence $x_n = n$, and y the sequence $y_n = -n$. The sets of accumulation points of x and y are both empty. But their sum $x+y$ is the constant sequence $(x+y)_n = 0$, and hence converges to 0. This shows that it is *not* the case that the accumulation points of $x + y$ must be a subset of the sum of those of x against those of y.

If we let x be the sequence where $x_n = 0$ if n is even and $x_n = n$ if n is odd; and let y be such that $y_n = 0$ if n is odd and $y_n = n$ if n is even. Both of these sequences claim 0 as a accumulation point. But the sequence $x + y$ has no accumulation points. This shows that it is *not* the case that the accumulation points x and y can always be combined to get one of $x + y$. ■

Exercise 4.10. Given x, y real-valued nets with the same index set. Suppose x converges to z, and w is an accumulation point of y. Prove that $z + w$ is an accumulation point of $x + y$.

§4.2.2 Open and closed sets.—It turns out that the concepts of open and closed sets, introduced in Definition 3.24, can be characterized in terms of limits.

Theorem 4.24. *A set $A \subseteq \mathbb{R}$ is open if and only if every real-valued net x with an accumulation point in A, is frequently in A.*

Proof. For the forward direction, suppose $z \in A$ is an accumulation point of x. Since A is supposed open, there exists an open interval $I \subseteq A$ containing z. By definition of accumulation points, this means that x is frequently in I and hence frequently in A.

For the reverse direction, we prove the contrapositive: suppose that A is not open, we will construct a net x that converges to some point in A that is not frequently in A. Since A is not open, there exists $z \in A$ such that no open interval containing z is contained in A. Construct the directed set $\mathbb{I}_z$ as in Example 4.7; this means that for every $I \in \mathbb{I}_z$ there exists $y \in I$ such that $y \notin A$. So the choice function c that makes such an assignment is a net (see Example 4.10) that converges to z (see Exercise 4.6). By construction the range of c is disjoint from A, and hence is infrequently in A. □

Corollary 4.25. *A set $A \subseteq \mathbb{R}$ is closed if and only if for every real-valued net x taking values in A has all of its accumulation points contained in A.*

Proof. For the forward direction we will argue its contrapositive: suppose a real-valued net x taking values in A has an accumulation point outside A, then the complement of A is such that x accumulates at a point of $\mathbb{R} \setminus A$, but the range of x is disjoint from $\mathbb{R} \setminus A$, and hence in particular x is infrequently in $\mathbb{R} \setminus A$. Hence by Theorem 4.24 $\mathbb{R} \setminus A$ cannot be open, and hence A cannot be closed.

For the reverse direction, by Theorem 4.24, it suffices to show that any net y with accumulation point in $\mathbb{R} \setminus A$ must be frequently in $\mathbb{R} \setminus A$. Suppose for contradiction that y has an accumulation point w in $\mathbb{R} \setminus A$ and is infrequently in $\mathbb{R} \setminus A$. This means that y is eventually in A, and hence there exists α such that $y_{\uparrow(\alpha)} \subseteq A$. By Lemma 4.5 the restriction y to $\uparrow(\alpha)$, which we will call y', is still a real-valued net. And we just showed that y' takes values in A.

The point w, however, must still be an accumulation point of y': let I be any open interval containing w, since y accumulates at w we know that for every β, $y_{\uparrow(\beta)} \cap I$ is non-empty. Suppose $\beta \in \uparrow(\alpha)$, then by transitivity we have $y'_{\uparrow(\beta)} = y_{\uparrow(\beta)}$. Thus we conclude that $y'_{\uparrow(\beta)} \cap I$ is also non-empty. Since this holds for every $\beta \in \uparrow(\alpha)$, and any interval I, we conclude that w is an accumulation point of y. But then we obtained our contradiction, since we assumed that A has the property "every real-valued net taking values in A has all its accumulation points contained in A". □

§4.3 Some convergence theorems

Definition 4.15 of convergence is a bit unwieldy, since to check that a net is converging, it requires us to correctly guess what the limit is. There are however some ways in which convergence can be inferred even when the limiting value is unknown. The first is the Monotone Convergence Theorem.

Similar to Definition 2.6 we shall say a net $x : A \to X$ where $(A, \leqq_A)$ is a directed set and $(X, \leqq_X)$ is a poset is *increasing* if $\alpha \leqq_A \beta \implies x_\alpha \leqq_X x_\beta$; *decreasing* if $\alpha \leqq_A \beta \implies x_\alpha \geqq_X x_\beta$; and *monotone* if it is either increasing or decreasing.

We also say that a net $x : A \to X$ is bounded (above/below) if its range in X is bounded (above/below).

Theorem 4.26 (Monotone Convergence)**.** *Let x be a non-empty real-valued net.*

- *If x is increasing and bounded above, then x converges to the supremum of its range.*
- *If x is decreasing and bounded below, then x converges to the infimum of its range.*

Proof. We shall only prove the increasing case. The decreasing case is similar.

Let z be the supremum of the range of x. Let (a, b) be an open interval about z. By the definition of supremum, since $a < z$ there exists α such that $a < x_\alpha \leq z$. Since x is increasing, whenever $\beta \geqq \alpha$ we must have $x_\alpha \leq x_\beta \leq z$, the second inequality is due to z being an upper bound of the range of x. Hence we've proven that $x_{\uparrow(\alpha)} \subseteq (a, b)$. Since (a, b) is arbitrary this shows that z is the limit of x. □

The next notion depends on some preliminary definitions.

Definition 4.27. *Given a bounded, non-empty, interval $I \subseteq \mathbb{R}$, its* width*, which we denote by $|I|$, is the real number $|I| := \sup I - \inf I$. Necessarily $|I| \geq 0$.*

Definition 4.28. *A real-valued net x is a* Cauchy net *if for any real number $w > 0$, there exists an open interval I with width $0 < |I| \leq w$ such that x is eventually in I.*

It is certainly worth comparing and contrasting Definitions 4.15 and 4.28. What the notion of Cauchy nets gain over the definition of convergent nets is that in defining a convergent net, we need to know a priori what the limit z should be, and that all the intervals involved should contain the point z; when defining Cauchy nets, the knowledge of the limit is no longer required, however, it is replaced by an additional condition concerning the *width* of the intervals.

The definition of a convergent net essentially says that there is a location where open intervals of any width, when placed there eventually contains the net. The definition of a Cauchy net essentially says that for any width, there is a location where an open interval of that width eventually contains the net.[2]

As you have already learned in previous math courses, generally swapping $\exists$ and $\forall$ can have significant impact on the meaning of mathematical statements. In this case, due to the completeness of the real numbers, it turns out the swapping doesn't change the meaning.

Theorem 4.29 (Cauchy's Criterion)**.** *A real-valued net x is convergent if and only if it is Cauchy.*

Proof. The forward implication, which goes from "$\exists\ldots\forall\ldots$" to "$\forall\ldots\exists\ldots$" is, as one expects, easier to prove. Let $w > 0$ be given. Then there exists an open interval I around the limit $z = \lim x$ such that $|I| \leq w$; to be concrete we can take $I = (z - w/2, z + w/2)$. Since x converges to z, we see that x is eventually in I. Since w is arbitrary, we've shown that x is Cauchy.

The reverse implication is more involved, and in this step requires the completeness of $\mathbb{R}$. We shall use the form in terms of Cantor's theorem (Exercise 3.7). We construct a family of non-empty *closed* intervals $I_j = [a_j, b_j]$ as follows:

1. Let (a_1, b_1) be the *open* interval of width $\leq \frac{1}{2}$ guaranteed by the Cauchy property. Let α_1 be such that $x_{\uparrow(\alpha_1)} \subseteq (a_1, b_1)$.
2. Recursively construct (a_{j+1}, b_{j+1}) from (a_j, b_j) and α_j: By the Cauchy property there exists (a_{j+1}, b_{j+1}) with width $\leq 2^{-j+1}$ such that x is eventually in (a_{j+1}, b_{j+1}). Since the index set is directed, we can choose $\alpha_{j+1} \geqq \alpha_j$ such that $x_{\uparrow(\alpha_{j+1})} \subseteq (a_{j+1}, b_{j+1})$; and by possibly replacing (a_{j+1}, b_{j+1}) with its intersection with (a_j, b_j) (which is non-empty as it contains $x_{\uparrow(\alpha_{j+1})}$; note that taking the intersection makes the interval shorter and does not violate the width condition), we can require $(a_{j+1}, b_{j+1}) \subseteq (a_j, b_j)$.

The corresponding family of closed non-empty intervals I_j are nested by the above construction, and therefore their intersection is non-empty by Cantor's theorem. I claim that if $z \in \cap_{j=1}^{\infty} I_j$, then $z = \lim x$.

Let (c, d) be an open interval around z. Let $w = \min\{z - c, d - z\} > 0$. By the Archimedean property[3] there exists j such that $2^{-j} \leq w$. Since the interval (a_j, b_j) contains z, and has width $\leq 2^{-j}$, we have that $(a_j, b_j) \subseteq (c, d)$. Since x is eventually in (a_j, b_j), we also have that x is eventually in (c, d). □

Exercise 4.11. The proof above relied on certain properties of of intervals that we have not explicitly justified. Please prove them here.

1. If two open intervals have non-trivial intersection, their intersection is an open interval.
2. If $c \in (a, b)$, then $(a, b) \subseteq (c + a - b, c + b - a)$.

Exercise 4.12. Prove that the real-valued net x being a Cauchy net is equivalent to: "For every $\delta > 0$, there exists α such that whenever $y, y' \in x_{\uparrow(\alpha)}$, $|y - y'| < \delta$." (This formulation is somewhat easier to wield in practice.)

[2] The definition of Cauchy net requires additional structure; to generalize the notion of a convergent net we just need to suitably replace "open intervals" by some collection of subsets of an abstract set X; this leads to general point-set topology. To generalize Cauchy nets we need to be able to "slide" those "open intervals" around. This requires being about to compare elements of this collection of subsets, and lead to the study of "uniform spaces".

[3] The completeness of $\mathbb{R}$ is used again here.

§4.4 Limit superior and limit inferior

In this section we will define the useful notion of limit superior and limit inferior, and discuss some of their properties.

Definition 4.30. *A real-valued net x is said to be* eventually bounded (above/below) *if there exists α_0 such that $x_{\uparrow(\alpha_0)}$ is bounded (above/below).*

Let x be a non-empty real-valued net. If x is eventually bounded above, we can define the net U, with the index set $\uparrow(\alpha_0)$ (where α_0 is given in the above definition), by

$$U_\alpha = \sup x_{\uparrow(\alpha)}. \tag{4.3}$$

Similarly, if x is eventually bounded below, we can define the net L by

$$L_\alpha = \inf x_{\uparrow(\alpha)}. \tag{4.4}$$

Notice that the net U is decreasing: if $\beta \geqq \alpha$, then $x_{\uparrow(\beta)} \subseteq x_{\uparrow(\alpha)}$, and hence $U_\beta \leq U_\alpha$. Similarly the net L is increasing.

By Theorem 4.26, this means that U and L both converge when they are defined.

Definition 4.31. *Let x be a non-empty real-valued net.*

- *If x is eventually bounded above, then its limit superior is defined as*

  $$\limsup x := \lim U$$

  where U is defined as in (4.3).
- *If x is eventually bounded below, then its limit inferior is defined as*

  $$\liminf x := \lim L$$

  where L is defined as in (4.4).

Exercise 4.13. For each of the nets described in Exercise 4.8, determine whether its limits superior and inferior exist, and find their values.

Exercise 4.14. Since $\mathbb{R}$ with its standard $\leq$ is totally ordered, it is also a directed set. For each of the following function $x : \mathbb{R} \to \mathbb{R}$, interpreted as a net in $\mathbb{R}$, find its lim sup and lim inf. (The answers are supposed to be obvious: try to give detailed proofs in each case as a practice; in particular, try to describe as well as you can the corresponding nets U and L.)

1. $x(s) = \cos(s) + \tan^{-1}(s)$.
2. $x(s) = s^3 \cos^2(s)$.
3. $x(s) = \sin(s) \ln(|s| + 1)$

Exercise 4.15. Suppose x and y are both real-valued nets, with the same index set, such that $y - x$ is eventually non-negative. Prove that:

1. If x is eventually bounded below, then so is y, and $\liminf x \leq \liminf y$.
2. If y is eventually bounded above, then so is x, and $\limsup x \leq \limsup y$.

Exercise 4.16. Prove that if x and y are both real-valued nets, with the same index set, that are eventually bounded above, then $\limsup(x+y) \leq \limsup x + \limsup y$. Show via example that it is possible for the inequality to be strict.

Proposition 4.32. *Given a real-valued net x, its limit superior and limit inferior, when they exist, are the largest and smallest (respectively) accumulation points of x.*

Proof. We demonstrate for the limit superior; the inferior case is similar.

Suppose x is eventually bounded above and set $z = \limsup x$. By definition for every open interval I containing z, the net U is eventually in I. This means that there exists α such that $U_{\uparrow(\alpha)} \subseteq I$. Fix $\beta \geqq \alpha$, since I is open, there exists an open interval $I_\beta \subsetneqq I$ such that $I_\beta \ni U_\beta$.

Since $U_\beta = \sup x_{\uparrow(\beta)}$, we find that for $\inf I_\beta < U_\beta$, there exists an element $y \in x_{\uparrow(\beta)}$ such that $\inf I_\beta < y \leq U_\beta$, and hence $y \in I_\beta \subseteq I$. Chaining everything together we find that this implies for every $\beta \succeq \alpha$, the set $x_{\uparrow(\beta)} \cap I \neq \emptyset$.

Now, let γ be an arbitrary index. Since the index set is directed, there exists $\beta' \in \uparrow(\alpha) \cap \uparrow(\gamma)$. Since $x_{\uparrow(\beta')} \cap I \neq \emptyset$, this means $x_{\uparrow(\gamma)} \cap I \neq \emptyset$. And hence we have shown that every open interval I containing z is frequented by x, showing that z is an accumulation point of x.

It remains to show that there can be no larger accumulation point of x. Suppose $w > \limsup x$, we will show that it cannot be an accumulation point of x. Since $w > \limsup x$, there exists disjoint open intervals $I_w \ni w$ and $I_x \ni \limsup x$. By definition the net U is eventually in I_x. This implies that there is an α such that $U_\alpha \in I_x$; by definition this means that $\sup x_{\uparrow(\alpha)} \leq U_\alpha < \sup I_x \leq \inf I_w$, and hence $x_{\uparrow(\alpha)}$ is disjoint from I_w, showing that x is infrequently in I_w and that w is not an accumulation point of x. □

Exercise 4.17. For each of the nets in Exercise 4.14, find the set of all accumulation points, and verify that the Proposition above indeed holds for them.

An immediate Corollary of the above proposition is the Bolzano-Weierstrass Theorem.

Theorem 4.33 (Bolzano-Weierstrass)**.** *Every non-empty bounded real-valued net x has an accumulation point.*

Proof via lim sup *and* lim inf. If x is bounded, by definition $\limsup x$ and $\liminf x$ exist. By Proposition 4.32, they are accumulation points. □

I will include here a second proof of the Bolzano-Weierstrass Theorem. The above prove makes use of the limits superior and inferior, which are concepts only well-defined when the set X has an order structure. For more general X (such as the case in general topological spaces), the following proof generalizes more readily.

Proof via Heine-Borel. Let the real-valued net x be bounded, this mean that there exists some numbers a, b such that the range of x is contained in the closed interval $[a, b]$.

We shall argue by contradiction. Suppose x has no accumulation points, this means that for every $p \in \mathbb{R}$ there exists an open interval $I_p \ni p$, such that x is infrequently in I_p. The assignment $p \mapsto I_p$ is an entourage mapping (see Definition 3.10). By Theorem 3.12, $\mathbb{R}$ has the Heine-Borel property, and hence there exists a finite subset $S \subseteq [a, b]$ such that $\cup\{I_p : p \in S\}$ covers $[a, b]$.

Now we have a contradiction: since x only takes values in $[a,b]$, we have that x is by hypothesis eventually in $\cup\{I_p : p \in S\}$; but by construction x is also infrequently in any of the I_p for $p \in S$. This contradicts Exercise 4.4. □

We note that by Corollary 4.25, if x takes values in some bounded, closed subset $K \subseteq \mathbb{R}$, the accumulation points found via the Bolzano-Weierstrass theorem will also be in K.

Theorem 4.34. *A real-valued net x converges if and only if it is eventually bounded and* $\limsup x = \liminf x$.

Proof. First we prove the forward implication. If x converges to z, then x is eventually in $(z-1, z+1)$, showing that x is eventually bounded. By Proposition 4.19 combined with Proposition 4.32 we find that $\limsup x = \liminf x = \lim x$.

For the reverse implication, let $z = \limsup x = \liminf x$. Let $I \ni z$ be an open interval. By definition we have that U and L are both eventually in I; let α be such that $U_\alpha, L_\alpha \in I$ (the common α can be found as the index set is directed). Then by definition we have that for every $y \in x_{\uparrow(\alpha)}$, $\inf I < L_\alpha \leq y \leq U_\alpha < \sup I$ showing that $x_{\uparrow(\alpha)} \subseteq I$. Since I is arbitrary this shows that z is the limit of x. □

Exercise 4.18. Prove the sandwich (a.k.a. squeeze) theorem: "Suppose x, y, z are real-valued nets with the same index set, such that both $y - x$ and $z - y$ are eventually non-negative. Then if z is eventually bounded above, and x is eventually bounded below, and $\limsup z = \liminf x = r$, then all three sequences converge to r.

Exercise 4.19. Consider the sequence (meaning x is a net with the index set being $\mathbb{N}$ with its standard ordering) $x : \mathbb{N} \to \mathbb{R}$ where $x_n = \sqrt[n]{n}$, show that $\lim x = 1$. (*Hint: first show that $x_n \geq 1$. Next use the squeeze theorem together with the fact (to be proven!) that* $(1 + \sqrt{2/n})^n \geq n$.)

Exercise Sheet: Week 4

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 4.1. Since $\mathbb{Q}$ is countable, this means there exists a bijective function $\ell : \mathbb{N} \to \mathbb{Q}$. Consider $\mathbb{N}$ with its standard ordering as a directed set. Prove that: for every enumeration ℓ of the rationals, considered as a net (sequence), every real number is an accumulation point.

Question 4.2.

1. Give an example of an *unbounded* net that is eventually bounded.
2. Can there be an example of an *unbounded* sequence that is eventually bounded? Why or why not?

Question 4.3. Let x be a real-valued net, let $|x|$ denote the net whose elements $|x|_\alpha = |x_\alpha|$. Prove that if x converges then so does $|x|$. Prove that the converse is false.

Question 4.4. The notion of "eventually" is extremely useful when considering nets. Let x, y be two nets $A \to X$, where X is some set and $(A, \leqq)$ is a directed order. We way that x and y are *eventually* equal if there exists $\alpha \in A$ such that the restriction of x and y to $\uparrow(\alpha)$ are equal. Prove that if x and y are eventually equal, then

1. x is eventually in S if and only if y is eventually in S;
2. x is frequently in S if and only if y is frequently in S;
3. x is infrequently in S if and only if y is infrequently in S.

This exercise shows that only "tail events" matter for nets. Even though unlike sequences, two tail sets in a net may have infinitely many elements in their symmetric difference, it is still the case that the convergence and accumulation properties of a net is determined entirely by its restriction to any tail set.

Question 4.5. Let x be the sequence given by

- $x_1 = 0$.
- x_m, when m is even, is $x_{m-1}/2$.
- x_m, when m is odd, is $x_{m-1} + \frac{1}{2}$.

Find $\limsup x$ and $\liminf x$.

Question 4.6. Let $a > 0$; construct a sequence by setting $x_1 = a$, and

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

1. Prove that if $a = 1$, then x is the constant sequence.

2. Prove that if $a > 1$, then x is a decreasing sequence and converges to $\sqrt{a}$.
3. Prove that if $a < 1$, then x is an increasing sequence and converges to $\sqrt{a}$.

(Hint: for part 2, (a) notice that if $a > 1$ then $\sqrt{a} < a$; (b) try to prove that $x - \sqrt{a}$ decreases monotonically to zero.)

Remark: the above exercise gives a way to generate a sequence of rational numbers that converges to $\sqrt{q}$ for any q rational and positive.

Problem Set 4

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

The first two questions concern the following definition.

> Let $(A, \leqq)$ be a directed set. We say that a subset $F \subseteq A$ is a *frequent* subset if every tail set of A intersects F.

Problem 4.1. Prove that:

1. the restriction of the ordering $\leqq$ to a frequent subset F is still directed. *(And hence $(F, \leqq)$ is a directed set.)*
2. if $F \subseteq A$ is a frequent subset of A, and $G \subseteq F$ is a frequent subset of F, then G is also a frequent subset of A.

Problem 4.2. Let $x : A \to X$ be a net. Suppose F is a frequent subset of A, and let $y : F \to X$ be the net given by the restriction of x to F. Let S be a subset of X. Prove that:

1. If y is frequently in S, then x is frequently in S.
2. If x is eventually in S, then y is eventually in S.
3. Give examples to show that the converse of *both* of the previous statements are false.

Problem 4.3.

1. Let x be a real-valued net. Prove that the set of all of its accumulation points must be a closed set. (*Hints: (a) it is easier to prove that its complement, that is, the set of points at which x does NOT accumulate, is open; (b) it is easier to use directly Definition 3.24 together with Definition 4.18, rather than try to argue with Theorem 4.24.*)
2. Let $[a, b]$ be an arbitrary closed interval, give an example of a *sequence* x such that the set of all of its accumulation points is *exactly* $[a, b]$.

Problem 4.4. Let x be a real-valued sequence. Suppose x satisfies the following recursion rules:

$$x_{n+1} = \begin{cases} \frac{3}{4} x_n, & n \text{ is even;} \\ x_n + 1, & n \text{ is odd.} \end{cases}$$

Find the set of all accumulation points of x.

(Hint: consider the sequences $y_n = x_{2n-1}$ and $z_n = x_{2n}$. What are their sets of accumulation points respectively? Can x accumulate at any other points?)

Reading Assignment 5

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We introduce the notion of metric spaces. We study their basic properties, and then discuss how the notion of nets and their convergence introduced last week in the context of real numbers, can be easily carried over to the more general setting of metric spaces. We define Cauchy nets in this more general context, and define the notion of Cauchy completeness. We finish with a discussion of the idea of compactness and some of its equivalent forms.

Contents

In this set of readings, we take a short detour to see how we can transfer the Dedekind completeness of the real numbers to the study of a more general class of sets: the *metric spaces*. The material in this set of readings is only loosely related to the subsequent ones; but it serves as a good way for us to revisit the notion of limits and convergence discussed in Week 4 and to further our understanding thereof. The concept of metric spaces also underlies a lot of higher analysis, so it is useful for you to at least be aware of the definitions going forward.

§5.1 Metric spaces

The notion of metric spaces is modelled on one of the important features of the real line $\mathbb{R}$; for two numbers $x, y \in \mathbb{R}$, we can say that they are a distance $|x-y|$ apart. A metric space is simply a set with a notion of a *distance*.

Definition 5.1. *A* metric space *is a set X together with a* distance function *(also called* metric*) $d : X \times X \to \mathbb{R}$, that satisfies the following properties:*

Positivity *Given $x_1, x_2 \in X$, their distance $d(x_1, x_2) \geq 0$.*
Non-degeneracy *The distance $d(x_1, x_2) = 0 \iff x_1 = x_2$.*
Symmetry *The distances $d(x_1, x_2) = d(x_2, x_1)$.*
Triangle inequality *Given x_1, x_2, x_3, their mutual distances satisfy*

$$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3).$$

Example 5.2. The set $\mathbb{R}$ with the absolute value distance $d(x, y) = |x - y|$ is a metric space. The absolute value function only outputs non-negative numbers, and it outputs 0 if and only if $x = y$. Symmetry is due to $x - y = -(y - x)$ and have the same absolute value, and triangle inequality boils down to the standard triangle inequality for the absolute value sign. ■

Exercise 5.1. The set $\mathbb{R}^2$ can be equipped with a distance function as follows. Let (x_1, x_2) and (y_1, y_2) be elements of $\mathbb{R}^2$. Their distance is $d((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| + |x_2 - y_2|$. Prove that this distance function makes $\mathbb{R}^2$ a metric space.

Example 5.3. The previous exercise can be generalized to $\mathbb{R}^k$ for arbitrary finite dimensions. Let $\bar{x} = (x_1, \ldots, x_k)$ and $\bar{y} = (y_1, \ldots, y_k)$. We can define the distance

$$d(\bar{x}, \bar{y}) = \sum_{j=1}^{k} |x_j - y_j|. \tag{5.1}$$

Pretty much the exact same proof as the previous exercise shows that this makes $(\mathbb{R}^k, d)$ a metric space. This metric is sometimes called the *Manhattan metric* or *Taxicab metric*, because if the points in $\mathbb{R}^2$ represent a location in a large city ruled by East-West streets and North-South avenues, then this metric measures how far a taxicab will drive to get from one point to the other (ignoring potential one-way rules). ■

Example 5.4. The *Euclidean* distance on $\mathbb{R}^k$ is $d(\bar{x}, \bar{y}) = \left(\sum |x_j - y_j|^2\right)^{1/2}$. This also satisfies the four conditions for metric space. Positivity, non-degeneracy, and symmetry are more-or-less obvious. Triangle inequality requires some work: it is a consequence of the Cauchy-Schwarz inequality, and a proof can be found[1] at `https://math.stackexchange.com/a/75213/1543`. ■

Exercise 5.2. Prove that if (X, d) is a metric space, and $Y \subset X$, then Y together with the restriction of d to $Y \times Y$ forms a metric space.

The distance function allows us to transfer properties of the real line to arbitrary metric spaces. The basic idea is that "open intervals" are to be replaced by "open balls". Compare the following definitions to Definition 3.24

Definition 5.5. *Given a metric space (X, d), an* open ball, centered at x with radius r *is the set $B(x, r) = \{y \in X : d(x, y) < r\}$. (By convention, when we speak of an open ball the radius is necessarily positive, and hence $B(x, r)$ is never empty.)*

Definition 5.6. *Given a metric space (X, d), a subset $S \subseteq X$ is said to be*

open *if for every $x \in S$, there exists $r > 0$ such that $B(x, r) \subseteq S$.*
closed *if $X \setminus S$ is open.*

Exercise 5.3. Prove that in a metric space (X, d): (*Hint: triangle inequality.*)

1. If $y \in B(x, r)$, then $B(y, \rho) \subseteq B(x, r + \rho)$
2. If $B(x, r)$ and $B(y, \rho)$ have non-empty intersection, then there exists z such that $B(x, r) \cup B(y, \rho) \subseteq B(z, 2\max\{r, \rho\})$.
3. If $x_1, x_2, \ldots$ is a sequence of points in X, such that $B(x_i, 2^{-i}) \ni x_{i+1}$, then $\{x_1, x_2, \ldots\} \subseteq B(x_1, 1)$.
4. If $y \in B(x, r)$, then there exists some $\rho \in (0, r)$ such that $B(y, \rho) \subseteq B(x, r)$.
5. If $d(x, y) \geq 2r$, then $B(x, r) \cap B(y, r) = \emptyset$.

Food for Thought 5.4. In Definition 3.24, we did not require that the intervals are "centered" at the point x in question; in Definition 5.6, we did. This turns out not to matter at all; you may wish to try and see if you can justify that the two versions are equivalent.

[1] I am omitting the proof since it is tangential to this course. The Cauchy-Schwarz inequality is fundamental to analysis in higher dimensional spaces, as well as functional analysis.

§5.2 Nets in metric spaces

We can discuss convergence of nets in metric spaces in a very similar fashion to nets in $\mathbb{R}$.

Example 5.7. This is a generalization of Example 4.10. Let $\mathbb{U}_x$ be the set of all open sets that contain x. This forms a directed set with $U \leqq V \iff U \supseteq V$. We can also let $\mathbb{B}_x$ be the set of all open *balls* centered at x. This also forms a directed set. (Note that this in fact is a total order, since $B(x,r) \leqq B(x,\rho) \iff r \geq \rho$.)

First we claim that in the terminology of Problem Set 4, $\mathbb{B}_x$ is a frequent subset of $\mathbb{U}_x$. It is a subset since all open balls are open sets (this is a consequence of Exercise 5.3, part 4). It is frequent since by Definition 5.6, if $U \in \mathbb{U}_x$, since $x \in U$, there exists $r > 0$ such that $B(x,r) \leqq U$.

Any choice function on $\mathbb{U}_x$ (or on $\mathbb{B}_x$) defines a net with values in X. ■

Definition 5.8. *Let (X,d) be a metric space, and let $x : A \to X$ be a net.*

- *We say that x converges to $z \in X$ if x is eventually in every open ball (of positive radius) centered at z.*
- *We say that x accumulates at $z \in X$ if x is frequently in every ball (of positive radius) centered at z.*

Exercise 5.5. Continuing Example 5.7, check that any choice function c on $\mathbb{U}_x$ gives a net that converges to x.

The following lemma reproduces Proposition 4.19.

Lemma 5.9. *If $x : A \to X$ is a convergent net, then $\lim x$ is the unique accumulation point of x.*

Proof. Suppose $y \neq \lim x$, then by non-degeneracy of the metric, $d(y, \lim x) > 0$, and hence setting $r = d(y, \lim x)/2$, we find that $y \notin B(\lim x, r)$. Since x is eventually in $B(\lim x, r)$, it must be infrequently in its complement (Exercise 4.3). Therefore x is infrequently in $B(y,r)$ (see Exercise 5.3, part 5), and hence cannot accumulate at y. □

Exercise 5.6. Let X be an arbitrary set. Let $d : X \times X \to \mathbb{R}$ be defined by

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

Prove:

1. (X,d) is a metric space.
2. If x is a convergent net in X, then x is eventually constant.

Theorem 4.24 and Corollary 4.25 carries over also. The corresponding proofs are omitted, since they can be obtained from the proofs of the $\mathbb{R}$ case with minimal changes.

Theorem 5.10. *Let (X,d) be a metric space.*

1. *A set $S \subseteq X$ is open if and only if every net x in X with an accumulation point in S is frequently in S.*
2. *A set $S \subseteq X$ is closed if and only if every net x in X taking values in S has all of its accumulation points in S.*

Exercise 5.7. Make sure you can carry forward the proof from the real case to the metric space case, for the theorem above.

§5.3 Completeness and Compactness

Since (X, d) is, in general, not ordered, it makes no sense to think about generalizing lim sup and lim inf, or the Monotone Convergence Theorem. We will present a generalization of the Bolzano-Weierstrass Theorem. But before we start, we need to discuss the notion of *completeness*.

As we saw in the case of the reals, a lot of the convergence properties of nets are powered by notions of completeness, such as the least upper bound property (Monotone Convergence Theorem); Cantor's theorem on nested intervals from Exercise 3.7 (Cauchy implies convergent); and the Heine-Borel property (Bolzano-Weierstrass Theorem). It is reasonable to expect that a form of completeness is required; and for metric spaces, the operative notion is that of *Cauchy completeness*.

Definition 5.11. *A net x taking values in a metric space (X, d) is a* Cauchy net *if, for every real number $r > 0$, there exists a point $z \in X$ such that x is eventually in $B(z, r)$.*

The definition of a Cauchy net is formulated similar to the real valued case: it says that we can slide a ball around in X with the fixed radius r, and will be guaranteed to hit some center z for which $B(z, r)$ will contain x eventually.

Lemma 5.12. *If x is a convergent net in a metric space, then x is Cauchy.*

Exercise 5.8. Prove the lemma. This is almost identical to the first part of the proof of Theorem 4.29.

As we saw in our discussion of Cauchy nets in the reals, the converse of this lemma, in the case of the real numbers, required using the Dedekind completeness of the real numbers. For metric spaces, the proof of the converse is much simpler, since we will just define it to be so.

Definition 5.13. *A metric space (X, d) is said to be* Cauchy complete *if every Cauchy net in X converges.*

Example 5.14. From Example 5.3 and Exercise 5.2, we see that $\mathbb{Q}^k$ equipped with the distance function (5.1) is a metric space. This metric space is *not Cauchy complete*. The construction in Question 4.6 of the Week 4 exercise sheet can be used to build, with $a = 2$, a Cauchy sequence of elements in $\mathbb{Q}^k$ (setting the points to be $(x_n, 0, 0, \ldots, 0)$) that does not converge (if it were to converge, it "would" converge to $(\sqrt{2}, 0, 0, \ldots, 0)$). ■

We conclude with a discussion of a generalization of the Bolzano-Weierstrass Theorem. It turns out that for sets that are not total orders, the correct generalization of a bounded set is the following:

Definition 5.15. *Let (X, d) be a metric space. A subset S is said to be* totally bounded *if, for every $r > 0$, there exists a finite subset $\Sigma \subseteq X$, such that $\{B(x, r) : x \in \Sigma\}$ covers S.*

Exercise 5.9. Prove: if $K \subseteq \mathbb{R}$ is bounded, then K is totally bounded. (*Note: on the real line, $B(x, r) = (x - r, x + r)$. You can either use the Archimedean property, or the Heine-Borel property, to extract the finiteness of the cover.*)

Theorem 5.16. *Let (X, d) be a complete metric space. If S is a* totally bounded *subset, and x a net that takes values in S, then x has at least one accumulation point.*

Proof. The proof is divided into several steps.

Step 1.

Let $r_k = 2^{-k}$, I claim that I can choose a sequence of points z_k in X such that

- $B(z_k, r_k) \supseteq B(z_{k+1}, r_{k+1})$ for all k. (So the sets are nested).
- x is frequently in each of $B(z_k, \frac{1}{2}r_k)$.

We construct the points by induction. By total boundedness, there exists a finite cover of S by open balls of radius $\frac{1}{2}r_1$. By Exercise 4.4, x is frequently in one of the balls, call its center z_1. By enlarging the radius back to r_1, we see that x is frequently in $B(z_1, r_1)$.

Suppose the points up to z_k has been constructed; notice that x is frequently in $B(z_k, \frac{1}{2}r_k)$. Since S is totally bounded, so must be any subset, an in particular so is $S \cap B(z_k, \frac{1}{2}r_k)$. Then for radius $\frac{1}{2}r_{k+1}$, there exists a finite cover by balls of this radius, and by Exercise 4.4 again x is frequently in one of these balls, call its center z_{k+1}. By construction $z_{k+1} \in B(z_k, \frac{1}{2}r_k)$ and hence by Exercise 5.3, $B(z_{k+1}, r_{k+1}) \subseteq B(z_k, \frac{1}{2}r_k + r_{k+1}) \subseteq B(z_k, r_k)$.

Step 2.

Construct a sequence y_k, by selecting from $B(z_k, r_k)$ an arbitrary image of x. I claim that this sequence is Cauchy. First observe that by construction, $y_{\uparrow(k)} \subseteq B(z_k, r_k)$ using the nested property of the construction in the previous step. If $r > 0$ is given, by the Archimedean property of the reals, there exists some K such that $r_K < r$. Therefore $y_{\uparrow(K)} \subseteq B(z_K, r_K) \subseteq B(z_K, r)$; this shows that y is eventually in some ball of radius r, and hence y is Cauchy.

By Cauchy completeness y converges to some $\zeta \in X$.

Step 3.

I claim that ζ is an accumulation point of X. Let $r > 0$ be given. Since ζ is the limit of y, there exists some K such that $y_{\uparrow(K)} \subseteq B(\zeta, \frac{1}{4}r)$. There exists some $J > K$ such that $r_J < \frac{1}{2}r$ by the Archimedean property. For this J, we have that

- $d(y_J, \zeta) < \frac{1}{4}r$, and
- $d(z_J, y_J) < \frac{1}{2}r_J < \frac{1}{4}r$.

And so $z_J \in B(\zeta, \frac{1}{2}r)$ by the triangle inequality, and $B(z_J, \frac{1}{2}r_J) \subseteq B(\zeta, \frac{3}{4}r)$ by Exercise 5.3. Since we know that x is frequently in $B(z_J, \frac{1}{2}r_J)$ by construction, we conclude that x is also frequently in $B(\zeta, r)$.

Since $r > 0$ is arbitrary, this shows that ζ is an accumulation point of X. □

Corollary 5.17. *In Theorem 5.16, if S was further assumed to be* closed, *then we can conclude that x has an accumulation point in S.*

Proof. This is a direct consequence of Theorem 5.10. □

In the proof of Theorem 5.16, the second step extracted a sequence y from the net x, and in doing so, we proved that the limit of the convergent sequence y ends up being an accumulation point of the net x. This is a part of a much more general construction of "taking subnets"; we will discuss this next week.

The property of "every net having an accumulation point" turns out to be extremely useful in analysis: it allows us to argue for the existence of solutions to certain problems, even though we may not know exactly how to construct the solutions. By virtue of its power, it deserves a name.

Definition 5.18. *Let (X,d) be a complete metric space. A subset K of X is said to be* compact *if every net x taking values in K has an accumulation point in K.*

Theorem 5.19. *Let (X,d) be a Cauchy complete[2] metric space. The following statements about a subset K are equivalent.*

1. *K is compact.*
2. *K is closed and totally bounded.*
3. *If $\mathscr{S}$ is a collection of open subsets of X, and $\mathscr{S}$ covers K, then there exists a finite subset $\mathscr{D} \subseteq \mathscr{S}$ that also covers K.*
4. *If $f : K \to \mathbb{R}$ is a positive function, then there exists a finite subset $S \subseteq K$ such that $\{B(x, f(x)) : x \in S\}$ covers K.*

Proof. We will prove $2 \implies 1 \implies 3 \implies 4 \implies 2$.

$2 \implies 1$ is Corollary 5.17.

$3 \implies 4$: given f, set $\mathscr{S} = \{B(x, f(x)) : x \in K\}$. Let $\mathscr{D}$ be the resulting finite set of open balls, and let S be their centers.

$4 \implies 2$. To show K is totally bounded, given $r > 0$, set $f(x) \equiv r$ to be the constant function. The set $\{B(x,r) : x \in S\}$ is the finite cover by radius r balls. To show that K is closed, we will show that its complement is open. Let $p \in X \setminus K$ be arbitrary. Let $f : K \to \mathbb{R}$ be given by $f(x) = \frac{1}{2}d(x,p)$. Since $p \notin K$, and the metric is non-degenerate, we have that $f(x) > 0$ always.

Therefore there is a finite subset S such that $\{B(x, f(x)) : x \in S\}$ covers K. I claim that $B(p, \min f(S))$ is an open ball contained in $X \setminus K$. First, since S is finite, $\min f(S)$ is well defined. As a consequence of part 5 of Exercise 5.3, the ball $B(p, \min f(S))$ is disjoint from $B(x, f(x))$ for any $x \in S$, and since K is covered by the latter balls, we have as claimed that $B(p, \min f(S)) \subseteq X \setminus K$, and K is closed.

$1 \implies 3$ will be proven by contrapositive. That is, we will prove $(\neg 3) \implies (\neg 1)$.

Suppose $\mathscr{S}$ is a collection of open subsets of X, *no finite subset of which* covers K. We first build a direct set: let A denote the set of *all finite subsets of* $\mathscr{S}$. This can be directed[3] by inclusion: $\alpha_1 \leqq \alpha_2 \iff \alpha_1 \subseteq \alpha_2$. (Note that $\alpha_{1,2}$ are sets, whose elements are open subsets of X.) Note that the ordering is a partial order, and it is directed since $\alpha_1 \cup \alpha_2$ is finite and succeeds both α_1 and α_2.

Since no finite subset of $\mathscr{S}$ covers K, this means that for $\alpha \in A$, the remainder $K \setminus (\cup\alpha)$ is not empty. So it is possible to define a function $c : A \to K$ such that $c(\alpha) \in K \setminus (\cup\alpha)$. This defines a net with values in K. It suffices to prove that c has no accumulation points.

We will accomplish this by showing that for every $x \in K$, there exists a radius r_x such that the ball $B(x, r_x)$ is visited infrequently by c. Since $\mathscr{S}$ covers K, given any $x \in K$ there is an open set $U_x \in \mathscr{S}$ such that $x \in U_x$. Since U_x is open, there exists an r_x such that $B(x, r_x) \subseteq U_x$. Let $\alpha_x = \{U_x\}$, the singleton set. By construction of c, if $\alpha \geqq \alpha_x$, then $c_\alpha \notin U_x$. This means that $c_{\uparrow(\alpha_x)} \cap U_x = \emptyset$, showing that c is in U_x infrequently, and hence also that c is in $B(x, r_x)$ infrequently. □

[2] In fact, for the equivalence of the first, third, and fourth statements, it is not necessary to assume that (X,d) is Cauchy complete; it is only when introducing statement two that completeness is useful.

[3] Note that this is not the inverse order we used before.

Problem Set 5

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Special rules for PS 5*: only two problems are listed below; only those two problems need to go through Eli Review. You are asked to submit, in addition, revisions to two problems from Problem Sets 1-4, when you submit your solutions on D2L for final grading. The revision will earn you new grade toward Problem Set 5 as well as replace the original grade you earned on the problem previously. Please be sure to clearly indicate the problem set number and question number selected on your work. While not necessary, I would also appreciate it if you transcribe/summarize the question statement on the document; it will help facilitate grading.*

Let (X,d) be a metric space, that is *not-necessarily Cauchy complete*. Let C denote the set of all Cauchy *sequences* on (X,d). (Notice that since X is not necessarily Cauchy complete, elements of C do not necessarily converge.)

Problem 5.1.

1. Let $c, c' \in C$. Construct a sequence $x : \mathbb{N} \to \mathbb{R}$ by setting
$$x_n = d(c_n, c'_n).$$
Prove that x is a real-valued Cauchy sequence (and hence converges).
2. Prove that if c and c' is such $\lim d(c_n, c'_n) = 0$, and that $a \in C$ is any other Cauchy sequence, then $\lim d(c_n, a_n) = \lim d(c'_n, a_n)$.

Define now the set $C_c = C/\sim$ be the set of equivalence classes (see Section 1.1.2 in Week 1, if you need a refresher on equivalence classes), where the equivalence relation is $c \sim c' \iff \lim d(c_n, c'_n) = 0$.

By part 2 of the previous problem, the following definition of a function $\delta : C_c \times C_c \to \mathbb{R}$ makes sense. Recall that $[c]$ is the equivalence class in C_c to which the element $c \in C$ belongs. Setting
$$\delta([c],[a]) = \lim d(c_n, a_n);$$
we see that value computed by the limit expression, by virtue of part 2 of the previous problem, does not depend on which representatives are taken from the classes $[c]$ and $[a]$.

Problem 5.2.

1. Prove that (C_c, δ) is indeed a metric space by checking the properties of a metric against the function δ.
2. Prove that if $\alpha : \mathbb{N} \to C_c$ is a Cauchy sequence, then α converges.
Hint 1: The main difficulty of this question is keeping track of which object is what. If α is a Cauchy sequence in C_c, then each entry α_n is itself an equivalence class, whose elements are Cauchy sequences in X. Let $c^{(n)}$ be a representative of the class α_n, then $c^{(n)}$ is itself a Cauchy sequence on X.
Hint 2: Your goal is to construct the limit, which is an equivalence class of elements of C. And so it is enough to find a single element of this equivalent class, which we can call $z : \mathbb{N} \to X$. I claim that the sequence z can be given by the diagonal formula:
$$z_k = c^{(k)}_k, \text{ the } k\text{th term of the representative of the } k\text{th equivalence class } \alpha_k.$$

Hint 3: Given the previous two hints, what you need to prove are:

- The sequence z described above is a Cauchy sequence in X, making it an element of C.
- The sequence α converges to the equivalence class $[z]$. This requires proving that for every $r > 0$, there exists N such that $\delta(\alpha_n, [z]) < r$ for any $n \geq N$.

As a final remark: $p, q \in X$. Let $a, b \in C$ be the constant sequences $a_n = p$ and $b_n = q$. The definitions above show that $\delta([a],[b]) = \lim d(a_n, b_n) = d(p, q)$. So (C_c, δ) contains within it a copy of (X, d). In fact, if you successfully completed part 2 of Problem 2, then you have proven (up to some general nonsense converting sequences to nets) that (C_c, δ) is Cauchy complete. The process of starting from an incomplete (X, d) and building the Cauchy complete (C_c, δ) is called "taking the Cauchy completion".

This has some similarity to taking the Dedekind completion. Dedekind completion of a total order is formed by considering a set whose elements are cuts of the original order. We were able to argue that cuts of this set of cuts have the requisite property to be Dedekind complete. Cauchy completion of a metric space is formed by considering a set whose elements are Cauchy sequences in the original metric space. We were able to argue that Cauchy sequences whose terms are drawn from this set of Cauchy sequences will converge to a Cauchy sequence.

This is a very general theme in mathematics.

Problem Set 5 Supplement

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 5.2 had an incomplete Hints 1 and 2; it appears that this caused more confusion then helped. Let me try to amend that by including the missing preparatory steps.

Lemma 5-s.1. Given any Cauchy sequence $a \in C$, there exists a representative $c \in [a]$ such that for every $k \in \mathbb{N}$, there exists $z \in X$ such that $c_{\uparrow(k)} \subseteq B(z, \frac{1}{k})$.

Proof. Fix $k \in \mathbb{N}$; let $\mathcal{N}_k$ denote the set of natural numbers N such that there exists $z \in X$ for which $a_{\uparrow(N)} \subseteq B(z, \frac{1}{k})$. The set $\mathcal{N}_k$ is not empty because $a \in C$ is Cauchy. It is also an upper set by construction. So since $\mathbb{N}$ is well-ordered we can choose a function $n : \mathbb{N} \to \mathbb{N}$ such that $n(k) = \min(\mathcal{N}_k \cap \uparrow(k))$. This function is guaranteed to be increasing.

Let the sequence c be given by $c_k = a(n(k))$. Since n is increasing, $n(\uparrow(k)) \subseteq \uparrow(n(k))$. Thus our definition guarantees that there exists $z \in X$ such that $c_{\uparrow(k)} \subseteq B(z, \frac{1}{k})$. Which in turn guarantees that c is Cauchy, since $\mathbb{R}$ has the Archimedean Property and for every $\epsilon > 0$ there exists $k \in \mathbb{N}$ such that $1/k < \epsilon$.

To show that $c \in [a]$, we need to show that $\lim d(c_k, a_k) = 0$. Let $\epsilon > 0$, choose $k \in \mathbb{N}$ such that $1/k < \frac{1}{2}\epsilon$. I claim that for every $m > n(k)$, $d(c_m, a_m) < \epsilon$. But this is because $c_m = a_{n(m)}$ with $n(m) \geq m$, and so there exists $z \in X$ such that $a_m, c_m \in B(z, \frac{1}{k})$ and by triangle inequality there distance is no more than twice the radius of the ball. □

Now, for the Cauchy sequence $\alpha : \mathbb{N} \to C_c$, each α_n is an equivalence class of Cauchy sequences in X. So for each α_n you can choose a representative $c^{(n)}$ that satisfies the properties of the Lemma above.

Now construct the sequence $z : \mathbb{N} \to X$ as described: letting $z(k) = c^{(k)}(k)$.

Here's a further hint to help prove that z is Cauchy: notice that

$$d(z_n, z_m) = d(c_n^{(n)}, c_m^{(m)}) \leq d(c_n^{(n)}, c_\ell^{(n)}) + d(c_\ell^{(n)}, c_\ell^{(m)}) + d(c_\ell^{(m)}, c_m^{(m)})$$

by triangle inequality; given an $r > 0$, how can you make sure each of the three terms on the RHS is $< \frac{1}{3}r$?

(Incidentally, the same hint should help you prove that $\alpha \to z$.)

Problem Set 5 Solution to 5.2.2

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Suppose α is a Cauchy sequence in C_c. By Lemma 5-s.1, we are allowed to choose, for each n, choose $c^{(n)} \in \alpha$ a representative such that $c^{(n)}_{\uparrow(k)}$ is contained in a ball of radius $\frac{1}{k}$.

Define $z_k = c^{(k)}_k$.

First we prove that z is a Cauchy sequence

Let $r > 0$. Since α is Cauchy, there exists N_1 such that $\alpha_{\uparrow(N_1)}$ is contained in a ball of radius $\frac{1}{8}r$. Enlarge N_1, if necessary, such that $\frac{1}{N_1} < \frac{1}{8}r$.

Let $n, m \geq N_1$. Then

$$d(z_n, z_m) = d(c^{(n)}_n, c^{(m)}_m) \leq d(c^{(n)}_n, c^{(n)}_\ell) + d(c^{(n)}_\ell, c^{(m)}_\ell) + d(c^{(m)}_\ell, c^{(m)}_m)$$

by triangle inequality. By the definition of $\delta(\alpha_n, \alpha_m)$, there exists L such that for every $\ell > L$,

$$d(c^{(n)}_\ell, c^{(m)}_\ell) < \delta(\alpha_n, \alpha_m) + \frac{1}{4}r.$$

By our choice of N_1, and the triangle inequality, we also have that

$$\delta(\alpha_n, \alpha_m) < \frac{1}{4}r.$$

By our choice of representatives, and triangle inequality, if we choose ℓ to be also $> \max\{m, n\}$ we have

$$d(c^{(n)}_n, c^{(n)}_\ell) < \frac{2}{n} \leq \frac{2}{N_1} \leq \frac{1}{4}r$$

and

$$d(c^{(m)}_n, c^{(m)}_\ell) < \frac{2}{m} \leq \frac{2}{N_1} \leq \frac{1}{4}r.$$

So combining everything, we have

$$d(z_n, z_m) < \frac{1}{4}r + \frac{1}{4}r + \frac{1}{4}r + \frac{1}{4}r = r.$$

And hence

$$z_{\uparrow(N_1)} \subseteq B(z_{N_1}, r)$$

proving that z is Cauchy.

Next we prove that $\alpha \to [z]$

We compute, for arbitrary n.

$$\delta(\alpha_n, [z]) = \lim_\ell d(c^{(n)}_\ell, z_\ell)$$

And

$$d(c_\ell^{(n)}, z_\ell) \leq d(c_\ell^{(n)}, c_n^{(n)}) + d(c_n^{(n)}, z_n) + d(z_n, z_\ell)$$

by the triangle inequality. The middle term vanishes by definition.

Let $r > 0$. Since z is Cauchy, there exists M such that $z_{\uparrow(M)}$ is contained in a ball of radius $r/5$. For any $m \geq \max\{M, \frac{5}{r}\}$, then by our choice of representatives, $c_{\uparrow(m)}^{(m)}$ is contained in a ball of radius no more than $r/5$.

Therefore, for any $\ell \geq m$, we have that

$$d(c_\ell^{(m)}, c_m^{(m)}) < \frac{2r}{5}$$

and

$$d(z_m, z_\ell) < \frac{2r}{5}.$$

And therefore for we have that for all $\ell \geq m$

$$d(c_\ell^{(m)}, z_\ell) < \frac{4r}{5}.$$

And therefore

$$\lim_\ell d(c_\ell^{(m)}, z_\ell) \leq \limsup_\ell d(c_\ell^{(m)}, z_\ell) \leq \frac{4r}{5} < r.$$

And hence we have shown that given $r > 0$, there exists M such that for every $m \geq \max\{M, \frac{5}{r}\}$, we have

$$\delta(\alpha_m, [z]) < r.$$

And hence $\alpha \to [z]$.

Reading Assignment 6

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We finish up our discussion of real-valued nets and their convergence by discussing two related, additional concepts. First we introduce the idea of subnets and subsequences. The notion of subnets affords us with the nice characterization that given any net, any accumulation point thereof can be written as the limit of a subnet. This gives another useful way of approaching the study of accumulation points. We follow this by an esoteric presentation of the concept of the infinite series. Traditionally, the notion of the infinite series is approached primarily through the sequence of its partial sums. We take a different route here and start by thinking about sums of infinite sets of numbers, using the net concept we already introduced. We then see that the distinction between absolute and conditional convergence, a stalwart of Calculus II classes around the country, can in fact be regarded as a manifestation of the distinction between limit points and accumulation points. Some theorems concerning absolute convergence are proven, and on the conditional convergence side, we prove the Riemann rearrangement theorem.

Contents

§6.1 Subnets

Given a net $x : A \to X$, sometimes it is instructive to consider a net whose values are drawn from those of x. The following definition is due to S. Willard.[1]

Definition 6.1. *Given a net $x : A \to X$, a net $y : B \to X$ is said to be a* subnet *of x, if there exists a function $\varphi : B \to A$, such that*

- $y = x \circ \varphi$

[1] In the literature there are many different *inequivalent* definitions of subnets. Willard's version is among the more restrictive, but still retains enough generality for applications. There are other less restrictive definitions due to J. L. Kelley, and to J. F. Aarnes and P. R. Andenæs; these less restrictive definitions can be easier to manipulate in some cases, and harder to manipulate in others. However, in terms of statements regarding convergence and accumulation of nets, statements that hold true for one type of subnets of the above also hold true for the others. It turns out that there are ways of converting between these three notions, though the conversion is often indirect and non-obvious. We will therefore limit ourselves to the more common version by Willard in these notes.

- *φ is increasing:* $\beta_1 \leqq_B \beta_2 \implies \varphi(\beta_1) \leqq_A \varphi(\beta_2)$
- *for every $\alpha \in A$, there exists $\beta \in B$ with* $\varphi(\beta) \geqq_A \alpha$.

Exercise 6.1. Check that if y is a subnet of x, and z a subnet of y, then z is a subnet of x.

Example 6.2. Note that if x is a sequence, and y is a subnet of x, *it is not true that y is a subsequence*.

Example 1: Suppose $x : \mathbb{N} \to \mathbb{R}$ is given by $x_n = n$. The following is a subnet of x:

$$y_n : 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, \ldots$$

Usual definitions of subsequences (as taught in calculus classes) require the corresponding φ function to be strictly increasing, and thus ruling out repeats. In the subnet definition, repeats are allowed.

Example 2: The fact that repeats are allowed, means that the index set used for a subnet may have larger cardinality than that of the net. A trivial example has x being any sequence, and y having its index set the real numbers, with φ the function that sends a real number β to the least natural number that is greater than or equal to β.

Example 3: Since there can exist increasing functions from a general directed set to a total order, we can have more complicated order relation for B than for A. For example, enlarge $\mathbb{N}$ by adding to it a second copy of the numbers 1 and 2; we will distinguish the two copies by denoting them as $1_1, 2_1$ and $1_2, 2_2$ respectively. The ordering we take to have $1_1 < 2_1 < 3 < 4 < \cdots$ and $1_2 < 2_2 < 3 < 4 < \cdots$ and with the $1_1, 1_2, 2_1, 2_2$ not otherwise comparable. This is still a directed set; call it $\tilde{N}$. If we set the function $\varphi : \tilde{N} \to \mathbb{N}$ to be

$$\varphi(1_1) = 1, \quad \varphi(1_2) = 2, \quad \varphi(2_1) = 3, \quad \varphi(2_2) = 4, \qquad \varphi(x)|_{\uparrow(3)} = x + 2$$

we see that this φ defines a subnet of any sequence, now with domain $\tilde{N}$. ■

Example 6.3. For a less trivial example, return to the example from Week 4 on the two different orderings of the natural numbers. For clarity, denote the standard total order by $(\mathbb{N}_+, \leq)$, and the divisibility partial order by $(\mathbb{N}_\times, \leqq)$. I claim that the identity function $\varphi(n) = n$ satisfies the conditions for generating a subnet (with domain $\mathbb{N}_\times$) out of a sequence (with domain $\mathbb{N}_+$).

We see that φ is increasing: if $n \leqq m$ in $\mathbb{N}_\times$, then m is a multiple of n, and hence $\varphi(n) = n \leq m = \varphi(m)$ in $\mathbb{N}_+$ also. And for any $n \in \mathbb{N}_+$, treating it as an element of $\mathbb{N}_\times$, we have $\varphi(n) = n \geq n$ satisfying the final property. ■

Exercise 6.2. Find a function $\varphi : \mathbb{N}_+ \to \mathbb{N}_\times$, that can generate a subnet indexed by $\mathbb{N}_+$ out of a net indexed by $\mathbb{N}_\times$.

Example 6.4. A special class of subnets, that more closely resemble subsequences, are those associated to *frequent* subsets. Let A be a directed set, and let $F \subseteq A$ be a frequent subset (see Problem Set 4). Part 1 of Problem 4.1 shows that F is a directed set. Let $\varphi : F \to A$ be the natural inclusion map. Then since the ordering on F is restricted from that of A, the inclusion map is increasing. And the definition of F being a frequent subset automatically satisfies the final property listed in Definition 6.1. Not all subnets can be expressed as one coming from a frequent subset: the differences vis-à-vis subsequences discussed in Example 6.2 also apply to subnets coming from frequent subsets. In the context of analysis on the real line or in metric spaces, these differences turn out to be insignificant. ■

The Example 6.4 leads naturally to the following proposition.

Proposition 6.5. *Let x be a net in X. Suppose $S \subseteq X$ is such that x is frequently in S. Then there exists a subnet y of x that is eventually in S.*

Proof. Denote by A the index set of x. Let $B \coloneqq \{\beta \in A : x_\beta \in S\}$. I claim that B is a frequent subset. But by definition of x being frequently in S, we know that $x_{\uparrow(\alpha)} \cap S \neq \emptyset$ for all α, this immediately means that $\uparrow(\alpha) \cap B \neq \emptyset$ for all α, and hence B is frequent. The subnet associated to B has all its terms taking values in S, and therefore is certainly eventually in S. □

Your work on Problem 4.2 can be generalized to statements about subnets.

Theorem 6.6. *If x is a net in X (with index set A), and y is a subnet (with index set B),*

- *If x is eventually in S, then y is eventually in S.*
- *If y is frequently in T, then x is frequently in T.*

Proof. Denote by φ the mapping such that $y = x \circ \varphi$. Notice that since φ is increasing, for any fixed $\beta \in B$, we have the inclusion $\varphi(\uparrow(\beta)) \subseteq \uparrow(\varphi(\beta))$. (On the LHS φ is the induced function on the power set, while on the RHS φ is the original function.) This implies

$$y_{\uparrow(\beta)} \subseteq x_{\uparrow(\varphi(\beta))}. \tag{6.1}$$

Now, suppose x is eventually in S, so that there exists α such that $x_{\uparrow(\alpha)} \subseteq S$. By definition, there exists β such $\varphi(\beta) \geqq_A \alpha$. By (6.1), we have $y_{\uparrow(\beta)} \subseteq x_{\uparrow(\varphi(\beta))} \subseteq x_{\uparrow(\alpha)} \subseteq S$, and hence y is eventually in S.

Next, suppose that y is frequently in T. Given $\alpha \in A$, by definition there exists $\beta \in B$ such that $\varphi(\beta) \geqq_A \alpha$. Since y is frequently in T, we know $y_{\uparrow(\beta)} \cap T$ is non-empty. By (6.1) again we see also $x_{\uparrow(\alpha)} \cap T \neq \emptyset$, and hence x is frequently in T. □

Corollary 6.7. *(On either the real line, or in a metric space.) Let x be a net and y be a subnet. Then if x converges to a point z, then so does y. If y accumulates at a point z', then so does x.*

Example 6.8. We can now revisit some of our discussions in the Week 4 readings in view of Corollary 6.7. Based on Example 6.3, for a given function $x : \mathbb{N} \to \mathbb{R}$, the net formed by regarding x as over $\mathbb{N}_\times$, can be thought of as a subnet of the sequence by regarding x as over $\mathbb{N}_+$. So returning to Exercise 4.8, we can check that indeed that the set of all accumulation points of the subnet x over $\mathbb{N}_\times$, is a subset of all accumulation points of the sequence x over $\mathbb{N}_+$. But the reverse inclusion may fail. ■

The results above are based on general symbol pushing. The following theorem, which asserts that accumulation points can be regarded as limits of subnets, require a bit more work.

Theorem 6.9. *Let X be either the real line, or a metric space. Let $x : A \to X$ be a net, and z an accumulation point of x. Then there exists a subnet y of x that converges to z.*

Proof. For the purpose of this proof, a "neighborhood" of z, when $X = \mathbb{R}$, is an open interval containing z; and when X is a metric space, it is an open ball with positive radius centered at z. Since x accumulates at z, every neighborhood N of z contains some element x_α.

Let B denote the set of all *ordered pairs* (N, α), where N is a neighborhood of z, and $\alpha \in A$ is such that $x_\alpha \in N$. Order B by setting

$$(N_1, \alpha_1) \leqq_B (N_2, \alpha_2) \iff \big(N_1 \supseteq N_2 \wedge \alpha_1 \leqq_A \alpha_2\big). \tag{6.2}$$

It is not too hard to see that this ordering is reflexive and transitive. It is directed because: given (N_1, α_1) and (N_2, α_2), we can set $M = N_1 \cap N_2$, which is still a neighborhood of z. Furthermore, since A is directed there is β that succeeds both α_1 and α_2, and since x accumulates at z there is an element $\gamma \in \uparrow(\beta)$ with $x_\gamma \in M$. So (M, γ) succeeds both (N_1, α_1) and (N_2, α_2) in B.

Let $\varphi : B \to A$ be the mapping $\varphi((N, \alpha)) = \alpha$. This mapping is increasing by virtue of (6.2). Now since x accumulates at z, given any $\alpha_0 \in A$ and any neighborhood N_1 of z, there is $\alpha_1 \in \uparrow(\alpha_0)$ such that $x_{\alpha_1} \in N_1$. And hence $\varphi((N_1, \alpha_1)) \geqq_A \alpha_0$. Thus $y = x \circ \varphi$ is a subnet.

It remains to show that $y \to z$. This is true by construction, since if N is a neighborhood of z, then for any element of the form (N, α) of B, any succeeding element (N', α') must have $x_{\alpha'} \in N' \subseteq N$. And hence $y_{\uparrow((N,\alpha))} \subseteq N$ as required for showing convergence. □

Food for Thought 6.3. When x is a sequence, and X a metric space or the real line, the construction in the previous proof gives a pretty horrendous subnet. It turns out in this special case we can do something different and construct y as a subsequence: meaning that the domain of φ is still $\mathbb{N}$ and φ is strictly increasing. The construction uses two ingredients: first is that $\mathbb{N}$ (the domain of the original sequence x) is well-ordered and infinite, and that the Archimedean property holds for the reals (the codomain of the distance function d). We can set $\varphi(1)$ arbitrarily, and $\varphi(k+1) = \min\{n : n > k \wedge d(x_n, z) < \frac{1}{k+1}\}$. You should verify that this indeed gives rise to a subsequence, and the subsequence converges to z.

To give an example showing that there are general nets taking values in a metric space, which do not contain a "subsequence" that converges to an accumulation point, we can let $A = \{\emptyset\}$ be the one point set with its universal ordering, and $x_\emptyset = z$. This net converges to z but does not have any subsequence converging to z by virtue of the fact that there are no strictly increasing (and hence injective) functions from $\mathbb{N} \to A$.

For less trivial examples that shows general nets may not have subsequences that converge to its accumulation points, this requires a study of topological spaces in more generality then we can fit within the scope of this course.

Combining Theorem 6.9 and Corollary 6.7, we get the following characterization:

Corollary 6.10. *Let x be a net (in the reals or in a metric space), then z is an accumulation point of x if and only if there exists a subnet y that converges to z.*

We close our discussion of general nets and subnet with a structural theorem on the set of accumulation points.

Theorem 6.11. *Let x be a net (in the reals or in a metric space), then the set of all of its accumulation points is a closed set.*

Exercise 6.4. Prove the above theorem following the steps here:

1. Let S denote the set of all accumulation points of x. Let $z \notin S$. It suffices to prove that there is a neighborhood (interval or ball) about z that is disjoint from S. (Why?)
2. There exists a neighborhood N about z that is visited infrequently by x.
3. Let $z' \in N$, then there exists a neighborhood N' about z' that is a subset of N.
4. Hence x does not accumulate at z'.

§6.2 Application to infinite sums

Since the real numbers form a field, it is closed under addition of two numbers. By induction this means that we can add any *finite* number of real numbers and get a number out. But can we add *infinitely many* real numbers? It turns out that one way to make sense of this is through the concept of nets and convergence.

Let $\mathcal{I}$ be an arbitrary set. Consider a function $\tau : \mathcal{I} \to \mathbb{R}$. Let A be the set of all *finite* subsets of $\mathcal{I}$, ordered by inclusion. A is a subset of the poset $2^{\mathcal{I}}$ and hence is a poset, it is also directed since if α_1, α_2 are two finite subsets of $\mathcal{I}$, so is the set $\alpha_1 \cup \alpha_2$ which succeeds both α_1 and α_2. Since we know how to add finitely many real numbers, we can construct a net $x : A \to \mathbb{R}$ where

$$x_\alpha = \sum_{i \in \alpha} \tau(i).$$

Since α is finite this sum is well-defined. We can then interpret the limit of the net x, if it exists, as the infinite sum of the numbers represented by τ.

Proposition 6.12. *If $\tau(i) \neq 0$ for* uncountably many *indices i, then the net x cannot converge.*

Proof. Write $\hat{\mathcal{I}}$ for the set of indices i for which $\tau(i) \neq 0$. For $k \in \mathbb{N}$, let $\mathcal{I}_k^+ = \{i \in \mathcal{I} : \tau(i) > \frac{1}{k}\}$ and $\mathcal{I}_k^- = \{i \in \mathcal{I} : \tau(i) < -\frac{1}{k}\}$. By the Archimedean property we find $\hat{\mathcal{I}}$ is the union of all the $\mathcal{I}_k^\pm$, which is a countably infinite list of sets. Were the $\mathcal{I}_k^\pm$ all finite, this would imply their union is countable. Since our hypothesis is that $\hat{\mathcal{I}}$ is uncountable, then at least one of $\mathcal{I}_k^\pm$ is infinite.

Without loss of generality, suppose the infinite set is $\mathcal{I}_{k_0}^+$. (The minus case can be treated similarly.)

Then, I claim the net x cannot be Cauchy. By the Archimedean property, there exists $N \in \mathbb{N}$ such that $N \cdot \frac{1}{k_0} > 2$. Let $\alpha \in A$ be a finite subset of $\mathcal{I}$. Since $\mathcal{I}_{k_0}^+$ is infinite, there exists a subset $\alpha' \subseteq \mathcal{I}_{k_0}^+ \setminus \alpha$ with N elements. We observe that

$$x_{\alpha \cup \alpha'} - x_\alpha = \sum_{i \in \alpha'} \tau(i) > N \cdot \frac{1}{k_0} > 2.$$

And hence $x_{\uparrow(\alpha)}$ cannot be contained within any interval of width less than 2. Since α is arbitrary, this shows that x is not Cauchy, and hence does not converge. □

Therefore, when it comes to infinite sums, it is only interesting to consider the case where there are only countably many numbers to add. And since the finite case is well-understood, we can assume that we have countably infinitely many non-zero numbers to add.[2] And this brings us to the study of **series**.

Assumption 6.13. *For the remainder of this section, we shall assume that we are adding a countably infinite list of non-zero numbers. We will fix the following notations.*

- *We let $\mathcal{I}$ be a countably infinite set, and $\tau : \mathcal{I} \to \mathbb{R} \setminus \{0\}$ our list of non-vanishing terms.*
- *We will denote by $A \subseteq 2^{\mathcal{I}}$ the set of all finite subsets of $\mathcal{I}$.*
- *We have the net $x : A \to \mathbb{R}$, where $x_\alpha = \sum_{i \in \alpha} \tau(i)$, which is well defined as a finite sum of real numbers.*

[2] The assumption that there are no zero terms is mainly for convenience; many of the constructions below can also be done when there are zero terms. However, frequently the case with finitely many and infinitely many zeros have to be treated separately as different cases, adding complications without adding clarity.

First, let us define a series. A series depends on a particular enumeration of $\mathcal{I}$ (remember, a enumeration of a countably infinite set is just a bijection from $\mathbb{N}$ to the set).

Definition 6.14. *Given a particular enumeration $\eta : \mathbb{N} \to \mathcal{I}$, the associated* series *is the sequence σ of* partial sums *of τ, whose terms are $\sigma_n = \sum_{j=1}^{n} \tau(\eta(j))$. The series is said to converge if the sequence σ converges, in which case we write $\sum_\eta \tau = \lim \sigma$.*

Notice that since the partial sums depend on the choice of enumeration η, the series limit may depend also on η. Hence notationally we indicate the sum is performed with respect to the enumeration η by indicating it in the subscript.

How does series convergence compare against the infinite sum described using the limit of x?

Lemma 6.15. *σ is a subnet of x.*

Proof. Let $\varphi : \mathbb{N} \to A$ be the mapping $n \mapsto \{\eta(1), \eta(2), \ldots, \eta(n)\}$. Then $\sigma = x \circ \varphi$. This functions φ is increasing since A is ordered by inclusion. And for any $\alpha \in A$, since it is finite, we know the natural number $\max \eta^{-1}(\alpha)$ exists. Then $\varphi(\max \eta^{-1}(\alpha)) \supseteq \alpha$, and hence φ satisfies all the conditions in Definition 6.1. □

Corollary 6.16. *If $\lim x$ exists, then for every enumeration η of $\mathcal{I}$, the corresponding series converges and $\sum_\eta \tau = \lim x$.*

Proof. This follows by combining Corollary 6.7, with Lemma 6.15. □

But is it possible that $\sigma \to z$ but x doesn't converge?

Example 6.17. Suppose that $\mathcal{I} = \mathbb{N}$, and we have chosen η to be the identity map. Let the terms of τ be given by

$$1, -1, \frac{1}{2}, \frac{-1}{2}, \frac{1}{2}, \frac{-1}{2}, \frac{1}{3}, \frac{-1}{3}, \frac{1}{3}, \frac{-1}{3}, \frac{1}{3}, \frac{-1}{3}, \frac{1}{4}, \frac{-1}{4}, \ldots$$

The corresponding sequence σ has its terms

$$1, 0, \frac{1}{2}, 0, \frac{1}{2}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{4}, 0, \frac{1}{4}, 0, \ldots$$

and clearly converges to 0.

I claim the net x is not Cauchy. Let $\alpha \in A$. Based on our construction, there exists some odd number $M > \max \alpha$, and a natural number N, such that

$$\tau(M) = \tau(M+2) = \tau(M+4) = \ldots = \tau(M+2(N-1)) = \frac{1}{N}.$$

Let $\beta = \{M, M+2, \ldots, M+2N-2\}$. Then we have

$$x_{\alpha \cup \beta} = x_\alpha + 1.$$

And hence $x_{\uparrow(\alpha)}$ cannot be contained within any interval of width 1/2. This proves that x is not Cauchy, and hence does not converge. ■

At the risk of giving the entire game away, let us introduce the following definition for *absolute convergence*.[3] Note that this definition is rather different from what you have previously encountered in your calculus course. The definition given here is more natural from our point of view having introduced nets. We will prove the equivalence with the notion familiar from your calculus class later in these notes.

Definition 6.18. *We say that the infinite sum of* τ converges absolutely *if the net* x *converges. In this case we write* $\sum_{\text{abs}} \tau = \lim x$.

We use the subscript "abs" to denote this is the absolute convergence that is independent of any enumeration of $\mathcal{I}$.

So another way to summarize the result so far is that if the infinite sum of τ converges absolutely, then the series of τ relative to any enumeration η converges, and $\sum_\eta \tau = \sum_{\text{abs}} \tau$. For the remainder of this section, we aim to answer two questions:

1. What are some sufficient conditions for absolute convergence in terms of series convergence?
2. What can we say when some series for τ converges, but the net x diverges?

Note that in the latter case, since σ is a subnet of x, this means x has some finite accumulation point. The following lemma is known as the "nth term test for divergence" in calculus textbooks.

Lemma 6.19. *If there exists an enumeration* η*, such that the corresponding* $\sum_\eta \tau$ *converges, then for every* $\epsilon > 0$*, the set* $\{i \in \mathcal{I} : |\tau(i)| \geq \epsilon\}$ *is finite.*

Exercise 6.5. Prove Lemma 6.19. (*Hint: most of the work is already done in the proof to Proposition 6.12.*)

Example 6.20. The fact that we speak of an enumeration and the accompanying series is crucial in Lemma 6.19. One may be tempted to suggest that "if x has an accumulation point, then for every $\epsilon > 0$, the set ..." is true. Unfortunately it isn't.

Consider the underlying set $\mathcal{I}$ to be the natural numbers $\mathbb{N}$. Let τ be the function $n \to (-1)^n$. Let $B \subseteq A$ be given by the set of all finite subsets that satisfy:

$$\beta \cap \{2k-1, 2k\} \neq \emptyset \implies \beta \supseteq \{2k-1, 2k\}.$$

B is a frequent subset of A, and hence the restriction of x to B is a subnet. But this subnet is the constant subnet: it equals 0 always. And hence x has an accumulation point 0. However, it is not true that $\{i \in \mathcal{I} : |\tau(i)| \geq \frac{1}{2}\}$ is finite. ∎

Exercise 6.6. For the example $\tau(n) = (-1)^n$ of the previous example, consider its net x, and determine the set of all accumulation points of x.

§6.2.1 Absolute convergence.—For convenience, we will denote by

$$\mathcal{I}^+ := \{i \in \mathcal{I} : \tau(i) > 0\}, \quad \mathcal{I}^- := \{i \in \mathcal{I} : \tau(i) < 0\}. \tag{6.3}$$

We observe that $\mathcal{I}^+$ and $\mathcal{I}^-$ are disjoint, and their union equals $\mathcal{I}$. Notice that since $\mathcal{I}$ is infinite, the two subsets cannot both be finite.

[3] For real valued sums, the definition is equivalent to the notion based on sums of the absolute values. There are settings in functional analysis in which these two notions are not equivalent; in those contexts the notion defined here is called "unconditional convergence".

Theorem 6.21. *Suppose there exists an enumeration η for which $\sum_\eta \tau$ converges, and suppose that exactly one of $\mathcal{I}^\pm$ is finite, then the infinite sum of τ converges absolutely and $\sum_{\text{abs}} \tau = \sum_\eta \tau$.*

Proof. Without loss of generality, assume that $\mathcal{I}^-$ is finite; this leaves $\mathcal{I}^+$ infinite. Let η be the enumeration, and σ the corresponding sequence of partial sums. Then there exists some $M \in \mathbb{N}$ such that $\tau \circ \eta(n) > 0$ for all $n > M$, using the finiteness of $\mathcal{I}^-$.

Since the sequence σ converges, it is Cauchy, and hence for every width $w > 0$ there exists an N and an open interval J with width w such that $\sigma_{\uparrow(N)} \subseteq J$. We can assume further that $N > M$.

Now let $\alpha = \eta(\{1, 2, \ldots, N\})$. I claim that $x_{\uparrow(\alpha)} \subseteq J$. This is because for any $\beta \supseteqq \alpha$, we have $x_\beta = x_\alpha + \sum_{i \in \beta \setminus \alpha} \tau(i)$. If we let $M' = \max \eta^{-1}(\beta)$, we see that

$$x_\alpha = \sigma_N \leq x_\beta \leq \sigma_{M'}.$$

This chain of inequality holds because for every $i \in \mathcal{I} \setminus \alpha$, $\tau(i) > 0$. Since σ_N an $\sigma_{M'}$ are both in J, so must x_β.

With this, we have proven that x is Cauchy, and hence converges. □

Theorem 6.22. *Let τ take only positive values, and suppose $\sum_{\text{abs}} \tau$ converges absolutely. If $\mu : \mathcal{I} \to \mathbb{R}$ is any function, such that for every i, $|\mu(i)| \leq \tau(i)$, then the infinite sum of μ also converges absolutely.*

Proof. Denote by $y_\alpha = \sum_{i \in \alpha} \mu(i)$ the finite sum, for $\alpha \in A$. We observe that since $|\mu| \leq \tau$, we find that for any $\beta \supseteqq \alpha$, we have that

$$|y_\beta - y_\alpha| \leq x_\beta - x_\alpha \tag{6.4}$$

by triangle inequality.

Let $w > 0$ be a width. Since x converges, there exists an open interval J with width $w/2$ and an index α such that $x_{\uparrow(\alpha)} \subseteq J$. And hence for any $\beta \supseteqq \alpha$, $x_\beta - x_\alpha \in [0, w/2)$. Applying (6.4), we see that $|y_\beta - y_\alpha| < w/2$ also, and hence $y_{\uparrow(\alpha)} \subseteq (y_\alpha - w/2, y_\alpha + w/2)$, an interval of width w. This shows that y is Cauchy, and hence converges. □

When μ takes also positive values, Theorem 6.22 is essentially the *comparison test* for series convergence.

Exercise 6.7. Putting together Theorem 6.21 and 6.22, prove the Calculus II statement (here stated in typical Calculus II notation): "If $\sum_{i=1}^\infty |a_i|$ converges, then $\sum_{i=1}^\infty a_i$ converges."

The previous exercise connects our notion of absolute convergence, to the more familiar calculus notion.

Theorem 6.21 can be generalized further.

Theorem 6.23. *Denote by τ^+ the restriction of τ to $\mathcal{I}^+$, and τ^- the restriction of τ to $\mathcal{I}^-$. Suppose the sums $\sum_{\text{abs}} \tau^+$ and $\sum_{\text{abs}} \tau^-$ both converge absolutely, then the infinite sum for τ converges absolutely and $\sum_{\text{abs}} \tau = \sum_{\text{abs}} \tau^+ + \sum_{\text{abs}} \tau^-$.*

Proof. We will denote by $x^\pm$ the nets corresponding to $\tau^\pm$. Let $w > 0$ be a width, then since $x^\pm$ both converge, there exists intervals $J^\pm$, each of width $w/2$, and indices $\alpha^\pm$ (being finite subsets of $\mathcal{I}^\pm$ respectively), such that $x^\pm_{\uparrow(\alpha^\pm)} \subseteq J^\pm$ respectively.

Now let $\alpha = \alpha^+ \cup \alpha^-$. Suppose $\beta \geqq \alpha$. Let $\beta^\pm = \beta \cap \mathcal{I}^\pm$. We have that $x_\beta = x^+_{\beta^+} + x^-_{\beta^-} \in J$, where we define $J = J^+ + J^-$. Note that J is an interval with width w. And thus we have shown that x is Cauchy, and hence converges. □

It is illustrative to ask what would happen if exactly one of the two sums $\sum_{\text{abs}} \tau^\pm$ diverge. The following theorem answers that.

Theorem 6.24. *Suppose exactly one of $\sum_{\text{abs}} \tau^\pm$ converges, then the net x has no accumulation points.*

Proof. We will assume without loss of generality that $\sum_{\text{abs}} \tau^-$ converges, The case for $\sum_{\text{abs}} \tau^+$ is similar. We note that the nets $x^\pm$ are both monotone; x^+ is increasing and x^- is decreasing. Hence $x_- \geq \lim x_-$. And furthermore, by the Theorem 4.26 on Monotone Convergence, for x^+ to fail to converge necessarily x^+ is unbounded.

I claim that for every $M \in \mathbb{R}$, there exists some $\alpha \in A$ such that for every $\beta \geqq \alpha$, $x_\beta > M$. This would show that every bounded open interval of $\mathbb{R}$ is infrequently visited by x, and hence x does not have any accumulation points.

To prove the claim, we notice that since x^+ is increasing and unbounded, there exists α^+ such that $x^+_{\alpha^+} > M - \lim x_-$. Observing that α^+ is a finite subset of $\mathcal{I}^+$ and hence a finite subset of $\mathcal{I}$, we see that we can take α^+ as an index for x. For any $\beta \in A$ that succeeds α_+, we can split β into $\beta^\pm = \beta \cap \mathcal{I}^\pm$. We find then $x_\beta = x^+_{\beta^+} + x^-_{\beta^-} \geq x^+_{\alpha^+} + x^-_{\beta^-} > M - \lim x_- + \lim x_-$, which proves the claim. □

§6.2.2 Conditional convergence.—Next we discuss the implications of τ admitting an enumeration such that $\sum_\eta \tau$ converges, but x does *not* converge. By Theorems 6.23 and 6.24, we know that for this to happen, necessarily both $\mathcal{I}^\pm$ are infinite, and both of the nets $x^\pm$ have to be unbounded.

Lemma 6.25. *Suppose x is divergent, but there exists an enumeration η such that the series σ converges. Then there exists enumerations $\nu_\pm$ of $\mathcal{I}^\pm$ respectively, such that $\tau \circ \nu_+$ is a decreasing function, and $\tau \circ \nu_-$ is an increasing function.*

Proof. We give the proof for the "+" case, the "−" case is similar.

Divide $\mathcal{I}^+$ into the subsets $\mathcal{I}^+_k$, where $k \in \mathbb{N}$. These subsets are defined by

$$\mathcal{I}^+_k := \{i \in \mathcal{I}^+ : k - 1 < 1/\tau(i) \leq k\}.$$

By Lemma 6.19, each one of the $\mathcal{I}^+_k$ are finite. So we can define ν_+ by listing first the elements of $\mathcal{I}^+_1$, in the order of decreasing τ value, followed by those of $\mathcal{I}^+_2$, and so on and so forth. □

Exercise 6.8. Prove that under the assumptions of Lemma 6.25, the enumerations $\nu_\pm$ are such that the partial sums $\sigma^\pm_n = \sum_{j=1}^n \tau \circ \nu_\pm(j)$ grow unboundedly.

Theorem 6.26 (Riemann rearrangement theorem)**.** *Suppose x is divergent, there exists an enumeration η_0 with a convergent series. Let $z \in \mathbb{R}$. Then there exists a possibly different enumeration η of $\mathcal{I}$ such that the corresponding series converges, with $\sum_\eta \tau = z$.*

Sketch of proof. We build η in a zig-zag procedure. Let $\eta(1) = \nu_+(1)$, and $\sigma_1 = \tau \circ \nu_+(1)$. At each step k,

- If $\sigma_k < z$, let $\eta(k+1)$ be the next unused ν_+ value, and set $\sigma_{k+1} = \sigma_k + \tau \circ \eta(k+1)$.

- If $\sigma_k \geq z$, let $\eta(k+1)$ be the next unused ν_- value, and set $\sigma_{k+1} = \sigma_k + \tau \circ \eta(k+1)$.

Since $\sigma^{\pm}$ both are unbounded, this process forever alternates: once η starts taking from ν_+, after finitely many steps it must start taking from ν_-, and vice versa. Therefore, every element of $\mathcal{I}^{\pm}$ is eventually used. The monotonicity of $\nu_{\pm}$ guarantees then $\limsup \sigma = \liminf \sigma = z$. □

Exercise 6.9. Justify why, in the proof above, that $\limsup \sigma = \liminf \sigma = z$.

Putting together all the results in the section, we obtain the following:

Theorem 6.27. *If there exists an enumeration η whose series converges, then the net x corresponding to the infinite sum of τ satisfies exactly one of the following:*

- *x converges; or*
- *the set of all accumulation points of x is $\mathbb{R}$.*

§6.3 Some convergence tests

We will take the notational Assumptions 6.13 for this section.

Define,

$$\mathcal{J}_0 := \{i \in \mathcal{I} : |\tau(i)| \geq 1.\} \tag{6.5}$$

and for $k \in \mathbb{N}$,

$$\mathcal{J}_k := \{i \in \mathcal{I} : |\tau(i)| \in [2^{-k}, 2^{1-k}).\} \tag{6.6}$$

We shall assume that each of these sets are finite. (For if any were infinite, by Lemma 6.19 together with Theorem 6.27, the infinite sum will have no convergent series.) We will denote by ρ_k, for $k \in \mathbb{N} \cup \{0\}$, the corresponding number of elements in $\mathcal{J}_k$.

Exercise 6.10. Prove the following: if the series $\sum_{k=1}^{\infty} 2^{-k}\rho_k$ converges, then the infinite sum $\sum_{\text{abs}} \tau$ converges absolutely.

This last exercise is, essentially, the *Cauchy condensation test* for convergence.

Exercise 6.11. Prove the following: if there exists $C > 0$ and $s < 1$ such that $\rho_k \leq C2^{sk}$ for all k, then $\sum_{\text{abs}} \tau$ converges absolutely.

This last exercise is essentially the *p-series test*.

Exercise Sheet: Week 6

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 6.1 (Partial Summation 1)**.** Let τ_n, τ'_n be real sequences. Denote by σ_n, σ'_n their partial sums: $\sigma_n = \sum_{j=1}^{n} \tau_j$ and $\sigma'_n = \sum_{j=1}^{n} \tau'_j$. Verify, by induction, that for $n > m$

$$\sigma_n\sigma'_n - \sigma_m\sigma'_m = \sum_{j=m}^{n-1} \Big(\tau_{j+1}\sigma'_j + \sigma_j\tau'_{j+1} + \tau_{j+1}\tau'_{j+1} \Big).$$

Question 6.2 (Partial Summation 2)**.** Let $\tau_n, \tau'_n, \sigma_n, \sigma'_n$ be as in the previous exercise. Verify, by induction, that for $n > m$,

$$\sigma_n\sigma'_n - \sigma_m\sigma'_m = \sum_{j=m+1}^{n} \tau_j\sigma'_j + \sum_{j=m}^{n-1} \sigma_j\tau'_{j+1}.$$

Question 6.3. Let τ, τ' be functions from $\mathcal{I} \to \mathbb{R}$, where $\mathcal{I}$ is countably infinite.

1. Suppose τ takes only non-negative values, and its infinite sum converges absolutely, prove that the infinite sum of the function τ^2 converges absolutely.
2. Suppose τ takes only non-negative values and its infinite sum converges absolutely, and suppose τ' is bounded, prove that the infinite sum of the function $\tau \cdot \tau'$ converges absolutely.
3. Suppose the infinite sums of the functions τ^2 and τ'^2 both converges absolutely, prove that the infinite sum of the function $\tau \cdot \tau'$ converges absolutely.

Question 6.4. Let $\tau : \mathbb{N} \to \mathbb{R}$ be the function $\tau(n) = (-1)^n$. Let x be the net of its finite sums (see Assumption 6.13). Find all accumulations points of x.

Question 6.5. Let $\tau : \mathbb{N} \to \mathbb{R}$ be the function $\tau(n) = (-1)^n$. Prove that there exists no enumeration $\nu : \mathbb{N} \to \mathbb{N}$ such that the series corresponding to $\tau \circ \nu$ converges.

Problem Set 6

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 6.1. Using the partial summation formulae from the exercise sheet, prove:

1. *(Abel's Test)* If the sequence a_n is bounded and monotone (not necessarily converging to zero), and if the series $\sum_{n=1}^{\infty} b_n$ converges, then the series $\sum_{n=1}^{\infty} a_n \cdot b_n$ converges.
2. *(Dirichlet's Test)* If the sequence a_n is positive, and monotonically decreases to zero, and if the sequence b_n is such that there exists $M > 0$ such that for every k, $|\sum_{n=1}^{k} b_n| < M$. Then the series $\sum_{n=1}^{\infty} a_n \cdot b_n$ converges.

(*Hint: in both cases, let* $\sigma_n = a_n$.)

Problem 6.2. Let $\tau : \mathcal{I} \to \mathbb{R}$, where $\mathcal{I}$ is countably infinite. Prove that the infinite sum of τ converges absolutely if and only if the infinite sum of $|\tau|$ converges absolutely.

Hints: first remember that our definition of absolute convergence is Definition 6.18. One direction of the argument is already proven in Theorem 6.22. It is enough to prove the reverse direction. You can consider the functions $\tau^{\pm}$ as in Theorem 6.23. The following steps may help:

1. Using Theorems 6.23 and 6.24, reduce the problem to the case where neither of the sums of τ^+ or τ^- converge absolutely.
2. Prove that in this case the corresponding nets $x^{\pm}$ are both unbounded.
3. Prove that in this case x cannot be Cauchy.

Problem 6.3. Let (X, d) be a Cauchy complete metric space. Let $z : \mathbb{N} \to X$ be a sequence. Let $\tau : \mathbb{N} \to \mathbb{R}$ be given by

$$\tau(n) = d(z_n, z_{n+1}).$$

1. Prove that if the infinite sum for τ converges absolutely, then the sequence z converges.
2. Give an example to show that the converse is false: namely, give an example of a Cauchy complete metric space (X, d), and a converging sequence z, such that the corresponding infinite sum for τ does not converge.

Problem 6.4. For this problem, you will prove Banach's fixed point theorem:

> Let (X, d) be a Cauchy complete metric space. Let $f : X \to X$ be a function such that there is a real number $\lambda \in (0, 1)$, such that for every $z, z' \in X$,
>
> $$d(f(z), f(z')) \leq \lambda d(z, z').$$
>
> Then there exists a unique $z_0 \in X$ solving $f(z_0) = z_0$.

You can follow the following steps.

1. Construct a sequence $y : \mathbb{N} \to X$ by picking $y_1 \in X$ arbitrary, and setting $y_{n+1} = f(y_n)$. Prove that this sequence converges. *(Hint: consider* $d(y_n, y_{n+1})$ *and apply the previous problem.)*
2. Prove that the limit satisfies $f(\lim y) = \lim y$.
3. Prove that there *cannot* be two distinct values z_0 and z_1 both satisfying $f(z_0) = z_0$ and $f(z_1) = z_1$.

Reading Assignment 7

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

We move on to discussion of continuity of functions, both from $\mathbb{R} \to \mathbb{R}$ and between metric spaces. We introduce some related concepts to continuity (semi-continuity, uniform continuity). We discuss properties of continuous functions, focusing on the intermediate value and extremal value theorems, and their generalizations.

Contents

§7.1 Continuity

§7.1.1 Basic definitions.—In a beginning calculus course, you have learned that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous if $f(x) = \lim_{a \to x} f(x)$. Now that we have actually learned about nets, we can actually say the same sentence, but with confidence and bravado. Remember first that, if $x : A \to X$ is a net, and $f : X \to Y$ is a function, then by function composition, $f \circ x : A \to Y$ is a net in Y.

Definition 7.1. *Let $S \subseteq \mathbb{R}$, and let $z \in S$. A function $f : S \to \mathbb{R}$ is said to be* continuous at the point z *if for every real-valued net x, which takes values only in S, and which converges to z, we have* $\lim f \circ x = f(\lim x) = f(z)$. *The function is said to be* continuous on S *if it is continuous at every point in $z \in S$.*

Example 7.2. A trivial example: the identity function $f(s) = s$ is continuous.

Another trivial example: the constant function $g(s) = c$ is continuous. This is the case because for any net x, $g \circ x$ is the constant net and converges. ∎

Example 7.3. Using the results of §4.2.1, we see that

1. Any polynomial is continuous at every point.

2. Any rational function (ratio of two polynomials; its domain is the set of all real numbers which are not a root of the denominator polynomial) is continuous at every point in its domain. ■

In your beginning calculus course, you often only considered continuity of functions defined on intervals: since the notion of $\lim_{x\to a} f(x)$ you used is based on approaching the point a on the interval; and for closed intervals you only speak of one-sided limits. The net definition above is much more flexible.

Example 7.4. Let $S = \{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots\}$. Then *any* function $f : S \to \mathbb{R}$ is continuous. This is because for every $p = \frac{1}{k} \in S$, we can set $\epsilon = \frac{1}{2k^2}$ and check that the open interval $(p-\epsilon, p+\epsilon) \cap S = \{p\}$. This means that if x is a net with values in S that converges to p, then x is eventually equal p. And hence $f \circ x$ is eventually equal to $f(p)$ (not just in some neighborhood of it). And the continuity at p holds.

In contrast, let $T = S \cup \{0\}$. Not every function $f : T \to \mathbb{R}$ is continuous. This is because there exists a net x taking values in T that converges to 0 but is not eventually constant. (Think: the sequence $x_n = \frac{1}{n}$.) So the condition of being continuous at 0 is not trivial. (On the other hand, any function is continuous at every p that is *not* zero, by the same argument as above for S.)

To give an example of a function that is discontinuous at 0, consider $f(x) = 1/x$ if $x \in S$ and $f(0) = 0$. Then the sequence $x_n = \frac{1}{n}$ converges to 0 in T, but $f(x_n) = n$, and this sequence diverges. ■

Before we continue, it is convenient to introduce the notion of a canonical net.

Exercise 7.1. Let $S \subseteq \mathbb{R}$, and $z \in S$. Let $\mathbb{I}_z$ be the set of all open intervals containing z. Consider the set

$$A := \{(y, I) \in S \times \mathbb{I}_z : y \in I\}, \quad \text{with ordering } (y_1, I_1) \leqq (y_2, I_2) \iff I_1 \supseteq I_2.$$

Let $x : A \to S$ be given by $x((y, I)) = y$.

1. Prove that A is a directed set, and hence x is a net taking values in S.
2. Prove that $x \to z$.

The net is called the *canonical net at z*.

Theorem 7.5. *A function $f : S \to \mathbb{R}$ is continuous at $z \in S$ if and only if for every open interval $J \ni f(z)$, there is an open interval $I \ni z$ such that $f(S \cap I) \subseteq J$.*

Proof. ($\Rightarrow$): Given that f is continuous, let $J \ni f(z)$. Let x be the canonical net at z (see Exercise 7.1). Since by Definition 7.1, $f \circ x \to f(z)$, there exists $(y, I) \in A$ such that $f(x_{\uparrow((y,I))}) \subseteq J$. Noting that for every $y' \in S \cap I$, $(y', I) \geqq (y, I)$, we see that for every $y' \in S \cap I$, $f(y') \in J$, proving the theorem.

($\Leftarrow$): Suppose $x : B \to S$ is an arbitrary net converging to z. We need to prove that $f \circ x$ converges to $f(z)$. Given $J \ni f(z)$ an open interval, by assumption there exists $I \ni z$ such that $f(S \cap I) \subseteq J$. Since $x \to z$ there exists $\beta \in B$ such that $x_{\uparrow(\beta)} \subseteq I$. But this implies $f(x_{\uparrow(\beta)}) \subseteq J$. This means $f \circ x$ is eventually in any open interval around $f(z)$, and hence converges to $f(z)$. □

The net description in Definition 7.1 is easy to understand: it says that if we approach the point z in the domain, then the corresponding images will also approach $f(z)$. The meaning of Theorem 7.5 requires a bit more interpretation. One way to think about it is in terms of error bars. The open interval J can be regarded as a required tolerance imposed on the output: it says in addition to $f(z)$, we can accept anything that is close enough to $f(z)$, in the sense of being in the interval J. The theorem statement then says that a function that is continuous at z is precisely one for which, given any degree

of tolerance of the output, we can find a degree of control of the input (being in the interval I) that guarantees that the output will be within the tolerance. What does this have to do with continuity? This ability to have a guarantee would fail if the function is so sensitive that changing the input by an arbitrarily small amount will result in the output tolerance no longer met. And this is precisely how we can imagine what a (jump) discontinuity should look like.

Exercise 7.2. Definition 7.1 pretty trivially (with the change of a few symbols) extends to the case that S is a subset of some metric space (X, d_X), and the codomain of the function is another metric space (Y, d_Y). Try to come up with, and prove, an analogous version of Theorem 7.5 for metric spaces, with balls replacing intervals.

The following is a pretty convenient characterization of continuous functions on metric spaces.

Theorem 7.6. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is continuous on X if and only if the induced power set function $f^{-1} : 2^Y \to 2^X$ maps closed subsets to closed subsets.*

Before giving the proof, the following lemma is useful.

Lemma 7.7. *If $f : X \to Y$ is any function. Then the induced power set map $f^{-1} : 2^Y \to 2^X$ maps closed sets to closed sets if and only if it maps open sets to open sets.*

Exercise 7.3. Prove the lemma.

Proof of Theorem 7.6. ($\Rightarrow$): Let $K \subseteq Y$ be closed, and consider $f^{-1}(K) = \{x \in X : f(x) \in K\}$. Let μ be an arbitrary net in $f^{-1}(K)$, let z be an accumulation point of μ. By replacing μ by a subnet, we can assume $\mu \to z$. Definition of continuity requires then $f \circ \mu \to f(z)$ also. But $f \circ \mu$ is a net taking values in a closed set K, and hence by Theorem 5.10 the limit $f(z) \in K$, meaning that $z \in f^{-1}(K)$ also. Since this works for all nets μ and all accumulation points z, this implies $f^{-1}(K)$ is closed by the same corollary.

($\Leftarrow$): By Lemma 7.7, we assume that the power set map f^{-1} maps open sets to open sets. Let μ be a net converging to $z \in X$. Let $r > 0$, consider the ball $B(f(z), r)$, which is open. Then $f^{-1}(B(f(z), r))$ is also open by hypothesis, and since $z \in f^{-1}(B(f(z), r))$, there exists $r' > 0$ such that $B(z, r') \subseteq f^{-1}(B(f(z), r))$. Since $\mu \to z$, μ is eventually in $B(z, r')$, which implies that $f \circ \mu$ is eventually in $B(f(z), r)$. Since r is arbitrary, we conclude $f \circ \mu \to f(z)$. □

§7.1.2 Other modes of continuity.—The following definition is a strengthened version of continuity for metric spaces.

Definition 7.8. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is said to be* uniformly continuous *if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that for $x, x' \in X$, the bound $d_X(x, x') < \delta \implies d_Y(f(x), f(x')) < \epsilon$.*

Food for Thought 7.4. To appreciate the difference between a function $f : X \to Y$ being uniformly continuous, versus only being continuous on X, contrast Definition 7.8 against the formulation you came up with for Exercise 7.2.

Thinking in terms of the tolerance / guarantee picture, the difference between continuity and uniform continuity is that for continuity, the tolerance is on a sliding scale: you can change the amount of control you require of the input based on the value of the preferred input. With uniform continuity, you require the amount of control to be the same for all possible inputs. The following example illustrates the difference.

Example 7.9. Let $f : (0,1) \to \mathbb{R}$ be given by $f(x) = 1/x$. This function is a rational function and hence is continuous.

We can also verify its continuity by using Theorem 7.5. Let $z \in (0,1)$. Let $(a,b) \ni f(z)$ be an open interval. Our goal is to find an open interval containing z whose image under f is contained in (a,b).

Consider the value $a' = \max\{a, 1\}$, then $(a', b) \subseteq (a,b)$, and $(\frac{1}{b}, \frac{1}{a'}) \subseteq (0,1)$ and $z \in (\frac{1}{b}, \frac{1}{a'})$ (since $a < \frac{1}{z} < b$). A direct computation shows that $f((\frac{1}{b}, \frac{1}{a'}) \cap (0,1)) \subseteq (a,b)$ and we've prove that f is continuous.

The function f however is not uniformly continuous. We will show that for $\epsilon = 1$, for any $\delta > 0$, there exists $x, x' \in (0,1)$ with $|x - x'| < \delta$ and $|f(x) - f(x')| \geq \epsilon$. Indeed, let $x = \min\{\delta, \frac{1}{2}\}$, and $x' = x/2$. Then $|x - x'| = \frac{1}{2}\min\{\delta, \frac{1}{2}\} < \delta$. And $|f(x) - f(x')| = |\frac{2}{x} - \frac{1}{x}| = \max\{\frac{1}{\delta}, 2\} > 1$. ∎

The function $f(x) = 1/x$ on the interval $(0,1)$ is continuous, since for any $z \in (0,1)$, changing the input a small amount will also only change the output a small amount. However, the function is not uniformly continuous because, for the same amount of change of input, as z gets closer to zero, the corresponding change of output gets larger and larger.

Exercise 7.5. Let $f : (0,1) \to \mathbb{R}$ be the function $f(x) = \cos(1/x)$. Prove using the definition that this function is not uniformly continuous. (To prove that it is continuous requires a precise definition of the cos function; so we will skip that here.)

Exercise 7.6. Let $f : [0,1) \to \mathbb{R}$ be such that, when restricted to the interval $(0,1)$, it agrees with the function $x \mapsto \sin(1/x)$. Prove that this function is discontinuous at 0 regardless of what value f takes there. (*Hint: this would be the case if there exists a positive-real-valued net $\mu \to 0$ such that $f \circ \mu$ does not converge.*)

Our experience from calculus courses have taught us that there are different ways that a function can be discontinuous. The exercise above shows one of the more severe: that the discontinuity cannot be "repaired" by shifting the value of the function at that point. Let us formalize this concept.

Definition 7.10. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is said to have a* removable discontinuity *at x_0 if f is discontinuous at x_0 but there exists $g : X \to Y$ that is continuous at x_0, such that $f(x) = g(x)$ for all $x \neq x_0$.*

Exercise 7.6 shows that any function on $[0,1)$ that agrees with $\sin(1/x)$ on $(0,1)$ is discontinuous at 0, and the discontinuity is not removable.

Exercise 7.7. Prove that if a function f has a removable discontinuity at x_0, then for every net $\mu \to x_0$, with μ taking values only in $X \setminus \{x_0\}$, the net $f \circ \mu$ converges in Y.

The converse of the above exercise is a bit harder, so I include the proof here.

Proposition 7.11. *If a function $f : X \to Y$ between metric spaces is such that, for every net $\mu \to x_0 \in X$, with μ taking values only in $X \setminus \{x_0\}$, the net $f \circ \mu$ converges in Y, then f is either continuous at x_0 or has a removable discontinuity there.*

Proof. It suffices to show that under the hypotheses, the limit of $f \circ \mu$ is independent of μ. If so, then the function g that is equal to f on $X \setminus \{x_0\}$ but equal to the common limit would be continuous at x_0, proving the proposition.

It suffices to show then that under the hypotheses, the limits of $f \circ \mu$ and $f \circ \nu$ agree, for any pair of

nets μ, ν both taking values in $X \setminus \{x_0\}$ and converging to x_0. Let A, B denote the underlying directed set of the nets μ, ν respectively. Let $C = A \times B \times \{0,1\}$, with ordering

$$(\alpha, \beta, i) \leqq (\alpha', \beta', i') \iff \alpha \leqq \alpha' \wedge \beta \leqq \beta'.$$

This can be easily checked to make C into a directed set.

Let the net $\lambda : C \to X$ be given by

$$\lambda(\alpha, \beta, i) = \begin{cases} \mu(\alpha) & i = 0 \\ \nu(\beta) & i = 1 \end{cases}.$$

Notice that $\lambda \to x_0$, since both μ, ν converge to x_0. Furthermore, λ takes values only in $X \setminus \{x_0\}$; hence by our hypothesis, $f \circ \lambda$ converges.

On the other hand, since $f \circ \lambda_{\uparrow((\alpha,\beta,i))} = f \circ \mu_{\uparrow(\alpha)} \cup f \circ \nu_{\uparrow(\beta)}$, this shows that $\lim f \circ \lambda = \lim f \circ \mu = \lim f \circ \nu$, as we desired. □

Food for Thought 7.8. The construction of the net λ from the nets μ and ν is a generalization of the "interlacing" construction for sequences, where we combine two sequence μ_n and ν_n into a sequence λ whose elements are $\mu_1, \nu_1, \mu_2, \nu_2, \ldots$. When the underlying directed sets of μ and ν are equal, the same interlacing procedure can be done. But when the underlying directed sets are different, the slightly more complicated argument above needs to be used. Notice that in the construction, neither μ nor ν are obviously subnets of λ. However, one can check that there exists nets μ', ν', such that μ' is a subnet of μ, and ν' is a subnet of ν, and both μ', ν' are subnets of λ.

For functions $f : (a, b) \to \mathbb{R}$, and $x \in (a, b)$, let $\mu : (a, x) \to (a, x)$ be the identity map, where the domain is equipped with the usual ordering; and let $\nu : (x, b) \to (x, b)$ be the identity map, where the domain is equipped with the reverse ordering. Then both μ, ν converge to x. The function f is said to have a limit from below (or from the left) at x if $f \circ \mu$ converges, and the function f is said to have a limit from above (or from the right) at x if $f \circ \nu$ converges.

Exercise 7.9. Let $f : (a, b) \to \mathbb{R}$, x, and μ, ν as in the paragraph above. Prove that:

1. f is continuous at x if and only if $\lim f \circ \mu = f(x) = \lim f \circ \nu$.
2. f has a removable discontinuity at x if and only if $\lim f \circ \mu = \lim f \circ \nu \neq f(x)$.

(*Hint: part 2 follows immediately from part 1, based on our definition of removable discontinuity. For part 1 you may wish to pass through Theorem 7.5.*)

The function f may be called *continuous from below* at x if $\lim f \circ \mu = f(x)$ and *continuous from above* at x if $\lim f \circ \nu = f(x)$.

For functions that take values in the real number line, we can take advantage of the real numbers being totally ordered to incorporate the concepts of lim sup and lim inf, and arrive at a weakening of continuity called "semi-continuity".

Definition 7.12. *Let (X, d_X) be a metric space, and $f : X \to \mathbb{R}$ a function.*

- *f is said to be* upper semi-continuous *at a point x_0 if for every net $\mu \to x_0$, its value $f(x_0) \geq \limsup f \circ \mu$.*
- *f is said to be* lower semi-continuous *at a point x_0 if for every net $\mu \to x_0$, its value $f(x_0) \leq \liminf f \circ \mu$.*

Exercise 7.10. Prove that a real-valued function f is continuous at x_0 if and only if it is both upper semi-continuous and lower semi-continuous at x_0.

Exercise 7.11. Returning to Exercise 7.6, for which values c would setting $f(0) = c$ make the function upper semi-continuous? How about lower semi-continuous?

We can get a characterization similar to Theorem 7.6 for semi-continuity.

Theorem 7.13. *Let (X, d_X) be a metric space, and $f : X \to \mathbb{R}$ a function. Denote by $f^{-1} : 2^{\mathbb{R}} \to 2^X$ its induced power set mapping.*

- *f is upper semi-continuous on all of X if and only if $f^{-1}(\uparrow(y))$ is closed for every $y \in \mathbb{R}$.*
- *f is lower semi-continuous on all of X if and only if $f^{-1}(\downarrow(y))$ is closed for every $y \in \mathbb{R}$.*

Proof (upper case). Notice that $f^{-1}(\uparrow(y))$ is closed if and only if $\{x \in X : f(x) < y\}$ is open.

($\Rightarrow$): Suppose $f(x_0) < y$. Let $\mu \to x_0$ be a net. By upper semi-continuity $\limsup f \circ \mu \leq f(x_0) < y$. By Proposition 4.32, this means all accumulation points of $f \circ \mu$ are less than y, meaning that $f \circ \mu$ is eventually in $(-\infty, y)$. This means μ is eventually in $\{x \in X : f(x) < y\}$. And hence by Theorem 5.10 the set $\{x \in X : f(x) < y\}$ is open.

($\Leftarrow$): We argue by contrapositive. Suppose there exists $x_0 \in X$ and $\mu \to x_0$ such that $\limsup f \circ \mu > f(x_0)$. Then for some $y \in (f(x_0), \limsup f \circ \mu)$, we have μ is frequently in $f^{-1}(\uparrow(y))$. Therefore there exists a subnet ν (see Example 6.4) of μ that takes values only in $f^{-1}(\uparrow(y))$. So we have exhibited a net ν taking values in $f^{-1}(\uparrow(y))$ whose limit x_0 is outside that set, and therefore by Theorem 5.10 the set $f^{-1}(\uparrow(y))$ cannot be closed. □

Food for Thought 7.12. Do not confuse semi-continuity with continuity from the left or right on an interval (see Exercise 7.9). There are nine ways to form pairs between the three properties {upper s.c., lower s.c., not s.c.} and {cont. from below, cont. from above, neither}. For each class, give an example of a function $f : \mathbb{R} \to \mathbb{R}$ that belongs only in that class, and not in the others.

§7.2 Interpolation and Extrapolation

§7.2.1 Intermediate Value Theorem.—A basic property of continuous functions on the real line is that if f hits values a and b, then it must hit any value in between. This is a form of "interpolation". As we shall see, this property is based on the Dedekind completeness of the real number line.

Theorem 7.14 (Intermediate Value)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Suppose $f(a) \neq f(b)$, then for any γ strictly between $f(a)$ and $f(b)$, there exists $c \in (a, b)$ such that $f(c) = \gamma$.*

Proof. We shall assume without loss of generality that $f(a) < \gamma < f(b)$; the case with the reverse inequality is similar. Let $U = \{x \in [a, b] : f(x) > \gamma\}$. Then U is non-empty since $b \in U$, and U is lower bounded since $a \notin U$. Hence $\inf U$ exists by the greatest lower bound property; we will denote by $c = \inf U$.

I first claim that $f(c) \geq \gamma$: let $\mathcal{I}_c$ be the directed set of open intervals containing c. Since $c = \inf U$, for every $J \in \mathcal{I}_c$, there exists an element $x_J \in J \cap U$. This defines a net converging to c. Since f is continuous, the net $f \circ x$ converges to $f(c)$. Since the net $f \circ x$ takes values in the closed set $\uparrow(\gamma)$, by Corollary 4.25, $f(c) \in \uparrow(\gamma)$ also. This also implies that $c > a$.

Next, the identity map $\mu : [a,c) \to [a,c)$ is a net (the domain using the usual ordering on $\mathbb{R}$); this net converges to c. Since $[a,c) \cap U = \emptyset$ (with c being $\inf U$), we know that $f \circ \mu$ takes values in the closed set $\downarrow(\gamma)$, and hence by Corollary 4.25, and the continuity of f, again we find $f(c) = \lim f \circ \mu \in \downarrow(\gamma)$. Hence $\gamma \leq f(c) \leq \gamma$ which implies $f(c) = \gamma$.

Since neither $f(a)$ nor $f(b)$ equals γ, this means $c \in (a,b)$. □

Exercise 7.13. Prove the one-dimensional Brouwer fixed-point theorem: "if $f : [a,b] \to [a,b]$ is continuous, then there exists $c \in [a,b]$ such that $f(c) = c$."

Corollary 7.15. *Let $I \subseteq \mathbb{R}$ be an interval, and $f : I \to \mathbb{R}$ is a continuous function. If we know $f(x) \neq \gamma$ for any $x \in I$, and $f(x_0) < \gamma$ for some x_0, then $f(x) < \gamma$ for all $x \in I$.*

Proof. Suppose the conclusion were false, that there is $x_1 \in I$ such that $f(x_1) > \gamma$ (by assumption $f(x_1) \neq \gamma$). Then Theorem 7.14 applies and there exists some c between x_0 and x_1 such that $f(c) = \gamma$, a contradiction. □

This formulation of the theorem relies on the fact that both the domain and codomain are totally ordered. The most natural generalization of the Intermediate Value Theorem to metric spaces relies on the notion of connectedness, which more closely resembles the corollary.

Definition 7.16. *A metric space (X,d) is said to be* connected *if there does* not *exist a partition*[1] *$\{X_1, X_2\}$ of X into closed subsets. The metric space is said to be* disconnected *otherwise.*

Exercise 7.14. Let $X = \mathbb{R} \setminus \{0\}$ and take $d(x,y) = |x-y|$. Prove that X is disconnected. (*Hint: this means that even though the set $(-\infty, 0)$ is* not *closed as a subset of $\mathbb{R}$, it is in fact closed as a subset of the metric space X.*)

Theorem 7.17. *Suppose (X, d_X) is a connected metric space, and (Y, d_Y) is disconnected, with partition $\{Y_1, Y_2\}$ by closed subsets. If $f : X \to Y$ is a continuous function, then $f(X)$ can only intersect one of Y_1 and Y_2.*

Proof. Suppose for contradiction $f(X)$ intersects both Y_1 and Y_2, then $f^{-1}(Y_1)$ and $f^{-1}(Y_2)$ are disjoint, and both non-empty. By Theorem 7.6, $f^{-1}(Y_1)$ and $f^{-1}(Y_2)$ are both closed. However, since every $x \in X$ has $f(x) \in Y_1 \cup Y_2$ (they forming a partition), every $x \in X$ is in one of $f^{-1}(Y_1)$ or $f^{-1}(Y_2)$. Therefore $\{f^{-1}(Y_1), f^{-1}(Y_2)\}$ form a partition of X by closed sets, which contradicts the assumption that X is connected. □

§7.2.2 Extensions from dense subsets.—The fact that continuous functions map nearby points to nearby points means that there is some degree of rigidity concerning how wild the function can be. To capture this idea we will introduce the notion of a dense subset.

Definition 7.18. *A subset S of a metric space (X,d) is said to be* dense *if for every $x \in X$ and $r > 0$, $B(x,r) \cap S \neq \emptyset$.*

Example 7.19. As a consequence of the Archimedean property, $\mathbb{Q}$ is dense in $\mathbb{R}$, so is the set of all irrational numbers, or the set of all rational multiples of $\sqrt{2}$. ■

Example 7.20. $\mathbb{N}$ and $\mathbb{Z}$ on the other hand, are not dense in $\mathbb{R}$. ■

[1] Remember that in a partition, neither component can be empty.

Proposition 7.21. *Let f, g be continuous functions from a metric space (X, d_X) to (Y, d_Y). Suppose the restriction of f and g to a dense subset $S \subseteq X$ agree, then $f = g$ everywhere.*

Proof. Let $x \in X$. That $\mu : (0,1] \to X$ be a net, where $(0,1]$ is equipped with the reverse (decreasing) ordering, such that $\mu_r \in B(x,r) \cap S$. This net is well-defined for all r since S is dense; and $\mu \to x$. We know $f \circ \mu = g \circ \mu$ since the restrictions of f and g to S agree, and μ takes values in S. Since both f and g are continuous, we have

$$f(x) = \lim f \circ \mu = \lim g \circ \mu = g(x).$$

This holds for every x, and hence $f \equiv g$. □

In the previous proposition, we assumed that there are already some given continuous functions on X, and concluded that their values on the whole of X is fixed, once we know the values on a dense subset. It turns out that it is also possible to *extend*[2] a continuous function, initially defined only on a dense subset S, to the entire space X.

Theorem 7.22. *Let (X, d_X) and (Y, d_Y) be metric spaces, with Y Cauchy complete. Let $S \subseteq X$ be a dense subset; note that d_X restricts to a metric on S. Given a* uniformly continuous function $\mathring{f} : S \to Y$, *then there exists a unique uniformly continuous function $f : X \to Y$ that extends $\mathring{f}$.*

Before proving the theorem, we record a lemma about Cauchy sequences.

Lemma 7.23. *If (X, d_X) and (Y, d_Y) are metric spaces, and $f : X \to Y$ is uniformly continuous, then for any Cauchy net μ in X, the net $f \circ \mu$ is Cauchy.*

Proof. Let $\epsilon > 0$. By uniform continuity, there exists $\delta > 0$ such that $d_X(x, x') < \delta \implies d_Y(f(x), f(x')) < \epsilon$. Since μ is Cauchy, this means there exists $z \in X$ such that μ is eventually in $B(z, \delta)$. By uniform continuity, we have that $f(B(z,\delta)) \subseteq B(f(z), \epsilon)$. And hence $f \circ \mu$ is eventually in $B(f(z), \epsilon)$, showing that it is Cauchy. □

To prove our theorem, for any $x \in X \setminus S$ we will approximate it by a net μ with values in S. This net is convergent, and hence Cauchy, and so $\mathring{f} \circ \mu$ is a Cauchy net that converges to some point in y, using the completeness of Y. We would like to set $f(x)$ to be this value. To do so, however, requires making sure that this procedure is well-defined: that if I chose a different net μ to start with, this won't give us a different function. Finally we have to check that the extension f is uniformly continuous; uniqueness follows from Proposition 7.21.

Proof of Theorem 7.22. Let $x \in X$, then by density there exists a net $\mu \to x$ such that μ takes values in S. Since μ is convergent, it is Cauchy in X. Since the metric on S is the restriction of that of X, this means that as a net in S, μ is still Cauchy. By Lemma 7.23 $\mathring{f} \circ \mu$ is Cauchy in Y, and hence converges because Y is complete. Suppose ν is another such net, we can analogously establish the limit of $\mathring{f} \circ \nu$. I claim that the two limits agree.

[2] The difference between these two notions are: The Proposition solves the "uniqueness problem" for extensions: it says that if an extension exists, then it is unique. The Theorem that follows solves the "existence problem" for extensions: it gives sufficient conditions for the extension to exist.

Suppose not, then $\epsilon_0 \coloneqq d_Y(\lim \mathring{f} \circ \nu, \lim \mathring{f} \circ \mu) > 0$. On the other hand, since μ, ν converge to the same value, for every $\delta > 0$, there exists α, β such that $\mu_{\uparrow(\alpha)} \cup \nu_{\uparrow(\beta)} \subseteq B(z, \delta)$ with $z \in S$. By uniform continuity, this means that for every $\epsilon > 0$, there exists a $y \in Y$ such that both $\mathring{f} \circ \mu$ and $\mathring{f} \circ \nu$ are eventually in $B(y, \epsilon)$. Choosing $\epsilon < \frac{1}{3}\epsilon_0$ we get a contradiction.

Hence the limit $\lim \mathring{f} \circ \mu$ is independent of the Cauchy net μ chosen, and we can define $f(x) = \lim \mathring{f} \circ \mu$.

We complete this proof by showing that f is uniformly continuous. Let $\epsilon > 0$, then there exists $\delta > 0$ such that for $s, s' \in S$, that $d_X(s, s') < \delta \implies d_Y(\mathring{f}(s), \mathring{f}(s')) < \epsilon$. Let $d(x, x') < \delta$. Let $\mu \to x$ and $\mu' \to x$ be nets with values in S. By our construction, for every $\hat{\delta} > 0$, we can choose indices α, α' such that each of the following distances

$$d_X(\mu_\alpha, x), \;\; d_X(\mu'_{\alpha'}, x'), \;\; d_Y(\mathring{f}(\mu_\alpha), f(x)), \;\; d_Y(\mathring{f}(\mu'_{\alpha'}), f(x')) \;\; < \hat{\delta}.$$

By triangle inequality

$$d_Y(f(x), f(x')) \leq d_Y(\mathring{f}(\mu_\alpha), f(x)) + d_y(\mathring{f}(\mu_\alpha), \mathring{f}(\mu'_{\alpha'})) + d_Y(\mathring{f}(\mu'_{\alpha'}), f(x'))$$

and

$$d_X(\mu_\alpha, \mu'_{\alpha'}) < d_X(\mu_\alpha, x) + d_X(x, x') + d_X(x', \mu'_{\alpha'}).$$

This latter inequality, for $\hat{\delta}$ sufficiently small, and with $d_X(x, x') < \delta$, implies that $d_X(\mu_\alpha, \mu'_{\alpha'}) < \delta$. This implies that

$$d_Y(f(x), f(x')) < 2\hat{\delta} + \epsilon.$$

Since $\hat{\delta}$ is arbitrary, we conclude that $d_X(x, x') < \delta \implies d_Y(f(x), f(x')) \leq \epsilon$, showing uniform continuity of f. □

Example 7.24. Each of the hypothesis in Theorem 7.22 are needed for the argument.

1. Let $X = \mathbb{R}$ and $S = Y = \mathbb{R} \setminus \{0\}$. The identity function $S \to Y$ is uniformly continuous, but does not have any continuous extension to a function $X \to Y$. This shows Cauchy completeness of Y is crucial.
2. Let $X = \mathbb{R} = Y$ and $S = \mathbb{R} \setminus \{0\}$. The function $\mathring{f} : S \to Y$ with $\mathring{f}(x) = 1/x$ is continuous but not uniformly continuous (see Example 7.9). It has no continuous extension to a function $X \to Y$. This shows that uniform continuity of $\mathring{f}$ is crucial. ■

Theorem 7.22 is extremely useful in functional analysis; there are lots of arguments that work by checking that a certain property holds for a dense subset of elements, and then asserting that the same property holds for all elements "by continuity".

Exercise 7.15. Prove that a function $\mathring{f} : (a, b) \to \mathbb{R}$ is uniformly continuous if and only if there exists an uniformly continuous function $f : [a, b] \to \mathbb{R}$ that extends $\mathring{f}$.

§7.3 Continuous functions and compact sets

Let us return now to a different property for sets in metric spaces: compactness. You may wish to review Definition 5.18 and Theorem 5.19 for the notions.

Theorem 7.25. *Let $f : K \to Y$ where K is a compact subset of a complete metric space, and Y is a metric space. If f is continuous, then f is uniformly continuous.*

Proof. Let $\epsilon > 0$. Then there exists a function $\delta : K \to \mathbb{R}$ such that $f(B(x, \delta(x)) \cap K) \subseteq B(f(x), \frac{1}{2}\epsilon)$, since f is continuous. Since $\delta(x) > 0$ for every x, by Theorem 5.19, there exists a finite subset $S \subseteq K$ such that $\{B(s, \frac{1}{2}\delta(s)) : s \in S\}$ covers K.

Let $\delta_0 = \frac{1}{2} \min \delta(S)$. Suppose $x, x' \in K$ is such that $d_X(x, x') < \delta_0$. Since $\{B(s, \frac{1}{2}\delta(s)) : s \in S\}$ covers K, there exists s_* such that $x \in B(s_*, \frac{1}{2}\delta(s_*))$, By Exercise 5.3 we see that $x' \in B(s_*, \delta(s_*))$, since $\delta(s_*) \geq \frac{1}{2}\delta(s_*) + \delta_0$. And hence $f(x), f(x') \in B(f(s_*), \frac{1}{2}\epsilon)$, from which we conclude $d_Y(f(x), f(x')) < \epsilon$ as desired. □

Example 7.26. In particular, this implies that any continuous function $f : [a, b] \to \mathbb{R}$ is uniformly continuous. However, on open intervals $f : (a, b) \to \mathbb{R}$ may be continuous and not uniformly continuous (such as $f(x) = 1/x$ on $(0, 1)$). ■

Theorem 7.27. *Let $f : X \to Y$ be a continuous function between two Cauchy complete metric spaces. If $K \subseteq X$ is compact, then $f(K)$ is also compact.*

Proof. Let $\mathcal{S}$ be an open cover of $f(K)$. Then the set $\mathcal{T} = \{f^{-1}(S) : S \in \mathcal{S}\}$ is a cover of K. The elements of $\mathcal{T}$ are open by Theorem 7.6 and Lemma 7.7. And hence $\mathcal{T}$ is an open cover of K. Let $\mathcal{T}'$ be the finite subcover guaranteed by Theorem 5.19; there exists a finite subset $\mathcal{S}'$ of $\mathcal{S}$ such that $\mathcal{T}' = \{f^{-1}(S) : S \in \mathcal{S}'\}$. Since $\mathcal{T}'$ covers K, we must also have $\mathcal{S}'$ covers $f(K)$. This shows that there exists a finite subcover of $f(K)$ given an open cover, and thus $f(K)$ is compact. □

An application of the above theorem is useful for constructing continuous inverses between spaces.

Theorem 7.28. *Let $f : K \to Y$ be continuous bijection, where K is compact and complete, and Y a Cauchy complete metric space. Then the inverse function f^{-1} is also continuous.*

Before proving the theorem, we introduce a useful lemma.

Lemma 7.29. *If K is a complete compact metric space, and $F \subseteq K$ is closed, then F is also compact.*

Proof. The result follows from Theorem 5.19. Since K is compact, then K is totally bounded. Any subset of a totally bounded set is also totally bounded, so F is totally bounded and closed, and thus F is also compact. □

Proof of Theorem 7.28. Since f is a bijection, it is invertible; for clarity denote the inverse function g. Furthermore, we can check that the induced power set mappings $g^{-1} : 2^K \to 2^Y$ and $f : 2^K \to 2^Y$ agree. So it suffices to show that the power set mapping sends closed sets to closed sets (by Theorem 7.6). Let $F \subseteq K$ be closed, then by the preceding Lemma, F is compact, and by Theorem 7.27, its image in Y is compact. By Theorem 5.19, this means that its image is closed. □

Another application of Theorem 7.27 is the Extremal Value Theorem.

Theorem 7.30 (Extremal Value Theorem)**.** *Let $f : K \to \mathbb{R}$ be continuous, where K is a compact subset of a metric space. Then $\sup f(K) \in f(K)$ and $\inf f(K) \in f(K)$. (In particular, this implies there exists $s \in K$ such that $f(s) = \sup f(K)$ and $t \in K$ such that $f(t) = \inf f(K)$; i.e. the extrema are attained.)*

Notice that this theorem would follow immediately from Theorem 7.27, provided we can prove that for any compact set $C \in \mathbb{R}$, the values $\sup C \in C$ and $\inf C \in C$.

Lemma 7.31. *For any compact $C \subseteq \mathbb{R}$, both $\sup C$ and $\inf C$ are elements of C.*

Proof. We include the proof for the case of the supremum; the infimum case is similar. Since compact sets are bounded by Theorem 5.19, we know that $\sup C$ exists.

Let $\mathbb{I}$ be the directed set of open intervals containing $\sup C$, ordered by reverse inclusion. By definition of $\sup C$, $C \cap I \neq \emptyset$ for every $I \in \mathbb{I}$. So there exists a net x taking values in C that converges to $\sup C$. But by definition of compactness, $\sup C$, being the unique accumulation point of x, must be in C also. □

Corollary 7.32. *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then the range of f is a closed interval.*

Exercise 7.16. Prove the Corollary.

§7.4 Examples and Counterexamples

It is easy to construct a function $f : \mathbb{R} \to \mathbb{R}$ that is nowhere continuous. Take $f = 1$ on $\mathbb{Q}$ and 0 otherwise. The density of $\mathbb{Q}$ in $\mathbb{R}$ shows that this function cannot be continuous.

Exercise 7.17. Prove the last assertion.

But can there be a function $f : \mathbb{R} \to \mathbb{R}$ that is continuous on a dense subset of $\mathbb{R}$, and discontinuous on also a dense subset of $\mathbb{R}$?

Example 7.33. Define the function

$$f(x) = \begin{cases} \frac{1}{b} & x \in \mathbb{Q} \wedge x = a/b \text{ in lowest terms;} \\ 0 & x \notin \mathbb{Q}. \end{cases}$$

I claim that this function is continuous at every irrational number, but discontinuous at every rational number.

The discontinuity at the rationals is easy to prove: let $x = a/b$ be a rational number in lowest terms; then $(0, 2/b)$ is an open interval containing $f(x)$. But every open interval I containing x contains at least one irrational number, and hence $f(I) \ni 0$ and thus $f(I) \not\subseteq (0, 2/b)$. By Theorem 7.5 then f is not continuous at x.

For the continuity at an irrational x, let J be an open interval around 0. Then there exists k such that $J \supseteq (-1/k, 1/k)$. Consider the set of rational numbers between $(x, x+1)$, whose lowest terms form has denominator $\leq k$. This set is finite and hence has a minimum q_+. Similarly, we find a rational number $q_- \in (x-1, x)$. In the open interval (q_-, q_+), by construction every rational number has a lowest-terms representation with denominator $> k$. And hence $f((q_-, q_+)) \subseteq [0, \frac{1}{k+1}] \subseteq J$. This shows continuity by Theorem 7.5. ■

A natural question to ask is whether the example can be reversed: does there exist a function that is continuous at every rational number, but discontinuous at every irrational number? The answer turns out to be no.

Lemma 7.34. *If $f : \mathbb{R} \to \mathbb{R}$ is a function, and let $C \subseteq \mathbb{R}$ denote the set of points at which f is continuous, then C can be written as the intersection of a countable family of open sets.*

Proof. Let $C_\epsilon := \{x \in \mathbb{R} : \exists I \ni x \text{ open}, f(I) \text{ has width } < \epsilon\}$. Notice that if $x \in C_\epsilon$, then by definition there exists $I \ni x$ such that $f(I)$ has width $< \epsilon$. But the same I also shows that every other $y \in I$ has the property, and hence $I \subseteq C_\epsilon$. Since I is an open interval, this means that C_ϵ is open.

By Theorem 7.5, the function f is continuous at x if and only if $x \in C_\epsilon$ for all $\epsilon > 0$. By the Archimedean property of the reals, this holds if and only if $x \in C_{1/k}$ for all $k \in \mathbb{N}$. And hence C is the intersection of a countable family of open sets. □

Quite clearly, the set of the irrational numbers is formed by $\cap\{\mathbb{R} \setminus \{q\} : q \in \mathbb{Q}\}$ and hence is the intersection of a countable family of open sets, and the previous example shows that there exists a function whose continuity set C is exactly the irrationals.

It turns out that the set of the rational number *cannot* be formed by such a procedure. The proof of this fact relies on the Baire Category Theorem which we likely won't have time to cover in this course.

Additionally, the converse of the Lemma above turns out to be also true; the construction is slightly technical, so we won't present it here to save space and time.

A common misconception among Calculus students is the assumption that a function $\mathbb{R} \to \mathbb{R}$ with the intermediate value property must be continuous.

Definition 7.35. *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be a* Darboux function *if for any $a, b \in \mathbb{R}$, for any γ between $f(a)$ and $f(b)$, there exists $c \in [a, b]$ such that $f(c) = \gamma$.*

The Intermediate Value Theorem states that "continuous implies Darboux". The converse turns out to be false. Many counterexamples exist, but a particularly spectacular one is the following.

Example 7.36 (Conway's base-13 function)**.** Define $f : \mathbb{R} \to \mathbb{R}$ via the following algorithm.

1. Expand x in base 13. We will use 🡐 to denote the "digit" equal to 10, 🡑 to be 11, and 🡘 to be 12. We will take the convention that repeating 🡘 at the end of the expansion is not allowed. (Like how instead of $0.\overline{999}$ we prefer 1 in base 10.) So the number
$$3🡘.2🡐5🡑 \to 39 + 12 + \frac{2}{13} + \frac{10}{13^2} + \frac{5}{13^3} + \frac{11}{13^4}.$$
2. Take the base 13 expansion of x, and remove the "decimal" point. Consider the resulting string.
   (a) If the symbols 🡘, 🡐, 🡑 appear infinitely many times, then $f(x) = 0$.
   (b) If the symbols appear finitely many times, and the last one to appear is 🡘, and
      i. the second to last to appear is 🡐, then throw away all leading "digits" before the final 🡐. Replace 🡐 $\to -$, and 🡘 $\to .$, and read the number as if in base 10. Example: if x in base 13 equals 1549🡐23🡐3154🡘9231542, then $f(x) = -3154.9231542$ in base 10.
      ii. the second to last to appear is 🡑, then do similarly as above, except replace 🡑 $\to +$.
   (c) If none of the above: $f(x) = 0$.

The amazing property is that for any non-empty open interval (a, b), $f((a, b)) = \mathbb{R}$. To see that: let $r \in \mathbb{R}$, expressed in decimal expansion as $\pm A.B$ where A is a finite string of decimal digits, and B is a possibly infinite string of decimal digits. By the Archimedean property, there exists a $p \in \mathbb{N}$ and $m \in \mathbb{Z}$ such that $(m/13^p, (m+1)/13^p) \subseteq (a, b)$. Since $m/13^p$ has a terminating base-13 expansion (at p "digits" after the decimal point), we can append to it first either 🡐 or 🡑 depending on whether the real number r is positive or negative, and then the string A, followed by 🡘, followed by the string B. This new base-13 number is guaranteed to be in the interval (a, b), and applying f to it, using the algorithm above, yields r.

Since $f((a, b)) = \mathbb{R}$ for any (a, b), we see that f is trivially a Darboux function. But the function is discontinuous on $\mathbb{R}$ in the strongest sense possible. ■

We conclude by giving a sufficient condition for a Darboux function to be continuous.

Theorem 7.37. *If $f : \mathbb{R} \to \mathbb{R}$ is a Darboux function, and if there exists a dense subset $D \subseteq \mathbb{R}$ such that for every $y \in D$, the set $f^{-1}(\{y\})$ is closed, then f is continuous.*

Proof. First, notice that if f is Darboux, then for any interval I, $f(I)$ is also an interval.

We argue by contradiction. Suppose that f is not continuous. Then there exists a point $x_0 \in \mathbb{R}$ and a net $\mu \to x_0$ such that $\limsup f \circ \mu > \liminf f \circ \mu$. (See also Exercise 7.10.) Therefore for every open interval I around x_0, $f(I) \supset (\liminf f \circ \mu, \limsup f \circ \mu)$. Since the latter is a non-degenerate open interval, and D is dense, there exists distinct $y_1, y_2 \in D$, such that for any open interval I containing x_0, the values $y_1, y_2 \in f(I)$.

Let $\mathbb{I}_{x_0}$ be the directed set of all open intervals around x_0. The above construction means we can construct two nets, both from $\mathbb{I}_{x_0} \to \mathbb{R}$, such that $f \circ \nu_1$ is the constant net y_1 and $f \circ \nu_2$ is the constant net y_2. In other words, we can construct ν_1, ν_2 such that ν_1 takes values only in $f^{-1}(\{y_1\})$ and ν_2 takes values only in $f^{-1}(\{y_2\})$. But since x_0 is the limit of both ν_1 and ν_2, and it cannot belong to both $f^{-1}(\{y_1\})$ and $f^{-1}(\{y_2\})$, at least one of the two sets must not be closed. This contradicts the definition of D. □

Exercise Sheet: Week 7

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 7.1. Let (X, d) be a metric space. Fix $x_0 \in X$. Let the function $f : X \to \mathbb{R}$ be defined by

$$f(x) = d(x_0, x).$$

Prove that $f(x)$ is uniformly continuous.

Question 7.2. A metric space (X, d) is said to be *totally disconnected* if for every $x \in X$, there exists $r > 0$ such that $B(x, r) = \{x\}$.

1. Prove that if (X, d_X) is totally disconnected, and (Y, d_Y) is any metric space, then any $f : X \to Y$ is continuous.
2. Prove that if (X, d_X) is a metric space, and (Y, d_Y) is a totally disconnected metric space, then any continuous function $f : X \to Y$ is constant.

Question 7.3. Use the previous exercise to produce a counterexample to Theorem 7.28, if we drop the requirement that K is compact and Y is complete: prove that $\mathbb{N}$ with the absolute value of difference metric is a totally disconnected metric space, and that $\mathbb{Q}$ with the absolute value of difference metric is an incomplete metric space. Then, prove that any enumeration of $\mathbb{Q}$ gives a continuous mapping from $\mathbb{N} \to \mathbb{Q}$, whose inverse is not continuous.

Question 7.4. Prove that if $f : [a, b] \to \mathbb{R}$ is an increasing function, then the points at which f is *discontinuous* is countable.

(*Hint: notice that given $c \in (a, b)$, then $f : (a, c) \to \mathbb{R}$ can be regarded as a monotone net bounded above, and hence the limit of f from below exists. Similarly the limit of f from above exists. Thus for f to be discontinuous at c, we must have that the left limit is strictly less than the right limit.*)

Question 7.5. We say a real valued net x *converges to* $+\infty$ if it is eventually in $\uparrow(z)$ for every $z \in \mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ (the codomain can also be an arbitrary metric space, but let's keep it simple). Prove:

1. Suppose f is such that $f \circ x$ converges for every real-valued net converging to $+\infty$, then the limit is unique. (See proof of Proposition 7.11.)
2. Let μ be the identity net $\mathbb{R} \to \mathbb{R}$. Prove that if $f \circ \mu$ converges, then $f \circ x$ converges for every real-valued net x converging to $+\infty$.

This allows us to define what it means for f to have a *horizontal asymptote*.

Problem Set 7

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 7.1. Let (X, d) be a Cauchy complete metric space, and fix $K \subseteq X$ a compact subset.

1. Let $y \in X$, prove that there exists a *closest point to* y in K. More precisely, prove that there exists $z \in K$ such that for any other $z' \in K$, the distances $d(z, y) \leq d(z', y)$.
2. Given an example showing that the closest point may be non-unique: namely, give metric space X, a compact subset K, and a point y, such that there exists two distinct closest points $z, z' \in K$. Use this example to briefly argue why it is in general not possible to find a continuous function $c : X \to K$ such that $c(y)$ is a closest point to y in K.
3. Prove, nonetheless, that the function

$$\rho_K : X \to \mathbb{R}, \quad \rho_K(y) = d(z, y) \text{ where } z \text{ is a closest point in } K \text{ to } y$$

   is uniformly continuous.

The function ρ_K is usually called "the distance to K"; it can be equivalently defined by

$$\rho_K(y) = \inf\{d(z, y) : z \in K\}.$$

Problem 7.2. Let $f : [a, b] \to [c, d]$ be a surjective function that sends $f(a) = c$ and $f(b) = d$. Prove that f is injective and continuous if and only if f is strictly increasing.

The following statement may be useful for the next Problem: if $\mathcal{C}$ is a family of closed subsets of a metric space, then $\cap\mathcal{C}$ is also closed. To prove that this is true: observe that if $\mathcal{O}$ is the family of the complements of the members of $\mathcal{C}$, then elements of $\mathcal{O}$ are open. Furthermore, $\cap\mathcal{C}$ is equal to the complement of $\cup\mathcal{O}$. And if $x \in \cup\mathcal{O}$, it belongs to some open set $O \in \mathcal{O}$, but then a ball around x also belongs to O, and so also to $\cup\mathcal{O}$. Hence $\cup\mathcal{O}$ is open.

Problem 7.3. Let $\mathcal{F}$ be a family of continuous functions from $[0, 1] \to \mathbb{R}$. Let f be the pointwise infimum of the family $\mathcal{F}$ (see Problem 2.1 on the Week 2 Problem Set; namely $f(x) = \inf\{g(x) : g \in \mathcal{F}\}$). Prove that

1. f is upper semi-continuous. (*Hint: it is easier to work with the descriptions of Theorem 7.13 of semi-continuity, and of Theorem 7.6 for continuity.*)
2. when $\mathcal{F}$ is finite, f is continuous. (*Hint: it is enough to prove, in view of the first part, that in this setting, f is also lower semi-continuous.*)

For the next question, let's introduce some definitions. First, let's generalize the notion of a canonical net to metric spaces. Let (X, d) be a metric space, and $z \in X$. Consider the set $\mathcal{B}$ of ordered pairs (y, B) where B is an open ball centered at z and $y \in B$. Order $\mathcal{B}$ by $(y_1, B_1) \leqq (y_2, B_2) \iff B_1 \supseteq B_2$. Let $\beta : \mathcal{B} \to X$ be given by $\beta((y, B)) = y$. In exactly the same way as Exercise 7.1, $\mathcal{B}$ is a directed set, and β is a net that converges to z. This is the canonical net in X that converges to z.

> **Lemma**. Let $\mu : A \to X$ be a net that converges to z, then there exists a net ν that is both a subnet of μ and a subnet of β, where β is the canonical net.

(The proof of the Lemma is not important for this Problem Set, you can answer the Problem below by using this lemma as a "black box"; I include the proof here for completeness.)

Proof. The proof is very similar to the proof of Theorem 6.9. Let N be the set of ordered pairs (α, B) such that B is an open ball centered at z and $\alpha \in A$ is such that $\mu_\alpha \in B$. Order N with

$$(\alpha_1, B_1) \leqq (\alpha_2, B_2) \iff B_1 \supseteq B_2 \wedge \alpha_1 \leqq \alpha_2.$$

As shown in the proof of Theorem 6.9, this N is a directed set.

Let $\nu : N \to X$ be given by $\nu((\alpha, B)) = \mu_\alpha$. I claim that this ν is a subnet of both μ and β.

To see that ν is a subnet of μ: we see that $\nu = \mu \circ \varphi$ where $\varphi : N \to A$ is given by $\varphi((\alpha, B)) = \alpha$. This mapping is increasing by definition. For any $\alpha \in A$, since $\mu \to z$, then for any fixed ball B around z, $\uparrow(\alpha) \cap B$ is non-empty, and hence there exists $\alpha' \in \uparrow(\alpha) \cap B$. Then $\varphi((\alpha', B)) \geqq \alpha$. This shows that ν is a subnet of μ

To see that ν is a subnet of β: we see that $\nu = \beta \circ \psi$, where $\psi : N \to \mathcal{B}$ is given by $\psi((\alpha, B)) = (\mu_\alpha, B)$. This function is increasing. For any $(y, B) \in \mathcal{B}$, since $\mu \to z$, there exists α such that $\mu_\alpha \in B$. Thus for this α, we have $\psi((\alpha, B) = (\mu_\alpha, B) \geqq_{\mathcal{B}} (y, B)$. And thus ν is a subnet of β. □

Problem 7.4. Using the Lemma above, prove the following:

> If (X, d_X) and (Y, d_Y) are metric spaces and $f : X \to Y$ a function, then f is continuous at a point $z \in X$ if and only if for the canonical net β that converges to z, the net $f \circ \beta$ converges.

Hints: One direction is trivial. The other direction you need to show that for every net $x \to z$, $\lim f \circ x$ exists and equals $f(z)$. First notice that since the constant net z is a subnet of β, if $f \circ \beta$ converges it has to converge to $f(z)$. Using the Lemma, prove first that every net $x \to z$ has a subnet y such that $\lim f \circ y = f(z)$. Then upgrade this to show that every net $x \to z$ is such that $\lim f \circ x = f(z)$.

Remark: this problem shows that something analogous to part 1 of Exercise 7.9 also hold for functions between metric spaces.

Reading Assignment 8

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

In this set of readings we go over the classical theorems related to differentiability of functions. We start by defining a notion of tangency between functions between metric spaces. Specializing the functions $\mathbb{R} \to \mathbb{R}$, we define differentiability in terms of "being tangent to an affine function". Some basic properties of derivatives are described, including the Leibniz and chain rules. After this introduction, we devote some significant amount of space to the Mean Value Theorem and its cousins. While the Mean Value Theorem itself is thoroughly a "one-dimensional theorem" and doesn't have higher dimensional analogues, some of the related facts do have nice generalizations: Fermat's Lemma and Rolle's Lemma generalize nicely (a particularly powerful version of the latter is the famous Mountain-Pass Theorem of functional analysis); and the Inverse Function Theorem is an important tool in higher dimensional analysis. Another nice application of the Mean Value Theorem (specifically of the version due to Cauchy) is L'Hôpital's rule for taking limits of indeterminate forms. We conclude these notes with a brief discussion of higher order derivatives, specifically the existence of the Taylor polynomial and its approximation properties, and the Taylor remainder theorem. For simplicity we limit our discussion to the case of the second derivative; but the higher derivative cases follow by induction.

Contents

§8.1 Detour: exponential functions

A very useful concept in analysis is the ability to raise an arbitrary non-negative real number x to an arbitrary real power α. In this section we go over quickly how this quantity is defined.

First, given $x \geq 0$, we can define x^n for any $n \in \mathbb{N}$ by repeated multiplication, and using that $\mathbb{R}$ is a field, and hence we have defined the function $f_n : \uparrow(0) \to \mathbb{R}$ that sends $x \mapsto x^n$. Using the ordered field

properties, we observe that this function is strictly increasing. It is not too hard to prove that f_n is continuous, and restricted to any compact interval $[0,a]$, the function f_n is a bijection $[0,a] \to [0,a^n]$. Hence by Theorem 7.28, the function f_n has a continuous inverse; we call this function $f_{1/n}$ the "n-th root" function.

This allows us to construct x^q for any $q \in \mathbb{Q}$, by first taking the root corresponding to the denominator, and then the power with respect to the numerator. Next one can show, by taking common denominators, that for any fixed $x > 0$, the mapping $q \mapsto x^q$ is monotonic: when $x = 1$, the mapping is constant. When $x < 1$ the mapping is strictly decreasing in q, and when $x > 1$ the mapping is strictly increasing in q. A bit of tedious technical work allows us to show that on bounded subsets, for fixed $x > 0$, the function $q \mapsto x^q$ is uniformly continuous on $\mathbb{Q}$. And hence we can extend it using Theorem 7.22 to a function that takes real valued arguments.

§8.2 Differentiability

§8.2.1 Tangency.—Before we talk about differentiability of functions, let's talk about a slightly more general concept that is applicable to general metric spaces. This will help us see exactly which of the properties of the real numbers $\mathbb{R}$ is needed to go beyond this general notion of tangency, to a notion of differentiability.[1] First, we introduce a notational convenience.

Definition 8.1. *Let (X,d) be a metric space, $x \in X$ a point, and $\alpha > 0$ a real number. We say that a function $f : X \to \mathbb{R}$ is in the set $\mathfrak{o}(x,\alpha)$ if, for every $\epsilon > 0$, there exists $r > 0$ such that restricted to $B(x,r)$, the function f satisfies $|f(y)| \leq \epsilon d(x,y)^\alpha$.*

Roughly speaking, this says that a function f is in the class $\mathfrak{o}(x,\alpha)$ if, as we approach x, the value of the function converges to 0 faster than $d(x,y)^\alpha$. More concretely, we have:

Proposition 8.2. *If $f \in \mathfrak{o}(x,\alpha)$, then for every net $\mu \to x$ taking values in $X \setminus \{x\}$, we have that* $\lim f \circ \mu \cdot d(x,\mu)^{-\alpha} = 0$.

Proof. Since $f \in \mathfrak{o}(x,\alpha)$, given any $\epsilon > 0$, there exists $r > 0$ such that restricted to $B(x,r)$ the quantity $|f(y)d(x,y)^{-\alpha}| < \epsilon$. Since $\mu \to x$, we have that μ is eventually in $B(x,r)$, and hence $f \circ \mu \cdot d(x,\mu)^{-\alpha}$ is eventually less than ϵ. This shows convergence. □

Exercise 8.1. Verify that if $f : X \to \mathbb{R}$ is in $\mathfrak{o}(x,\alpha)$ for some $\alpha > 0$, then f is continuous at x.

Exercise 8.2. Prove that:

1. If $f \in \mathfrak{o}(x,\alpha)$, then $f \in \mathfrak{o}(x,\beta)$ for all $\beta \in (0,\alpha]$.
2. If $f, g \in \mathfrak{o}(x,\alpha)$, then $f + g \in \mathfrak{o}(x,\alpha)$.
3. If $f \in \mathfrak{o}(x,\alpha)$ and $g \in \mathfrak{o}(x,\beta)$, then $f \cdot g \in \mathfrak{o}(x,\alpha+\beta)$.
4. If $f \in \mathfrak{o}(x,\alpha)$, and g is bounded on $B(x,r)$ for some $r > 0$, then $f \cdot g \in \mathfrak{o}(x,\alpha)$.

Exercise 8.3. For $\alpha > 0$, let $f_\alpha : \mathbb{R} \to \mathbb{R}$ be the function $f_\alpha(x) = |x|^\alpha$. Prove that $f_\alpha \in \mathfrak{o}(0,\beta)$ for all $\beta \in (0,\alpha)$; and $f_\alpha \notin \mathfrak{o}(0,\gamma)$ for any $\gamma \geq \alpha$.

Example 8.3. In Exercise 8.1 we saw that any $\mathfrak{o}(x,\alpha)$ function is continuous. Is the converse true? Namely: is every continuous function at a point x necessarily in $\mathfrak{o}(x,\alpha)$ for some $\alpha > 0$?

[1] This approach has also been used by Cheeger to develop a notion of "derivatives" for more general metric spaces that do not admit an affine structure.

The answer is *no*. One example is the function $g(x) = -1/\ln(|x|)$ when $x \neq 0$, and $g(0) = 0$. Assuming certain things that we haven't proven rigorously yet (how is $\ln(x)$ defined, and L'Hôpital's rule), it can be seen that g is continuous at 0, but $g(x)$ is not in $\mathfrak{o}(0,\alpha)$ for any $\alpha > 0$.

The function $g_\alpha : \mathbb{R} \to \mathbb{R}$ defined by $g_\alpha(x) = |x|^\alpha \cdot g(x)$ can also be seen as a function that is in $\mathfrak{o}(0,\alpha)$, but not in $\mathfrak{o}(0,\gamma)$ for any $\gamma > \alpha$. These are important families of examples in analysis: it shows that attempts to classify *asymptotic behavior* using "power law" type functions are not sufficiently sensitive to capture "logarithmic corrections". Another similar example you have previously encountered in Calculus II is the convergence of the improper integral $\int_1^\infty |x|^{-p}\, dx$; it is know that the threshold in the power law realm is that convergence happens if and only if $p > 1$. But if we refine the family be considering integrals of type $\int_2^\infty x^{-p}\ln(x)^{-q}\, dx$, we see that when $p = 1$ there is some subtlety with respect to the q. The analysis can be further extended by thinking about integrals of functions of type $x^{-p}\ln(x)^{-q}\ln(\ln(x))^{-r}$ and so on. ■

What is the idea behind tangency? Think back to functions from $\mathbb{R} \to \mathbb{R}$; if the graphs of two functions were to intersect transversely, then at the point of intersection we expect the values of the function to grow linearly. And if the graphs of two functions were to touch but not cross, we expect the rate at which the two curves separate, near the point of intersection, to be negligible.

We capture this with the following definition.

Definition 8.4. *Let (X, d_X) and (Y, d_Y) be metric spaces. Two functions f and g, both from X to Y, are said to be* tangent at $x_0 \in X$, *if the function $X \ni x \mapsto d_Y(f(x), g(x)) \in \mathbb{R}$ is an element of $\mathfrak{o}(x_0, 1)$.*

Example 8.5. Any function $f : X \to \mathbb{R}$ that is in $\mathfrak{o}(x_0, 1)$ is tangent at x_0 to the constant function $g(x) = 0$.

More generally, if $f : X \to Y$ is such that the function $x \mapsto d_Y(f(x), f(x_0))$ is in $\mathfrak{o}(x_0, 1)$, then f is tangent to the constant function $g(x) = f(x_0)$ at x_0. ■

Exercise 8.4. Prove that tangency is an equivalence relation.

Exercise 8.5. Prove that if x_0 is an isolated point[2] then *any two functions f* and g from X to Y which agree at x_0 are tangent at x_0

Food for Thought 8.6. Notice that we did not make any assumptions on the functions f, g in Definition 8.4. Ask yourself, among functions $\mathbb{R} \to \mathbb{R}$: which are tangent to the following functions?

1) $x \mapsto |x|$ 2) $x \mapsto \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ 3) $x \mapsto \begin{cases} 1/x & x \neq 0 \\ 0 & x = 0 \end{cases}$

§8.2.2 Definition of Differentiability.—What allows us to define differentiability on $\mathbb{R}$ but not on general metric spaces? The answer is that, while on a general metric space, we can compare two arbitrary functions and decide whether they are tangent, on $\mathbb{R}$, we can compare any function to those drawn from a special family: the *affine functions*.[3]

Definition 8.6. *The set of* affine functions *on* $\mathbb{R}$, *denoted*[4] Aff *is the set of all functions of the form*

[2] A point x_0 in a metric space X is said to be isolated if there exists $r > 0$ such that $B(x, r) = \{x\}$.

[3] In prior courses you probably have used the term "linear" to describe such functions. However, when you work on functions between general vector spaces, a "linear" function must send the origin to itself, while an "affine" function can also have a translation. So we will use the affine terminology to fit with the more general settings.

[4] Normally I would put some information about the domain and codomain of the functions in the notation; but in this course we will only be treating affine functions from $\mathbb{R}$ to $\mathbb{R}$ so we can suppress the notation.

$x \mapsto m \cdot x + b$ *for* $m, b \in \mathbb{R}$. *The number* m *we refer to as the* slope *of the affine function.*

Exercise 8.7. Verify that a function $f \in \mathsf{Aff}$ if and only if there exists $m \in \mathbb{R}$ such that for every $x, x' \in \mathbb{R}$, the difference $f(x) - f(x') = m(x - x')$.

The graphs of affine functions being given by straight lines, we see that the following definition essentially says that a function is differentiable at a point, if it has a "tangent line" at that point.

Definition 8.7. *Let* $S \subseteq \mathbb{R}$. *A function* $f : S \to \mathbb{R}$ *is said to be* differentiable at the point $x_0 \in S$ *if it is tangent at* x_0 *to (the restriction to* S *of) an element of* Aff. *(In other words, there exists* $\ell \in \mathsf{Aff}$ *such that* $f - \ell \in \mathfrak{o}(x_0, 1)$.*)*

In this definition, we are fairly general about what the set S can be. It turns out the notion of differentiability is not very useful if x_0 is an isolated point of S. By Exercise 8.5 if x_0 is isolated, then any element $\ell \in \mathsf{Aff}$ with $\ell(x_0) = f(x_0)$ is tangent to f.

Proposition 8.8. *If* $x_0 \in S$ *is* not *isolated, then at most one element of* Aff *can be tangent to any given function* f *at* x_0.

Proof. By Exercise 8.4 tangency is an equivalence relation. Thus it suffices to show that when x_0 is not isolated, and ℓ_1, ℓ_2 two distinct elements of Aff, then they are not tangent at x_0. We may write elements of Aff in point-slope form, so $\ell_i = m_i(x - x_0) + b_i$. If $b_1 \neq b_2$, then $\ell_1(x_0) - \ell_2(x_0) = b_1 - b_2 \neq 0$ and so the difference is not in $\mathfrak{o}(x_0, 1)$. Supposing $b_1 = b_2$ and $m_1 \neq m_2$, then $\ell_1(x) - \ell_2(x) = (m_1 - m_2)(x - x_0)$. Since x_0 is not isolated, for every $r > 0$ there exists $x_r \in (S \cap B(x_0, r)) \setminus \{x_0\}$. Thus if we take $\epsilon < |m_1 - m_2|$, we see that for no values of r is the restriction of $\ell_1 - \ell_2$ to $S \cap B(x_0, r)$ bounded by $\epsilon|x - x_0|$, and hence $\ell_1 - \ell_2 \notin \mathfrak{o}(x_0, 1)$. □

Definition 8.9. *Let* $S \subseteq \mathbb{R}$. *Suppose* x_0 *is not an isolated point and* $f : S \to \mathbb{R}$ *is differentiable at* x_0. *By* the derivative of f at x_0 *we refer to the slope* m *of the unique element of* Aff *that is tangent to* f *at* x_0.

Furthermore, if S *contains no isolated points, and* $f : S \to \mathbb{R}$ *is differentiable at every point of* S, *by* the derivative of f *we refer to the function* $f' : S \to \mathbb{R}$ *such that* $f'(x)$ *is the slope of the unique element of* Aff *that is tangent to* f *at* x.

Example 8.10. By definition, any function $f \in \mathsf{Aff}$ is differentiable at every point in $\mathbb{R}$. If $f(x) = mx + b$, then $f'(x) = m$. ■

Exercise 8.8. Let $S \subseteq \mathbb{R}$, and suppose $f : S \to \mathbb{R}$ is differentiable at a point $x_0 \in S$. Prove that $f : S \to \mathbb{R}$ is continuous at x_0. (*Hint: treat the case* x_0 *is isolated separately from the case* x_0 *is not isolated.*)

Exercise 8.9. Prove, using Definition 8.7, that the following functions $\mathbb{R} \to \mathbb{R}$ are *not* differentiable at the origin. (Do not use calculus facts such as the derivative of sin or the chain rule.)

1) $x \mapsto |x|$ 2) $x \mapsto \begin{cases} x \sin(\frac{1}{x}) & x \neq 0 \\ 0 & x = 0 \end{cases}$

Exercise 8.10. In this exercise we will consider functions of the form $g(x) = |x|^{1+f(x)}$ when $x \neq 0$, and $g(0) = 0$. Determine whether $g(x)$ is differentiable at 0 for each of the given $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ below. (You don't need to give a formal proof: try to use all the tools you have learned in Calculus classes to help

you decide.)

1) $f(x) = 1/|x|$ 2) $f(x) = |x|$ 3) $f(x) = x^2$
4) $f(x) = \ln(|x|)$ 5) $f(x) = 1/\ln(|x|)$ 6) $f(x) = 1/\ln(|\ln(|x|)|)$

Theorem 8.11. *Let $S \subseteq \mathbb{R}$ and f, g are functions $S \to \mathbb{R}$. Suppose $x_0 \in S$ is not isolated, and f and g are both differentiable at x_0 with derivatives m_f and m_g respectively. Then:*

1. *The function $h = f + g$ is differentiable at x_0 with derivative $m_f + m_g$.*
2. *The function $h = f \cdot g$ is differentiable at x_0 with derivative $m_f g(x_0) + f(x_0)m_g$.* [Leibniz Rule]

Proof. Let $\ell_f(x) = m_f(x-x_0)+f(x_0)$ and $\ell_g(x) = m_g(x-x_0)+g(x_0)$. The hypothesis is equivalent to $f-\ell_f$ and $g-\ell_g$ both being functions in $\mathfrak{o}(x_0, 1)$. Hence by Exercise 8.2, we have $f+g-(\ell_f+\ell_g) \in \mathfrak{o}(x_0, 1)$ and the result for sum follows.

For the product, we have the following identity:

$$fg - [m_f g(x_0) + f(x_0)m_g](x-x_0) - f(x_0)g(x_0) = \underbrace{(f-\ell_f)g}_{\in \mathfrak{o}(x_0,1)} + \underbrace{(g-\ell_g)\ell_f}_{\in \mathfrak{o}(x_0,1)} + \underbrace{m_f m_g (x-x_0)^2}_{\in \mathfrak{o}(x_0,1)}.$$

Since g and ℓ_f are continuous at x_0, they are bounded in a small ball around x_0. And hence by Exercise 8.2 we conclude the entire right hand side is in $\mathfrak{o}(x_0, 1)$ and the theorem follows. □

Exercise 8.11. Suppose $x_0 \in S$ is not isolated, and $f : S \to \mathbb{R}$ is differentiable at x_0 with derivative m. Assume $f(x) \neq 0$ for all $x \in S$. Prove that the function $\frac{1}{f}$ is also differentiable at x_0 with derivative $-m/f(x_0)^2$. (*Hint: show that the quantity $f(x_0)^2 - (f(x) - m(x-x_0))f(x)$ is $\mathfrak{o}(x_0, 1)$.*)

Theorem 8.12 (Chain rule)**.** *Let $S \subseteq \mathbb{R}$ with x_0 a non-isolated point in S. Suppose $f : S \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is such that f is differentiable at x_0 with derivative m_f; and g is differentiable at $y_0 := f(x_0)$ with derivative m_g. Then $g \circ f$ is differentiable at x_0 with derivative $m_g m_f$.*

Proof. Let $\ell_f(x) = m_f(x - x_0) + y_0$, and $\ell_g(y) = m_g(y - y_0) + g(y_0)$. Then $\ell_g \circ \ell_f = m_g m_f (x - x_0) + g(y_0)$. We compute

$$g \circ f - \ell_g \circ \ell_f = (g - \ell_g) \circ f + \ell_g \circ f - \ell_g \circ \ell_f = (g - \ell_g) \circ f + m_g(f - \ell_f).$$

The second term is clearly in $\mathfrak{o}(x_0, 1)$. For the first term, let $\epsilon > 0$, then by Definition 8.1 there exists r_y such that $|g(y) - \ell_g(y)| < \frac{\epsilon}{2m_f}|y - y_0|$ for all $y \in B(y_0, r_y)$. On the other hand, there exists r_x such that $|f(x)-\ell_f(x)| < m_f|x-x_0|$ for all $x \in B(x_0, r_x)$. This implies that in $B(x_0, r_x)$ we have $|f(x)-y_0| < 2m_f|x-x_0|$. And hence, with $r'_x = \min\{r_x, r_y/2m_f\}$, we have that in $B(x_0, r'_x)$

$$|g(f(x)) - \ell_g(f(x))| < \frac{\epsilon}{2m_f}|f(x) - y_0| < \epsilon|x - x_0|.$$

And this shows that $(g - \ell_g) \circ f$ is $\mathfrak{o}(x_0, 1)$. □

An immediate consequence of Theorems 8.11, 8.12, Exercise 8.11, and Example 8.10 above is that polynomial functions and rational functions are differentiable at every point of their domains.

In Example 8.10, we saw that any function $f \in \mathsf{Aff}$ with $f(x) = mx + b$ is differentiable on $\mathbb{R}$ with $f'(x) \equiv m$. The following gives the converse.

Proposition 8.13. *Suppose $f : [a,b] \to \mathbb{R}$ is differentiable on $[a,b]$ with $f'(x) \equiv m$. Then $f \in \mathsf{Aff}$.*

Proof. First we reduce to the case $m = 0$. By Theorem 8.11, if we let $g(x) = f(x) - mx$, then g is also differentiable at every point in $[a,b]$ with derivative identically 0. Since Aff is closed under addition of elements, if we can prove that $g \in \mathsf{Aff}$ then so must f.

We will in fact prove that g is the constant function. It is enough to show $g(a) = g(b)$ under our hypotheses; for we can then apply the result to the subinterval $[a,b']$ for any $b' \in (a,b]$ to get the constancy.

To show $g(a) = g(b)$: since g is differentiable at every point with derivative 0, this means that for every $\epsilon > 0$ and for every $x \in [a,b]$ there exists some r_x such that, for every $y \in B(x,r_x) \cap [a,b]$, we have $|g(y) - g(x)| < \epsilon|y - x|$. Consider the set $\mathscr{S} = \{B(x, \frac{1}{2}r_x) : x \in [a,b]\}$, this is an open cover of the compact set $[a,b]$ and hence has a finite subcover centered at points $\{x_1, \ldots, x_K\}$; we enumerate the points x_i to be strictly increasing. Notice that $|x_i - x_{i+1}| < \frac{1}{2}(r_{x_i} + r_{x_{i+1}})$; otherwise the halfway point between them will not be covered by the open balls. And hence either $x_i \in B(x_{i+1}, r_{x_{i+1}})$ or vice versa. By construction we also must have that $a \in B(x_1, r_{x_1})$ and $b \in B(x_K, r_{x_K})$. Thus we can write

$$g(b) - g(a) = g(b) - g(x_K) + g(x_K) - g(x_{K-1}) + \cdots + g(x_2) - g(x_1) + g(x_1) - g(a).$$

Each paired difference is between the values of g evaluated at points belonging to the same $B(x_i, r_{x_i})$ ball, with one of the points being the center. And hence we have

$$|g(b) - g(a)| < \epsilon|b - x_K| + \epsilon|x_K - x_{K-1}| + \cdots + \epsilon|x_2 - x_1| + \epsilon|x_1 - a|.$$

Since $a \le x_1 < x_2 < \ldots < x_K \le b$, we have that the sums collapse and we have $|g(b) - g(a)| < \epsilon|b - a|$. Since ϵ is arbitrary, we conclude that $|g(b) - g(a)| = 0$. □

Food for Thought 8.12. By considering closed subintervals, the previous proposition can be seen to also hold for *all* intervals, not just closed and bounded ones. Can the proposition be generalized to sets that are *not* intervals? Explain how, or explain why not.

Exercise 8.13. Prove that if f, g are functions from an interval $[a,b] \to \mathbb{R}$, both differentiable on $[a,b]$, such that $f' = g'$, then f and g differs by a constant.

Definition 8.14. *Let $S \subseteq \mathbb{R}$ be a set with no isolated points. A function $f : S \to \mathbb{R}$ is said to be* continuously differentiable *if f is differentiable at every point in S, and the function $f' : S \to \mathbb{R}$ is continuous. This property is sometimes written $f \in \mathcal{C}^1(S;\mathbb{R})$.*

Food for Thought 8.14. In your previous calculus classes, the distinction between a function f being merely differentiable on every point of S, versus it being *continuously* differentiable, may not have been emphasized. Observe that most of the theorems in these readings apply to differentiable functions, without requiring the derivative be continuous.

Example 8.15. For a standard example of a function that is differentiable on every point of its domain, but not continuously so, consider $f : (-1,1) \to \mathbb{R}$ given by $f(x) = x^2 \sin(\frac{1}{x})$. First we check that f is differentiable at 0. In fact, since $|\sin(a)| \le 1$ we have that $f \in \mathfrak{o}(0,1)$. This implies that f is differentiable at 0 with derivative 0.

Taking for granted the "calculus fact" that $\sin'(x) = \cos(x)$; our Theorems 8.11 and 8.12 allows us to compute the derivative of f away from the origin:

$$f'(x)|_{x \neq 0} = 2x \sin \frac{1}{x} - \cos \frac{1}{x}.$$

This shows that away from 0, the derivative f' is continuous. However it is pretty easy to see that for the identity net $\mu : (-1, 0) \to (-1, 0)$, we have $\limsup(f') \circ \mu = 1$ and $\liminf(f') \circ \mu = -1$, so f' cannot be continuous at 0.

This example is a good one to have in your head why trying to ask whether properties should hold for all differentiable functions, or just for continuously differentiable functions. ■

§8.3 Mean Value Theorems and Applications

§8.3.1 Mean value theorems.—The mean value theorems are a family of theorems that allow us to compare the value of the derivative of a function on an interval, with the values of the functions on the endpoints. The heavy lifting is done by the following lemma of Fermat.

Lemma 8.16 (Fermat's stationary point lemma)**.** *Let $f : (a, b) \to \mathbb{R}$ be such that (a) f is differentiable at some $c \in (a, b)$ and (b) $f(c) \geq f(x)$ for any $x \in (a, b)$. Then f has derivative 0 at c.*

Proof. By assumption f is differentiable at c, so $f - \ell_f \in \mathfrak{o}(c, 1)$ for some $\ell_f = m(x - c) + f(c)$. It suffices to show that the slope m must necessarily be zero. Suppose $m \neq 0$, let $\epsilon = \frac{1}{2}|m|$, then there exists $r > 0$ such that $|f(x) - \ell_f(x)| \leq \epsilon|x - c|$ for all $x \in B(c, r)$ and $B(c, r) \subseteq (a, b)$. This implies, on $B(c, r)$, we have

$$m(x - c) - \frac{1}{2}|m(x - c)| \leq f(x) - f(c) \leq m(x - c) + \frac{1}{2}|m(x - c)|.$$

Suppose $m > 0$, then this means $f(c + r/2) - f(c) \geq \frac{1}{4}mr > 0$, contradicting the fact that $f(c)$ is the maximum value. Suppose $m < 0$, then this means $f(c - r/2) - f(c) \geq \frac{1}{4}|m|r > 0$, again contradicting the fact that $f(c)$ is the maximum. Hence $m = 0$. □

Notice that Fermat's lemma applies also to local minima, by replacing f with $-f$.

Exercise 8.15. Prove the following variations of Fermat's lemma:

1. If $f : (a, c] \to \mathbb{R}$ is differentiable at c and attains its maximum (minimum) there, then the derivative at c is non-negative (non-positive).
2. If $f : [c, b) \to \mathbb{R}$ is differentiable at c and attains its maximum (minimum) there, then the derivative at c is non-positive (non-negative).

Fermat's lemma implies the basic form of a mean value theorem, which states that if f is constant on the "boundary", then there must be an interior critical point.

Corollary 8.17 (Rolle's lemma)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous, and suppose f is differentiable on (a, b). Then: if $f(a) = f(b)$ then there exists $c \in (a, b)$ with $f'(c) = 0$.*

Proof. By Corollary 7.32, $f([a, b]) = [c, d]$ a closed interval. If $c = d$, then necessarily f is the constant function and $f'(x) = 0$ for all $f \in (a, b)$. Suppose $c \neq d$, then at least one of the two is different from $f(a)$. And hence there exists $p \in (a, b)$ where $f(p)$ attains an extremum value. By Fermat's lemma $f'(p) = 0$. □

Food for Thought 8.16. Which properties of continuous functions did we use in proving this lemma? Obviously to identify the extremum point within the interval we used that $[a, b]$ is compact. But do we need that $[a, b]$ is connected? Why or why not?

Exercise 8.17. Prove that if f, g are two functions that are both continuous on $[a,b]$ and differentiable on (a,b), such that $f(a) = g(a)$ and $f(b) = g(b)$, then there exists $c \in (a,b)$ such that $f'(c) = g'(c)$.

Theorem 8.18 (Cauchy's Mean Value Theorem)**.** *Let f, g be functions from $[a,b] \to \mathbb{R}$; suppose they are both continuous on $[a,b]$ and differentiable on (a,b). Then there exists $c \in (a,b)$ such that*

$$f'(c)[g(a) - g(b)] = g'(c)[f(a) - f(b)].$$

Proof. Let $\tilde{f}(x) = [g(a)-g(b)][f(x)-f(a)]$ and $\tilde{g}(x) = [f(a)-f(b)][g(x)-g(a)]$. By Theorem 8.11 we have $\tilde{f}, \tilde{g}$ are differentiable on (a,b) with derivatives $\tilde{f}' = [g(a)-g(b)]f'$ and $\tilde{g}' = [f(a)-f(b)]g'$. Notice that $\tilde{f}(a) = \tilde{g}(a) = 0$, and $\tilde{f}(b) = \tilde{g}(b) = -[f(a)-f(b)][g(a)-g(b)]$. So by Exercise 8.17 there exists $c \in (a,b)$ such that $\tilde{f}'(c) = \tilde{g}'(c)$, which is exactly the equality we seek. □

Corollary 8.19 (Mean Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous, and suppose f is differentiable on (a,b). Then there exists $c \in (a,b)$ such that $f'(c) = [f(a) - f(b)]/(a-b)$.*

Proof. Apply Theorem 8.18 with $g(x) = x$. □

Exercise 8.18. Using Corollary 8.19, prove that for a differentiable function $f : (a,b) \to \mathbb{R}$:

1. If $f'(x) \geq 0$ for all $x \in (a,b)$, then f is increasing on (a,b).
2. If $f'(x) \leq 0$ for all $x \in (a,b)$, then f is decreasing on (a,b).

For the next result, we need to introduce an important concept in analysis.

Definition 8.20. *A function $f : X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is said to be* uniformly Lipschitz continuous *(usually written $f \in \mathcal{C}^{0,1}(X;Y)$) if there exists some M such that $d_Y(f(x), f(x')) \leq M d_X(x,x')$ for all $x, x' \in X$. The infimum of all values M for which the inequality holds is called the* Lipschitz constant *of the function f.*

In problem set 6, the Banach fixed point theorem was proven for Lipschitz functions Lipschitz constant $\lambda < 1$. It is a significant strengthening of the concept of continuity, as seen in the following exercise:

Exercise 8.19. Prove that if $f : X \to Y$ is uniformly Lipschitz, then f is uniformly continuous.

The importance of Lipschitz continuity is that it is the "next best thing" for functions that fail to be differentiable. As evidenced by the following theorem.

Theorem 8.21. *Let I be an interval (bounded or unbounded). If $f : I \to \mathbb{R}$ is differentiable on I, such that $|f'| \leq M$ on I, then f is uniformly Lipschitz continuous with Lipschitz constant at most M.*

Proof. The function f, being differentiable on I, is continuous by Exercise 8.8. For any $x, y \in I$ with $x \neq y$, we have by Corollary 8.19 that there exists c between x, y such that $f'(c)(x-y) = f(x) - f(y)$. Thus we have that $|f(x) - f(y)| \leq M|x-y|$ for all $x, y \in I$. □

Exercise 8.20. Show that the assumption that I is an interval in Theorem 8.21 is necessary, by giving an example of a set S that is not an interval, a function $f : S \to \mathbb{R}$ that is differentiable on S, with bounded derivative, such that f is not uniformly continuous.

Exercise 8.21. Prove that, for $\alpha > 0$, the function $x \mapsto |x|^\alpha$ is uniformly Lipschitz continuous on $\mathbb{R}$ if and only if $\alpha = 1$. Prove that on the interval $(-1,1)$, the function $x \mapsto |x|^\alpha$ is uniformly Lipschitz continuous if and only if $\alpha \geq 1$.

§8.3.2 Darboux's Theorem and Consequences.—A related result to the mean value theorems is the following.

Theorem 8.22 (Darboux's Theorem)**.** *Suppose $f : [a,b] \to \mathbb{R}$ is differentiable on $[a,b]$. Then the derivative f' is a Darboux function. (See Definition 7.35.)*

As a consequence of Darboux's theorem, if one wishes to search for more examples of Darboux functions which are not continuous, one can start by examining non-continuous functions that arise as derivatives.

Proof. If suffices to show that for every γ strictly between $f'(a)$ and $f'(b)$, there exists $c \in [a,b]$ such that $f'(c) = \gamma$. Consider the function $g(x) = f(x) - \gamma x$. Notice that g is differentiable. There are two cases.

Case 1: $f'(a) < \gamma < f'(b)$. Since g is differentiable, by Exercise 8.8, the function g is continuous. And by Corollary 7.32 $g([a,b])$ contains a minimum. By Exercise 8.15, this minimum cannot be attained at a or b since $g'(a) = f'(a) - \gamma < 0$ and $g'(b) = f'(b) - \gamma > 0$. And hence it is attained at some $c \in (a,b)$. By Lemma 8.16 $g'(c) = 0$ and hence $f'(c) = \gamma$.

Case 2: $f'(a) > \gamma > f'(b)$. The argument is the same as in case 1, except we look for the maximum value of g instead of the minimum, using that $g'(a) > 0$ and $g'(b) < 0$ to rule out the end points. □

Notice that Darboux's Theorem is automatic if the derivative were continuous. But it still holds without the continuity assumption.

Example 8.23. Return to example $f(x) = x^2 \sin \frac{1}{x}$, which we recall is differentiable but not continuously so at 0, we see that in fact for every $\gamma \in (-1,1)$, and for every $\epsilon > 0$, there exists $c \in (0,\epsilon)$ such that $f'(c) = \gamma$, which is in agreement with Darboux's theorem.

More generally, this also shows that derivatives cannot have jump discontinuities. In particular, if $f : (a,b) \to \mathbb{R}$ is differentiable on (a,b), and $c \in (a,b)$ and $\mu \to c$ is a net, then $\limsup f' \circ \mu \geq f'(c) \geq \liminf f' \circ \mu$. ■

For the subsequence discussion, the following Corollary of Darboux's Theorem is useful.

Corollary 8.24. *Let I be an interval, and $f : I \to \mathbb{R}$ real valued function differentiable on I. Suppose $f'(x) \neq 0$ for all $x \in I$, then f is strictly monotonic on I.*

Proof. By Theorem 8.22 we either must have $f'(x) > 0$ or $f'(x) < 0$ for all x. (If not, then there exists $a, b \in I$ with $f'(a)f'(b) < 0$; but Theorem 8.22 implies there exists c between a and b with $f'(c) = 0$, a contradiction.) Therefore by Exercise 8.18 the function f is monotone. It suffices to show that f is strictly monotone: suppose not, then there exists $a < b$ in I with $f(a) = f(b)$. But by Corollary 8.17 this implies there exists $c \in (a,b)$ with $f'(c) = 0$, again a contradiction. □

This allows us to give a proof of the one-dimensional inverse function theorem[5].

[5]The general inverse function theorem for functions between $\mathbb{R}^k \to \mathbb{R}^k$, or even between general Banach spaces, is much more involved. This is because in the higher dimensional cases we don't have an analogue of the order structure of $\mathbb{R}$, which helps simplify a lot of arguments in the one-dimensional case.

Theorem 8.25 (Inverse function theorem)**.** *Suppose $f : [a,b] \to \mathbb{R}$ is differentiable on $[a,b]$, such that $f'(x) \neq 0$ for any $x \in [a,b]$. Then*

- *There exists $g : f([a,b]) \to [a,b]$ such that $g \circ f$ is the identity map.*
- *The function g is differentiable on its domain, with*

$$g'(y) = \frac{1}{f'(g(y))}.$$

The following technical lemma is useful.

Lemma 8.26. *Let $m, X, Y \in \mathbb{R}$, with $Y \neq 0$, and $\epsilon \in (0,1)$. Then*

$$|Y - mX| < \epsilon |Y| \implies |Y - mX| < \frac{\epsilon}{1-\epsilon}|mX|.$$

Proof. We treat separately $Y > 0$ and $Y < 0$.

Case $Y > 0$: then $|Y - mX| < \epsilon Y \implies (1-\epsilon)Y < mX < (1+\epsilon)Y \implies \frac{m}{1+\epsilon}X < Y < \frac{m}{1-\epsilon}X$
and so $mX - \frac{m\epsilon}{1+\epsilon}X < Y < mX + \frac{m\epsilon}{1-\epsilon}X \implies |Y - mX| < \frac{\epsilon}{1-\epsilon}|mX|.$

Case $Y < 0$: then $|Y - mX| < -\epsilon Y \implies (1+\epsilon)Y < mX < (1-\epsilon)Y \implies \frac{m}{1-\epsilon}X < Y < \frac{m}{1+\epsilon}X$
and thus $mX + \frac{m\epsilon}{1-\epsilon}X < Y < mX - \frac{m\epsilon}{1+\epsilon}X \implies |Y - mX| < \frac{\epsilon}{1-\epsilon}|mX|.$ □

Proof of Theorem 8.25. By Corollary 8.24, the function f is strictly monotonic, and hence injective; therefore the inverse g exists. By Corollary 7.32, we know that $f([a,b])$ is a closed interval, and by Theorem 7.28, we have that g is also continuous. It remains to show that g is differentiable.[6]

Let $y_0 \in f([a,b])$, and set $x_0 = g(y_0)$. Differentiability of f implies that for every $\epsilon > 0$, there exists r such that $x \in B(x_0, r) \cap [a,b]$ implies $|f(x) - f(x_0) - f'(x_0)(x - x_0)| < \epsilon |f'(x_0)||x - x_0|$. (Since $|f'(x_0)| > 0$, we insert this factor on the RHS for convenience in algebra later.) This means that

$$\forall y \in f([a,b]) : g(y) \in B(g(y_0), r) \cap [a,b] \implies \left| g(y) - g(y_0) - \frac{1}{f'(g(y_0))}(y - y_0) \right| < \epsilon |g(y) - g(y_0)|.$$

By Lemma 8.26, provided $\epsilon \in (0,1)$ this gives

$$\left| g(y) - g(y_0) - \frac{1}{f'(g(y_0))}(y - y_0) \right| < \frac{\epsilon}{(1-\epsilon)|f'(g(y_0))|}|y - y_0|.$$

Since g is continuous, there exists some ρ such that $y \in B(y_0, \rho) \implies g(y) \in B(g(y_0), r)$.

Observe that $\epsilon \mapsto \frac{\epsilon}{(1-\epsilon)|f'(g(y_0))|}$ is strictly increasing bijection from $(0,1) \to (0,\infty)$. And hence, putting everything together: Given any $\tilde{\epsilon} > 0$, there exists $\epsilon \in (0,1)$ such that $\frac{\epsilon}{(1-\epsilon)|f'(g(y_0))|} = \tilde{\epsilon}$. For this ϵ, our work in the previous paragraph shows that there exists $\rho > 0$ such that

$$y \in B(y_0, \rho) \implies \left| g(y) - g(y_0) - \frac{1}{f'(g(y_0))}(y - y_0) \right| < \tilde{\epsilon} |y - y_0|.$$

This shows that g is differentiable at y_0 with derivative $1/f'(g(y_0))$, and the theorem is proved. □

[6] For the higher dimensional case, the main additional difficulty is verifying that f is locally a bijection; the actual computation of the value of the derivative is largely the same as the computation show here. On the real line we can take advantage of the monotonicity to get a very simple proof of bijectivity.

§8.3.3 L'Hôpital's Rule.—A very useful tool for evaluating limits is L'Hôpital's Rule.

Theorem 8.27. *Let $I = (a,b)$ or $\uparrow(a)$, considered as a directed set with the usual ordering. Let f, g be functions mapping $I \to \mathbb{R}$, such that they are both differentiable on I. Assume:*

- *$g'(x) \neq 0$ for all $x \in I$;*
- *as nets, $\lim f = 0 = \lim g$;*
- *the net $\lim(f'/g') = \alpha$.*

Then the net f/g is well-defined, and has limit $\lim f/g = \alpha$.

Proof. By Corollary 8.24, we have that g is strictly monotonic, and since $\lim g = 0$ we have that $g \neq 0$ for any $x \in I$. And hence the function $x \mapsto f(x)/g(x)$ is well-defined.

To show the limit property, observe that since g is strictly monotonic, for $c < d$ we have by Theorem 8.18 that for some $s \in (c,d)$,

$$\frac{f(c) - f(d)}{g(c) - g(d)} = \frac{f'(s)}{g'(s)}. \tag{8.1}$$

Since $\lim f'/g' = \alpha$, for any $\epsilon > 0$, then there exists $T \in I$ such that for $s > T$, $f'(s)/g'(s) \in (\alpha - \epsilon/2, \alpha + \epsilon/2)$. By (8.1), we have if $T \leq c < d$, then

$$\frac{f(c) - f(d)}{g(c) - g(d)} \in (\alpha - \epsilon/2, \alpha + \epsilon/2).$$

Consider the expression on the left with c fixed and d an element of the directed set $\uparrow(c) \setminus \{c\}$, the left hand side is convergent (since $f(c), g(c)$ are constants, and $\lim f = \lim g = 0$). So by taking limits we find

$$\frac{f(c)}{g(c)} \in [\alpha - \epsilon/2, \alpha + \epsilon/2] \subset (\alpha - \epsilon, \alpha + \epsilon).$$

And hence we see that for all $c > T$, $f(c)/g(c) \in (\alpha - \epsilon, \alpha + \epsilon)$, which shows that $f/g \to \alpha$. □

Exercise 8.22. Fully justify the statement that $g' \neq 0 \wedge \lim g = 0 \implies g \neq 0$ used in the first step of the proof above.

Exercise 8.23. In one of the final steps of the previous proof, we used the following fact: let x be a real valued net and suppose x is eventually in an interval I, then $\inf I \leq \liminf x \leq \limsup x \leq \sup I$. (Notice that when I is an open interval, it is possible for x to accumulate out of I.)

1. Prove this fact.
2. Give an example showing that when I is open, it is *not* possible to upgrade the statement to $\inf I < \liminf x \leq \limsup x < \sup I$.

§8.4 Second derivatives

Given $f : (a,b) \to \mathbb{R}$, suppose it is differentiable on (a,b), we can ask whether f' is also differentiable (at a point or on the whole interval). Notice that a necessary condition for f' to be differentiable is that f is continuously differentiable. In this section we will prove two theorems concerning these second derivatives. (These theorems can be also extended to higher order derivatives by induction, but we will just illustrate the principles and prove only the second order case here.)

Our first theorem says that if the second derivative exists at a point, then the function can be well-approximated by a quadratic function.

Theorem 8.28. *Suppose $f \in \mathcal{C}^1((a,b);\mathbb{R})$, such that f' is differentiable at $c \in (a,b)$ with derivative m_2. Then the function*

$$x \mapsto f(x) - f(c) - f'(c)(x-c) - \frac{1}{2}m_2(x-c)^2$$

is in $\mathfrak{o}(c,2)$.

Proof. Let $g(x) = f(x) - f(c) - f'(c)(x-c) - \frac{1}{2}m_2(x-c)^2$. Since polynomials are continuously differentiable, we have that g' is continuous, and we know $g'(x) = f'(x) - f'(c) - m_2(x-c)$. Since we assumed f' has derivative m_2 at c, this means $g' \in \mathfrak{o}(c,1)$. And so for every $\epsilon > 0$ there exists $r > 0$ such that for every $x \in B(c,r)$ we have $|g'(x)| < \epsilon|x-c|$.

Since g is differentiable, we can apply Corollary 8.19 to g and get, for some s between c and x:

$$g(x) = g(x) - g(c) = g'(s)(x-c) \implies |g(x)| < \epsilon|s-c||x-c| < \epsilon|x-c|^2.$$

(We used that if $x \in B(c,r)$ then $s \in B(c,r)$ also since s is between c and x.) □

Unfortunately, the converse is not true, and unlike first order derivatives which are characterized by having linear approximations, second order differentiation is not equivalent to having quadratic approximations.

Example 8.29. Consider the function $f(x) = x^3\sin(1/x)$. Since $|f(x)| \leq x^3$, we have that $f(x) \in \mathfrak{o}(0,2)$. Furthermore, a direct computation shows that $f'(0) = 0$ and $f'(x)|_{x\neq 0} = 3x^2\sin(1/x) - x\cos(1/x)$ and so f' is continuous.

However, f' is not differentiable at the origin (see Exercise 8.9). ■

Food for Thought 8.24. The higher order version of Theorem 8.28 is essentially stating the existence of the $k+1$ degree Taylor polynomial approximation of a function that is k-times continuously differentiable on an interval, with the $k+1$st derivative existing at one point.

Theorem 8.30. *Suppose $f \in \mathcal{C}^1((a,b);\mathbb{R})$, such that f' is differentiable (not necessarily continuously) on (a,b). Denote by f'' the derivative of f'. Then for $c,d \in (a,b)$, there exists s between c and d such that*

$$f(d) = f(c) + f'(c)(d-c) + \frac{f''(s)}{2}(d-c)^2.$$

Proof. Consider the function $g(x) = f(x) - f(c) - f'(c)(x-c) - \frac{f(d)-f(c)-f'(c)(d-c)}{(d-c)^2}(x-c)^2$, which we note is continuously differentiable, with $g(c) = g'(c) = 0$, and $g(d) = 0$. The derivative g' is also differentiable, as g is formed by adding to f a polynomial. Applying Corollary 8.17 to g we have that there exists t between c and d such that $g'(t) = 0$. Applying it again to g' between c and t we find s such that $g''(s) = 0$. The point at which $g''(s) = 0$ satisfies

$$0 = f''(s) - 2 \cdot \frac{f(d) - f(c) - f'(c)(d-c)}{(d-c)^2}$$

since the second derivative of $(x-c)^2$ is the constant 2. Rearranging this inequality gives exactly the claimed approximation. □

Food for Thought 8.25. The higher order version of Theorem 8.30 is the Taylor approximation theorem with remainder. As we can see, it is really a generalization of Corollary 8.19 to higher derivatives.

Exercise Sheet: Week 8

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 8.1. Suppose $f : X \to \mathbb{R}$ is uniformly Lipschitz continuous (here X is some metric space), and $f(x_0) = 0$. Prove that $f \in \mathfrak{v}(x_0, \alpha)$ for all $\alpha < 1$.

Question 8.2. Let f, g be two functions on $(-1,1) \to \mathbb{R}$.

1. Suppose that
   - f, g are both continuously differentiable on $(-1,1)$.
   - $f(0) = g(0) = 0$.

   Prove that the function $f \cdot g$ has a *second* derivative at 0. (Namely, prove that $(f \cdot g)'$ is differentiable at 0.)
2. Produce counterexamples to show that the assumptions are necessary:
   (a) Give an *explicit* example of f, g with $f(0) = g(0) = 0$, where f and g are differentiable on $(-1,1)$ but not necessarily continuously so, such that $(f \cdot g)'$ is *not* differentiable at 0.
   (b) Give an *explicit* example of f, g both continuously differentiable on $(-1,1)$, but with $f(0)$ and $g(0)$ not necessarily vanishing, such that $(f \cdot g)'$ is *not* differentiable at 0.

(*Hint: it may be helpful to do part 2* ***first*** *so you can see what the possible obstructions are and how the assumptions factor in.*)

Question 8.3. In this exercise we will touch on a notion of differentiability that is stronger than plain differentiability, but weaker than continuous differentiability.

> **Definition.** Let $S \subseteq \mathbb{R}$. A function $f : S \to \mathbb{R}$ is said to be *strongly differentiable* at $x_0 \in S$ if there exists an affine function $\ell \in \mathsf{Aff}$ such that, for every $\epsilon > 0$, there exists $r > 0$ such that the restriction of $f - \ell$ to $B(x_0, r)$ is uniformly Lipschitz with Lipschitz constant $< \epsilon$.

1. Prove that the function $f : (-1,1) \to \mathbb{R}$ given by $f(x) = x^2 \sin(x^{-4})$ when $x \neq 0$ and $f(0) = 0$ is differentiable at 0, but not strongly differentiable at 0.
2. Prove that if a function is strongly differentiable at x_0, then it is differentiable at x_0. (This should be a one line proof.)
3. Give an example of a function $f : (-1,1) \to \mathbb{R}$ that is strongly differentiable at 0, but not continuously differentiable on any open interval that contains 0.
   (Hint: try to built a piecewise linear function so that within any open interval that contains 0, the function will have a corner and fail to be differentiable there.)
   (It turns out that it is also true that if a function is continuously differentiable on S, then it is strongly differentiable at every $x_0 \in S$; the proof is more involved and I only just mention this in passing.)

Question 8.4. Let $f : \uparrow(0) \to \mathbb{R}$ be such that

- f is differentiable on its domain.
- $f(0) = 0$.
- f' is an increasing function.

Prove that the function $g(x) = f(x)/x$ is increasing.

(*Hint: prove that g is differentiable when $x > 0$ and that $g'(x) \geq 0$. For the latter: apply Corollary 8.19 to the expression you found for g'.*)

Problem Set 8

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 8.1. In this problem you will prove an "alternative version" of L'Hôpital's rule. In the version in the notes, we do not require f and g to be defined at the end point of the interval, nor do we require it to be differentiable there. But we require f, g to be differentiable on the interval itself, and that $\lim f'/g'$ exists. Here, we will require the opposite conditions.

> Suppose f, g are functions on $(-1, 0]$, such that $f(0) = g(0) = 0$, the functions are differentiable at 0 (*note: we don't assume differentiability anywhere else!*) with $g'(0) \neq 0$. Then
>
> - there exists $c \in (-1, 0)$ such that g is non vanishing on $(c, 0)$;
> - considering the net f/g over the index set $(c, 0)$ with the usual ordering, $\lim f/g = f'(0)/g'(0)$.

1. Prove the Theorem stated above. (*Hint: L'Hôpital's rule does not apply. Work with your bare hands using that f and g can each be written as the sum of a linear function with an $\mathfrak{o}(0,1)$ function.*)
2. Give an example of a pair of functions on $(-1, 0]$ to which the version of L'Hôpital's rule in the course readings can be used, but the version stated in this problem cannot.
3. Give an example of a pair of functions on $(-1, 0]$ to which the version of L'Hôpital's rule stated in this problem can apply, but not the version in the course readings.

Problem 8.2. Suppose f is a real-valued differentiable function on $\uparrow(0) \subseteq \mathbb{R}$. Prove that: if $f(0) = 0$ and there exists $M > 0$ such that $|f'(x)| \leq M|f(x)|$ for all x, then $f \equiv 0$.

Hint: one possible approach is via the continuity argument from Lemma 3.14. Let $S = \{x : f(y) = 0 \forall y \in [0, x]\}$. Then S is obviously non-empty and initial. For closure: use the fact that f is continuous. For continuation, use the Corollary 8.19.

Problem 8.3. Let $S \subseteq \mathbb{R}$, and suppose $f : S \to \mathbb{R}$ is such that $f(x) = mx + g(x)$ where $g(x)$ is uniformly Lipschitz with Lipschitz constant $k < |m|$.

1. Prove that $f(x)$ is injective.
2. Let $f^{-1} : f(S) \to S$ be the inverse mapping. Prove that $f^{-1}(y) = \frac{1}{m}y + h(y)$ where $h(y)$ is uniformly Lipschitz with Lipschitz constant no more than $\frac{k}{|m|-k} \cdot \frac{1}{|m|}$.

Problem 8.4. Using the result of the previous problem, prove the following version of the inverse function theorem.

> Let $S \subseteq \mathbb{R}$, and suppose $f : S \to \mathbb{R}$ is strongly differentiable at $x_0 \in S$ with non-zero derivative. (For a definition of "strongly differentiable", see this week's Exercise Sheet.) Then there exists a radius $r > 0$ such that
>
> - f is injective when restricted to $B(x_0, r) \cap S$
> - the inverse function $f^{-1} : f(B(x_0, r) \cap S) \to S$ is strongly differentiable at $f(x_0)$, with derivative $1/f'(x_0)$.

(*Hint: take a look at the proof of Theorem 8.25. The proof has two parts: showing invertibility and then showing differentiability. The two parts corresponds to the two statements you are asked to prove here, and each relies on one of the two statements in the previous problem.*)

Reading Assignment 9

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

This week we target the Riemann integral. It turns out that there are multiple different means to defining what it means for a function to be integrable. The simplest version is that due to Riemann, which essentially approaches the problem by discretizing the function and replacing the problem of "finding the area under the curve" with the evaluation of a finite sum. Of course, the area of the discretized function can only be an approximation, and so the main difficulty becomes the justification that, by taking finer-and-finer-grained discretizations, the approximate areas actually converge. We will capture this process by considering a net whose values are the approximate areas thus computed, and whose index set is some measure of the graininess of the discretizations. This week we target the definability of this notion, together with some technical aspects such as the question of "which functions are Riemann integrable?" In the next set of readings we will connect the Riemann integral to differentiation, and give some applications.

Contents

§9.1 Riemann Integral

§9.1.1 Riemann Sums.—In Week 6, we discussed how to sum an infinite list of numbers; there we mentioned the impossibility of summing an uncountably-infinite list of numbers. The process of integration gets around it by putting in a *weight*, which traditionally you associate to the width of a subinterval. In this section, we describe the "finite" case of this procedure; and in the next section we will take the limit.

Definition 9.1. *Given a closed bounded interval* $[a,b] \subsetneq \mathbb{R}$*:*

- *A* tagged subinterval *is an ordered pair* (τ, I) *where* $I \subseteq [a,b]$ *is a closed interval and* $\tau \in I$*; we refer to* τ *as the tag of the interval* I*.*
- *A* tagged division *of* $[a,b]$ *is a finite set* $\mathcal{T}$ *of tagged subintervals such that*
  1. *the union of all the subintervals appearing in* $\mathcal{T}$ *equals* $[a,b]$*;*
  2. *the sum of the widths of all the subintervals appearing in* $\mathcal{T}$ *equals* $b-a$*.*

Exercise 9.1. Prove that if $\mathcal{T}$ is a tagged division, and (τ_1, I_1) and (τ_2, I_2) are elements of $\mathcal{T}$, then $I_1 \cap I_2$ is either the empty set or the set of only one element. (*Hint: since the union covers* $[a,b]$*, the total width is at least* $b-a$*; if the intersection has more than two elements, then the overlap has positive width.*)

Each tagged subinterval represents a vertical rectangular strip: it has as its base the interval I, and its height $f(\tau)$ for the given function $f : [a,b] \to \mathbb{R}$. Because of Exercise 9.1, for a tagged division, two rectangles at most share a side; they cannot have an overlap with positive area. So we can regard the total (signed) area of the rectangles as an approximation of the area under the curve for the function f. We make this a definition. (Recall that for an interval I, the notation $|I|$ refers to its width, namely $\sup I - \inf I$.)

Definition 9.2. *Given* $f : [a,b] \to \mathbb{R}$*, and a tagged division* $\mathcal{T}$ *of* $[a,b]$*, the corresponding* Riemann sum *of* f *is*

$$\mathsf{S}_{\mathcal{T}} f \coloneqq \sum_{(\tau,I)\in\mathcal{T}} f(\tau)\cdot|I|.$$

(The finite sum is over all the tagged subintervals that make up $\mathcal{T}$*.)*

Food for Thought 9.2. One thing one may remember from calculus courses is the distinction between an "improper" versus a "proper" integral. In calculus classes these are usually described as being based on the function becoming unbounded, or discontinuous at some points, or based on the function being integrated over a domain of infinite width. These are very much exposed by the Riemann sum formulation in Definition 9.2. First consider the situation where the function has an unbounded domain. Then necessarily the analogous definition of the Riemann sum would need to handle either (i) one interval with infinite width in which case $|I|$ is ill-defined, or (ii) an infinite sum. The first option leaves the formula ill-defined. The second option required contending with all the complication we learned in Week 6 concerning infinite summation. And hence we prefer to limit to "proper" settings where the domain is a bounded interval. We will consider the case where the function has an unbounded range later on after we introduce the Riemann integral.

Exercise 9.3. Consider the function $f : [-1,1] \to \mathbb{R}$ given by

$$f(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ +1 & x > 0 \end{cases}$$

Check that each of the following is indeed a tagged division of $[-1,1]$, and compute the Riemann sum of f relative to it.

1. $\mathcal{T}_1 = \left\{\left(-\frac{1}{2}, [-1,-\frac{1}{2}]\right), \left(0, [-\frac{1}{2}, 0]\right), \left(1, [0,1]\right)\right\}$.
2. $\mathcal{T}_2 = \left\{\left(-\frac{1}{2}, [-1,0]\right), \left(0, [0,0]\right), \left(1, [0,1]\right)\right\}$.
3. $\mathcal{T}_3 = \left\{\left(-\frac{1}{2}, [-1,-\frac{1}{2}]\right), \left(-\frac{1}{2}, [-\frac{1}{2}, 0]\right), \left(0, [0,1]\right)\right\}$.

The definition of the Riemann sum approximation is pretty easy to parse. One particular interpretation of the Riemann sum, useful for numerical analysis, is to regard the placement of the tags of the tagged division as a way of sampling the outputs of the function f. In this sense we can see the tagged division process as essentially replacing the function f by a "discretized" version.

Slightly more difficult is understanding in what sense we wish to take the limit, so that the limiting object will actually reflect the area under the curve. Returning to Exercise 9.3, we can compare the tagged divisions $\mathcal{T}_1$ and $\mathcal{T}_3$: there we see that with the same underlying interval (widths), one can get drastically different values for the Riemann sum if the value of the function f varies greatly within the interval. This leads us to the next section.

§9.1.2 Riemann Integral.—In this subsection we will formulate the Riemann integral as a limit of the Riemann sums relative to tagged divisions; the main question is in what sense should the limit be taken? We will answer using nets; the formulation of the Riemann integral as a limit was actually one of the motivating questions that led to the development of the nets concept. Some of the formulation in this section may seem a bit unnecessarily complicated; these are done to set the stage for later, when we will introduce the generalization of Riemann integrals called Henstock integrals.

Definition 9.3. *The* width *of a tagged division $\mathcal{T}$, denoted $|\mathcal{T}|$, is the width of its widest subinterval. (In symbols: $|\mathcal{T}| = \max\{|I| : (\tau, I) \in \mathcal{T}\}$.)*

The idea behind introducing the idea of the width, is that (i) for "nice" functions (such as continuous ones), we expect the amount of variation of a function to be less over shorter intervals; this would mean that in the Riemann sum $\mathsf{S}_{\mathcal{T}} f$, we expect the terms corresponding to tagged subintervals with smaller width to give better approximations; (ii) for functions with bad discontinuities (such as the jump discontinuities seen in Exercise 9.3, while shortening the interval on which the jump happens will *not* change the amplitude of the jump, the reduction of the width $|I|$ will still lessen the contribution of the discontinuous point to the Riemann sum $\mathsf{S}_{\mathcal{T}} f$.

Thus, Riemann integration is built on the idea of taking the limit as the width of the tagged divisions go to zero.

Food for Thought 9.4. A useful observation is that for a closed bounded interval $[a, b] \subsetneq \mathbb{R}$, for any $\delta > 0$, there exists a tagged division $\mathcal{T}$ with $|\mathcal{T}| < \delta$. And hence "taking the limit as the width goes to zero" is not a vacuous procedure.

Consider the following set of ordered pairs, with the first element a positive real and the second element a tagged division of $[a, b]$:

$$\mathfrak{r}([a,b]) := \{(\delta, \mathcal{T}) : \delta > |\mathcal{T}| > 0\} \tag{9.1}$$

We establish an ordering on $\mathfrak{r}([a, b])$ by

$$(\delta, \mathcal{T}) \leqq (\delta', \mathcal{T}') \iff \delta \geq \delta'. \tag{9.2}$$

By the Archimedean property of the real numbers, together with Food for Thought 9.4, we find that $\mathfrak{r}([a, b])$ is a directed set.

Definition 9.4. *A function $f : [a, b] \to \mathbb{R}$ is said to be Riemann integrable (abbreviated $f \in \mathcal{R}([a, b])$), if the net $\rho[f] : \mathfrak{r}([a, b]) \to \mathbb{R}$ given by $\rho[f]_{\delta,\mathcal{T}} = \mathsf{S}_{\mathcal{T}} f$ converges. We denote the value of the limit by $\int_a^b f(x)\, dx$.*

Example 9.5. Let's prove that the function given in Exercise 9.3 is Riemann integrable.

Let $\mathcal{T}$ be any tagged division of $[-1,1]$, and let its width $|\mathcal{T}| = w$. Consider the intervals I that correspond to tags sitting in $(0,1]$; the union of all these intervals must be itself an interval (why?). This union has total width at least $1-w$ and at most $1+w$. And hence these tags contribute a total between $1-w$ and $1+w$ to the Riemann sum $\mathsf{S}_{\mathcal{T}} f$.

Next consider the intervals I that correspond to tags sitting in $[-1,0)$; their union is again an interval, with total width between $1-w$ and $1+w$. Hence the tags contribute a total between $-w-1$ and $w-1$ toward $\mathsf{S}_{\mathcal{T}} f$. A tag (if any) that sits at 0 contributes nothing to the Riemann sum.

Thus we conclude that for an arbitrary tagged division of $[-1,1]$, we have that

$$-2w \leq \mathsf{S}_{\mathcal{T}} f \leq 2w$$

after adding the contributions from the tags in $[-1,0)$ to the tags in $(0,1]$. Now we are in a position to prove convergence of the corresponding net. Let $\epsilon > 0$, set $\delta_0 = \frac{1}{3}\epsilon$, and let $\mathcal{T}_0$ have width less that δ_0 per Food for Thought 9.4. Then for every $(\delta, \mathcal{T}) \in \uparrow((\delta_0, \mathcal{T}_0))$, we have $|\mathsf{S}_{\mathcal{T}} f| \leq 2\delta < \epsilon$. This shows that the net $\rho[f]$ converges and $\int_{-1}^{1} f(x)\,dx = 0$. ∎

Example 9.6. Not all functions are Riemann integrable. Let $f : [0,1] \to \mathbb{R}$ with $f(x) = 1$ when $x \in \mathbb{Q}$ and 0 otherwise.

Let $(\delta, \mathcal{T}) \in \mathfrak{r}([0,1])$. We can construct a new tagged division $\mathcal{T}_1$, using the same set of intervals as $\mathcal{T}$, such that the tag for any non-degenerate subinterval is a rational number; we can also construct a new tagged division $\mathcal{T}_2$ using the same intervals as $\mathcal{T}$, such that the tag for any non-degenerate sub-interval is a irrational number. Since the width of a tagged division only depends on the underlying intervals, we have that both $(\delta, \mathcal{T}_1), (\delta, \mathcal{T}_2) \in \mathfrak{r}([0,1])$ and they both succeed $(\delta, \mathcal{T})$.

We have however that $\mathsf{S}_{\mathcal{T}_1} f = 1$ and $\mathsf{S}_{\mathcal{T}_2} f = 0$. And so for any $(\delta, \mathcal{T}) \in \mathfrak{r}([0,1])$ we've shown that $\sup \rho[f]_{\uparrow(\delta,\mathcal{T})} = 1$ and $\inf \rho[f]_{\uparrow(\delta,\mathcal{T})} = 0$; and thus the net $\rho[f]$ cannot converge. ∎

Exercise 9.5. A *necessary condition* for a function to be Riemann integrable is that it is bounded on the interval $[a,b]$. (This explains why integrals with unbounded integrands are also considered to be "improper".) In this exercise you will prove the contrapositive: "if $f : [a,b] \to \mathbb{R}$ is unbounded, then $f \notin \mathcal{R}([a,b])$". The following steps give an outline: Given $(\delta, \mathcal{T}) \in \mathfrak{r}([a,b])$.

1. Prove that there exists a sequence $t : \mathbb{N} \to [a,b]$ such that
   (a) $|f(t_i)| > i$.
   (b) every t_i lies in the *same* non-degenerate subinterval I_* of $\mathcal{T}$.
2. Define $\mathcal{T}_i$ to be the tagged division that is formed by replacing the tagged subinterval (τ_*, I_*) of $\mathcal{T}$ with (t_i, I_*) (notations from the previous step).
3. Prove that $(\delta, \mathcal{T}_i) \in \mathfrak{r}([a,b])$, and the set $\{\mathsf{S}_{\mathcal{T}_i} f : i \in \mathbb{N}\}$ is unbounded.
4. Use the previous step to conclude that $\rho[f]$ cannot converge.

§9.1.3 Some technical lemmas.—In the remainder of this section we will record some technical lemmas that will help streamline subsequent discussions.

Definition 9.7. *A tagged division $\mathcal{T}'$ is said to be a* refinement *of $\mathcal{T}$ if for each subinterval I' that appears in $\mathcal{T}'$, there exists a subinterval I of $\mathcal{T}$ such that $I' \subseteq I$.*

In particular, none of the subintervals of $\mathcal{T}'$ can "cross boundaries" and be part of two distinct subintervals of $\mathcal{T}$ (ignoring the cases of the degenerate subintervals).

Sometimes for our arguments it is convenient to ignore the tags.

Definition 9.8. *Given a tagged division $\mathcal{T}$ of an interval $[a,b]$ and a bounded function $f : [a,b] \to \mathbb{R}$, the* upper and lower Darboux sums *corresponding to $\mathcal{T}$ are*

$$\overline{\mathsf{S}}_{\mathcal{T}} f \coloneqq \sum_{(\tau,I)\in\mathcal{T}} (\sup f(I)) \cdot |I|;$$
$$\underline{\mathsf{S}}_{\mathcal{T}} f \coloneqq \sum_{(\tau,I)\in\mathcal{T}} (\inf f(I)) \cdot |I|.$$

Rather obviously $\underline{\mathsf{S}}_{\mathcal{T}} f \le \mathsf{S}_{\mathcal{T}} f \le \overline{\mathsf{S}}_{\mathcal{T}} f$.

Lemma 9.9. *If $\mathcal{T}'$ is a refinement of $\mathcal{T}$, and f is a bounded function, then*

$$\underline{\mathsf{S}}_{\mathcal{T}} f \le \underline{\mathsf{S}}_{\mathcal{T}'} f \le \overline{\mathsf{S}}_{\mathcal{T}'} f \le \overline{\mathsf{S}}_{\mathcal{T}} f.$$

Exercise 9.6. Prove the lemma.

Exercise 9.7. If $\mathcal{T}'$ is a refinement of $\mathcal{T}$, prove that $|\mathcal{T}'| \le |\mathcal{T}|$.

Lemma 9.10. *If $\mathcal{T}_1$, $\mathcal{T}_2$ are two tagged divisions, then there exists a tagged division $\mathcal{T}$ that is a refinement of both $\mathcal{T}_1$ and $\mathcal{T}_2$.*

Proof. It suffices to specify the subintervals of $\mathcal{T}$, since Definition 9.7 makes no restrictions on the tags. We can simply let the underlying subintervals of $\mathcal{T}$ be those intervals of the form $I_1 \cap I_2$ where I_1 is a subinterval of $\mathcal{T}_1$ and I_2 is one of $\mathcal{T}_2$. □

Lemma 9.11. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two tagged divisions, and f a bounded function, then*

$$|\mathsf{S}_{\mathcal{T}_1} f - \mathsf{S}_{\mathcal{T}_2} f| \le \overline{\mathsf{S}}_{\mathcal{T}_1} f - \underline{\mathsf{S}}_{\mathcal{T}_1} f + \overline{\mathsf{S}}_{\mathcal{T}_2} f - \underline{\mathsf{S}}_{\mathcal{T}_2} f.$$

Proof. By Lemma 9.10, there exists a common refinement $\mathcal{T}$. By triangle inequality we have $|\mathsf{S}_{\mathcal{T}_1} f - \mathsf{S}_{\mathcal{T}_2} f| \le |\mathsf{S}_{\mathcal{T}_1} f - \mathsf{S}_{\mathcal{T}} f| + |\mathsf{S}_{\mathcal{T}} f - \mathsf{S}_{\mathcal{T}_2} f|$. By Lemma 9.9 we have that both $\mathsf{S}_{\mathcal{T}_1} f$ and $\mathsf{S}_{\mathcal{T}} f$ are between $[\underline{\mathsf{S}}_{\mathcal{T}_1} f, \overline{\mathsf{S}}_{\mathcal{T}_1} f]$ and hence their difference is bounded by the width of this interval. Similarly for the difference between $\mathsf{S}_{\mathcal{T}_2} f$ and $\mathsf{S}_{\mathcal{T}} f$. □

Using the ideas given above, we can reformulate Cauchy's criterion for the convergence of the net $\rho[f]$ in to the following theorem.

Theorem 9.12 (Darboux's criterion)**.** *A bounded function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if for every $\epsilon > 0$, there exists $\delta > 0$ such that for every tagged division $\mathcal{T}$ with width less than δ, the difference $\overline{\mathsf{S}}_{\mathcal{T}} f - \underline{\mathsf{S}}_{\mathcal{T}} f < \epsilon$.*

Proof. Recall that by definition, $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if the net $\rho[f]$ is Cauchy. The latter requires that for every ϵ one can find a tail set of $\rho[f]$ that fits within an interval of width ϵ, and is thus equivalent to requiring, for every $\epsilon > 0$, there to exist some $\delta > 0$ such that for every $\mathcal{T}, \mathcal{T}'$ with width less than δ, $|\mathsf{S}_{\mathcal{T}} f - \mathsf{S}_{\mathcal{T}'} f| < \epsilon$. We will show the equivalence of this final inequality with the bound on the differences of the upper and lower Darboux sums.

($\Leftarrow$) follows from Lemma 9.11.

($\Rightarrow$) holds because given any $\mathcal{T}$, for every $\eta > 0$, there exists a tagged division $\mathcal{T}_u$ with the same intervals but with the tags satisfying $f(\tau) > \sup f(I) - \eta$; and there exists a tagged division $\mathcal{T}_l$ with the same intervals but with the tags satisfying $f(\tau) < \inf f(I) + \eta$. For this pair we can compare

$$\overline{\mathsf{S}}_{\mathcal{T}} f - \underline{\mathsf{S}}_{\mathcal{T}} f - |\mathsf{S}_{\mathcal{T}_u} f - \mathsf{S}_{\mathcal{T}_l} f| \in [0, 2\eta(b-a)).$$

And so taking $\eta \to 0$ we get the desired estimate. □

§9.2 Riemann Integrability and Properties of the Riemann Integral

As shown in Example 9.6, not every function defined on bounded intervals are integrable. In this section we examine some theorems giving sufficient conditions for a function to be Riemann integrable. The simplest one of our sufficient conditions is:

Theorem 9.13. *If $f : [a,b] \to \mathbb{R}$ is continuous, then $f \in \mathcal{R}([a,b])$.*

Proof. Since $[a,b]$ is compact, by Theorem 7.25, the function f is uniformly continuous and bounded. Thus for every $\epsilon > 0$, we can choose δ such that whenever $x, x' \in [a,b]$ has $|x - x'| < \delta$, then $|f(x) - f(x')| < \frac{\epsilon}{b-a}$. For this δ, we see that whenever $|\mathcal{T}| < \delta$,

$$\overline{\mathsf{S}}_{\mathcal{T}} f - \underline{\mathsf{S}}_{\mathcal{T}} f = \sum_{(\tau,I)\in\mathcal{T}} (\sup f(I) - \inf f(I)) \cdot |I| < \sum_{(\tau,I)\in\mathcal{T}} \frac{\epsilon}{b-a} \cdot |I| = \epsilon. \tag{9.3}$$

The Theorem then follows by Theorem 9.12. □

This simple theorem, however, does not cover the case of Example 9.5; the function f there is not continuous. Comparing the two arguments, we see that there are two ways in which we can make use of the width δ of the tagged division $|\mathcal{T}|$:

1. Where the function f is continuous, by taking δ small enough we can guarantee the function f has small (size ϵ) variation over each subinterval of $\mathcal{T}$; this implies that different samplings of f using the same sets of subintervals will give approximately the same Riemann sum.
2. Where the function f is discontinuous, by taking δ small enough we constrain the large variation of f to take place on only a small interval, so that again different samplings of f using the same sets of subintervals will give approximately the same Riemann sum.

The latter is possible in Example 9.5 because there aren't too many discontinuity points. Before discussing the general theorem, let's give an illustrative example to show the power of the two ideas above.

Example 9.14. Let us revisit the function we first saw in Example 7.33, given by

$$f(x) = \begin{cases} \frac{1}{b} & x \in \mathbb{Q} \wedge x = a/b \text{ in lowest terms;} \\ 0 & x \notin \mathbb{Q}. \end{cases}$$

Recall that this function is discontinuous at all the rational numbers, and continuous at all the irrationals. Let's prove that its restriction to $[0,1]$ is Riemann integrable.

In fact, I claim that its Riemann integral is 0. We will use the following fact: let $b \in \mathbb{N}$, then the set $\{x \in [0,1] : f(x) \geq 1/b\}$ is a finite set. In fact, since there are at most $b+1$ numbers in $[0,1]$ with denominator b, we see that the size of the set is no bigger than $\frac{1}{2}(b+2)(b+1)$.

Let $\epsilon > 0$. Then there exists $b \in \mathbb{N}$ such that $\frac{1}{b} < \epsilon/2$. Set $\delta > 0$; we will keep δ as a variable for now and determine a value for it, in terms of ϵ, later. Let $\mathcal{T}$ be a tagged division with width $|\mathcal{T}| < \delta$. Split $\mathcal{T} = \mathcal{T}_g \cup \mathcal{T}_b$, where $\mathcal{T}_g$ contains only those subintervals on which $f(x) \leq \frac{1}{b}$, and $\mathcal{T}_b$ those subintervals which contain a point x such that $f(x) > \frac{1}{b}$. With this splitting we can write

$$\mathsf{S}_{\mathcal{T}} f = \sum_{(\tau,I)\in\mathcal{T}_g} f(\tau)\cdot|I| + \sum_{(\tau,I)\in\mathcal{T}_b} f(\tau)\cdot|I|.$$

For the sum over $\mathcal{T}_b$, since there are at most $\frac{1}{2}(b+1)b$ intervals that can contain a point x such that $f(x) \geq 1/(b-1)$, and since that we know $f(x) \in [0,1]$ by its definition, we have

$$\sum_{(\tau,I)\in\mathcal{T}_b} f(\tau)\cdot|I| < \sum_{(\tau,I)\in\mathcal{T}_b} 1\cdot\delta \leq \frac{1}{2}(b+1)b\delta.$$

For the sum over $\mathcal{T}_g$, on this set we have that $f(\tau) \leq 1/b$ by definition, and so

$$\sum_{(\tau,I)\in\mathcal{T}_g} f(\tau)\cdot|I| \leq \sum_{(\tau,I)\in\mathcal{T}_g} \frac{1}{b}\cdot|I| \leq \frac{1}{b}.$$

So if we choose $\delta \leq \frac{\epsilon}{b(b+1)}$, then

$$0 \leq \mathsf{S}_{\mathcal{T}} f < \frac{1}{b} + \frac{1}{2}(b+1)b\delta < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus we have shown that for any $\mathcal{T}$ with width $|\mathcal{T}| < \frac{\epsilon}{b(b+1)}$, for our net

$$\rho[f]_{\uparrow(\frac{\epsilon}{b(b+1)},\mathcal{T})} \subseteq [0,\epsilon)$$

and this shows that $\rho[f] \to 0 = \int_0^1 f(x)\,dx$. ∎

It turns out that the ideas behind the previous example can be strengthened to a general theorem.

Definition 9.15. *A subset $S \subseteq \mathbb{R}$ is a* null set *if for every $\epsilon > 0$, there exists a countable collection $\mathscr{U}$ of open intervals, such that*

1. *$\cup\mathscr{U} \supseteq S$;*
2. *the (possibly infinite) sum $\sum_{I\in\mathscr{U}} |I|$ converges (absolutely) and evaluates to be less than ϵ.*

Proposition 9.16. *Any countable set is a null set.*

Proof. If S is countable, then there exists a surjection $\nu : \mathbb{N} \to S$. Let $\mathscr{U} = \{(\nu(n) - \epsilon 2^{-2-n}, \nu(n) + \epsilon 2^{-2-n}) : n \in \mathbb{N}\}$. Notice that $\sum_{I\in\mathscr{U}} |I| = \frac{1}{2}\epsilon$. □

Exercise 9.8. Prove that if $S_1, S_2, \ldots$ is a countable family of null sets, then their union $\cup_{i=1}^{\infty} S_i$ is also a null set. (*Hint: to cover the union with countably many open intervals whose total length is at most ϵ, first find a cover of S_1 with total length $\frac{1}{2}\epsilon$, add to it a cover of S_2 with total length $\frac{1}{4}\epsilon$, etc.*)

Example 9.17. Not all null sets are countable. One example of an uncountable null set is the *Cantor middle thirds* set. This set is formed by a recursive construction. Let $C_1 = [0,1]$. Let $C_2 = [0,1] \setminus (\frac{1}{3}, \frac{2}{3}) = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$.

Given C_n which is a finite union of closed intervals, let C_{n+1} be formed by taking each closed interval of C_n, and removing the middle third of it (in the form of an open interval). We notice that C_{n+1} is formed by 2^n closed intervals, each of length 3^{-n}.

The Cantor set C is the infinite intersection $\cap_{n=1}^{\infty} C_n$ of this family. That C is non-empty follows from Cantor's nested intersection theorem. That it has uncountably many points can be proven using Cantor's diagonal argument. That it is null follows from noting that, by slightly enlarging the closed intervals in C_{n+1} (by lengthening each one by a factor of 1/100, say), we can get a cover of C by finitely many open intervals whose lengths add up to be no more than $\frac{101}{100}\left(\frac{2}{3}\right)^n$. ■

Theorem 9.18 (Lebesgue Criterion for Riemann Integrability)**.** *A bounded function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if the set of points on which it is discontinuous is a null set.*

The proof of the Lebesgue Criterion is long and technical; I will defer it to a separate section later on in this set of notes.

Example 9.19. On the Week 7 exercise sheet you proved that an increasing function $f : [a,b] \to \mathbb{R}$ can have at most countably many points of discontinuity. Combining that with Proposition 9.16 and Theorem 9.18, we see that every bounded monotone function $f : [a,b] \to \mathbb{R}$ is Riemann integrable. ■

Proposition 9.20. *If $f : [a,b] \to [m,M]$ is Riemann integrable, and $g : [m,M] \to \mathbb{R}$ is continuous, then $g \circ f \in \mathcal{R}([a,b])$.*

Proof. We will use Theorem 9.18. Let C denote the subset of $[a,b]$ on which f is continuous. Then $g \circ f$ is continuous on C (for every $z \in C$, let μ be a net converging to z, then $f \circ \mu$ converges to $f(z)$ since f is continuous at z, and hence $g \circ f \circ \mu$ converges to $g(f(z))$ since g is continuous). And hence the set of *discontinuous* points of $g \circ f$ must be a subset of that of f; the latter being a null set implies the former is a null set too. (Any open cover of a set is an open cover of its subsets.) □

Exercise 9.9. Given an example of a function $f : [a,b] \to [m,M]$ and a function $g : [m,M] \to \mathbb{R}$ such that the set of discontinuous points for $g \circ f$ is a *proper* subset of the set of discontinuous points for f.

Example 9.21. A consequence of Proposition 9.20 is that if $f \in \mathcal{R}([a,b])$, then so is $|f| \in \mathcal{R}([a,b])$; here we take $g(y) = |y|$ which is uniformly continuous on $\mathbb{R}$.

Notice that the converse *does not hold*. Consider function f given in Example 9.6; the function $h(x) = f(x) - \frac{1}{2}$ with the same domain is such that $|h|$ is the constant function $1/2$ and hence is integrable, but h itself is not integrable. ■

Proposition 9.22.

1. *If $f, g \in \mathcal{R}([a,b])$, and $c,d \in \mathbb{R}$, then $cf + dg \in \mathcal{R}([a,b])$ and $\int_a^b cf + dg\ dx = c\int_a^b f\ dx + d\int_a^b g\ dx$.*
2. *If $f, g \in \mathcal{R}([a,b])$ and $f(x) \le g(x)$ for all $x \in [a,b]$, then $\int_a^b f\ dx \le \int_a^b g\ dx$.*

Proof. Both claims follow from the properties of nets. Notice that $\mathsf{S}_{\mathcal{T}}(cf + dg) = c\mathsf{S}_{\mathcal{T}} f + d\mathsf{S}_{\mathcal{T}} g$ by properties of finite sums, so the nets $\rho[cf + dg] = c\rho[f] + d\rho[g]$ and the first result follows from §4.2.1. The second result follows from noting that $\rho[f] \le \rho[g]$ pointwise as nets. □

Exercise 9.10. Check that if $f \in \mathcal{R}([a,b])$, then $|\int_a^b f\ dx| \leq \int_a^b |f|\ dx$. (Recall that $|f|$ is integrable by Example 9.21.)

Exercise 9.11. Suppose $f, g \in \mathcal{R}([a,b])$.

1. Prove that for any $p > 0$, the function $|f|^p \in \mathcal{R}([a,b])$.
2. Prove that $f \cdot g \in \mathcal{R}([a,b])$. (*Hint: you can use the polarization identity* $4fg = (f+g)^2 - (f-g)^2$.)

Exercise 9.12. Prove Schwarz's inequality: if $f, g \in \mathcal{R}([a,b])$, then

$$\left| \int_a^b fg\ dx \right|^2 \leq \int_a^b f^2\ dx \cdot \int_a^b g^2\ dx.$$

(The integrability of the functions involved are dealt with in the previous exercise. Instead of working with f and g directly, you should work with $\tilde{f} = f/(\int_a^b f^2\ dx)^{1/2}$ and similarly $\tilde{g}$. First prove that under these assumptions $|\int_a^b \tilde{f}\tilde{g}\ dx| \leq 1$. This last inequality is a consequence of the arithmetic-mean-geometric-mean inequality $|xy| \leq \frac{1}{2}(x^2+y^2)$.)

Proposition 9.23. *Given interval $[a,b]$ and $c \in (a,b)$. Let $f : [a,b] \to \mathbb{R}$, and write f_1 for its restriction to $[a,c]$ and f_2 for its restriction to $[c,b]$. Then $f \in \mathcal{R}([a,b])$ if and only if $f_1 \in \mathcal{R}([a,c])$ and $f_2 \in \mathcal{R}([c,b])$.*

Exercise 9.13. Prove the previous proposition. It is perhaps easiest to use Theorem 9.18. Use $D \subset [a,b]$ for the set of points at which f is discontinuous, $D_1 \subseteq [a,c]$ that of f_1, and $D_2 \subseteq [c,b]$ that of f_2. Prove that $D = D_1 \cup D_2$.

§9.3 Proof of the Lebesgue Criterion (Theorem 9.18)

§9.3.1 Preliminary Discussions.—Given $f : [a,b] \to \mathbb{R}$ a bounded function, define the function $\operatorname{osc} : 2^{[a,b]} \to \mathbb{R}$ given by

$$\operatorname{osc}(S) = \sup f(S) - \inf f(S). \tag{9.4}$$

Given $x \in [a,b]$, let $\mathbb{I}_x$ denote the set of open intervals containing x, ordered by reverse inclusion, which makes it a directed set. The mapping $I \mapsto \operatorname{osc}(I \cap [a,b])$ is a net in $\mathbb{R}$; it is monotone decreasing as moving to subsets only decreases the supremum and increases the infimum. Hence by the Monotone Convergence Theorem it converges. Define

$$\omega_f(x) \coloneqq \lim(\mathbb{I}_x \ni I \mapsto \operatorname{osc}(I \cap [a,b])) \tag{9.5}$$

to be the limit.

Lemma 9.24. *The function f is continuous at x if and only if $\omega_f(x) = 0$.*

Proof. We will use Theorem 7.5.

($\Leftarrow$) Let $J \ni f(x)$ be an open interval, and let $w = \min\{\sup J - f(x), f(x) - \inf J\}$. There exists $I \ni x$ such that $\operatorname{osc}(I \cap [a,b]) < w$ since $\omega_f(x) = 0$. Since $f(x) \in f(I \cap [a,b])$, we have $(f(x)-w, f(x)+w) \supseteq f(I \cap [a,b])$, and hence $f(I \cap [a,b]) \subseteq J$. This shows f is continuous at x.

($\Rightarrow$) If f is continuous at x, then for every $\epsilon > 0$, there exists $I \ni x$ such that $f(I \cap [a,b]) \subseteq (f(x) - \epsilon/2, f(x) + \epsilon/2)$, and hence $\operatorname{osc}(I \cap [a,b]) \leq \epsilon$. Thus the net defining ω_f converges to 0. □

It is convenient to take advantage of the Archimedean property of real numbers and introduce the following notations for the sets of discontinuity points of the function f:

$$D_k := \{x \in [a,b] : \omega_f(x) \geq \frac{1}{k}\} \tag{9.6}$$

$$D := \bigcup \{D_k : k \in \mathbb{N}\} \tag{9.7}$$

Lemma 9.25. *D is a null set if and only if for every $k \in \mathbb{N}$, the set D_k is a null set.*

Proof. ($\Rightarrow$) Since D_k are subsets of D, any cover of D by open intervals of total length $< \epsilon$ is also a cover of D_k by open intervals of total length $< \epsilon$.

($\Leftarrow$) Follows from Exercise 9.8. □

Corollary 9.26. *If S is such that every countable cover of S by open intervals has total length bigger than or equal to ϵ, and T is a subset of S formed by removing countably many points, then every countable cover of T by open intervals also has total length bigger than or equal ϵ.*

Proof. Fix a countable cover of T by open intervals, and denote the total length of of this cover by ℓ.

A countable set is a null set, so for every $\delta > 0$ there exists a countable cover of $S \setminus T$ with total length $< \delta$. Adding this to a countable cover of T, this gives a countable cover of S, which by hypothesis has total length at least ϵ. Hence it must be that $\ell \geq \epsilon - \delta$. Since δ is arbitrary, this means that $\ell \geq \epsilon$. □

Lemma 9.27. *D_k is compact.*

Proof. Since $D_k \subseteq [a,b]$, it is bounded. It is enough to prove that it is closed. Let ν be a net in D_k, and $x \in [a,b]$ an accumulation point of ν.

Let I be an open interval containing x. Then for some α, the term $\nu_\alpha \in I$. We have by definition $\operatorname{osc}(I \cap [a,b]) \geq \omega_f(\nu_\alpha) \geq \frac{1}{k}$. Since this hold for all intervals I, this means that the limit $\omega_f(x) \geq \frac{1}{k}$ also. □

§9.3.2 Necessity.—In this section we will prove that if D is *not* a null set, then f cannot be Riemann integrable. By Lemma 9.25, for D to not to be a null set, there must be some $k \in \mathbb{N}$ for which D_k is not a null set. There exists some $\zeta > 0$ such that for every cover of D_k with countably many open intervals, the cover has total length at least ζ.

We will argue via Theorem 9.12, and show that for every tagged division, $\overline{\mathsf{S}}_{\mathcal{T}} f - \underline{\mathsf{S}}_{\mathcal{T}} f \geq \frac{\zeta}{k}$.

Let $\mathcal{T}$ be an arbitrary tagged division of $[a,b]$. Let D_k' be formed by removing from D_k all elements that appear as an endpoint of a subinterval in $\mathcal{T}$. Let $\mathcal{T}_b$ be the subset of $\mathcal{T}$ whose subinterval contains an element of D_k'. Let $\mathcal{I}_b$ be the set of the *open versions* of subintervals of $\mathcal{T}_b$. By our construction D_k' is covered by $\mathcal{I}_b$, and hence the sum of the widths of the elements of $\mathcal{I}_b$ is at least ζ, by Corollary 9.26. For each (open) interval I in $\mathcal{I}_b$, since it contains an element of D_k', we find $\operatorname{osc}(I) \geq 1/k$, and hence the same holds for the corresponding closed interval in $\mathcal{T}_b$.

We can now compute

$$\overline{S}_{\mathcal{T}} f - \underline{S}_{\mathcal{T}} f = \sum_{(\tau,I)\in\mathcal{T}} \operatorname{osc}(I)\cdot|I| = \sum_{(\tau,I)\in\mathcal{I}_b} \operatorname{osc}(I)\cdot|I| + \underbrace{\sum_{(\tau,I)\notin\mathcal{I}_b} \operatorname{osc}(I)\cdot|I|}_{\geq 0} \geq \sum_{(\tau,I)\in\mathcal{I}_b} \frac{1}{k}\cdot|I| = \frac{1}{k}\sum_{I\in\mathcal{I}_b}|I| \geq \frac{\zeta}{k}.$$

§9.3.3 Sufficiency.—In this section we will prove that if D is a null set, then f is Riemann integrable. We will again pass through the Darboux criterion Theorem 9.12.

Since f is assumed to be bounded, we can assume $|f| < M$ everywhere.

Our strategy will be to divide $[a,b]$ into two subsets: first is a subset which will intersect regions where ω_f is large; on this set the difference of the Riemann sums will be made small by using that D is a null set and so the region has small total width. On the complement, the function ω_f is small, and the difference of the Riemann sums will be made small by taking advantage of that.

Step 1, some constructions.—

Given the ϵ that we aimed for, observe that there exists $k \in \mathbb{N}$ such that $\frac{1}{k} < \frac{\epsilon}{4(b-a)}$.

By Lemma 9.25, the corresponding D_k is a null set. And hence there is a covering of D_k by countably many open intervals such that the total length of the intervals is less than $\frac{\epsilon}{12M}$. Since D_k is compact according to Lemma 9.27, we may assume the covering of D_k is by *finitely many* open intervals; call this covering $\mathcal{I}_k$.

Denote by δ_0 the width of the *smallest* of this finite collection.

Denote by $K = [a,b] \setminus \cup\mathcal{I}_k$, since the latter is an open set, K is closed and bounded, and hence compact.

Finally, let $\mathcal{J}_k$ be a finite collection of open intervals, each one formed by taking one open interval of $\mathcal{I}_k$, and tripling its width by expanding in equal amounts above and below the interval. Thus $\mathcal{J}_k$ is a cover of D_k by open intervals, with total length less than $\frac{\epsilon}{4M}$.

Our construction of $\mathcal{J}_k$ and K means that if I is a closed subinterval of $[a,b]$ of width less than δ_0, then either $I \subseteq K$ or $I \subseteq J$ for some $J \in \mathcal{J}_k$.

Exercise 9.14. Prove that last statement. (*Hint: it suffices to show that if I has $|I| < \delta_0$ and $I \not\subseteq K$, then $I \in J$ for some J of $\mathcal{J}_k$; remember that $\mathcal{J}_k$ is formed by lengthening the intervals of $\mathcal{I}_k$.*)

Step 2, estimating the oscillation on K.—

By construction, if $x \in K$, we have that $\omega_f(x) < \frac{1}{k}$, and hence there exists an value ℓ_x such that $\operatorname{osc}((x-3\ell_x, x+3\ell_x)\cap[a,b]) < \frac{2}{k}$. Since the set $\{(x-\ell_x, x+\ell_x) : x \in K\}$ forms an open cover of a compact set, we can pick a finite subcover centered at points $x_1, \ldots, x_m$.

Set $\delta = \min\{\ell_{x_1}, \ldots, \ell_{x_m}, \delta_0\}$. By construction if I is a closed subinterval of $[a,b]$ with width less than δ, we find that either I is entirely contained in an element J of $\mathcal{J}_k$ or I is entirely contained in one of $(x_i - 3\ell_{x_i}, x_i + 3\ell_{x_i})$. In the latter case, we have that $\operatorname{osc}(I) < \frac{2}{k}$.

Step 3, estimating the difference of upper and lower sums.—

Let $\mathcal{T}$ be a tagged division with width less than δ. We can compute

$$\overline{S}_{\mathcal{T}} f - \underline{S}_{\mathcal{T}} f = \sum_{(\tau,I)\in\mathcal{T}} \operatorname{osc}(I)\cdot|I| = \sum_{(\tau,I)\in\mathcal{T}\wedge I\subseteq K} \operatorname{osc}(I)\cdot|I| + \sum_{(\tau,I)\in\mathcal{T}\wedge I\not\subseteq K} \operatorname{osc}(I)\cdot|I|.$$

The first term we have the smallness of oscillation

$$\sum_{(\tau,I)\in\mathcal{T}\wedge I\subseteq K} \operatorname{osc}(I)\cdot|I| < \frac{2}{k}\sum_{(\tau,I)\in\mathcal{T}}|I| = \frac{2(b-a)}{k} < \frac{\epsilon}{2}.$$

The second term we notice that since f is bounded, $\operatorname{osc}(I) < 2M$. And hence we have

$$\sum_{(\tau,I)\in\mathcal{T}\wedge I\not\subseteq K} \operatorname{osc}(I)\cdot|I| < 2M \sum_{(\tau,I)\in\mathcal{T}\wedge I\not\subseteq K}|I|.$$

For the latter sum, based on our construction, each I is contained in some interval J of $\mathcal{J}_k$. So we can split them up: for each $J\in\mathcal{J}_k$, first add up the lengths of those I contained in J.

$$\sum_{(\tau,I)\in\mathcal{T}\wedge I\not\subseteq K}|I| \le \sum_{J\in\mathcal{J}_k}\sum_{(\tau,I)\in\mathcal{T}\wedge I\subseteq J}|I|.$$

This is an $\le$ sign because it is possible that one I belongs to two different Js, so the double counting will increase the total length. Since the Is are drawn from a tagged division, this means that the any given pair I, I' have negligible overlap, and so we must have

$$\sum_{(\tau,I)\in\mathcal{T}\wedge I\subseteq J}|I| \le |J|.$$

Therefore, putting the last few inequalities together, we get

$$\sum_{(\tau,I)\in\mathcal{T}\wedge I\not\subseteq K}\operatorname{osc}(I)\cdot|I| \le 2M\sum_{J\in\mathcal{J}_k}|J| < 2M\cdot\frac{\epsilon}{4M} = \frac{\epsilon}{2}.$$

Putting everything together we see that this shows $\overline{S}_{\mathcal{T}} f - \underline{S}_{\mathcal{T}} f < \epsilon$, and hence we can apply Darboux's criterion to show convergence of the Riemann integral.

Exercise Sheet: Week 9

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Instructions: *Please work together within your small group to first address any lingering questions you may have about the assigned readings for this week. Here are some additional topics and questions for discussion. You are not required to limit the scope of your break-out room to these questions, nor are you required to touch on every item in the list below. However, the items are listed in order of relevance to your written problem set, and so I would suggest working through the questions below linearly.*

Question 9.1. Prove that if $f \in \mathcal{R}([a,b])$ and $g(x) = f(x+c)$, then $g \in \mathcal{R}([a-c, b-c])$ and that $\int_a^b f(x)\,dx = \int_{a-c}^{b-c} g(x)\,dx$.

*Important note: we have not proven **any** statements about changes of variables yet. So you cannot prove this by just using a "u-substitution" etc. You have to get your hands dirty. What you can do:*

1. First prove that you can get a one-to-one mapping from element of the directed set $\mathfrak{r}([a,b])$ to $\mathfrak{r}([a-c, b-c])$.
2. Prove that if $(\delta, \mathcal{T}) \in \mathfrak{r}([a,b])$ is mapped to $(\delta', \mathcal{T}') \in \mathfrak{r}([a-c, b-c])$, then $\rho[f]_{\delta,\mathcal{T}} = \mathsf{S}_{\mathcal{T}} f$ and $\rho[g]_{\delta',\mathcal{T}'} = \mathsf{S}_{\mathcal{T}'} g$ are equal in values.
3. The previous step plus a little work shows that $\rho[f]$ and $\rho[g]$ are subnets of each other. And hence if one converges, the other converges to the same limit.

Question 9.2. If $f \in \mathcal{R}([a,b])$, Proposition 9.23 shows that $f \in \mathcal{R}([a,c])$ for every $c \in [a,b]$. Prove that for every $c \in (a,b)$ and $\epsilon > 0$, there exists $\delta > 0$ such that if $\gamma \in (c-\delta, c+\delta)$, then

$$\left| \int_a^c f(x)\,dx - \int_a^\gamma f(x)\,dx \right| < \epsilon.$$

Question 9.3.

1. Prove that if $f : [a,b] \to \mathbb{R}$ is continuous and *not identically zero*, then $\int_a^b |f(x)|\,dx > 0$.
2. Prove that the analogous statement without the continuous assumption is false.

Problem Set 9

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Problem 9.1.

1. Prove that if $f : [a,b] \to \mathbb{R}$, and $f^3 \in \mathcal{R}([a,b])$, then $f \in \mathcal{R}([a,b])$. (There exists a complete proof that is only about 3 sentences long, using what is available in the lecture notes.)
2. Explain why the same method will *not* work if instead of assuming $f^3 \in \mathcal{R}([a,b])$, we had assumed $f^2 \in \mathcal{R}([a,b])$. (I am not asking for a counterexample; I am asking for you to examine the proof you gave in part one and point out which steps will no longer work with the f^2 assumption, and why.)

Problem 9.2.

1. Let $0 < a < b$. Prove that if $f \in \mathcal{R}([a,b])$ and $g(x) = f(\lambda x)$, for some $\lambda > 0$, then $g \in \mathcal{R}([\lambda^{-1}a, \lambda^{-1}b])$ with integral $\int_a^b f(x)\,dx = \lambda \int_{\lambda^{-1}a}^{\lambda^{-1}b} g(x)\,dx$.
   (*Hint: do the same thing you did for the first exercise problem, find a one-to-one mapping between* $\mathfrak{r}([a,b])$ *and* $\mathfrak{r}([\lambda^{-1}a, \lambda^{-1}b])$*, and establish a relation between the corresponding nets* $\rho[f]$ *and* $\rho[g]$.)
2. Using the previous part, prove that for $s, t > 1$,

$$\int_1^{st} \frac{1}{x}\,dx = \int_1^{s} \frac{1}{x}\,dx + \int_1^{t} \frac{1}{x}\,dx.$$

   (In other words, you are asked to prove the well-known identity $\ln(st) = \ln(s) + \ln(t)$.)

Problem 9.3. Let $\varphi : [a,b] \to [c,d]$ be a strictly increasing, uniformly Lipschitz continuous bijection. Let $f : [c,d] \to \mathbb{R}$ be a function.

1. Prove that f is discontinuous at y if and only if $f \circ \varphi$ is discontinuous at $\varphi^{-1}(y)$. (Theorem 7.28 may be useful.)
2. Prove that if $S \subseteq [a,b]$ is a null set, then $\varphi(S) \subseteq [c,d]$ is also a null set.
3. Prove that if $f \circ \varphi \in \mathcal{R}([a,b])$, then $f \in \mathcal{R}([c,d])$.

(*Remark: the converse of part 3 is not true.*)

Problem 9.4. Given $f \in \mathcal{R}([a,b])$, prove that for every $\epsilon > 0$ there exists a continuous function $g : [a,b] \to \mathbb{R}$ such that

$$\int_a^b |f - g|\,dx < \epsilon.$$

(*Hint: it is easiest to let g be piecewise affine. Darboux's criterion and its proof can serve as inspirations.*)

Reading Assignment 10

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

In this shortened week, we will talk about indefinite integrals and their relation to derivatives. The highlights are the fundamental theorems of calculus, and a version of integration by parts whose proof does not depend on Leibniz rule and differentiability.

Contents

§10.1 Indefinite Integrals and Derivatives

§10.1.1 Variable boundaries.—Our starting point for this section is Proposition 9.23. Repeated application of that proposition tells us that if $f \in \mathcal{R}([a,b])$ and $[c,d] \subseteq [a,b]$, then $f \in \mathcal{R}([c,d])$ also. This enables the following definition. For convenience we will take the convention that

$$\text{if } a > b, \text{ then } \int_a^b f(x)\,dx := -\int_b^a f(x)\,dx. \tag{10.1}$$

Definition 10.1. *Let $f \in \mathcal{R}([a,b])$, and $c \in [a,b]$, the* indefinite Riemann integral of f based at c *is the function $F : [a,b] \to \mathbb{R}$ given by $F(x) = \int_c^x f(y)\,dy$.*

Exercise 10.1. Prove that if $f \in \mathcal{R}([a,b])$ and $c \in [a,b]$, then $\int_a^c f(x)\,dx + \int_c^b f(x)\,dx = \int_a^b f(x)\,dx$. (*Hint: Prove that if $\mathcal{T}_1$ is a tagged division for $[a,c]$ and $\mathcal{T}_2$ a tagged division for $[c,b]$, then $\mathcal{T}_1 \cup \mathcal{T}_2$ is a tagged division for $[a,b]$.*)

Exercise 10.2. Using the previous exercise, prove that if F is the indefinite Riemann integral of f based at c and G is the indefinite Riemann integral of f based at d, then $F - G$ is a constant. Compute that constant.

Proposition 10.2. *If $f \in \mathcal{R}([a,b])$, and F its indefinite Riemann integral based at some $c \in [a,b]$, then F is uniformly Lipschitz continuous.*

Proof. By Exercise 9.5, the function f is bounded. Let $M \in \mathbb{R}$ be such that $|f| < M$ on $[a,b]$. Let $x, y \in [a,b]$ with $x < y$, by Exercise 10.1 we find that $F(y) - F(x) = \int_x^y f(z)\,dz$. By Exercise 9.10, we have

$|F(y) - F(x)| \leq \int_x^y |f(z)|\, dz$. By Proposition 9.22 we find $|F(y) - F(x)| < \int_x^y M\, dz = M|y - x|$. Since y, x are arbitrary, this shows that F is uniformly Lipschitz with constant M. □

However, in spite what you may have remembered from your calculus course, *differentiation and integration are not fully opposites of each other*. For general Riemann integrable functions, uniform Lipschitz continuity is the best we can do: the indefinite Riemann integral of a function need *not* be differentiable.

Example 10.3. Return to the function given in Exercise 9.3, which we proved to be Riemann integrable in Example 9.5. Let F be its indefinite integral based at 0, we can actually compute to find $F(x) = |x|$. This function we have already established is not differentiable at 0. ■

Example 10.4. Even when the indefinite Riemann integral turns out to be differentiable, the derivative need not agree with the original function. Let f be the function that is identically zero except at the origin, where we set $f(0) = 1$. The function f has a single point of discontinuity, and hence is Riemann integrable on any bounded interval; let's take $[a, b] = [-1, 1]$. Its indefinite Riemann integral based at 0 turns out to be $F(x) \equiv 0$. This being a constant function is differentiable, but with derivative equaling zero everywhere. And thus we have $F(x) = \int_0^x f(y)\, dy$, but $F' \not\equiv f$. ■

§10.1.2 Fundamental Theorems of Calculus.—In view of the examples give above, in this section we give some instances where we can actually recognize differentiation and integration as "opposite operations". These are generally grouped[1] under the banner of "fundamental theorems of calculus".

Continuing our discussion from the previous section, we have:

Theorem 10.5 (Fundamental Theorem of Calculus: derivatives of integrals)**.** *Let $f \in \mathcal{R}([a, b])$ and F an indefinite Riemann integral of f. If f is continuous at $x_0 \in [a, b]$, then F is differentiable at x_0 and $F'(x_0) = f(x_0)$.*

Proof. It suffices to show that $F(x) - F(x_0) - f(x_0)(x - x_0) \in \mathfrak{o}(x_0, 1)$. Since f is continuous at x_0, for every $\epsilon > 0$ there exists $r > 0$ such that $|f(x) - f(x_0)| < \epsilon$ on $B(x_0, r)$. Since

$$F(x) - F(x_0) - f(x_0)(x - x_0) = \int_{x_0}^{x} f(s) - f(x_0)\, ds,$$

the same argument used in the proof of Proposition 10.2 shows that $F(x) - F(x_0) - f(x_0)(x - x_0)$ is uniformly Lipschitz continuous with Lipschitz constant ϵ on the ball $B(x_0, r)$. This in fact shows that F is not only differentiable at x_0 with derivative $f(x_0)$, but that F is in fact *strongly* differentiable there. (See Week 8 Exercise Sheet.) □

Corollary 10.6 (Existence of primitives)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then there exists $F \in C^1(\mathbb{R}; \mathbb{R})$ such that $F' = f$. (A function F such that $F' = f$ is called a* primitive *of f.)*

Proof. By Theorem 9.13, the continuous $f \in \mathcal{R}([a, b])$ for any $[a, b] \subseteq \mathbb{R}$. Define $F(x) = \int_0^x f(s)\, ds$. Then the result follows from Theorem 10.5. □

[1] There are at least two commonly used statements when it comes to Riemann integrals, referred to frequently as the first and second theorems. Unfortunately, there's no universal agreement of which comes first and which comes second…

Example 10.7. As an application, we will prove the following statement: if $f \in \mathcal{R}([a,b])$ and $f(x) > 0$ for all x, then $\int_a^b f(x)\,dx > 0$.

Notice that Proposition 9.22 would only give you that $\int_a^b f(x)\,dx \geq 0$. This is because the argument there relies on comparing the underlying nets of Riemann sums behind the Riemann integral and, when you take limits of two nets, one having strictly larger values than the other, the best you can conclude in a non-strict inequality (think $\lim \frac{1}{x}$ as $x \to \infty$).

Let us prove that the inequality must be strict by contradiction.

First notice that by Proposition 9.22 as already mentioned, our assumptions imply that for any subinterval $[c,d] \subseteq [a,b]$ that $\int_c^d f(x)\,dx \geq 0$. Hence, if we assume $\int_a^b f(x)\,dx = 0$ for contradiction, the indefinite integral F of f based at a will be identically zero on the interval $[a,b]$. (The only way two non-negative numbers sum to zero is if both of them vanishes.)

On the other hand, by Lebesgue's criterion (Theorem 9.18), the set of discontinuity points of f is a null set and hence there exists at least one point $c \in [a,b]$ at which f is continuous. (Sketch of proof: suppose f is discontinuous everywhere, and there exists a countable cover of the discontinuous points of total length less than $|b-a|/2$, then since $[a,b]$ is a compact there is also a finite sub-cover, whose total length is still less than $|b-a|/2$; but that's absurd since $[a,b]$ has length $|b-a| > |b-a|/2$.)

Notice that F is constant, and hence differentiable everywhere with derivative 0. At c, by the fundamental theorem, $F'(c) = f(c)$. This contradicts the assumption that $f(c) > 0$. Hence the statement is proved. ∎

Notice that as seen in our examples, it is possible for F to be differentiable at a point of discontinuity of f. We next turn our attention to the other direction: are derivatives of differentiable functions automatically Riemann integrable? The answer is no.

Example 10.8. The function given by $f(x) = x^2 \sin(\frac{1}{x^2})$ when $x \neq 0$, and $f(0) = 0$, is differentiable but *not* continuously differentiable. In fact, one can check that f' is *unbounded* near the origin. Therefore f' cannot be integrable on any non-degenerate interval containing the origin. ∎

In the previous example the function fails to be integrable because it is unbounded. A natural follow-up question is: what if we require the derivative to be bounded? Are such functions automatically Riemann integrable? The answer, unfortunately again, is no. The construction will have to rely on creating a differentiable function whose bounded derivative is nonetheless discontinuous on a non-null set. Since derivatives are Darboux, such a construction is bound to be tricky. We will not go into the details into such a construction; interested readers can look up *Volterra's Function*, perhaps the best known of examples of this type.

The most efficient way of dealing with this issue, then, is to just assume the derivative is Riemann integrable.

Theorem 10.9 (Fundamental Theorem of Calculus: integrals of derivatives)**.** *Let $f : [a,b] \to \mathbb{R}$ be differentiable. If $f' \in \mathcal{R}([a,b])$ then for every $x, y \in [a,b]$, $f(y) - f(x) = \int_x^y f'(s)\,ds$.*

Proof. For convenience we will assume $y > x$; if $y = x$ the equality holds trivially as $0 = 0$; and when $x > y$ we will swap the symbols using the convention in (10.1).

What we will show is that for every $\epsilon > 0$, there exists some $\delta > 0$ such that for every tagged division $\mathcal{T}$

of $[x,y]$ with $|\mathcal{T}| < \delta$, the difference $|f(y) - f(x) - \mathsf{S}_{\mathcal{T}} f'| < \epsilon$. To do so, we will actually prove that for every tagged division $\mathcal{T}$, there exists a refinement $\mathcal{T}'$ such that $\mathsf{S}_{\mathcal{T}'} f' = f(y) - f(x)$. Supposing for a moment this is true. Given $\epsilon > 0$, since f' is assumed to be integrable, we can appeal to Darboux's Criterion (Theorem 9.12) that there exists some $\delta > 0$ such that every tagged division $\mathcal{T}$ with width less than δ satisfies $\overline{\mathsf{S}}_{\mathcal{T}} f' - \underline{\mathsf{S}}_{\mathcal{T}} f' < \epsilon/2$. Let $\mathcal{T}'$ be appropriate refinement. Then we have

$$|f(y) - f(x) - \mathsf{S}_{\mathcal{T}} f'| \leq |\mathsf{S}_{\mathcal{T}'} f' - \mathsf{S}_{\mathcal{T}} f'| < \epsilon;$$

in the final inequality we made use of Lemmas 9.9 and 9.11.

It remains to demonstrate the existence of $\mathcal{T}'$. We will in fact let $\mathcal{T}'$ have the *same* subintervals as $\mathcal{T}$, just with different tags; this makes it a refinement of $\mathcal{T}$. Let I be a subinterval of $\mathcal{T}$, we will choose as its tag for $\mathcal{T}'$, a point $\tau \in I$ that satisfies

$$f'(\tau)(\sup I - \inf I) = f(\sup I) - f(\inf I).$$

This τ exists thanks to the Mean Value Theorem (Corollary 8.19). For these choices, we have that

$$\mathsf{S}_{\mathcal{T}'} f' = \sum_{(\tau,I)\in\mathcal{T}'} f'(\tau)|I| = \sum_{(\tau,I)\in\mathcal{T}'} f(\sup I) - f(\inf I).$$

Using that the subintervals of a tagged division can be laid end-on-end to reassemble into the full interval $[x,y]$, we see that this sums telescopes and evaluates to $f(y) - f(x)$. □

§10.1.3 Integration By Parts.—Let's begin with an exercise.

Exercise 10.3. Prove the integration by parts formula for derivatives: Suppose F, G are differentiable real-valued functions defined on $[a,b]$, with derivatives f, g respectively. If $f, g \in \mathcal{R}([a,b])$, then

$$F(b)G(b) - F(a)G(a) = \int_a^b F(x)g(x)\,dx + \int_a^b f(x)G(x)\,dx.$$

(Make sure to justify why the functions $F \cdot g$ and $G \cdot f$ are Riemann integrable.)

The following Theorem is a strengthened version, that no longer requires differentiability of F and G.

Theorem 10.10. *Let $f, g \in \mathcal{R}([a,b])$, and denote by $F(x) = \int_a^x f(y)\,dy$ and $G(x) = \int_a^x g(y)\,dy$. Then*

$$F(b)G(b) = \int_a^b f(x)G(x)\,dx + \int_a^b F(x)g(x)\,dx.$$

(Notice that $F(a) = G(a) = 0$.)

Proof. The functions $f \cdot G$ and $F \cdot g$ are Riemann integrable, since they are products of a Riemann integrable function with a continuous function: the set of points at which $f \cdot G$ is discontinuous is a subset of the set of points at which f is discontinuous, and hence we can apply Lebesgue's Criterion (Theorem 9.18); similarly for $F \cdot g$. It remains to show that the integrals converge to the claimed values.

We will do so by showing that for every $\epsilon > 0$, we can prove that $|F(b)G(b) - \int_a^b f(x)G(x)\ dx - \int_a^b F(x)g(x)\ dx| < \epsilon$. The proof is essentially an extension of the summation-by-parts formulae you worked on in the exercise sheet for Week 6.

First, notice that if we have a tagged division[2] $\mathcal{T}$, we can write the telescoping sum

$$F(b)G(b) = \sum_{(\tau,I)\in\mathcal{T}} F(\sup I)G(\sup I) - F(\inf I)G(\inf I).$$

A little bit of algebra yields

$$\begin{aligned} F(b)G(b) &= \sum_{(\tau,I)\in\mathcal{T}} (F(\sup I) - F(\inf I))G(\sup I) + F(\inf I)(G(\sup I) - G(\inf I)) \\ &= \sum_{(\tau,I)\in\mathcal{T}} G(\sup I)\cdot \int_I f(x)\ dx + F(\inf I)\cdot \int_I g(x)\ dx. \end{aligned} \tag{10.2}$$

Here we used the short hand $\int_I f(x)\ dx = \int_{\inf I}^{\sup I} f(x)\ dx$. Exercise 10.1 was used to convert differences of Riemann integrals $\int_a^{\sup I} f(x)\ dx - \int_a^{\inf I} f(x)\ dx$ to $\int_I f(x)\ dx$. Similarly we can write

$$\int_a^b f(x)G(x)\ dx = \sum_{(\tau,I)\in\mathcal{T}} \int_I f(x)G(x)\ dx, \qquad \int_a^b F(x)g(x)\ dx = \sum_{(\tau,I)\in\mathcal{T}} \int_I F(x)g(x)\ dx. \tag{10.3}$$

Since both f and g are Riemann integrable, they are both bounded. Let M be such that $|f|, |g| < M$ on $[a,b]$. By Proposition 10.2, both F and G are uniformly Lipschitz continuous with Lipschitz constant less than M. And so for any interval I, we have

$$\left|G(\sup I)\cdot \int_I f(x)\ dx - \int_I G(x)f(x)\ dx\right| \le \int_I \underbrace{|G(\sup I) - G(x)|}_{<M|I|}\underbrace{|f(x)|}_{<M}\ dx < M^2|I|^2.$$

And we also have a similar expression for $|F(\inf I)\cdot \int_I g(x)\ dx - \int_I F(x)g(x)\ dx|$. Applying these to (10.2) and (10.3) we find that

$$\begin{aligned} &\left|F(b)G(b) - \int_a^b f(x)G(x)\ dx - \int_a^b F(x)g(x)\ dx\right| \le \\ &\qquad \sum_{(\tau,I)\in\mathcal{T}} \left|G(\sup I)\cdot \int_I f(x)\ dx - \int_I G(x)f(x)\ dx\right| + \left|F(\inf I)\cdot \int_I g(x)\ dx - \int_I F(x)g(x)\ dx\right| \\ &\qquad\qquad < \sum_{(\tau,I)\in\mathcal{T}} 2\cdot M^2|I|^2. \end{aligned} \tag{10.4}$$

[2] We will not actually use the tags at all in this proof, since we will not directly handle the Riemann sums; it is just convenient to use the subintervals of a tagged division since we already introduced the notion.

Given $\epsilon > 0$, let $\delta_0 = \epsilon \cdot (2M^2(b-a))^{-1}$. Then as long as $|\mathcal{T}| < \delta_0$, we are guaranteed by (10.4) that

$$\left|F(b)G(b) - \int_a^b f(x)G(x)\,dx - \int_a^b F(x)g(x)\,dx\right| < \sum_{(\tau,I)\in\mathcal{T}} 2M^2\delta_0|I| = \epsilon. \qquad \square$$

Problem Set 10

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

***Special rules for PS 10**: only two problems are listed below; only those two problems need to go through Eli Review. You are asked to submit, in addition, revisions to two problems from Problem Sets 6-9, when you submit your solutions on D2L for final grading. The revision will earn you new grade toward Problem Set 5 as well as replace the original grade you earned on the problem previously. Please be sure to clearly indicate the problem set number and question number selected on your work. While not necessary, I would also appreciate it if you transcribe/summarize the question statement on the document; it will help facilitate grading.*

Problem 10.1. In this question you will extend the Mean Value Theorem for derivatives to *weighted* integrals.

1. First, the unweighted case: let f be a real-valued continuous function on the interval $[a,b]$. Prove that there exists $c \in [a,b]$ such that $f(c)\cdot(b-a) = \int_a^b f(x)\,dx$. (You can prove this using the mean value theorem for derivatives.)
2. Next, the weighted case: let f be a real-valued continuous function on the interval $[a,b]$, and let $w \in \mathcal{R}([a,b])$ satisfy $w(x) \geq 0$ for all $x \in [a,b]$. Prove that there exists some $c \in [a,b]$ such that

$$f(c)\cdot \int_a^b w(x)\,dx = \int_a^b f(x)w(x)\,dx.$$

   (*Hint: since w is not continuous, you can no longer use the fundamental theorem of calculus. Instead, consider the continuous function $x \mapsto f(x)\cdot\int_a^b w(x)\,dx$ and try to apply the intermediate value theorem.*)

Problem 10.2. For this question, you will prove a version of what is often called the *second mean value theorem for definite integrals*. Suppose $w : [a,b] \to \mathbb{R}$ is differentiable, increasing, and with $w' \in \mathcal{R}([a,b])$. Let $f \in \mathcal{R}([a,b])$. Prove that there exists $c \in [a,b]$ such that

$$w(b)\cdot\int_c^b f(x)\,dx + w(a)\cdot\int_a^c f(x)\,dx = \int_a^b w(x)f(x)\,dx.$$

(*Hint: use the integration by parts formula for integrable functions.*)

(*Remark: the assumption that w is differentiable with integrable derivative is actually not necessary for this result to hold; but without differentiability of w the proof is significantly messier.*)

Reading Assignment 11

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

Summary

In this final week, our goal is to introduce two generalizations of the Riemann integral. The first, the Henstock integral, "fixes" many of the shortcomings of the Riemann integral. For example, its definition bypasses the need for an "improper integral" on a bounded interval; and it also clarifies the role played by null sets in integration theory. The second, the Stieltjes integral, we introduce because it is a good stepping stone toward learning measure theory, and it is a crucial ingredient for defining line integrals (integrals along curves in higher dimensional spaces).

Contents

§11.1 The Henstock Integral

In this section we will discuss a generalization of the Riemann integration called *the Henstock Integral*.[1] One of the main motivations of the original introduction of the Henstock integral was to allow us to consider certain "improper" Riemann integrals as actually integrable. Let's return first to the two basic modes in which an integral can be improper. First is when the domain of the integral is an unbounded interval. In this case, we see that there's not much we can do if we want to use Riemann sums: it simply is not possible to form a Riemann sum with finitely many subintervals, since at least one of the subintervals will be infinite in extent and the Riemann sum will fail to make sense. Second is when the domain of the integral is bounded, but the integrand is unbounded.

Example 11.1. A good example is the function $f(x) = 1/\sqrt{x}$ on $(0,1]$ extended to have $f(0) = 0$. By Exercise 9.5 we see that this function is not Riemann integrable on $[0,1]$; but for every $x_0 \in (0,1]$, we have (since the restriction of f is continuous) $f \in \mathcal{R}([x_0,1])$. Furthermore, the limit (here I am using calculus notation) $\lim_{x_0\to 0^+} \int_{x_0}^1 f(x)\,dx$ exists and converges.

The fact that $f \notin \mathcal{R}([0,1])$ boils down to the fact that for every $\delta > 0$ and every $N > 0$, there exists a tagged division $\mathcal{T}$ with width less than δ such that $\mathsf{S}_{\mathcal{T}} f > N$, due to the unboundedness. The idea

[1] This has been independently (re)discovered and developed by many people over the past century; the various definitions were often not shown to be equivalent until much later. And hence this concept is also referred to, in various places, as the Henstock–Kurzweil integral, the generalized Riemann integral, the gauge integral, the Denjoy integral, the Denjoy–Perron integral, the Luzin integral, or the Perron integral.

of Denjoy (who first formulated a version of the Henstock integral) basically boils down to this: in the definition of the Riemann integral, a tagged division with width δ only requires the tag τ to be within the interval I. Is there a way we can enforce that certain (τ, I) combinations are not allowed? For example, suppose there is a way to ensure that for the first subinterval $[0, t]$ that appears in $\mathcal{T}$, the only compatible choice of tag τ is the origin, then that tagged subinterval will no longer make arbitrarily large contributions to the Riemann sum. In this way, the compatibility restriction actually may allow the Riemann sums to remain bounded. ■

A different way of understanding the same problem is to think about the process of the Riemann sum approximation. Our idea has always been to replace the area under the curve by the area of a discretized version of the function f. Riemann's method measures the coarseness of the approximation using the *largest* of the widths that appears in the tagged division. This treats every point in the domain equally, with the same width restriction applied to all the points.

However, just by inspecting graphs of functions, it appears that this is not necessarily very effective. For continuous functions on bounded closed intervals, by uniform continuity this method works reasonably well: since we can choose a $\delta > 0$ for every $\epsilon > 0$ that guarantees our discretization with widths less than δ will approximate the original function at every point with error no more than ϵ. For discontinuous functions, however, one may want to put "more resolution" near where the function is discontinuous to capture the jumps and oscillations there, and put "less resolution" elsewhere where the functions are nearly constant.

In this lens, the problem with the function $f(x) = 1/\sqrt{x}$ is that on the interval $[0, t]$ it has infinite variation. So perhaps it is possible to place some restrictions on the association of tags to intervals to account for the need of much higher resolution near the origin and much lower resolution away from the origin?

To do so, we can circle back to an earlier idea.

Definition 11.2. *If $\eta : [a, b] \to 2^{\mathbb{R}}$ is an entourage mapping, we say that the tagged division $\mathcal{T}$ is* η-fine *if, for every $(\tau, I) \in \mathcal{T}$, we have $I \subsetneq \eta(\tau)$.*

Exercise 11.1. Prove that for any entourage mapping η, there exists a tagged division $\mathcal{T}$ of $[a, b]$ that is η-fine. (You've done almost this exact same exercise for Exercise Sheet 3, question 3.1.)

We shall interpret the entourage mapping as prescribing the "worst resolution" (or if you like, the "minimum zoom level") allowed at a point, if we wish to use that point as a tag. From this point of view, an entourage mapping η has better resolving power than $\tilde{\eta}$, if for every $x \in [a, b]$, $\eta(x) \subseteq \tilde{\eta}(x)$. We will define the directed set

$$\mathfrak{h}([a, b]) := \{(\eta, \mathcal{T}) : \eta \text{ is an entourage mapping }, \ \mathcal{T} \text{ is } \eta\text{-fine}\} \tag{11.1}$$

with the ordering

$$(\eta, \mathcal{T}) \leqq (\eta', \mathcal{T}') \iff [\forall x \in [a, b] : \eta'(x) \subseteq \eta(x)]. \tag{11.2}$$

(Notice that given η_1 and η_2, we can define $\eta(x) = \eta_1(x) \cap \eta_2(x)$ to show directedness.) And similarly to the case of Riemann integration, we will define the net $\varsigma[f] : \mathfrak{h}([a, b]) \to \mathbb{R}$ by

$$\varsigma[f]_{\eta, \mathcal{T}} = \mathsf{S}_{\mathcal{T}} f. \tag{11.3}$$

Proposition 11.3. *$\varsigma[f]$ is a subnet of $\rho[f]$.*

Proof. Given an entourage mapping η, denote by $|\eta|_{[a,b]} = \sup\{|\eta(x)| : x \in [a,b]\}$, the maximum width of all intervals that appear in the output of η. Then necessarily if $\mathcal{T}$ is η-fine, we have $|\mathcal{T}| < |\eta|_{[a,b]}$. Let's define $\varphi : \mathfrak{h}([a,b]) \to \mathfrak{r}([a,b])$ by the mapping $\varphi((\eta, \mathcal{T})) = (|\eta|_{[a,b]}, \mathcal{T})$. We check straightforwardly that $\rho[f]_{|\eta|_{[a,b]},\mathcal{T}} = \varsigma[f]_{\eta,\mathcal{T}}$; and that φ is an increasing map of directed sets.

Finally, to guarantee the frequency of $\varphi(\mathfrak{h}([a,b]))$, notice that that for any $\delta_0 > 0$, we can choose η_0 such that $\eta_0(x) = (x - \delta_0/2, x + \delta_0/2)$. By construction $|\eta_0|_{[a,b]} \leq \delta_0$ and hence $\varphi((\eta_0, \mathcal{T})) \geqq (\delta_0, \mathcal{T})$. □

Definition 11.4. *We say a function $f : [a,b] \to \mathbb{R}$ is* Henstock Integrable, *written $f \in \mathcal{H}([a,b])$, if the net $\varsigma[f]$ converges. We will write using $\int_a^b f(x)\,dx = \lim \varsigma[f]$.*

Food for Thought 11.2. By virtue of Proposition 11.3, if a function is Riemann integrable then it is also Henstock integrable, with the same integral. Hence in Definition 11.4 we can use the same notation $\int_a^b f(x)\,dx$ to denote the limit of $\varsigma[f]$ when it converges. Based on what we know about nets, this also means that if $f \in \mathcal{H}([a,b])$, then the net $\rho[f]$ has finite accumulation points. This places some minor restrictions on what functions can be Henstock integrable.

Example 11.5. The function $f : [0,1] \to \mathbb{R}$ with $f(x) = 1$ when $x \in \mathbb{Q}$ and 0 otherwise has been established as not Riemann integrable (see Example 9.6). We shall show that it is Henstock integrable with integral 0. Fix $q : \mathbb{N} \to \mathbb{Q}$ an enumeration of the rationals. Let $\epsilon > 0$. Set η be such that

- $\eta(q_n) = (q_n - \epsilon 2^{-2-n}, q_n + \epsilon 2^{-2-n})$;
- if $x \notin \mathbb{Q}$, let $\eta(x) = (x-1, x+1)$.

Notice that η is much more "zoomed in" near the rational numbers, compared to the irrational numbers. Let $\mathcal{T}$ be any η-fine tagged division. Then

$$\varsigma[f]_{\eta,\mathcal{T}} = \sum_{(\tau,I)\in\mathcal{T}\wedge\tau\in\mathbb{Q}} \underbrace{f(\tau)}_{=1}\,|I| + \sum_{(\tau,I)\in\mathcal{T}\wedge\tau\notin\mathbb{Q}} \underbrace{f(\tau)}_{=0}\,|I|.$$

So we have

$$0 \leq \varsigma[f]_{\eta,\mathcal{T}} \leq \sum_{n\in\mathbb{N}} |\eta(q_n)| \leq \frac{\epsilon}{2} < \epsilon.$$

This shows $\varsigma[f]$ is eventually in any ϵ neighborhood of 0, and hence converges to it. ■

Food for Thought 11.3. Notice that what Henstock integrability does is *allow* us to zoom in more on some points compared to others. It does not force us to zoom in on any particular pre-defined set. The idea is that as long as the set of "problematic points" is sufficiently sparse, then this zoom in procedure can cure a lot of non-convergences that happen for Riemann integrals.

Exercise 11.4. The idea behind Example 11.5 can be extended to show that if $f : [a,b] \to \mathbb{R}$ is such that the set $\{x \in [a,b] : f(x) \neq 0\}$ is a null set, then $f \in \mathcal{H}([a,b])$ with integral 0. Try to prove this following the outline below. (Also ask yourself, "what is the purpose of step 2?")

1. Let $D_n = \{x \in [a,b] : |f(x)| \in (n-1, n]\}$. Argue that for each $n \in \mathbb{N}$, the set D_n is a null set.
2. Given $\epsilon > 0$, for each $n > 0$, there exists a countable cover $\mathcal{E}_n$ of D_n by open intervals whose total length is less than $\epsilon/(n2^{2+n})$. Define the entourage mapping η such that
   - If $f(x) = 0$, then $\eta(x) = (x-1, x+1)$.
   - If $x \in D_n$, then choose $\eta(x)$ arbitrarily from amount $\mathcal{E}_n$, provided that $x \in \eta(x)$.
3. Show that if $\mathcal{T}$ is η-fine, then $\mathsf{S}_{\mathcal{T}} f = \sum_{n\in\mathbb{N}} \sum_{(\tau,I)\in\mathcal{T}\wedge\tau\in D_n} f(\tau)|I|$; show that this sum cannot be greater than ϵ.

Exercise 11.5.

1. Prove that Proposition 9.22 still holds if we replace $\mathcal{R}([a,b])$ by $\mathcal{H}([a,b])$.
2. As a consequence, show that if $f \in \mathcal{H}([a,b])$, and g differs from f only on a null subset of $[a,b]$, then $g \in \mathcal{H}([a,b])$ and $\int_a^b f(x)\,dx = \int_a^b g(x)\,dx$.

A fantastic property of Henstock integration is that we can significantly strengthen one half of the fundamental theorem of calculus.

Theorem 11.6 (Fundamental Theorem of Calculus for Henstock Integrals)**.** *Let $f : [a,b] \to \mathbb{R}$ be differentiable on $[a,b]$. Then $f' \in \mathcal{H}([a,b])$ and $\int_a^b f'(x)\,dx = f(b) - f(a)$.*

Proof. Given $\epsilon > 0$, let $\epsilon' = \epsilon/(b-a)$. Since f is differentiable at every point, there is a function $r : [a,b] \to \mathbb{R}$ taking positive values so that $|f(y) - f(x) - f'(x)(y-x)| \le \frac{1}{2}\epsilon'|y-x|$ for all $y \in B(x, r(x))$. Let $\eta(x) = B(x, r(x))$; this is an entourage mapping. Given a tagged division $\mathcal{T}$, we can write $f(b) - f(a) = \sum_{(\tau,I)\in\mathcal{T}} f(\sup I) - f(\inf I)$. And hence

$$\left|f(b) - f(a) - \mathsf{S}_{\mathcal{T}} f'\right| \le \sum_{(\tau,I)\in\mathcal{T}} \left|f(\sup I) - f(\inf I) - f'(\tau)|I|\right|.$$

If $\mathcal{T}$ is η-fine, we have

$$f(\sup I) - f(\inf I) - f'(\tau)|I| = \underbrace{f(\sup I) - f(\tau) - f'(\tau)(\sup I - \tau)}_{|-| \le \frac{1}{2}\epsilon'(\sup I - \tau)} + \underbrace{f(\tau) - f(\inf I) - f'(\tau)(\tau - \inf I)}_{|-| \le \frac{1}{2}\epsilon'(\tau - \inf I)}.$$

This shows

$$\left|f(b) - f(a) - \mathsf{S}_{\mathcal{T}} f'\right| \le \sum_{(\tau,I)\in\mathcal{T}} \frac{1}{2}\epsilon'|I| = \frac{1}{2}\epsilon'(b-a) = \frac{1}{2}\epsilon < \epsilon.$$

And hence $\varsigma[f] \to f(b) - f(a)$ as claimed. □

Example 11.7. Let $f : [-1,1] \to \mathbb{R}$ be given by $f(x) = x^2 \sin(1/x^3)$ when $x \neq 0$, and $f(0) = 0$. This function is differentiable on the entirety of $[0,1]$; it has $f'(0) = 0$. However, the derivative is not continuous at 0. (See also the similar Example 8.15.) In fact f' is unbounded on any open interval containing 0. Therefore f' is not Riemann integrable. But by the previous theorem, we know that f' is Henstock integrable on any subinterval of $[-1,1]$. ■

Lemma 11.8. *Let $f : [a,b] \to \mathbb{R}$, and take $c \in [a,b]$. Suppose $f \in \mathcal{H}([a,c])$. Then*

$$f \in \mathcal{H}([a,b]) \iff f \in \mathcal{H}([c,b]).$$

And when either hold, we have $\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$.

Proof. Unlike the analogous statement for Riemann integrals, we don't have something so convenient as the Lebesgue criterion this time. So we have to roll up our sleeves and do things by hand.

First we prove the direction ($\Rightarrow$). From our assumptions, given any $\epsilon > 0$ there must exist some entourage mapping η such that if $\mathcal{T}_0$ is an η-fine tagged division of $[a,b]$, then $|\mathsf{S}_{\mathcal{T}_0} f - \int_a^b f(x)\,dx| < \epsilon/2$.

Similarly, by possible replacing η by a successor, we can assume that if $\mathcal{T}_1$ is an η-fine tagged division of $[a,c]$, then $|\mathsf{S}_{\mathcal{T}_1} f - \int_a^c f(x)\,dx| < \epsilon/2$.

Now let $\mathcal{T}_2$ be an arbitrary η-fine tagged division of $[c,b]$, and $\mathcal{T}_1$ be an η-fine tagged division of $[a,c]$. Then $\mathcal{T}_1 \cup \mathcal{T}_2$ forms an η-fine tagged division of $[a,b]$. Since $\mathsf{S}_{\mathcal{T}_1 \cup \mathcal{T}_2} f = \mathsf{S}_{\mathcal{T}_1} + \mathsf{S}_{\mathcal{T}_2}$, we find that the computations from the previous section shows that necessarily

$$\left|\mathsf{S}_{\mathcal{T}_2} f - \int_a^b f(x)\,dx + \int_a^c f(x)\,dx\right| < \epsilon,$$

and we establish both integrability and the value of the integral.

For the reverse direction ($\Leftarrow$), it is not the case that an arbitrary tagged division of $[a,b]$ splits nicely into a tagged division of $[a,c]$, together with another one of $[c,b]$. So a bit more work is required.

Let η_c be an entourage mapping satisfying

$$\eta_c(x) = \begin{cases} (a-1, c) & x \in [a,c), \\ (c-1, c+1) & x = c, \\ (c, b+1) & x \in (c,b]. \end{cases}$$

Suppose $f \in \mathcal{H}([a,c]) \cap \mathcal{H}([c,b])$. Then for every $\epsilon > 0$, there exists some entourage mapping η_a such that for any η_a-fine tagged division $\mathcal{T}_a$ of $[a,c]$ we have $|\mathsf{S}_{\mathcal{T}_a} f - \int_a^c f\,dx| < \epsilon/2$. Similarly we can define η_b. Let η be an entourage mapping that succeeds all three of η_a, η_b, and η_c.

Suppose $\mathcal{T}$ is a tagged division of $[a,b]$ that is η-fine. By the construction, this means $\mathcal{T}$ is also η_c-fine, and hence any tagged subinterval (τ, I) with $c \in I$ must have $\tau = c$. Therefore, we can define $\mathcal{T}'$ such that any tagged subinterval (τ, I) with $c \in I$ is replaced by the pair of subintervals $(c, I \cap [a,c])$ and $(c, I \cap [c,b])$. This construction is such that $\mathsf{S}_{\mathcal{T}'} f = \mathsf{S}_{\mathcal{T}} f$, and $\mathcal{T}'$ is still η-fine.

On the other hand, every tagged subinterval in $\mathcal{T}'$ is such that it is either entirely contained in $[a,c]$, or it is entirely contained in $[c,b]$. Therefore we can write $\mathcal{T}' = \mathcal{T}_a \cup \mathcal{T}_b$ such that $\mathcal{T}_a$ is an η-fine tagged subdivision of $[a,c]$, and $\mathcal{T}_b$ an η-fine tagged subdivision of $[c,b]$. Since $\eta_a, \eta_b \leqq \eta$, we get that

$$\left|\mathsf{S}_{\mathcal{T}'} f - \int_a^c f\,dx - \int_c^b f\,dx\right| = \left|\mathsf{S}_{\mathcal{T}_a} f - \int_a^c f\,dx + \mathsf{S}_{\mathcal{T}_b} f - \int_c^b f\,dx\right| < \epsilon.$$

This shows that $f \in \mathcal{H}([a,b])$ with $\int_a^c f\,dx + \int_c^b f\,dx$ as its integral. □

As discussed earlier in Example 11.1: Riemann integrability is not preserved under "taking limits of domain". More precisely, given $f : [a,b] \to \mathbb{R}$ such that $f \in \mathcal{R}([a,c])$ for every $c \in [a,b)$, and even assuming $\lim_{c \to b^-} \int_a^c f(x)\,dx$ converges, this does not imply $f \in \mathcal{R}([a,b])$. This turns out to not be an issue for Henstock integrals. (So there is no such thing as an "improper Henstock integral on a bounded interval".)

Theorem 11.9 (Hake's Theorem)**.** *Given $f : [a,b] \to \mathbb{R}$ such that $f \in \mathcal{H}([a,c])$ for every $c \in [a,b)$, and such that $\lim_{c \to b^-} \int_a^c f(x)\,dx = L$ converges, then $f \in \mathcal{H}([a,b])$ with $\int_a^b f(x)\,dx = L$.*

Proof. This proof is a bit technical. So I will break it into parts.

Preliminary Steps Let $\xi : \mathbb{N} \to [a,b)$ be a strictly increasing sequence that converges to b, with $\xi_1 = a$. The precise choice of the ξ_n does not matter much, but it gives us a concrete foundation to build upon. By Lemma 11.8 and induction, we have that $f \in \mathcal{H}([\xi_n, \xi_{n+1}])$ for every n.

Consider the function $\tilde{f}$ which is identical to f on $[a,b)$ but with $\tilde{f}(b) = 0$. This function is also in $\mathcal{H}([a,c])$ for every $c \in [a,b)$, and with the same limit L as f. By Exercise 11.5 part 2, since a singleton set is a null set, either f and $\tilde{f}$ are both Henstock integrable, or neither of them are. And when they are integrable, they would have the same integral. In other words, the precise value of $f(b)$ does not matter for this Theorem. So we can assume without loss of generality that $f(b) = 0$.

Strategy Outline Given an $\epsilon > 0$, we will construct an entourage function η defined on $[a,b]$ such that any η-fine tagged division $\mathcal{T}$ would satisfy $|\mathsf{S}_{\mathcal{T}} f - L| < \epsilon$. We will build η in such a way that any η-fine tagged division $\mathcal{T}$ of $[a,b]$ restrict to each of the $[\xi_n, \xi_{n+1}]$ subintervals as a tagged division of it. Furthermore, using the fact that $f \in \mathcal{H}([\xi_n, \xi_{n+1}])$, we can build η such that this tagged division of $[\xi_n, \xi_{n+1}]$ is such that the difference between the Riemann sum and the integral is small compared to ϵ, in a way that is summable (we can add up the errors from each of the sub-intervals and still get something less than ϵ). The tricky part, however, is that since $\mathcal{T}$ is finite, there will be one final subinterval of it that doesn't follow the above mould. And most of the analysis in this proof has to do with that one subinterval.

Construct η Let $\epsilon > 0$. For each $n \in \mathbb{N}$, there exists an entourage function η_n such that for every η_n-fine tagged division $\mathcal{T}_n$ of $[\xi_n, \xi_{n+1}]$, we have

$$\left| \mathsf{S}_{\mathcal{T}_n} f - \int_{\xi_n}^{\xi_{n+1}} f(x)\, dx \right| < \frac{1}{2^{n+2}} \epsilon. \tag{11.4}$$

Additionally, by the convergence of $\int_a^c f(x)\, dx$ to L as $c \to b$ from below, there exists $N \in \mathbb{N}$ such that:

$$\text{for every } k \geq N, \quad \left| \int_a^{\xi_k} f(x)\, dx - L \right| < \frac{1}{4}\epsilon \tag{11.5}$$

$$\text{for every } y' \geq y \geq \xi_N, \quad \left| \int_y^{y'} f(x)\, dx \right| < \frac{1}{16}\epsilon. \tag{11.6}$$

Define our η by requiring that

$$\eta(x) = \begin{cases} \eta_1(a) \cap (a-1, \xi_2) & x = a, \\ (\xi_{n-1}, \xi_{n+1}) \cap \eta_n(\xi_n) \cap \eta_{n-1}(\xi_n) & x = \xi_n, \\ (\xi_n, \xi_{n+1}) & x \in (\xi_n, \xi_{n+1}), \\ (\xi_N, b+1) & x = b. \end{cases} \tag{11.7}$$

This construction mirrors the use of the "common successor to η_a, η_b, η_c" construction in the proof of Lemma 11.8. The key property is that if a tagged division $\mathcal{T}$ of $[a,b]$ is η-fine, then any tagged

subinterval that contains one of the ξ_n must have ξ_n as its tag, *unless it also contains* b. And hence the same procedure as in the proof of Lemma 11.8 means that if any tagged subinterval (τ, I) contains one of ξ_n, and not b, we can replace it by the *two* tagged subintervals $(\xi_n, I\cap[\xi_{n-1},\xi_n])$ and $(\xi_n, I\cap[\xi_n,\xi_{n+1}])$ and create a new tagged division *with the same Riemann sum as* $\mathcal{T}$. Since we are interested in the value of $\mathsf{S}_{\mathcal{T}} f$, we can assume that this replacement has already been made, and every tagged subinterval that contains ξ_n but not b are either entirely contained in $[\xi_{n-1},\xi_n]$ or $[\xi_n,\xi_{n+1}]$.

Analysis of $\mathcal{T}$ Now let $\mathcal{T}$ be a tagged division of $[a,b]$ that is η-fine, with the replacement procedure of the previous paragraph completed. By design $\eta(x) \not\ni b$ for any $x \neq b$; and since $\mathcal{T}$ is finite, this means that one element of $\mathcal{T}$ is (b, I_b), where I_b is non-degenerate. This means that there exists some $K \geq N$ such that $\xi_K < \inf I_b \leq \xi_{K+1}$.

Thus $\mathcal{T}$ can be grouped into the union of

$$\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_{K-1} \cup \mathcal{S}_K \cup \{(b, I_b)\}.$$

Here $\mathcal{T}_n$ is a tagged division of $[\xi_n, \xi_{n+1}]$. The odd set out $\mathcal{S}_K$ contains all those tagged subintervals within $[\xi_K, \xi_{K+1}]$. When $\inf I_b < \xi_{K+1}$, those don't form a tagged division of the whole subinterval $[\xi_K, \xi_{K+1}]$. The Riemann sum of $\mathcal{T}$ thus evaluates to

$$\mathsf{S}_{\mathcal{T}} f = \sum_{n=1}^{K-1} \mathsf{S}_{\mathcal{T}_n} f + \mathsf{S}_{\mathcal{S}_K} f. \tag{11.8}$$

Notice that we can omit the final interval since the tag is at b and we have assumed that $f(b) = 0$.

Applying the triangle inequality we can write

$$\begin{aligned}\left|\mathsf{S}_{\mathcal{T}} f - L\right| &= \left|\sum_{n=1}^{K-1}\left(\mathsf{S}_{\mathcal{T}_n} f - \int_{\xi_n}^{\xi_{n+1}} f(x)\,dx\right) + \int_a^{\xi_K} f(x)\,dx - L + \mathsf{S}_{\mathcal{S}_K} f\right| \\ &\leq \underbrace{\sum_{n=1}^{K-1}\left|\mathsf{S}_{\mathcal{T}_n} f - \int_{\xi_n}^{\xi_{n+1}} f(x)\,dx\right|}_{<\frac{1}{4}\epsilon} + \underbrace{\left|\int_a^{\xi_K} f(x)\,dx - L\right|}_{<\frac{1}{4}\epsilon} + \left|\mathsf{S}_{\mathcal{S}_K} f\right|. \end{aligned} \tag{11.9}$$

It remains to prove that $|\mathsf{S}_{\mathcal{S}_K} f| < \frac{1}{2}\epsilon$. To do so, notice that $\mathcal{S}_K$ is a tagged division of $[\xi_K, \inf I_b]$. Since f is also Henstock integrable on $[\inf I_b, \xi_{K+1}]$, there exists a tagged division $\mathcal{S}'_K$ of $[\inf I_b, \xi_{K+1}]$ such that it is both η-fine *and* satisfies

$$\left|\int_{\inf I_b}^{\xi_{K+1}} f(x)\,dx - \mathsf{S}_{\mathcal{S}'_K} f\right| < \frac{1}{16}\epsilon.$$

The union $\mathcal{S}_K \cup \mathcal{S}'_K$ is a tagged division of $[\xi_K, \xi_{K+1}]$ that is η_K-fine; and hence by (11.4) we have that

$$\left|\int_{\xi_K}^{\xi_{K+1}} f(x)\,dx - \mathsf{S}_{\mathcal{S}_K \cup \mathcal{S}'_K} f(x)\right| < \frac{1}{2^{K+2}}\epsilon \leq \frac{1}{8}\epsilon.$$

So writing

$$\mathsf{S}_{\mathcal{S}_K} f = \underbrace{\mathsf{S}_{\mathcal{S}_K \cup \mathcal{S}'_K} f - \int_{\xi_K}^{\xi_{K+1}} f(x)\,dx}_{|-|<\frac{1}{8}\epsilon} + \underbrace{\int_{\xi_K}^{\inf I_b} f(x)\,dx}_{|-|<\frac{1}{16}\epsilon} + \underbrace{\int_{\inf I_b}^{\xi_{K+1}} f(x)\,dx - \mathsf{S}_{\mathcal{S}'_K} f}_{|-|<\frac{1}{16}\epsilon},$$

where for the middle term we used (11.6), we find $|\mathsf{S}_{\mathcal{S}_K} f| < \frac{1}{4}\epsilon$, and the theorem is proved. □

Example 11.10. As a consequence, the function $f(x) = 1/\sqrt{x}$ on $(0,1]$ with $f(0) = 1$ is an element of $\mathcal{H}([0,1])$: on any interval $(c,1]$ with $c \in (0,1)$ we find $f \in \mathcal{R}([c,1])$ since it is continuous and bounded; and hence $f \in \mathcal{H}([c,1])$ also with the same $\int_c^1 f(x)\,dx$. Since the improper integral $\int_0^1 f(x)\,dx$ is well-defined, applying the previous theorem we find $f \in \mathcal{H}([0,1])$. ■

The passage from Riemann to Henstock integration is, however, not a situation with only gains and no losses. A consequence of Proposition 9.20 is that if $g : \mathbb{R} \to \mathbb{R}$ is continuous, then given any Riemann integrable $f : [a,b] \to \mathbb{R}$, the function $g \circ f \in \mathcal{R}([a,b])$. Theorem 11.9 shows however the same statement cannot be true for Henstock integrable functions.

Exercise 11.6. Let f be as in Example 11.10, where we showed $f \in \mathcal{H}([0,1])$. Take $g(x) = x^3$. Prove that $g \circ f \notin \mathcal{H}([0,1])$.

One may try to argue that problem in the previous exercise is caused by g not being uniformly continuous. The following example puts that argument to rest.

Example 11.11. Consider the function $h(x) = -3x^{-2}\cos(1/x^3)$ on $(0,1]$ and $h(0) = 0$. Observe that $h(x) + 2x\sin(1/x^3) = f'(x)$ where f is as in Example 11.7. Since $2x\sin(1/x^3)$ is continuous on $[0,1]$, it is Riemann integrable there; since f' is Henstock integrable on $[0,1]$, by Exercise 11.5 part 1, we see that $h \in \mathcal{H}([0,1])$. However, $|h| \notin \mathcal{H}([0,1])$. The integral $\int_c^1 |h(x)|\,dx$ has size approximately c^{-1} and diverges as $c \to 0$ from above. ■

Food for Thought 11.7. In some ways, the relation between the Riemann integral and the Henstock integral is analogous to the relation between an absolutely convergent sum of countably infinitely many numbers, versus the convergent series associated to a particular enumeration (see Week 6). The convergence of an absolutely convergent sum is described by the convergence of some net; the series associated to an enumeration is given by the convergence of some subnet. And as we have seen in Proposition 11.3, the Henstock integral is given in terms of a subnet $\varsigma[f]$ of the net $\rho[f]$ that defines Riemann integration. You have proven if a sum is absolutely convergent, then the sum of absolute values also converges; this is well-known to be false in general for convergent series. Here we see that Riemann integrable functions are such that their absolute values are also Riemann integrable, but this breaks down when talking about Henstock integrals.

§11.2 The Stieltjes Integral

In this final section of the course, we will change tack a bit and look at the Stieltjes integral. Whereas the Henstock Integral looks at what would happen if we used a different way of ordering tagged divisions, the Stieltjes Integral looks at what would happen if we used a different way of performing Riemann sums. As such, the Stieltjes modification to integration *can be applied to both the Riemann and Henstock integrals*.

§11.2.1 Basic definition.—In the standard Riemann sum, the value of the tag τ is the value $f(\tau)$, and the weight assigned to each tag is the width of the interval I that it represents. The Stieltjes modification asks: what if we assign, to each subinterval $I \subseteq [a,b]$, a different weight? There are various degrees of generalities at which one can pursue this angle: if one makes the assignment of weights essentially arbitrary, one gains a lot of freedom of what can be represented by the Riemann sum, but it becomes a lot harder to understand the convergence of the Riemann sums. The most commonly considered case assigns this weight with the help of an increasing function.

Definition 11.12. *Let $\alpha : \mathbb{R} \to \mathbb{R}$ be an increasing (not necessarily strict!) function, and let $\mathcal{T}$ be a tagged division of some closed, bounded interval $[a,b]$. Given a function $f : [a,b] \to \mathbb{R}$, the* Riemann-Stieltjes sum with weight *α with respect to $\mathcal{T}$ of f is*

$$\mathsf{S}_{\mathcal{T}}^{(\alpha)} f := \sum_{(\tau,I)\in\mathcal{T}} f(\tau)(\alpha(\sup I) - \alpha(\inf I)).$$

Food for Thought 11.8.

1. The Riemann sum $\mathsf{S}_{\mathcal{T}} f$ can be obtained as the special case $\alpha(x) = x$.
2. Verify that if α is a monotone function, and $\mathcal{T}$ is a tagged division of $[a,b]$, then
$$\sum_{(\tau,I)\in\mathcal{T}} |\alpha(\sup I) - \alpha(\inf I)| = |\alpha(b) - \alpha(a)|.$$
3. Prove that if α and β are two increasing functions, with $\alpha - \beta$ being constant, then $\mathsf{S}_{\mathcal{T}}^{(\alpha)} f = \mathsf{S}_{\mathcal{T}}^{(\beta)} f$.
4. Prove that if $\beta = \lambda\alpha$ with $\lambda \geq 0$, then $\mathsf{S}_{\mathcal{T}}^{(\beta)} f = \lambda \mathsf{S}_{\mathcal{T}}^{(\alpha)} f$.
5. The assumption that α is increasing is for simplicity of exposition. Much of the theory developed below can be applied also to the cases where α is a decreasing function also. In fact, this can also be extended to the case of functions $\alpha : \mathbb{R} \to \mathbb{R}$ that admit a decomposition as a sum $\alpha_+ + \alpha_-$, where α_+ is increasing, and α_- is decreasing. (Functions admitting this decomposition are said to have "bounded variation".)

We can analogously define the nets $\rho[f]^{(\alpha)} : \mathfrak{r}([a,b]) \to \mathbb{R}$ and $\varsigma[f]^{(\alpha)} : \mathfrak{h}([a,b]) \to \mathbb{R}$ using the Riemann-Stieltjes sum in place of the Riemann sum. When they converge, we say that f is "Riemann-Stieltjes integrable (with weight α)" and "Henstock-Stieltjes integrable (with weight α)" respectively, and use the symbolic shorthands $f \in \mathcal{R}([a,b],\alpha)$ and $f \in \mathcal{H}([a,b],\alpha)$ for the two cases. Notice that since the definition of the subnet only depends on the transition mapping φ, the exact same proof of Proposition 11.3 shows that for any fixed α, the net $\varsigma[f]^{(\alpha)}$ is again a subnet of $\rho[f]^{(\alpha)}$, and hence we still have the inclusion

$$\mathcal{R}([a,b],\alpha) \subseteq \mathcal{H}([a,b],\alpha). \tag{11.10}$$

To differentiate the Stieltjes integrals from their Riemann/Henstock counterparts, we will denote the limit of $\varsigma[f]^{(\alpha)}$ and $\rho[f]^{(\alpha)}$, when they converge, with the notation

$$\int_a^b f(x)\, d\alpha.$$

Note that instead of using dx to denote the integration element, we now use $d\alpha$.

Food for Thought 11.9. As a consequence of the definition in terms of nets, Proposition 9.22 also generalizes to $\mathcal{R}([a,b],\alpha)$ and $\mathcal{H}([a,b],\alpha)$ functions.

Exercise 11.10. Prove that if α,β are two increasing functions on $\mathbb{R}$, and $f\in\mathcal{R}([a,b],\alpha)\cap\mathcal{R}([a,b],\beta)$, then $f\in\mathcal{R}([a,b],\alpha+\beta)$ with $\int_a^b f(x)\,d(\alpha+\beta)=\int_a^b f(x)\,d\alpha+\int_a^b f(x)\,d\beta$. The same statement is true for the Henstock-Stieltjes integrals.

§11.2.2 Basic properties of the Riemann-Stieltjes integral.—For simplicity of discussion, we will only consider the Riemann-Stieltjes integrals in this and the next section.

The precise sets of integrable functions in the Riemann-Stieltjes sense depend strongly on the weight function α. A convenient and common set of integrable functions are those f which are continuous on $[a,b]$.

Theorem 11.13. *If f is continuous on $[a,b]$, then $f\in\mathcal{R}([a,b],\alpha)$ for any increasing function α.*

We omit the proof here; it is largely the same as the proof of Theorem 9.13.

Proposition 11.14. *If there exists $c\in(a,b)$ such that both α and f are discontinuous at c, then $f\notin\mathcal{R}([a,b],\alpha)$.*

Proof. It suffices to show that there exists some $\epsilon>0$, such that for every $\delta>0$, there exists tagged divisions $\mathcal{T},\mathcal{T}'$, both with width less than δ, such that $|\mathsf{S}^{(\alpha)}_{\mathcal{T}}f-\mathsf{S}^{(\alpha)}_{\mathcal{T}'}f|>\epsilon$.

We will choose $\epsilon=\frac{1}{2}|\omega_f(c)\cdot\omega_\alpha(c)|>0$ (cf. (9.5)). For any $\delta>0$, we can choose a tagged division $\mathcal{T}_0$ of $[a,b]$, with $|\mathcal{T}_0|<\delta$, such that c is not on the boundary of any I; therefore c is contained in exactly one interval I_c of $\mathcal{T}_0$. We will define $\mathcal{T},\mathcal{T}'$ to be identical to $\mathcal{T}_0$ except for their tags τ_c and τ_c' for the interval I_c. Notice that

$$|\mathsf{S}^{(\alpha)}_{\mathcal{T}}f-\mathsf{S}^{(\alpha)}_{\mathcal{T}'}f|=|f(\tau_c)-f(\tau_c')|\cdot[\alpha(\sup I_c)-\alpha(\inf I_c)].$$

Since α is monotone, the difference $\alpha(\sup I_c)-\alpha(\inf I_c)\geq\omega_\alpha(c)$. Since I_c contains an open interval around c, by the definition (9.5), there exists $y_+,y_-\in I_c$ such that $|f(y_+)-f(y_-)|>\frac{1}{2}\omega_f(c)$. Set τ_c to be one and τ_c' to be the other. The desired conclusion follows. □

Exercise 11.11. Extend Proposition 11.14 to the case where the common discontinuity point c occurs at the endpoint (either a or b).

One of the main uses of the Stieltjes formulation is to allow "atoms" in the integral. When defining the Riemann integral, the value of f at any specific point doesn't matter.

Example 11.15. If $f\in\mathcal{R}([a,b])$ and $g:[a,b]\to\mathbb{R}$ is equal to f except at a point $c\in[a,b]$, then we note that if $\mathcal{T}$ is a tagged division with width less than δ, then at worst $|\mathsf{S}_{\mathcal{T}}f-\mathsf{S}_{\mathcal{T}}g|<\delta|f(c)-g(c)|$. By choosing smaller and smaller δ, we see that $\rho[g]$ converges to $\int_a^b f(x)\,dx$ also. ■

The presence of jump discontinuities in the weight α changes the argument. Firstly, from Proposition 11.14 we see that if α has a jump discontinuity at c, then changing the value of f at c will turn a $\mathcal{R}([a,b],\alpha)$ function into one that is not integrable. Secondly, we have the following result.

Proposition 11.16. *Let α be an increasing function such that $\alpha(x)=0$ when $x<0$ and $\alpha(x)=1$ when $x>0$.*

Then for any $a < 0 < b$ and $f : [a,b] \to \mathbb{R}$ that is continuous at 0, we have

$$\int_a^b f(x)\, d\alpha = f(0).$$

Proof. Let $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(0)| < \epsilon$ for every $x \in (-\delta, \delta)$. Let $\mathcal{T}$ be a tagged division with width $< \delta$. Then either:

- There exists a tagged subinterval (τ_0, I_0) with $\inf I < 0 < \sup I$. In this case $\mathsf{S}_{\mathcal{T}}^{(\alpha)} f = f(\tau_0)$, which is within ϵ of $f(0)$.
- There exists tagged subintervals (τ_-, I_-) and (τ_+, I_+) with $\sup I_- = \inf I_+ = 0$. In this case $\mathsf{S}_{\mathcal{T}}^{(\alpha)} f = \alpha(0) f(\tau_-) + (1 - \alpha(0)) f(\tau_+)$ is between $f(\tau_-)$ and $f(\tau_+)$. Since both of these values are within ϵ of $f(x)$, we conclude that so is $\mathsf{S}_{\mathcal{T}}^{(\alpha)} f$. □

Food for Thought 11.12. This actually makes the Riemann-Stieltjes integral a good setting for studying *probability theory*. The domain $[a,b]$ represents the continuum of possible outcomes of an "experiment", and the increasing function α with $\alpha(a) = 0$ and $\alpha(b) = 1$ is the "cumulative distribution function". Given an interval $I \subseteq [a,b]$, the value $\sup \alpha(I) - \inf \alpha(I) \in [0,1]$ is the probability the outcome of the experiment lies within the interval I.

The ability to have atoms means that we can gather both continuous and discrete probabilities under one umbrella. For example, for the throw of a regular six-sided die, we can let the domain of possible outcomes be the interval $[0,6]$, with $\alpha(x) = \frac{1}{6}\lfloor x \rfloor$, where the $\lfloor x \rfloor$ is the "integer part" function.

We can let $f : [a,b] \to \mathbb{R}$ represent the "payout" one receives, depending on the possible outcomes of the experiment. And in this sense the integral $\int_a^b f(x)\, d\alpha$ would represent the "expected compensation" given the probability distribution α and the payout function f.

§11.2.3 Change of variables and integration by parts.—We conclude with several theorems on practical applications of the Riemann-Stieltjes integral. The first few codify the change of variables formula (a.k.a. u-substitution) you've seen in a calculus course.

Theorem 11.17. *Let $\alpha : [a,b] \to \mathbb{R}$ be increasing and continuous, and $f : [c,d] \to \mathbb{R}$, where the interval $[c,d] = \alpha([a,b])$. If $f \in \mathcal{R}([c,d])$, then $f \circ \alpha \in \mathcal{R}([a,b], \alpha)$, with*

$$\int_a^b f \circ \alpha \; d\alpha = \int_c^d f(x)\, dx.$$

Proof. Given $\delta > 0$, since α is continuous on $[a,b]$, it is uniformly continuous, and there is $\delta' > 0$ such that if $x, y \in [a,b]$ has $|x - y| < \delta'$, then $|\alpha(x) - \alpha(y)| < \delta$.

Thus if $\mathcal{T}'$ is a tagged division of $[a,b]$ with width $|\mathcal{T}'| < \delta'$, then

$$\mathcal{T} = \{(\alpha(\tau'), \alpha(I')) : (\tau', I') \in \mathcal{T}'\}$$

is a tagged division of $[c,d]$ with width $|\mathcal{T}| < \delta$.

Observe that $\mathsf{S}_{\mathcal{T}} f = \mathsf{S}_{\mathcal{T}'}^{(\alpha)}(f \circ \alpha)$. This implies that $\rho[f \circ \alpha]^{(\alpha)}$ is a subnet (in the Kelley sense, not the Willard sense) of $\rho[f]$, and hence the desired conclusion follows. (Since we have only been working

with Willard subnets so far, if you don't feel confident in this argument, use the fact that since $\rho[f]$ converges, for every $\epsilon > 0$ there exists $\delta > 0$ such that if $|\mathcal{T}| < \delta$ then $|\mathsf{S}_{\mathcal{T}} f - \int_c^d f(x)\,dx| < \epsilon$. Use the results from the previous two paragraphs to show that there exists $\delta' > 0$ such that if $|\mathcal{T}'| < \delta'$ then $|\mathsf{S}^{(\alpha)}_{\mathcal{T}'}(f \circ \alpha) - \int_c^d f(x)\,dx| < \epsilon$ also.) □

Food for Thought 11.13. Theorem 11.17 also holds for Henstock-Stieltjes integrals: the conclusion can be modified to read "if $f \in \mathcal{H}([c,d])$, then $f \circ \alpha \in \mathcal{H}([a,b],\alpha)$." We can modify the proof as follows: We observe that if η is an entourage function defined on $[c,d]$, then $\eta' := \alpha^{-1} \circ \eta \circ \alpha$ is an entourage function defined on $[a,b]$. (Here the α^{-1} is the induced power-set mapping which maps open intervals to open intervals.) From this, we see that from every η'-fine tagged division $\mathcal{T}'$ of $[a,b]$, the constructed $\mathcal{T}$ must also be a η-fine tagged division of $[c,d]$. And this shows that $\varsigma[f \circ \alpha]^{(\alpha)}$ is a subnet of $\varsigma[f]$ and the proof follows.

The following Theorem should be contrasted against Problem 9.3 on the Week 9 Problem Set.

Theorem 11.18. *Let $\alpha : [a,b] \to \mathbb{R}$ be increasing and differentiable, with $\alpha' \in \mathcal{R}([a,b])$; and let $f : [a,b] \to \mathbb{R}$ be a bounded function. Then $f \in \mathcal{R}([a,b],\alpha)$ if and only if $f \cdot \alpha' \in \mathcal{R}([a,b])$; when this holds, the integrals*

$$\int_a^b f(x)\alpha'(x)\,dx = \int_a^b f(x)\,d\alpha.$$

Proof. Notice that the nets $\rho[f]^{(\alpha)}$ and $\rho[f \cdot \alpha']$ are indexed by the same set $\mathfrak{r}([a,b])$. The theorem follows if we can show that for every $\epsilon > 0$, there exists $\delta > 0$ such that for every tagged division $\mathcal{T}$, the sums $|\mathsf{S}^{(\alpha)}_{\mathcal{T}} f - \mathsf{S}_{\mathcal{T}}(f \cdot \alpha')| < \epsilon$. This will imply that the two nets either both converge or both diverge, and when they converge, they converge to the same limit.

Since α is differentiable, by the Mean Value Theorem (Corollary 8.19) for each I there is a $\gamma(I) \in I$ such that $\alpha(\sup I) - \alpha(\inf I) = \alpha'(\gamma(I))|I|$. So we can write

$$\mathsf{S}^{(\alpha)}_{\mathcal{T}} f - \mathsf{S}_{\mathcal{T}}(f \cdot \alpha') = \sum_{(\tau,I)\in\mathcal{T}} f(\tau) \cdot \Big(\alpha(\sup I) - \alpha(\inf I) - \alpha'(\tau)|I|\Big) = \sum_{(\tau,I)\in\mathcal{T}} f(\tau) \cdot \Big(\alpha'(\gamma(I)) - \alpha'(\tau)\Big)|I|.$$

Since we assumed f is a bounded function, we can set $M = \sup|f|([a,b])$. Then by the triangle inequality we find

$$\Big|\mathsf{S}^{(\alpha)}_{\mathcal{T}} f - \mathsf{S}_{\mathcal{T}}(f \cdot \alpha')\Big| \le \sum_{(\tau,I)\in\mathcal{T}} M \cdot |\alpha'(\gamma(I)) - \alpha'(\tau)| \cdot |I| \le M \cdot \Big(\overline{\mathsf{S}}_{\mathcal{T}}\alpha' - \underline{\mathsf{S}}_{\mathcal{T}}\alpha'\Big). \tag{11.11}$$

Now, given $\epsilon > 0$, by Darboux's Criterion (Theorem 9.12) applied to the integrable function α' there exists $\delta > 0$ such that if $|\mathcal{T}| < \delta$, the difference $\overline{\mathsf{S}}_{\mathcal{T}}\alpha' - \underline{\mathsf{S}}_{\mathcal{T}}\alpha' < \epsilon/M$. Plugging this into (11.11) we find the needed estimate. □

Combining the previous two theorems, we obtain the following result. When f is continuous, this can also be derived using the chain rule for differentiation and the fundamental theorems of calculus; here we can address also the case where f is integrable but not continuous.

Corollary 11.19 (Change of Variables). *Let $u \in \mathcal{C}^1([a,b];\mathbb{R})$, with $u'(x) > 0$ for all $x \in [a,b]$. Then*

$$f \in \mathcal{R}(u([a,b])) \iff (f \circ u) \cdot u' \in \mathcal{R}([a,b])$$

and $\int_{u(a)}^{u(b)} f(x)\,dx = \int_a^b f(u(x)) \cdot u'(x)\,dx$.

Proof. ($\Rightarrow$) Since $u'(x) > 0$, the function u is increasing by Corollary 8.24. And since u' is continuous, it is in $\mathcal{R}([a,b])$. And the result follows after chaining together Theorems 11.17 and 11.18.

($\Leftarrow$) By Theorem 8.25, there exists a differentiable function $v : u([a,b]) \to [a,b]$ such that $u \circ v(x) = x$, with $v' = 1/u' \circ v$. The continuity and positive of u' implies that $v' > 0$ and v' is continuous. And so v is increasing and $v' \in \mathcal{R}([a,b])$. Denote by $g(x) = f(u(x)) \cdot u'(x)$. Then $g(v(x))v'(x) = f(x)$, and hence the result follows by applying the ($\Rightarrow$) direction to the pair g, v. □

We will end our short tour of the Stieltjes integral with a generalization of the integration by parts formula from Theorem 10.10. Previously we required that F and G can be expressed as the indefinite Riemann integrals of f and g. Using Stieltjes-integrals we can make sense of integration by parts formula for arbitrary increasing, continuous functions. (In fact, using Food for Thought 11.8, part 5, this can also be extended to arbitrary continuous functions of bounded variation.)

Theorem 11.20. *If $\mu, \nu : [a,b] \to \mathbb{R}$ are both continuous and increasing, then*

$$\mu(b)\nu(b) - \mu(a)\nu(a) = \int_a^b \mu(x)\,d\nu + \int_a^b \nu(x)\,d\mu.$$

Proof. First we remark that by Theorem 11.13, the integrals make sense. So all it is required is to check that the values of the two sides are equal. We will follow largely the strategy of our proof of Theorem 10.10.

First, notice that if one of μ or ν is constant, then the result follows directly. Suppose $\mu(x) = \mu_0$ is constant. Then the Riemann-Stieltjes sum $\mathsf{S}_{\mathcal{T}}^{(\mu)} f$ for any f vanishes. Then the desired equality reduces to checking

$$\mu_0(\nu(b) - \nu(a)) = \mu_0 \int_a^b d\nu$$

which follows directly from the definition of the Riemann-Stieltjes integral. So we can assume that $\mu(b) - \mu(a)$ and $\nu(b) - \nu(a)$ are both positive from here on.

It suffices to show that for every ϵ, there exists $\delta > 0$ such that whenever $\mathcal{T}$ has width less than δ, then

$$\left|\mu(b)\nu(b) - \mu(a)\nu(a) - \mathsf{S}_{\mathcal{T}}^{(\nu)}\mu - \mathsf{S}_{\mathcal{T}}^{(\mu)}\nu\right| < \epsilon.$$

To do so, we first rewrite $\mu(b)\nu(b) - \mu(a)\nu(a)$ using a telescoping sum, with respect to an arbitrary tagged division:

$$\mu(b)\nu(b) - \mu(a)\nu(a) = \sum_{(\tau,I)\in\mathcal{T}} \Big(\mu(\sup I) - \mu(\inf I)\Big)\nu(\sup I) + \mu(\inf I)\Big(\nu(\sup I) - \nu(\inf I)\Big).$$

And so we have

$$\mu(b)\nu(b)-\mu(a)\nu(a)-\mathsf{S}_{\mathcal{T}}^{(\nu)}\mu-\mathsf{S}_{\mathcal{T}}^{(\mu)}\nu = \sum_{(\tau,I)\in\mathcal{T}}\Big(\mu(\sup I)-\mu(\inf I)\Big)\Big(\nu(\sup I)-\nu(\tau)\Big)+\Big(\mu(\inf I)-\mu(\tau)\Big)\Big(\nu(\sup I)-\nu(\inf I)\Big). \quad (11.12)$$

Now, given $\epsilon>0$, by the continuity (and hence uniform continuity) of μ and ν we find that there exists some $\delta>0$ such that whenever $|x-y|<\delta$,

$$|\mu(x)-\mu(y)|<\frac{\epsilon}{2(\nu(b)-\nu(a))}, \qquad |\nu(x)-\nu(y)|<\frac{\epsilon}{2(\mu(b)-\mu(a))}.$$

So if $|\mathcal{T}|<\delta$, by (11.12) we find

$$\begin{aligned}\Big|\mu(b)\nu(b)-\mu(a)\nu(a)-\mathsf{S}_{\mathcal{T}}^{(\nu)}\mu-\mathsf{S}_{\mathcal{T}}^{(\mu)}\nu\Big| &\le \sum_{(\tau,I)\in\mathcal{T}}\Big(\mu(\sup I)-\mu(\inf I)\Big)\Big|\nu(\sup I)-\nu(\tau)\Big|+\Big|\mu(\inf I)-\mu(\tau)\Big|\Big(\nu(\sup I)-\nu(\inf I)\Big)\\ &< \frac{\epsilon}{2(\mu(b)-\mu(a))}\sum_{(\tau,I)\in\mathcal{T}}\Big(\mu(\sup I)-\mu(\inf I)\Big)+\frac{\epsilon}{2(\nu(b)-\nu(a))}\sum_{(\tau,I)\in\mathcal{T}}\Big(\nu(\sup I)-\nu(\inf I)\Big)=\epsilon.\end{aligned}$$

This concludes the proof. □

Problem Set 11

MTH 327H: Honors Intro to Analysis *(Fall 2020)* **Willie WY Wong**

In Example 11.11, it is asserted that $\int_c^1 |h(x)|\,dx$ has size approximately c^{-1} and hence $|h| \notin \mathcal{H}([0,1])$. Let's complete that argument in the next two problems.

Problem 11.1 (Converse of Hake's Theorem)**.** Given $f : [a,b] \to \mathbb{R}$, such that for every $c \in [a,b)$, we have $f \in \mathcal{H}([a,c])$. Let $L_c = \int_a^c f(x)\,dx$. Suppose the limit as $c \to b$ from below of L_c does not exist. Prove that $f \notin \mathcal{H}([a,b])$.

Hint: if $\lim L_c$ doesn't exist, then either it is unbounded, or it is bounded but with $\limsup L_c \neq \liminf L_c$. Your goal is to show that

- (if L_c is unbounded) for every entourage mapping η and for every $M > 0$, there exists an η-fine tagged division $\mathcal{T}$ such that $|\mathsf{S}_{\mathcal{T}} f| > M$.
- (if L_c is bounded) there exists some $\epsilon > 0$ (say, $\frac{1}{3}(\limsup L_c - \liminf L_c)$) such that for every entourage mapping η, there exists η-fine tagged divisions $\mathcal{T}_1, \mathcal{T}_2$ such that $|\mathsf{S}_{\mathcal{T}_1} f - \mathsf{S}_{\mathcal{T}_2} f| > \epsilon$.

Problem 11.2. Writing $L_c = \int_c^1 |h(x)|\,dx$ where h is defined as in Example 11.11. Prove that L_c is unbounded as $c \to 0^+$. (*Hint: try to find a function k such that $0 \leq k \leq |h|$, such that k is Riemann integrable on each $[c,1]$, and whose integral is reasonably easy to evaluate, at least along a sequence of points c_n that converges to 0.*)

Problem 11.3. Modify the argument of §9.1.3, and the proof of Theorem 9.13, to write out a proof of Theorem 11.13 (that if α is increasing and f is continuous, then f is Riemann-Stieltjes integrable with respect to α).

Problem 11.4. Prove the mean value theorem for Riemann-Stieltjes integrals, namely, that if f is a continuous function on $[a,b]$ and α is an increasing function on $[a,b]$, then there exists $c \in [a,b]$ such that

$$f(c) \cdot (\alpha(b) - \alpha(a)) = \int_a^b f(x)\,d\alpha.$$

(The argument should be extremely similar to what you did for Problem 10.1.)