

Introduction to Real Analysis MATH 5200-5210

Theodore Kilgore

Date of most recent revision is

September 11, 2019

Contents

Preface	ix
1 Some Basic Tools	1-1
1.1 Sets	1-1
1.1.1 Operations upon sets	1-6
1.1.2 More concerning set operations	1-9
1.1.3 The Cartesian Product of two sets	1-13

1.1.4	Relations and Functions	1-13
1.2	Logic	1-17
1.2.1	Propositions, operations, and truth tables	1-17
1.2.2	Quantifiers	1-22
1.2.3	Negations	1-26
2	Integers and Rational Numbers	2-1
2.1	Introductory Remarks	2-2
2.2	Basic properties of the integers	2-4
2.2.1	Algebraic properties of the integers	2-4
2.2.2	Order properties of the integers	2-7
2.2.3	The Well Ordering Principle and Mathematical Induction	2-10
2.2.4	Decimal and other representations of integers	2-16
2.2.5	Other properties of the integers	2-18
2.3	A small excursion	2-19
2.4	Construction of the rational numbers	2-23
2.5	Inadequacy of the rational numbers	2-28

3	Building the real numbers	3-1
3.1	Introduction	3-2
3.2	Sequences	3-12
3.3	Sequences of rational numbers	3-14
3.4	The real numbers	3-22
3.5	Completeness	3-32
4	Series	4-1
4.1	Finite series and sigma notation	4-2
4.2	Tools – the binomial theorem	4-7
4.3	Infinite series	4-9
4.4	Tools – the geometric series	4-13
4.5	A small excursion – the definition of e	4-14
4.6	Some discussion of the exponential function	4-19
4.7	More on convergence of series	4-21
4.7.1	The Root Test and the Ratio Test	4-21
4.7.2	Conditional Convergence and the Alternating Series Test	4-27

5	Topological concepts	5-1
5.1	Basics	5-1
5.2	A brief discussion of topology	5-4
5.2.1	Base for a topology	5-6
5.2.2	Functions and Continuity	5-7
5.2.3	Topological Properties	5-8
5.2.4	Relative Topologies	5-9
5.3	Connectedness	5-11
5.4	Compactness	5-12
5.5	Metric Spaces	5-14
5.6	Some general results	5-17
6	Functions, limits, and continuity	6-1
6.1	Functions	6-2
6.2	Limits of functions	6-6
6.3	Continuity	6-13
6.4	Uniform continuity	6-20
6.5	\limsup and \liminf	6-21

7	Cardinality	7-1
7.1	Finite and countable sets	7-1
7.2	Uncountable sets	7-6
8	Representations of the real numbers	8-1
8.1	Introduction	8-2
8.2	Decimal representation	8-3
8.3	Binary representation	8-5
8.4	Other representations	8-7
8.5	The Cantor Set	8-8
9	Conclusion of MATH 5200, MATH 5210 begins	9-1
10	The Derivative and the Riemann Integral	10-1
10.1	The Derivative	10-2
10.2	Integrals	10-6
10.2.1	The integral of a bounded non-negative function on a bounded closed interval	10-7
10.2.2	The integral of a bounded function on a bounded closed interval	10-13
10.2.3	The effect of unnatural ordering on integration	10-14
10.2.4	The Riemann integral	10-16

10.2.5	The linearity of the integral	10-18
10.2.6	Shortcuts can go wrong, when defining the integral	10-19
10.3	The Fundamental Theorem of Calculus	10-21
10.4	Improper integrals and the integral test	10-25
10.5	A problem with the Riemann integral	10-30
11	Vector and Function Spaces	11-1
11.1	Normed vector spaces	11-2
11.2	Inner Products	11-11
11.3	Norms for continuous functions	11-17
11.4	Banach Spaces	11-21
11.5	Linear Transformations and Continuity	11-24
11.6	Weierstrass Approximation Theorem	11-28
11.7	Linear operators defined by integral kernels	11-34
11.8	Periodic functions and the Fourier series	11-35
11.9	The Féjér operator	11-40
11.10	Weierstrass Theorem for periodic functions	11-43

12 Finite Taylor-Maclaurin expansions	12-1
12.1 Introductory remarks	12-1
12.2 Finite Taylor expansions of a function	12-2
13 Functions given as series	13-1
13.1 Introductory remarks	13-1
13.2 Functions defined as series	13-4
13.3 Convergence of a power series	13-5
13.4 Differentiation and integration of power series	13-10
13.5 Double sums	13-15
13.6 The rest of the story	13-16
14 Integrals on Rectangles	14-1
14.1 Integrals defined on Rectangles	14-2
14.2 Fubini's Theorem	14-5
15 The Stieltjes Integral	15-1
15.1 A broader view of integration	15-1
15.2 The Riemann-Stieltjes Integral	15-8

Preface

This document was begun as a text for MATH 5200, Spring 2012 and for MATH 5210, Fall 2012. Since then, some of the topics in it have been expanded or have been subjected to stylistic revision. Continued attention has been devoted to the seemingly never-ending task of removing typographical errors. The document is presented to the students and to their future instructors so that everyone interested in the matter can know what has been done in the course. The current version is “published” at www.auburn.edu/~kilgota for downloading.

Just in case that future revisions may be desirable, the students are encouraged to keep the “book” in a binder.

That ought to make it easy to add any new pages, and also make it easy to incorporate any portions which are revised or expanded. Since the text is distributed in the form of a PDF file, students can keep an electronic copy and print any part as needed. Or, if desired, a student can merely keep an electronic copy and bring it to class on a laptop, netbook, or other electronic device. In any event, the pagination is done chapter-by-chapter, thereby making it possible to revise or expand any portion of the text without need to revise the pagination of subsequent chapters.

For additional convenience, the text is distributed in two formats which differ only in the dimensions of the pages. One of the two formats uses standard letter-sized pages, suitable for printing or for viewing on a desktop computer with a large monitor. The second format uses short and wide pages, presented in landscape mode. When thus arranged, the pages can be conveniently viewed with no need for scrolling. Thus, this second format is intended exclusively for viewing and not for printing.

Some perspectives for the students

Typically, the students entering an introductory course in real analysis have taken the calculus sequence and one or both of an introductory differential equations course and an introductory linear algebra course. To a too great extent, these courses present a very rushed and too informal introduction to huge areas of mathematics. In these courses, the contents and the emphases are predetermined by external constraints. These are service courses. Students are expected to learn a bewildering multitude of topics in a huge rush for future use in various other subjects, such as engineering or science. Inevitably, the logical and methodical development which is the very heart and soul of mathematics has

been sacrificed to institutional and curricular needs. That is unfortunate, precisely because mathematics cannot be learned well without due attention to its logical and methodical development. Indeed, some including the greatest mathematicians, dating all the way back to Pythagoras, Plato and Euclid, have maintained that without rigorous and methodical logical development the subject matter of mathematics cannot be learned and assimilated beyond a very basic level.

In fact, the logical and methodical development of mathematics has evolved and is valued by mathematicians precisely because it is a valuable tool for learning and comprehending the subject. Not only this, but also logical and methodical development is the best antidote that exists for the varieties of naive and intuitive assumptions which masquerade as common sense but which on closer examination turn out to be false. Many have fallen victim to such fuzzy thinking. Even mathematicians, indeed even some famous mathematicians, have indulged in some of the prevailing prejudices of their times and are on record as having believed mathematical “facts” which on closer examination turned out not to be facts at all. As a prime example, Euclid is justly celebrated for the systematization of geometry. His famous postulate on parallel lines showed great insight, in that he explicitly listed it as an assumption. Nevertheless, efforts to prove it continued for centuries. Based upon an abiding belief that it just had to be true and that the proof had somehow escaped Euclid. As a result, a lot of effort was spent in a vain effort to figure out a proof based on his other assumptions. But Euclid also made a few mistakes of his own, in the direction of treating certain things as “obvious” which in fact are not and failing to list them as axioms or postulates.

Other examples of fuzzy and imprecise thinking relate to a naive and fuzzy understanding of what we mean by a limit. Although I am not able to chase down the precise citation, I distinctly remember having read in one of the dialogues

of Socrates a scene in which Socrates starts with a naive description of a limit process, propounded by someone who is debating with him, and uses the “definition” to “prove” that 2 and $\sqrt{2}$ must be exactly equal to each other. A more modern variation of a similar argument can be used to “demonstrate” that π is equal to 2.

Also, there are not a few mathematicians who without close examination rejected the idea that there could be such a thing as a function which is defined and continuous but is nowhere differentiable. Nowadays, we know better. But such naive and unexamined assumptions were far too prevalent for far too long and have only been dispelled by serious attention to logical development of mathematics.

The point behind the above reflections is that there is no substitute for clear, logical, and methodical thinking. It is the only tool available to us mere mortals which we can use to avoid logical traps and pitfalls.

Students who – in the face of the institutional constraints described several paragraphs above – have tried to understand logical development and unifying principles have almost certainly been rewarded immediately, even when sitting in service courses. For, in general such students get better grades even in the service courses and tend better to retain what they have learned. These students are very different from those who merely try to cram for the next test by memorizing a seemingly never-ending list of seemingly isolated facts and formulas. Those who adopt the second approach, thinking it is somehow easier, too often forget what they learned even before the final exam has rolled around and rarely seem to remember much of the material when they need it later on.

Here, in the introductory course sequence in real analysis, logical development and unifying principles take center stage. We will thoroughly revisit some of the basics of mathematics which have been so cavalierly glossed over in previous courses. The emphasis will be upon logical and sequential development of a fundamental part of mathematics. Much

more of depth, systematically presented and learned. Much less upon breadth.

A great proportion of the material in this course will be presented in problem form for the students to work out. Most of the problems will actually consist of proofs. Many of them will be major results which will be used later on. And when working the problems, the students should only use material which has preceded the problem in this text. The intent behind this approach is to present new material systematically and in sequence, building up one's knowledge in logical steps. One needs to be clear at each stage about what is known and what is not yet known. Such an approach may seem more difficult at the beginning, but it almost immediately brings clarity and organization to learning. The result is greater comprehension, confidence, and better retention of what has been learned.

The organization of the material

Chapter 1 is envisioned as a review of relevant portions of set theory and logic. If the students have previously completed MATH 3100, then it is hoped that they should have seen all of this material previously. Students coming into MATH 5210 ought to know everything in Chapter 1, too, of course. But they should also make sure that they are aware of what is in the text. For, among other things Chapter 1 contains definitions, concepts, and terminology which are used later in the course.

After Chapter 1, which is intended as a review, the first topic in MATH 5200 is a step-by-step logical construction of the real number system. We start out with what we know of the integers, constructing then the rational numbers, and finally we build the set of real numbers. Most of the students have probably not seen a detailed, step-by-step logically

rigorous construction of the real number system in their previous course work. Moreover, such a construction is usually not done in more advanced courses, either. Typically, those more advanced courses only present a quickly-studied list of descriptive axioms and seem to presume that the students have learned all about what the real numbers are through some process of osmosis. It is therefore appropriate for the students to see an actual construction, and this course is the appropriate place to deal with the topic.

In Chapter 2 we will begin the project of constructing the real numbers by dealing first with relevant basic properties of the integers. Then the rational numbers will be developed as a quotient field over the integers. In the course of this development, the important concept of an equivalence relation is introduced, as it is well illustrated by what we mean when we say that two rational numbers are equal. After this, some care is given to describing the logical and practical need to expand the system of rational numbers to a larger system. A crucial property which the set of rational numbers does not have is called “completeness.” As a particular example of the lack of completeness, some discussion is given to the fact that there is no rational number whose square is exactly equal to 2 and to the problems and difficulties which we consequently face.

Then in Chapter 3 the real numbers are constructed. Constructed, not merely described, which is not the same thing. We start with what we know of the rational numbers. Without going outside of the set of rational numbers, we carefully lay groundwork for a systematic expansion of them to form a new, larger set which will be called the set of real numbers when we get finished. The method we use is to describe what is meant by the limit of a sequence, and it is noted immediately that sequences of rational numbers exist which one might think “ought to have limits” but which in fact fail to have limits within the set of rational numbers. Then, a systematic method for dealing with this problem is introduced,

based upon the concept of Cauchy sequences. After certain basic properties of Cauchy sequences of rational numbers have been proven, and only then, we can proceed. Care must be taken during this process. It is not logically permissible to assume, for example, that desired properties of the real numbers hold true within the set of rational numbers. Such assumptions are simply false and are moreover prominent among the reason why we are doing the construction. The method of construction which is used here is to define the set of real numbers as a set of equivalence classes of Cauchy sequences, defining two such sequences to be equivalent if their term-by-term difference forms a sequence with limit zero. This method emphasizes an obvious analogue to the construction of the rational numbers as equivalence classes, as has already been done. Then it is shown that any attempt to “complete the completion” of the rational numbers does not take us outside of what we have already constructed. Thus, we have finished the construction of a new set, which we call the set of real numbers. This new set of real numbers satisfies the Statement of Completeness, which says that every Cauchy sequence of real numbers must converge, and converge to a real number. It should be emphasized that the Statement of Completeness is a result and an outcome of the construction, not in any way an “axiom.” After the Statement of Completeness has been established, several exercises deal with various logically equivalent statements. In particular, there is an exercise in which one is supposed to prove a statement which is often featured in other texts as “The Completeness Axiom” and is therefore familiar to many of us.

Throughout this text and course, and in particular in Chapter 3, emphasis and stress will be laid upon a logical and sequential development. In this chapter, then, the construction of the real numbers is completed.

The intent of Chapter 4 is to cover rigorously and well some basic topics in computational mathematics. In particular, we deal with series and their convergence or divergence. In the previous chapter, we have developed and constructed

the set of real numbers. Now we need to learn how to use it. Thus, in Chapter 4 we will cover the sigma notation for a sum. We will actually derive some of the summation formulas which appeared so mysteriously in the calculus text in the chapters on integration. We will also learn the Binomial Theorem and how to compute the sum of a geometric series. We will have a basic definition of the number e and, using that definition, show how its definition can lead to its expression as an infinite series. Chapter 4 is very important for what comes after it. Again, though, we strictly follow a sequential and step-by-step development of its topics. All that is done in Chapter 4 – and much is done there – is done only using the definition of the limit of a sequence introduced in Chapter 3. Most particularly we have not done any systematic treatment of such things as differentiation and its application to such things as L’*h*ôpital’s Rule, nor integration. These topics will be introduced in a later chapter. Thus, it is not permissible to use any of these in Chapter 4. The only tools which it is permissible to use are developed either in previous sections of Chapter 4, or in previous chapters. Those tools which we can use will be seen to be simple and basic in their nature, but unexpectedly powerful and far-reaching in their application.

Chapter 5 covers things which are often called topological properties of the real number system, relating to such things as open and closed sets, accumulation points, interior points, and the consequences and ramifications associated therewith. This chapter also introduces the general concept of a topology upon a given set and the definition of a metric space. Some basic properties of topological spaces and metric spaces are introduced, too. Everything in this chapter will find application later in the course or in its successor course, MATH 5210.

Building on Chapter 5, Chapter 6 discusses the concepts of continuity and limits for real-valued functions of a real variable.

Chapter 7 deals with countability and uncountability, in particular showing that the set of real numbers is uncountable, whereas the set of rational numbers is countable.

Chapter 8 deals with the representation of real numbers as decimal expansions, or expansions to other bases than ten. One purpose of this chapter is to put upon a firm, logical foundation the description, often made informally in lower-level mathematics courses, that the set of real numbers consists of the set of all decimal expansions, whether repeating or non-repeating. In Chapter 8, also, an interesting set of real numbers known as the Cantor set is briefly introduced and described. The immediate reason which lies behind the introduction of the Cantor set is to dispel an often-encountered naive assumption, that an uncountable set must contain an interval at least somewhere inside of it. The Cantor set is an uncountable set, but it contains no interval. Of course, there are many other interesting properties of the Cantor set, too. Further study of those properties is left to do in other courses.

Chapter 9 is a “placeholder” to remind everyone that MATH 5200 has ended, and MATH 5210 begins.

At this point, the following should be said, with emphasis:

The material of MATH 5200 is important as background for students entering a course such as MATH 5210. In particular, new results depend upon previous results in any sequential development of a subject. Therefore, the student should be familiar with what was done in the first nine chapters of this text while going forward with the material designated here for MATH 5210.

Math 5210 begins with Chapter 10, on differentiation and on the Riemann and Riemann-Darboux definitions of the integral. At the end of that chapter, it is pointed out by way of an example that there is a clear deficiency of the

Riemann integral. Namely, we cannot, except under very restricted circumstances, assert that the limit of the successive Riemann integrals of a sequence of functions will always be equal to the integral of the limit of the sequence of functions. It is pointed out that this is a deficiency, in that many situations even in applied mathematics require us to be able to do that with impunity. The resolution of this problem is left as unfinished business, to be addressed in further study, in more advanced courses in analysis.

Chapter 11 deals with normed vector spaces and questions of compactness and completeness. It deals with the equivalence of norms on a given finite-dimensional space and the non-equivalence of various norms which can be put on infinite-dimensional spaces such as $C[a, b]$, the set of all continuous functions defined upon the closed interval $[a, b]$. Several norms are introduced. The concept of an inner product space is introduced. The Weierstrass Approximation Theorem is proven. Fourier expansions of continuous periodic functions are discussed.

Chapter 12 introduces finite Taylor or Maclaurin approximations.

Chapter 13 discusses power series, the interval of convergence, and discusses functions which are defined by their series expansions within their radius of convergence. As all of the tools are now in place in previous chapters and sections of the text, the exponential function is defined in terms of its standard series expansion. This series expansion was already mentioned in Chapter 4, where little could be done with it except to note that the series defines a function wherever it converges, and to prove that in fact the series can be seen to converge for every real input x . Further we could not go in Chapter 4 because we had not yet defined the derivative and the integral. The (natural) logarithm is defined, too, in Chapter 14 as the inverse function of the exponential function, and the basic properties of the logarithm function are addressed.

Also in Chapter 13, a discrepancy is pointed out in the apparent lack of connection which can exist in that a function can often be defined upon a much larger domain than the interval upon which its series expansion can be seen to converge. It is further pointed out that the resolution of this mysterious behavior lies in the behavior of the function in the complex plane, not in the behavior of the function on the real line. The students are encouraged to take a course in complex analysis in order to comprehend “the rest of the story.”

Chapter 14 defines the integral of a function defined upon a closed and bounded rectangle in \mathbf{R}^2 and, in a sequence of problems, develops a proof for Fubini’s Theorem.

Chapter 15 extends the Riemann integral by defining the Riemann-Stieltjes integral and presents some of the basic properties of this new integral.

Expectations

As to the internal organization of the course, a heavy emphasis will be placed upon classroom participation, problem solving, and classroom presentations by the students of the results of their efforts. Classroom discussions and classroom presentations by students will comprise a large portion of the grade, too.

The problems which are presented here are intended to relate directly to the fundamental theme of the course. Indeed, many of these problems will be used as essential building blocks. By the time that the course is completed, most of the problems presented will have actually been presented by one or another student during class time. Some of the exercises, in the interests of democracy and fairness, will have been presented by the instructor. Some of the problems

may be assigned as homework to be written out, or may be used as test problems. But, it is intended that all students who take the course should learn how to handle essentially every problem which is in the book.

An obvious corollary of the previous two paragraphs is that regular class attendance and active participation are very important, not only for learning the material well but also for the respect due to others.

It should also be clear that there are some partially unspoken rules about how homework problems and exercises should be approached. The proper approach is, clearly, that when doing a given exercise or showing that a given result is true, one really does need to stick to using only results which have been previously stated or proved either in the text or in previous exercises. Appeal to external authority in the form of “This result is true because the book of so-and-so, which I found in the library, says so.” should be avoided. For, the book of so-and-so has used material which is similar to what we are doing in order to come to that result. Appeals to other areas of mathematics are similarly to be avoided. As a prime example of what this means, one of the purposes of the logical development contained here is to provide a logically rigorous foundation of the area called Calculus. In such a situation, it is hardly appropriate to call upon results taught in Calculus in order to “prove” things upon which Calculus is based.

Another principle which is important is the proper application of credit. Mathematicians often collaborate in order to do research, and moreover there is nothing wrong at all if two or more students work together to solve a problem. But ethics requires one to be open and public about the matter if collaboration has taken place, if help has come from some other source, or if materials outside the course, such as books in the library or online resources, have been used while solving a problem. It is not forbidden to consult others to read books, or to look things up online. Nevertheless, it ought to be obvious that the student benefits the most who can come up with his or her own solutions to the problems,

relying only upon inner resources.

Finally, it is the hope of the instructor that the students, on completing this course, will have had an experience which is positive, if at times challenging, and that the students will have assimilated and mastered enough of the subject for future success.

Chapter 1

Sets and logic

1.1 Sets

Mathematics is built up as an axiomatic structure. This means, that we start out with some basic assumptions which either look reasonable to us, or they are assumptions which we would like to take on in order to explore what kind of

system comes out as a consequence. One of the basic tools for doing our work is the concept of a set. But nobody knows what a set is. Funny, that. Well, actually, what this means is, a definition of what a set is has no place *inside* our formal system, precisely because it is something which we must *presuppose* and therefore *logically precedes* the formal system. If one perceives that we have already descended into thinking which is not serious, then one might try to play the “dictionary game” in order to see what the problem is. The dictionary game is played by looking up a word in the dictionary. Then one looks up all of the words used in the definition of the word which was looked up. This procedure is repeated until one comes back to one of the words previously defined. Usually, it does not take very many steps before this happens.

Well, then, what is a set? We can make an *informal* definition easily enough. A set is a collection of objects, things, concepts, or whatever. Why can we not make this formal? Well, perhaps we could. But then we would require a rigorous definition of “collection,” “objects,” “things,” “concepts,” and so on. As already stated, we have to start somewhere, and we might as well start here.

Things are either in a given set, or they are not. The things which are in the set are called *elements* of the set. If we have decided that the name of the set is S , and x is some object, then we write $x \in S$ to mean that x is an element of S . If it is not true that $x \in S$, then we write $x \notin S$. Now that we are getting all formalistic, we can restate the first sentence of this paragraph as saying that an element is either in a given set, or it is not. A set can be defined either by explicitly listing the elements in it, or it can be defined by giving some kind of characterization which completely describes the elements in it, in other words defining the set by some defining property of the elements. Here are some rather stock examples:

1. $\{1, 2, 3\}$ describes the set which contains the three listed integers (it is customary to use the curly brackets when giving a definition of a set). Also notice that the order in which the elements are listed does not matter. For example, the set $\{3, 1, 2\}$ is the same set.
2. $\{x|x \text{ is an even integer}\}$
3. $\{x : x \text{ is an even integer}\}$ (same as previous)
4. $\{x|x = 2k, \text{ for some integer } k\}$ (equivalent to previous two – how do we define an even integer, anyway?)

The astute, or those who have already seen things like this before, may be aware that there can already be problems at this basic level. We have to be careful, else we cannot even pretend to build a logical structure on what we have already done. Clearly, it is possible to put words together to come up with a description which stumbles over something basic. The problem which arises from that is, how do we know that we are doing such a thing on a given occasion, or not? Ideally, we would like to be able to put some kind of an airtight system in place which automatically disallows such an act, even if it is committed by the well-intentioned who would prefer always to stick to common sense. A rather obvious problem arises, now that we have agreed we know all about sets, is the following:

$$S = \{x|x \notin S\}$$

For, if an x is given to us to test and decide whether it is in S or not, then, if it is, then it is not, and if it is not, then it is.

Exercises

1.1.1 This exercise relates to the example just above. One might suppose that the difficulties which it presents need somehow to be avoided. Perhaps the way out is to agree that the example somehow does not give a proper definition of a set? Has anything already been stated as a property of sets, which this “definition” violates? If so, find it in the previous text.

In the alternative, it may be that in fact nothing at all has been violated, but the situation is nonetheless strange. Is there something which can be done about the matter? Or, have we successfully defined a set after all, but just a very strange one?

Another, more serious matter is implicit here. If you are looking for a hint about how to solve the previous problem, then do not look further. No hint will be given in what follows. But we do have yet another problem, independent of the previously posed one. Namely, we have to be careful about where our elements come from as well as where our sets come from. Consider the following definition, that

$$S = \{Y | Y \notin Y\},$$

in which the Y in the definition is, by its usage, a *set*. This would seem to make sense, until one looks more closely and asks that just how the definition can be applied to S itself. For, if S is not an element of itself, then $S \in S$. But, alternatively, if $S \in S$, then S must satisfy the definition of S which says then that $S \notin S$. This is the famous Russell’s paradox, named after Bertrand Russell, who stumbled upon it approximately 100 years ago.

There are several solutions to Russell's paradox and to similar logical conundrums. One of the most practical, which we will adopt here, is to realize something which goes all the way back to Aristotle's logic. Aristotle recognized the relevance of a "universe of discourse" which, in terms of set theory is a "universal set" which serves as the context for the discussion of everything inside of it. That is, the universal set is not the set which includes everything. Rather, the universal set represents the result of prior agreement on what is relevant to the discussion. We are on quite safe ground if we declare (for example) that on a given occasion we are speaking of integers, cards in a standard deck, or whatever. Then, if we want to speak of sets of integers, those sets are not integers. If we wish to speak of poker or bridge hands, then those hands are not individual cards and cannot be confused with such. There are other, more sophisticated approaches, but this is the one that we will adopt.

Another reason that we use a universal set is that it quite often makes a difference what the universal set is, what is the correct answer to a given question. For example, is 11 divisible by 4? Yes? Or, no? The answer to that question depends very much upon context, in other words, upon the universal set which is presumed to underlie the question.

Homework or (not unduly long because this is not a class in metaphysics or epistemology, but in mathematics) class discussion:

Exercises

1.1.2 Other logical paradoxes? Think of some examples. Discuss.

1.1.3 Does the employment of a universal set really solve all our problems? Or are we still glossing something over?

Is it possible to have a poorly defined universal set and to land right back in the soup again?

1.1.4 There once was a town famous for its barber. The only men in the town whom the barber did not shave were those who shaved themselves. Who shaved the barber? Is this another example of a paradox? Or is some other kind of problem hidden behind the question?

1.1.1 Operations upon sets

Sets, of course, are used as a tool in mathematics. Therefore, we need to pay attention to some of the basic operations used upon them and to familiarize ourselves with some of the notation. Let us first agree that we have chosen a universal set, U , for the purposes of the ensuing discussion.

If A and B are any two subsets of U , then we can define three operations on each of them. First, it is possible to consider the **complement** of each of these sets:

We agree that, given the set A , then concerning each element $x \in U$ it is the case that either $x \in A$, or $x \notin A$. We have already agreed that

$$\{x|x \in A\}$$

is merely an alternative and equivalent definition of A . But, then, what about the set

$$\{x|x \notin A\}?$$

This set is the complement of A . Several notations are in use for this set. It can be denoted with

$$U \setminus A$$

or sometimes as

$$U - A$$

$$U \sim A$$

or, sometimes, as just $\sim A$, or \tilde{A} or A^c . Similarly we can construct the complement of B or any other subset of U .

Now, the above discussion raises an obvious question, that what is $U \setminus U$? Does this make sense? Well, certainly it does if we want to agree that every subset of U has a complement. Of course, one could say that every subset except U has a complement. However, we shall see that it becomes very inconvenient if one tries to formulate general statements and rules and then to start listing exceptions. Thus, perhaps we ought to say that we have just defined the **empty set** which we will denote by Φ . There is the obvious problem that there might be more than one empty set, and that problem needs to be resolved. We will defer the discussion of that until we have presented the other set operations.

Given our two sets A and B it is clear that we can consider the set

$$\{x | x \in A \text{ and } x \in B\}$$

which is called the **intersection** of A and B . Notice that the word “and” here means that the statements to the left and right of it are both simultaneously true. Notice that the intersection of the two given subsets A and B may not contain any elements at all. There is nothing, after all, which says that it has to. In such an eventuality, we could say that the intersection is empty, and we do. We would be tempted to define that empty intersection as the set Φ which has already been discussed. But there is that little problem which needs to be cleared up. Observe that if we cannot do so, then our set theory is already a kind of a mess because we can’t really say that any two sets *have* an intersection, even if there is a universal set lurking in the background.

Also there is the set

$$\{x|x \in A \text{ or } x \in B\}$$

which is called the **union** of A and B . Notice that here the word “or” signifies that at least one of the statements to its left and right is true. In particular, we do not intend to exclude the possibility that both are true. Naturally, a situation could arise in which we intend to say “one or the other but not both” but then we would need to write that down with some other words, which exactly specify what we mean to say.

We now need to consider possible relationships between our two sets A and B . It might be, for example, that every element of A is also an element of B . In that case, we say that A is a **subset** of B . The notation for this is

$$A \subseteq B$$

Of course, we could also write

$$B \supseteq A$$

which we might express as “ B contains A ” which obviously means the same thing.

There is also the possibility, of course, that the two sets A and B are in fact identical, and then we write

$$A = B.$$

Exercises

1.1.5 Given two subsets A and B of some universal set U , show that

- i. If $A = B$, then $A \subseteq B$ and $B \subseteq A$
- ii. The converse of the previous statement is true, too.

1.1.6 If Φ_1 and Φ_2 are two subsets of a universal set U and neither of them has any elements, then in fact $\Phi_1 = \Phi_2$. (Hint: Use the previous problem).

1.1.2 More concerning set operations

The set operations obey certain rules. The following are left as exercises for the students.

Exercises

1.1.7 Associative Laws. Given any three subsets A , B , and C of a universal set, U , we have

$$A \cap (B \cap C) = (A \cap B) \cap C \quad \text{and} \quad A \cup (B \cup C) = (A \cup B) \cup C$$

1.1.8 Distributive Laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

1.1.9 Commutative Laws

$$A \cap B = B \cap A \quad \text{and} \quad A \cup B = B \cup A.$$

There are also laws which relate specifically to U and Φ . Continuing our convention that A , B , C are arbitrarily chosen subsets of U , we have

$$A \cup \Phi = A \quad \text{and} \quad A \cap \Phi = \Phi$$

and

$$A \cup U = U \quad \text{and} \quad A \cap U = A.$$

There are also some rules relating to complements. First of all, we extend the notion of the complement to include the **relative complement**. That is, we define the notation $B \setminus A$ to mean the set

$$B \cap (U \setminus A).$$

De Morgan's Laws state that

$$\sim (A \cap B) = (\sim A) \cup (\sim B) \quad \text{and} \quad \sim (A \cup B) = (\sim A) \cap (\sim B).$$

Now, in view of the fact that union and intersection operations each obey the associative law, we can in fact define the union and intersection of more than two sets. First, suppose that we have some sets A_1, \dots, A_n . Then we can define

$$\bigcup_{i=1}^n A_i = \left(\bigcup_{i=1}^{n-1} A_i \right) \cup A_n$$

and similarly define

$$\bigcap_{i=1}^n A_i = \left(\bigcap_{i=1}^{n-1} A_i \right) \cap A_n.$$

These are examples of **inductive** definitions. The way a definition like this works is, we know what $A_1 \cup A_2$ is, and we can work up from there to any value of n (similar remarks if the operation was “ \cap ” instead).

We can also replace the “ n ” by “ ∞ ” if we want to. How? Well, we cannot use an inductive definition there, so we had better introduce a concept which gets us away from counting entirely, the **index set**.

Let us consider that we have a collection of subsets of U called \mathcal{C} . Then, \mathcal{C} itself is a set. It is not a subset of U , obviously, but we can say that it lives in some other universal set, for example the set of all subsets of U , which is called the power set of U . Well, then, the elements of \mathcal{C} are subsets of U . We can define

$$\bigcup_{C \in \mathcal{C}} C = \{x | x \in C \text{ for at least one } C \in \mathcal{C}\}$$

and similarly define

$$\bigcap_{C \in \mathcal{C}} C = \{x | x \in C \text{ for all } C \in \mathcal{C}\}.$$

Observe that we have gotten away completely from the need to *count* the sets when thus defining the union and the intersection of the sets in the collection \mathcal{C} .

Exercises

1.1.10 Show that these general definitions of the union and the intersection are compatible with the definitions previously given, for two sets.

1.1.11 What form is taken by De Morgan's laws, as applied to arbitrary collections of sets? Prove that these generalized De Morgan's laws are valid.

Now, if we need more of general set theory, we will try to introduce it in the course of doing the rest of our business. We will move soon to a brief discussion of some of the fundamentals of logic, but before doing so we really should define Cartesian product of two sets and formally introduce the concepts of "relation" and "function." All three of these definitions, technically speaking, are concepts of set theory. But they are of central importance in mathematics.

1.1.3 The Cartesian Product of two sets

Let two sets A and B be given. The **Cartesian product** of these two sets is then denoted by $A \times B$ and consists of all ordered pairs (a, b) in which $a \in A$ and $b \in B$. The students are presumably quite familiar with the original Cartesian product which is represented by two crossed axes and the thus generated plane used for graphing.

1.1.4 Relations and Functions

Any subset of a Cartesian product $A \times B$ is considered to be a **relation**, since it describes a relationship or a linkage between certain elements of either of the sets A or B to elements in the other set.

A **function from A to B** may be defined as a relation on the two sets which has the two additional properties:

- i. Every element of A may be found in some element of the relation.
- ii. If $a \in A$ and $b_1, b_2 \in B$, then (a, b_1) and (a, b_2) are not both present in the function unless $b_1 = b_2$.

Some would say that the above represents the “graph” of a function, however, and would say instead that the function consists of a **rule which associates to each element of A one and only one element of B** . In view of the fact that a function is often given by such a rule, procedure, or computation, this definition of function is perhaps more intuitive. One often denotes the rule by a symbol such as f and writes $f : A \rightarrow B$.

In certain respects, more precise terminology is needed here. Perhaps unfortunately, this more precise terminology is not universally agreed upon. If we have two sets A and B and a function $f : A \rightarrow B$, it is generally agreed that the function f is defined upon all of A and that A is called the **domain** of f . That is, the function f “uses all of A .” The set B is usually called the **range** of f . Careful examination of the definition above, however, will show that nothing in it requires all of B to be used. Some, such as many algebra and calculus books, may add the additional restriction that all of the set B needs to be in use, or, worse, tend to require that in practice without ever actually saying so. However, others would say that there is nothing wrong at all. The definition of a function is perfectly good as stated above. But, if one wishes to refer by name to the set of those elements in B which are actually of importance here, then one needs to speak of the **image of f** , not the range of f . The image is then the set

$$\text{Im}(f) = \{b \in B \mid b = f(a) \text{ for some } a \in A\}.$$

We will follow this usage here. The consequence is that, in our usage, the range (the set B) can be any set containing

the image. If the image of f and range of f happen to be the same set, then the function f is said to be **onto** by some mathematicians. Others, more concerned about grammatical correctness than linguistic purity, say that f is **surjective**, or is a **surjection**.

A further source of possible confusion is the frequently used expression “a real-valued function of a real variable” to denote a function whose domain is *any* subset of the real numbers (side comment: we officially do not know what those are, as yet!) and whose image is any subset of the real numbers. For, often this fact is important all by itself. And then it is possible to give all those exercises, just like in the precalculus book or the calculus book, where the students are supposed to find the domain of the function and the “range” of the function (that is, what we have just defined as the image of the function).

Now, let us notice that indeed the prescription of the domain is part of the definition of a function. Consider the function given by $f(x) = x^2$. If we take the domain to be the largest set possible, then the domain consists of all real numbers. The range can be any set containing the image, which comprises all non-negative real numbers. This function is not invertible, which means that there is no function going back from the image to the domain which reverses the action of f . However, if we took the domain to be the set of all non-negative numbers, instead, then the resulting function *is* invertible. Thus, these two functions are not identical even though given by the same computational procedure. It can, however, be said that the second function can be obtained from the first by **restriction of the domain**. This is a term of the art, which is why it has been put in boldface.

A function $f : A \rightarrow B$ for which $f(a_1) = f(a_2)$ cannot happen if $a_1 \neq a_2$ is called **one to one**, or, by those who like fancy words, **injective**, or to be an **injection**. Note that in the example mentioned in the previous paragraph are seen

two functions. One of them is one to one, and the other is not, even though the same computational formula is used for both. To complete the introduction of necessary terminology, a function which is both one to one and onto is called a **one to one correspondence** or is said to be **bijective**. Finally, the word **mapping** is often used as a synonym for the word “function” on occasions when it seems more descriptive of what is happening, but the formal definition is identical.

One might ask why, exactly, have we not agreed that a function could have several output values instead of only one? The answer to this question is that the cumulative experience of mathematicians leads to the conclusion that the definition is most useful the way it is, and not some other way. The definition of function is, after all, a tool for making mathematics easier to do, not an end in itself. An example might serve to illustrate what can happen if we allow more than one value to “come out” when just one value is “put in”:

Let the function f be defined by $f(x) = \sqrt{x}$ and g defined by $g(x) = \sqrt{x+1}$. It is advantageous to be able to say that $f(x) + g(x)$ will define for each individual x which is in the domain of both of f and g a “new function” which we could (and do) call $f + g$. And, indeed, there is no problem in computing $\sqrt{x} + \sqrt{x+1}$ if both square roots exist.

But if (contrary to standard usage!!!) someone wanted to allow “plus or minus the square root of a nonnegative number” to be a function, then how many possible results are there to the computation $\pm\sqrt{x} + (\pm\sqrt{x+1})$? Clearly, there are now four possible outcomes, not two. The obvious consequence is that a door has been opened which leads to a rapid and uncontrollable descent into nonsense and chaos. One of the important reasons why the concept of function was developed in the first place was to make our lives simpler. But this kind of thing would violate the clarity which we are seeking. Such problems are easily avoided by insisting that a function has to have unambiguous outcomes.

1.2 Logic

1.2.1 Propositions, operations, and truth tables

A proposition is a statement which is either true or false. This means, among other things, that whatever statements can not be classified as either true or false might not be fitted very easily into the scheme and are thus excluded. We can denote a proposition either by issuing the related statement, or we can also label propositions with letters, such as p, q, r .

Now, let p, q, r stand for propositions, as described above. We can connect propositions by symbols representing the connectors “and” and “or” and “not.” We will use the following notation, which is fairly standard:

$p \vee q$	means	p or q
$p \wedge q$	means	p and q
$\sim p$	means	not p

A compound statement or proposition is a statement in which several substatements are given, linked by such connectors as are described here, or one more which was not listed in the previous table of connectors, namely

$p \Rightarrow q$	which means	p implies q
-------------------	-------------	-----------------

We also use parentheses in order to group things together, in a manner similar to what is done with algebraic expressions. For example, the compound propositions $\sim p \vee q$ does not mean the same thing at all as $\sim (p \vee q)$. In the first instance, only p is negated, but in the second instance the entire compound statement $p \vee q$ is negated. However,

parentheses can also be used to improve legibility without changing the meaning. As an example of this, some would prefer to write $(\sim p) \vee q$ rather than $\sim p \vee q$. The meaning is the same with or without the parentheses, but sometimes parentheses can add clarity even if the meaning is the same without them.

Now that we have our notation in place, we need to look for ways to analyse compound statements to see whether they have an intended logical meaning, or not. One of the most basic ways to do this is with a truth table, which schematically lists all possible combinations of true and false for the component propositions, and for the compound statement, too. Here are some examples:

- The statement $\sim p$ is of course easiest. The table for it is

p	$\sim p$
T	F
F	T

- The table for $p \vee q$ is

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

- The table for $p \wedge q$ is

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

- The table for $p \Rightarrow q$ is (discuss)

p	q	$p \Rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

We say that two statements are **equivalent** if they have the same truth table.

Exercises

1.2.1 Construct the truth tables for $q \vee p$ and for $q \wedge p$. The idea is, of course, to see that these come out equivalent to two of those given above, respectively.

1.2.2 Set up a truth table for three statements, p, q, r and then construct the entries for $(p \vee q) \vee r$ and for $p \vee (q \vee r)$. The two should look identical, of course.

1.2.3 Is $p \vee (q \wedge r)$ equivalent to $(p \vee q) \wedge (p \vee r)$?

1.2.4 Construct the truth table for $(\sim p) \vee q$. Is it identical to any of those above?

1.2.5 De Morgan's laws apply in the context of propositions as well as in the context of sets. Here, they say that $\sim (p \vee q)$ is equivalent to $(\sim p) \wedge (\sim q)$, and that $\sim (p \wedge q)$ is equivalent to $\sim p \vee \sim q$. Construct a truth table for each of the two De Morgan's laws, which demonstrates the equivalence.

1.2.6 How might one go about writing $\sim (p \Rightarrow q)$ in terms of the other listed operations? Using only the operations \sim and \vee and \wedge , construct something involving p and q which has a truth table identical to the table for $\sim (p \Rightarrow q)$.

1.2.7 We have two connectors, namely \vee and \wedge , as well as the “not” operation, \sim . Is this more than is actually needed? Could we get away with just two of them? For example, is it possible using some cunning sequence of \wedge and \sim operations to write something equivalent to $p \vee q$ by using only those other two operations?

1.2.8 Show using a truth table that $p \vee \sim p$ always has to come out as true, and that $p \wedge \sim p$ always comes out as false. A statement which is always true, no matter what truth values of its respective inputs are, is called a **tautology**. A statement which is always false, no matter what the truth values of its respective inputs are, is called a **contradiction** or a **self-contradiction**.

The main purpose in the following exercises is to introduce certain standard nomenclature regarding implications and to describe the related equivalences and non-equivalences. When one is trying to do a proof, it is often helpful to remember which of these reformulations are useful and which are not.

Exercises

1.2.9 The statement $\sim q \Rightarrow \sim p$ is called the **contrapositive** of $p \Rightarrow q$. Show by appeal to a truth table that they are equivalent.

1.2.10 The statement $q \Rightarrow p$ is called the **converse** of $p \Rightarrow q$. Show by appeal to a truth table that they are **not** equivalent.

1.2.11 The statement $\sim p \Rightarrow \sim q$ is called the **inverse** of $p \Rightarrow q$. Show by appeal to a truth table that they are **not** equivalent.

1.2.12 Each of the inverse and the converse of $p \Rightarrow q$ is the contrapositive of the other.

Now, in principle we could do all logic relating to propositions or statements by appeal to truth tables. However, notice that when we put three statements into a truth table, things already got a bit tedious. We had to have eight rows in order to take all possibilities into account. If we have four statements to glue together, then sixteen rows are required, and if we had five then thirty-two rows are required. This kind of thing gets old really fast, which is one very

good reason why we don't just quit using logical arguments and bring in truth tables instead. Nevertheless, truth tables certainly have their uses for showing some basic things.

1.2.2 Quantifiers

There is an important class of statement-like things which we use all the time, but which do not exactly fit into the scheme above, of propositions or statements which are either true or false. In dealing with sets, we have in fact already touched upon such things. A construction such as $x \in A$ may not be such a proposition or sentence as we have considered, at all. If it is asserted about some particular, specific x , then in principle it can be decidable as true or false. But if x is an element of the universal set which is chosen quite at random, or, better, which will be chosen at random after we make the statement, then we do not know and cannot decide whether $x \in A$ is true, or not. An example of this might be that we pick an integer at random, and we want to investigate whether it is even, or not. We can inspect the integer to see if it is even or odd, but unless and until we do inspect the number, the statement that it is even is undecidable. For, the truth or falsity of the statement depends upon the number. A statement of the sort described, the truth or falsity of which depends upon some other input or inputs, is an **open** statement.

In connection with open statements, it is usually possible to save the situation by asserting the statement to be true for all x , some x (taken to be equivalent to at least one x), or for no x at all. In symbolic language, we have the quantifiers

$\forall x$ For all x

$\exists x$ There exists x

Now, let $p(x)$ be an open statement. We can say things such as $\forall x p(x)$ to mean that $p(x)$ is true for all x . We can say that $\exists x p(x)$ if we mean that there is at least one x for which $p(x)$ is true.

A reasonable question is, how do we formulate the negations of statements when quantifiers are present? We should notice the following are true (whether read from left to right or from right to left)

$\sim (\forall x p(x))$ means the same as $\exists x (\sim p(x))$

$\sim (\exists x p(x))$ means the same as $\forall x (\sim p(x))$

and that, in fact, nothing else would work. Also, once understood, the procedure for constructing the negation of a statement with quantifiers in it can be made completely mechanical, which is nice when things get complicated.

The principles just outlined work also if there is an open statement which requires more than one input variable, and thus more than one quantifier. One can work through one quantifier at a time. To negate such a statement, first change the outermost quantifier and negate everything which comes after. Then go to the next quantifier, and continue in like fashion.

Also, in constructing statements which involve more than one input and more than one quantifier, the order in which the quantifiers are stated can very much affect the meaning. For an example that every reader probably knows, consider the difference between “Everybody gets a Ring of Power” (\forall precedes \exists) and “One Ring to rule them all” (\exists precedes \forall).

More relevant to our mathematical interests is the meaning of statements involving quantifiers in mathematics. A

very important example of a statement with several quantifiers is the definition of continuity of a real-valued function of a real variable. Let us assume that the function is called f and is defined upon a set D , its domain. Then, f is said to be continuous at the point x in D provided that the following condition is met:

For all $\epsilon > 0$ there is $\delta > 0$ such that for all $t \in D$, $|t - x| < \delta$ implies $|f(t) - f(x)| < \epsilon$. This may be slightly shortened in our new symbolic language as

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad \forall t \in D, \quad (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

We also say that f is continuous on D if it is continuous at each individual $x \in D$. That is,

$$\forall x \in D \quad \forall \epsilon > 0 \quad \exists \delta > 0 \quad \forall t \in D, \quad (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

If D is a closed interval $[a, b]$, then we can say, of course,

$$\forall x \in [a, b] \quad \forall \epsilon > 0 \quad \exists \delta > 0 \quad \forall t \in [a, b], \quad (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

Exercises

1.2.13 Negate the statement which expresses what is meant by the continuity of f at x .

1.2.14 Let us suppose that D is an interval $[a, b]$. Then, why does the statement below mean something other than the statement that f is continuous at each $x \in [a, b]$? What is actually said here?

$$\forall \epsilon > 0 \exists \delta > 0 \forall t \in [a, b] \text{ and } \forall x \in [a, b], (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

What if D is an open or a half-open interval? Can you think of a function which is continuous on D but does not satisfy this new property?

1.2.15 Why is the statement below different from all of the previous statements? What does it actually say about f ?

$$\exists \delta > 0 \forall \epsilon > 0 \forall t \in [a, b] \text{ and } \forall x \in [a, b], (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

To conclude this section, we should notice that the definition of continuity at a single point x given above, which is the definition actually used by mathematicians in their own work, is *not* the same as the definition which is given in the typical calculus book. For, if the domain D is for example the set of positive integers, then the statement of the definition given here implies automatically that f is continuous upon D , whereas according to the calculus book that is not so. Mercifully, the definition used in the typical calculus book *is* equivalent to the above if the domain D is an open interval, which is what the calculus book is mostly interested in. The discrepancy between the two definitions of continuity is pointed out here to avoid future confusion, and also to point out but one example of the shortcuts which have to be made in order to get the job done in the time allotted for teaching calculus – we actually feel forced to engage in “white lies.”

We should also notice that a statement with a “for all” quantifier in it does not imply the existence of anything in particular, for it may in fact be true precisely upon all the elements in the empty set. This may seem strange, forced, and a little bit tricky. Nevertheless, it does go along with ordinary understanding. A sign in a park which says “All dogs in this park must be kept on a leash” certainly does not imply that there is always at least one dog in the park.

But the observation above about the “for all” quantifier does have another consequence. Namely, the negation of a statement with “for all” in it is a statement with “there exists” in it, which is a bald assertion that something exists. What do we mean by saying that something exists? That’s a good question. To take what ought to be a rather everyday example, we all believe that irrational numbers exist. Otherwise, we probably would not be meeting together in a classroom to study mathematics. But have you, the reader, ever really met an irrational number? Probably not. What, then, ought it to mean if one says that irrational numbers exist?

1.2.3 Negations

Before concluding our review of logic, we should give some special emphasis to the negation of propositions and the negation of open statements which contain quantifiers. To be able to formulate negations properly is an essential skill when formulating proofs precisely because we often find it easier or more natural to perform a proof by contradiction, or by proving the contrapositive of the given statement.

Therefore, this section is a compendium of what has already been said about this important topic, repeating the discussion of negations which already has been done in the previous sections for emphasis and for the sake of additional

clarity.

One who has mastered the techniques of formulating negations understands completely that to form a negation of a complicated compound statement can be reduced to a completely mechanical procedure, once the procedure has actually been learned. One who has actually mastered the methods of negation can produce negations without worry, trepidation, or errors. Moreover, since this is a course in mathematics, which is a human activity which has its underpinnings in logic, it is absolutely necessary to learn the procedure for formulating negations of compound statements, to learn it well, and to be able to carry out that procedure whenever a situation demands it. The intent of this section is, then, to review and to reiterate the procedure for producing the negation of a statement.

First, we need to review the negation of compound propositions. The first column in the table below gives a compound proposition which is to be negated. The second column gives the obvious negation of it. The third column gives an equivalent for the negation, which is carried out one more step:

original	negation	more basic form of negation
$p \vee q$	$\sim (p \vee q)$	$\sim p \wedge \sim q$
$p \wedge q$	$\sim (p \wedge q)$	$\sim p \vee \sim q$
$p \Rightarrow q$	$\sim (p \Rightarrow q)$	$p \wedge \sim q$
$\sim p \vee q$	$\sim (\sim p \vee q)$	$p \wedge \sim q$

The above equivalences are established already in Subsection 1.2.1, above, where appeal was made to truth tables

in order to justify the equivalences. On review of that section, you would notice two things. First, the table just above is in fact just a little bit redundant because some lines in it can be derived from others. Second, the table above would indicate that $p \Rightarrow q$ and $(\sim p) \vee q$ are in fact logically equivalent statements, on the grounds that they have the same negation. That should not be a surprise. The two were seen back in Subsection 1.2.1 to share the same truth table.

To those who are new to the subject of logic, the negation of an if-then statement, which is represented in logical symbols as $p \Rightarrow q$ often seems to cause confusion, trouble, and unease. They have seen that by appeal to truth tables, the statement $p \Rightarrow q$ is logically the same as the statement $\sim p \wedge q$ and therefore the negation of $p \Rightarrow q$ can be nothing other than the negation of $\sim p \wedge q$. That is, the negation of $p \Rightarrow q$ can be nothing other than $p \wedge \sim q$. A possible reason for the confusion is that some of us apparently have trouble with believing what the truth table is telling them, perhaps not believing that what is in the truth table reflects their real-life experiences.

More serious thought on the matter ought to clear up any possible confusion. Let us look at just one such if-then statement which is taken from everyday life and experience and pick apart what we actually mean by it:

If there is ice on the water in the birdbath, then the temperature is or has recently been below freezing.

The above if-then statement has two parts. We could denote by p the statement “there is ice on the water in the birdbath” and by q the statement “the temperature is or has recently been below freezing.” Then, the if-then statement says the same thing as $p \Rightarrow q$.

Surely, we all agree that the above if-then compound statement is true, at least in the context that we clearly mean

it. That is, whatever ice there might be in the birdbath is present due to natural causes which are related to the surrounding temperature, not caused by some kind of human activity such as putting ice cubes into the birdbath or affixing a refrigeration unit to same. The qualification “has recently been” of course is somewhat elastic, as often occurs in everyday life. We agree to interpret it in a reasonable manner, not perversely. But let us now analyse what we really mean by this if-then statement from everyday life, as related to the possible truth or falsity of the two components of it, the hypothesis p and the conclusion q :

- The first thing to understand is, our everyday experience and everyday speech agree that the statement taken as a whole is true. Someone who is making this statement does not need to expect that there is actually ice on the water in the birdbath, and the hearers do not need to expect that, either. Somebody could say this if-then statement in mid-July, after all, and in Central Alabama there is almost surely no ice on top of any water in the birdbath. Thus, we would surely agree that the validity of this if-then statement is unaffected by the calendar. Applying this observation in the logical context, we are not caused to doubt the statement $p \Rightarrow q$ in the situation that both p and q are false.
- We could also be in mid-January, right after a cold front has moved in, and there is no ice on the birdbath because the temperature has just dropped steeply and the water in the birdbath has not had sufficient time to react, by freezing. This situation does not cause us to be skeptical about the if-then statement, either. The statement $p \Rightarrow q$ therefore does not become false merely because p is false (no ice on the water in the birdbath), even if q is true (the temperature is below freezing at the moment of observation).

- It also does not cause us to doubt the validity of the if-then statement if there is ice on the birdbath and in fact the weather is quite cold. We do agree that the cold weather is what has created the ice. In this case, both p and q are true, and we accept that $p \Rightarrow q$ is true. However, let us again notice that we did not reject the statement in the two previously enumerated scenarios, either.
- The only situation which would cause us to doubt would be if there is ice on the water in the birdbath at the same time that the temperature is above freezing and has remained above freezing for a reasonably length of time. Such could happen, of course, if the ice had been formed by some artificial means. But we have excluded that possibility. In this case, p is true, and q is false. And if we see this happening with our own eyes, we consider the statement $p \Rightarrow q$ to be false.

The point of the above discussion is to relate the meaning of the logical implication $p \Rightarrow q$ to what we mean by an if-then statement in ordinary life and ordinary speech. In fact, we **do** accept the validity of an if-then statement unless we are confronted by a situation in which p is true and, simultaneously, q is false. And exactly that is the situation in which we deem the statement $p \Rightarrow q$ to be false. Thus, by careful analysis of an everyday example the table entry above for the negation of $p \Rightarrow q$ is seen to be correct. By the negation we actually do mean $p \wedge (\sim q)$.

We now review the negation of statements which contain quantifiers, followed by an open statement at the end. Here, the rule is very simple. There are two steps:

First, reverse **all** of the quantifiers one by one. This means, change each \forall to \exists , and change each \exists to \forall . In doing this, of course, “all” means all, not some of them, not all but one of them, but all of them. And do not move quantifiers from one place to another before doing this step. As we have seen, that can cause the meaning of the entire statement to change.

Second, form the negation of the open statement which came after all of the quantifiers.

Also note that when doing the negation of a statement with quantifiers in it the location or usage of such things as “such that” may easily switch themselves around from one part of the long statement with quantifiers in it to some other place in the long statement. Things like “such that” actually do not contribute much of anything to logical meaning. They are present mainly as lubricants, intended to help us human beings better to say or understand things. The best advice, therefore, is to write down the long statement without using these little lubricants. Just use quantifiers. Then do the negation as described above. Finally, put the whole thing back into words.

Our standard example of a statement was the definition of the continuity of a function f which is defined upon a domain D at a point x . That definition was

For all $\epsilon > 0$ there is $\delta > 0$ such that for all $t \in D$, $|t - x| < \delta$ implies $|f(t) - f(x)| < \epsilon$.

When this definition was put into its symbolic form, it was seen to be:

$$\forall \epsilon > 0 \quad \exists \delta > 0 \quad \forall t \in D, \quad (|t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon).$$

in which the “such that” was suppressed.

Now, completely in accord with the analysis above, the negation of this statement is

$$\exists \epsilon > 0 \quad \forall \delta > 0 \quad \exists t \in D, \quad (|t - x| < \delta \quad \text{and} \quad |f(t) - f(x)| \geq \epsilon).$$

If we would like to translate this resulting negation back into words, then it ought to read as

There exists an $\epsilon > 0$ such that for all $\delta > 0$ there exists a $t \in D$ such that $|t - x| < \delta$ and $|f(t) - f(x)| \geq \epsilon$.

Notice that, when the statement of the negation is put back into words, there are two places where “such that” has been put in, not just one as in the original. And neither of these is in the location where the “such that” was situated in the original. Note also that a couple of times the word “a” has been inserted. None of these things actually have anything to do with the logical content. The sole reason for adding them is in order to improve the human readability of the final product. That is a serious reason, too, of course.

The bottom line regarding the negation of a statement which contains quantifiers is that the safest thing to do is to start by rewriting the statement using the standard quantifier symbols of symbolic logic. When doing this or after having done this, the filler words such as “a,” “an,” “the,” and “such that” may be safely ignored. Then the next step is to write out the negation by following steps one and two above. Finally, if it is needed or desirable to have the negation back in words again, put in the needed filler words to make it readable and understandable for humans.

Chapter 2

Construction of the rational numbers

2.1 Introductory Remarks

the subject of mathematics from its foundations, one of the things which we should do is to perform the steps of the construction of the real number system. We will do this in steps which we will complete in the next chapter. Here, we lay the foundations for that chapter by reviewing the most basic and relevant properties of the set of integers and then by doing a formal construction of the set of rational numbers as what is called a *quotient field* over the integers. Unfortunately, we will not be able to carry out all aspects of the program for this chapter due to obvious constraints of time. Thus, rigor and step-by-step development will not universally be followed in this chapter at the expense of all other goals. In particular, we will not be able to prove everything we need of the properties of the integers, simply because that would require us to spend most of a semester doing nothing else. Nevertheless, enough of the construction will be given or indicated that the difficulties and peculiar problems involved should be made clear to the student. That is, if we do not have time systematically to explore all answers to all possible questions, which in fact we will not (and, one should realize, some questions may not have answers!), then at least we should be aware of the questions. One reason for that is, some of the questions that can or could be posed have in fact given rise to various important areas of mathematics. Of course, we continue the pattern already used in the previous chapter. There will be problems interspersed in the material for the student to solve, and further progress may use or even depend upon the solution of some of those problems.

2.2 Basic properties of the integers

The set of integers will be denoted by the standard symbol \mathbf{Z} . Two algebraic operations are defined upon \mathbf{Z} , addition and multiplication. However, the set of integers also satisfies several other properties as well. We discuss these properties in more detail in the following sections.

2.2.1 Algebraic properties of the integers

The addition satisfies the associative and commutative laws, and there is a zero, and there are additive inverses. The multiplication is associative and commutative, and there is a multiplicative unit. The two operations of addition and multiplication are linked in the distributive law. More specifically, where x, y, z signify arbitrary integers, then we can list the above properties more systematically. It remains to observe that the “equals sign” means that what on either side of it may freely be replaced by what is on the other. As an example of what this means, consider that item A4 below says that the expression “ $x + (-x)$ ” may be freely replaced by the expression “0.” We are all accustomed to that, of course, because of endless drill on simplifying algebraic expressions and solving equations in school exercises. But due to that mental conditioning it is less obvious that also the expression “0” may be freely replaced by “ $x + (-x)$ ” where x is any particular integer. And sometimes one needs to do just that in order to get the job done.

The entire list of the algebraic laws which the integers obey now follows. It is understood, of course, that x and y and z can refer to arbitrary integers.

$$\text{A1. } x + (y + z) = (x + y) + z$$

$$\text{A2. } x + y = y + x$$

$$\text{A3. There is an integer called zero, denoted by } 0, \text{ satisfying } x + 0 = x$$

$$\text{A4. To each integer } x \text{ there corresponds an integer } (-x), \text{ such that } x + (-x) = 0.$$

$$\text{M1. } x(yz) = (xy)z$$

$$\text{M2. } xy = yx$$

$$\text{M3. There is an integer } 1, \text{ not equal to } 0, \text{ which satisfies } 1x = x$$

$$\text{M4. } x(y + z) = xy + xz$$

Before stating the exercises which follow, it should be emphasized very much that in doing them **nothing** should be assumed except what is stated in A2 through A4 and M1 through M4, above, and the immediately preceding discussion of what is meant by “equals.” The exercises can be done, indeed must be done, based upon these properties and upon nothing else. If the student has not done similar exercises previously, they may not all be that easy. For, one of the hardest things to do is to put aside the “knowledge” based upon previous experience and to meet the problems with

nothing in hand except for the above-listed properties. Nobody can learn to play a guitar without passing through a stage of sore muscles and fingertips. Nobody can learn mathematics without learning to concentrate and to put forth mental effort and struggle, nor, even more to the point, without learning to prove the next step in a chain of logical development by appeal to nothing whatsoever except for that which came before in that same chain.

Exercises

2.2.1 Based upon the properties A1 through A3, the zero is unique. That is, if A1 and A2 hold true, there can be no other integer besides 0 for which the property A3 holds true.

2.2.2 Given x , the integer $(-x)$ is unique, based upon the properties A1 through A4.

2.2.3 If x and y and z are any three integers such that $x + z = y + z$, then $x = y$. This is often referred to as the Law of Cancellation.

2.2.4 Based upon M1 through M3, the integer 1 is unique.

2.2.5 Based upon all of the properties and in particular upon M4, $0x = 0$ for every integer x . Note that this is **not** the same thing as to say, “If x and y are any two integers such that $xy = 0$, then either $x = 0$ or $y = 0$.” This second statement does also happen to be true within the set of integers. But as we shall see from later discussion it does **not** follow from the list of algebraic axioms given above this problem set.

2.2.6 For every integer x it is the case that $(-x) = (-1)x$. Note that, once this matter has been settled, we can simply write $-x$ instead of $(-x)$ and $y - x = y + (-x)$.

2.2.7 $(-(xy)) = (-x)y = x(-y)$

2.2.8 If $x = y$, then $zx = zy$, where z can be any integer (still true, but not so interesting, if $z = 0$). The statement that if $zx = zy$ with $z \neq 0$, then $x = y$ would obviously follow if we know that, whenever the product of two integers is zero, then one of the two has to be zero. But please see the remark in Exercise 2.2.5 above.

Now, certainly we agree that the integers satisfy the properties listed above. However, these properties do not fully describe the integers. The properties above are considered by mathematicians in the category of *algebraic* properties. The same properties listed above also hold in many other situations, and in still other situations they almost hold, with only one or two of them failing. Let us try to think of some examples of this. If we do, then perhaps we can add some problems to those just above.

2.2.2 Order properties of the integers

One of the big differences that the integers have from some other systems which satisfy the above algebraic axioms is the fact that the integers also satisfy order relations “ $<$ ” and “ \leq .” For, any two integers can be compared using either of these. An efficient way to describe and to establish these relations is to postulate the existence of two sets, P (set of positive integers) and N (set of negative integers). The properties of these sets are that

- I1. The union of N and P and $\{0\}$ is all of the integers, and the three sets are pairwise disjoint.
- I2. An integer n is in P if and only if $-n \in N$
- I3. The sum of any two elements of P is in P
- I4. The product of any two elements of P is in P

We will say that $x < y$ (or $y > x$) if and only if $y - x \in P$. We will show in exercises that “ $<$ ” forms a *strict* order relation on the integers. That is,

- O1. $x < x$ is never true.
- O2. If $x < y$, then $y < x$ is false.
- O3. If $x < y$ and $y < z$, then $x < z$.

Also very important is the fact that this order relation pertains to any two integers. The Law of Trichotomy states that

- O4. Given any two integers x and y , either $x < y$ or $x = y$ or $y < x$.

Remark: It is the property O4 which guarantees that the ordering of the integers allows us to compare **any two** integers. An order relation which allows one to compare any two elements of a given set is a total ordering. One may not have this, and still have an order relation, that is a relation satisfying O1, O2, and O3, or perhaps satisfying the properties NO1 through NO3 which are listed below. The most obvious examples of such **partial orderings** are based on properties of set inclusion.

As far as the set of integers is concerned, the properties O1 through O4 will be seen in the exercises to follow from the properties I1 through I4.

Exercises

2.2.9 $1 \in P$, and more generally $z^2 \in P$ for all integers $z \neq 0$.

2.2.10 If x and y are integers satisfying $xy = 0$, then either $x = 0$ or $y = 0$.

2.2.11 Show the properties O1 through O4, based upon I1 through I4.

2.2.12 If $x > y$, then

(a) $x + z > y + z$

(b) $xz > yz$ if $z > 0$

(c) $xz < yz$ if $z < 0$ (order of the inequality is reversed)

2.2.13 Using the order properties listed in 2.2.2 or, in the alternative, the order properties O1 through O4, it should now follow that if $zx = zy$ with $z \neq 0$, then $x = y$ (See Exercise 2.2.5 for a previous discussion of this topic).

It should be obvious that “ \leq ” is also an order relation of a slightly different kind. It is clearly related to the previous order relation, with only small changes in what the describing axioms say. They would now have to read something like

NO1. $x \leq x$ is always true.

NO2. $x \leq y$ and $y \leq x$ are both true if and only if $y = x$

NO3. If $x \leq y$ and $y \leq z$, then $x \leq z$.

In fact, it might even be possible to combine the first two properties in one. However, observe that the property O4, which one would definitely wish to keep, is much harder to state in terms of “ \leq ” than it was previously.

2.2.3 The Well Ordering Principle and Mathematical Induction

Now, since we have listed the order properties of the integers, we note one other important property, related specifically to the *positive* integers:

The **Well Ordering Principle** states that any non-empty subset of the positive integers must have a smallest element in it.

It should be clear to the reader that the well ordering property is distinct from the order properties listed in the previous subsection, and it is one of the things which make the set of integers quite different and distinct from some other sets with otherwise similar properties. Among the first consequences of the Well Ordering Principle, we list the following, in the form of exercises.

Exercises

2.2.14 As I am sure we are all aware, it is generally agreed that 1 is the smallest positive integer, there being no positive integer which is less than 1. Based upon the properties of the integers previously described in this section, and not upon what “everybody knows” can you prove this? (Hint: Suppose that the set of integers n such that $0 < n < 1$ were not empty. Then by the Well Ordering Principle it would have a least element. Call that least element by the name N . Then, based upon the laws of inequalities previously established, can you show that $N^2 < N$ would follow?)

2.2.15 Let n be any integer. Show that it follows from the result stated in the previous problem that there are no integers between n and $n + 1$.

Further to illustrate the use of the Well Ordering Principle, we show the following:

Proposition: Given any two positive integers m and n , there exist two unique integers $q \geq 0$ and r , satisfying the two conditions that $m = nq + r$ and that $0 \leq r < n$.

Proof: First, we notice that if $n = 1$, the proof is obvious. In that case, we have $m = m \cdot n$ exactly, with $r = 0 < n$. Having taken note separately of this simple case, we therefore can proceed with the rest of the proof on the assumption that $n > 1$.

Now, with $n > 1$, let us define

$$K = \{x | x = ny, \text{ for some positive integer } y \text{ and } x > m\}.$$

To see that $K \neq \Phi$, note that $mn > m$ and hence $mn \in K$. This can be seen to follow from the fact that $n > 1$, in the following steps:

$n > 1$ implies that $n - 1 > 0$ (definition of what is meant by $n > 1$).

In turn, $mn - m > 0$ because $m(n - 1) > 0$ (being the product of two positive integers), and we have established that $mn \in K$, showing that K is not empty.

Now, to continue the proof we invoke the Well Ordering Principle in order to notice that K has a smallest element. Since every element of K is a multiple of n by some other integer, we can assume that this smallest element of k may be represented as $x_0 = nk_0$. By the definition of k_0 , we see that $nk_0 > m$, whereas $n(k_0 - 1) \leq m$. It is further clear that $k_0 > 0$. For, $nk_0 \in K$ implies that $nk_0 > m > 0$ and it is not true that $0 = n \cdot 0 > m > 0$. Consequently, since $k_0 > 0$ it follows by Exercise 2.2.14 that $k_0 - 1 \geq 0$.

If we now define $q = k_0 - 1$, then we can also define $r = m - nq$, and it follows that $r \geq 0$. Also, since $nq + r = m < nk_0$, it is seen that $r < nk_0 - nq$, whence $r < n$.

Finally, we must address the question of unicity. Let us suppose that there were to exist $q' \geq 0$ and r' which satisfy $0 \leq r' < n$ and that $m = nq' + r'$, possibly in addition to the q and r already found, above. We need to show that the only way this can happen is that, in fact, $q' = q$ and $r' = r$.

First, we show that $q' = q$. If $q' \neq q$, then either $q' > q$, or $q' < q$. That $q' > q$ is not possible. For, then by Exercise 2.2.14 we have $q' \geq q + 1$, and then $nq' \in K$ and hence $nq' > m$. If on the other hand it were the case that $q' < q$, then we would have $m = nq + r = nq' + n(q - q') + r = nq' + r'$, and we see that then $r' = n(q' - q) + r$ would have to hold true. But then it would also follow from Exercise 2.2.14 that $q - q' \geq 1$, whence $q' - q \leq -1$, and $n(q' - q) < -n$. And then $r' = n(q' - q) + r < -n + r < 0$, which violates the restriction that $r' > 0$. Thus, $q' = q$.

Now, since $q' = q$, we have both $m = nq + r$ and $m = nq + r'$. Hence, both $r' = m - nq$ and $r = m - nq$, and the equality $r' = r$ must follow. The proof is completed.

Related to the Well Ordering Principle for the positive integers is the **Principle of Mathematical Induction**:

Let $P(n)$ be a statement which depends upon the positive integer n . If the following two conditions hold, then $P(n)$ is true for all positive integers n

- (i) $P(1)$ is true.
- (ii) For any $n \geq 1$ the implication $P(n) \Rightarrow P(n + 1)$ is true.

Exercises

2.2.16 The Well Ordering Principle for the positive integers and the Principle of Mathematical Induction are

logically equivalent. Note that both directions in this proof require the result of Problem 2.2.14. Thus, one needs to show that Problem 2.2.14 follows from the Principle of Mathematical Induction as well as from the Well Ordering Principle.

2.2.17 Report to the class about the following very interesting question:

We have seen in Exercise 2.2.14 that we needed to prove that there is no integer n for which $0 < n < 1$. The proof used the Well Ordering Principle. Could it also be true that the Well Ordering Principle would logically be implied if one were to assume that 1 is the first positive integer?

In fact, while it is tempting to believe that this just has to be the case, any attempt to prove it must fall short. To see why, a good place to start is at the article “Peano Axioms” on Wikipedia. A brief explanation is that the axioms of algebra and the axioms of order on the integers deal with integers as *elements* of the set of integers. An axiom added to these which would say that there is no integer between any integer n and $n + 1$ would partake of the same logical nature. The problem is that a set with precisely this set of axioms could exist which contains all of the integers we think that we “know” and also contains an element which is greater than all of the integers we think that we “know.”

But (taking just one of our two equivalent statements) the Principle of Mathematical Induction says that a *set* of positive integers defined in a certain way will be equal to the set of *all* positive integers. This statement characterizes the set of integers in a way which statements which merely involve elements can not do. Thus, in fact either the

Principle of Mathematical Induction or the Well Ordering Principle must be taken as an axiom in addition to the algebraic axioms and the order axioms, in order to describe and characterize fully the set of integers.

Before moving on, we note the possibility of stating definitions which have an inductive component. We have already done this, in fact, in Section 1.1.2, where we defined

$$\bigcup_{i=1}^n A_i = \left(\bigcup_{i=1}^{n-1} A_i \right) \cup A_n$$

and

$$\bigcap_{i=1}^n A_i = \left(\bigcap_{i=1}^{n-1} A_i \right) \cap A_n.$$

We should also note that for a number a and for an integer $n > 1$ it is similarly appropriate to define a^n by

$$a^1 = a \quad \text{and} \quad a_n = a^{n-1}a \quad \text{for} \quad n > 1$$

Moreover, we can define $n!$ (read as “n factorial”) by

$$0! = 1 \quad \text{and} \quad n! = n \cdot (n-1)! \quad \text{for} \quad n \geq 1$$

The above are but two of many possible examples in which explicit definitions can be given, which avoid the use of “dot, dot, dot” in the middle of a definition.

2.2.4 Decimal and other representations of integers

A standard way to represent integers is in terms of powers of ten. This representation is so familiar to us that we often have to stop and think of what is meant. For example, to write 536 actually means $5 \times 10^2 + 3 \times 10^1 + 6 \times 10^0$, which is then the same thing as $5 \times 100 + 3 \times 10 + 6 \times 1$. With some justification, then, the number ten is called a **base** for the standard notation for integers which is used in the above example.

Careful reflection ought to lead us to reflect that in the above description of the “standard” representation of numbers the only things which cause us to use the number ten as our building block are our prior conditioning and long force of habit, which in turn seem to be based mainly upon the fact that almost all of us are born with ten fingers. Indeed, we do not even universally stick to base ten in everyday life. We buy and sell eggs by the dozen, not by tens, and we even have an expression for a dozen dozens, namely “one gross.” We also agree that 2 teaspoons make a fluid ounce, and 8 fluid ounces make a cup, and two cups make a pint, with two pints making a quart, and four quarts making a gallon. These measurements of volume are evidently based upon successive powers of two, not ten.

As should be clear from the previous paragraph, there is nothing particularly sacred about the number ten. We could just as well represent our numbers in terms of two, not ten. If we were to do this, then, using 0 as a placeholder as we are already accustomed to do, we would do something like the following. In which a “B” after the number signifies that it is to be read in terms of base 2, not base ten:

decimal		binary
0	=	0 <i>B</i>
1	=	1 <i>B</i>
2	=	10 <i>B</i>
3	=	11 <i>B</i>
4	=	100 <i>B</i>
5	=	101 <i>B</i>
6	=	110 <i>B</i>
7	=	111 <i>B</i>
8	=	1000 <i>B</i>

One may continue in like fashion, with, for example, 16 corresponding to 10000*B* because $16 = 2^4$; 17 to 10001*B*; 32 to 100000*B* because $32 = 2^5$, 34 to 100010*B*, and so on.

The obvious disadvantage to using 2 as a base for our counting is that the representations for our numbers get very long very fast. But nevertheless base 2 arithmetic is very important these days because it reflects almost perfectly what a computer's internal circuits are doing.

Another base which is often used around computers is base 16, or "hexadecimal." For this, we need some new digits. We can start out with the standard 0, 1, ..., 9. But then we have to say that 10 = A*h*, 11 = B*h*, 12 = C*h*, 13 = D*h*, 14 = E*h*, 15 = F*h*, and then finally 16 is 10*h*. Continuing, one might note that $50h = 5 \times 16 = 80$ and $FFh = 15 \times 16 + 15 = 255$, which is one less than $100h = 16^2 = 256$.

Finally, let us note that any integer $p > 1$ can be used as a base. The way to do this is to represent any given number by its expansion in terms of powers of the number p . For example, if $p = 3$ we can represent 10 as $3^2 + 1$, and therefore 10 is 101, represented in base 3. And similarly 19 is $2 \times 3^2 + 1$ and therefore is 201 if represented in base 3.

Exercises

2.2.18 Represent 536 in terms of base 2, base 3, and base 16 as well as in terms of base 10. Represent 21 in base 2 and base 16 as well as in terms of base 10 and base 3.

2.2.5 Other properties of the integers

Finally, we mention that certain other properties of the integers will be taken here as known, without further discussion. Some of these properties are in themselves quite important in their own right, but their proofs are more appropriately learned elsewhere.

Examples of such properties include the existence of prime numbers and results related to them, such as

- The list of prime numbers is not finite; there is no largest prime.
- If p is a prime number and p is a divisor of the product ab , where a and b are both integers, then either p is a divisor of a , or p is a divisor of b . The previous statement may or may not be true if p is not prime.

- The Prime Factorization Theorem, which states that every positive integer can be factored as a product of primes, and further states that if the primes are listed in their natural order then the factorization is unique.
- The Euclidean Algorithm, and its uses for finding the greatest common divisor and the least common multiple of any two positive integers.

We leave this section with the observation that the listed properties, among some others, are very important for day-to-day use in the manipulation of integers. Most particularly, they are important in handling a sum of fractions with different denominators and writing the result in lowest terms. For, in doing such problems it is necessary to find the least common multiple of the denominators or else the problem can easily become quite unmanageable.

Exercises

2.2.19 If you have taught the arithmetic of fractions or have reason to believe that you might ever have to teach such material, what approach would you take?

2.3 More about the algebraic properties of the integers: a small excursion

It is stated in the previous section that the set of integers has two algebraic operations and that these two algebraic operations obey certain laws. It is further stated that there are properties relating to order. Here, we play a little bit of

theme and variations. One purpose of this section is to show the power of abstraction in mathematics, which consists of the distillation of general principles from a number of specific examples or manifestations, followed by the exploration of the consequences of those general principles taken in isolation from other factors. Abstraction has been one of the driving principles behind the development of mathematics in the last one hundred fifty years or so, and it has been very successful in keeping things manageable while the field has grown so greatly in size and complexity.

Consider the following examples:

Exercises

2.3.1 We construct a set containing precisely two elements, which we will call 0 and 1. We construct rules for addition and multiplication, as follows

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

Show that with the indicated algebraic operations the set $\{0, 1\}$ is an algebraic system which obeys all of the laws of algebra which have been prescribed for the integers. It goes without saying, of course, that the order properties fail to work. Nevertheless, the result shown in Exercise 2.2.13 is valid in this system.

2.3.2 In a manner similar to the previous example, we can construct an algebraic system defined upon the set $\{0, 1, \dots, n-1\}$, in which the sum of any two numbers is their usual sum, unless that sum exceeds $n-1$, in which case we take the remainder upon division by n . Parenthetically, one might note that the effect is to identify n with 0. We define the multiplication similarly. As an example, of this, we take n to be 12. Then an everyday specimen of this system is seen on a clock face. Show that the algebraic operations for the integers all hold here. Show also that it is sometimes possible for two numbers to be non-zero, but their product in this 12-based system can come out zero nevertheless. Can you account for the fact that the result in Exercise 2.2.13 does not hold here, but it does hold in Exercise 2.3.1? Are there values of n (other than 12, obviously, and in addition to 2) for which the product of two non-zero numbers is never zero?

2.3.3 Consider the set of all subsets of a given universal set, X . We call this set of subsets of X the **power set** of X , and it is usually denoted by $\mathcal{P}(X)$. We define two operations upon $\mathcal{P}(X)$. One of them is intersection, which we already know. That one will play the role of the multiplication. The second operation, which will serve as the addition, is the **symmetric difference** defined by

$$A \triangle B = (A \cup B) \setminus (A \cap B).$$

Parenthetically, this operation corresponds to an exclusive “or” operation in logic and to the XOR bitmap operation on computers. Here, however, the problem is to show that $\mathcal{P}(X)$ satisfies the listed algebraic properties of the integers, using these two indicated operations.

The axioms which define the basic algebraic properties of the integers are those which, in the modern field of algebra, are taken to define an algebraic structure called a **commutative ring with unit**, the unit being, of course, 1. It is obviously possible to come up with structures which have some but not all of these properties, but inside of which it is still meaningful and relevant to do algebra-like activities. Consider the following further examples:

Exercises

2.3.4 The set of all functions which are continuous on $[0, 1]$ but which all satisfy $f(0) = 0$ (nothing here corresponds to the number 1, so this set seems to be a ring without unit). Also, again it is possible to have two functions in this set whose product is zero, even though neither of the two functions are zero.

2.3.5 If we restrict the functions in the previous example to be polynomials, then it is no longer possible to have two non-zero polynomials whose product is identically zero (why?).

2.3.6 The set of all 2 by 2 matrices with integer coefficients, with the usual rules for addition and multiplication of them by one another. Here, the commutative law for multiplication clearly fails, which in turn requires that the distributive law for multiplication, M4, has to be amended and written as a “left distributive law” and a “right distributive law” in order fully to describe the situation. Also, the product of two non-zero two by two matrices with integer coefficients can come out to be zero.

We can devote a little bit of time to presentations and proofs of some of the properties of some of the previous examples. Any volunteers?

2.4 The construction of the rational numbers as a quotient field

The construction of the rational numbers from the integers is the classic example of what is called in abstract algebra the construction of a quotient field over the integers. To perform the construction, one begins with the Cartesian product $\mathbf{Z} \times (\mathbf{Z} \setminus \{0\})$. Upon this set, an equivalence relation is then imposed, namely we will say that $(m, n) \sim (p, q)$ if and only if $mq = np$.

Note 1: The Cartesian product of two sets was defined in the previous chapter, in the subsection 1.1.3.

Note 2: What is an equivalence relation? The definition is seen in the problem immediately below.

Note 3: Clearly, the notation (m, n) is a different notation for what we all agree is the common fraction m/n . The reason why we are using the funny notation is to highlight the logical steps which are involved in the construction of the rational numbers. We will go ahead and write our fractions as fractions very soon. In the meantime, it may help you to relate to the standard properties of the rational numbers, which we are in the business of establishing, if you think of them using the fraction notation. But be mindful that in order to make sure that the standard properties do work out we do need to work within the indicated Cartesian product set $\mathbf{Z} \times (\mathbf{Z} \setminus \{0\})$ until we have finished that task.

Exercises

2.4.1 The advertised equivalence relation is indeed an equivalence relation. We say that “ \sim ” is an equivalence relation on a set S , provided that

E1. $s \sim s$ for all $s \in S$

E2. $s \sim t$ if and only if $t \sim s$

E3. $s \sim t$ and $t \sim u$ imply together that $s \sim u$

Now, we define the algebraic operations of addition and multiplication upon our set $\mathbf{Z} \times (\mathbf{Z} \setminus \{0\})$. The addition is defined by

$$(m, n) + (p, q) = (mq + np, nq),$$

and the multiplication is defined by

$$(m, n)(p, q) = (mp, nq).$$

We will now define the set of rational numbers \mathbf{Q} to be the quotient set $(\mathbf{Z} \times (\mathbf{Z} \setminus \{0\}))/\sim$. That is, the set of rational numbers Q is the set of **equivalence classes** which is generated by the above equivalence relation. In \mathbf{Q} , of course, it is customary to write (m, n) as $\frac{m}{n}$. We will start to do that quite soon. But the immediate problem, is that we have to show that this all actually works. In particular, it is necessary to show that the algebraic operations which we would like to use are valid. As a trivial, concrete example of what we have to do, we have to show not only how to

add two fractions such as $\frac{1}{2}$ and $\frac{1}{3}$, but we have to show that the procedure still works if we take any other fractions which are respectively equivalent to the two given ones, by providing a sum which is equivalent to the sum of the two given fractions. We also have to do the same thing for multiplication. Hence, the following exercises.

Exercises

2.4.2 Show that the addition is well-defined. That is, show that if

$$(m_1, n_1) \sim (m_2, n_2) \text{ and } (p_1, q_1) \sim (p_2, q_2),$$

$$\text{then } (m_1, n_1) + (p_1, q_1) \sim (m_2, n_2) + (p_2, q_2).$$

The purpose of this exercise is to show that the addition defined in the entire set $\mathbf{Z} \times (\mathbf{Z} \setminus \{0\})$ is compatible with the definition of \mathbf{Q} , that is, most particularly, the proposed method for addition does not do violence to the equivalence relation that underlies the definition of \mathbf{Q} .

2.4.3 Show that the multiplication is well-defined.

2.4.4 Prove the following two statements:

- (a) The algebraic properties A1 through A4 and M1 through M4 which held for the integers also hold for \mathbf{Q} .

- (b) In addition to part (a), there are multiplicative inverses for all elements (m, n) for which $m \neq 0$. The result of this exercise is to show that \mathbf{Q} has the algebraic properties of a field (see just below these exercises for the definition of a field).

2.4.5 Establishing an order relation on the rational numbers can be done by at least two methods which both lead in the end to the same result. In this problem, we explore one of these two methods, and in the next problem we look at the other of the two.

We can define sets of rational numbers which are similar in their functionality to those sets P and N of integers which we have already defined. Let us call them \mathcal{P} and \mathcal{N} . We will say that the ordered pair (m, n) which represents a rational number is in \mathcal{P} if and only if $mn > 0$ and is in \mathcal{N} if and only if $mn < 0$. Show:

- (a) that these sets \mathcal{P} and \mathcal{N} are compatible with the equivalence class structure of the rational numbers.
- (b) that the properties I1 through I4 which are listed in Section 2.2.2 can be successfully extended to the new sets.

We can then define an order relation on all of \mathbf{Q} by agreeing that $(m, n) < (p, q)$ if $((p, q) - (m, n)) \in \mathcal{P}$.

2.4.6 In this problem, we extend the order relation already defined for the set of integers directly, without bothering to construct first the sets \mathcal{P} and \mathcal{N} that were constructed in the previous problem, Exercise 2.4.5.. We say that $(m, n) < (p, q)$ if both $n > 0$ and $q > 0$ and it is true that $mq < np$. Show the following:

- (a) Independently of Exercise 2.4.5 but only based upon the definition in the previous sentence, the order relation just now defined makes sense on all of \mathbf{Q} and in fact defines an order relation between any two elements of \mathbf{Q} which are not equal to each other. For, every pair (m, n) in $\mathbf{Z} \times (\mathbf{Z} \setminus \{0\})$ is equivalent to another pair in which the second number is positive (that is, if $n < 0$ we can use $(-m, -n)$ in any such comparison instead of (m, n)).
- (b) The result in part (a) is also compatible with the result of Exercise 2.4.5 and could have been derived from Exercise 2.4.5 instead of having been defined in isolation from Exercise 2.4.5.

2.4.7 Show that the integers \mathbf{Z} can be embedded into \mathbf{Q} , by the rule that $z \mapsto (z, 1)$ (or, equivalently at this point, $z \mapsto \frac{z}{1}$). Show further that this mapping preserves the algebraic operations on \mathbf{Z} .

2.4.8 Show that the definition for the order relation on \mathbf{Q} which was given in Exercise 2.4.5 above is compatible with the order relation defined previously upon \mathbf{Z} , under the natural embedding defined in problem 2.4.7.

We have given a construction of the set of rational numbers, starting from the set of integers, and we have shown that the set of rational numbers satisfies the axioms A1, A2, A3, and A4 and M1, M2, M3, M4, as does the set of integers, and in addition we have a new property of multiplication exists which was not previously present. It was proved in Problem 2.4.4. We will call it M5, and it says

M5. For each non-zero rational number q there is a corresponding rational number q^{-1} such that $(q)(q^{-1}) = 1$.

The axioms A1, A2, A3, A4, and M1, M2, M3, M4, M5 are the axioms of an algebraic structure called a **field**.

Exercises

2.4.9 Prove that the system described in Problem 2.3.1 is also an algebraic field.

2.4.10 Prove that the properties listed in Problem 2.2.12 remain true in the set of rational numbers.

When the proof of this problem is completed, we have shown that the set of rational numbers consists of an **ordered field**, a property which the system discussed in Problem 2.3.1 obviously can not share.

2.4.11 In addition to the properties shown in the previous problem, we have the following property:

If $x \in \mathbf{Q}$ and $y \in \mathbf{Q}$ with $x > y$ and $y > 0$, then $0 < 1/x < 1/y$

We have also shown that the integers may be naturally embedded into the rational numbers (Problem 2.4.7), and that the order properties of the rational numbers (introduced in Problems 2.4.5 and 2.4.8) naturally and compatibly extend the order properties of the integers which were given in I1, I2, I3, I4 and in O1, O2, O3, O4.

2.5 Inadequacy of the rational numbers

This goes way back.

When the theorem of geometry known nowadays as the Pythagorean Theorem was proved, Pythagoras and his disciples are said to have sacrificed a hundred oxen in celebration of the great discovery. Unfortunately for their world

view, which had as a doctrine the idea that all of nature was built on ratios of numbers, it was seen very quickly that there is no rational number which is equal to the square root of two. That number, of course, would express the length of the hypotenuse of an isosceles right triangle with both sides having the length one unit. In the language used by the Greek geometers, the legs and the hypotenuse of said triangle were seen to be “incommensurable” which is an equivalent statement. The historical accounts agree that this was discovered the very next day after the big result was proved and the oxen were slaughtered. After agreeing on this, the historical accounts also agree that the entire school of Pythagoras was sworn to secrecy about the second discovery. For, as said above, this discovery struck at the heart of their world view and if widely publicized would have made them look foolish. Accounts differ about what happened to the disciple who made the second discovery. Some of the accounts say that he was executed. Needless to say and whether the hapless discoverer was or was not executed, the truth of the matter could not be concealed for very long. It was rediscovered fairly soon.

Exercises

2.5.1 Show that there can be no rational number which is the square root of 2. (Hint: Start by assuming that there is such a number, and observe that if it does exist it can be expressed as a fraction in lowest terms. From this starting point, show that both the numerator and the denominator of the fraction are necessarily divisible by two, and thus the fraction, which was assumed to be in lowest terms, is not in lowest terms and can not be put in lowest terms. Therefore, the initial assumption, that the fraction exists, must be false.)

In view of the previous account and the previous exercise, there is at least one quantity which we would think ought to be expressible as some kind of a number, but it is seen that one must somehow step outside the set of rational numbers to do so. As often happens in the development of mathematics, one little example seems to indicate that there is at least potentially a multitude of similar examples. As far as the deficiencies of the rational number system are concerned, it is quite clear that the only way out of such problems is, if possible, to construct a larger number system in which such things can not happen. Most ideally, that new number system needs to incorporate the rational numbers inside it, too, in similar fashion to the way in which the integers can be seen as embedded in the rational numbers. In the next chapter, we carry out the project of constructing this larger system.

Chapter 3

Construction of the real numbers from the rational numbers

3.1 Introduction

The title of this chapter is “Construction of the real numbers from the rational numbers.” Many other treatments of the real numbers do not do a construction, but only assume that they exist and provide a set of descriptive axioms. What

we are about to do is to **construct** the set of real numbers, which is not the same thing. The construction is not terribly difficult, but it is of crucial importance while doing the construction that one does not use hidden assumptions about what will come out at the end before one has completed the construction. Most particularly, one can not assume that the real numbers exist before we construct them, nor that the rational numbers have certain properties which they do not have, in order to do the construction of the real numbers. For, it is the obvious shortcomings of the set of rational numbers which motivate the construction of the real numbers in the first place. We begin the chapter with some informal discussion of just what the problem is with the rational numbers, which motivates the desire to construct a larger set which will include them.

Informally, one might wish for the existence of a set containing the rational numbers, and additionally any other “number” which one can compute or can approximate by a rational number, within any previously given error tolerance. That is, starting with some kind of description of a property which defines any particular new “number,” it should be possible to obtain a rational number which approximates the desired new “number” within the previously given error tolerance.

Our main objective in this chapter will be to make logically rigorous the informal description which was alluded to in the previous paragraph and actually to do the construction not for just one number at a time but in a general way, which covers all possible eventualities. The goal is a logically rigorous formal construction. Using as the starting point the set of rational numbers, which we have constructed in the previous chapter, the procedure will culminate in the completed construction of a larger set, which we will call the set of real numbers. The new set of real numbers will have a very nice property, which will be called *completeness*. It will be clear that the set of rational numbers does not have

the property of completeness. Indeed, it will be seen that this lack is one of the main reasons why we take the trouble to construct the set of real numbers.

As it is clear that the construction of this new set of real numbers must needs be a sequence of logical steps, it will be essential to follow a logical sequence in doing those steps. It goes without saying that before we have actually finished the construction of the set of real numbers we cannot assume that its hoped-for nice properties related to completeness already hold for the set of rational numbers. Those properties do not hold for the set of rational numbers, and that is the reason why we are motivated to do the construction. To make a very down-to-earth comparison, let us suppose that we had an architect's plan for a very nice house. Among other things, the plan depicts a very nice roof on top of the very nice house. But we all know that the architect's plan, which envisions a roof on the house, will not protect our heads from rain before the house is built. The house must be built first, step by step, and only after the construction is completed will the actual roof be part of the house which is not yet built. Our situation here is very similar.

As an example of what we might try to do, let us consider a very simple way to obtain a rational number which provides an estimate for $\sqrt{2}$ within any previously prescribed degree of accuracy. Before getting started with any such procedure, though, let us remind ourselves that we have actually proved in Exercise 2.5.1 that there is no rational number whose square is exactly equal to 2. And we do not as yet know anything about any larger set. Thus, in obedience to the strict requirements of logic, we explicitly do not claim at this point to know that there really is some kind of number which really exists and which is exactly equal to what we have denoted by the symbol " $\sqrt{2}$." At this point we are merely using this terminology as a matter of convenience, nothing more. Saying this again for emphasis, the reason why we do not claim (yet) that there is such a thing as "the square root of two" is a very good reason. Namely, we do not yet

in fact know of any kinds of numbers other than rational numbers because we have not yet come up with any general procedure for constructing such numbers. Such knowledge will come later on, after we have learned systematically and step by step how to construct a suitable set of such numbers. Thus, in conformity with our need to proceed in a logical fashion, we cannot anticipate beforehand what that set is and what its properties are before we have actually built the set. Until then, we must keep working strictly within the set of rational numbers, carefully laying the groundwork for the construction of that larger set of real numbers, and collecting the tools which will be used for that construction.

Thus, again, there is no rational number whose square is exactly 2. Nevertheless, if one is given any prescribed tolerance for the error, it is possible to construct rational numbers whose squares approximate 2 within that previously prescribed degree of accuracy. The method which we will use is a simple but powerful method which is often called the **bisection method**. This method, or some small variation on it to fit the method into some similar context, is often useful.

To begin a search for a rational number which fits the desired outcome of having a square which is closer to two than any previously prescribed error tolerance, one might try an overshoot and undershoot procedure which can be repeated as many times as one wants or needs to do. To get the procedure started, one might take notice that $a_0 = 1$ is too small because $1^2 = 1 < 2$, and $b_0 = 2$ is too big because $2^2 = 4 > 2$. After this, a natural thing to do is to check what happens at the value $(a_0 + b_0)/2 = 3/2$, which is halfway between.

Now, in keeping with our intention to come up with a procedure which can be generalized and, if possible, repeated, we need at least abstractly to consider two possibilities:

The first possibility is that squaring $(a_0 + b_0)/2 = 3/2$ results in a number which is more than 2. If this is what has

happened, then we know that $a_0 = 1$ is too small and $3/2$ is too large. In fact, this is what has happened. Therefore, we let $a_1 = a_0 = 1$, and we choose $b_1 = 3/2$ because then $b_1^2 > 2$. The initial step is now completed, and we are ready to begin the next step in the procedure.

The second possibility is that squaring $(a_0 + b_0)/2$ could possibly have given us a result which is less than 2. This in fact is not what has happened when we started with $a_0 = 1$ and $b_0 = 2$, but it is clear that such could have happened if some other values of a_0 and b_0 had been used. If such had occurred, we would have been able to see by actual computation that $(a_0 + b_0)/2$ is too small and b_0 is too large. Therefore, under such circumstances we would have let $a_1 = (a_0 + b_0)/2$, and we would choose $b_1 = b_0$.

Abstractly speaking, there is a third possibility which we might need to consider in some similar problems, but not in this particular one. Namely, it could be possible that $(a_0 + b_0)/2$ is exactly equal to 2. However, as we already know, there is no rational number whose square is exactly equal to 2. Therefore, in the present context this simply can not happen. In the application of the method of bisection to some other problem, however, we might need to consider this third possibility, too. If such a thing were to occur, we would be finished with our search.

The initial step has now been completed, and we are ready to begin the next step in the procedure. For, we have two numbers a_1 and b_1 such that $a_1^2 < 2$ and $b_1^2 > 2$. The second step then begins by computing $(a_1 + b_1)/2$. Then after squaring and comparing to 2, we decide whether to pick $a_2 = a_1$ and $b_2 = (a_1 + b_1)/2$, or else to pick $a_2 = (a_1 + b_1)/2$ and $b_2 = b_1$.

In the actual event, of course, what has happened is that $(3/2)^2 = 9/4 > 2$, and the first possibility is the one which has actually occurred. Therefore, we are left with $a_1 = a_0 = 1$ and $b_1 = (a_0 + b_0)/2 = 3/2$. The next step is then to

consider the average $(a_1 + b_1)/2$, compare its square to 2, and use the comparison to choose two numbers a_2 and b_2 which satisfy $a_2^2 < 2$ and $b_2^2 > 2$. Then, one must keep going, doing this procedure as many times as are needed, in order to obtain the desired degree of accuracy.

In the exercise which is immediately below, you will be asked to prove that the procedure for which the initial steps are described above can be continued indefinitely, as many times as one might wish. Moreover, you will be asked to show that after n steps we have an interval $[a_n, b_n]$ which is of length 2^{-n} . And moreover that $a_n^2 < 2$ and that $b_n^2 > 2$, and that both $b_n - a_n$ and $b_n^2 - a_n^2$ tend to zero, too, as we do more steps. Under such circumstances, we are strongly tempted to believe that inside of all of these progressively smaller intervals $[a_n, b_n]$ there is some kind of a number-like entity which is “exactly equal” to something which we would like to call $\sqrt{2}$.

Exercises

3.1.1 Describe as an algorithm the procedure which is informally described above, which can be used to estimate the square root of 2. Take the point of view that you are going to write a repetitive loop which can be used, in the spirit of mathematical induction, to start after n steps have been carried out, and then to arrive at the setup for the next step, that is, the $(n + 1)$ st step. Specifically:

Assume that you have started with numbers $a_0 = 1$ and $b_0 = 2$. Then define $c_0 = (a_0 + b_0)/2$. Describe the procedure for choosing whether in the next step we should choose $a_1 = a_0$ and $b_1 = c_0$, or whether we should choose $a_1 = c_0$ and $b_1 = b_0$. Now, assume that you have found numbers a_n and b_n such that $a_n^2 < 2$ and $b_n^2 > 2$. Show how

to find the numbers a_{n+1} and b_{n+1} , so that one is then ready to carry out the next step in the procedure.

At the risk of belaboring the obvious, the above argument does **not** compute a rational number whose square is “exactly equal” to 2. For, as we have already seen in Exercise 2.5.1, there can be no such rational number. Nevertheless, if it is repeated a sufficient number of times the above procedure will certainly produce any desired degree of accuracy even though at every step we are staying inside the set of rational numbers. In a way, the question of whether or not there is any number of any kind whose square is “exactly equal” to 2 can now be answered by saying, “Well, perhaps that depends upon what you mean by saying that such a number exists whose square is exactly equal to 2. If you are willing to believe that a construction of a process which gives us arbitrary accuracy is a proof for existence, then the question is answered.” In fact, something like this is what we are going to do. But one needs to be really careful.

One of our potential problems is that the above-described method of bisection is far from being the only method for obtaining a rational approximation for “ $\sqrt{2}$ ” which satisfies the requirement for producing arbitrary, pre-prescribed accuracy when repeated a sufficient number of times. There are lots of other methods, too. The numerical values which come out of the successive approximation steps in those other methods might or might not be identical to those which come out of the steps in the bisection algorithm described above. For that matter, the steps in the bisection algorithm itself would give us different numbers at each step in it if we had started out with different choices for a_0 and b_0 . We could have started out, for example, with $a_0 = 0$ and $b_0 = 29/20$ and then the numbers coming out would not be the same. This is not even to mention that very different methods can be used, too. For example, there is a very old procedure which vaguely resembles long division but is even worse than the worst problem in long division that a grade-school student could imagine in a nightmare. Though quite a bit more complicated than the bisection algorithm,

that method similarly provides arbitrary accuracy in the computation of a square root. Many who studied arithmetic in the days before the calculator was invented were forced to learn that procedure. The only thing that is nice about it is that in the successive steps the numbers which come out are the actual decimals which appear. One obtains in the first step 1. Then in the second step 1.4, then in the third step 1.41, and so on. All by hand calculation, believe it or not. The upshot is that, if we would adopt the idea that the quantity which we wish to pin down is to be identified, somehow, with a procedure for approximating it, we would have to ensure (among other things) that all conceivable procedures for approximating the square root of 2 would actually have to approximate the same thing. This and other difficulties need to be solved, and not only when we are seeking the square root of 2 but in general and for all occasions where similar questions might arise.

Much of the rest of this chapter will deal with laying out the necessary groundwork for the construction of the real number system. In view of the sort of thing we did in the actual example we just investigated, which prescribed a sequential approximation method for what we will soon learn is a certain, specific real number, the first step is to start with some discussion of sequences. We will do that in the next section of this chapter.

Before we get started with the rest of the chapter, though, we need to remind ourselves of a few things, which bear repeating. What we are doing is to **construct** a number system which expands the system of rational numbers in certain appropriate ways. We will need to develop a method for the construction, then to carry out the construction, and then, finally, to show that the construction has done what we wanted. We must not and can not adopt the attitude that we “have” this new number system or that we “know” what it is until we have finished this construction. Moreover, we must adopt a method which will do the job and then stick to it until the task is shown to be completed. As a couple of

examples of what this statement means, consider the following:

- “The real numbers consist of all the numbers between minus infinity and infinity.” is quite unsuitable as a definition of the set of real numbers. One very good reason why this definition falls short is that we have not previously introduced the concepts of “infinity” and “minus infinity.” Thus, we do not know what those words mean. Even if we did know what those words mean, or think that we do, such a “definition” also bypasses and therefore leaves completely open the important question of what kind of numbers these are. Thus, such a “definition” clearly begs the question of what the real numbers actually are, by assuming without saying so that we already know what they are. The rational numbers, too, are between these two extremes. In addition to the rational numbers, are there any other numbers between these two extremes? Yes, or no? Why, or why not? In other words such a “definition” may be all very fine for helping some people sort of understand what a real number is, or some might think so. But it is a statement which actually has little if any logical meaning, is therefore not a statement with actual mathematical content, and is therefore totally unhelpful for our purposes.
- There are other perfectly legitimate methods for constructing the system of real numbers besides the approach we are taking here. Some of those methods seem easy, natural, and intuitive from some standpoint, but complications arise when one gets into the details. One of them is called the method of Dedekind cuts. It is the topic of a problem later on in this chapter. It seems conceptually simple to start with, but it gets really nasty as soon as one would wish to define algebraic operations on real numbers. For this reason among others, we do not use that method here for defining the set of real numbers.

- Another of the alternative approaches is to define the set of real numbers as equal to the set of all (possibly infinite) decimal expansions (leaving aside the rather serious question of what one might mean by an “infinite” decimal expansion). As we will learn later on, this description actually is in a sense true, and therefore this definition could possibly be tightened up into a valid approach to the definition of the set of real numbers. There are some natural features to recommend it, too. For example, it is a description which can seem intuitive even to students who have just gotten through learning something about decimals. That is not a bad thing at all. But on closer inspection there are several basic problems. The first problem is that some numbers actually have not one but two distinct but equal decimal expansions (Which numbers might those be? Do you happen to know?). Another problem is that, in fact, decimal expansions are sequences of a certain particular form. So why shouldn’t one start out by learning about sequences? It may after all end up being easier. Another problem is, how is one supposed to do arithmetic with nonterminating decimals, anyway? How ought such a thing actually work? After all, we are only human, and we cannot add together an infinite number of digits because we are literally unable to do any such thing; each such step requires some finite length of time and we do not live forever. Even worse, what happens when one needs to do carrying during addition, or borrowing during subtraction? And still worse, how might one propose a rule or method for the multiplication or the division of one nonterminating decimal by another? Such questions about the mechanics of the algebraic operations are all very good and legitimate questions. And, so long as one has defined the real numbers as the set of all decimal expansions and has not started with a more flexible definition, mere hand-waving will not magically make these questions go away. These issues are the main reasons why the decimal representation of the real numbers will be discussed in Chapter 8, where we will **prove** that the

set of real numbers is the same as the set of all decimal expansions. The material presented in Chapter 8 will logically depend upon the contents of the current chapter, not the other way around. It goes without saying that we cannot assume in this chapter some results which will be done in the subsequent Chapter 8 in order to “prove” the results in this chapter, and then turn right around and use what we claim to have “proved” in this chapter in order to get the results which are presented in Chapter 8. To do that kind of thing is to indulge in circular reasoning, which consists of nothing more nor less than to assume what one is going to prove in order to prove it. We are all supposed to know better than to do such a thing.

3.2 Sequences

Formally, a **sequence** of elements from any set S is a function from the natural numbers with range in S . The sequence is often represented by $\{s_n\}_{n=1}^{\infty}$, or more briefly, if the context makes the matter clear, as just $\{s_n\}$. Of course, more informally, a sequence may “start” at 0 or at some other integer, even a negative one. A **finite sequence** is a function defined on the set $\{1, \dots, n\}$ for some natural number n with its range in the set S . Such a finite sequence, therefore, can be represented as $\{s_n\}_{n=1}^n$. The individual elements s_n from S are called the **terms** of the sequence.

Remark:

We are using the grouping symbols “{” and “}” to denote sequences as well as sets. The difference between sequences and sets should be clear from context in the future. But the difference between a set and a sequence is that for a set the

order in which the elements are listed does not matter, while for a sequence the order of the terms does matter. Also, terms in a sequence can repeat because it is their position which matters, not their value. This statement applies both to infinite sequences and finite sequences. We will look at three examples:

Example 1: Consider the finite sequence $\{s_n\}_{n=1}^3$ defined by $s_1 = 1$ and $s_2 = 2$ and $s_3 = 3$. Then the sequence is the **ordered** list of the numbers 1, 2, and 3, in that order. This is not the same thing as the **set** $\{1, 2, 3\}$. The set $\{1, 2, 3\}$ is equal to the set $\{3, 1, 2\}$. But the reordering of the elements has resulted in a different sequence from the one we started with. Namely, if the second of these two sets were to be viewed as a sequence instead of merely as a set, then its first element is 3 and its second element is 1 and its third element is 2.

Example 2: A sequence with three terms in it could be defined by $\{s_n\}_{n=1}^3$, with $s_1 = 1$ and $s_2 = 1$ and $s_3 = 1$. The set of numbers which are listed in the sequence is then, obviously, the set $\{1\}$ which contains only one element. A similar remark would hold if we were to define $s_n = 1$ for all $n \in \mathbf{N}$.

Example 3: An infinite sequence can be defined by $s_n = (-1)^n$. Then the n th term of the resulting sequence is $s_n = -1$ whenever n is odd, and it is $s_n = 1$ whenever n is even. The values of all of the terms in this sequence lie in the set $\{-1, 1\}$ which is a set containing exactly two elements.

A **subsequence** of a sequence defined upon the set S may be formally viewed as a composition of a function from the natural numbers to S (a sequence) with a (strictly) increasing function from the natural numbers to the natural

numbers. The effect is to “skip” some of the terms of the original sequence. That is, if we are given a sequence $\{s_n\}_{n=1}^{\infty}$, we may write a subsequence of the given sequence as $\{s_{n_k}\}_{k=1}^{\infty}$ or simply as $\{s_{n_k}\}$. When we do this, then there is the obvious requirement that $n_{k+1} > n_k$ must hold true for every k . As a simple example of a sequence and a subsequence, we might consider sequence $\{s_n\}_{n=1}^{\infty}$ to be defined by the rule $s_n = 1/n$. One could now define a subsequence by (for example) specifying that $n_k = 2k$, and then $s_{n_k} = 1/2k$. The resulting subsequence is then a sequence of fractions with even denominators only. As another example of a subsequence of the same original sequence, one might let n_k be the k th prime number, listing the prime numbers in increasing order. Then the subsequence which would result would list the successive terms in the original sequence for which the denominator is a prime number.

3.3 Sequences of rational numbers

Now, after the above general definition of what a sequence is, let us assume that we are working in particular with sequences of rational numbers.

We define the **limit** (if it exists) of a sequence $\{s_n\}$ of rational numbers as a rational number s for which:

Given $\epsilon > 0$, there exists a number N , such that, if $n \geq N$, then $|s_n - s| < \epsilon$.

If s is the limit of the sequence $\{s_n\}$, we also write

$$\lim_{n \rightarrow \infty} s_n = s.$$

Remark 1 on the definition of the limit of a rational sequence:

It should be kept mind that we have given no definition to any larger set of numbers which contains the rational numbers. For that reason, the above definition lives in a restricted context. The number ϵ must perforce be rational. The numbers s_n must be rational. **The number s must be rational, too.** For, we have given no formal definition for more than this. Emphasizing: the limit s needs to be rational. We can not claim at this point that some given sequence of rational numbers “has a limit which is a real number” because at this point we have given no definition of what that means! Indeed, we what we are about to do is that we are going to use certain carefully derived properties of sequences of rational numbers in order to *construct* the set of real numbers.

Remark 2 on the definition of the limit of a rational sequence:

The limit defined above can not be ∞ or $-\infty$, either, because, whatever these symbols might represent (and we have as yet given them no formal definition!), neither of the two is a rational number. Thus, for now either the limit of a given sequence of rational numbers is a rational number, or else the given sequence has no limit. As to the use of the symbol “ ∞ ” in the expression “ $\lim_{n \rightarrow \infty}$ ” though, what it clearly indicates is that there is no upper bound upon the value of n . From this point of view, it does not represent any kind of quantity at all, but rather a denial of a limitation. The discussion of what it might mean to say that $\lim_{n \rightarrow \infty} s_n = \infty$ is deferred to the end of the next chapter, since its presence here would only take us off on a tangent.

A closely related concept to that of the limit of a sequence is that of a **Cauchy sequence**. The sequence of rational numbers $\{s_n\}$ **has the Cauchy property**, or is a **Cauchy sequence** provided that:

Given $\epsilon > 0$, there exists a number N such that if $m \geq N$ and $n \geq N$, then $|s_m - s_n| < \epsilon$.

Note that again we are operating under restrictions. The numbers s_n and the number ϵ must still be rational. We are working completely within the context of the rational numbers. We have not gone outside that context. Indeed, *after* we have completed the construction of the real numbers we will then need to give two very similar-looking definitions again in that new, expanded context. When we will have completed the construction of the set of real numbers, a construction which still lies in our future, the definition of “has a limit” will then superficially look the same as the definition above. And the definition of “is a Cauchy sequence” will superficially appear to be the same, too. But those two definitions which we will encounter then will not be the same as the two definitions above because all of the quantities in them which have to be rational numbers here will be allowed to take on real-number values there. And there are plenty of real numbers which are not rational numbers.

Exercises

3.3.1 In the introductory section of this chapter, an informal description was given for the construction of progressively better and better rational approximations to what we would like to denote by the symbol $\sqrt{2}$. In Exercise 3.1.1 you were asked to provide an actual algorithm for carrying out that construction. You were asked there to show how to construct two sequences of rational numbers which could be called $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$. Both of these sequences appear to be “trying to converge to $\sqrt{2}$,” the first one from below and the second one from above. You will again need to use mathematical induction in order to show that these two sequences have the following properties:

- The sequence $\{a_n\}$ is **non-decreasing**. That is, $a_0 \leq a_1 \leq \dots$
- $(a_n)^2 < 2$ for all $n = 0, 1, 2, \dots$
- The sequence $\{b_n\}$ is **non-increasing**. That is, $b_0 \geq b_1 \geq \dots$
- $(b_n)^2 > 2$ for all $n = 0, 1, 2, \dots$
- $b_n - a_n = 2^{-n}$ for all $n = 0, 1, 2, \dots$

Finally, show based upon the above items that both of these sequences of rational numbers, the sequence $\{a_n\}$ and the sequence $\{b_n\}$ are Cauchy sequences. This means that you should show each of them satisfies the definition of a Cauchy sequence of rational numbers which is given above. You should not go outside the set of rational numbers (for example, by attempting to invoke the existence of a limit for the two sequences) because we have not yet constructed the set of real numbers. We know that neither sequence has a limit which is a rational number, and therefore we cannot logically assume in the present context that either one of the two sequences actually has a limit.

Very important remark: In view of the previous problem, it is not valid to assume that every Cauchy sequence of rational numbers has a limit, so long as the limit must also be a rational number. Furthermore, the only kind of numbers that we formally know about, at this point, is rational numbers. And consequently the limit must be a rational number, too, if there is a limit. Moreover, the previous problem actually constructs not one but two Cauchy sequences of rational

numbers, neither of which has a limit which is a rational number. Making an obvious generalization, we see that any given, particular sequence of rational numbers – even if it is a Cauchy sequence – may very well not have a limit which is a rational number. What we are about to do is to **use** this fact in order to **construct** a larger set which we will call the set of real numbers. It will turn out to be true, in the end, that every Cauchy sequence of **real** numbers will have a limit which is also a **real** number. And furthermore it will be seen that the rational numbers can be identified with a proper subset of the real numbers. When this construction has been done and not before the construction has been done, we will then know that every Cauchy sequence of rational numbers must have a limit, too, inside of the larger set of real numbers, even if it does not have a limit which is inside the set of rational numbers. **But we cannot assume such a thing before we have completed the actual construction of the real numbers. For, before we have done the construction of the set of real numbers we can not claim to know what the real numbers are, nor can we claim to know what their properties are.**

Exercises

3.3.2 Formulate precisely the negation of the statement $\lim_{n \rightarrow \infty} s_n = L$. If you are not sure how to do this, then please see the discussion at the end of Chapter 1 about how to formulate the negation of a statement which contains quantifiers. Notice also that the statement “It is not true that $\lim_{n \rightarrow \infty} s_n = L$ ” does not say anything at all about whether or not the sequence $\{s_n\}$ has a limit, or not. Therefore, your answer to this problem ought to contain no such implication, either.

3.3.3 Formulate precisely the negation of the statement, “The sequence $\{s_n\}$ is a Cauchy sequence.” If you are not sure how to do this, then please see the discussion at the end of Chapter 1 about how to formulate the negation of a statement which contains quantifiers.

3.3.4 If a sequence of rational numbers has a limit, then it does not have another, distinct limit.

3.3.5 Given two sequences of rational numbers $\{a_n\}$ and $\{b_n\}$, we can define

–their sum as the sequence $\{S_n\}$, where $S_n = a_n + b_n$.

–their product as the sequence $\{P_n\}$, where $P_n = a_n b_n$.

–their quotient as the sequence $\{Q_n\}$, where $Q_n = a_n/b_n$, provided that none of the b_n are zero.

Given a rational number q and a sequence $\{a_n\}$, we can also construct the sequence $\{qa_n\}$.

Now, assume that $\{a_n\}$ and $\{b_n\}$ have limits respectively equal to the rational number a and to the rational number b . Show that, in all of these cases, the limit of the new sequence is what should be expected. However, an additional condition must hold before the quotient sequence $\{Q_n\}$ can be said to have a limit. What is that condition?

3.3.6 Who was the mathematician named Cauchy, and why does the concept of a Cauchy sequence carry his name? Do some reading, and report to the class about this topic.

3.3.7 Show that every Cauchy sequence of rational numbers is bounded. Note that in order to prove this we cannot assume that every Cauchy sequence of rational numbers has a limit. For, in view of Exercise 3.3.1, a Cauchy sequence of rational numbers may not have a limit which is a rational number. Therefore, do not assume that a

Cauchy sequence of rational numbers has a limit in order to do this problem. Do not use anything other than the definition of a Cauchy sequence which is stated above this set of problems.

3.3.8 If $\{s_n\}$ is a Cauchy sequence of rational numbers and $\{s_n\}$ does not have 0 as its limit, then there is $b > 0$ and there is some N (depending upon b) such that for all $n \geq N$ we have either $s_n > b$, or $s_n < -b$. In either of these two possible cases there can be no more than a finite number of terms whose value is zero. Note that we cannot assume here that every Cauchy sequence of rational numbers has a limit. In view of Exercise 3.3.1, a Cauchy sequence of rational numbers may not have a limit which is a rational number, and we do not know yet about any other kinds of numbers.

Hint: A good start toward doing this problem is to do Exercise 3.3.2 first if you have not yet done it. Then proceed by combining the output of Exercise 3.3.2 with the statement that the sequence is a Cauchy sequence.

What Problem 3.3.8 is saying is, either the values s_n are “ultimately positive” or they are “ultimately negative” and in either situation there is some integer N for which the terms are bounded away from zero for all $n \geq N$. Note that the requirements here are only that $\{s_n\}$ is a Cauchy sequence, and that the sequence does not have 0 as its limit. This language does not say that the Cauchy sequence $\{s_n\}$ has a limit, because at this point our “universe of discussion” consists of the set of rational numbers. We do not know, as yet, what a real number is. And there are lots of Cauchy sequences of rational numbers which have no limit among the rational numbers. See Exercise 3.3.1 and the discussion which follows it.

Exercises

- 3.3.9** As we have done in exercise 3.3.5, let us start with two sequences of rational numbers $\{a_n\}$ and $\{b_n\}$, defining
- their sum as the sequence $\{S_n\}$, where $S_n = a_n + b_n$.
 - their product as the sequence $\{P_n\}$, where $P_n = a_nb_n$.
 - their quotient as the sequence $\{Q_n\}$, where $Q_n = a_n/b_n$, provided that none of the b_n are zero.

Show that, if the sequences $\{a_n\}$ and $\{b_n\}$ are assumed to be Cauchy sequences, then so are the sequences $\{S_n\}$, $\{P_n\}$, and (under reasonable conditions which you should determine) $\{Q_n\}$ (Exercise 3.3.7 is useful here in dealing with the sequence $\{P_n\}$, and Exercise 3.3.8 is needed in dealing with the sequence $\{Q_n\}$). Note that we cannot assume here that any of the sequences $\{a_n\}$, $\{b_n\}$, $\{S_n\}$, $\{P_n\}$, or $\{Q_n\}$ have limits. The stated exercise does not say that they do, and until we have actually completed our construction of the real numbers we can not claim to know any such thing. And, what is more, the results of this Exercise will be used in our near-future construction of the real numbers. Therefore, please do not assume that any of the sequences in this Exercise have limits. Just stick to using the definition of a Cauchy sequence. To do otherwise is to descend rapidly into circular logic, which has no place here.

3.3.10 Show that every convergent (that is, possessing a limit) sequence of rational numbers is a Cauchy sequence.

3.3.11 If $\{s_n\}$ is a Cauchy sequence of rational numbers, then every subsequence of it is a Cauchy sequence. Again,

please note that we cannot assume here that every Cauchy sequence of rational numbers has a limit. In view of Exercise 3.3.1, a Cauchy sequence of rational numbers may not have a limit which is a rational number. And we will use this result in order to develop the set of real numbers, which has not yet been introduced or defined.

3.3.12 The sequence of rational numbers $\{s_n\}$ has a limit s if and only if every subsequence of $\{s_n\}$ has limit s , too.

3.4 The real numbers as Cauchy sequences of rational numbers

Based upon the groundwork which has been carefully laid out in the previous section, we are now in a position to introduce a formal definition for the set of real numbers. This new set of real numbers will often be denoted by the symbol \mathbf{R} .

As we did while constructing the rational numbers, we begin by defining an equivalence relation. This time, the equivalence relation is defined upon the set of all Cauchy sequences of rational numbers. The equivalence relation can be defined by saying that, for any two sequences $\{a_n\}$ and $\{b_n\}$, we have $\{a_n\} \sim \{b_n\}$ if and only if

$$\lim_{n \rightarrow \infty} (a_n - b_n) = 0.$$

Exercises

3.4.1 The advertised equivalence relation is indeed an equivalence relation.

We now can define the set of real numbers as the set of all equivalence classes of Cauchy sequences of rational numbers, using the equivalence relation described immediately above.

To extend the algebraic and order relations of the rational numbers to the set of real numbers, let us suppose that r and s are two real numbers. This means that, just as a rational number actually was an equivalence class containing many individual specimens, the same is true about r and s . For each of r and s there corresponds a sequence of rational numbers, $\{r_n\}$ a member of the equivalence class for r and $\{s_n\}$ a member of the equivalence class for s . Then:

- The equivalence class of 0 is the set of all the Cauchy sequences of rational numbers which have limit equal to 0
- The equivalence class of 1 is the set of all the Cauchy sequences of rational numbers which have limit equal to 1.
- The sum $r + s$ may be taken to be the equivalence class of the sum of any sequences $\{r_n\}$ from the equivalence class of r and $\{s_n\}$ from the equivalence class of s .
- The product of r and s is the equivalence class of the product of any two sequences $\{r_n\}$ from the equivalence class of r and $\{s_n\}$ from the equivalence class for s .

- If s is not zero, then no sequence $\{s_n\}$ which represents it has limit 0. If the sequence which has been chosen to represent s also has no terms in it which are equal to 0, then it is not a problem to construct the quotient sequence whose respective terms are $\frac{1}{s_n}$. If, however, some of the terms in $\{s_n\}$ do happen to be zero, then there are at most a finite number of such terms (see Exercise 3.3.8, above). For those terms, which are finite in number, we can replace $\frac{1}{s_n}$ (which otherwise does not exist if $s_n = 0$) by anything we like. For example, we could handle this by arbitrarily declaring the corresponding term in the quotient sequence to be equal to 1. For, in no event are there more than a finite number of such terms. By this means, and thus modifying the rule for the construction of a quotient sequence for a sequence which represents s , we can define s^{-1} to be the equivalence class of any sequence such that $s(s^{-1}) = 1$, where 1 also signifies an equivalence class of sequences, too. Namely, it is the equivalence class of all those sequences of rational numbers which have the limit 1.

Exercises

3.4.2 The addition of real numbers which is proposed above is well-defined.

3.4.3 The multiplication of real numbers which is proposed above is well-defined.

3.4.4 The results in this problem have already been foreshadowed above. Please work out any details which need to be covered. If the real number s is not equal to zero, then there is a representative $\{s_n\}$ from the equivalence class which defines s and which satisfies the condition that $s_n \neq 0$ for all n . More than this, any sequence $\{s_n\}$

representing s has the property that $\{n : s_n = 0\}$ is a finite set, and therefore the sequence $\{s_n\}$ is always equivalent to another sequence in which there are no zero terms.

3.4.5 For any sequence $\{s_n\}$ which represents a real number $s \neq 0$, we define the n th term of a sequence for s^{-1} by specifying that it is $1/s_n$ if $s_n \neq 0$ and it is 1 in the situation that $s_n = 0$. With this definition for division, the real number s^{-1} is well-defined.

3.4.6 Show that the set of real numbers, with the algebraic operations which we have defined upon it, is an algebraic field (cf. Problem 2.4.4).

3.4.7 Show that the rational numbers can be naturally embedded into the real numbers, with algebraic operations preserved.

We now wish to extend the order relations defined upon the rational numbers to the real numbers. To this end, we declare a real number r to be positive, provided that there is a Cauchy sequence $\{r_n\}$ of rational numbers representing r , and the sequence $\{r_n\}$ is “ultimately positive.” That is, there is N such that, if $n \geq N$, then $r_n > \delta$, for some rational number $\delta > 0$. We declare the real number r to be negative if $-r$ is positive. Furthermore, we say for any two rational numbers r_1 and r_2 , that $r_2 > r_1$ provided that $r_2 - r_1$ is positive, and we write $r_2 - r_1 > 0$.

Remark: The following exercises are to be done using the construction of the real numbers which we have performed here. Most particularly, they are not to be done by appeal to the often-used description of the real numbers as all of

those numbers which can be represented by a decimal expansion. Although that statement happens to be true, we have **not** used it, nor have we at this point shown that it is correct. It will be seen later on that a decimal expansion of a real number comprises a representation of that number by a certain, specific kind of sequence generated by an infinite series. Infinite series will be discussed in the next chapter of this text. To establish the equivalence of the real numbers to the set of all decimal expansions will thus depend both upon the present chapter and the next chapter after this one, and will be presented later on. This present material precedes that discussion and is fact a logical prerequisite for that future discussion. Therefore, please stick to using things which we have already done when working the problems currently presented here, not results which we will do in the future, results which will use the material which is here as their foundation.

Exercises

3.4.8 (cf. Problem 3.4.6.) There is a natural embedding of the rational numbers into the real numbers which preserves the ordering of the rational numbers, as well as the algebraic operations upon them. Hint: A rational number can be represented as a constant sequence whose terms are all equal to the given rational number.

3.4.9 Between zero and any positive real number, there is a rational number. (Note that what is actually meant here is, between zero and any positive real number, there is a rational number, placed in its location by the natural embedding of the rational numbers into the real numbers.)

3.4.10 Between zero and any positive rational number, there is an irrational (meaning, other than rational) real number. Indeed, between any two rational numbers there is an irrational number and also, for that matter, a rational number.

3.4.11 Every real number is either positive, negative, or zero.

3.4.12 For any two real numbers r_1 and r_2 , either $r_1 > r_2$ or $r_2 > r_1$, or else $r_1 = r_2$.

3.4.13 In the sense of the order relation which is defined upon the set of real numbers

- There is not a largest real number.
- There is not a largest rational number.
- There is not a largest integer.
- For each real number $r > 0$ there is an integer n such that $\frac{1}{n} < r$.

The last of the above items is often referred to as the “Archimedean” property of the set of real numbers.

3.4.14 Between any two irrational numbers there is a rational number.

The real numbers are said to be the *completion* of the rational numbers. What this exactly means will be discussed in more detail in the next section. Here, let us only note that to give the full justification of this statement, we must first

of all extend the definition of the limit of a sequence and the definition of a Cauchy sequence to real numbers. To do this, we need only to drop the previous restrictions that ϵ had to be rational, that the terms of the sequence itself had to be rational, and that the limit had to be rational. Other than this, the definition remains at this point identical to the one which we have already been using. Here follow the formal statements.

We define the limit (if it exists) of a sequence $\{s_n\}$ of real numbers as a real number s for which:

Given $\epsilon > 0$, there exists a number N , such that, if $n \geq N$, then $|s_n - s| < \epsilon$.

If s is the limit of the sequence $\{s_n\}$, we also write

$$\lim_{n \rightarrow \infty} s_n = s.$$

To re-emphasize, the above “new” definition for the limit of a sequence of real numbers essentially repeats the same words and symbols which were used in the definition for the limit of a sequence of rational numbers. The difference is that here all the symbols which previously had to be rational numbers are now permitted to be real numbers. In other words, we now allow ϵ to be a real number, the terms in the sequence to be real numbers, and the limit itself to be a real number. It should also be clear that, in this new context, every Cauchy sequence of rational numbers now has a limit, equal to that real number which the given sequence of rational numbers represents.

The restatement of the definition of a Cauchy sequence of real numbers is, after this, perhaps not surprising. The sequence $\{s_n\}$ of real numbers has the Cauchy property, provided that:

Given $\epsilon > 0$, there exists a number N such that if $m \geq N$ and $n \geq N$, then $|s_m - s_n| < \epsilon$.

In view of what has come before, we can essentially repeat the contents of a previous set of exercises in this new context, too. The proofs of these results will not differ in any essential details, either, from those for the very similar results for rational numbers.

Exercises

3.4.15 If a sequence of real numbers has a limit, then it does not have another, distinct limit.

3.4.16 Given two sequences of real numbers $\{a_n\}$ and $\{b_n\}$ which have limits respectively equal to a and to b , we can define

–their sum as the sequence $\{S_n\}$, where $S_n = a_n + b_n$.

–their product as the sequence $\{P_n\}$, where $P_n = a_n b_n$.

–their quotient as the sequence $\{Q_n\}$, where $Q_n = a_n/b_n$, provided that none of the b_n are zero.

Given a real number r and a sequence $\{a_n\}$, we can also construct the sequence $\{ra_n\}$.

Show that, in all of these cases, the limit of the new sequence is what should be expected, if the original sequences $\{a_n\}$ and $\{b_n\}$ have limits. However, what additional condition, if any, must hold before the quotient sequence $\{Q_n\}$ can be said to have a limit?

3.4.17 Show that every Cauchy sequence of real numbers is bounded. Note that we are almost ready here to assert that every Cauchy sequence of real numbers has a limit which is a real number, but not quite. Since we have not yet actually proved that, please do not assume it in order to prove this result.

3.4.18 Taking into consideration the definitions in problem 3.4.16, show that, if the sequences $\{a_n\}$ and $\{b_n\}$ are assumed to be Cauchy sequences of real numbers, then so are the sequences $\{S_n\}$, $\{P_n\}$, and (under reasonable conditions which you should determine) $\{Q_n\}$ (Problem 3.4.17 is useful here). Note that we are almost ready here to assert that every Cauchy sequence of real numbers has a limit which is a real number, but not quite. Since we have not yet actually proved that, please do not assume such a thing in order to prove this result.

3.4.19 Show that every convergent (that is, possessing a limit) sequence of real numbers is a Cauchy sequence.

3.4.20 If $\{s_n\}$ is a Cauchy sequence of real numbers, then every subsequence of it is a Cauchy sequence; if $\{s_n\}$ has a limit s , then every subsequence of $\{s_n\}$ has limit s , too. Note that we are almost ready here to assert that every Cauchy sequence of real numbers has a limit which is a real number, but not quite. Since we have not yet actually proved that, please do not assume it in order to prove this result.

Now, as the final step of the construction of the real numbers, we show that the set of real numbers is indeed *complete*. We use the word “complete” in this context because we confronted a situation with the rational numbers in which things were missing, things which we probably believed all along ought to be present. We have provided logical steps by which we were able to supply those things which seemed to be missing. So, what we have done is a **completion** of the rational numbers and have obtained a bigger set of numbers. So now the situation needs to be investigated, whether there is anything which seems to be missing in this new set of real numbers, or not? What will occur if we “complete the completion”? More exactly, if we start with the set of real numbers and do what we already did when starting with the

rational numbers, are we going to get some still bigger new set, or will we get the real numbers back again, with nothing new added? The answer to this question is that, if we follow the same procedures as before, then in fact we do not get anything new. That is what we mean by saying that the set of real numbers is complete. However, it should be clear that this statement also needs a proof.

Therefore, the final result which needs to be shown is that every Cauchy sequence of *real* numbers converges to a *real* number and not to some new entity. Because it addresses this very question, the following problem is thus essential to all that comes after it in this Chapter.

Exercises

3.4.21 Let $\{r_n\}$ be a Cauchy sequence of real numbers. Then $\{r_n\}$ has a limit which is a real number. (Note: The proof of this statement has been preceded by a certain method for the construction of the real numbers. Namely, every real number can be represented by a Cauchy sequence of rational numbers, and thus each of the real numbers r_n which is a term in the sequence $\{r_n\}$ has such a representation, too. Thus, to complete this problem it is a requirement to use the construction of the real numbers which we have followed. You need to construct a Cauchy sequence $\{q_n\}$ of *rational* numbers which is equivalent to the given sequence $\{r_n\}$, in the sense that $\lim_{n \rightarrow \infty} (q_n - r_n) = 0$. Then from the fact that it is possible to do this, explain why it follows that to find the limit of the given Cauchy sequence $\{r_n\}$ of *real* numbers, one does not need to go outside of the newly constructed system of real numbers to do another job of completion). Note that in this problem we actually are supposed to prove something which we

have heretofore scrupulously avoided saying, namely that every Cauchy sequence of real numbers actually does have a limit which is a real number. From now on, we are free to use that statement.

3.5 Equivalent formulations of completeness in \mathbf{R}

We have taken the approach of constructing the real numbers, by starting with the rational numbers and studying Cauchy sequences of rational numbers. Then what we have shown in Problem 3.4.21 is the following:

Statement of Completeness of \mathbf{R} Every Cauchy sequence of real numbers converges to a real number.

It is precisely this statement, that every Cauchy sequence of real numbers converges to a real number, which is what we mean here by saying that the real numbers are complete. The statement that the real numbers are the completion of the rational numbers means that the rational numbers are embedded in the real numbers, that the real numbers are complete, and that any proper subset of the real numbers which contains the rational numbers will fail to be complete. It should be fairly obvious at this point to see that the Statement of Completeness must be true, since **what we have exactly done is to use the set of Cauchy sequences of rational numbers to construct the real numbers**.

Now, the main point of the remainder of this section is to take notice that the Statement of Completeness, as stated in the form above, implies several other statements. In fact, it will be seen later on that each one of the statements which are found in the upcoming problem set is in fact logically equivalent to the Statement of Completeness, and thus all of these statements are logically equivalent to each other. The Exercises explore some of the relationships between

some of the formulations which will be seen in the future to be equivalent to the Statement of Completeness. However, before actually stating the Exercises we need to define two new terms which are used in them:

The **supremum** (if any) of a set A is written $\sup A$ and is defined to be that (real) number s satisfying the two conditions

- (i) For all $a \in A$, $a \leq s$
- (ii) For each $t < s$, there exists $a \in A$ such that $t < a$.

Similarly the **infimum** (if any) of a set A is written $\inf A$ and is defined to be that (real) number i satisfying the two conditions

- (i) For all $a \in A$, $i \leq a$
- (ii) For each $j > i$, there exists $a \in A$ such that $j > a$.

It should also be noticed that no non-empty set can have two suprema, nor two infima (Remark about language: the words supremum and infimum are in fact Latin nouns, and their properly constructed plurals are indeed as given here). It should also be noted, of course, that a set which is not bounded above cannot have a supremum, also that a set which is not bounded below cannot have an infimum. Also, it should be noted that the empty set has no supremum, because part (ii) of the definition would require the empty set to be non-empty. And for similar reasons the empty set cannot

have an infimum. Nevertheless, the empty set is, in a trivial sense, bounded both above and below by any arbitrary real number.

Remark:

Perhaps it is also appropriate to notice that the supremum of a set S may have been referred to in some previous course as the “least upper bound” of the set, sometimes abbreviated as “lub S ” and that the infimum may have been referred to as the “greatest lower bound” of the set S , sometimes abbreviated as “glb S ,” and some of you may have actually seen these terms and abbreviations, just as some of us may have encountered as first-graders “and” and “take away” operations which were later on given their proper names of “addition” and “subtraction.” Please be assured that the terms “supremum” and “infimum” are the standard terms in mathematics.

In the four Exercises starting from Exercise 3.5.1, please use the Statement of Completeness in the form given above. That is, show in each of these Exercises that the statement “every Cauchy sequence of real numbers has a limit” logically implies what is asserted in the Exercise. Also, before doing them it might be a good idea to read and digest what is being said in all of these four Exercises, and also what is said, down through the Remark which follows these four Exercises.

Exercises

3.5.1 A **monotone** sequence is a sequence $\{s_n\}$ in which it is either true that $s_n \leq s_{n+1}$ for all n (non-decreasing), or that $s_n \geq s_{n+1}$ for all n (non-increasing).

Show that every bounded monotone sequence has a limit. Show this based upon our Statement of Completeness (above). As a hint, I can think of two very different proofs for this. Feel free to choose either of the two approaches described below.

The first of the two approaches is a frontal attack which starts with a proof that every bounded monotone sequence is a Cauchy sequence. It is easier to see how to do this if one proves the contrapositive. That is, assume that a sequence is monotone and is not a Cauchy sequence. Then, based upon the logical negation of the statement that the given monotone sequence is a Cauchy sequence, show that it cannot be bounded. The monotonicity is essential here; there are lots of bounded sequences which are not Cauchy sequences, after all, but they are not monotone. So do make sure that your proof actually uses the monotonicity.

The second approach would use a variant of the bisection algorithm to converge on something which you then would need to show is equal to the claimed limit of the sequence.

Please also notice that many real analysis texts and many real analysis courses, wishing to present a description of the set of real numbers instead of an actual construction, assume the content of the next Exercise 3.5.2 as an “axiom.” But while it is indeed true that one can prove using Exercise 3.5.2 that every bounded monotone sequence must converge, and the proof is available from many sources including sources online, please note that such is not appropriate here. We have proven the Statement of Completeness, which says that every Cauchy sequence of real numbers has a limit, and we have to work with that. We have neither started our study of the real numbers by listing Exercise 3.5.2 as an axiom nor have we previously proved that what Exercise 3.5.2 says is actually true.

3.5.2 Prove based upon our Statement of Completeness above that every non-empty subset of \mathbf{R} which is bounded above has a supremum. HINT: A good way to approach this problem is to assume that we have an upper bound b_0 for the nonempty set and a number a_0 which is in the nonempty set. Then construct a variant of the bisection algorithm, which will “zero in” on something which looks as though it ought to be the supremum of the set. Then prove that what you have come up with really is the supremum. But take care when constructing your bisection algorithm. There are differences between this situation and what we were doing when trying to zero in on $\sqrt{2}$. And also, while you are doing this construction, do not ruin an otherwise valid proof by assuming that an arbitrary non-empty subset of the real numbers can be represented as the range of a sequence. For, as we will actually prove later on, such an assumption is in general false.

Note on Exercise 3.5.2 : You are asked to show that the statement in the problem follows logically from the Statement of Completeness which was given above. If for any reason you think that what this problem asks you to do is impossible or is somehow nonsensical because it asks you to prove what you think you have heard or read somewhere else is some kind of an “axiom” and everybody knows that you can’t prove an axiom so this problem must be mis-stated or must simply be wrong, then you are mistaken and for more than one reason. If you are having such an experience, then it might be helpful to re-read the title of this section, and the paragraphs in this section which precede this set of exercises, and to digest what those words are saying. Also it might help if you read the Remark which is below this problem set, which says some of the same things again.

3.5.3 Every non-empty subset of \mathbf{R} which is bounded below has an infimum.

- 3.5.4** (a) Every nested sequence of bounded closed intervals has a non-empty intersection. More exactly:
Let $\{I_n\}$ be a sequence of bounded closed intervals, such that $I_{n+1} \subseteq I_n$ for each n . Then

$$\bigcap_{i=1}^{\infty} I_n \neq \Phi.$$

To get started on the proof, note that if we have a sequence of bounded intervals as given in the problem, then each I_n can be represented as $I_n = [a_n, b_n]$, and as a result of this we have two sequences of numbers $\{a_n\}$ and $\{b_n\}$. And the first of these sequences is non-decreasing and the second is non-increasing. If you have done Exercise 3.5.1 then you can certainly use it here to show that each of these sequences has a limit. But then you need to show something more, too. You need to show that if $\lim a_n = a$ and $\lim b_n = b$, then it also follows under the given conditions that $a \leq b$, which one needs to know in order to be sure that the intersection of the given sequence of intervals is indeed nonempty.

- (b) Finally, the conclusion in (a) need not be true if the word “bounded” is omitted from the hypotheses. Find or construct an example of a nested sequence of unbounded closed intervals for which the intersection is, in fact, empty.

And there are even more equivalent formulations of the Statement of Completeness, not listed here. Some of these additional formulations will be explored in future exercises, even in subsequent sections or chapters of this text.

Remark: Quite often, mathematics courses whose content depends upon the properties of the real number system **do not contain a construction of the real numbers**. Rather, it is quite common that the real numbers are taken as a “given” and the course starts off with a list of their properties which is purely descriptive in nature. The listed properties, which describe the set of real numbers, are then called by the name “Axioms.” **This has not been our approach.** Instead, what has been done in the preceding pages is to perform an actual **construction** of the real numbers by performing the **completion** of the set of rational numbers. This is something other than a mere description or listing of whatever properties the real numbers are supposed to have. Unavoidably, when the merely descriptive approach has been taken, the Statement of Completeness or one of several logically equivalent statements is asserted as one of the descriptive axioms. What is usually done in those descriptive treatments is to use the statement which is seen in Exercise 3.5.2, and to dignify that statement with the name “The Completeness Axiom.” No proof is then given for that statement precisely because it has been taken as an axiom. There are three things which the student should note well at this point:

- i. The Statement of Completeness, given above as the statement that every Cauchy sequence of real numbers converges to a real number is **not** an axiom. It is a theorem. The proof of that theorem was the topic of Exercise 3.4.21. That proof depended upon the construction of the real numbers which preceded the Statement of Completeness. For this reason, we did not take and we did not need to take the Statement of Completeness as an axiom. And as the final part of doing the construction we actually showed that the Statement of Completeness is true for the set which we had constructed. Less thorough treatments of the real numbers which do not do a construction, but which merely describe, have no choice but to state as an “axiom” something which is logically

equivalent to our Statement of Completeness. For, completeness, after all, is one of the essential properties of the real number system.

- ii. Quite independently of whether or not the real numbers are constructed or merely described, all of the statements in the previous set of exercises are in fact logically equivalent to one another, and also to our Statement of Completeness. That means, it is possible to assume any one of them as a hypothesis and prove any other of them as a conclusion. Whether or not any particular one of them is deemed to be an axiom has no relationship to their logical equivalence, which is a fact in itself. Nevertheless, our starting point needs to be the Statement of Completeness which arises naturally from our construction of the set of real numbers, and not one of the logically equivalent statements which someone else may have taken as a descriptive axiom.
- iii. Do not make the mistake of assuming that the statement in any one of the above problems is true while attempting to prove it. Especially, the temptation to view the statement in Exercise 3.5.2 as “The Completeness Axiom” while attempting to work the problem is very great because that statement is listed as “The Completeness Axiom” in many books and articles on real analysis – and also because it is treated as an axiom in many online materials. Again, those sources intend to describe the real numbers instead constructing them. If you succumb to such a temptation, it will severely impede your understanding. It will probably vitiate all of your efforts to complete Exercise 3.5.2. Why is that? Because you will believe that what you are **supposed to prove** in Exercise 3.5.2 is an “axiom” in spite of everything that is being said right here. If you have fallen into this trap, you will somehow think that Exercise 3.5.2 is somehow requesting you to use its conclusion to prove itself, which it most definitely is

not, and which would obviously be an absurd thing to ask anyone to do. Or else, you will try in the middle of the proof to smuggle in the “Completeness Axiom” by the back door while trying to prove it. Again, the reason for this confusion: many books, including even the typical calculus text, and many online resources, do indeed refer to a “Completeness Axiom” and state it in the same way as what you are supposed to prove in Exercise 3.5.2. Those books intend merely to describe the number system and then to get on with some other business. Please do understand that our Statement of Completeness was a **theorem** which follows from **an actual construction** of the real number system. In our context, then, the name “Completeness Axiom” is a misnomer. The real number system which has been constructed in the preceding pages certainly does have the property of completeness, but that property does not come in the form of an axiom.

Now, it is also true that other constructions of the real number system are feasible (see Exercise 3.5.8 below). If the set of real numbers has been constructed by some other method of construction, then the Statement of Completeness might very well have a natural expression which naturally derives from **that** construction, just as our Statement of Completeness naturally derives from our construction. Moreover, even if the real numbers are merely presented as some pre-existing entity which one wishes to describe with a set of “axioms” there still remains the need to show that all of the different possible expressions of the “Completeness Axiom” are logically equivalent. For example, if a “Completeness Axiom” has been stated in the form which is given in Exercise 3.5.2, it is then clearly necessary to prove as a theorem that every Cauchy sequence of real numbers converges, that every bounded monotone sequence of real numbers has a limit, and any or all of the other statements which are equivalent to the stated “Completeness Axiom” so that these other statements are ready for use when needed – and they will be needed. The following set of Exercises explores some

of these “reverse” implications and some other related topics.

Exercises

3.5.5 Every non-empty subset of \mathbf{R} which is bounded above has a supremum if and only if every non-empty subset which is bounded below has an infimum.

3.5.6 If every non-empty subset of \mathbf{R} which is bounded above is presumed to have a supremum, then it follows that every bounded nondecreasing sequence has a limit. A similar statement is true about sets which are bounded below and nonincreasing sequences. Note that, in this problem, one is showing that the statements about bounded monotone sequences which are the conclusions of Exercise 3.5.1 can be derived from the conclusion stated in Exercise 3.5.2.

3.5.7 If every bounded monotone sequence is assumed to converge, it follows that every Cauchy sequence also converges. Note that what we are doing here is to prove that the conclusion of Exercise 3.5.1 implies our Statement of Completeness. Hints for this problem: Show first that every sequence contains a monotone subsequence, then second, that a Cauchy sequence is bounded and therefore that any subsequence of a Cauchy sequence must also be bounded, and third that, if a Cauchy sequence has any convergent subsequence, then the original Cauchy sequence must itself converge to the same limit as the subsequence. If now the subsequence is monotone and hence converges, then the original Cauchy sequence must converge, too, and to the same limit.

3.5.8 As there are several characterizations of completeness, so there are alternative methods for the construction of the real numbers, other than the method of dealing with Cauchy sequences as we have done in the preceding sections of this chapter. One of those alternative methods for constructing the set of real numbers is the method of *Dedekind cuts*. Do a class presentation outlining this method. The Wikipedia article “Dedekind Cut” is an excellent place to start, as it contains an accurate description of this method. But be prepared to discuss the following questions:

- What are the benefits of this method? (if you see any benefits)
- What are its deficiencies or disadvantages? I perceive at least two deficiencies. First, it is often more difficult actually to define a Dedekind cut than it might appear in the abstract. Second, actually to define the operations of algebra on the collection of Dedekind cuts of the rational numbers is unexpectedly complicated, and the outcome seems to be forced and unnatural. But nothing says you have to agree with me. What do you think about these objections?
- Assuming that one has constructed the set of real numbers by defining the set of real numbers to be the set of all Dedekind cuts of the rational numbers, then which of the several equivalent statements which we have considered is the most natural one to use for describing the completeness of the set of real numbers? And does that statement follow from the method of construction which was used?

Finally, it should be noticed that there is a characterization of the real numbers which is often stated in textbooks used in more introductory classes in mathematics than this one. Namely, the real numbers are often described as the set of all numbers which can be described by a decimal expansion. The justification of this statement is a project in itself.

For, to prove that this is indeed true requires a lot of machinery which we have not developed. Nevertheless, it is also a true and accurate characterization of the real number system. Some aspects of this characterization will be described or delved into, later on. For the here and now, let us only note that a decimal expansion of a real number can be viewed as a sequence of a special type which converges to the real number in question. To see this, let us assume without loss of any actual generality that the real number in question is somewhere in the interval $(0, 1]$. Then, we have a decimal expansion for the number r which looks like

$$r = .a_1a_2a_3a_4 \dots$$

in which each of the numbers a_k is one of $\{0, 1, \dots, 9\}$. Then, what we have in terms of sequences and limits is the statement that

$$r = \lim_{n \rightarrow \infty} A_n$$

in which $\{A_n\}$ is the sequence of fractions defined by $A_n = \frac{a_1 \dots a_n}{10^n}$ where the numerator $a_1 \dots a_n$ is the integer which has the given digits in its (standard) decimal expansion. Consequently, the representation of a real number as a decimal expansion is but a special case of the representation of a real number as a Cauchy sequence. Indeed:

Exercises

3.5.9 Let $\{A_n\}$ any sequence whose terms are of the form

$$A_1 = \frac{a_1}{10}$$

and for $n > 1$

$$A_n = A_{n-1} + \dots + \frac{a_n}{10^n}$$

in which for each n the integer a_n satisfies $0 \leq a_n \leq 9$. Then $\{A_n\}$ is a Cauchy sequence.

Chapter 4

Series finite and infinite

4.1 Finite series and sigma notation

Let us start with a finite sequence of numbers $\{a_1, \dots, a_n\}$. Then, the notation

$$\sum_{k=1}^n a_k$$

denotes the sum of the numbers in the sequence. The number denoted by k is called the *index of summation* or, sometimes, the “counter” or the “pointer.” Notice that its presence here is important because of its role, not its name or its particular symbol. Instead of k , one could have called it i or j or m upon whim, without changing the meaning at all. It should also be clear that one can systematically change the upper and lower limits for the index which are given in the sum and at the same time can systematically relabel the terms in the sum, and the sum comes out to be the same thing. As a couple of simple examples of what the previous sentence is saying, it is clearly true that

$$\sum_{k=1}^n a_k = \sum_{k=0}^{n-1} a_{k+1} = \sum_{k=2}^{n+1} a_{k-1}$$

because the same things are being added together. Moreover, though the description above considers a sequence of consecutive numbers whose corresponding indices or counters start off with 1, the corresponding counters could obviously start with any integer and proceed from there. Indeed, in many applications, especially those involving successive powers of a number, it is quite usual to start the indicated sum with the initial index value 0 instead of 1 (see for example Problem 4.1.4 below).

Because of its flexibility in expression, the sigma notation is quite powerful and makes a lot of things easy which are otherwise not at all obvious. But the sigma notation is often introduced so suddenly (usually in the middle of a calculus course which is already moving too fast) that some of the students come away from the experience with phobias. Let us try to overcome those phobias in case that any remain. First, let us note that the following two properties are obvious:

- $\sum_{k=1}^n a_k = a_1 + \sum_{k=2}^n a_k = a_n + \sum_{k=1}^{n-1} a_k$
- More generally, when $1 \leq m < n$ we have

$$\sum_{k=1}^m a_k + \sum_{k=m+1}^n a_k = \sum_{k=1}^n a_k$$

Here follow some introductory exercises. Note that, for really rigorous proofs of some of them, you may wish to use Mathematical Induction or the Well Ordering Principle.

Exercises

4.1.1 Show that $\sum_{k=1}^n a_k + \sum_{k=1}^n b_k = \sum_{k=1}^n (a_k + b_k)$.

4.1.2 Further, show that the distributive law holds, namely that if c is any number, we have $c \sum_{k=1}^n a_k = \sum_{k=1}^n ca_k$.

4.1.3 Exercise on “the gentle art of index sliding”: Give an actual proof based upon Induction or the Well Ordering principle, that

$$\sum_{k=1}^n a_k = \sum_{k=0}^{n-1} a_{k+1} = \sum_{k=2}^{n+1} a_{k-1}$$

4.1.4 (Sum of the finite geometric series) Use the properties given in the previous exercises to show that, when r is any number, we have

$$(1-r) \sum_{k=0}^n r^k = \sum_{k=0}^n r^k - \sum_{k=0}^n r^{k+1} = 1 + \sum_{k=1}^n r^k - \sum_{k=1}^n r^k - r^{n+1} = 1 - r^{n+1}$$

From this, find the formula which gives $\sum_{k=0}^n r^k$ whenever $r \neq 1$.

Note: An agreement on a small point regarding notation is relevant here. The value of r^0 is equal to 1 for all values of r which are not zero. We agree in the present context that $r^0 = 1$ even when $r = 0$. Without any such carefully restricted context, the expression “ 0^0 ” can not be assigned any precise meaning. For, there are two ways to assign a definition and the two ways do not lead to the same outcome.

4.1.5 Note that $\sum_{k=1}^n 1 = n$, and note that $\sum_{k=1}^{n+1} k^2 - \sum_{k=1}^n k^2 = (n+1)^2$. The first series on the left can now be rewritten using “index sliding,” obtaining

$$\sum_{k=1}^{n+1} k^2 = \sum_{k=0}^n (k+1)^2 = \sum_{k=0}^n k^2 + \sum_{k=0}^n 2k + \sum_{k=0}^n 1$$

Therefore, from

$$\sum_{k=1}^{n+1} k^2 - \sum_{k=1}^n k^2 = (n+1)^2$$

one obtains (noticing that all positive powers of zero are zero),

$$\sum_{k=1}^n k^2 + \sum_{k=1}^n 2k + n + 1 - \sum_{k=1}^n k^2 = \sum_{k=1}^n 2k + n + 1 = (n+1)^2$$

. Verify the above steps for yourself, and complete the problem by finding a formula for $\sum_{k=1}^n k$.

Note that Problem 4.1.5 actually gives a constructive, step by step derivation of a summation formula. The student may well have been *given* the resulting formula or a similar one and asked to prove by mathematical induction that the formula, as given, is correct. But mathematical induction is not very helpful, at all, if one does not know the formula and needs to *find* it.

Exercises

4.1.6 The previous problem can be generalized to a method for finding the summation formula for $\sum_{k=1}^n k^p$ in which p is any positive integer. Show in detail how this can be done when $p = 2$, and explain what is needed in order to obtain the formula corresponding to any higher value of p .

4.2 Tools – the binomial theorem

The binomial theorem states that, for a given positive integer n

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k},$$

in which

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is often read as “n choose k” from another application of the binomial theorem, in probability.

Exercises

4.2.1 Prove the Binomial Theorem (Hint: try using induction).

As a concluding remark about the Binomial Theorem, it should be noted that it can be combined with the methods developed in Exercise 4.1.5 and in Exercise 4.1.6 in order to develop a general formula for $\sum_{k=1}^n k^p$ in which p is any positive integer. Unfortunately, and despite the fact that this would be a topic interesting in itself, at this point the development of such a general formula would take us a bit off the track of the rest of this chapter. And so we put that matter aside.

4.3 Infinite series

An infinite series is written formally as

$$\sum_{n=1}^{\infty} a_n,$$

where $\{a_n\}$ is some sequence of numbers, which in this connection are called *terms* of the series. Closely related to the series is the sequence of *partial sums* $\{A_N\}_{N=1}^{\infty}$, where A_N is defined by

$$A_N = \sum_{n=1}^N a_n.$$

We say that the series *converges* or *is convergent* provided that

$$\lim_{N \rightarrow \infty} A_N$$

exists and is finite (this statement is for clarity only, as we have not at this time defined limits which are other than finite). A series which is not convergent is often called *divergent*, or it is said that the series *diverges*.

Students in calculus classes often find the difference between “has a limit” and “converges” to be confusing. There really ought to be no confusion, provided only that the student realizes the reason we have these two slightly different

concepts in the first place. Namely, the use of the term “converges” instead of “has a finite limit” harks back to the use of infinite series and infinite sequences as computational procedures. For, when one is actually using a procedure for computation it is neither interesting nor useful if the results of the procedure increase without bound and “go to infinity” or decrease without bound and “go to minus infinity” or just jump around without actually settling down on a limit. In fact, such results usually indicate that something needs to be done over again because nothing good has come out of the procedure.

Exercises

4.3.1 If $\sum_{n=1}^{\infty} |a_n|$ converges, then $\sum_{n=1}^{\infty} a_n$ also converges. A given series $\sum_{n=1}^{\infty} a_n$ for which $\sum_{n=1}^{\infty} |a_n|$ is known to converge is said to **converge absolutely**.

Hint for the proof: Having been through the construction of the real numbers in the previous chapter, we now know that every convergent sequence of real numbers is a Cauchy sequence, and that every Cauchy sequence of real numbers converges (to something, which is a real number, even though we may not know explicitly and precisely what number it is). Therefore show that the sequence of partial sums of the series $\sum_{n=1}^{\infty} a_n$ is a Cauchy sequence, on

the grounds that the series $\sum_{n=1}^{\infty} |a_n|$ is known to have a finite limit. The Triangle Inequality for absolute value is helpful here.

The result shown in the preceding problem is often restated as “A series which converges absolutely must itself converge.”

Before moving on to other matters, let us notice that the converse of this statement is obviously not universally valid. It is quite possible for a given series $\sum_{n=1}^{\infty} a_n$ containing some mixture of positive and negative terms to converge while the series $\sum_{n=1}^{\infty} |a_n|$ diverges. In such a situation, the series $\sum_{n=1}^{\infty} a_n$ is said to **converge conditionally**. There are many examples of series which converge conditionally, and we will discuss some of them later on in this chapter.

Exercises

4.3.2 Consider two series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$. If there is N such that $0 \leq a_n \leq b_n$ holds for all $n \geq N$, then

- (i) if the second series converges, then the first also converges
- (ii) if the first series diverges, then the second also diverges.

4.3.3 Given two series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$. If both converge, then

$\sum_{n=1}^{\infty} (a_n + b_n)$ also converges, and moreover

$$\sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} (a_n + b_n)$$

Similarly, if k is any non zero number, then $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} k a_n$ converges. Moreover, when both converge, we have

$$\sum_{n=1}^{\infty} k a_n = k \sum_{n=1}^{\infty} a_n$$

The above formula is also true when $k = 0$ if $\sum_{n=1}^{\infty} a_n$ converges, of course, but is hard to interpret if $\sum_{n=1}^{\infty} a_n$ diverges

(The left side is obviously zero then, but just what exactly does the right side then mean?).

4.4 Tools – the geometric series

A particularly interesting and basic infinite series is the geometric series

$$\sum_{k=0}^{\infty} r^k.$$

Exercises

4.4.1 The geometric series diverges for $|r| \geq 1$, while for $|r| < 1$ its sum is equal to $\frac{1}{1-r}$. Note: An integral part of this problem is to show rigorously, **by appeal to things done previously in this text only**, that

$$\lim_{n \rightarrow \infty} r^n = 0 \text{ whenever } |r| < 1.$$

4.4.2 Any repeating decimal defines a rational number, and any rational number can be represented as a repeating decimal.

4.5 A small excursion – the definition of e

Taking a historical view of mathematics and its applications, the number e arose from the study of compound interest. If the period of compounding is decreased to zero, then one is presumed to have “continuous compounding,” and the question naturally arose as to what happens then. As a very simple case, suppose one has annual compounding at 100% interest, starting with \$1.00 principal. Then, after one year one has \$2.00. If one compounds quarterly, then one has four periods per year, and the interest rate is therefore 25% per period. The amount in dollars accumulated after one year (ignoring such inconveniences as roundoff to the nearest cent, which we intend to ignore here for the purposes of higher mathematics) is

$$\left(1 + \frac{1}{4}\right)^4.$$

More generally, if the compounding is with n periods per year, then the amount accumulated after one year at 100% interest, starting with one dollar, is

$$\left(1 + \frac{1}{n}\right)^n.$$

At some point, some clever money lender noticed that he could get a lot of word-of-mouth publicity among potential investors and steal a march on his competition by shortening the period of compounding, and he also did some calculations which seemed to indicate that there seems to be some kind of ceiling on what comes out of this sequence, so he could see that the buzz thus generated was not really going to cost him very much money. Based upon these origins and taking

into account what we know now, it is natural to try to define a number, which is usually denoted by the symbol e , as

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n.$$

This number is called e in honor of the eighteenth-century mathematician Euler. To put the matter into modern terms, he gave a careful description of it and a proof that it actually exists, after being given a royal grant by Catherine the Great, Tsarina of Russia, to explore the topic of compound interest.

Of course, the above definition of e is only an attempt at a definition unless and until it can be shown actually to define something. To this end, it is necessary to show that the sequence in the definition actually does have a limit and that this limit actually is a finite number. The details of proving these two things are examined in the first two of the following three problems.

Exercises

4.5.1 Let $s_n = \left(1 + \frac{1}{n}\right)^n$. Show that $\{s_n\}$ is an increasing sequence. To see this, use the Binomial Theorem to expand $\left(1 + \frac{1}{n}\right)^n$ and $\left(1 + \frac{1}{n+1}\right)^{n+1}$ and carefully compare the indexed terms from 0 to n . Specifically, let us represent

$$\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^n a_k$$

and

$$\left(1 + \frac{1}{n+1}\right)^{n+1} = \sum_{k=0}^{n+1} b_k$$

in which a_k and b_k are the terms which arise from the respective expansions using the Binomial Theorem. Assuming that $n \geq 2$, show that $a_0 = b_0 = 1$ and $a_1 = b_1 = 1$ and that for any k from 2 up to n it is true that $a_k < b_k < \frac{1}{k!}$. Thus, it follows that

$$s_n = \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^n a_k < \sum_{k=0}^n b_k < \sum_{k=0}^{n+1} b_k = \left(1 + \frac{1}{n+1}\right)^{n+1} = s_{n+1}.$$

4.5.2 To show that the same sequence $\{s_n\}$ is bounded above and to get a good upper bound, notice that in the problem above it is shown already (using the notation introduced in that problem), that for any $n \geq 2$ we have $a_0 = 1$ and $a_1 = 1$ and that for any k from 2 up to n we have $a_k < \frac{1}{k!}$. From this it follows for arbitrary $n > 1$ that

$$\left(1 + \frac{1}{n}\right)^n < \sum_{k=0}^n \frac{1}{k!}.$$

From this, it follows in turn that

$$\left(1 + \frac{1}{n}\right)^n < \sum_{k=0}^{\infty} \frac{1}{k!},$$

provided that the series on the right of the inequality can be shown to converge. The convergence of this series will be established in the next exercise, and in the course of establishing the convergence an estimate for the error will be established, too.

4.5.3 The series $\sum_{k=0}^n \frac{1}{k!}$ converges. Indeed, for each $N > 0$ the difference between the series and its N th partial sum can be estimated as follows:

$$0 < \sum_{k=0}^{\infty} \frac{1}{k!} - \sum_{k=0}^N \frac{1}{k!} = \sum_{k=N+1}^{\infty} \frac{1}{k!} < \frac{1}{(N+1)!} \sum_{j=0}^{\infty} (N+1)^{-j} = \frac{1}{N!} \cdot \frac{1}{N}.$$

Justify the steps above. Also, show that if $N = 1$, we get the estimate that

$$2 < \sum_{k=0}^{\infty} \frac{1}{k!} < 3.$$

Moreover, note that we have an error estimate for the partial sum of this series which in fact improves rapidly if N increases.

4.5.4 Continuing the line of reasoning used in the immediately previous problems, you ought to be able to show (still using only the tools presented in this chapter!) that in fact

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}$$

Hint: Think of N as a fixed but arbitrary number, and n as a number which we will allow to be much bigger than N . Then, given any $\epsilon > 0$ it is possible to show that for sufficiently large n (obviously bigger than something which is already larger than N) we have

$$\sum_{k=0}^N \frac{1}{k!} - \epsilon < \left(1 + \frac{1}{n}\right)^n$$

from which, for all n sufficiently large, we can say that

$$\sum_{k=0}^N \frac{1}{k!} \leq \left(1 + \frac{1}{n}\right)^n.$$

Since this is true for every N , the claimed equality is established.

4.6 Some discussion of the exponential function

Without going into all the details at this time, let us note before moving on to other topics that it is possible to generalize what was done in the previous section to the discussion of a function of x defined by

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

and similarly it is possible to see that the value of this function can actually be calculated by the infinite series

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

The argument by which this can be shown is similar to what has been done in the previous section, with a few refinements.

What is not clear at first, of course, is that the series just given is also equal to e^x , where e is the number described in the previous section. For, to establish this would require the proof of the identity

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^{nx}$$

This identity can indeed be shown without using any of the tools of “calculus” other than the concept of the limit of a sequence or sum of a series, as has been the practice in this chapter. Probably, the most effective way to carry out the

proof is to do it in stages. First, one may assume that x is an integer. Second, one may show that it is true when x is a fraction of the form $\frac{1}{m}$. Then in the third step one may prove it is true when x is any rational number expressed in the form $x = \frac{m}{n}$. Then finally one can show that the result can be “completed” and shown true when x is any real number. The manner of doing this “completion” would use, essentially, constructions of sequences which are quite similar to those developed in the previous chapter in order to get the job done.

Note, however, that the above identity and trivial themes and variations on it lie behind many of the standard exercises in the calculus book on L’*h*ôpital’s Rule. We can not use L’*h*ôpital’s Rule here, of course, to prove anything at all about such limits. One reason for our inhibitions is that we have not at this point discussed derivatives, not even limits of functions, and so we are certainly not free to use such things. But there is a much more fundamental reason, too. Namely, the typical calculus book has not really defined the exponential function at all and indeed has not even taken the trouble to give a rigorous definition of e . Instead, the book has just done some vague mumble and hand-waving about such matters. In most of the books currently in use, this vague mumble and hand-waving about e takes place in some very early chapter. Quite often, that very early chapter is even a chapter which is skipped during the course because there is too much other stuff which needs to get done, and the contents of that chapter are thus ignored both by the teacher and the students. Taking a somewhat cynical point of view, it seems it is apparently hoped (and, alas, not unreasonably so in practice) that by the time the students come to learn about L’*H*ôpital’s Rule they have forgotten just how vague the mumble and hand-waving were concerning that funny number e – if indeed they were actually asked in the first place to look at that introductory chapter. Therefore, the arguments found in the Calculus book which, based upon mumble and hand-waving, purport to establish results via L’*h*ôpital’s Rule about limits involving the exponential

and logarithm functions are circular arguments by their fundamental nature. It is merely hoped that the students are too inattentive to catch on to what is being done to them.

Also note that such is not what we have done here.

4.7 More on Convergence of Series

4.7.1 The Root Test and the Ratio Test

Two more tests which are often used for showing the convergence or divergence of an of infinite series. These tests are called the Root Test and the Ratio test. The most general of the two is the Root Test. A complete statement of the Root test would depend upon the concept of the “limit superior” of a sequence, which we have not yet introduced. Because of this limitation, we present here the Root Test in the more restricted version which is often found in calculus books. This more restricted version is certainly valid. There is nothing wrong with it. But the more restricted version applies to fewer series than does the full version. We also give the Ratio Test.

The Root Test and the Ratio Test are specifically related to series with positive terms. It goes without saying, therefore, that they can be used to study the question of absolute convergence of any series $\sum_{n=1}^{\infty} a_n$ in which the terms are not all of the same sign. In such a context, the Root Test or the Ratio Test should be applied to the related series

$\sum_{n=1}^{\infty} |a_n|$. If the test shows convergence of this second series, then the first one converges, too, and converges absolutely.
See Exercise 4.3.1.

Exercises

4.7.1 In this exercise, you should prove a weak version of the Root Test, the version which is presented in most calculus texts. A stronger version of the Root Test will be presented in Exercise 6.5.8, after we have introduced the concepts which are needed in order to express it.

Consider an infinite series $\sum_{n=1}^{\infty} a_n$ in which all of the terms (or all of the terms for $n \geq N$, some N) are non-negative. Then, let $r = \lim_{n \rightarrow \infty} (a_n)^{\frac{1}{n}}$, provided that the limit exists. Prove the following:

- If $r < 1$ the series converges.
- If $r > 1$ the series diverges.
- If $r = 1$ then this test tells absolutely nothing about the convergence of the series. Why? And give an example of a convergent series and a divergent series for each of which $r = 1$.

4.7.2 (the Ratio Test) Consider an infinite series $\sum_{n=1}^{\infty} a_n$ in which all of the terms (or all of the terms for $n \geq N$, some N) are positive. Then, let $r = \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}$, provided that the limit exists. Prove the following:

- If $r < 1$ the series converges.
- If $r > 1$ the series diverges.
- If $r = 1$ then this test tells absolutely nothing about the convergence of the series. Why? And give an example of a convergent series and a divergent series for each of which $r = 1$.

A hint for the proof of both of the preceding exercises is to do a comparison test (see Problem 4.3.2) of the given series with a geometric series $\sum_{n=0}^{\infty} \rho^n$ (or, better said, a comparison of the terms in the “tail” of the given series with the corresponding terms in the “tail” of the geometric series, consisting in each case of all the terms beyond some sufficiently large index). If $r < 1$, then ρ can be chosen as any number satisfying $r < \rho < 1$, and if $r > 1$ then choose ρ to be any number such that $1 < \rho < r$. And note that when $r = 1$ it is not possible to do either of these.

The above hint shows that the Root Test and the Ratio Test depend, both of them, upon a comparison with a geometric series. In the previous discussion of e we did in fact use an argument based upon the geometric series to establish error estimates and thereby to prove convergence.

Also, a few words about the use of these two theorems. The Ratio Test is often easier to set up, but it is less general. Even in the more restricted form seen in this section the Root Test is more general. The Ratio Test is completely

destroyed if, for example, a lot of the terms a_n could be zero. The problem becomes especially bad if the occurrence of zero terms happens in some apparently irregular fashion. The Root Test is also a lot quicker to do, provided that one is actually able to compute the limit involved, But it is sometimes not easy to compute such limits. To this end, the following problems are relevant. Note that it is forbidden while working these problems to use L'Hôpital's rule or any other tool involving derivatives. We have not even defined derivatives yet. Therefore we surely can not use them for anything at all. We also have not formally developed the exponential function, nor have we even begun to define or discuss the logarithm function. Please stick to using machinery which is developed in the preceding pages, only.

Exercises

4.7.3 (this problem deals with a loose end which needs to be tied up, and now is the time) Formulate an appropriate definition for what we ought to mean by the statement $\lim_{n \rightarrow \infty} a_n = \infty$.

4.7.4 If $a > 0$, then $\lim_{n \rightarrow \infty} a^{\frac{1}{n}} = 1$. Hint: Show first that if $a > 1$ we have a decreasing sequence which is bounded below by 1. Show that, from this, it follows that when $a > 1$ the limit must exist and must be greater than or equal to 1. Find a way to complete the argument by showing that the limit (which is already shown to exist) can not, in fact, be any number which is strictly greater than 1.

Similarly, if $a < 1$ show that we have an increasing sequence which is bounded above by 1. The rest of the argument is quite similar to one for the case that $a > 1$.

4.7.5 $\lim_{n \rightarrow \infty} n^{\frac{1}{n}} = 1.$

Hint 1: As the first step, show that this sequence is decreasing for $n \geq N$. This means, find that N , for one thing. Also notice that $n^{\frac{1}{n}} \geq 1$ for all n . Therefore, the series has a limit, and the limit is not less than 1. Complete the problem by showing that indeed the limit is 1. One possible way to do this is to notice that, once it is known that there is a limit, it must be true that $\lim_{n \rightarrow \infty} n^{\frac{1}{n}} = \lim_{n \rightarrow \infty} (2n)^{\frac{1}{2n}}$. The right side may now be rewritten as $\lim_{n \rightarrow \infty} (2)^{\frac{1}{2n}} \cdot (\lim_{n \rightarrow \infty} (n)^{\frac{1}{n}})^{\frac{1}{2}}$. Now use the previous problem to compute $\lim_{n \rightarrow \infty} (2)^{\frac{1}{2n}}$. The result is that the limit which we want to compute is equal to its own square root. The limit is therefore one of 0 or 1 or ∞ . Now explain why the limit cannot be 0 and cannot be ∞ . Fill in all details.

An alternative proof for this result follows entirely different lines from what is suggested in Hint 1. This alternative proof is outlined in

Hint 2: Since clearly $n^{\frac{1}{n}} \geq 1$ for all $n > 1$, let us write $n^{\frac{1}{n}} = 1 + b_n$, with $b_n \geq 0$ for every $n > 1$. Then, by taking the n th power of both sides, we have $n = (1 + b_n)^n$. Expand the right side using the Binomial Theorem, and then notice that the equation becomes an inequality if we subtract 1 from both sides and then toss away every term on the right except the term involving b_n^2 . That is, that $n - 1 > \binom{n}{2} b_n^2$. Insert the correct value of $\binom{n}{2}$. Then solve the inequality for b_n , showing that b_n is bounded above for each n by a certain sequence which tends to 0 as $n \rightarrow \infty$. Fill in all details.

4.7.6 $\lim_{n \rightarrow \infty} (n!)^{\frac{1}{n}} = \infty$. Hint: Show that if $n > K$, where K is any fixed but arbitrary integer (think of it as large) then $n! > K! \cdot K^{n-K} = (K^{-K} K!) K^n$. Now notice that the factor in parentheses is a constant. Thus, by Problem 4.7.4 we have $\lim_{n \rightarrow \infty} (K^{-K} K!)^{\frac{1}{n}} = 1$, and it follows that there exists N such that when $n \geq N$ one has $(K^{-K} K!)^{\frac{1}{n}} > \frac{1}{2}$. One might infer from this that $\lim_{n \rightarrow \infty} (n!)^{\frac{1}{n}}$ is bounded below by $\frac{K}{2}$. Fill in the details, and complete the proof.

Remark: Another test for convergence or divergence which is applicable to some infinite series which have positive terms is called the Integral Test. This particular test will be discussed later, in Section 10.4. Before presenting it we will have to go through the development of the integral, and we will also need to introduce improper integrals.

Exercises

4.7.7 The series $\sum_{n=2}^{\infty} \frac{1}{n^2 - 1}$ is one of those rare series for which we can directly compute the sum. To carry out that computation, the first step is to split $\frac{1}{n^2 - 1}$ into two fractions by using partial fractions. Then separate the partial sum which stops at N into a difference of two finite series. Now use index sliding to rewrite the partial sum

$\sum_{n=2}^N \frac{1}{n^2 - 1}$ as a telescoping sum. Finally, evaluate the sum of the series, which is the limit of the partial sums.

4.7.8 The series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ is a p -series with $p = 2$. In Exercise 10.4.5 we will see that this series converges because the value of p is greater than 1. But here we can use a more direct argument to prove the convergence. Namely, one can use a direct comparison test with the series in the previous problem. Explain, providing all details.

4.7.9 Compute the limits involved in applying the Root test and the Ratio test to each of the two series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ and $\sum_{n=1}^{\infty} n^2$. And what sort of convergence/divergence behavior does each of these two series have, in fact?

4.7.2 Conditional Convergence and the Alternating Series Test

In a previous section, in or near Exercise 4.3.1, the concepts of absolute and conditional convergence were introduced. We have seen just above that there are several extensive and detailed results about absolute convergence. What about results for conditional convergence? Well, that is a bit harder to prove, except in some very special circumstances.

Without doubt, there are many series for which it is easy to show that they do not converge absolutely, and thus if they converge at all then they must converge conditionally, but no clear path toward such a proof seems to open itself for that, either. In exactly one circumstance, though, we do have a result which shows convergence of a series in which not all the terms are of the same sign, and it is rather nice. The result is called the **Alternating Series Test**. It also comes with an error estimate for the difference between the sum of the series and its N th partial sum.

Theorem: *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of non-negative numbers. Then the series*

$$\sum_{n=1}^{\infty} (-1)^{n-1} a_n$$

will converge provided that the following two conditions are met:

- (i) $a_n \rightarrow 0$ as $n \rightarrow \infty$
- (ii) $a_n \geq a_{n+1}$ for all n .

Moreover, if the above two conditions are met, then the magnitude of the difference between the sum of the series and the N th partial sum of the series is not greater than a_{N+1} . Furthermore, the N th partial sum exceeds the sum of the series if N is odd and is less than the sum of the series if N is even.

Exercises

4.7.10 Prove the above result. As a hint, consider that the sequence of partial sums can be rewritten by grouping the terms two at a time in two different ways.

4.7.11 The series

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

is called the **harmonic series**. Show that for every $k \geq 1$ it is true that

$$\sum_{j=2^k+1}^{2^{k+1}} \frac{1}{j} > \sum_{j=2^k+1}^{2^{k+1}} \frac{1}{2^{k+1}} = \frac{1}{2}.$$

Based upon this result, show that the harmonic series diverges.

4.7.12 Show that the **alternating harmonic series** defined by

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

converges, as it meets the criteria of the Alternating Series Test.

4.7.13 Consider the sequence $\{a_n\}$ defined by $a_n = 10/n$ when n is odd and $a_n = 1/n$ when n is even, and the series

$$\sum_{n=0}^{\infty} (-1)^{n-1} a_n$$

based upon this sequence $\{a_n\}$. Does this series meet all of the conditions which are required in the Alternating Series Test? Also, find an answer to the question of whether this series converges or diverges.

Concluding Remarks: The Alternating Series Test is a test for convergence. It applies to any series in which the terms of the series satisfy the hypotheses. All by itself, the Alternating Series Test says nothing at all, neither one way or the other, about whether a series converges conditionally or absolutely. Thus:

1. If a given series can be seen to converge by appeal to the Alternating Series Test but fails to converge absolutely, then that series does converge conditionally.
2. If a given series can be seen to converge by appeal to the Alternating Series Test and also can be seen (by some other test) to converge absolutely, then the series does converge absolutely. The question of whether the Alternating Series Test applies to the given series is orthogonal to the question of absolute convergence. The Alternating series test does not violate the concept of absolute convergence. Neither does absolute convergence affect the possibility that the Alternating Series test may also pertain to a given series.

3. If a series is seen to converge by the Alternating Series Test then the error estimate which is part of the Alternating Series Test also pertains to that series. Again, the matter of whether the Alternating Series Test pertains to the series has nothing whatsoever to do with the question of whether the given series converges absolutely, or does not converge absolutely. The error estimate which is part of the Alternating Series Test is a very nice and easily applied error estimate, though, isn't it? And error estimates are, as a general rule, hard to come by. For that reason, if the error test associated with an alternating series is applicable, it is often used even if the series we are looking at does happen to converge absolutely.

Chapter 5

Topological concepts in analysis

5.1 Basics

We have already seen that limit processes are essential to the very definition of the real numbers. We now explore related concepts more systematically, intending to use later on the concepts and terminology which are developed in this

chapter.

Given a subset S of \mathbf{R} , we say that a point x is an *accumulation point* of S , provided that for every $\delta > 0$ there exists a point $s \in S$, such that $0 < |x - s| < \delta$.

We say that x is an *interior point* of S , provided that there exists $\delta > 0$ such that $(x - \delta, x + \delta) \subset S$.

Related to these two definitions we say that a set S is *open*, provided that each point of S is an interior point of S . The set S is *closed* if S contains all of its accumulation points.

In connection with the definitions of accumulation points and interior points, there is some fairly standard notation. Given a set A ,

- the notation A° denotes the set of interior points of A .
- the notation A' denotes the set of accumulation points of A .

Another commonly used, related concept and its notation is

- the closure of A , which is $A \cup A'$, written \overline{A}

The set \overline{A} is also often referred to as the set of *limit points* of A .

Exercises

5.1.1 Above, we have defined what we mean by “ x is an interior point of the set S ” and what we mean by “ x is an accumulation point of the set S .” Formulate the logical negation of each of these statements. Check the summary

section at the end of Chapter 1 on how to construct negations, and be certain that you have followed the procedures which are described there.

5.1.2 According to the above definitions of open and closed sets, both \mathbf{R} and Φ (the empty set) are open, and both are closed.

5.1.3 Any singleton set (i. e. any set containing only one real number in it) is closed, on the grounds that it has no accumulation points.

5.1.4 Every open interval (bounded or unbounded) is an open set.

5.1.5 Every closed interval (bounded or unbounded) is a closed set.

5.1.6 The union of the open sets in any collection of open sets is open.

5.1.7 The intersection of any finite number of open sets is open.

5.1.8 The intersection of an infinite number of open sets may fail to be open. Such a denial of a general statement requires an example. Please provide one.

5.1.9 The complement of any open set is closed, and the complement of any closed set is open.

5.1.10 The intersection of closed sets in any collection of closed sets is closed, and the union of any finite number of closed sets is closed. This problem can be proved easily by using the previous three problems. It can also be proved using directly the definitions of “open” and “closed.” You should learn how to handle both proofs.

5.1.11 The union of an infinite number of closed sets may fail to be closed. Such a denial of a general statement requires an example. Give one.

5.1.12 Give several examples of sets which are neither closed nor open.

5.2 A brief discussion of topology

Interesting in itself, the discipline of topology provides a large number of abstractions and generalizations which are quite useful in analysis. A very brief introduction to topology and some of its basic subject matter will therefore be useful for us. It goes without saying that a deeper study of topology is useful, too, and is highly recommended. What will be covered in the next few sections will definitely not serve as a substitute for the serious study of topology. Rather, these coming sections will serve to introduce some needed perspective, some needed generalizations of what we have done in the previous section, and some concepts and definitions which will serve us well in our further study of analysis.

First, we must define what is a topology.

Given a set X , a collection \mathcal{T} of subsets of X is called a *topology* on X if

- $X \in \mathcal{T}$, and $\Phi \in \mathcal{T}$
- $(\bigcup_{\alpha \in A} Y_\alpha) \in \mathcal{T}$ provided that $Y_\alpha \in \mathcal{T}$ for each $\alpha \in A$.
- $(\bigcap_{k=1}^n Y_k) \in \mathcal{T}$, provided that $Y_k \in \mathcal{T}$ for each $k \in \{1, \dots, n\}$.

A set X with an associated topology \mathcal{T} defined upon it is often referred to as a **topological space**. Often, too, when one wishes to emphasize the presence of a topology, or even more so when one wishes to refer to a specific topology upon the set X , then one refers to the “pair” (X, \mathcal{T}) . The implication is that a given set X could have more than one topology defined upon it. Upon reflection, this implication is certainly true, too, because the above definition of what a topology upon X needs to satisfy has indeed a short list of the requirements.

As has already been intimated, the above definition is the beginning of the mathematical area called General Topology. Among other questions of interest in that area, the properties of a given topology on a given set X are described and analysed. Much of what we have done here specifically for \mathbf{R} and its subsets can be and is abstracted to this very general context, in which it is assumed that we have some set X with a topology defined on it. As already stated, there may be many ways to define a topology upon some given set. Thus, the topology defined upon the real number system which is described in the previous section of this chapter is **the usual topology on \mathbf{R}** . It is certainly possible to define other topologies upon \mathbf{R} , and we will see an example or two, later on.

5.2.1 Base for a topology

The reader may note that method of defining the open sets which comprise the usual topology on \mathbf{R} , laid down in Section 5.1 is not the method which was taken in the previous section. Indeed, the method which we used in Section 5.1 was in a sense opposite to the general definition given in Section 5.2. Recall that in Section 5.1 we defined first the concept of an interior point and an accumulation point, and after that a subset of \mathbf{R} was said to be open if all of its elements are interior points and was said to be closed if it contained all of its accumulation points. Then it was shown that the resulting open sets satisfy the properties which define a topology on \mathbf{R} . And moreover the closed sets thus defined were seen always to be the complements of open sets. A natural question is whether this procedure, too, can be generalized in order to generate a topology on a set X . To this end, let a given, specific collection \mathcal{B} of subsets of X be given. This collection of sets will be called a **base** if it satisfies the two conditions

- B1. The original set X is contained in the union of all the sets in \mathcal{B} .
- B2. If B_1 and B_2 are any two sets from \mathcal{B} , then for each x in their intersection there is a set $B_3 \in \mathcal{B}$ which contains x and is contained in $B_1 \cap B_2$.

Given a base \mathcal{B} of subsets a topology on the set X can then be generated in the same manner as we have generated the usual topology on \mathbf{R} in Section 5.1. There, recall, we started with the collection of all symmetric open intervals of the form $(x - \delta, x + \delta)$, a collection which satisfies the properties of a base. Then we defined the concepts of interior point and accumulation point, and we defined a set to be open if it contains all its interior points, or, alternatively, closed if it

contains all of its accumulation points. This type of construction very much continues to be relevant in a more general context.

Exercises

5.2.1 (trivial) Given a topological space (X, \mathcal{T}) , the collection \mathcal{T} is a base for itself.

5.2.2 Given a set X and a base \mathcal{B} , then the topology on X which is generated by \mathcal{B} comprises the intersection of all those topologies which could be defined upon X which contain the collection \mathcal{B} .

5.2.3 The collection of all intervals of the form $(x - \delta, x + \delta)$ is a base for the usual topology on \mathbf{R} , as claimed above this problem. Also, if we require $\delta > 0$ to be rational we still have a base for the usual topology.

5.2.4 The collection of all open intervals is a base for the usual topology on \mathbf{R} . Even more interesting, the collection of all open intervals *which have rational endpoints* is a base for the usual topology on \mathbf{R} , as well.

5.2.2 Functions and Continuity

If X and Y are two sets for which each has a topology defined on it, and $f : X \rightarrow Y$ is a function between the two, then f is called **continuous** if it is true that the inverse image under f of every open subset in Y is an open subset of X . That is, if $f^{-1}[Q]$ is open in X whenever Q is open in Y .

Remark:

We have already discussed the definition of continuity of a real-valued function which is defined upon a subset of \mathbf{R} toward the end of Chapter 1. The definition was given there primarily as a case study in the use of quantifiers and in the negation of statements with quantifiers in them. As mentioned back in Chapter 1, the definition given there is the standard definition for the continuity of a real-valued function of a real variable. The definition from general topology, as given above, is in fact in full accord with the definition given in Chapter 1 where we assumed that X and Y are subsets of \mathbf{R} and their respective topologies are determined by the “usual topology” upon \mathbf{R} . **The definition of continuity which is given in the typical calculus book is not equivalent to the definition just above, which is why the definition which is in the calculus book is not used anywhere in mathematics outside of the calculus course.** The brief discussion of continuity which is touched on in this paragraph will be continued in the next chapter, in much greater detail.

5.2.3 Topological Properties

Now, continuing the brief discussion of general topology, what are called **topological properties** are those properties of any set X (with a given topology on it, of course) which are preserved by any continuous and invertible function to a set Y (with a topology on it), for which the inverse function of f is also continuous. Such functions are called **homeomorphisms**. It should be clear that when such a function is seen to exist between two sets X and Y with a topology upon each, then from the topological point of view the two sets along with their respective topologies, which

we say are **homeomorphic**, are in some way essentially “equal.” Some topological properties are described in the rest of this chapter. Connectedness and compactness, which will be defined quite soon, are among the important topological properties of certain subsets of real numbers and are also important concepts in general, in the study of topological spaces (sets on which a topology has been defined).

Interestingly, though, and rather strangely given its extreme importance in construction and the further study of the real number system and given the fact that it will come up again even in further discussions in this chapter dealing with topological ideas, completeness is **not** a topological property. A very easy counterexample will suffice to show that the completeness property does not need to hold in two sets X and Y with topologies on them, even when there is a homeomorphism between the two.

Consider the two functions $f(x) = \tan x$, using as the domain only the interval $(\frac{-\pi}{2}, \frac{\pi}{2})$, and its inverse function $\arctan x$ defined on \mathbf{R} . These are, both of them, continuous functions. Moreover, in \mathbf{R} every Cauchy sequence converges, which we have taken as the defining property of completeness. But the interval $(\frac{-\pi}{2}, \frac{\pi}{2})$ is clearly not complete. For, any sequence whose terms are all in the interval $(\frac{-\pi}{2}, \frac{\pi}{2})$ but which is converging to $\frac{\pi}{2}$, for example, is clearly a Cauchy sequence. But its limit is not in the interval $(\frac{-\pi}{2}, \frac{\pi}{2})$, which therefore is not complete.

5.2.4 Relative Topologies

Finally, we mention a concept which is a very natural one, but its use makes much easier the application of many topological concepts. This is the concept of a **relative topology**. Consider a topological space (X, \mathcal{T}) , and let S be any

subset of X . Then the relative topology on S is the topology which consists of all subsets of S which can be represented as the intersection of S with some set from \mathcal{T} .

Exercises

5.2.5 Let X be a topological space with topology \mathcal{T} , and let S be a subset of X . Then the relative topology on S is indeed a topology (that is, the relative topology on S satisfies the defining properties of a topology).

5.2.6 Describe the relative topology induced on a closed interval $[a, b]$ by the usual topology on \mathbf{R} .

5.2.7 The statement that a real-valued function f of a real variable is continuous has been defined in terms of an ϵ and a δ , at the end of Chapter 1, and the definition given in Chapter 1 will be further discussed in the next Chapter. Show that f satisfies this ϵ and δ definition at every point in its domain if and only if it is true that $f^{-1}[Y]$ is an open subset of the domain of f whenever Y is an open subset of \mathbf{R} . This problem assumes, of course, that we are using the usual topology on \mathbf{R} , and that the topology defined on the domain of f is the relative topology upon that set.

5.2.8 Let (X, \mathcal{T}) and (Y, \mathcal{U}) be two topological spaces, and let $f : X \rightarrow Y$. Let \mathcal{B} be a base for the topology on X and let \mathcal{C} be a base for the topology on Y . Then the following statement, one would think, ought to be equivalent to the statement that f is continuous at some $x_0 \in X$:

Given any $C \in \mathcal{C}$ such that $f(x_0) \in C$, there exists a set $B \in \mathcal{B}$ such that for all $x \in B$, $f(x) \in C$. Show that f is

continuous at each $x \in X$ under this definition if and only if f is a continuous function from (X, \mathcal{T}) to (Y, \mathcal{U}) , in the sense of the definition given in Section 5.2.2.

5.3 Connectedness of intervals

A set S in a topological space X is said to be **connected** if S cannot equal the union of two disjoint, nonempty sets which are open in the relative topology on S .

An important property of \mathbf{R} itself is that \mathbf{R} cannot be represented as the union of two non-empty disjoint open sets. We say, then, that \mathbf{R} is *connected*. You are asked to prove this statement in an Exercise immediately below.

For any subset S of \mathbf{R} , we say that S is *connected*, provided that there do not exist two open sets Y_1 and Y_2 , such that

$$(Y_1 \cap S) \cup (Y_2 \cap S) = S$$

and at the same time both $Y_1 \cap S \neq \Phi$ and $Y_2 \cap S \neq \Phi$.

Exercises

5.3.1 \mathbf{R} is connected.

5.3.2 Any subinterval of \mathbf{R} is connected.

5.3.3 Any connected subset of \mathbf{R} is either a singleton set or an interval.

5.4 Heine-Borel Theorem, Bolzano-Weierstrass Theorem, and Nested Closed Interval Theorem

We say that a collection of sets $\{X_\alpha | \alpha \in A\}$ is a *covering* for a set S if

$$S \subset \bigcup_{\alpha \in A} X_\alpha$$

We say that the collection is an *open* covering for S , if each set in the collection is an open set. If $B \subset A$ and

$$S \subset \bigcup_{\alpha \in B} X_\alpha$$

then the collection $\{X_\alpha | \alpha \in B\}$ is called a *subcovering*. If the set B is finite, then we have a *finite* subcovering.

The Heine-Borel theorem states:

Theorem: *Every open covering of a closed and bounded interval $[a, b]$ has a finite subcovering.*

A related result is the Bolzano-Weierstrass theorem, which states:

Theorem: *Every bounded infinite subset of \mathbf{R} has an accumulation point.*

Yet another related result is the Nested Closed Interval Theorem, which states:

Theorem: *Let $\{I_n\}_{n=1}^{\infty}$ be a sequence of bounded closed intervals such that $I_{n+1} \subset I_n$ for $n = 1, \dots$. Then*

$$\bigcap_{n=1}^{\infty} I_n \neq \Phi.$$

In fact, each of these statements is related in an intimate manner to the Statement of Completeness for the real numbers, which we have already seen can take several equivalent forms.

Exercises

5.4.1 Give a proof for the Heine-Borel Theorem.

5.4.2 Give a proof for the Bolzano-Weierstrass Theorem.

5.4.3 Give a proof for the Nested Closed Interval Theorem which directly uses the Heine-Borel Theorem instead of using the Statement of Completeness (see Exercise 3.5.4).

5.4.4 The Heine-Borel Theorem, the Bolzano-Weierstrass Theorem, the Nested Interval Theorem, and the Statement of Completeness are equivalent to one another, if the underlying set being described by them is \mathbf{R} .

A set for which the Heine-Borel theorem is true is referred to as **compact**. As mentioned above, to be compact or not compact is a topological property of the given set, and completeness is not. Clearly, \mathbf{R} is not compact, but it is complete. But \mathbf{R} is called **locally compact** because of what is said in the following problem.

Exercises

5.4.5 A subset of \mathbf{R} is compact if and only if it is closed and bounded.

5.5 Metric Spaces

A set S with a topology defined upon it is called a **metric space** if the topology upon it is defined by a metric.

A **metric** is a real-valued function d defined upon $S \times S$ which has the following properties:

m1. $d(x, y) \geq 0$ for all $x, y \in S$, with $d(x, y) = 0$ if and only if $x = y$.

m2. $d(x, y) = d(y, x)$ for all $x, y \in S$

m3. $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in S$ (often referred to as the Triangle inequality).

A moment's reflection should convince the reader that the definition of a metric is essentially abstracted from the concept of distance. When one has a set S with a metric d defined upon it, it is possible to define very easily an associated **metric topology**. We need only to define the **open ball at x of radius δ** as

$$B(x, \delta) = \{s \mid d(x, s) < \delta\}$$

and then a set is open in the metric topology on S if and only if it is equal to the set of its interior points. Quite unsurprisingly, a point x is an interior point of some subset X of S means that there exists some $\delta > 0$ such that $B(x, \delta) \subseteq X$. The way to extend or adapt the definitions of such things as accumulation points, limit points, and closed set is similarly unsurprising. The fact that the collection \mathcal{T} of all sets which are open under the definition just now given is indeed a topology on the set S ought at this point to be unsurprising, too. What we have done, after all, is to construct a base for the topology on the metric space.

Exercises

5.5.1 Justify the remark just above, that the collection of all the open balls $B(x, \delta)$ is a base for a topology on a metric space. Show that as we have defined things the ball $B(x, \delta)$ is an open set, in the sense that every point in it is an interior point.

Finally, two miscellaneous observations:

A set with a topology defined upon it is said to be **metrizable** provided that it is possible to produce a metric upon the set which reproduces the given topology. Criteria which guarantee metrizability are interesting.

A function p which satisfies all the properties of a metric, except only that $p(x, y) = 0$ may not imply $x = y$ is called a **pseudometric**.

Exercises

5.5.2 Let S be any set. then the collection of subsets of S containing exactly the two sets Φ and S is a topology upon S . This topology is not metrizable unless S contains exactly one element. It is often referred to as the “trivial” topology or as the “indiscrete” topology on S . Needless to say, this is not a very interesting topology.

5.5.3 Let S be any set. Then the power set of S is a topology upon S , and that topology is always metrizable. It is often called the “discrete” topology on S for the obvious reason that every subset of S is open, including in particular any subset consisting of only one point. In spite of the fact that this topology is metrizable (how?), it is not a very interesting topology, either.

5.5.4 Is a closed and bounded subset of a metric space necessarily compact? It is clear that something like this ought to be true. But what that something is, is not exactly obvious and would take us too far afield. You are encouraged to take a course in general topology, where the matter will be addressed. To get some idea of what the

problem is, consider that the underlying set S could be an open interval and that the usual topology upon that open interval (inherited from the usual topology on \mathbf{R}) clearly turns the open interval into a metric space. The same open interval is then bounded by its length, and clearly it is closed as a subset of itself. However, these properties do not cause the open interval to become compact. Give an actual example that demonstrates what the problem is.

5.5.5 This example is intended to overcome some of the everyday prejudices and unconscious hidden assumptions which can bite us from behind when we are trying to do mathematics. Suppose that in \mathbf{R} we define the distance between any two non-zero numbers x and y to be $|x - y|$, and the distance between 0 and any non-zero number x to be $|x| + 1$, and the distance from 0 to itself to be 0. Show that the resulting distance function is a metric on \mathbf{R} .

Needless to say, the metric just defined is quite incompatible with the algebraic operations and the order relations defined upon \mathbf{R} . But that is the whole point of the example, actually.

5.5.6 Show that \mathbf{R} , under the rather weird topology defined in Problem 5.5.5, is not connected.

5.6 A useful collection of general results

Topology, as already seen, is a discipline which abstracts the basic properties of open sets and continuous functions. By doing this job for us, it makes possible many sweeping statements which are quite useful. Most of the properties which are listed in the following problems are not difficult to prove, in the context of general topology. Nevertheless,

the properties are very important because they occur again and again in many contexts. In particular, we will find immediate uses for almost all of them in future chapters of these lecture notes.

Exercises

5.6.1 A subset S of a compact set X is also compact if and only if S is a closed subset of X .

5.6.2 The continuous image of a compact set is compact. That is, if (X, \mathcal{T}) is a topological space and (Y, \mathcal{U}) is another, and if $f : X \rightarrow Y$ is continuous, then the set $f[X]$, the image of X , is compact in Y .

5.6.3 The continuous image of a connected set is connected. That is, if (X, \mathcal{T}) is a connected topological space and (Y, \mathcal{U}) is another topological space, and if $f : X \rightarrow Y$ is continuous, then the set $f[X]$, the image of X , is connected in Y .

5.6.4 Let f be a continuous real-valued function defined on a compact set (X, \mathcal{T}) . Then the set $f[X]$ is compact, and hence both closed and bounded. Consequently, the set $f[X]$ contains a maximum and a minimum.

5.6.5 The min-max and intermediate value theorems (cf. Exercises 6.3.12 and 6.3.13) for a continuous real-valued function defined on a closed and bounded interval $[a, b]$ will follow, once it has been established that the interval $[a, b]$ is both compact and connected. This problem is revisited in Chapter 6, which contains a more in-depth discussion of continuous real-valued functions of a real variable.

The apparent simplicity of the previous problem is in fact somewhat superficial. The heavy lifting is in showing that the closed and bounded interval is both compact and connected in the usual topology on \mathbf{R} , after all. Moreover, neither to prove that a closed and bounded interval is compact, nor to prove that it is connected, can be done by abstraction and generalization alone.

Chapter 6

Real-valued functions of a real variable, limits, and continuity

6.1 The definition of a function

Much of what follows in this section has already been discussed briefly in Chapter 1. However, we have not done much with functions in the meantime, and so it is perhaps good to review these topics.

Let S and X be any two (non-empty) sets. A function from S to X is a rule which associates to each $s \in S$ a unique element $x \in X$. The rule or procedure for choosing x , given s , can be given a name or symbol, for example f . Then, we write $f(s) = x$. The set S is then called the *domain* of f , and X is called the *range* of f , and the subset $f[S] = \{x \in X | f(s) = x, \text{ some } s \in S\}$. is called the *image* of f .

Remarks:

- The reader is hereby warned that the terminology used here is not completely standard in all areas of mathematics, or even in all treatments of the same subdiscipline in mathematics and, because of this, even these lecture notes themselves might occasionally lapse from pure consistency. Some people, for example, would call the set $f[S]$ the “range” and de-emphasize the role of the set X completely. This is typical of algebra and calculus books, for example, in giving instructions which read “Find the domain and the range of the following functions.” The only thing I can say is, when you encounter this situation (and you will or already may have!), you will need to adjust to it.
- When we are dealing with real-valued functions of a real variable, the set S , which we have called the domain, may or may not be explicitly given. Nevertheless, it is an essential part of the definition of a function. We say more about this matter in the continuing discussion.

Often, the domain of a real-valued function of a real variable is not explicitly given. Then, we could say that the domain is given *implicitly* and, for example, expect students in a College Algebra course to find it. Suppose we gave

them a function such as

$$f(x) = \frac{\sqrt{x^2 - 1}}{\sqrt{x - 1}}$$

and asked them to find the domain. We would mean by this that they should find all values of x for which the computation can be performed without impediment.

Exercises

6.1.1 Come to think of it, what **is** the domain of that function?

On the other hand, the same algebraic formula can be used in different contexts to mean entirely different things, and the results can be quite different.

Example 1:

Consider $f(x) = x^2$. The natural implicit domain of this function is clearly $(-\infty, \infty)$. This is a simple example, too, of a function which is not *invertible*, by which we mean that there is no function g such that $g(f(x)) = x$ for each x in the domain of f .

On the other hand, consider the formula $A = A(s) = s^2$. Obviously, we are concerned here with computing the area of a square of side length s . And squares do not have sides of negative length. Therefore, the formula could equally well state that, given a square of area A , we could compute $s = s(A) = \sqrt{A}$. The function $A(s)$ is therefore invertible

and behaves in this essential respect in a manner entirely different from $f(x) = x^2$. The reason for this is that $A(s)$ has domain $\{s|s > 0\}$ (or we could argue about $s = 0$, too, if we wanted to get into a quibble). Whatever the outcome of the quibble, the point is that two functions may have the same algebraic formula but be different because they have different domains.

Example 2:

It is also possible to state the domain from the outset. We could (and soon will) state that we are interested in functions defined on a certain, given subset of \mathbf{R} , such as some interval I . Then, the domain is declared to be I , and the issue of whether a particular function might be given by a formula which may or may not be possible to use upon some number outside of I is totally irrelevant. We have declared the interval to be I , and we are not concerned about other things. When the domain of immediate interest is a proper subset of the “natural” domain of a function, the terminology “restriction of the domain” is in common use.

Example 3:

Consider the interval $[0, 1]$ and the function

$$f(x) = \frac{1}{x - 4}$$

defined upon that interval. Then, upon that interval, the function is continuous, has a maximum value and a minimum value. The fact that the function behaves very badly indeed in the close vicinity of $x = 4$ is not relevant to the discussion. The domain is $[0, 1]$. We said so. On the other hand, the function

$$g(x) = \frac{1}{x}$$

is **not** defined upon the interval $[0, 1]$.

6.2 Limits of real-valued functions of a real variable

The statement that

$$\lim_{x \rightarrow a} f(x) = L$$

means that a is an accumulation point of the domain of f and that, given any $\epsilon > 0$, there is $\delta > 0$ such that, for all x in the domain of f , $0 < |x - a| < \delta$ implies that $|f(x) - L| < \epsilon$.

It should be noted that this definition of “limit” is **not** the same definition as that used in introductory calculus books.

Some calculus books are very sloppily written and do not trouble themselves to mention any relationship whatsoever between the point a in the definition and the domain of f . This cavalier handling of the limit definition opens the door to logical absurdities and is therefore dangerous. Such sloppiness is often regarded as a valid reason for a textbook committee to reject the book, no matter what other merits the book might or might not have. The next exercise demonstrates what can follow from such a messed up definition.

Exercises

6.2.1 Explain the following:

If one were to use a “definition” of $\lim_{x \rightarrow a} f(x)$ which neither requires nor logically implies that the point a is an accumulation point of the domain of f , then the obviously nonsensical statement that

$$\lim_{x \rightarrow 3} \sqrt{1 - x^2} = 29$$

would logically follow from the messed-up “definition,” and moreover under that same “definition” the same limit would also be equal to 37, or to any other number one might imagine. Even worse, it would also follow from that “definition” that the moon is made of green cheese and cows jump over it.

The embarrassing conclusions of the previous problem follow, of course, because one can take $\delta = 1$ in the limit definition and then the final if-then statement is vacuously satisfied. To avoid such embarrassment, most calculus books require a much stronger condition, namely that for some $\delta > 0$ there exists an interval $(a - \delta, a + \delta)$, such that every point in it, except perhaps for a itself, is contained in the domain of f . This requirement is, however, excessively stringent. For, it requires the special treatment of points which should not require special treatment, as may be seen in the next exercise.

Exercises

6.2.2 Under our definition of limit, which requires the point a to be an accumulation point of the domain of the function f , it is true that

$$\lim_{x \rightarrow 1} \sqrt{1 - x^2} = 0.$$

However, if the definition of limit requires the “stronger condition” that there must exist some interval $(a - \delta, a + \delta)$, such that every point in it, except perhaps for a itself, is contained in the domain of f (as is done in some calculus books), then the above limit does not exist (Why not? Explain.) and one is immediately forced to speak of the “left-hand limit” instead, in order to talk about any kind of limit at all for this expression as $x \rightarrow 1$.

6.2.3 Not to assume from the previous problem that the left-hand limit and the right-hand limit do not have their appropriate uses. They do. Construct correct definitions for the left-hand limit

$$\lim_{x \rightarrow a^-} f(x)$$

and the right-hand limit

$$\lim_{x \rightarrow a^+} f(x)$$

which are in the spirit of our definition for

$$\lim_{x \rightarrow a} f(x).$$

6.2.4 Show that if both of

$$\lim_{x \rightarrow a^-} f(x) \text{ and } \lim_{x \rightarrow a^+} f(x)$$

exist and they are both equal to a common value L , then also

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

Also show that

$$\lim_{x \rightarrow a} f(x) = L$$

implies only that at least one of

$$\lim_{x \rightarrow a^-} f(x) \text{ and } \lim_{x \rightarrow a^+} f(x)$$

must exist and be equal to L . If both of them exist, then both of them must equal L . But one of the two may not exist.

6.2.5 Show that, if

$$\lim_{x \rightarrow a^-} f(x) = \lim_{x \rightarrow a^+} f(x) = f(a).$$

then also

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

There is, of course, a reason why we do not tell our calculus students the truth about the definition of limit. The reason is that first they would have to understand what an accumulation point is, and rightly or wrongly we do not want to tell them. But this decision has bad consequences as well as good, in that other, related topics become more convoluted.

Exercises about limits follow. In these exercises we will assume that we have two functions f and g , such that

$$\lim_{x \rightarrow a} f(x) = L \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = M.$$

Exercises

6.2.6 Under the right circumstances involving the domains of f and of g and the point a , it is the case that

$$\lim_{x \rightarrow a} (f + g)(x) = L + M$$

Describe the “right circumstances” and prove the result. Hint: The “right circumstances” involve some attention to the domain of the function $f + g$.

6.2.7 Similar problem about $f \cdot g$.

6.2.8 Similar problem about f/g . And what else do we need here as a minimal assumption?

6.2.9 State an appropriate definition for

$$\lim_{x \rightarrow \infty} f(x) = L.$$

Your definition has to contain a requirement analogous to “the point a is an accumulation point of the domain of f ” which is applicable in this context.

6.2.10 State an appropriate definition for

$$\lim_{x \rightarrow \infty} f(x) = \infty.$$

Again, your definition has to contain a requirement analogous to “the point a is an accumulation point of the domain of f ” which is applicable in this context.

6.2.11 Show that if $\lim_{x \rightarrow a} f(x) = L_1$ and $\lim_{x \rightarrow a} f(x) = L_2$, then $L_1 = L_2$.

Also explain why this statement is nonsense and can neither be proved nor disproved if the definition of “limit” were not correctly formulated in the first place. To do this part of the problem, consider what happened in Exercise 6.2.1 and explain why similar problems would arise here.

6.2.12 (problem on substitution in limits) Suppose that $\lim_{x \rightarrow b} h(x) = a$ and that g is some other function. We define the composite function $g \circ h$ by $(g \circ h)(t) = g(h(t))$.

- (a) Using the notations that the domain of h is $D(h)$ and the image of h is $D(h)$ and corresponding notation for g , describe the conditions which must be satisfied in order that the function $g \circ h$ can exist. In particular, describe the largest possible domain for $g \circ h$.
- (b) Let it be assumed that $\lim_{t \rightarrow b} (g \circ h)(t) = L$. Give the conditions which must be satisfied in order that this statement is true. In particular, what condition does this impose upon the point b ?
- (c) Give the conditions which must be satisfied in order that $\lim_{x \rightarrow a} g(x) = L$. In particular, what does this require in regard to the point a ?

- (d) What is required in order that $\lim_{t \rightarrow b} (g \circ h)(t) = \lim_{x \rightarrow a} g(x)$? In particular, what conditions does this impose upon the points a and b ? Are the conditions upon the point a identical to those which were required in part (c)?
- (e) Can it happen that $\lim_{t \rightarrow b} g(h(t))$ and $\lim_{x \rightarrow a} g(x)$ both exist, but the two are not equal? If so, then describe the circumstances and give an example.
- (f) Can it happen that $\lim_{t \rightarrow b} g(h(t))$ exists but $\lim_{x \rightarrow a} g(x)$ does not exist? If so, then describe the circumstances and give an example.

Concluding Remark: Our assumptions in the definition of a limit about the proper relationship of the point a to the domain of f are now in place. One does need to be really careful about these assumptions and about the resulting theorems. Wouldn't everyone agree?

6.3 Continuity of a function

We say that a real-valued function f defined upon a subset S of \mathbf{R} is **continuous at** a , provided that given any $\epsilon > 0$, there is $\delta > 0$ such that for all x in the domain of f , the condition $|x - a| < \delta$ implies $|f(x) - f(a)| < \epsilon$. We further say that the function f is **continuous on** S if f is continuous at each $x \in S$.

Remark: As was the definition of limit different from what is usually presented in introductory calculus, so is the definition of continuity different. To see what one of the major differences is, suppose that a is an *isolated point* of S , the domain of f . That is, suppose that there is some $\delta > 0$ such that $(a - \delta, a + \delta)$ contains no point of S other than a itself. Then f is ipso facto continuous at a according to our definition. But, all calculus books say that f is continuous at a if $\lim_{x \rightarrow a} f(x) = f(a)$.

For those calculus books which require that the function f must be defined at all points in a set $(a - \delta, a + \delta) \setminus \{a\}$ in order for $\lim_{x \rightarrow a} f(x)$ to exist, that limit cannot possibly exist if a is an isolated point of S . That is appropriate. Our definition of limit, of course, does not allow the limit to exist as $x \rightarrow a$, either. So far, so good.

But the calculus book then goes on to define “continuous at a ” in terms of “has a limit at a .” If the definition of limit has required that there must exist a set $(a - \delta, a + \delta) \setminus \{a\}$ contained completely within the domain of f , then no function f can be continuous at a point a if a is an isolated point of its domain. For that matter, a function such as $f(x) = \sqrt{1 - x^2}$ can not be continuous at the points ± 1 either unless some fine print is immediately added to the definition of continuity to cover such cases. In other words, that book’s too-restrictive definition of what a limit is has caused further unnecessary complications in the definition of continuity.

Even worse than the above, if the calculus book is one of those irresponsible ones which have cavalierly put no restriction at all upon the point a in its definition of limit, then the situation descends into complete chaos. We have already seen that if the point a is not in the domain of the function f , then $\lim_{x \rightarrow a} f(x)$ can be anything whatsoever, and it is also logically implied that the moon is made of green cheese. Also, for that matter, it is implied that $f(a)$ exists

even if it does not, and that it is not equal to itself even if it does exist. The situation is equally bad if a is an isolated point of the domain of the function, and for exactly the same reasons that apply if it is an isolated point which is not in the domain; the limit in question is not uniquely defined and is both equal to and not equal to $f(a)$. If one wants to be completely logical, as the subject matter obviously demands from us, then a definition which leads to such logical absurdity is obviously not in the best interest of anyone at all.

However, since our definition of the continuity of f at a is *independent* of our definition of $\lim_{x \rightarrow a} f(x)$, the continuity of a function at an isolated point of its domain is automatic and has nothing to do with any limit. You will be asked to prove this below, in Exercise 6.3.8.

As with the definition of limit, there are also reasons why the definition of continuity given here is the standard definition used by mathematicians. The points of difference between the definition presented here and the definition presented in the better class of calculus texts relate to the different treatment of isolated points of the domain of a function and to the special treatment which the calculus book must use at points where the function is in fact defined only on one side of the point. Because of these differences, the calculus book's definition leads to additional complications. Our definition, in contrast, is simple and clear. Furthermore (as you will be asked to prove in Exercise 6.3.7), when the topology on \mathbf{R} is the usual topology, which was defined in Section 5.1, our definition is logically equivalent to the definition which was given in Section 5.2.4, that a function is continuous if and only if the inverse image of every (relatively) open set in the image (or in the range, which could be any set containing the image!) is (relatively) open in the domain. That definition of continuity is a general statement and covers all conceivable situations where one might

want to discuss continuity. But if we were to settle for other definitions which could disagree with that one in this or that small particular or in this or that context, then we might end up with a very long list of small particulars. That would unnecessarily complicate our lives and complicate our work. Thus, the definition which has been presented here for the continuity of a real-valued function of a real variable is the standard mathematical definition of continuity, used everywhere else in mathematics other than in calculus texts.

Exercises

6.3.1 As already emphasized, more responsible calculus books require as part of the definition of

$$\lim_{x \rightarrow a} f(x)$$

that there must exist some $\delta > 0$ such that $(a - \delta, a + \delta) \setminus \{a\}$ is a subset of the domain of f .

These more responsible calculus books also say that f is continuous at a provided that

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

The irresponsible calculus books generally say this, too, but their “definition” of what is a limit omits something really essential, as we have seen, and thus does not make any sense to start with.

- (a) Show that, if f is continuous at a according to the definition used in the more reasonable and careful calculus books, then it follows that f is continuous according to our definition (stated in the first paragraph of this Section).
- (b) Also show that the converse of this statement in part (a) is not always true. That is, our definition that f is continuous at a does not necessarily require that

$$\lim_{x \rightarrow a} f(x) = f(a)$$

must be true.

6.3.2 The function f can be continuous at a point a and at the same time all three, or any two, or any one, or none of the expressions

$$\lim_{x \rightarrow a^-} f(x), \quad \lim_{x \rightarrow a^+} f(x), \quad \lim_{x \rightarrow a} f(x),$$

must necessarily exist. Describe the circumstances in which each of these situations can occur.

6.3.3 If

$$\lim_{x \rightarrow a^-} f(x) = \lim_{x \rightarrow a^+} f(x) = f(a).$$

then it follows that f is continuous at a .

6.3.4 If the point a is an interior point of the domain of f then indeed all three of the limits

$$\lim_{x \rightarrow a^-} f(x), \quad \lim_{x \rightarrow a^+} f(x), \quad \lim_{x \rightarrow a} f(x),$$

must exist if f is continuous at a , and indeed all three of them must be equal to $f(a)$.

6.3.5 If it is given that the function f is continuous at a and a is an accumulation point of the domain of f intersected with $(-\infty, a)$, then

$$\lim_{x \rightarrow a^-} f(x) = \lim_{x \rightarrow a} f(x) = f(a).$$

6.3.6 If it is given that the function f is continuous at a and a is an accumulation point of the domain of f intersected with (a, ∞) , then

$$\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a} f(x) = f(a).$$

6.3.7 In the context of the usual topology on the set of real numbers, which was described in Section 5.1, prove the equivalence of our definition of continuity for a real-valued function of a real variable and the topological definition, as advertised in the paragraph just before this problem section. Note: this problem requires definition of the concept of a “relative topology” or “induced topology.” The definition was introduced in Section 5.2.4 and may be restated here as follows:

Let S be a subset of a set X which has a topology defined on it. Then a subset Y of S is *relatively open in S* provided that Y can be written as $Y_1 \cap S$, for some set Y_1 open in X .

In Exercise 5.2.5, you have been asked to show that the relative topology on the set S is indeed a topology on S , which we could say is ‘inherited’ from or induced by the (previously existing) topology on X .

6.3.8 Let S be any subset of \mathbf{R} , and f be any function whose domain is S . Then (according to our definition of continuity) f is automatically continuous at any isolated point of S . But according to the standard calculus book definition of continuity, which requires that f must be defined on an open interval $(a - \delta, a + \delta)$, no f can be continuous at any isolated point in S . This problem continues the discussion above. Produce any details of the proofs which were not addressed in that discussion.

6.3.9 According to our definition of continuity, a sequence is automatically continuous at every positive integer, whereas according to the standard calculus book definition of continuity this cannot be the case.

6.3.10 Suppose that f and g are two functions which are continuous at a . Then their sum and product are continuous at a . So is the quotient f/g , provided that $g(a) \neq 0$. Prove these statements. Also specify what, if anything, one to assume here about the domains of the functions f and g in order to make these statements to be truthful?

6.3.11 The image of a continuous real-valued function defined upon a finite closed interval is bounded.

6.3.12 The image of a continuous real-valued function defined upon a finite closed interval has a maximum and a minimum.

6.3.13 The image of a continuous real-valued function defined upon a finite closed interval contains every value between its maximum and its minimum.

6.3.14 The previous three problems, taken together, imply the Statement of Completeness for the real number system (see the Section 5.4 about the Heine-Borel and Bolzano-Weierstrass and Nested Closed Interval theorems, and recall that we already know several equivalent formulations of the Statement of Completeness, from Chapter 3).

6.4 Uniform continuity

Let the function f be defined upon a set S . Then f is **uniformly continuous on S** means that, given $\epsilon > 0$ there is $\delta > 0$ such that whenever $x, y \in S$ and $|x - y| < \delta$, it is true that $|f(x) - f(y)| < \epsilon$. Note that this definition is similar to the definition that f is continuous on the set S . But the difference between the two definitions lies in the different ordering of the quantifiers. Recall that “there exists for all” is not the same thing as “for each there exists.” The comparison given back in Chapter 1 was between “One ring to rule them all” and “Everybody has a ring.” Big difference. Thus, the effect is to make the definition of uniform continuity a much stronger statement than to say that f is continuous at each individual $x \in S$. Further referring back to Chapter 1 for a comparison, one might look again at Exercise 1.2.14. For, that Exercise in fact dealt with the definition of uniform continuity without actually saying so.

Exercises

6.4.1 A function which is continuous on a bounded closed interval is uniformly continuous on that interval.

6.4.2 If a function is continuous on an interval which is not bounded or is not closed, then the function may or may not be uniformly continuous. Give some examples.

6.4.3 It should be fairly obvious that the definition of uniform continuity may be formulated in the context of a function defined on a metric space, not just if the function is defined on some subset of \mathbf{R} as above. Formulate such a definition.

6.5 Limits superior and inferior

Limits superior and inferior and the related concept of a cluster point pertain both to sequences and to functions. First, let us define what these mean in relation to a sequence.

Let $\{s_n\}$ be a sequence. Then s is said to be a **cluster point** of the sequence if given any $\epsilon > 0$, then for any N there is an $n \geq N$ for which $|s_n - s| < \epsilon$.

Note that a sequence could have many cluster points. The concept of a cluster point is therefore not exactly the same as the concept of a limit.

The limit superior of a sequence $\{s_n\}$ is denoted as $\limsup s_n$ or $\overline{\lim} s_n$. It is defined as $\lim_{n \rightarrow \infty} \sup_{k \geq n} s_k$.

The limit inferior of the sequence $\{s_n\}$ is defined as $\lim_{n \rightarrow \infty} \inf_{k \geq n} s_k$ and is denoted as $\liminf s_n$ or as $\underline{\lim} s_n$.

The first exercise in the series of exercises immediately following ought to help in clarifying what is going on with the limit superior and the limit inferior of a sequence, and it might be quite helpful in doing most of the rest of the problems.

Exercises

6.5.1 Let $\{s_n\}$ be any sequence of real numbers. Let us define two sequences of (possibly extended) real numbers which are related to it. We shall define the sequence $\{U_n\}$ by $U_n = \sup_{k \geq n} s_k$, and we define the sequence $\{L_n\}$ by

$L_n = \inf_{k \geq n} s_k$. Then

- (i.) $\limsup s_n = \infty$ if and only if $U_n = \infty$ for all n .
- (ii.) If $\limsup s_n < \infty$, then $\{U_n\}$ is a non-increasing sequence of real numbers.
- (iii.) $\liminf s_n = -\infty$ if and only if $L_n = -\infty$ for all n .
- (iv.) If $\liminf s_n > -\infty$, then $\{L_n\}$ is a non-decreasing sequence of real numbers.

6.5.2 Let $\{s_n\}$ be any sequence, and let the sequences $\{U_n\}$ and $\{L_n\}$ be defined as in Problem 6.5.1. Then, allowing for the possibilities that a limit can be ∞ or $-\infty$ or some finite number, it is the case that $\lim_{n \rightarrow \infty} U_n$ always exists and is equal to $\limsup s_n$. Similarly, $\lim_{n \rightarrow \infty} L_n$ must always exist and must equal to $\liminf s_n$.

6.5.3 Is the set of cluster points of a sequence necessarily a closed set?

6.5.4 Show that $\limsup s_n$ may alternatively be defined as $\inf_n \sup_{k \geq n} s_k$ or as $\inf_n \sup_{k > n} s_k$.

6.5.5 Show that $\liminf s_n$ may alternatively be defined as $\sup_n \inf_{k \geq n} s_k$ or as $\sup_n \inf_{k > n} s_k$.

6.5.6 Is it true that $\limsup s_n$ is the greatest cluster point of the sequence $\{s_n\}$?

6.5.7 The statement $\lim_{n \rightarrow \infty} s_n = s$ is true if and only if $\limsup s_n = \liminf s_n = s$.

6.5.8 (Root Test for convergence of a positive term series) Let $\sum_{n=0}^{\infty} a_n$ be a series in which $a_n \geq 0$ for all n (or, indeed for all $n \geq N$, some N). If $r = \limsup (a_n)^{\frac{1}{n}} < 1$ then the series converges. (Hint: if the above is true, then the series can be compared directly to a geometric series $\sum \rho^n$ in which ρ is any number such that $r < \rho < 1$).

Moreover, if $r > 1$ then the given series diverges.

Moreover, if $r = 1$ this test provides us with precisely no new information at all concerning the convergence of the given series. Explain why this is true.

6.5.9 Show that the statement “Every Cauchy sequence converges” follows from the results in Problem 6.5.1 and from the statement “Every non-empty subset of the real numbers which is bounded above has a supremum” in the following steps:

First, show that the statement “every non-decreasing sequence which is bounded above must have a limit” must follow from what is given just above. To show this, show that the supremum of the range of the non-decreasing sequence must be its limit.

Second, show that every non-empty subset of the real numbers which is bounded below has an infimum must also follow if we have assumed that every non-empty subset of the real numbers which is bounded above must have a supremum.

Third, show that the statement “every non-increasing sequence which is bounded below must have a limit” must follow. To complete this, it suffices to show that the infimum of the range of a non-increasing sequence has to be its limit.

Now, let $\{s_n\}$ be a Cauchy sequence. Then it is bounded (this is a repetition of a previously given problem). Therefore, the sequences $\{U_n\}$ and $\{L_n\}$ defined in the previous problem, above, both have (finite) limits.

Finally, show that, since $\{s_n\}$ is a Cauchy sequence, it must follow that the statement $\lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} U_n$ is true, or, alternatively, that $\lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} L_n$ is true, establishing that $\lim_{n \rightarrow \infty} s_n$ must exist.

In a similar manner, we can define the limit inferior and the limit superior of a function f at a point a . The limit superior of f at a is usually denoted by $\overline{\lim}_{x \rightarrow a} f(x)$, and it is given by $\inf_{\delta > 0} \sup_{0 < |x-a| < \delta} f(x)$.

The limit inferior of f at a is usually denoted by $\underline{\lim}_{x \rightarrow a} f(x)$, and it is given by $\sup_{\delta > 0} \inf_{0 < |x-a| < \delta} f(x)$.

Finally, p is a cluster point of the range (image) of f at a provided that for all $\epsilon > 0$ and for all $\delta > 0$ there exists $x \neq a$ in the domain of f , such that $|f(x) - p| < \epsilon$.

Exercises

6.5.10 Show that $\overline{\lim}_{x \rightarrow 0} \sin \frac{1}{x} = 1$ and that $\underline{\lim}_{x \rightarrow 0} \sin \frac{1}{x} = -1$ and further that the set of cluster points of the image of this function as $x \rightarrow 0$ is the entire interval $[-1, 1]$.

Chapter 7

Cardinality

7.1 Finite and countable sets

A set is said to be **finite** if it is the empty set (which contains no elements at all), or if it can be put into a one-to-one correspondence with a set of the form $\{1, \dots, n\}$, in which case the given set contains n elements.

A set is called **infinite** if it is not finite.

A set is called **countable** if it is finite, or if it can be put into one-to-one correspondence with the set of natural numbers, \mathbf{N} . It should be obvious from this definition that \mathbf{N} itself is countable, and is not finite. A set which is not countable is called **uncountable**.

Here follows a small group of exercises about countable sets. None of them are very difficult, or at least they do not seem so to the instructor.

Exercises

7.1.1 The set \mathbf{N} itself is countable and infinite.

7.1.2 Any subset of \mathbf{N} which is not finite can be put into one-to-one correspondence with all of \mathbf{N} . Hint: Use the Well Ordering Principle.

7.1.3 Any subset of \mathbf{N} is countable, even if not finite.

7.1.4 The set \mathbf{Z} of all integers is countable.

7.1.5 The set of even integers in \mathbf{N} is countable and infinite.

7.1.6 The set of square integers is countable and infinite (this was explicitly noticed by Galileo, who was probably also not the first to notice it).

7.1.7 Since there is no largest prime number, the set of prime numbers is infinite and is also obviously countable (The proof that there is no largest prime number seems to date back at least to the ancient Greek mathematician Eratosthenes).

7.1.8 Any subset of any countable set is countable.

Less obvious than the proofs of any of the above exercises is the fact that the set of rational numbers, \mathbf{Q} , is countable. As a first step toward proving this, and for the sake of making the argument a little bit simpler, let us prove first that the set of rational numbers which are non-negative is a countable set:

We recall that any rational number $q > 0$ can be given in lowest terms in the form $q = m/n$, where m and n are positive integers. Therefore, the number q can be mapped uniquely to the integer $2^m \cdot 3^n$. By the Prime Factorization Theorem for positive integers, this mapping is one-to-one. In addition, we can let the rational number 0 be mapped to 1. Then the mapping thus defined is one-to-one, in that no two non-negative rational numbers can be mapped to the same integer. Finally, the mapping which we have just now defined may actually be regarded as a sequential counting of the set of non-negative rational numbers by invoking the Well Ordering Principle in order to represent the image of our mapping as a strictly increasing sequence (see Exercise 7.1.2 above, where this argument has already been invoked).

Exercises

7.1.9 Can you find a simple way to adapt the above argument in order to show that the entire set \mathbf{Q} is countable? Hint: If a rational number is negative, then it can be written in lowest terms in the form $-m/n$ in which m and n

are both positive integers. In such a situation, we can map $-m/n$ to the integer $2^m 3^n 5^1$ and the mapping defined above is extended to all of \mathbf{Q} and remains unique.

An alternative method of proving that the set of positive rational numbers is countable is to prove that all of $\mathbf{N} \times \mathbf{N}$ is countable, in which we do not even consider the fact that rational number q is identified with a pair (m, n) that gives q in lowest terms. To see that $\mathbf{N} \times \mathbf{N}$ is countable, visualize the set as graphed using a vertical and horizontal axis. Then one sees the set as represented by a grid of dots in the first quadrant (think of all the places on a piece of standard graph paper where a vertical and a horizontal line intersect). Then, observe that for each positive integer k there are exactly k pairs (m, n) such that $m + n = k + 1$. That is, $(1, 1)$ is the only pair in which $m + n = 2$, and then $(2, 1)$ and $(1, 2)$ are the only two pairs in which $m + n = 3$, and then $(3, 1)$ and $(2, 2)$ and $(1, 3)$ are the only three pairs in which $m + n = 4$, and the same will be true for every value of k . By this method, we can successfully count all of the ordered pairs (m, n) in $N \times N$ in a “diagonal” fashion. Finally, we may invoke the result of Exercise 7.1.8 above, and it follows that the set of all positive rational numbers is also countable because it can be identified with the proper subset of $N \times N$ for which it is true that m/n is a fraction in lowest terms.

Important Remark: Notice that any of the schemes described above for counting the rational numbers will be highly incompatible with the standard order relation defined upon the rational numbers, way back in Chapter 2. But there is no contradiction of what was done back in Chapter 2, either. Here, we are not describing the rational numbers as an ordered set, in which the order relation on the set is compatible with the algebraic operations of addition, subtraction, multiplication, and division. To the contrary, we are only showing how schemes can be set up which will **count** them, which is something entirely different. Indeed, it is surely impossible to perform any counting operation upon \mathbf{Q} which

would preserve the order relation previously defined in Chapter 2. The simple reason is that, in the sense of the ordering in Chapter 2 it is manifestly true that between any two rational numbers there is another one. In the sense of the order relation in Chapter 2, then, there cannot be any such thing as a rational number and then a “next” one. Obviously not. But, again, the sole objective here is merely to show that the rational numbers can be counted.

Exercises

7.1.10 (The “salt and pepper” function) Let any method be established which counts the rational numbers in the interval $[0, 1]$. That is, we can form a sequence $\{q_n\}$ in which every rational number in $[0, 1]$ is equal to q_n for appropriate n . Let the function f_n be defined by $f_n(q_n) = 1$ and $f_n(x) = 0$ for all other values of x in $[0, 1]$. Then the function f defined by $f(x) = \sum_{n=1}^{\infty} f_n(x)$ has the property that $f(x) = 1$ whenever x is rational, and $f(x) = 0$ whenever x is irrational.

7.1.11 Show that the function f defined in the previous problem is continuous at no point in its domain.

7.2 Uncountable sets, and the real numbers are uncountable

In Chapter 1, there was brief mention of the power set of a given set. It was stated there that the power set of the given set is the collection of all subsets of the given set. Let us describe the size of the power set, relative to the given set.

Exercises

7.2.1 Suppose that the set S is finite and contains exactly n elements. Then the power set of S , which we will denote by $\mathcal{P}(S)$, contains exactly 2^n elements.

Here, we can extend this result, showing that the power set of \mathbf{N} itself contains somehow “more” elements than does \mathbf{N} , that is, that $\mathcal{P}(\mathbf{N})$ is uncountable. The way that we do this, of course, is to show that any such counting scheme which is alleged to work, in fact does not work. To show that any such scheme cannot work, it furthermore suffices to show that no such scheme can even complete the counting of the infinite subsets of \mathbf{N} , on the grounds that if those cannot be counted, then it is certainly not possible to count the finite subsets in addition to the infinite ones.

We prove that the collection of infinite subsets of \mathbf{N} is uncountable:

Let S_1, S_2, \dots be any alleged counting of the infinite subsets of \mathbf{N} . For convenience, let us write for each k the set S_k in double subscript notation as a strictly increasing sequence $\{s_{kn}\}_{n=1}^\infty$. Now consider the “diagonal” set D described by the sequence $\{s_{nn}\}$. From D , construct a new set, D_{new} , given by the sequence $\{d_n\}$, in which $d_{11} = s_{11} + 1$ and for each $n > 1$,

$d_n = \max \{s_{nn} + 1, d_{n-1} + 1\}$. Then D_{new} , by its construction, cannot be the same set as any of the sets S_k . For, D_{new} is also given in the form of a strictly increasing sequence of elements d_k , and d_k cannot be equal to s_{kk} for any value of k .

Now, we end with the main result of this chapter of these notes, by showing that the set \mathbf{R} is uncountable. To do this, it clearly suffices to show that the interval $(0, 1]$ is uncountable. In turn, to do this, it is necessary to understand the **binary expansion** of a real number. It is part of the final examination in this class to show that every real number r in the interval $(0, 1]$ can be represented in the form of an infinite series, as

$$r = \sum_{n=1}^{\infty} \frac{c_n}{2^n}, \quad (7.1)$$

in which each c_n is either 0 or 1. The representation (7.1) is unique for “most” values of r . It is part of the final exam to characterize exactly the numbers for which the series representation is not unique. But (spoiler alert) there are exactly two possibilities for non-uniqueness of the expansion for a given real number. The first of the two possibilities is that its expansion (7.1) can start to repeat $c_n = 0$ after a certain point (a terminating expansion). The second possibility is that the expansion (7.1) for the same number r could after a certain point start to repeat $c_n = 1$. As an example, the number 1 can be represented in terminating fashion as, simply, itself (obviously a terminating expansion). But 1 can also be represented in a non-terminating series form as an infinite sum of form (7.1) by choosing $c_n = 1$ for all of the indices n .

Now, let us agree here for our immediate purposes that all terminating expansions are disallowed. If we have thus agreed to exclude all terminating expansions, then the series representation in the form (7.1) given above for each real number in $(0, 1]$ is unique. Moreover, if we have made this agreement then each number in $(0, 1]$ clearly corresponds to the set of indices n in its binary expansion (7.1) for which c_n is 1. Not only this, but also it is true that every infinite subset S of \mathbf{N} corresponds to that real number in $(0, 1]$ whose expansion (7.1) has $c_n = 1$ when and only when $n \in S$.

The proof now follows from what was already proved above, namely that the collection of infinite subsets of \mathbf{N} is uncountable.

Concluding remark: As we have seen here, the entire set \mathbf{Q} of rational numbers is representable as a sequence and therefore is countable. But the set \mathbf{R} of real numbers is not countable, meaning that it is impossible to represent all of \mathbf{R} as a sequence. Moreover, there are also many subsets of \mathbf{R} which cannot be represented as a sequence. For example, no interval in \mathbf{R} can be represented as a sequence. It is therefore factually incorrect to say that an arbitrary subset of \mathbf{R} can be represented as a sequence.

Chapter 8

Representations of the real numbers by expansions

8.1 Introduction

The purpose of this chapter is to present some other standard ways of representing the real numbers. The results contained here will be presented in problem form, as their proofs would depend upon the concepts developed previously

during the course.

8.2 The decimal representation of the real numbers

It is often stated in more elementary courses in mathematics, even at the pre-university level, that the real numbers consist of every number which can be represented as a decimal expansion, whether terminating or non-terminating, and if non-terminating then either repeating or non-repeating. This statement is true, of course. But there are some drawbacks to its use as a basic characterization of the number system. One of those drawbacks is that some numbers have more than one representation. Thus, these numbers would need to be characterized. Another, more serious problem is that true comprehension of the concept of a non-terminating decimal expansion clearly involves at a very basic level the understanding of the concept of a sequence and of the limit for a sequence. It is even better if the student has an understanding of series representations. The results, however, can be summarized in the following problem. The problem is stated only for the interval $[0, 1)$ because the extension of its results to real numbers outside of that interval is presumed obvious.

Exercises

8.2.1 Give a separate proof for each of the following:

- (a) Every infinite series of the form

$$\sum_{n=1}^{\infty} \frac{a_n}{10^n} \tag{8.1}$$

in which the coefficients $\{a_n\}$ satisfy the restriction that a_n will be an integer, and $0 \leq a_n \leq 9$, represents a number in the interval $[0, 1)$, with the exception that if $a_n = 9$ for all n then the sum of the series is exactly 1.

- (b) Every real number in the interval $[0, 1)$ can be represented using a decimal expansion, that is, can be represented by a series of the form (8.1).
- (c) An expansion of the form (8.1) is said to be **terminating** if there is some N such that $a_n = 0$ for all $n > N$. What property characterizes those numbers in $[0, 1)$ which have terminating expansions?
- (d) If a number in $[0, 1)$ has a terminating decimal expansion, show that it also has exactly one nonterminating expansion. Describe how that non-terminating expansion is constructed.
- (e) Show that the decimal representation of a number in $[0, 1)$ is unique if and only if the number has no terminating expansion.

Note: A proper proof of all of the statements in the above problem will require the systematic use of the material in the previous chapters. Naturally, since the description in the problem involves a particular kind of series it is good to know about series. Also useful here is the Statement of Completeness in one or perhaps in more than one of its several equivalent formulations.

8.3 The binary representation of the real numbers

In the above discussion of the representation of the real numbers as decimals, nothing was particularly of interest about the number 10, except that we use it for such purposes by force of habit. Indeed, in some circumstances it is not even advantageous to use 10 as a base at all. Inside a computer, for example, every number is represented by a string of ones and zeroes. This has pretty much seeped into everyday culture at this point, and most people are quite aware of this, at least as far as integers are concerned. Computer scientists, computer engineers, and software engineers are quite accustomed to thinking in base 2 or in base 16 (hexadecimal). But what about numbers between 0 and 1? Can they also be represented in a form similar to the above? Yes, they can. We can't call the results "decimals" of course, because "decimal" means that we are using base 10.

Exercises

8.3.1 Give a separate proof for each of the following:

- (a) Every infinite series of the form

$$\sum_{n=1}^{\infty} \frac{b_n}{2^n} \tag{8.2}$$

using appropriate choices for the coefficients $\{b_n\}$, with the restriction that either $b_n = 0$ or $b_n = 1$ is true for every n , represents a number in the interval $[0, 1)$, with the exception that if $b_n = 1$ for all n then the sum of the series is exactly 1.

- (b) Every real number in the interval $[0, 1)$ can be represented using a binary expansion, that is, can be represented by a series of the form (8.2).
- (c) An expansion of the form (8.2) is said to be **terminating** if there is some N such that $a_n = 0$ for all $n > N$. What property characterizes those numbers in $[0, 1)$ which have terminating binary expansions?
- (d) If a number in $[0, 1)$ has a terminating expansion, of the form (8.2), show that it also has exactly one nonterminating expansion. Describe how that non-terminating expansion is constructed.
- (e) Show that the representation of a number in $[0, 1)$ by a binary expansion is unique if and only if the number has no terminating expansion.

8.3.2 Represent the fractions $1/3$ and $2/3$ and $1/5$ in their binary expansions. Show work, and explain why your work is valid.

8.4 Representation of the real numbers using any base

The above results clearly generalize. Let p be any integer greater than 1. You, the student, probably already know how to write any integer in its expansion base p , and I am assuming that you do. But also the following is true.

In the above discussion of the representation of the real numbers as decimals, nothing was particularly of interest about the number 10, except that we use it for such purposes by force of habit. Indeed, in some circumstances it is not even advantageous to use 10 as a base at all. Inside a computer, for example, every number is represented by a string of ones and zeroes. This has pretty much seeped into everyday culture at this point, and most people are quite aware of this, at least as far as integers are concerned. But what about the numbers between 0 and 1? Can they also be represented in a form similar to the above? Yes, they can.

Exercises

8.4.1 Give a separate proof for each of the following:

(a) Every infinite series of the form

$$\sum_{n=1}^{\infty} \frac{c_n}{10^n} \tag{8.3}$$

in which the coefficients $\{c_n\}$ satisfy the restriction that c_n will be an integer, and $0 \leq c_n \leq p-1$, represents a

number in the interval $[0, 1)$, with the exception that if $c_n = p - 1$ for all n then the sum of the series is exactly 1.

- (b) Every real number in the interval $[0, 1)$ can be represented by a series of the form (8.3).
- (c) An expansion of the form (8.3) is said to be **terminating** if there is some N such that $c_n = 0$ for all $n > N$. What property characterizes those numbers in $[0, 1)$ which have terminating expansions of the form (8.3)?
- (d) If a number in $[0, 1)$ has a terminating expansion of the form (8.3), show that it also has exactly one nonterminating (8.3) expansion. Describe how that non-terminating expansion is constructed.
- (e) Show that the representation of a number in $[0, 1)$ in the form (8.3) is unique if and only if the number has no terminating expansion.

8.5 The Cantor set

The Cantor set is a somewhat infamous construction in mathematics. Frequently, it has been used as a counterexample to very natural conjectures which turn out instead to be based upon naive intuitions.

For example, now that we have seen that the real numbers in the interval $(0, 1]$ are uncountable, it is clear that that in fact many subsets of \mathbf{R} can not be represented as sequences. In particular, no interval can be represented as a sequence. The reason is that the set of numbers used in a sequence is, by construction, a countable set. Again, this statement is clear enough and at this point should be obvious to all of us.

But the natural, human tendency after learning that an interval is not countable is to incline toward the belief that perhaps an uncountable set has to be an interval. Or perhaps it has to contain something like an interval somewhere inside it. This natural and naive impression turns out to be quite false, and the standard counterexample is the Cantor set.

The Cantor set may be defined in the following way:

Let us start with the representation of the real numbers in the interval $[0, 1]$ using base three. That is, we can write each number x which is in this interval as

$$x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$$

The usual rules, of course, restrict a_n to be either 0 or 1 or 2, depending upon what is needed in order to express x correctly. But we also notice that some values of x have two expansions, one of them terminating and the other having $a_n = 2$ for all n beyond some specific N . For some examples of this, we notice that we can write

$$1 = \sum_{n=1}^{\infty} \frac{2}{3^n}$$

and

$$1/3 = \sum_{n=2}^{\infty} \frac{2}{3^n}$$

or as

$$1/3 = \frac{1}{3} + \sum_{n=2}^{\infty} \frac{0}{3^n}.$$

Moreover, we can write the fraction $2/3$ in two different ways. One of them is the obvious, terminating one, and the other represents it as

$$2/3 = \frac{1}{3} + \sum_{n=2}^{\infty} \frac{2}{3^n}.$$

Now, the Cantor set is defined as

$$C = \{x \mid x = \sum_{n=1}^{\infty} \frac{a_n}{3^n} \text{ and } x \text{ has a representation in which } a_n \neq 1 \forall n\}$$

From the above construction of the Cantor set, it ought to be clear that it contains 0 and 1, but lots of “holes” are left in the interval. For example, just as a start it contains $1/3$ and $2/3$, but the open interval $(1/3, 2/3)$ is removed, as it were, in the first step of constructing C . Indeed, another way to construct C (which has more intuitive visual content) is to remove that interval, and what remains is the union of the two intervals $[0, 1/3]$ and $[2/3, 1]$. In the second step, remove the middle third of each of these two intervals. In the third step, remove the middle third of each of the intervals which remained after the second step, and keep on in like fashion. However, the presence of some of the numbers which

do remain in the Cantor set can be somewhat counterintuitive. It is easily verified, for example, that the fraction $1/4$ is in the set because it most definitely has a ternary representation in which the coefficients a_n are either 0 or 2.

The Cantor set C can be shown to contain no interval whatsoever, and it has many other properties which we have not discussed here. Indeed, it has been the shipwreck of many plausible but false hypotheses during the development of the modern concepts used in mathematics. But a most fundamental and important property of the Cantor set is that it is clearly not countable.

To see that the Cantor set is uncountable, let x be an element of it. Recalling that in the above definition we used coefficients a_n which are either 0 or 2, let $b_n = a_n/2$ for each n . Then we let x correspond to the number

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

It is not difficult to see that this correspondence is one-to-one and that it maps the Cantor set C onto the entire interval $[0, 1]$. We have established a one-to-one correspondence between the Cantor set and another set which we already know is uncountable.

Most particularly, this construction demonstrates that an uncountable set need not be a set which resembles an interval, nor even a set which contains an interval as a subset.

Finally, as a historical note it should be mentioned that the Cantor set was described not only by Georg Cantor in 1882, but also by at least three other mathematicians who actually published the description of it before he did. Those were Henry J. S. Smith (1874), Paul du Bois-Reymond (1880), and Vito Volterra (1881).

Chapter 9

Conclusion of MATH 5200

This chapter, as its title indicates, marks the end of MATH 5200 and the beginning of its successor course MATH 5210.

Chapter 10

Some basics for differentiation and integration

10.1 The derivative of a function

We assume that a real-valued function f of a real variable is given. At any point a in its domain the derivative of f at a is defined as

Clearly, the quantity $f'(a)$ defined in the above expression may or may not exist for some given function f and some given point a in its domain. If the expression does exist, then we say that f is differentiable at a . The definition requires that a is in the domain of f . Moreover, it should be remembered that the definition includes the use of a limit, and the definition of that limit further requires the point a to be an accumulation point of the domain of f . The potentially unaware are reminded that these conditions do not require the point a to be an interior point of the domain of f and do not even require that the domain of f contains an interval which contains the point a . As an extreme example, a function could be defined on a and upon a sequence tending to the point a , and the external requirements for checking whether or not the above limit exists at a are then met.

Exercises

10.1.1 The above definition often occurs in an alternate form, namely that

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

Show that this statement of the definition is in fact equivalent to what is given above.

10.1.2 Show that if f is differentiable at a , then f is continuous at a . Give a counterexample which shows that the converse of this statement need not be true.

10.1.3 Suppose that the function f is differentiable at c . Then, if $f'(c) > 0$ and if c is an accumulation point of the set constructed by intersecting the domain of f with (c, ∞) , then there is a $\delta > 0$ such that at each point x in the domain of f which lies in $(c, c + \delta)$ we have $f(x) > f(c)$. If c is an accumulation point of the domain of f intersected with $(-\infty, c)$, then there is a $\delta > 0$ such that at each point y in the domain of f which lies in $(c - \delta, c)$ we have $f(y) < f(c)$. Formulate similar statements if $f'(c) < 0$.

We say that a function is differentiable on a set S if the function is differentiable at each individual point a in S . The set S , of course, must be a subset of the domain of f . Note also that the definition of the derivative at each point in S additionally requires that every point in S is an accumulation point of the domain of f , too. Any interval, of course, has the property that each point in it is an accumulation point of the interval.

Exercises

10.1.4 Let f be continuous on the interval $[a, b]$ with maximum value M and minimum value m . Further, let f be differentiable on (a, b) . Then show that $f'(c) = 0$ at any point $c \in (a, b)$ for which $f(c) = m$ or $f(c) = M$. In addition to the above possibilities, may there also exist points $c \in (a, b)$ for which $f'(c) = 0$ while $f(c)$ is neither equal to m nor equal to M ?

10.1.5 (Rolle's Theorem) Let the function f be continuous on a closed interval $[a, b]$ and differentiable on the interval (a, b) . If it is also true that $f(a) = f(b) = 0$, then there must exist some point c with $a < c < b$ at which $f'(c) = 0$.

10.1.6 (Mean Value Theorem of Lagrange) Let the function f be continuous on a closed interval $[a, b]$ and differentiable on the interval (a, b) . Then there must exist some point c with $a < c < b$ at which $f'(c) = m$, where m is the slope of the line passing through the points $(a, f(a))$ and $(b, f(b))$.

10.1.7 The derivative of a constant function defined upon an interval is identically zero, as I suppose we all know. Show that the converse of this statement is also true, namely that if f is a function defined upon an interval and $f'(x) = 0$ at every x in the same interval, then f is a constant function.

10.1.8 (Cauchy's mean value theorem) Let f and g both be defined and continuous on the interval $[a, b]$ and differentiable on (a, b) . Then there is a point c satisfying $a < c < b$ at which

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

Cauchy's mean value theorem is used to prove l'Hôpital's Rule and is quite useful in some other contexts, too.

HINT: Consider the function

$$h(x) = [f(x) - f(a)][g(b) - g(a)] - [g(x) - g(a)][f(b) - f(a)].$$

Show that it is possible to apply Rolle's theorem. Apply it. Then do some simple algebra.

10.1.9 Under the hypotheses used in the previous problem, show the particular case of L'hôpital's rule which states that, if $f(a) = g(a) = 0$ and $g'(x) \neq 0$ for $a < x < b$, then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$$

10.1.10 Give proofs for the standard rules of differentiation: the Sum and Product Rules, the Quotient Rule, and the Chain Rule.

10.1.11 (on inverse functions and their derivatives) Suppose that f is a differentiable function on an interval $[a, b]$ and $f'(x) > 0$ for every x in the interval. Then show that f has an inverse function f^{-1} which is defined upon the interval $[f(a), f(b)]$. Further show how the derivative of f^{-1} is related to the derivative of f . Exercise 6.2.12 may be helpful in evaluating the limit which is involved in defining the derivative of f^{-1} .

10.2 The Riemann-Darboux integral

The integral is motivated by the concept of area, with important differences which will be seen later during our development of the topic. However, the initial development of an integral does indeed depend upon the geometric concept of area. In particular, the development of the integral depends upon the fact that the area of a region in \mathbf{R}^2 which is

partitioned into two or more pairwise disjoint subsets is equal to the sum of the areas of the subsets in the partition, also that the area of a subset is non-negative and is monotone, in the sense that if S_1 and S_2 are two sets with $S_1 \subseteq S_2$, then the area of S_1 is not greater than the area of S_2 . We also take as a given the formula for the area of a rectangle.

10.2.1 The integral of a bounded non-negative function on a bounded closed interval

Let us begin our discussion of the integral by assuming we have a bounded, non-negative function f defined upon a (bounded) closed interval $[a, b]$. The function f , note, is **not** assumed to be continuous. The idea is that, somehow, we wish to compute the area which is enclosed below by the x -axis, on the left by the vertical line $x = a$, on the right by the vertical line $x = b$, and above by the graph of $y = f(x)$. It may turn out, of course, that it is not possible actually to carry out this objective for the particular, given function f . Thus, we wish to construct a procedure which will either work, or will visibly not work, depending ultimately upon the nature of the given function. The procedure which we will follow is credited originally to Darboux, and the resulting integral is called the Darboux integral. The Riemann integral will be described later on. The construction of the Riemann integral is somewhat different, but it is the same in the sense that the results are the same. That is, the function f has a Darboux integral if and only if it has a Riemann integral, and the values of the integrals computed by both procedures must coincide if both exist.

To this end, we first construct **partitions** of the interval $[a, b]$.

A partition of the interval $[a, b]$ is any set of the form $\{t_0, \dots, t_n\}$, in which it is agreed that

$$a = t_0 < t_1 < \dots < t_{n-1} < t_n = b. \quad (10.1)$$

Note also that in the representation just above, the points t_1 and t_{n-1} are separately listed only for emphasis, to make it clear what is happening. In particular, the above description of the points in the partition is not intended to exclude the possibility that $n = 1$ or $n = 2$.

Based upon any given partition P as defined above, we now define two sums, the *upper Darboux sum* U (or if we need to be precise, then $U(P)$) and the *lower Darboux sum* L (or if we need to be precise, then $L(P)$) for the function f by

$$U = \sum_{k=1}^n (t_k - t_{k-1}) \sup_{x \in [t_{k-1}, t_k]} f(x) \quad (10.2)$$

$$L = \sum_{k=1}^n (t_k - t_{k-1}) \inf_{x \in [t_{k-1}, t_k]} f(x) \quad (10.3)$$

Also, given the partition P the quantity $\max_{1 \leq k \leq n} (t_k - t_{k-1})$ will be referred to as the **mesh** of the partition P . Note that the terminology “norm” of P is also in common use to mean the same thing.

Very loosely, the idea is that we can define the integral of a bounded, non-negative function f on the interval $[a, b]$ as the common limit of the lower sums and the upper sums, when those sums are based upon any sequence of partitions in which there are “more and more” partition points. This idea, of course, needs to be carried out very carefully. There are many technicalities. The very first technicality, of course, is that we had better require the mesh of the partition to tend to zero as the number of partition points tends to infinity. Otherwise, the result may not be at all what we intend. Indeed, the following two problems depict situations which we obviously wish to exclude:

Exercises

10.2.1 For any n , it is possible to partition the interval $[0, 1]$ using $n + 1$ partition points for which (see (10.1)) $t_{n-1} = \frac{1}{2}$ and the mesh of all these partitions is equal to $\frac{1}{2}$.

10.2.2 Even worse than in the previous problem, it is possible to construct a sequence of partitions P_n of $[0, 1]$, each containing $n + 1$ points as explained in (10.1), and the mesh of the partition P_n tends to 1 as $n \rightarrow \infty$.

Thus, our second attempt to define the integral of a bounded, non-negative function on the interval $[a, b]$ will say that if for any sequence of partitions P_n for which the mesh of the partitions tends to 0, the associated lower and upper sums converge always to a common limit, which is the integral of f on $[a, b]$.

The above is, in fact, a good definition for the integral. But there is still a bit of work to be done in order to make sure of that. For one thing, the definition says **any** sequence, and it wants the common limit of the upper and lower sums to come out the same. Could it be possible that we have some other sequence of partitions with the same required properties, and the common limit associated with the other sequence of partitions might not exist or might exist but be something different? Well, we would hope not and we would expect not. But the second attempt is a bit unwieldy for proving this. Therefore, let us introduce the concept of a **refinement** of a partition:

The partition S_m consisting of ordered points $a = s_0 < s_1 < \dots < s_m = b$ is a refinement of the partition P_n given by $a = t_0 < t_1 < \dots < t_n = b$ provided that $\{t_0, \dots, t_n\}$ is a subset of $\{s_0, \dots, s_m\}$.

Exercises

10.2.3 If the partition S_m is a refinement of the partition P_n and if f is bounded and non-negative, then $L(P_n) \leq L(S_m) \leq U(S_m) \leq U(P_n)$.

Hint: Note that the simplest case which can occur is that $m = n + 1$ and the set of partition points in S_n has been constructed by adding just one new point to the partition points in P_n . Show that the result is true in this particular case, and then explain how this case actually suffices to prove the more general statement.

10.2.4 If P_n and S_m are partitions of an interval $[a, b]$ containing respectively $n + 1$ points and $m + 1$ points, show that there is another partition T , based upon the set $\{t_0, \dots, t_n\} \cup \{s_0, \dots, s_m\}$, which is a refinement of both P_n and S_m . What is the maximum possible number of points needed to construct T ?

10.2.5 Let \mathbf{C} be the collection of all possible partitions of a given interval $[a, b]$, and let f be a bounded non-negative function defined on $[a, b]$. Show that

$$\sup_{P \in \mathbf{C}} L(P) \leq \inf_{P \in \mathbf{C}} U(P) \quad (10.4)$$

With the help of the problems just above, we can reformulate the definition of the integral of a non-negative bounded function f defined upon a closed interval $[a, b]$. The integral will exist if and only if the two sides of the inequality (10.4) are equal to one another, and then the integral of f is equal to that common value. That is, we have

$$\sup_{P \in \mathbf{C}} L(P) = \inf_{P \in \mathbf{C}} U(P) \quad (10.5)$$

For this integral we of course use the well known notation

$$\int_a^b f(x) dx,$$

and its value is equal to the value of the two equal expressions in (10.5).

Finally, the integral of a bounded non-negative function defined on an interval $[a, b]$ defines the **area** of the region in \mathbf{R}^2 which is described by $\{(x, y) | a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}$, which in turn is often described as “the region in the plane which is bounded by the lines $x = a$, $x = b$, $y = 0$, and by the graph of $y = f(x)$.” And, indeed, the value of the integral does coincide with the area of the described region in case that the region is simple enough that we know some other way to compute the integral. For example, if the region is a triangle with a horizontal and vertical side and a slanting hypotenuse, then the integral we have defined does indeed give the enclosed area.

Furthermore, the above analysis shows that in order to compute an integral by direct appeal to the definition just given, it suffices to use partitions which contain equally spaced points in the interval $[a, b]$. That is, if one already knows that the function can be integrated (meaning that (10.5) is known to be true), then one can just go ahead and use successive partitions defined by the points t_0, \dots, t_n which are given for each $k \in \{0, \dots, n\}$ by the explicit formula

$$t_k = a + (b - a) \cdot \frac{k}{n}$$

in order to compute the integral. Furthermore, if the upper sums U and the lower sums L actually do converge to a

common value for some particular sequence of partitions (including here as one possibility that the chosen partitions all use equally spaced points), then the equation (10.5) must hold, too, and the integral must exist.

For the following two exercises, the reader is referred to the introductory section of Chapter 4 for help with certain needed formulas, in particular to Exercises 4.1.5 and 4.1.6. The formulas which are derived in these two problems of course exist “out there in the literature” but it is good to know a way to derive them.

Exercises

10.2.6 Using the definition of the integral, find the integral of the function $f(x) = x$ on the interval $[0, 1]$

10.2.7 Using the definition of the integral, find the integral of the function $f(x) = x^2$ on the interval $[0, 1]$

10.2.8 Show that the function which is defined on $[0, 1]$ by assigning to each irrational number the output value 0 and to each rational number the output value 1 has no Darboux integral because the value of $U(P)$ for any partition P is always 1 while the value of $L(P)$ is always 0.

10.2.9 Show that if f is a non-negative bounded function which is continuous on the interval $[a, b]$, then $\int_a^b f(x) dx = 0$ implies that f is identically zero on $[a, b]$. Give an example to illustrate why we need the continuity of f .

10.2.2 The integral of a bounded function on a bounded closed interval

The most natural method for extending the definition of the integral of a non-negative bounded function defined on an interval $[a, b]$ now needs to be extended. Here, we assume that our function f is an arbitrary bounded function. That is, f is not assumed here to be non-negative. To handle this new situation, what we do in fact is to set up everything exactly as before, with no deviation whatsoever from what we have already done. That is, given any partition P of the interval we take the definitions of the upper sum $U(P)$ and the lower sum $L(P)$ literally, exactly as given in (10.2 and (10.3). After this, the same path of development follows as before, when the function was presumed to be non-negative. There is one important difference, though: the integral of a function is no longer so intimately connected with area because the terms in the sums $U(P)$ and $L(P)$ do not necessarily agree with the computation of the area of a rectangle. For example, if we have two adjacent partition points t_{k-1} and t_k , then the computation of the k th term in U is

$$(t_k - t_{k-1}) \sup_{x \in [t_{k-1}, t_k]} f(x)$$

and this will clearly be negative if

$$\sup_{x \in [t_{k-1}, t_k]} f(x) < 0.$$

Moreover, if f is actually negative across the interval $[a, b]$, then the integral of f on that interval will certainly be non-positive and may well end up being negative, too. Also, if a function f is positive on some parts of an interval $[a, b]$ and is negative on some other part, then the correspondence of the integral to an area is totally lost.

Exercises

10.2.10 The integral of $f(x) = x$ on the interval $[-1, 1]$ is zero. (Danger, Will Robinson! The Fundamental Theorem of Calculus will come at the end of this chapter. So, it is forbidden to use it here!)

10.2.11 Now that we know how to integrate any bounded function, it becomes possible to prove that $\int_a^b f(x) dx$ always exists, for every monotone function f defined on an interval $[a, b]$.

10.2.12 Now that we know how to integrate any bounded function, it becomes possible to prove that $\int_a^b f(x) dx$ always exists, for every continuous function f defined on an interval $[a, b]$.

10.2.3 The effect of unnatural ordering on integration

In the previous two subsections, we have defined what we mean by

$$\int_a^b f(x) dx.$$

Namely, we have used partitions of the interval $[a, b]$ and have described the integral by the use of partitions of $[a, b]$ and the construction of the sums U and L . In order to carry out our project we made sure that all of our partitions used

points which were indexed from left to right. That is, we always used $t_0 = a$ and $t_n = b$ with the rest of the points t_k being similarly ordered from left to right as k runs from 0 to n .

Now, it may seem unnatural or perverse to suppose that the interval is partitioned in a nonstandard manner, such that $t_0 = b$ and $t_n = a$ and the partition points are all ordered from right to left instead of left to right. But we do need to consider this situation, too. Suppose, then, that the partition has been thus set up, perhaps by a bad, willful, or ignorant child. Then, note that if we use the same formula (10.2) for the upper sum U and the same formula (10.3) for the lower sum L , then the result of the unnatural numbering is to create a sign change in every term. For, in this situation the value of the factor $(t_k - t_{k-1})$ switches from being positive to being negative. The effect is, then, to reverse the sign of the entire sum given in each of the formulas (10.2) and (10.3). Clearly, what happens is that if $\int_a^b f(x) dx$ is known to exist (having been computed by use of well-behaved standard partitions using ordering from left to right) then what we have obtained by using the perverse ordering which proceeds from right to left should be described as $\int_b^a f(x) dx$, and we see that

$$\int_b^a f(x) dx = - \int_a^b f(x) dx \quad (10.6)$$

Exercises

10.2.13 The discussion in this subsection shows that, if f is a bounded function on the interval $[a, b]$, then the

formula (10.6) holds. Can you use this to prove that

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx$$

is true for any bounded function f , so long as the interval $[\min\{a, b, c\}, \max\{a, b, c\}]$ is contained within the domain of f ? Show that in the “standard” case that $a \leq c \leq b$, we already knew that this is true without needing to appeal to (10.6).

10.2.4 The Riemann Integral

To define the Riemann integral of a function f defined on an interval $[a, b]$, we use partitions of the interval exactly as defined before. Then, given any partition P we choose points x_1, \dots, x_n , each of which is related to the partition points t_0, \dots, t_n by the requirement $t_{k-1} \leq x_k \leq t_k$ for $k = 0, \dots, n$. Then a **Riemann sum** based upon the partition P is any sum of the form

$$\sum_{k=1}^n (t_k - t_{k-1}) f(x_k) \tag{10.7}$$

The Riemann integral of the function f is then said to exist if any sequence of Riemann sums for which the mesh of the partitions used tends to zero will always give rise to the same limit. This definition explicitly includes the assumption

that the points x_1, \dots, x_n may be **any** points which satisfy the above-listed order requirements, with no other restriction. In the following exercises, the equivalence of the Riemann integral and the Darboux integral is shown.

Exercises

10.2.14 Let f be any function which is defined and bounded on $[a, b]$. Let any partition P of $[a, b]$ be given. Let $U(P)$ and $L(P)$ be as described in (10.2) and (10.3) and let R be any Riemann sum defined on the partition P , as described in (10.7). Show that

$$L(P) \leq R \leq U(P)$$

Further show that if f is integrable in the sense of Darboux, then f is also integrable in the sense of Riemann, and the value of the integral is the same no matter which method was used to obtain it.

10.2.15 Let f be any function which is defined and bounded on $[a, b]$. Let any partition P of $[a, b]$ be given. Let $U(P)$ and $L(P)$ be as described in (10.2) and (10.3). Show that, given $\epsilon > 0$ there exists a Riemann sum R such that $R > U(P) - \epsilon$ and also there exists a Riemann sum R (possibly different from the previous one) such that $R < L(P) + \epsilon$.

10.2.16 Show that if the function f which is defined and bounded on an interval $[a, b]$ is integrable in the sense of Riemann, then it is integrable in the sense of Darboux, and the value of the integral is the same no matter which method was used to obtain it.

10.2.5 The linearity of the integral

As developed in the previous sections, the process of integration is linear. In context, this means exactly the following two statements:

- (i) Given any two functions f and g which are defined and bounded on an interval $[a, b]$, it is true that

$$\int_a^b (f + g)(x) dx = \int_a^b f(x) dx + \int_a^b g(x) dx$$

- (ii) Given any function f which is defined and bounded on an interval $[a, b]$ and given any real number c , it is true that

$$\int_a^b (cf)(x) dx = c \int_a^b f(x) dx$$

Exercises

10.2.17 Prove the statements (i) and (ii) which comprise the definition that the integral is linear.

10.2.18 Explain why the fact that the integral of a function which is bounded and negative on an interval must be a negative number is compatible with linearity, whereas to require that a negative function has an integral which is

“equal to the area trapped between the function and the x -axis” would destroy the property of linearity. Which of these two conflicting goals would we prefer to sacrifice to the other, since we obviously can not have both?

10.2.6 Shortcuts can go wrong, when defining the integral

In some quarters, it seems to be currently fashionable to use “left sums” and “right sums” when defining the integral instead of the upper sums and lower sums described in (10.2) and (10.3). The “left sum” is of course to be computed by doing a Riemann sum in which the points x_1, \dots, x_n are always chosen at the left endpoint of the subintervals in which they reside, and the “right sum” is computed by choosing the points x_1, \dots, x_n are chosen at the corresponding right endpoints. Moreover, in a further attempt to simplify the topic of integration, only partitions consisting of equally spaced points are considered in the construction of these sums.

The idea of defining integrals in terms of a left sum and a right sum instead of an upper sum and a lower sum is apparently connected to an opinion that the upper sum and the lower sum are too difficult for the students to grasp, and this “simplification” thus arises from the noble desire to make the subject of calculus more comprehensible to the audience, one of the motivations often cited by a movement for “calculus reform.”

In some restricted circumstances, there is no terrible harm in the shortcuts described above. If one starts out by assuming that the function to be integrated is continuous, then it ought to be clear that its Riemann integral or Darboux integral must exist, and clearly the “left” Riemann sums and the “right” Riemann sums will also converge to the value of the integral. The same must be true, too, if the function is not necessarily assumed to be continuous but is assumed

to be monotone. But serious difficulties can arise in some other circumstances.

Exercises

10.2.19 Consider the function defined in Problem 10.2.8. Show that for any partition P_n consisting of the equally spaced points defined by $t_k = k/n$ for $k = 0, \dots, n$ the left sum always has value 1 and the right sum is always the same, and both of these sums are equal to $U(P_n)$. However (as was the result in the mentioned problem) this function is not integrable in spite of the fact that the left and right sums on equally spaced partition points seem to be well-behaved.

10.2.20 For the same function defined in 10.2.8 the left sum and the right sum defined on a set of evenly spaced partition points are either equal to zero or tend to zero on any interval of the form $[0, p]$, and also zero on the interval $[p, 1]$ if p is any irrational number. Note that, if one could actually get away with doing this kind of thing while defining the integral, then one would end up with a result which violates Problem 10.2.13 by asserting that

$$\int_0^p f(x) dx = \int_0^p f(x) dx + \int_p^1 f(x) dx$$

must hold, which would boil down to $0 + 0 = 1$, an evident absurdity.

It should be clear that the embarrassments which exist in the previous two problems are all caused by the attempt to make an “easy” shortcut in the definition of the integral. Specifically, the attempt to define the integral in terms of left and right sums constructed on partitions using equally spaced partition points has serious limitations and shortcomings. If this method of “definition” for the integral is to be used, its application must be restricted to functions which are continuous or monotone. Functions which are neither continuous nor monotone may quite possibly not be integrable, and this fact can not be reliably determined using such a restricted definition of the integral.

Exercises

10.2.21 Some calculus texts do not resort to defining the Riemann integral by referring only to the left sum and the right sum. Rather, they define the Riemann integral as we have done, above, using arbitrary points x_1, \dots, x_n for which each x_k is situated in the partition interval $[t_{k-1}, t_k]$. But only partitions of equally spaced points are used in the definition. Is this a good definition of the integral, or not?

10.3 The Fundamental Theorem of Calculus

In this section, we show a connection between the integration and differentiation. First, we need to prove a basic result, the Mean Value Theorem for integrals.

Exercises

10.3.1 (Mean Value Theorem for Integrals) Let us assume that the function f is continuous on a closed interval $[a, b]$. Then there is $c \in [a, b]$ such that

$$\int_a^b f(x) dx = (b - a)f(c) \quad (10.8)$$

The result just above is also true if $b < a$ and f is continuous on the interval $[b, a]$

10.3.2 Can you construct an example of a function f for which $\int_a^b f(x) dx$ exists (and is finite), but f is not continuous, and the result (10.8) fails to be true?

We also notice that, if f is any integrable function on a bounded interval which contains a , then for any x in the same interval we can define another function $F(x)$ by

$$F(x) = \int_a^x f(t) dt \quad (10.9)$$

The function f does not need to be continuous in order for the above definition to work, only integrable on the interval in question. However, we will assume that f is continuous in what follows. Also, note that in this definition for F there

is no problem in the definition if $x < a$ (assuming, of course, that f can be integrated on the interval $[x, a]$). Also note that the value of $F(a)$ is clearly 0.

Now, the so-called First Fundamental Theorem of Calculus states that, when f is continuous on an interval containing x and $F(x)$ is the function defined above, in (10.9), then

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

We recall the definition for the derivative $F'(x)$ which is found in Exercise 10.1.1.

The First Fundamental Theorem of Calculus can now be proved by using the steps which are laid out sequentially in the next three Exercises:

Exercises

10.3.3 For the function F defined in (10.9) it is the case that

$F(x+h) - F(x) = \int_x^{x+h} f(t) dt$, provided that the domain of f contains an interval which contains both x and $x+h$, as well as a .

10.3.4 We assume that f is continuous and that its domain contains an interval which contains both x and $x+h$, as well as a . If $h \neq 0$, then, using Exercise 10.3.1, between x and $x+h$ there is a point c (depending on h of course)

such that

$$\frac{1}{h} \int_x^{x+h} f(t) dt = f(c)$$

10.3.5 Under the hypotheses of the previous problem, if $h \rightarrow 0$ then we must have $c \rightarrow x$, and therefore the function F which is defined in (10.9) satisfies

$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \frac{1}{h} \int_x^{x+h} f(t) dt = f(x).$$

This problem completes the proof of the First Fundamental Theorem of Calculus.

The Second Fundamental Theorem of Calculus involves the same ideas, but in a slightly different way:

Suppose that the function f is continuous on an interval $[a, b]$ and that F is *any* function such that $F'(x) = f(x)$ for every $x \in [a, b]$. Then

$$\int_a^b f(x) dx = F(b) - F(a).$$

Exercises

10.3.6 Prove the Second Fundamental Theorem of Calculus. To complete the proof you will probably need the First Fundamental Theorem of Calculus and Problem 10.1.7. If you have not done that problem before now, do it, too.

10.4 Improper integrals and the integral test for a series

An **improper integral** is an integral of a function which is not defined at an endpoint of the interval of integration. That is, either the function fails to be definable in any meaningful way at the endpoint in question, or the interval of integration is unbounded. In such an eventuality, the integral may exist and be computable as a limit of integrals defined upon closed intervals, or, depending upon the function in question, it may not. In any event, if the function f is defined upon the interval $(a, b]$ we define the (improper) integral in terms of a one-sided limit, thus:

$$\int_a^b f(x) dx = \lim_{c \rightarrow a^+} \int_c^b f(x) dx.$$

Note that if in fact $f(a)$ does exist, then this is not a definition of something new. But it does define something new if $f(a)$ does not exist. We also can have a function which is not defined at b , and in that case we use a similar left hand limit, moving the upper limit of the integration upwards toward b . Note also that if the function f is defined only on (a, b) then we must perform a limit operation at both ends of the interval. It should be quite clear as to why these two

limits must be taken entirely separately from one another. The same is true, too, for any function which is not defined at some point c for which $a < c < b$. In such a case, in order to arrive at a correct value for $\int_a^b f(x) dx$, it is absolutely necessary to compute separately the two improper integrals from a to c and from c to b in isolation from one another, avoiding the possibility that if they are somehow computed together the badness of one of them might accidentally cancel out the badness of the other. If either of these two fails to exist, then the entire integral can not exist, either.

In usage which is similar to the usage for infinite series, an improper integral which can actually be computed and results in an answer that is finite is called **convergent** or is said to **converge**, and an integral which produces such an output as ∞ or $-\infty$ or worse is said to be **divergent** or is said to **diverge**. The usage is similar to the usage for infinite series for similar reasons. We often use these integrals to compute or to estimate something, and if the result of the computation turns out to be something infinite, one usually concludes that something has gone wrong with the calculation. Which might include, of course, something wrong with the setup and the assumptions behind it.

A somewhat different type of improper integral can be seen if the upper limit of integration is ∞ or if the lower limit is $-\infty$, or both. Assuming that if f is defined for all $x \geq a$, we define

$$\int_a^\infty f(x) dx = \lim_{b \rightarrow \infty} \int_a^b f(x) dx,$$

with an obvious corresponding definition if f is defined on $(-\infty, b]$.

Clearly, if the interval of integration is $(-\infty, \infty)$ or is any other kind of interval which requires the use of improper

integration at both ends of the interval, then the two limits in question need to be computed separately from one another. One obvious way to do this is to split the integral in two at any point between those two endpoints and to compute the integrals of the left half and the right half entirely separately from one another. For, an attempt to combine the two calculations can lead to yet another proof of the Law of Unintended Consequences. For example, the improper integral

$$\int_{-\infty}^{\infty} x \, dx$$

is clearly divergent. But it is equally obvious that

$$\lim_{b \rightarrow \infty} \int_{-b}^b x \, dx = 0,$$

showing what nonsense can occur if the requirement to compute limits separately toward ∞ and $-\infty$ is not observed.

Now, one of the standard applications of the improper integral to ∞ is the Integral Test for the convergence of an infinite series which has positive and decreasing terms. Let us suppose that we have a series

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

and we can find a decreasing function f which satisfies $f(n) = a_n$ for all n . Then it is clear that we have for each $n \geq 1$

$$a_n \geq \int_n^{n+1} f(x) dx \geq a_{n+1}$$

From this observation, it is possible to see that the series will converge if and only if

$$\int_1^{\infty} f(x) dx < \infty.$$

Moreover, it should also be clear that the partial sums of the given series can be used to estimate the value of the improper integral, and that the integral can be used to estimate the sum of the series, and also the error incurred by taking a partial sum of the series.

Exercises

10.4.1 Consider the integral $\int_0^1 \frac{1}{x^p} dx$. For which values of p do these integrals converge, and for which do they diverge?

10.4.2 Consider the integral $\int_1^{\infty} \frac{1}{x^p} dx$. For which values of p do these integrals converge, and for which do they diverge?

10.4.3 Could it be that $\int_{-\infty}^{\infty} x \, dx$ is equal to 0 because $\int_{-a}^a x \, dx = 0$ for all $a > 0$? Why or why not?

10.4.4 Could it be that $\int_{-1}^1 \frac{1}{x} \, dx$ is equal to 0 because $\int_{-1}^{-a} \frac{1}{x} \, dx + \int_a^1 \frac{1}{x} \, dx = 0$ for all $a > 0$? Why or why not?

10.4.5 A series of the form $\sum_{n=1}^{\infty} \frac{1}{n^p}$ is called a p -series. For which values of p do these integrals converge? Notice that one of the results of this problem is that we now “know both ways” that $\int_1^{\infty} \frac{1}{x} \, dx$ has to diverge, because we already know that the harmonic series diverges, independently of the Integral Test for convergence. See Exercise 4.7.11. Thus, in accord with the Integral Test as given above, we now know that the integral must diverge as well as the series.

10.4.6 Show that for all values of p we have $\lim_{n \rightarrow \infty} \left(\frac{1}{n^p}\right)^{\frac{1}{n}} = 1$, demonstrating that the Root Test for convergence can not resolve the question of convergence or divergence of any p -series whatsoever.

10.5 A problem with the Riemann integral

In Exercise 10.2.8 we have mentioned a really bad function, defined on the interval $[0, 1]$ by specifying its value as 1 at each rational number in that interval and 0 otherwise. There, we have also shown that the function thus defined does not have a Riemann integral. This might not be very surprising, and at first sight it might not seem to be a very big deal. Who would want to integrate such a stupid function, anyway, when any fool could see that it can't be done?

But, on second thought the fact that this function can not be integrated demonstrates a serious deficiency of the Riemann integral.

To see what the problem is, let us recall that set of all the rational numbers in $[0, 1]$ is a countable set. That means, we can count them. Let us assume that any such counting scheme has been established and agreed upon. Then, relative to our agreed-upon counting scheme, we can represent all of those rational numbers as a sequence $\{q_k\}$. Then we can define a sequence of functions $\{f_k\}$, each defined on $[0, 1]$ by specifying $f_k(q_k) = 1$ and $f_k(x) = 0$ for all other values of x in $[0, 1]$.

Now for each n we can further define

$$F_n(x) = \sum_{k=1}^n f_k(x).$$

Exercises

10.5.1 Show that each function F_n is Riemann integrable, and show that

$$\int_0^1 F_n(x) dx = 0$$

10.5.2 Show that for each $x \in [0, 1]$ the limit of the sequence $\{f_n(x)\}$ is 1 if x is rational, and 0 otherwise. That is, the limit of the sequence of the functions F_n is the “salt and pepper” function which we can not integrate.

10.5.3 Show that

$$\lim_{n \rightarrow \infty} \left(\int_0^1 F_n(x) dx \right) \neq \int_0^1 \left(\lim_{n \rightarrow \infty} F_n(x) \right) dx$$

The above state of affairs indicates, by way of example, that we can not ever be sure of our ability safely to reverse the order of integration and taking a limit. It is not that we cannot integrate “all functions,” nor the rather weird “salt and pepper” function in particular. Rather, the problem is that the set of Riemann integrable functions lacks completeness in a manner quite analogous to the incompleteness of the set of rational numbers. Namely, we had a sequence of functions which all could be integrated, and the sequence has a reasonably defined limit which can not be integrated. We started within the set of integrable functions and we constructed a sequence which takes us outside. That’s bad. That is very bad. The reason it is bad is the same reason that the incompleteness of the rational numbers was bad. We cannot with safety and confidence work any problem in which we would like to compute or estimate an

integral by constructing a sequence of functions which converge to the function to be integrated. We cannot safely do this even if each of the functions in the sequence can be integrated. Of course, we can still carry out such a procedure in very restricted and confined circumstances. But those circumstances are very restricted and confined indeed and are far from being valid in all situations where we would wish them to be.

Therefore, the example which is described above is one of the standard examples which has motivated the development of one of the central topics in 20th century mathematics. The efforts to find a solution to the problem occupied the attention of many good mathematicians. The topic is also a central topic in the graduate analysis sequence MATH 7200-7210. Those of you who are pursuing graduate studies in mathematics may quite possibly see the material in that course. Here, we only point out what the problem is.

Chapter 11

Normed Vector Spaces and Function Spaces

11.1 Normed vector spaces

Normed vector spaces are often also called *normed linear spaces*. A normed vector space is a vector space with a *norm*

defined upon it. We assume here that the algebraic definition of a vector space is familiar to the reader; we will also assume here that the field of scalars is \mathbf{R} . An alternative way of saying this is that the vector space is a “vector space over \mathbf{R} .” It is possible of course to define a vector space over any other field of scalars, but we will not study any of those other vector spaces here.

A *norm* defined upon a vector space V is a function from V to \mathbf{R} . As to notation, the norm of a vector $v \in V$ is represented by $\|v\|$. A norm must satisfy

N1. $\|v\| \geq 0$, for every $v \in V$, and $\|v\| = 0$ if and only if $v = 0$ ($= 0_V$).

N2. $\|\alpha v\| = |\alpha| \cdot \|v\|$ for every $\alpha \in \mathbf{R}$ and every $v \in V$.

N3. $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.

A norm introduces a concept of distance into a vector space, with the distance between any two vectors v and w in V defined as $\|v - w\|$. This distance is, in fact, a metric defined upon the vector space. The metric is compatible with the algebraic operations on V and is “translation invariant,” for, one may note that if v , w , and x are any three vectors in V , then

$$\|v - w\| = \|(v - x) - (w - x)\| \quad (11.1)$$

which shows that the distance between v and w is the same as the distance between $v - x$ and $w - x$, and for that matter the distance between v and w is the same as the distance between $v - w$ and 0 . In other words, the norm operates in V

in a manner very similar to that of absolute value in \mathbf{R} . These observations are definitely true, but they are one of the things which distinguish norm-induced metrics from metrics in general. To recall what could have happened, look again at Exercise 5.5.5. The metric defined there is most definitely a metric, but it is also most definitely not compatible with any metric defined by a norm on \mathbf{R} because (11.1) does not hold for the metric in Exercise 5.5.5.

In a normed vector space, a topology (a collection of open sets, recall) can be constructed in a manner which is quite similar to what has been developed already in \mathbf{R} . Let $S \subset V$. Then v is an *interior point* of S , provided that there is $\delta > 0$ such that the set $\{w : \|v - w\| < \delta\}$ is entirely contained in S . On the other hand, v is said to be an *accumulation point* of S , provided that for every $\delta > 0$ the set $\{w : \|v - w\| < \delta\}$ contains at least one vector x in S , with $x \neq v$. And now we declare S to be open if and only if each vector $s \in S$ is an interior point of S . We also declare that S is closed if and only if S contains all of its accumulation points. It will be seen in the following exercises that the topology induced by a norm on a vector space is “built” in a manner to what was done to build the usual topology on \mathbf{R} in Section 5.1. Indeed, many of the arguments are quite similar.

Exercises

11.1.1 In a normed vector space V , the set $B_{v,\delta} = \{w : \|w - v\| < \delta\}$ is called the **open ball** with center at v and radius δ , or just the open ball at v of radius δ . Also, the set $B_{0,1}$ is often referred to as “the open unit ball in V .” Show that the open ball at any $v \in V$ of radius δ is indeed an open set.

11.1.2 For a fixed v , again, the set $\overline{B}_{v,\delta} = \{w : \|w - v\| \leq \delta\}$ is called the *closed* ball at v , radius δ . The set $\overline{B}_{0,1}$ is often referred to as “the closed unit ball in V .” Not surprisingly, the set $\overline{B}_{v,\delta}$ is a closed set. Prove it.

11.1.3 Under certain circumstances, two norms which are defined on a given vector space V could quite possibly give rise to the same topological structure on V . Suppose that we have two norms defined on V which we will denote by $\|\cdot\|_a$ and $\|\cdot\|_b$. We say that the two norms are **equivalent** if and only if there exist two positive constants c_1 and c_2 such that $c_1\|v\|_a \leq \|v\|_b \leq c_2\|v\|_a$ is true for every $v \in V$. Show that the two norms are equivalent if and only if every set which is open in the topology defined by either of the norms is also open in the topology defined by the other.

11.1.4 Show that equivalence of norms is an equivalence relation, as defined in Problem 2.4.1.

11.1.5 Show that any norm is a continuous function in the topology defined by it. The definition that a function $f : V \rightarrow \mathbf{R}$ is continuous at a particular vector $v \in V$ is, of course, that for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $w \in V$ the condition $\|v - w\| < \delta$ implies the conclusion that $|f(w) - f(v)| < \epsilon$. And we also say that the function f is continuous if it is continuous at every $v \in V$. Thus, it suffices here to show that, given an arbitrary $v \in V$ and given $\epsilon > 0$ there is $\delta > 0$ such that if $w \in V$ and $\|w - v\| < \delta$ then $|||w|| - \|v||| < \epsilon$.

11.1.6 Let two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ be given on a vector space V , and assume that there is a constant C such that $\|v\|_a \leq C\|v\|_b$ for all $v \in V$. Show that for all $w \in V$ and for all $\epsilon > 0$ there is $\delta > 0$ such that for all $v \in V$, if $\|v - w\|_b < \delta$ then $|||v\|_a - \|w\|_a| < \epsilon$. That is, $\|\cdot\|_a$ is a continuous function on V in the topology defined by $\|\cdot\|_b$.

It is helpful to study some examples of normed vector spaces. The simplest vector space which is not \mathbf{R} itself is \mathbf{R}^2 . There are many ways to put a norm on this space. The most usual way is to identify the vectors in \mathbf{R}^2 geometrically with their “endpoints,” which are points in the Cartesian plane. Then, the norm of a vector can simply be defined as its length, according to the standard Distance Formula. Specifically, for $v = (x, y)$, we define a norm $\|v\|_2$ by

$$\|v\|_2 = \sqrt{x^2 + y^2}.$$

However, there are other ways to define a norm on \mathbf{R}^2 . For $v = (x, y)$ we could define a norm $\|v\|_1$ by

$$\|v\|_1 = |x| + |y|$$

or a norm $\|v\|_\infty$ by

$$\|v\|_\infty = \max\{|x|, |y|\}.$$

The funny subscripts are standard here. Their meaning will be made clear later. There are lots of other ways to define a norm on \mathbf{R}^2 , too.

Exercises

11.1.7 Show that on \mathbf{R}^2 the three norms just introduced are indeed norms.

11.1.8 Sketch the closed unit ball (i. e. closed ball at 0 with radius 1) in \mathbf{R}^2 which is associated with the 1-norm, the 2-norm, and the ∞ -norm.

11.1.9 Show that a subset S of \mathbf{R}^2 is open in the topology induced by the 1-norm if and only if S is open in the topology induced by the 2-norm if and only if S is open in the topology induced by the ∞ -norm (HINT: Stare at the pictures you sketched in exercise 11.1.8, and use what you can see from your sketch to show that the three norms are in fact equivalent).

11.1.10 In Exercise 11.1.2 we have seen that the closed unit ball is closed. If we use the ∞ -norm in \mathbf{R}^2 then the closed unit ball is the set $[-1, 1] \times [-1, 1]$. Generalize this result to show that any rectangle $[a, b] \times [c, d]$ is also a closed set in \mathbf{R}^2 .

11.1.11 An obvious question at this point is that of how many of the topological properties of the real numbers carry over to \mathbf{R}^2 . For example, one might ask whether or not the closed unit ball in \mathbf{R}^2 is compact. Taking into account exercise 11.1.9, it will not matter which norm we use, of the three that we have looked at. In this case, perhaps the ∞ -norm is easiest to visualize because of the rectangular nature of the “ball” of a given radius. So, let us assume that we have an open covering of the closed unit “ball” in this space, which is then the Cartesian product $[-1, 1] \times [1, 1]$. Can the proof of Heine-Borel theorem for a closed and bounded interval in \mathbf{R} be bootstrapped to show that the given open covering for $[-1, 1] \times [1, 1]$ must have a finite subcovering?

11.1.12 Generalize the previous problem by showing that any rectangle of the form $[a, b] \times [c, d]$ is also compact in \mathbf{R}^2 .

11.1.13 If you have succeeded in solving Exercise 11.1.11, then the way ought to be apparent, by which essentially the same proof can be generalized to show that the n -dimensional cube $[-1, 1]^n$, which is the closed unit ball in \mathbf{R}^n under the natural extension of the ∞ -norm defined above in \mathbf{R}^2 , is compact as well. Provide the details.

An interesting question is whether or not the closed unit ball is compact in any normed vector space, without restriction on the dimension. As it happens, the answer is “no.” Why do you think that might be the case? The next problem will be used later on in order to shed some light on this issue.

Exercises

11.1.14 A compact set S in any metric space (and hence in any normed vector space) must have the Bolzano-Weierstrass property, that any infinite subset of S must have an accumulation point. HINT: If the infinite subset of S has no accumulation point, then it is closed. And for each point x in the infinite subset there is an open set \mathcal{O}_x which contains x but does not contain any other point in the infinite subset. Further, the complement of the infinite subset is open. The union of that complement with all of the sets \mathcal{O}_x is an open covering for S . Find the contradiction and fill in any missing details.

11.1.15 Similar to what is in the problem about the 1-norm and the 2-norm and the ∞ -norm, but much harder to prove. All possible vector space norms on \mathbf{R}^2 , not just the three which have been defined above, must give rise to the same topology on \mathbf{R}^2 . That is, a set which is open in the topology defined by any given norm on \mathbf{R}^2 is open in the topology defined by any other norm on \mathbf{R}^2 . Hint: Show that the given norm is equivalent to the ∞ -norm, because you will need to use the result of Problem 11.1.11 in order to show that one of the two required constants exists.

11.1.16 Unsurprising in view of the previous problem, but perhaps even harder to prove, is the fact that a set which is open in the topology defined by any given norm on \mathbf{R}^n is open in the topology defined by any other norm on \mathbf{R}^n . Hint: Again compare the given norm to the ∞ -norm and show that they are equivalent, but this time you need to use the result in Problem 11.1.13.

11.1.17 Generalize the result in the previous problem, showing that all norms on a given vector space V of dimension n are equivalent.

A good start on this problem is to define a norm which is similar to the ∞ -norm. Let $\{v_1, \dots, v_n\}$ be a basis for V . Then any $v \in V$ can be written uniquely in the form

$$v = \sum_{k=1}^n \alpha_k v_k$$

for appropriate choices of the coefficients $\alpha_1, \dots, \alpha_n$. As a start, then, show a norm on V , quite analogous to the ∞ -norm on \mathbf{R}^n is defined by taking $\max\{|\alpha_1|, \dots, |\alpha_n|\}$. Let us denote this norm by $\|\cdot\|_u$ (the subscript denotes that it is a “uniform” norm, similar to the ∞ -norm though not necessarily the same).

Now, let $\|\cdot\|$ be any other norm on V . Let v be any vector such that $\|v\| = 1$. Then v can be represented as above, and we can also assume that the basis elements v_1, \dots, v_n have norm $\|v_k\| = 1$ for $k = 1, \dots, n$. Show, using the Triangle Inequality, that $\|v\| \leq n\|v\|_u$.

To get the second inequality needed to prove the equivalence, it is necessary to use Problem 11.1.6 and to prove that the set $\{v \in V, \|v\|_u \leq 1\}$ is compact in the topology induced by $\|\cdot\|_u$.

11.1.18 We have seen that every norm which can be defined upon a finite-dimensional vector space V is equivalent to every other norm which can be defined upon V . But, no matter how a norm is defined on V the closed unit ball $\overline{B}_{0,1}$ must be a **convex set**. That is, if v_1 and v_2 are any two vectors which are in $\overline{B}_{0,1}$, then $\{v : v = (1-t)v_1 + tv_2, 0 \leq t \leq 1\} \subset \overline{B}_{0,1}$. Note that if $V = \mathbf{R}^n$, then the set $\{v : v = (1-t)v_1 + tv_2, 0 \leq t \leq 1\}$ is a line segment. The proof of this exercise is an immediate and rather trivial consequence of the Triangle Inequality (N3), but the fact that the closed unit ball in any normed vector space must necessarily be a convex set is well worth noting.

11.2 Inner Products and Related Norms

Quite possibly, in Exercise (11.1.7) you found it unexpectedly difficult to prove that the Triangle Inequality holds for the 2-norm. There was computational detail and messiness, some rather nasty things which seem unavoidable. If one stops to think that similar problems, but worse, would crop up if one were hoping to prove that the 2-norm also works equally well in \mathbf{R}^n , or in other less expected contexts, too, then it is a very good idea to take a step backwards from the situation and to see if there is any more general method of approach. What confronts one, in other words, is an occasion in which the mathematical principles of abstraction and generalization become quite valuable.

Given a vector space V over \mathbf{R} , a function $\langle v, w \rangle$ defined for all ordered pairs $(v, w) \in V \times V$ with output in \mathbf{R} is called an **inner product on V** , and V is called an **inner product space** if the following properties hold:

IP1 $\langle v, v \rangle \geq 0$ for all $v \in V$, with $\langle v, v \rangle = 0$ if and only if $v = 0_V$.

IP2 $\langle v, w \rangle = \langle w, v \rangle$ for all $v \in V$ and for all $w \in V$.

IP3 $\langle cv, w \rangle = c \langle v, w \rangle$ for all $v \in V$, all $w \in V$, and all $c \in \mathbf{R}$.

IP4 $\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle$ for all $v_1 \in V$, all $v_2 \in V$, all $w \in V$.

The classic instance of the above definition is, of course, the dot product of two vectors in \mathbf{R}^n , which definitely satisfies all of the above properties. But there are lots of other inner products, and the reason for the general definition is that whatever follows from this general definition must follow in every particular case.

Now, nothing was said about any norm on V . But one of several motivations for the above general definition of an inner product is that

$$\|v\| = (< v, v >)^\frac{1}{2}$$

will invariably give rise to a norm on v . To show that this is true, we recall the defining properties of a norm and show that each of them must hold for the (still prospective) norm which we have just defined:

N1. $\|v\| \geq 0$, for every $v \in V$, and $\|v\| = 0$ if and only if $v = 0$ ($= 0_V$).

N2. $\|\alpha v\| = |\alpha| \cdot \|v\|$ for every $\alpha \in \mathbf{R}$ and every $v \in V$.

N3. $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$.

Exercises

11.2.1 Show that N1 just above follows from IP1.

11.2.2 Show that N2 just above follows from IP3 combined with IP2.

We have not yet proven N3, the triangle inequality. In order to show that it holds, we need an intermediate result, which is important in its own right and bears the name of several famous mathematicians. It has been referred to as the Cauchy inequality, then, as mathematical history became more exact, the Cauchy-Schwarz inequality, and then with

still more precision the Cauchy-Buniakowsky-Schwarz inequality. One wonders whether deeper investigation will add even more names to the list of its independent discoverers. In any event and whatever the inequality ought to be called, we have:

Theorem: *Given a vector space V over \mathbf{R} and an inner product defined upon it, the inequality*

$$| \langle v, w \rangle | \leq (\langle v, v \rangle)^{\frac{1}{2}} (\langle w, w \rangle)^{\frac{1}{2}}$$

holds for any two vectors $v \in V$, $w \in V$.

Proof:

To see that this is true, the first thing to do is to pick off an obvious case. Namely, if either $v = 0_V$ or $w = 0_V$ or both, then all of the results of the theorem are obviously true. Therefore, in what follows we need only to consider the situation that neither $v = 0$ nor $w = 0$.

Now, Consider $\langle v - w, v - w \rangle$. We have, using the properties IP1 and IP2 and IP4,

$$\begin{aligned} 0 \leq \langle v - w, v - w \rangle &= \langle v, v - w \rangle - \langle w, v - w \rangle \\ &= \langle v, v \rangle - 2 \langle v, w \rangle + \langle w, w \rangle \end{aligned}$$

from which

$$2 \langle v, w \rangle \leq \langle v, v \rangle + \langle w, w \rangle$$

If v and w are any two elements of V for which $\langle v, v \rangle = \langle w, w \rangle = 1$, then we have

$$2 \langle v, w \rangle \leq 2$$

from which

$$\langle v, w \rangle \leq 1$$

Moreover, this inequality can clearly be extended to say as well that

$$|\langle v, w \rangle| \leq 1$$

Now, we can return to the more general case. If $v \neq 0$, then let

$$v_1 = (\langle v, v \rangle^{-\frac{1}{2}})v$$

and similarly for $w \neq 0$ let

$$w_1 = (\langle w, w \rangle^{-\frac{1}{2}})w$$

Then by property IP3 it follows that $\langle v_1, v_1 \rangle = \langle w_1, w_1 \rangle = 1$ and since $|\langle v_1, w_1 \rangle| \leq 1$ we have

$$|(\langle v, v \rangle^{-\frac{1}{2}})v, (\langle w, w \rangle^{-\frac{1}{2}})w \rangle| \leq 1.$$

But, then we have

$$(\langle v, v \rangle^{\frac{-1}{2}})(\langle w, w \rangle^{\frac{-1}{2}})|\langle v, w \rangle| \leq 1$$

and it follows that

$$|\langle v, w \rangle| \leq (\langle v, v \rangle)^{\frac{1}{2}}(\langle w, w \rangle)^{\frac{1}{2}}$$

and the theorem follows.

Almost certainly, the “dot product” in \mathbf{R}^n is the original example of an inner product from which the more general and abstract concept of an inner product is descended. The reader may recall that for two vectors v and w in \mathbf{R}^n , the formula

$$\langle v, w \rangle (\langle v, v \rangle^{\frac{-1}{2}})(\langle w, w \rangle^{\frac{-1}{2}}) = \cos \theta$$

is often cited, in which θ is said to be the angle between the two vectors. In more general situations, it may seem strange to refer to an “angle” between two elements of a vector space which has no close resemblance to \mathbf{R}^n even if an inner product can exist there. Perhaps this could be done anyway and there would be no apparent harm, but it really can seem strange. However, one thing clearly is important. The concept of “at right angles” does seem to be truly useful and universal, and we make the following definition:

Definition: If v and w are elements of a vector space V which has an inner product and if $\langle v, w \rangle = 0$, then we say that v and w are **orthogonal**.

Now, we return to the task of proving that when we have defined $\|v\| = \sqrt{\langle v, v \rangle}$ for all $v \in V$ we have actually been successful in defining a norm on V . What remains to be shown is property N3, the Triangle Inequality. Specifically, we need to prove that if v and w are any two vectors in the inner product space V , then

$$\|v + w\| \leq \|v\| + \|w\|.$$

Clearly, since every term occurring is non-negative, the above is equivalent to

$$\|v + w\|^2 \leq (\|v\| + \|w\|)^2 = \|v\|^2 + 2\|v\|\|w\| + \|w\|^2$$

However, $\|v + w\|^2 = \langle v + w, v + w \rangle$ by definition, and we have already seen that

$$\langle v + w, v + w \rangle = \langle v, v \rangle + 2\langle v, w \rangle + \langle w, w \rangle$$

Replacing on the right side of the inequality above $\|v\|^2$ by $\langle v, v \rangle$ and $\|w\|^2$ by $\langle w, w \rangle$, we see that the inequality is true if and only if

$$2\langle v, w \rangle \leq 2\|v\|\|w\|.$$

Since this is true by the Cauchy-Buniakowsky-Schwarz inequality, the result follows. The proposal to define the norm of a vector v in an inner product space V by $\|v\| = \sqrt{\langle v, v \rangle}$ is shown always to work, dependent only upon the fact that we have an inner product.

An inner product space is special, in that many concepts of geometric nature seem to generalize nicely from their original versions, with hardly any glitch. Here is a small example.

Exercises

11.2.3 (the Pythagorean Theorem) Let V be an inner product space, and let v and w be any two orthogonal vectors in V . Then

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2$$

11.3 The normed vector space $C[a, b]$

We define $C[a, b]$ the set of all functions defined and continuous on the closed interval $[a, b]$.

Exercises

11.3.1 Show that $C[a, b]$ is a real vector space.

On the vector space $C[a, b]$, it is possible to impose many norms. We mention here three of them:

The *uniform norm* on $C[a, b]$ is defined by

$$\|f\| = \sup_{x \in [a, b]} |f(x)|$$

This is indeed the standard norm on the space $C[a, b]$, and we shall soon see why. However, to distinguish it from the two other norms we are going to introduce, we will write it here as $\|f\|_\infty$.

The *least squares norm* or 2-norm is given by

$$\|f\|_2 = \left(\int_a^b |f(x)|^2 dx \right)^{\frac{1}{2}}.$$

The *integral norm* or 1-norm is given by

$$\|f\|_1 = \int_a^b |f(x)| dx.$$

Exercises

11.3.2 Show that the uniform norm on $C[a, b]$ is indeed a norm, as advertised. Note that “sup” in the definition can in fact be replaced by “max.” Why? Note also that the most difficult of the properties to establish is the Triangle Inequality (N3). Give all relevant details.

11.3.3 Show that the integral norm is indeed a norm, as advertised. Note that the most difficult thing to prove rigorously here is that, if the integral of a non-negative continuous function on a closed interval is zero, then the function itself is zero everywhere. This property depends very much on continuity, incidentally. It obviously need not be true if the function is not continuous. A rigorous proof will depend upon the fact that, if a continuous function is non-negative on all of $[a, b]$ and positive at some point $c \in [a, b]$, then there is a rectangle which lies under

the graph of the function and above the x -axis; therefore the integral of the function must be strictly greater than the area of that rectangle.

11.3.4 Show that the least squares norm is indeed a norm, as advertised, on the space $C[0, 1]$. Note that the most difficult thing to prove here, after the previous problem is already done, is the Triangle Inequality (N3). This will require one to prove first that

$$\langle f, g \rangle = \int_0^1 f(x)g(x) dx$$

succeeds in defining an inner product upon $C[0, 1]$ and additionally to show that the proposed definition of the least squares norm coincides with the norm which is induced by this inner product. The proof of the Triangle Inequality will then need to use the Cauchy-Schwarz (Cauchy-Buniakowsky-Schwarz? Who else??) inequality. Explain what this inequality states about two continuous functions f and g defined upon $[0, 1]$, and then finish the problem.

Many norms exist on $C[a, b]$ besides those mentioned. One class of norms is the so-called p -norms. For $1 \leq p < \infty$ the p -norm is defined by

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}$$

For $p = \infty$, the norm $\|f\|_\infty$ is the same as the uniform norm of f , already defined above.

The p -norms are also meaningful in finite-dimensional vector spaces, of course. In this case the definitions are based upon finite sums, not upon integrals.

We will not study these p -norms systematically here, as that study would lead us too far afield. However, the corresponding notation has already been introduced in the special cases that p is 1, 2, or ∞ , with the promise that the notation would be clarified later on – which has just been done.

The first problem in studying the p -norms, of course, would be to prove that one has actually defined a norm. We have only done this in the cases that $p = 1$, $p = 2$, and $p = \infty$. It ought to be easy to see that the only real difficulty for the other permissible values of p is in showing the Triangle Inequality, also known in this context as the Minkowski Inequality. One of the ways to prove the Minkowski Inequality, in turn, depends upon Hölder's Inequality, which is a generalization of the Cauchy-Buniakowski-Schwarz inequality. But that is for the future.

Exercises

11.3.5 Show that if $p < 1$ the above definition of a “ p -norm” does not give us a norm in \mathbf{R}^2 (a counterexample will do, here), even though the defining expression is well defined. HINT: Draw a sketch of the “unit ball” when $p < 1$ and find two vectors in it for which the direction of inequality in the Triangle Inequality is reversed.

11.3.6 Show that $\|f\|_p \leq \|f\|_\infty$ for all $f \in C[0, 1]$ when $1 \leq p < \infty$. Also show, if you can, that $\|f\|_p \rightarrow \|f\|_\infty$ as $p \rightarrow \infty$.

11.4 Banach Spaces

The concept of completeness in normed vector spaces is quite important. As we have done in the case of \mathbf{R} itself, we say that a normed vector space V is complete if and only if every Cauchy sequence in V converges. A Cauchy sequence in V is, of course, first of all a sequence, say $\{v_n\}_{n=1}^{\infty}$, in which the terms in the sequence are elements of V . Then, the sequence is a Cauchy sequence provided that, given $\epsilon > 0$ there exists N such that for all $m, n \geq N$ we have $\|v_m - v_n\| < \epsilon$. In other words, the definition is the same as before, except the sequence elements are vectors and the absolute value is replaced by the norm in V .

A normed vector space which is complete is often referred to as a *Banach Space* in honor of the Polish mathematician Stefan Banach, one of the pioneers in the study of complete normed vector spaces.

Exercises

11.4.1 Show that any finite-dimensional normed vector space over \mathbf{R} is complete.

11.4.2 In $C[a, b]$, a sequence of functions $\{f_n\}$ is said to converge **uniformly** to a function f if $\|f_n - f\|_{\infty} \rightarrow 0$. A sequence of functions $\{f_n\}$ is said to converge **pointwise** to a function f if $f_n(x) \rightarrow f(x)$ at each individual $x \in [a, b]$. Show that if $\{f_n\}$ converges to f uniformly and $f_n \in C[a, b]$, then also $f \in C[a, b]$.

11.4.3 In contrast to the previous problem, pointwise convergence of a sequence of functions $\{f_n\}$ in $C[a, b]$ does not necessarily imply uniform convergence because it need not imply that the pointwise limit function is continuous.

There are many fairly standard and obvious counterexamples. As just one of these counterexamples, consider the sequence defined by $f_n(x) = x^n$. Each of these functions is in $C[0, 1]$. But the pointwise limit of this sequence as $n \rightarrow \infty$ is not continuous. Give the details which explain just how this happens.

11.4.4 Show that the space $C[a, b]$ is complete under the uniform norm, in the sense that every Cauchy sequence of continuous functions has a limit, which is another function, and that limit function is also continuous on $[a, b]$. The steps of this problem are to show, first, that there is a pointwise limit function for the Cauchy sequence, and then to show that the pointwise limit function is, under these restricted circumstances, a uniform limit and is a continuous function.

11.4.5 Show (by counterexample) that the space $C[a, b]$ is not complete under the integral norm, nor under the least-squares norm. Note that this shows we have a very different situation in $C[a, b]$ from what happens in \mathbf{R}^2 . There, recall, we had three norms which were similar to the three norms defined here. In \mathbf{R}^2 all three of the norms turned out to be equivalent. But that does not happen in $C[a, b]$.

11.4.6 Construct or describe, if you can, a sequence of functions $\{f_n\}_{n=1}^{\infty}$ for which every f_n is in $C[0, 1]$, $\|f_n\|_{\infty} = 1$ for all n , and for which $\lim_{n \rightarrow \infty} \|f_n\|_1 = 0$.

11.4.7 The goal in this problem is to show that the unit ball in $C[0, 1]$ is not compact when the norm is the standard uniform norm. To show this, it will suffice to construct an open covering for the unit ball and to show that, for that

open covering there is no finite subcovering. The open covering which we will use can be described as follows: for each $f \in C[0, 1]$ such that $\|f\| \leq 1$, choose the open ball of radius $1/4$. That is, the set of all functions g such that $\|g - f\| < \frac{1}{4}$. Now construct a sequence of non-negative functions $\{f_n\}_{n=1}^{\infty}$ (similar, in a way, to which you were supposed to construct in Exercise 11.4.6) which have the following properties:

- a. $\|f_n\| = 1$ for all n .
- b. $\{x | f_m(x) \neq 0\} \cap \{x | f_n(x) \neq 0\} = \emptyset$ if $m \neq n$, so that
- c. $\|f_m - f_n\| = 1$ whenever $m \neq n$.

Having done this construction, show that no two functions from this sequence of functions can be elements of just one set from the given open covering. From this, it follows that no finite collection of sets from the given open covering can cover more than a finite number of the functions f_n and therefore can not cover the unit ball in $C[0, 1]$.

11.4.8 Again, the goal in this problem is to show that the unit ball in $C[0, 1]$ is not compact when the norm is the standard uniform norm. But here we use a different method. Namely, one should construct a sequence of distinct functions in $C[0, 1]$, each of norm 1, which taken together comprise an infinite set which has no accumulation point, thus denying the conclusion of Exercise 11.1.14. The sequence which is described in Exercise 11.4.3 should suffice.

11.5 Linear Transformations and Continuity

First, we should introduce some standard terminology. Let V and W be two vector spaces. A **linear transformation** from V to W is a function $T : V \rightarrow W$ which obeys the properties of linearity already defined elsewhere. Namely, we have for any $v_1, v_2 \in V$ that

$$T(v_1 + v_2) = Tv_1 + Tv_2$$

and for any $\alpha \in \mathbf{R}$ and any $v \in V$ that

$$T(\alpha v) = \alpha Tv.$$

We have also already seen that vector spaces can be very large, as for example $C[a, b]$ which is a vector space in which the “vectors” are, in fact, functions. A function which acts upon functions as inputs and produces some kind of output, usually but not always another function, is often called an **operator**. Thus, a linear transformation in which the space V above is a space whose elements are functions is often called a **linear operator**. There is no logical reason for this, because a linear operator is a linear transformation, after all. But this different name is very much in common use in such contexts.

Also, a linear transformation in which the space W is equal to \mathbf{R} is often called a **linear functional**. Here are some examples:

1. Let $V = \mathbf{R}^2$. Map each vector (x, y) to $x - y$.

2. Let $V = C[a, b]$. Map each function f in this space to $f(a)$.
3. Map each $f \in C[a, b]$ to $\int_a^b f(x) dx$
4. For f differentiable on $[a, b]$, compute $f'(b)$.

Continuity or the lack thereof is also important when considering a linear transformation $T : V \rightarrow W$, in the situation that both V and W have norms. Quite unsurprisingly, we agree of course that a function $f : V \rightarrow W$ is continuous at a given $v_0 \in V$ if and only if

$$\forall \epsilon > 0 \exists \delta > 0, \forall v \in V, \|v - v_0\|_V < \delta \Rightarrow \|f(v) - f(v_0)\|_W < \epsilon$$

We only need to be careful to distinguish above which norm is applicable where, since the norm defined in V and the norm defined in W in fact have nothing to do with each other. We also say, of course, that a function which is defined upon V is continuous if it is continuous at each individual $v \in V$.

There is the question, then, what the above quite standard definition of continuity implies in the special case that the function f is not just any function but is a linear transformation, T . In fact, the consequences are far-reaching:

Exercises

11.5.1 Let V and W be two normed vector spaces, and let $T : V \rightarrow W$, Then the following statements are equivalent.

- i. T is continuous at every $v \in V$
- ii. T is continuous at 0_V
- iii. T is continuous at some particular $v_0 \in V$
- iv. There is a constant M such that $\|Tv\|_W \leq M$ for all v such that $\|v\|_V \leq 1$

In the context of the previous problem, it makes sense to refer to the set of all continuous linear operators from a normed vector space V with range W , another normed vector space, as the set of **bounded** linear operators from V to W . Further, inasmuch as this set of bounded linear operators from V into W is itself a vector space, it makes sense to define a norm upon it, too. For such a linear transformation $T : V \rightarrow W$ we can define $\|T\|$ by

$$\|T\| = \sup_{\|v\|_V \leq 1} \|Tv\|_W$$

Exercises

11.5.2 Show that $\|T\|$ as defined above also satisfies

$$\|T\| = \sup_{\|v\|_V=1} \|Tv\|_W = \sup_{\|v\|_V \neq 0} \frac{\|Tv\|_W}{\|v\|_V}.$$

The above definition for the norm of a linear transformation or operator is quite useful in itself for problems in “hard” analysis which involve actual number-crunching. But just how far do the properties of linearity carry us? This is examined in the final two problems of this section:

Exercises

11.5.3 If the normed vector space V is n -dimensional, where n is finite, then any linear transformation $T : V \rightarrow W$, where W is any other normed vector space, is continuous.

Note: If you prove everything which is behind this problem it will be quite lengthy. You may go ahead and take for granted the relevant facts of a purely algebraic nature, inasmuch as those facts are supposed to have been presented with proofs in a linear algebra course. You may treat it as known, therefore, that all vector spaces of dimension n are algebraically isomorphic to one another, and to \mathbf{R}^n . Also that the dimension of the image of a linear transformation does not exceed the dimension of the domain space. You may also assume that all norms on V give rise to mutually equivalent topologies, as do all norms defined on the image of T , which is a subspace of W . and the same is true for that matter on \mathbf{R}^n .

11.5.4 If the normed vector space V is not finite dimensional, then a linear transformation T defined upon V may or may not be continuous. For example, differentiation is obviously a linear operation which takes differentiable functions to other functions. It is not the case at all that this operation has to be continuous. Give an example which clarifies this statement. Probably your best approach is to show that differentiation is not a bounded linear operation

if defined, for example, on the set of continuously differentiable functions on $[0, 1]$. We can see that this set is a subspace of $C[0, 1]$ and hence we can take its norm to be the usual sup norm on $C[0, 1]$. The action of differentiating any continuously differentiable function then results in an output which is another continuous function. However, it can be seen that there is no upper bound for the norm of the derivative of a function whose norm is 1. Thus, to complete the problem it suffices to give an example of a sequence of continuously differentiable functions which demonstrates this fact.

11.6 Weierstrass Approximation Theorem

The Weierstrass Approximation Theorem states that any continuous function defined on a closed interval can be approximated with arbitrary accuracy by a polynomial. Its strength and importance are underscored by the fact that the theorem says and assumes absolutely nothing about the differentiability of the function to be approximated, and its proof does not assume differentiability, either. In fact, it is the case that functions can be constructed which are continuous on an interval $[a, b]$ but which are differentiable at no point in that interval. The construction of such a function is a bit off-topic here, but it can be done. This fact should be kept in mind when judging the strength of the following result. It says that even such nowhere differentiable functions can be approximated by the procedure which is given in the proof.

Theorem (Weierstrass): *Let f be a function which is defined and continuous on a closed interval $[a, b]$, and let $\epsilon > 0$ be*

given. Then there is a polynomial P such that $\|f - P\|_\infty < \epsilon$.

The proof of this result which will be given below is due to Sergei Bernstein. In order to prove it, he developed a sequence of approximating polynomials based upon the given function, which have afterwards been called the Bernstein polynomials in his honor. The construction of the Bernstein polynomials is given for functions in $C[0, 1]$, and therefore this proof of the Weierstrass theorem would seem to depend upon the particular choice of underlying interval. Therefore, the first thing which needs to be shown is:

Exercises

11.6.1 If the Weierstrass approximation theorem is known to be true on the interval $[0, 1]$, then it is also true, by change of variables, on any closed and bounded interval $[a, b]$.

We define the Bernstein polynomial of a function f at a point x by

$$B_n(f; x) = \sum_{k=0}^n \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$

Note that the Bernstein polynomial B_n of a function is another function, in this case a polynomial of degree at most n . Thus B_n can be regarded as a function whose domain and whose range are vector spaces of functions. Such a function

as B_n which takes functions as inputs is called an *operator*, as described in the previous section. And this terminology seems to fit the situation even better if the outputs are functions, too, as is the case here. If the operator is in addition also a vector space linear transformation, it is called a *linear operator*, as also discussed in the previous section.

Exercises

11.6.2 Show that the Bernstein polynomial operator is linear. That is,

$$B_n(\alpha f + \beta g; x) = \alpha B_n(f; x) + \beta B_n(g; x)$$

11.6.3 Show that the Bernstein polynomial operator is *monotone*. That is, if $f \leq g$ (meaning that $f(x) \leq g(x)$ for all $x \in [0, 1]$), then it follows that $B_n(f; x) \leq B_n(g; x)$ for all $x \in [0, 1]$, too.

11.6.4 Show that $B_n(1; x) = 1$.

11.6.5 Show that $B_n(x; x) = x$.

11.6.6 Show that $B_n(x^2; x) = (1 - \frac{1}{n})x^2 + \frac{1}{n}x$.

11.6.7 It is not immediately relevant to the proof which we are about to do, but $\|B_n\| = 1$.

We now give a proof of the Weierstrass theorem for $C[0, 1]$:

Let $f \in C[0, 1]$, and let $\epsilon > 0$ be given. Let x be chosen randomly from within $[0, 1]$. In what follows, then, x will be considered as fixed but arbitrary. We need to show that, independently of x , there is an integer $N > 0$ such that, for $n \geq N$, $|f(x) - B_n(f; x)| < \epsilon$. Due to the obvious complexity of this problem, we must take an indirect approach.

First, since f is continuous on the closed interval $[0, 1]$, we know that in fact f is uniformly continuous. The uniform continuity of f guarantees that there is $\delta > 0$ independent of x such that for all $y \in [0, 1]$ we have

$$|f(x) - f(y)| < \frac{\epsilon}{2}$$

whenever $|x - y| < \delta$. Furthermore, if y is such that $|x - y| \geq \delta$, then we have the obviously crude estimate that $|f(x) - f(y)| \leq 2M$, where we take $M = \|f\|$ (we are using here the uniform norm). Therefore, for $|x - y| \geq \delta$ we can correctly say that

$$|f(x) - f(y)| \leq 2M \frac{(x - y)^2}{\delta^2}.$$

Now, as one of these estimates is true for $|x - y| < \delta$ and the other is true for $|x - y| \geq \delta$, it is possible to say that on the entire interval $[0, 1]$

$$|f(x) - f(y)| \leq \frac{\epsilon}{2} + 2M \frac{(x - y)^2}{\delta^2},$$

or, equivalently, that

$$-\frac{\epsilon}{2} - 2M \frac{(x-y)^2}{\delta^2} \leq f(x) - f(y) \leq \frac{\epsilon}{2} + 2M \frac{(x-y)^2}{\delta^2}. \quad (11.2)$$

Now, taking y as the variable and x and therefore $f(x)$ as constant, and using the properties of the Bernstein polynomial operator given in the previous problem set, we have

$$B_n(-\frac{\epsilon}{2} - 2M \frac{(x-y)^2}{\delta^2}; y) \leq B_n(f(x); y) - B_n(f; y) \leq B_n(\frac{\epsilon}{2} + 2M \frac{(x-y)^2}{\delta^2}; y).$$

That is,

$$|B_n(f(x); y) - B_n(f; y)| \leq B_n(\frac{\epsilon}{2}; y) + \frac{2M}{\delta^2} B_n((x-y)^2; y)$$

is true for arbitrary y . Since x is fixed but arbitrary and y is the variable, $f(x)$ is a constant as far as y is concerned. Consequently, $B_n(f(x); y)$ is in fact simply equal to $f(x)$ regardless of the value of y .

Now, to estimate the terms on the right in (11.2), we use the linearity of B_n . First of all we have

$$B_n(\frac{\epsilon}{2}; y) = \frac{\epsilon}{2},$$

and using again the fact that x remains at a fixed but arbitrary value,

$$B_n((x-y)^2; y) = B_n(x^2; y) - B_n(2xy; y) + B_n(y^2; y)$$

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

whence for all $y \in [0, 1]$

$$|B_n(f(x); y) - B_n(f; y)| \leq \frac{\epsilon}{2} + \frac{2M}{\delta^2} \left(x^2 - 2xy + \left(1 - \frac{1}{n}\right)y^2 + \frac{1}{n}y\right).$$

From this, it follows in particular when y is put equal to x that

$$\begin{aligned} |f(x) - B_n(f; x)| &\leq \frac{\epsilon}{2} + \frac{2M}{\delta^2} \left(-\frac{1}{n}x^2 + \frac{1}{n}x\right) \\ &\leq \frac{\epsilon}{2} + \frac{M}{2n\delta^2}, \end{aligned}$$

and this is true independently from the original choice of $x \in [0, 1]$. It follows that, if n is chosen to be any integer greater than $\frac{M}{\delta^2\epsilon}$, then

$$|f(x) - B_n(f; x)| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

and the proof is complete.

11.7 Linear operators defined by integral kernels

A kernel is an function of the form $K(x, t)$, and based upon it one can define an operator L for continuous functions on an interval $[a, b]$, using the formula

$$Lf(x) = \int_a^b f(t)K(x, t) dt$$

Clearly, if Lf is to be defined at every x for every $f \in C[a, b]$, then certain reasonable restrictions need to be put on K . These reasonable restrictions include that $K(x, t)$ is defined on $[a, b] \times [a, b]$, that $K(x, t) = K(t, x)$ and that $|K(x, t)|$ is integrable with respect to t for any fixed $x \in [a, b]$.

The function $K(x, t)$ is called a **positive kernel** if $K(x, t) \geq 0$ always.

Exercises

11.7.1 The operator L defined by $Lf(x) = \int_a^b f(t)K(x, t) dt$ is, in fact, linear.

11.7.2 If $|K(x, t)|$ is integrable in t for every x , then

$Lf(x) = \int_a^b f(t)K(x, t) dt$ is defined for every $f \in C[a, b]$.

11.7.3 If $K(x, t)$ is integrable in t for every x and is positive, then

$Lf(x) = \int_a^b f(t)K(x, t) dt$ is a monotone linear operator (see Exercise 11.6.3 above for the definition that a linear operator is monotone).

11.8 Periodic functions and the Fourier series

The space $C_{2\pi}$ is defined to be the set of all 2π -periodic functions. It is obviously valid to view this set as a subset of $C[-2\pi, 2\pi]$ which is closed under addition and scalar multiplication, and hence a vector space in its own right. Moreover, any norm which is imposed upon $C[-\pi, \pi]$ naturally and obviously induces a norm on $C_{2\pi}$.

The class may be familiar with the construction of the Fourier series of a function which is 2π -periodic. To review briefly, the Fourier series of such a function f is a series of the form

$$a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx),$$

in which the coefficients are given by

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) dt, \tag{11.3}$$

and for $k \geq 1$

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos kt \, dt \quad (11.4)$$

and

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin kt \, dt. \quad (11.5)$$

As will be discussed below, these formulas for the coefficients depend upon the following observations about trigonometric integrals. In all of these problems, n and k are assumed to be integers.

Exercises

11.8.1 For $k > 0$, $\int_{-\pi}^{\pi} \cos kx \, dx = 0$, and $\int_{-\pi}^{\pi} \sin kx \, dx = 0$

11.8.2 For $n > 0$ and $k > 0$, $\int_{-\pi}^{\pi} \sin nx \cos kx \, dx = 0$

11.8.3 For $n > 0$ and $k > 0$, with $n \neq k$, both $\int_{-\pi}^{\pi} \cos nx \cos kx \, dx = 0$ and $\int_{-\pi}^{\pi} \sin nx \sin kx \, dx = 0$

11.8.4 For $k > 0$, $\int_{-\pi}^{\pi} \cos kx \cos kx \, dx = \pi$ and $\int_{-\pi}^{\pi} \sin kx \sin kx \, dx = \pi$

It should be noted as well that nothing much was said about f except that, obviously, we should be able to integrate f or else we might well not be able to compute some of the above coefficients. But many functions can be integrated which are not continuous. It is even possible that the function f might have vertical asymptotes or otherwise might behave quite strangely and it is still possible to compute its Fourier coefficients by means of improper integrals. Indeed, ever since the time when the Fourier expansion of a function was first developed, there have been questions about such matters as “Give a complete and accurate description of the set of functions for which we can compute the Fourier expansion” and then after that “Give a complete and accurate description of those functions for which the sequence of its finite Fourier expansions converges to the function.” And the answer to the second question may depend, obviously, upon “In what sense do you intend to describe convergence”? These are in fact deep questions, and the attempts to answer them have motivated a very large proportion of the development of modern mathematics, ever since the questions were proposed approximately 200 years ago. For now, we will be less ambitious. We will go ahead and assume that the function f in question is actually continuous, and we will work with the uniform norm in $C_{2\pi}$. Later is – well, later is later.

Moreover, just because it is possible to construct the Fourier series in a manner which will be explained below, no assumptions should be made as to whether the series converges or diverges, neither at any particular x nor in the sense of norm convergence. If the series does not converge, then it still exists as a formal construction.

The method for the formal construction of the Fourier series is itself of interest and also brings up many problems which need a solution. One begins with the (possibly false) assumption that the series exists and is somehow equal to

the function. That is, we pretend that the equation

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) \quad (11.6)$$

is somehow true. Even if that assumption is false, we press on regardless. Then, supposing that the series does exist and is equal to the function, we can integrate both sides of (11.6) over any interval of length 2π . When we do that, every term on the right vanishes except for the term a_0 for which we get $2\pi a_0$, and (11.3) follows. Similarly, if we multiply both sides of (11.6) by $\cos nx$ or by $\sin nx$ every term on the right again integrates to 0 except for the one involving $a_n \cos nx$ or $b_n \sin nx$, and we get the formulas (11.4) and (11.5) which were given above. Note that hidden behind this procedure there is in fact a very big question. Namely, we have tacitly assumed that the integral of an infinite series is equal, somehow, to the term by term integration of the same series. This is a statement which is obviously not universally valid, and the need is underscored for determining the most general possible circumstances in which it really is a valid procedure. At this time, however, we are merely doing a formal construction, which avoids that nasty question.

Now, having described what the Fourier series is, we define the finite Fourier series S_n to be the n th partial sum. That is, we get to the terms indexed by n and we stop. That is,

$$S_n f(x) = a_0 + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx), \quad (11.7)$$

Using trigonometric identities, it is possible to rewrite the formula for the finite Fourier series, representing it as one integral with an integral kernel. The procedure for doing this is given in the following exercises.

Exercises

11.8.5 Given $f \in C_{2\pi}$ (that is, the space consisting of all continuous 2π -periodic functions), show that the finite Fourier expansion $S_n f$ can be represented in integral form as

$$S_n f(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \left\{ \frac{1}{2} + \sum_{k=1}^n \cos k(x-t) \right\} dt$$

The expression in curly braces is an integral kernel for the operator S_n . It is called the Dirichlet kernel, $D_n(x, t)$.

HINT: Put the definitions for the coefficients given in (11.3) and (11.4) and (11.5) into the formula (11.7) in the place of the coefficients. Then write the result as one integral and simplify by using the identity for $\cos(x-t)$. Note that things like interchanging the order of summation and integration are perfectly feasible here because one is dealing with a finite sum.

11.8.6 Show the identity

$$\frac{1}{2} + \sum_{k=1}^n \cos ku = \frac{\sin \frac{2n+1}{2}u}{2 \sin \frac{1}{2}(u)}$$

Again, when $u = x - t$ the expression seen above is the Dirichlet kernel $D_n(x, t)$ for the finite Fourier series. It is given on the right in closed form.

It follows that $S_n f(x)$ can be written as

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \frac{\sin \frac{2n+1}{2}(x-t)}{2 \sin \frac{1}{2}(x-t)} dt$$

11.9 The Féjér operator

The proof of the Weierstrass approximation theorem depended heavily upon the fact that for each n the operator B_n is a monotone linear operator. It should also be clear from the proof that a monotone linear operator has some very interesting properties.

Here, we define another monotone linear operator and show that when it is represented in integral form it has a *positive kernel*.

The new operator will be σ_n , defined for all $n \geq 1$ and for any $f \in C_{2\pi}$ by the formula

$$\sigma_n f = \frac{1}{n} \sum_{k=0}^{n-1} S_k f.$$

This operator is called the Féjér operator, or the Féjér mean of the Fourier series. At first sight, it is not obvious that σ_n is a positive operator. That requires some proof, which we now provide.

We begin with the observation that, using trigonometric identities, it is possible to represent the operator σ_n in integral form because it is defined in terms of a sum of other operators which we already know have integral kernel representations. Thus, the obvious way to prove that σ_n is a positive operator is to try to write its kernel function in closed form, as was done already with the closed-form expression for the Dirichlet kernel. Indeed, the resulting expression for the kernel function of σ_n is a positive kernel.

Exercises

11.9.1 Given $f \in C_{2\pi}$ we can write

$$\sigma_n f(x) = \frac{1}{n\pi} \left\{ \sum_{k=0}^{n-1} \int_{-\pi}^{\pi} f(t) D_k(x, t) \right\} dt$$

11.9.2 Show the trigonometric identity

$$\sum_{k=0}^{n-1} D_k(x, t) = \frac{\sin^2 \frac{n}{2}(x-t)}{2 \sin^2 \frac{1}{2}(x-t)}$$

so that

$$\sigma_n f(x) = \frac{1}{n\pi} \int_{-\pi}^{\pi} f(t) \frac{\sin^2 \frac{n}{2}(x-t)}{2 \sin^2 \frac{1}{2}(x-t)} dt$$

The integral kernel which was found in this problem is called the Féjér kernel.

It follows from the representation in Problem 11.9.2 that the operator σ_n is a monotone operator. Based in part on this, it can be used to prove an analogue of the Weierstrass theorem for 2π -periodic functions. It is quite interesting to notice that the definition of $\sigma_n f$ is based upon integration of f instead of upon point evaluation of f , whereas the Bernstein polynomial operator depends completely upon point evaluation of f . The result is that the function f which is to be approximated by σ_n in fact only needs to be integrable. That is a far less stringent requirement than what is required for point evaluation. For, point evaluation requires continuity before it can make any sense at all. This is certainly food for thought for the future, as it seems that the Féjér operator can be applied to many more functions than can the Bernstein operator, so perhaps it can be used to approximate lots of those functions, too. The alert students will realize, of course, that the proof of the Weierstrass theorem also depended heavily upon uniform continuity. That clearly applies for a function in $C_{2\pi}$ as well, as a function in $C[0, 1]$, of course. But if one were going to try to generalize, creating a similar argument by extending the class of functions to which the procedure applies, then the question for the future is whether there is any appropriate substitute for uniform continuity when the norm is not the uniform norm and the function is not presumed to be continuous. The answer to that question happens to be “yes, under the right circumstances” but we are not ready at this time to go there. Thus, in what follows we will assume that the function f

to be approximated lies in the safe territory of $C_{2\pi}$ with the uniform norm.

In order to go forward, we will need the results of the following exercises:

Exercises

11.9.3 Compute $\sigma_n(1)(x)$ and $\sigma_n(\cos x)$ and $\sigma_n(\sin x)$.

11.9.4 Show that σ_n is a linear operator.

11.9.5 Show that σ_n is a monotone operator.

11.9.6 (not actually needed in what follows, but interesting) $\|\sigma_n\| = 1$.

11.9.7 Show that for $0 \leq |x| \leq \frac{\pi}{2}$ it is true that

$$|x| \leq \frac{\pi}{2} |\sin x| \tag{11.8}$$

11.10 Weierstrass Theorem for periodic functions

Theorem: *Given $f \in C_{2\pi}$ and given $\epsilon > 0$ there is a trigonometric polynomial T such that $\|f - T\| < \epsilon$*

The proof of this result follows steps which are quite similar to the proof of the Weierstrass Theorem for algebraic polynomial approximation of a function in $C[0, 1]$, but instead of using the Bernstein operator one uses the operator σ_n . The proof is appropriate for a presentation by a graduate student enrolled in MATH 6210, or for presentation in a lecture.

Chapter 12

Finite Taylor-Maclaurin expansions

12.1 Introductory remarks

In the previous chapter, we have looked at some finite-dimensional and infinite-dimensional vector spaces. At the end of the chapter, we considered the two spaces $C[a, b]$ and $C_{2\pi}$, both with the uniform norm. We showed that any function

in $C[a, b]$ can be approximated with arbitrary accuracy by an algebraic polynomial and similarly that any function in $C_{2\pi}$ can be approximated with arbitrary accuracy by a trigonometric polynomial. Most emphatically, both of these results had nothing to do with differentiability. It was explicitly not assumed that the function to be approximated was differentiable; it was only assumed to be continuous.

Here, we narrow the focus considerably. The purpose of this chapter is to begin an investigation into the question of approximating a particular function, given knowledge about the value of the function and its derivatives at a single point. The first attempt to do this is the finite Taylor-Maclaurin expansion. We will soon see that the method has some serious limitations if we do not actually know more about what we are doing, which will lead naturally to the consideration of functions defined by infinite series.

12.2 Finite Taylor expansions of a function

Let us suppose that we have a function f which is known to be differentiable at least $n + 1$ times inside of an open interval which contains the point a . We will use the notation, convenient for what we are about to do, that $f^{(0)}$ means the function with no derivatives taken, and that if $k > 0$ the notation $f^{(k)}$ signifies the k th derivative.

Let us further assume that $f^{(0)}(a), \dots, f^{(n+1)}(a)$ are all known. Then the **finite Taylor expansion** of f at a is the

expression

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

The special case that $a = 0$ occurs quite often, and the name for the expansion which results when $a = 0$ is the **finite Maclaurin expansion** of f . We do not need to study this separately in what follows, as it is indeed a special case of something more general.

We hope to be able to use this finite sum to approximate f . If we wish to do anything constructive or at all useful, we need some way to estimate the error. There are several ways to do this, but we look at one of the most common which is called the Lagrange form of the error. We do this in the following exercises. In doing them, we assume some familiarity with the basic rules of differentiation, of course.

Exercises

12.2.1 Show that the n th derivative of $(x-a)^n$ is $n!$ and the $(n+1)$ st derivative of the same, is zero.

12.2.2 Assume that both a and x are in the open interval upon which f is known to be $n+1$ times differentiable. Consider the fraction defined by

$$\frac{f(x) - \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k}{(x-a)^{n+1}}.$$

Show that this fraction is equal to

$$\frac{f^{(n+1)}(c)}{(n+1)!}$$

for some c which lies between a and x . HINT: Repeat Cauchy's Mean Value Theorem several times in order to reach this result.

12.2.3 Using the result of the preceding exercise, the error in the approximation of $f(x)$ by its finite Taylor expansion around a is

$$f(x) - \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k = \frac{f^{(n+1)}(c)}{(n+1)!} (x-a)^{n+1}$$

This is referred to as the **Lagrange form** of the error. Whether the error is given in this form or in some other form, the error will be referred to as $R_n(f, x)$.

The reader should note several things at this point.

First of all, the expression for $R_n(f, x)$ which is given in the preceding exercise is, in fact, theoretically exact. The only problem with using it – which indeed is sometimes a major problem – is that one does not know and in general can not know exactly where the point c is. One only knows that c lies somewhere between a and x . It follows that, even though the expression is exact, the exactness is only theoretical. In practice we cannot get our hands upon c except in very simple problems. Consequently, we can not compute the error but only estimate it. Error, after all, is something

which almost by definition we do not exactly know. If we were to know it exactly then it would not really be error, would it? For, then we could compute our approximating sum and then just add the error (or subtract it, if appropriate) and we would magically arrive exactly at the desired result. No, formulas for the error $R_n(x)$ which are theoretically exact are only useful in estimating the error, as will be made more clear below.

Second, we can not merely assume without proof that the finite Taylor expansion of a given function f will automatically converge at a given x when expanded around a given a . To rephrase this, even if we know that all the assumptions which went into the derivation of the formula for the finite Taylor expansion are true for all n , then we still do not know that the error will converge to 0 as $n \rightarrow \infty$. Neither, by looking at some estimate for $R_n(x)$, do we necessarily know that it will not converge, either.

Third, because of our lack of detailed knowledge about the precise location of c we have to assume the worst about c if we want to have a trustworthy estimate for the error. If even when we assume the worst about c in the estimate given in Exercise 12.2.3 the value of $R_n(f, x)$ tends to 0 as $n \rightarrow \infty$, then we can be sure that we can actually succeed in the effort to estimate $f(x)$ to a given, prescribed degree of accuracy. Remember, error is what we don't exactly know. We can only estimate it. If we actually knew the exact value of the error, then it would be pointless to use such a complicated procedure to find values for a function, would it not? It should go without saying that similar remarks would hold true about any other formula which gives a theoretically exact expression for $R_n(x)$.

Fourth, there may arise situations in which we are suspicious that the finite Taylor expansion of f is misbehaving badly as $n \rightarrow \infty$, and we would like to prove that the procedure is not behaving well. But again because of our lack of detailed knowledge about the precise location of c we may not be able to show that the procedure is, in fact, misbehaving.

For, if we wish to show that the procedure is **not** working we would need to assume the best about c , which is probably $c = a$, or close to it, and then see if the lower estimate for the error is misbehaving even if we make such a favorable assumption. Unfortunately, when one makes this assumption the error estimate usually would appear to converge to 0 quite nicely even though common sense ought to be telling us that convergence is improbable, or even impossible. The point is, appeal to a formula which gives $R_n(x)$ is usually a poor way to attempt to show that divergence is occurring, even if it is strongly suspected.

Now, the above observations do not say that there are no other sources of information, no other way to tell what is happening. There are other sources of information. There are other ways to tell what is happening. But the point is that we need to learn about some of those, too. What we have seen is not the full story. We will do that in the next chapter. But before doing that we should look at a couple of examples which illustrate the above points.

Exercises

12.2.4 Consider the function $f(x) = (1 - x)^{-1}$. Find a general formula for its n th derivative. Note that this is most efficiently done by repeated use of the Chain Rule, not by repeated use of the Quotient Rule, which is a most inefficient way to proceed. Use this information to write the finite Taylor-Maclaurin expansion out to n terms, and also the Lagrange form of the remainder $R_n(x)$, (the error).

12.2.5 We actually already know about the function $f(x) = (1 - x)^{-1}$ that it is the value obtained from the

summation of the geometric series

$$\sum_{k=0}^{\infty} x^k$$

We also know from some very basic considerations, discussed in detail in a previous chapter, that this infinite series must converge to the advertised function whenever $|x| < 1$ and must diverge for all other values of x . Consider what happens at the value $x = 9/10$. Pretend temporarily that you do not know we are dealing with a convergent infinite series, and investigate the convergence or divergence of the finite Taylor-Maclaurin expansion, based upon nothing but the error estimate which you obtained in the previous problem. Based solely upon that error estimate and not upon knowledge from other sources, such as our familiarity with the geometric series, explain why one cannot say for certain when $x = 9/10$ what the error term is actually doing as $n \rightarrow \infty$.

12.2.6 Same problem as the previous one, but let's see what we can conclude from that Lagrange form of the error estimate about convergence or divergence if we put $x = -9$? And what is happening in reality, based upon our knowledge of the infinite geometric series?

12.2.7 Do the previous three problems if the function is $f(x) = \log(1 - x)$ and watch the same problems occur. For the purpose of this exercise, you can go ahead and assume that you have a calculus-course level of knowledge about the logarithm function, but do take note that what is called here \log was probably denoted by \ln in the calculus

book. Also, for purposes of the comparison of what happens at the two points -9 and $9/10$ is concerned, please be aware that the infinite series expansion of $\log(1 - x)$ is valid if and only if $x \in [-1, 1)$.

Clearly, there is only one conclusion which can be drawn from the previous exercises. Namely, sometimes the Lagrange form of the error estimate for the finite Taylor or Maclaurin expansion of a function can give us useful information, and sometimes it can not. In many circumstances, more knowledge is required than what can be gleaned from the Lagrange form of the error term. One of the possibilities is to look for some other form in which to represent the error estimate for the finite Taylor-Maclaurin expansion. Other representations of the error term do indeed exist, and some of those other representations do indeed work better for the functions which we have just used as horrible examples. One point of these exercises is to serve as a lead-in to those other forms of the error estimate. But another point is to make sure that we are aware of potential shortcomings. We need to agree that more awareness of what to expect when procedures similar to a Taylor-Maclaurin expansion are put into effect is needed. In the next chapter, we will take a systematic look at functions which are defined by a **power series**, which is an infinite series whose terms involve successive powers of the input, x .

Meanwhile, there are at least two other expressions for the error estimate $R_n(f, x)$.

Exercises

12.2.8 Report on the Integral expression for $R_n(x)$ and Cauchy's form for $R_n(x)$ (there is a good discussion of these two items at the Wikipedia page called "Taylor's Theorem").

12.2.9 Does either the integral expression for $R_n(x)$ or Cauchy's expression for $R_n(x)$ solve the problems which have been unearthed in the previous exercises relating to the geometric series?

Chapter 13

Functions given as series

13.1 Introductory remarks

When considering the approximation of a function by a finite sum in the previous chapter, we saw that the error estimate used there can fail to give meaningful results at some locations where the function itself is actually defined. When we

are caught by this sad state of affairs, we can neither show convergence to the function value nor divergence from the function value. This situation may obtain in spite of the fact that good error estimates are necessary in order to see whether or not the finite Taylor expansion of the function is actually providing good and useful approximation, or not. We saw, for example, that the available error estimate for the finite Taylor expansion of $f(x) = \log(1 - x)$ (note that inside of mathematics “log” is the function which the calculus book has denoted by \ln) behaves particularly badly at certain values of x . Two particular instances were discussed, which taken together show that there is a serious difficulty. First, the attempt to compute $\log 10$, which requires the use of $x = -9$, caused a problem. If one were hoping to show that the procedure works, then one would need to assume that the worst possible thing is occurring and if the procedure still converges then all is good. But if one is assuming the worst, then one is led to an estimate for the error $R_n(f, x)$ which explodes as $n \rightarrow \infty$. Therefore, we can not prove that the procedure is working (which, by the way, it really is not, but the point is that the bad behavior of a worst-possible error estimate based upon Lagrange’s formula for the error $R_n(x)$ **does not prove** the divergence. On the other hand, if one assumes the best possible behavior of the error $R_n(f, x)$ then the procedure would seem to behave quite nicely. For, the best possible error estimate does indeed converge to 0 as $n \rightarrow \infty$. Hence, we can not prove by this means that the finite Taylor expansion with $x = 9$ is **not** working, either.

The clever might consider the idea of computing $\log \frac{1}{10}$ instead of $\log 10$. Then one must use $x = \frac{9}{10}$ instead of $x = -9$ as before, and one might hope that things will work out a little bit better. Indeed, to some extent things do seem to work out better. The finite Taylor expansions do seem to be converging to **something**. Nevertheless, if one sticks to the method of error estimation which is based upon Lagrange’s expression for $R_n(x)$ obtained in the previous

chapter then one is no better off at all. That is, by appeal to nothing but Lagrange's expression for $R_n(f, x)$ which is presented in the previous chapter, one can still prove neither convergence nor divergence.

Exercises

13.1.1 Indeed, the logarithm function is not the only function for which the error estimate given in the previous chapter is not working very well. This problem even occurs with the geometric series, which is perhaps the most basic series of all.

- (i) Show that the finite Maclaurin expansion of $f(x) = \frac{1}{1-x}$ out to n indeed agrees with the n th partial sum of the geometric series

$$\sum_{k=0}^{\infty} x^k$$

- (ii) Note that we already know this series converges when $-1 < x < 1$. But nevertheless you can not use the error estimate of the previous chapter to prove this. Determine for which values of x within the interval $(-1, 1)$ the error test of the previous chapter fails to demonstrate convergence for this series.

The above examples demonstrate clearly that we need to take a step back and look at other background information about what is happening here. To do this, we need to discuss what is happening with a series representation of a function in the first place.

13.2 Functions defined as series

In Chapter 4, among other things several crucial aspects of the theory of convergence and divergence of an infinite series have already been presented. In particular, the concepts of absolute and conditional convergence have been introduced in Exercise 4.3.1. The series which were discussed there were almost exclusively series whose terms involved numbers and not variables. The exception to this general rule was Section 4.6, in which there was brief mention of the series

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

It was pointed out back in Section 4.6 that, in fact, this series could be shown to give the same computational result as the definition of a function of x defined by

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n,$$

so long as we would stick to rational values of x . Here, we need to generalize this discussion. A series in which the terms involve successive powers of x is called a **power series** and the first question to resolve, naturally, is the question of convergence. Well, actually, even before that we need to agree about precisely what is meant by a series of the form

$$\sum_{k=0}^{\infty} a_k x^k$$

or alternatively (and this is actually not something different because it can be achieved from what is above by a substitution)

$$\sum_{k=0}^{\infty} a_k (x - a)^k$$

where a is some fixed number. We can and do assume that this is a series with terms in it which are numbers, given any particular x . Thus, the question of convergence or divergence of the series becomes dependent upon the choice of x . We give details about this matter in the next section. Before we get started on that, though, we should note that we have to agree about the meaning of the term $a_0 x^0$. We agree that this signifies, simply, a_0 for **all** values of x , including when $x = 0$. This does need explicit mention because in some other situations, it is well known that 0^0 can not be unambiguously defined. But in connection with power series it is a convention universally agreed on, that $a_0 x^0 = a_0$ for all values of x without exception.

On to the next section.

13.3 Convergence of a power series

Given any specific $x \in \mathbf{R}$ the series

$$\sum_{k=0}^{\infty} a_k x^k$$

becomes a numerical series, and we can test for its absolute convergence by considering instead the convergence of the series

$$\sum_{k=0}^{\infty} |a_k x|^k.$$

The Root Test in its most general form is given in Exercise 6.5.8 and previously in Exercise 4.7.1 in a less general form. The Ratio Test is also given in Exercise 4.7.2. Each of these tests, if it can actually be applied to a given series, will give the same results regarding convergence as can be derived from the general form of the Root Test in Problem 6.5.8, but on some occasions the generalized form of the Root Test is the only one which can be applied. Using it, we note that the series

$$\sum_{k=0}^{\infty} a_k x^k.$$

must be absolutely convergent if

$$\limsup |a_n x^n|^{\frac{1}{n}} < 1$$

and if the \limsup is greater than 1 then the series diverges. Also, if the \limsup is exactly equal to 1 the question of convergence or divergence remains unresolved by appeal to the Root Test alone.

From the above analysis, it is clear that the convergence of a power series is or at least may be dependent upon x .

First of all, if $x = 0$ the above \limsup is zero which is less than 1, guaranteeing convergence. And moreover, substituting $x = 0$ into the series shows that the series does indeed converge because it is a sum of exactly one term beyond which all other terms are seen to be zero.

More generally, the above observations lead to the following:

If

$$0 < \limsup |a_n|^{\frac{1}{n}} < \infty,$$

then let r be defined by

$$r = \frac{1}{\limsup |a_n|^{\frac{1}{n}}}$$

Then the series converges provided that $|x| < r$, diverges if $|x| > r$, and we are left to figure out what happens at each of the two points $x = \pm r$.

If $\limsup |a_n|^{\frac{1}{n}} = \infty$, then the given series must diverge for all $x \neq 0$ but it does converge when $x = 0$, of course, to the value of a_0 . It does no harm to say in this case that $r = 0$. But if $\limsup |a_n|^{\frac{1}{n}} = 0$, then convergence clearly occurs for all x . In this case, we say that $r = \infty$.

The number r calculated above is called the **radius of convergence** of the given series. Absolute convergence always occurs when $x = 0$, and if $r > 0$ then absolute convergence will always take place in the open interval $(-r, r)$. At the endpoints of this interval, anything can happen. The given series might diverge at both ends, or converge at both ends, converge only at one end and not at the other, and any convergence which takes place at an endpoint may be

absolute, or may be conditional. In short, anything can happen there, and individual cases must be examined. However, it is the case that on the complement of the interval $[-r, r]$ the series will diverge, and will diverge in very bad and dramatic fashion.

Finally, for all of the above observations there are corresponding statements which can be made when the original series was of the form

$$\sum_{k=0}^{\infty} a_k (x - a)^k$$

The only differences with the previous case are that convergence is automatic if $x = a$, and that the interval of convergence is centered at a instead of at 0.

More generally, it ought to be clear that one can do “ u -substitution” in any power series. That is, if f is a function which is given by a series, such as

$$f(x) = \sum_{k=0}^{\infty} a_k x^k,$$

then one can replace x by u and obtain

$$f(u) = \sum_{k=0}^{\infty} a_k u^k$$

The radius and the interval of convergence of the new series will obviously need to be adjusted in accord with the

substitution when this is done. There are many possible uses of this trick. At the moment, let us illustrate it by applying the method to four fairly simple examples:

We already know the geometric series. If we turn it into a power series by writing x in place of r , then we get

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k.$$

Now, we can make a substitution and get a series for $1/(1+x)$ by the substitution $x \rightarrow -x$, obtaining

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

The radius of convergence does not change. After doing this, we can replace x by x^2 and again the radius of convergence does not change. We get the series representation

$$\frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k}.$$

Finally, let us give an example of a series where the radius does change.

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

13.4 Differentiation and integration of power series

Differentiating or integration of a function which is given by a series representation may be done term by term, using the familiar formula that gives the derivative of a constant as 0 and gives the derivative of $a_n x^n$ as $n a_n x^{n-1}$ for n any integer such that $n \geq 1$. That is, if

$$f(x) = \sum_{k=0}^{\infty} a_k x^k$$

(which is, of course, defined on the interval of convergence of the series on the right), then it will be seen that

$$f'(x) = \sum_{k=1}^{\infty} k a_k x^{k-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k.$$

The radius of convergence does not change when this is done, though convergence at an endpoint of the interval of convergence may be lost. At this stage, these statements are asserted, not proved. Their proofs will come in Exercises 13.4.1 and 13.4.2, which follow below.

Similarly,

$$\int_0^x f(t) dt = \sum_{k=0}^{\infty} a_k \int_0^x t^k dt = \sum_{k=0}^{\infty} \frac{a_k}{k+1} x^{k+1}$$

The radius of convergence does not change due to integration, though convergence at an endpoint of the interval of convergence may be gained. Again, these statements are asserted here, not proved. Again, their proofs are the topic of Exercises 13.4.1 and 13.4.2.

Exercises

13.4.1 Show that the procedures described above for differentiation and integration of a power series do not change the radius of convergence. Hint: Use the Root Test as given in Exercise 6.5.8, and use the results of Exercise 4.7.5.

13.4.2 Show that the formulas given above for differentiation and integration of a series are in fact the valid procedures to use for obtaining the derivative and the integral of the function f defined by the original series. Note that, once we have shown this, it follows that the function defined by the original series is also continuous, in accord with Exercise 10.1.2,

13.4.3 Suppose that the series which defines the function is

$$\sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Show that the derivative of the function defined by this series is equal to itself. This function, which was already mentioned back in Chapter 4, is called the **exponential function**. We now know that this function is both continuous and differentiable. Also back in Chapter 4, a proof was outlined which shows that the function which is thus defined agrees with e^x for all rational values of x . The series above is valid for all values of x , and we can thus define $f(x) = e^x$ to be the function defined by the series for all x .

13.4.4 Show that for the exponential function there is an inverse function. Further, find the domain of this inverse function, and find its derivative. This inverse function has not previously received any formal treatment in this text, and thus we do not officially claim to know anything about it. This inverse function is called the **logarithm** $\log x$ or the **natural logarithm** $\ln x$. To do this Exercise, you will need the results in Exercise 10.1.11. Please note that you can not use implicit differentiation here. Implicit differentiation is a perfectly valid approach. But, alas, a proof that implicit differentiation actually works is not covered in this course. It is not possible to cover everything. A systematic treatment of implicit differentiation would require an over-long excursion into the calculus of functions of more than one variable. Sorry.

The inverse function of the exponential function is called the **logarithm** $\log x$ or the **natural logarithm** $\ln x$. What

you are supposed to do in this problem is to prove that this function exists, to find its domain, to show that it is differentiable, and to show that its derivative is what you have been taught that it is.

Remark on notation: Note that the notations $\log x$ and $\ln x$ are both in common use for this function. Mathematical purists use the notation $\log x$ almost exclusively, on the grounds that there really is no “logarithm” other than the inverse of the exponential function, and further that there is not even any “exponential function” which can be intelligently defined other than with reference to the one which has been defined above, for reasons which are similar to the reasons why we measure angles in radians and avoid using degrees when discussing trigonometric functions. Thus, since almost all mathematicians are purists, the notation $\log x$ instead of $\ln x$ prevails almost exclusively in the field of mathematics outside of service courses, all other so-called “logarithm functions” being then defined only relative to it and in terms of it. The notation $\ln x$ is in common use in many other fields, where it is intended to denote the “natural” logarithm function of x . Part of the reason why these other fields do this is that logarithms to the base 10 (“common” logarithms) were historically used for heavy numerical calculations and still appear in some applications. Thus, in those fields $\log x$ means $\log_{10} x$ which is the inverse function of 10^x , and $\ln x$ means the same as what they call $\log_e x$, which is the inverse function of e^x , as described above.

13.4.5 In accord with the previous problem, show that the derivative of $\log(1+x)$ is $\frac{1}{1+x}$. Then show that then the series representation of $\ln(1+x)$ can be found by integration, in accord with Exercise 13.4.2. Find this series and fully describe its interval of convergence, including what happens at the endpoints. Describe the discrepancy, if any, between the domain of the function and the interval of convergence of the series.

13.4.6 The derivative of $f(x) = \arctan x$ is $\frac{1}{1+x^2}$. Thus, the series representation of $\arctan x$ can also be found by integration. Find this series representation and fully describe its interval of convergence, including what happens at the endpoints.

13.4.7 Show that the result of the previous problem can be used to get a series which converges to $\frac{\pi}{4}$, which if multiplied by 4 will give us π to arbitrary accuracy. Show that the Alternating Series Test can give us an error estimate for this calculation. Would anyone actually want to use this as a method for calculating π to several decimal places? Why or why not?

13.4.8 Show by using the formula for the tangent of the sum of two angles that

$$\arctan 1 = \arctan \frac{1}{2} + \arctan \frac{1}{3}$$

Give an error estimate which is valid for N th partial sum of each of the series expansions for $\arctan \frac{1}{2}$ and $\arctan \frac{1}{3}$, and for the N th partial sum of what is obtained when the two separate partial sums are added together. Does this approach seem to be more efficient than the one in the previous problem? Find the smallest value of N which is needed if we intend to approximate π , good to four decimal places.

13.5 Double sums

Suppose that we have two absolutely convergent series, $\sum_{i=0}^{\infty} a_i$ and $\sum_{j=0}^{\infty} b_j$. Then, we may notice that the product of the two series, given by

$$\left(\sum_{i=0}^{\infty} a_i \right) \left(\sum_{j=0}^{\infty} b_j \right)$$

may be rewritten as

$$\sum_{i=0}^{\infty} \left(a_i \sum_{j=0}^{\infty} b_j \right),$$

and, in turn, the result is equal to

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i b_j$$

That is, the product of the two series can be represented as a double sum. This double sum can be validly manipulated in a number of different ways. Here, we only need to notice one of them. Namely, one can add up the above double sum

by adding up along the diagonals where $i + j$ is constant. When we do this, the double sum is seen to be equal to

$$\sum_{n=0}^{\infty} \sum_{k=0}^n a_k b_{n-k}.$$

An immediate application is to establish one of the basic properties of the exponential function.

Exercises

13.5.1 Provide the justifications which are needed in order to carry out the steps given above, which are used to write a product as a double sum.

13.5.2 Use the series definition of the exponential function in order to show that $e^x e^y = e^{x+y}$

13.6 The rest of the story

The entire previous discussion is based upon the assumption that we have a function which is “given by a series.” So long as we discuss a function which is given by a series, the above results are certainly true, that the derivative and the integral of the series will give the derivative and the integral of the function, and all within the radius of convergence,

of course. It is also true that if a function is given by its series then the Taylor or Maclaurin expansion for the function will recover exactly the series by which the function is given.

But what if we ask about this the other way around. Let us suppose that a function is continuously differentiable as many times as one wants, inside of some open interval $(a - \delta, a + \delta)$, centered of course at a . Does it then follow that this function must necessarily have a Taylor or Maclaurin series which “gives” the function on all of that interval? Does it then follow that the function is “given” by its Taylor or Maclaurin expansion around the point a ? Unfortunately, the answer to this question is, in general, negative. This situation seems to be unreasonable and thus causes many problems for the students in classes like calculus, where the students are so eager for cut-and-dried answers that sometimes they try to ignore the hard facts and just invent cut-and-dried answers instead. Probably, the situation does not make us happy here, either. But in the context which we have been working, it simply is what is happening and has to be accepted as such. Facts, after all, are facts. Here are a couple of examples.

Exercises

13.6.1 We have already derived the series expansion for $f(x) = (1 + x^2)^{-1}$ and we may remember, or can easily compute, the radius of convergence, which is 1. But have another look. What, exactly, is the problem, here? Why is this happening? The function is, after all, infinitely many times continuously differentiable for $-\infty < x < \infty$. So what is going on? So why is the interval of convergence not equal to the whole real line? Could it be relevant, that the function is not defined at either of the complex numbers $\pm i$, and by strange coincidence the distance from either

of these points in the complex plane from 0 happens to be one unit? What does that have to do with anything if we are just working with real values of x ? Give an answer if you can.

13.6.2 The function $f(x) = e^{-\frac{1}{x^2}}$ is defined for all $x \neq 0$, and is continuous on all of \mathbf{R} if we additionally define its value at 0 to be zero. And in like fashion the n th derivative of this function exists and is continuous whenever $x \neq 0$, and is continuous on all of \mathbf{R} if we additionally define its value at 0 to be zero. Clearly, then, the Taylor series for this function is the series in which all of the coefficients are zero, and therefore the output of that series at any x , or of any partial sum of that series at any x , is 0. But the function itself is clearly not identically equal to zero. It follows that this function is not given by its Maclaurin series. What is the problem? Why is this happening? A hint may be seen in that, if one allows x to take on a pure imaginary value, then the limit of the function value as $x \rightarrow 0$ is not zero at all, but is infinite instead. Thus, if we let x to represent a complex number instead of a real number, then any attempt to assign a defined value for this function at 0 will fail to make the function continuous there.

In both of the examples above, it made things a bit clearer if one looks in the complex plane instead of just considering what happens on the real line. Indeed, the answer does lie in that direction. One needs to study complex analysis, and then one can learn the rest of the story.

Chapter 14

Integrals on Rectangles and Fubini's Theorem

14.1 Integrals defined on Rectangles

Let us assume that we have a closed rectangular region R in \mathbb{R}^2 , bounded by the lines $x = a$, $x = b$, $y = c$, and $y = d$, and that we have a function $f(x, y)$ which is defined on the rectangle R . Then, we can construct an integral of f on the

rectangle in a manner which we proceed to describe.

As we have constructed partitions of an interval, we may construct partitions s_0, \dots, s_m of the interval $[a, b]$ and t_0, \dots, t_n of the interval $[c, d]$. Then, based upon these, we can construct a partitioning of the rectangular region R by means of the grid which is generated by the lines $x = s_i$ for $i \in \{0, \dots, m\}$ and the lines $y = t_j$ for $j \in \{0, \dots, n\}$. We may, for the sake of definiteness, label each of the smaller closed rectangles thus generated by the point at its upper right hand corner, so that the closed rectangle R_{ij} will be that rectangle which is bounded on the left by $x = s_{i-1}$, on the right by $x = s_i$, below by $y = t_{j-1}$, and above by $y = t_j$. We can define the mesh of this rectangular partition by any reasonable manner, such as taking the length of a diagonal of the rectangle, or by taking the mesh to be $\max\{s_i - s_{i-1}, t_j - t_{j-1}\}$, or by taking the mesh to be $s_i - s_{i-1} + t_j - t_{j-1}$. Any of these will work equally well, for the intended purpose.

Now, we can define upper and lower Darboux sums associated with the given partition and with the function $f(x, y)$ to be integrated, as

$$U = \sum_{i=1}^m \sum_{j=1}^n (s_i - s_{i-1})(t_j - t_{j-1}) \sup_{(x,y) \in R_{ij}} f(x, y) \quad (14.1)$$

and

$$L = \sum_{i=1}^m \sum_{j=1}^n (s_i - s_{i-1})(t_j - t_{j-1}) \inf_{(x,y) \in R_{ij}} f(x, y) \quad (14.2)$$

Having defined these sums, we would hope to be able to complete the rest of the definition of the integral of f over the region R in a manner very similar to what was done before, for the integral of a function of one variable. Namely, we

would wish to say that, if any sequence of partitions is taken for which the mesh tends to zero, the associated lower and upper sums always have the same limit, or, it would be nice to be able to say that all lower sums for a given function are always less than all upper sums, motivating the definition of integrability of f by saying that the integral exists when the supremum over all possible lower sums is equal to the infimum over all lower sums, just as was done before. That is, just as was done before we would like to prove the two-dimensional version of the inequality (10.5) and then to say that the integral exists in case that the two sides of this inequality are equal. In order to make this work, of course, we need to define what is meant by a refinement of a two-dimensional partition such as we are dealing with, here.

When we have done with the above preliminaries, we will have an integral which we will denote as

$$\int \int_R f(x, y) dA \tag{14.3}$$

in which A signifies “area.”

Exercises

14.1.1 Construct an intelligent and intelligible definition of a refinement of a partition, in the present context.

14.1.2 Show that the inequality (10.4) holds in the present context.

14.1.3 Show that if the function $f(x, y)$ is a continuous function on the rectangle R , then the integral denoted in (14.3) exists. Note that in order to prove this one needs the results contained in Exercises 5.6.4 and 11.1.12.

14.1.4 Can we similarly take a definition for the integral denoted in (14.3) by means of a Riemann-style definition? Show how this can be done, using sums of the form (10.7) (Riemann sums) as a starting point, and show that the Riemann method for the definition is in fact equivalent to the Riemann-Darboux development given above.

14.2 Fubini's Theorem

The definition of the integral given in the previous section ought to seem clear enough, but there is that little problem of how to compute the value of a given integral. If we would have no better way than to compute sums and take limits, then probably we would all rather be excused. So, is there any other way? Well, yes, there often is. Under the right conditions, we can integrate first by either x or by y , and then by the other. But to show that this actually works, we need to do quite a few preliminaries. In what follows, let us assume that the function $f(x, y)$ is continuous. In what follows, we will also use the labeling conventions laid out in the previous section. The function is presumed to be defined on a closed and bounded rectangular region, R , which will be defined exactly as in the first paragraph of the previous section.

First, we need some very basic results about “partial” integration.

Exercises

14.2.1 Let f be continuous on the closed rectangular region R . Then for any fixed y with $c \leq y \leq d$ the integral

$$\int_a^b f(x, y) dx$$

exists and defines a function of y . Let us call it $h(y)$. Moreover, the function $h(y)$ is a continuous function of y .

14.2.2 (this is not really a separate problem, as it can be done merely by switching labels in the previous problem) Let $f(x, y)$ be continuous on the rectangle R . Then for each fixed x such that $a \leq x \leq b$ the integral

$$\int_c^d f(x, y) dy$$

exists and defines a function of x . Let us call this function $g(x)$. Moreover, the function $g(x)$ which we have just defined is a continuous function of x .

Having done these preliminaries, we can state the theorem of Fubini:

Theorem: Let f be a continuous function defined upon a closed and bounded rectangle, R , bounded by $x = a$, $x = b$, $y = c$, and $y = d$. Let the functions $g(x)$ and $h(y)$ be defined as in Exercises 14.2.2 and 14.2.1. Then all three of the integrals $\int \int_R f(x, y) dA$ and $\int_a^b g(x) dx$ and $\int_c^d h(y) dy$ exist, and they are all equal to each other.

Exercises

14.2.3 Prove the Theorem of Fubini. Note that there are complications. There are several intermediate results which are needed. Moreover, some of those intermediate results are “obviously true” in that we have indeed proven things which are very similar but which are not identical. If your proof of Fubini’s Theorem will require the use of any such “fact” which is “obviously true” but which is not previously proved or is not given as an Exercise, then you need to prove it here. On the other hand, if the intermediate result which you need is found in some previously presented Exercise or theorem in the text, then please refer to that result by its location as well as by quoting it. Also, it bears saying again with emphasis that one should please stick to what is in this text. In more advanced courses in analysis, the integral is generalized and the class of integrable functions is also generalized. When that (which is more or less another year of careful step-by-step development that *begins* from what is covered in this course!) has been done one of the things which is done after that is to prove Fubini’s Theorem again in a much more general context. Please be aware that a “proof” which depends upon all of that material which you have not yet seen is not acceptable here.

Indeed, the proof of Fubini’s Theorem seems to use everything we have learned in a year of real analysis, and even a little bit more. No wonder it is always just presented in calculus courses, without proof.

Chapter 15

The Stieltjes Integral

15.1 A broader view of Integration

This section is intended to do two things. First, it is intended to motivate the definition of the Riemann-Stieltjes integral in the next section of this Chapter. The Riemann-Stieltjes integral is also the last topic in this text. Second,

this section will describe some problems which remain unsolved and unaddressed. It is hoped that this list of so far unsolved problems will help students who intend to go forward from here, giving them a sense of perspective about what they are doing in some of their future courses, and why.

Let us begin by looking again at the evolutionary development of the Riemann integral, which we have already studied.

The Riemann, or the Riemann-Darboux, integral took its motivation from the need to compute areas. The concept very soon got extended from that basis, as it was seen that to use the integral exclusively as a method for computing area would in fact be very inefficient. One might recall that the initial definition of the Riemann-Darboux integral was for a function which is defined on an interval $[a, b]$ and was nonnegative on that entire interval. At this point, the integral gave us a general method for computing the area above the x -axis and below the graph of $y = f(x)$ which lay between the vertical lines $x = a$ and $x = b$. But then the next thing done was to drop the assumption that the function is nonnegative. However, the same definition for the integral was kept. After this, the integral could no longer be said to provide the area between the graph of $y = f(x)$ and the x -axis. At this point, a choice had to be made. Either one kept the same definition, which allowed one to say that for two functions f and g and for any constant c the two equations

$$\int_a^b (f(x) \, dx + g(x)) \, dx = \int_a^b f(x) \, dx + \int_a^b g(x) \, dx$$

and

$$\int_a^b cf(x) dx = c \int_a^b f(x) dx$$

were true, or else one had to keep to an understanding that the integral would always give us area. To keep to a requirement that the integral would always compute area, it would have been necessary to redefine the integral to cover all of those cases where $y = f(x)$ might be negative instead of positive. Even worse, the above two properties are the properties of *linearity*, which are too nice to throw away. For, that linearity is essential to the proof of the Fundamental Theorem of Calculus. And without the Fundamental Theorem of Calculus the only way to compute an integral would be by direct application of the definition, which involved the limits of certain sums. Without doubt, almost everyone would prefer to be excused. And thus linearity triumphed over strict adherence to the computation of area.

The above remarks related specifically to the Riemann integral. Now let us look at some broader perspectives, which will end with a return to the topic of the integral again at the end and lead up to the next section of this Chapter.

During the past century, or, taking things a bit further back, of the past one and a half centuries, has been a constant search for similar features between entities and concepts which have arisen in different areas of mathematics.

Consequently, there has been a natural tendency toward abstraction, which means that one looks at many particular cases and examples and tries to distill out the common core of those different things or different areas, to study the commonalities as such, to try to draw conclusions by trying to develop axiomatic systems which assume only those commonalities, and then to study that system as such. Without these constant efforts at abstraction, the field of mathematics would surely by now have become totally unmanageable. A non-exhaustive list of entire areas of

mathematics which owe their very existence to the tendency toward abstraction would include two areas which very much overlap with real analysis. These two are the fields of general topology and linear algebra.

As to general topology, it is obvious that one of the major motivations which led to its development lay in the study of certain properties of the real number system and of functions defined upon the set of real numbers, followed by a realization that the questions which need to be answered for real-valued functions of a real variable come up again and again in other contexts, with only small differences. General topology has gone far beyond its origins, but in turn it has provided a framework of concepts and terminology which make it much easier to present or to learn those concepts in a course on analysis. The concepts and terminology were already useful in dealing with real-valued functions of a real variable and became indispensable when dealing with normed vector spaces, both n -dimensional and infinite-dimensional. The concepts and terminology of general topology also were also important in developing integration on rectangles, and in proving Fubini's theorem.

The importance of linear algebra in analysis starts with the study of finite-dimensional vector spaces and continues with the fact that much of our work in studying functions involves replacing them with similar functions which are somehow "close" to the original ones, but which are much simpler and more accessible. For example, the finite Fourier expansion and other bounded linear operations provide approximation by trigonometric polynomials of degree at most n which comprise a finite-dimensional vector space. There are many other examples, too. Also, sometimes the linear operators or a sequence of linear operators map functions or approximations of them into an infinite-dimensional space with a simpler structure. We have seen this put into practice using the Weierstrass theorem, proved both with the Bernstein polynomial operators and with the Féjér operator σ_n . We have also looked at expansions in terms of Fourier

series. Tying together the potential connection between linear algebra and integration, one should note that these last two approximation operators are directly based upon integrals, whereas the first, though sharing the property of linearity, is not in any obvious way based upon integration. Finally, one should take note that the simplest of all bounded linear operators are the bounded linear functionals, and the more complicated operators are, generally speaking, developed by pasting together some linear functionals.

It should also be noted that one major motivation for development in mathematics has nothing to do with abstraction per se. It is the constant and ever-increasing need for more clever problem-solving techniques and computational techniques which arise in applied mathematics. A classic example of this motivation at work was the development of Fourier series. Fourier himself was an applied mathematician, rather shamelessly indifferent to theoretical questions of such things as convergence or divergence of an infinite series, or the validity of asserting that the integral of the sum of an infinite series of functions is the sum of the integrals of the functions involved. But we have already seen that with the Riemann integral there is an obvious counterexample. A new integral is needed, the Lebesgue integral, which will obviously not be discussed in this course. It should merely be said here that Fourier has kept the theoreticians busy all the way to the present, and even today not all the questions which his work exposed have been answered.

Now, let us summarize the previous discussion by providing a short list of questions which naturally arise, specifically relating to integration. Most of them will not be answered here, but all but one of them have already cropped up in previous chapters. None of the questions in this list have been or will be answered in this course. Rather these questions are important both in theory and in computational practice in applied mathematics. Many of them will take a central role in subsequent courses. There is also one question which will appear at the end of this list. It has not cropped up in

previous chapters, but rather it is the topic of the rest of this chapter, and it may well also be of importance in future courses, too. Here is the list:

- When it is possible validly to reverse the order of integration and summation?
- Can a better integral be developed, for which the pointwise limit of a sequence of integrable functions is also integrable? One should also note that the pointwise limit of a sequence of functions might not even be defined at every relevant x . When might such a situation affect the answer to this question, and when does it not?
- The space $C[a, b]$ is complete, if endowed with the usual norm on it, the uniform norm. Yet, the same space is not complete if one uses the 1-norm or the 2-norm. Completeness is a nice property, and sometimes we really do need to invoke it (for example, when trying to come up with reasonable answers to the previous two questions!). Can anything be done about this? If so, then what?

And, finally the last question, which points toward the next section of this chapter:

- The Riemann integral of a continuous function defined on a closed interval $[a, b]$ defines a bounded linear functional on $C[a, b]$. That is hardly a profound observation once one has learned what the words mean. Also, just from the definition of the integral it is obvious that many other linear functionals can be defined from this same basic integral, too. One might for example integrate all the functions on a certain subinterval of the given interval. Or

one might integrate all the functions after multiplying them by some fixed function $w(x)$ called a weight function. There are lots of other possibilities, too.

Thus, an interesting question arises:

Can **every** bounded linear functional defined on $C[a, b]$ be represented by some kind of integral?

The answer to the last question above is, of course, “no” unless one stretches the definition of the integral just a bit. One of the obvious reasons for the failure is, of course, the fact that the evaluation of a function at a single point can not be represented by the Riemann-Darboux integral, which was constructed in such a way as to obliterate what is happening just at one point. Nevertheless, the answer to the question is “yes” if one adopts a more general point of view about just what is an integral. For this reason and several other very good reasons, the further development of the theory of integration is one of the most important parts of the field of real analysis. The full development which is needed will certainly not be done here; the whole project is far too big for an introductory course.

Therefore, here we do no more than to define one of the natural extensions of the concept of the integral, and to look at some of the consequences of that new definition.

15.2 The Riemann-Stieltjes Integral

The definition of the Riemann-Stieltjes integral results from a small change in the definition of the Riemann integral, which is found here in Section 10.2.4. The definition depended upon the construction of Riemann sums, as defined in (10.7) and the text immediately preceding that statement.

To construct a **Riemann-Stieltjes sum**, one assumes that a partition P is given, consisting of the points $a = t_0 < \dots < t_n = b$, and we also assume that points x_1, \dots, x_n are chosen, correspondingly in each of the subintervals $[t_{i-1}, t_i]$ of the interval $[a, b]$ upon which it is intended to define the integral. Again, as in the previous definition of the Riemann integral, no other restrictions are put upon the points x_1, \dots, x_n . So far, it looks as though we were defining the Riemann integral again. The difference will be seen in what comes next.

Now, it is assumed that there are two functions, f , the **integrand**, and g , the **integrator**. The ultimate goal will be to integrate f with respect to g . The Riemann-Stieltjes sum associated with the given partition, the given points x_1, \dots, x_n , and the two functions f and g will be given by

$$\sum_{k=1}^n (g(t_k) - g(t_{k-1}))f(x_k) \tag{15.1}$$

Then, the Riemann-Stieltjes integral of f with respect to g is said to exist if any sequence of sums of the form (15.1) will always converge, and converge to the same value, provided that the mesh of the partitions used tends to zero, and

this is true no matter how the intermediate points x_1, \dots, x_n are chosen. For this integral, we use the notation

$$\int_a^b f dg$$

Note that it is also possible to develop sums similar to the previously used Darboux sums, of the form

$$U = \sum_{k=1}^n (g(t_k) - g(t_{k-1})) \sup_{x \in [t_{k-1}, t_k]} f(x) \quad (15.2)$$

and

$$L = \sum_{k=1}^n (g(t_k) - g(t_{k-1})) \inf_{x \in [t_{k-1}, t_k]} f(x). \quad (15.3)$$

Some of the basic properties of Riemann-Stieltjes integral are investigated in the following exercises. When doing them, please stick to the definition. Do not attempt to use properties of the Riemann-Stieltjes integral which you may have picked up somewhere through osmosis, or through wishful thinking, which are not given here unless also provide proofs of any such properties which you intend to use.

Exercises

15.2.1 If $g(x) = x$, then the Riemann-Stieltjes integral of f with respect to g on an interval $[a, b]$ coincides with the Riemann integral of f on $[a, b]$.

15.2.2 Show that the Riemann-Stieltjes integral with respect to a fixed function g is linear. That is, $\int (\alpha f_1 + \beta f_2) dg$ and $\int (cf) dg$ come out to what they are supposed to if the integral is linear. You are of course free to combine two equations defining linearity into one equation, as is done in the next exercise.

15.2.3 Is the Riemann-Stieltjes integral also linear in g with respect to a fixed function f ? That is, is it always true that $\int f d(\alpha g_1 + \beta g_2)$ be split into $\alpha \int f dg_1 + \beta \int f dg_2$? (Note that an equivalent definition of linearity is used here, which treats addition and scalar multiplication together in one equation. You are of course free to use instead two equations, one dealing with addition and the other dealing with scalar multiplication, as was done in the previous problem.)

15.2.4 If $\int_a^b f dg$ exists, and $a < c < b$ is it necessarily true that $\int_a^b f dg = \int_a^c f dg + \int_c^b f dg$? Prove, or give a counterexample.

15.2.5 If f and g' are both continuous, then $\int_a^b f dg = \int_a^b f(x)g'(x) dx$

15.2.6 Let $g(x) = 0$ for $0 \leq x < \frac{1}{2}$ and $g(x) = 1$ for $\frac{1}{2} \leq x \leq 1$

- (a) Assume that f is continuous on $[0, 1]$. What can you say about the value of $\int_0^1 f dg$? Carefully use the definition of the Riemann-Stieltjes integral to answer this question.
- (b) What, if anything, can you say about the value if f is bounded, but it is only assumed that f is continuous at $\frac{1}{2}$?